Using Crowdsourcing to Investigate Perception of Narrative Similarity

Dong Nguyen
University of Twente
Enschede
The Netherlands
d.nguyen@utwente.nl

Dolf Trieschnigg
University of Twente
Enschede
The Netherlands
d.trieschnigg@utwente.nl

Mariët Theune
University of Twente
Enschede
The Netherlands
m.theune@utwente.nl

ABSTRACT

For many applications measuring the similarity between documents is essential. However, little is known about how users perceive similarity between documents. This paper presents the first large-scale empirical study that investigates perception of narrative similarity using crowdsourcing. As a dataset we use a large collection of Dutch folk narratives. We study the perception of narrative similarity by both experts and non-experts by analyzing their similarity ratings and motivations for these ratings. While experts focus mostly on the plot, characters and themes of narratives, non-experts also pay attention to dimensions such as genre and style. Our results show that a more nuanced view is needed of narrative similarity than captured by story types, a concept used by scholars to group similar folk narratives. We also evaluate to what extent unsupervised and supervised models correspond with how humans perceive narrative similarity.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Narratives; folktales; similarity; crowdsourcing

1. INTRODUCTION

Measuring the similarity between documents is essential in many applications. For example, clustering systems are inherently dependent on the used similarity measure. However, for many tasks it is unclear what an appropriate similarity measure should be. Multiple dimensions might play a role (e.g. topic, genre), and different users might not agree on which dimensions are important. So far, most research on text similarity has focused on topical or semantic similarity, thereby ignoring dimensions that might be important from a user's perspective. Research investigating *how humans perceive similarity* between documents has been scarce so far.

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661918.

Understanding how humans perceive similarity is useful in many situations. It could guide the development of similarity metrics to correspond better with human perception. Clustering systems could benefit by knowing along which dimensions documents should be clustered. And, it could aid in the creation of more suitable datasets [2] and evaluation metrics [13].

In this paper, we study perception of similarity in the domain of folk narratives (such as fairy tales and urban legends). With the increasing digitization of folk narratives [1, 16, 19], there is a need for better search and clustering systems [12]. However, so far little is known about *how humans perceive narrative similarity.* For example, take the following two narratives (summaries are shown):

> **Narrative 1** Some men sat around a fire. Nine cats came to sit near the fire, and the men got nervous. One of the men threw fire at the cats with a fire shovel. The next day, nine women in the village lay in bed with burned buttocks.

> **Narrative 2** Every afternoon a large black cat came to sit by the fire in the kitchen. The people knew about a witch in the neighborhood. One afternoon the cat came again. The woman threw a pan with hot oil at the cat's neck. The next day, the neighbor wore a white scarf, she had burned her neck.

The characters in both narratives are witches, humans and cats. Although the exact events are different, both narratives share a story line: The cats are actually witches, who are recognized by their wounds in their human form. Some people might even recognize that these narratives served a common purpose: demonstrating that witches are real.

Folktale researchers could recognize these narratives as belonging to the same story type (titled '*Witch hurt as animal; woman turns out to be wounded the next day*', SINSAG 0640). A *story type* represents a collection of similar stories. Story types are used by scholars to organize folk narratives and defined in catalogues such as SINSAG and ATU. For example, a well-known story type is '*Little Red Riding Hood*' (ATU 333). The many variations of this story (e.g., with different endings) are classified with the same story type.

The above example illustrates that similarity between narratives can be based on various dimensions (e.g. characters, plot, theme/purpose, story types). The goal of our study is to shed light on how narrative similarity is perceived. Which dimensions do people consider when judging narrative similarity, and do non-experts pay attention to different dimensions than experts?

Empirical studies on narrative similarity have only been done on a small scale so far (e.g., [9, 15]). This study is the first large-scale empirical study on narrative similarity. We collect data from a large number of non-experts using crowdsourcing by asking them to rate similarity between narrative pairs. Data on how experts judge narrative similarity was collected in two ways: 1) By asking experts directly to rate similarity, in the same way as data obtained from non-experts. 2) By using the story types that the narratives are (manually) classified with.

To summarize, our contributions are as follows:

- We show how crowdsourcing can be used to collect data for studying perception of similarity (Section 4).
- We identify the dimensions that play a role in perception of narrative similarity (Section 5).
- We show that non-experts and experts have a different perception of narrative similarity and that story types do not fully correspond with non-expert perception of narrative similarity (Section 5).
- We show that automatic methods correspond reasonably well with judgements of the crowd (Section 6).

2. RELATED WORK

In this section we discuss related work on empirical studies of similarity, narrative similarity, and crowdsourcing.

Empirical studies of similarity. Our study follows recent research on the human perception of similarity in various domains, for example images [21], style of paintings [14], text [2], multimedia files [31], music [17, 29] and videos [4]. Some of these studies also investigated which dimensions play a role in human judgement of similarity, for example of multimedia files [31] and preference judgements of search result lists [13]. The influence of structure, style, and content on text similarity have been studied by Bär et al. [2]. Compared to these previous studies, we collect and compare judgements from both experts and non-experts.

Narrative similarity. Narratives have traditionally been studied by focusing on their plot structure. This is also reflected in research focusing on narrative similarity [20]. Scholars have typically approached this by developing formal systems to represent and find analogies between plot structures of narratives. The approaches rely on in-depth annotations of story structures by humans and as a result have stayed either theoretical [20] or have only been tested on small amounts of data (e.g. 1 narrative pair [8], or 26 Aesop fables [7]).

A different line of work involves a more computational approach but uses shallower features. For example, automatic classification of folk narratives [23] or jokes [10]. These methods focus on lexical similarity and do not study which dimensions play a role in perception of similarity. In addition, their ground truth labels provide only a binary view of similarity.

Two recent studies investigated perception of narrative similarity, but on a very small scale (16 narrative pairs [15], variations of two stories [9]). Their results have suggested that non-experts also focus on dimensions other than structural similarity [9], and that humans are more likely to rate narratives as similar if they have a common summary [15].

Crowdsourcing. Crowdsourcing enables the collection of large amounts of data with low costs using platforms such as Amazon Mechanical Turk and Crowdflower. We target Dutch workers in our study, who are fast and of high quality [24]. Recent studies have explored how the crowd can be used to infer taxonomies [6] and clusterings from data [11]. However, such approaches need judgements for each item. An alternative approach is to use the crowd to learn a similarity metric, which can then be applied on large, growing collections (e.g., [32]). Our study follows the latter line of thought, by aiming to obtain insight into perceived similarity and develop automatic methods to measure similarity.

3. BACKGROUND ON FOLKTALES

Dutch Folktale Database. In this study, we use narratives from the Dutch Folktale Database. The database contains over 40.000 folk narratives [19], collected through various methods, including fieldwork and from social media. All narratives have been manually annotated with metadata such as a summary, keywords, language, story type and named entities.

Genres. In this study, we confine ourselves to the most frequent genres in the Dutch Folktale Database [22]:

- *Fairy tales* are set in an unspecified time and place, with often a happy ending and magical elements.
- *Legends* are situated in a known place and recent past, with human characters but also supernatural elements such as witches.
- *Urban legends* take place in modern times and are claimed to have happened. They are often about hazardous or embarrassing situations.
- *Jokes* are stories told to entertain each other, frequently ending with a punch line.

Narratives belonging to the same story type do not necessarily occur in only one genre. For example, '*Little Red Riding Hood*' can be told as a fairy tale or as a joke.

Story types. Story types are used by scholars to categorize similar folk narratives. A story type represents a collection of similar stories often with recurring plot, motifs or themes [27]. For example, the story type '*Little Red Riding Hood*' (ATU 333, [26]) is about a young girl who visits her grandmother, but then is eaten by a wolf disguised as her grandmother. Variations of a story emerge as stories are retold in different cultures and by different narrators. For example, in some variations the girl manages to escape from the wolf, and in other variations the story is transformed into a joke.

Story types are defined in catalogues created by various scholars. For example, the SINSAG catalogue focuses specifically on legends. While story types have been useful to organize narratives, they also suffer from limitations [5]. They provide a simplified (binary) view of similarity and story types do not always group narratives on the same level of specificity. Some story types are very specific, grouping narratives that share a common plot (e.g. '*Little Red Riding Hood*'). Other story types group narratives that share a common structure (e.g. repetition), and the broadest category of story types only share a common theme (e.g. '*Anecdotes about Lawyers*').

4. DATA COLLECTION

In this section we describe the collection of the narrative similarity judgements.

4.1 Preprocessing

We selected a subset of the narratives from the Dutch Folktale Database. We restricted the set to narratives that were easily readable (based on writing style and length) and had all the required metadata to support our analyses. More specifically, we only kept narratives with the following requirements: 1) Written in Standard Dutch [28] 2) With an annotated story type, genre and collector 3) Of intermediate length (between 10 and 250 tokens).

4.2 Task Design

To collect data, we designed a human intelligence task (HIT). The task was given to both experts and non-experts. Data from non-experts was collected by posting the HITs on Crowdflower, a crowdsourcing platform. We asked workers to judge the similarity between pairs of stories. We provided as few instructions as possible (e.g. by *not* mentioning terms like plot), so that workers were not influenced by us to pay attention to certain dimensions. Small pilot experiments were carried out while developing the design of the HIT. Each HIT consisted of 6 pairs of narratives (5 pairs + 1 pair with gold labels), and several survey questions. Each HIT was initially judged by 3 workers. We collected additional judgements for narrative pairs with large standard deviations of the judgements, such that all HITs received 3 to 5 judgements. We paid 40 US dollar cents for each HIT.

Survey. To study the influence of characteristics of people on how they perceive narrative similarity, we included several survey questions:

- Gender (male/female)
- Age (in years)
- Location (one of the Dutch provinces, or other)
- Highest completed education[1]
- How often do you read a book?[2]
- What kind of books do you read?[3]
- How often do you watch a movie?[2]

Similarity judgements. Workers were presented with pairs of narratives for which they were asked to rate the similarity on a scale from 1 (no similarity) to 5 ((almost) the same). A similar scale was used in related studies [18, 33]. Workers were also asked to provide a short motivation for their rating in a free text field for each narrative pair. The order of the displayed narrative pairs within a HIT was randomized.

Gold labels. To improve the detection of spammers, we manually created 12 narrative pairs with 'gold labels'. Workers who provided ratings deviating from these labels were identified as potential spammers. We created pairs with high similarity by copying an existing story and making small edits in spelling, punctuation, word order, etc. For such pairs, we expected a similarity judgement of 4 or 5. We also selected pairs with very low similarity by manually selecting stories that had nothing in common (e.g. plots and characters are completely different). For such pairs, we expected a similarity judgement of 1 or 2. We did not inform workers about their performance on the pairs with gold labels.

[1]No education, Elementary school, Pre-vocational secondary education, Senior general secondary education/pre-university secondary education, Secondary vocational education, Higher professional education, University education
[2]Daily, several times a week, several times a month, never
[3]Fiction, non-fiction, both

4.3 Pair Selection

Selecting pairs at random would generate many pairs with little similarity. Therefore, we control the selection of pairs as follows:

1. Similarity between narratives classified with the *same* story type and *same* genre.
2. Similarity between narratives classified with the *same* story type but *different* genre.
3. Similarity between narratives with the *same* genre, but *different* story types.

Under conditions 1 and 2, the narratives are the same based on their story types. We include pairs by varying the lexical similarity of these pairs based on cosine similarity. A threshold (based on data analysis, see below) was calculated to distinguish between low, mid and high similarity. We include an equal number of pairs from each bin.

Under condition 3, we only include pairs that have a high cosine similarity. We assume that pairs with low or mid similarity are less interesting, since they have little lexical similarity and are also not similar based on their story types.

Thresholds. We first group all narratives by story type. For each story type, we randomly select a pair of narratives and calculate the cosine similarity. Based on the samples, we take their 33% and 67% boundaries to define the thresholds to distinguish between low, mid and high cosine similarity.

Same story type, different genre. We select pairs of narratives that are classified under the *same* story type but under different genres. We first generate candidate pairs:

```
For each story type:
    Group all narratives by genre
    If #genres > 1:
        Sample pairs across genre (up to 3 per bin)
```

The final selection is made by sampling from all the candidate pairs, given the desired distribution for the cosine similarity bins and the number of pairs to include.

Same story type, same genre. We study similarity between pairs belonging to the same genre and story type, but with varying levels of cosine similarity.

```
For each genre:
    For each story type:
        Sample up to 3 pairs per cosine bin
```

The final selection is made by sampling from the candidate pairs ensuring an equal distribution across cosine similarity bins, given a desired genre distribution and the total number of pairs needed.

Same genre, different story types. We also select pairs belonging to different story types but with a high cosine similarity. We create candidate pairs as follows:

```
For each genre:
    For each story type:
        Select up to 3 pairs with high cosine
        similarity and with one of the
        narratives belonging to this story type.
```

The final selection is made by sampling from the candidate pairs given a desired genre distribution.

4.4 Groups

The designed HITs were given to two different groups: crowdworkers and folk narrative researchers.

Crowdworkers. We posted the tasks on CrowdFlower and targeted workers from the Netherlands. The jobs ran between April 4, 2014 and April 27, 2014. We launched the jobs in several batches, to prevent workers from doing the task too many times and to ban spammers in between. Potential spammers were identified by the following criteria:

- Inconsistent demographics. Most workers completed multiple HITs. We assumed workers with inconsistent demographics information to be spammers.
- Time spent on judgement. Workers who spent less than 3 minutes on a HIT (based on data analysis).
- Gold labels. Workers whose judgements did not match the gold labels.
- Motivation. We manually inspected the answers on the motivation questions. Spammers answered with random characters, by copying parts of the narratives or by always answering with the same sentence.

We manually checked if workers identified using these criteria were spammers. Such workers were excluded from the dataset and blocked for all next HITs. We collected in total 923 HITs (150 workers). 619 HITs (80 workers) were kept after filtering spammers. Figure 1 shows the average times spent on a HIT for workers (median: 677.5 seconds).

Folktale Researchers. We also asked three senior folktale researchers (all with a researcher/lecturer position) to do the same task. We selected 40 narrative pairs, ensuring that we included at least 2 pairs from each bin according to our sampling method described above. HITs were the same as presented to the crowdworkers, but without the pairs with gold labels (thus resulting in 5 narrative pairs per HIT).

4.5 Statistics

The statistics of the collected data are shown in Table 1.

Statistic	Crowdworkers	Experts
# unique narrative pairs	1002	40
# completed HITs	619	24
# persons	80	3

Table 1: Dataset statistics

5. ANALYSIS

In this section, we analyze the collected data. We start with studying the demographics of the workers and then continue with an analysis of their similarity judgements.

5.1 Workers

Demographics. Workers are mostly men (66%), but are relatively spread across different ages and education levels. The workers are spread throughout the Netherlands, but most workers come from the west of the Netherlands (where the population density is higher as well).

Reading and Watching Movies. Table 2 summarizes the users' reading and movie-watching behaviour. Most people read both fiction and non-fiction (52, 65%), and some read only fiction (20, 25%). A small fraction only reads non-fiction (8, 10%). We code the education responses, and movies and reading behaviour by converting each category to an integer. We find that the education level is highly correlated with frequencies of reading a book (Spearman's $\rho = .424$, $p < 0.001$). The education level is negatively correlated with the frequency of watching movies (Spearman's $\rho = -.229$, $p < 0.05$). Watching movies and reading books is not correlated (Spearman's $\rho = -.086$, not significant).

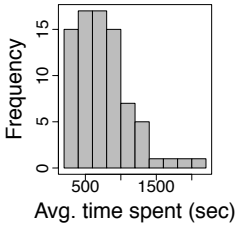


Figure 1: Average time spent on task

How often	B	M
Never	11	0
Couple of times a month	44	39
Multiple times a week	18	37
Daily	7	4

Table 2: Frequencies of reading books (B) and watching movies (M)

5.2 Understandability Ratings

Workers also indicated how well they understood the pair of narratives on a scale from 1 (not understandable) to 5 (well understandable) (Figure 2). Manual inspection of pairs with lower ratings, revealed that crowdworkers had difficulty with language use that was less standard (e.g., dialects, slang, uncommon words), unconventional style and structure. Narratives from more modern genres, urban legends and jokes, are understood better than narratives from the older genres, legends and fairy tales (Table 3).

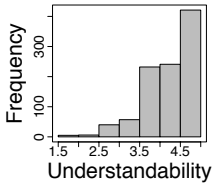


Figure 2: Understandability ratings

Genre	U
Urban legends	4.47
Jokes	4.33
Legends	4.12
Fairy tales	4.06

Table 3: Understandability (U) and genre

For each worker, we calculate the worker's understandability bias, by calculating the average difference between the worker's score and the average of the scores. We find that higher educated workers tend to give *lower* understandability scores (Spearman's ρ = -.249, p < 0.05). While this may seem counterintuitive, they also vary more in their understandability ratings (Spearman's ρ = .278, p < 0.05). We found no significant correlations with reading or movie watching behaviour. In the remainder of this paper, we only keep narrative pairs that received an average understandability rating of 3.5 or higher (removing 104 pairs).

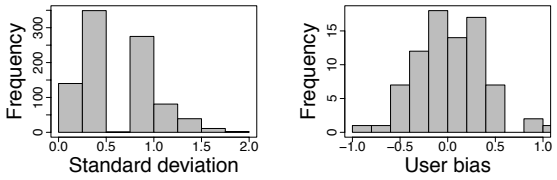
5.3 Narrative Similarity Ratings

We first analyze the agreement between the judgements. Next, we study the similarity judgements for different conditions. Finally, we study the similarity dimensions by analyzing the free-text motivations.

5.3.1 Agreement

Crowd Workers. We first analyze the agreement between crowdworkers. We have in total 80 workers and 898 pairs. For each pair, we have 3 to 5 judgements.

Figure 3 shows a histogram of the standard deviations of the judgements for each narrative pair. We find that 85% of the pairs have a standard deviation less than 1. We also calculate a user bias. For each worker, it is the average over the differences between a judgement made by a worker and the overall mean of the judgements for each narrative pair. Only 17.5% of the workers are on average more than 0.5 points off the mean.



(a) Std per narrative (b) User bias

Figure 3: Similarity judgements crowd

We calculate several agreement metrics (see Table 5). Following [18], we calculate a measure of inter-rater correlation. For each narrative pair, we select at random a judgement and correlate it with the average of the other similarity judgements. We also calculate a pairwise agreement. For each narrative pair, we check the agreement between all pairs of workers. The reported pairwise agreement is the number of pairs workers agreed on divided by the total number of pairs, and is similar to the value reported in [33]. We also find that mapping the scores to a lower number of categories (1-2, 3, 4-5) leads to a higher pairwise agreement of 0.517.

Metric	Crowd	Experts
Spearman correlation	.556	.778
Pearson correlation	.572	.796
Pairwise agreement	.335	.423

Table 5: Agreement crowd and experts

We also analyzed the influence of demographics on agreement. We find that people who read books daily (3) or never (0) tend to agree more within their group (pairwise agreement of 0.433 and 0.444, see Table 6). We also tested excluding groups with lower reading frequencies. Only including workers who often read ($\geq$ 2) leads to higher agreement (0.374) than including all workers. No clear trends were observed with watching movies or education.

	Reading frequency			
Criteria	0	1	2	3
$\geq$	.335	0.333	0.374	0.444
=	.433	0.310	0.340	0.444

Table 6: Pairwise agreement and reading frequency

Experts. We calculated the agreement between experts in the same way as with the crowd. Table 5 shows the calculated metrics. We find that experts achieve higher inter-annotator agreement than the crowd, probably because their reasoning involves story types and they agree more on which dimensions are important (see also the next section). We also study to what extent individual experts and the average of the experts correspond with the crowd judgements (Table 7). Averaging the expert judgements leads to a higher correlation with the crowd judgements.

	E1	E2	E3	Avg. expert
Crowd	.654	.683	.633	.744

Table 7: Spearman correlation of individual experts (E1-3) and average expert with the crowd

5.3.2 Analysis

The average similarity for each condition (see Section 4.3) is shown in Table 4.

Story types. We first investigate how story types correspond with judgements by the crowd. We would expect narratives belonging to the same story type to receive higher ratings than narratives belonging to different story types.

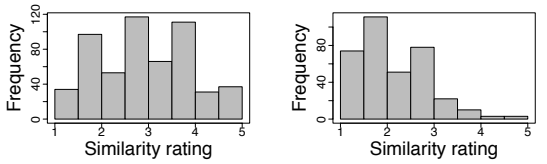
Narrative pairs (with high cosine similarity) with the same story type indeed receive higher ratings than pairs (with high cosine similarity) with different story types (Table 4).

Figure 4 shows a histogram of the similarity ratings for narrative pairs belonging to the *same* story type. If story types would correspond strongly with perceived similarity by non-experts, we would see a skewed distribution with most of the ratings being a 4 or 5. Instead, most of the ratings are in the middle and the perceived similarities of the narratives belonging to the same story type vary widely. Figure 4 also shows a histogram of the similarity ratings for narrative pairs belonging to *different* story types. Here, we do see a skewed distribution, with most scores being low (e.g. 1 or 2).

Thus, although narratives belonging to the same story type tend to be perceived as more similar than narratives belonging to different story types, story types do not explain all of the observed variation in similarity judgements by non-experts. This suggests that story types ignore dimensions that non-experts do find important.

	Urban legends		Jokes		Legends		Fairy tales		All	
Same story type, same genre										
Low cosine	2.900	(0.109)	2.119	(0.160)	2.503	(0.133)	2.343	(0.191)	2.501	(0.077)
Mid cosine	3.375	(0.134)	2.743	(0.139)	2.793	(0.112)	3.150	(0.268)	3.008	(0.078)
High cosine	3.972	(0.089)	3.550	(0.172)	3.536	(0.173)	3.806	(0.194)	3.719	(0.078)
Different story type, same genre										
High cosine	2.095	(0.072)	2.174	(0.070)	2.346	(0.092)	2.106	(0.119)	2.181	(0.042)
Same story type, different genre										
Low cosine			2.226 (0.094)							
Mid cosine			2.721 (0.110)							
High cosine			3.504 (0.121)							

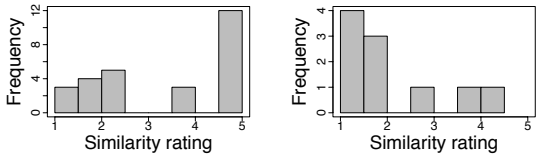
Table 4: Mean and standard errors of similarity scores per condition.



(a) Same story type (b) Different story type

Figure 4: Similarity judgements - crowd

Figure 5 shows figures based on expert judgements. The figures reflect that experts use story types in their research and give more extreme scores than non-experts. Pairs of narratives belonging to the same story type are mostly rated with a 5, narratives belonging to different story types often with a 1 or 2. However, we do observe variation indicating that other aspects influence their judgements as well. For example, based on their feedback, we find that experts tend to rate pairs of narratives from broad story types (e.g., based on theme) lower than story types defined based on plots.



(a) Same story type (b) Different story type

Figure 5: Similarity judgements - experts

Genres. Table 4 also shows the scores for narratives with the same story type but classified under *different genres*. We find that narrative pairs with different genres tend to receive a lower similarity judgement than pairs belonging to the same genre. In the next section we study the influence of genre using the provided free-text motivations.

Cosine similarity. In Table 4 we observe that within each genre, a higher cosine bin results in a higher average similarity judgement. In a later section, we experiment how the similarity judgements correspond with various supervised and unsupervised similarity metrics.

5.4 Dimensions of Narrative Similarity

For each similarity judgement, we also asked for a free-text answer with a motivation. In this section, we study the importance of different similarity dimensions (e.g., plot, characters) based on these motivations.

Crowdworkers. Motivations given for the narrative pair in the introduction are shown in Table 9. The first worker only mentions a similarity in the characters, the other workers also see a similarity in the plot. Note that other dimensions could have (unconsciously) influenced the workers as well, but they did not mention them.

Not much except they are about a cat *Given score: 2. Dimensions: Characters*
Both narratives are about witches and black cats. Furthermore in both stories the cat gets injured and as a result the woman is also injured. The narratives look very much like each other, but the content differs. Therefore I give it 4 out of 5. *Given score: 4. Dimensions: Characters, Plot*
easy to read, both narratives are about cats who are actually witches who sit at the fire and are thrashed there *Given score: 4. Dimensions: Style, Characters, Plot*

Table 9: Translated motivations by crowdworkers

We randomly selected 192 narrative pairs and included all motivations for these pairs (total: 589). The most frequent dimensions were identified after annotating subsets of the data. Each motivation was then manually annotated (Table 8) by one coder. A second coder annotated a subset of 64 narratives. Cohen's κ ranged from moderate (e.g., plot: $\kappa = 0.59$, style: $\kappa = 0.66$, theme: $\kappa = 0.59$) to high (e.g., genre: $\kappa = 0.80$, characters: $\kappa = 0.88$, number of details: $\kappa = 1.00$).

The characters, plot, genre and theme were mentioned the most. However, a variety of other dimensions were mentioned as well (e.g., style, number of details). No explanation was given in 18% of the motivations.

For each dimension (except 'none' and 'other'), we annotated whether a *difference* and/or *similarity* was mentioned, e.g. 'plots are different' (Table 8). When workers mentioned the characters, plot or theme, they tended to focus on the similarities between the narratives. However, when they referred to the amount of detail, workers only stated differences.

Dimension	Description	Crowd				Experts			
		M	Sim	Diff	P	M	Sim	Diff	P
Characters	The characters or important objects in a narrative (e.g., a princess, a ring)	.43	.88	.17	.80	.51	.74	.44	1.00
Plot	The sequence of events in a narrative	.37	.67	.49	.76	.54	.53	.62	1.00
Genre	For example '*both narratives are jokes*'	.21	.82	.18	.58	.14	.69	.31	1.00
Theme	The central topic / moral (e.g. paranormal events)	.28	.86	.15	.71	.36	.88	.13	1.00
Setting	Where the story is set. This can be more general (e.g. a castle) or a geographic location (e.g. Paris)	.04	.39	.61	.20	.01	1.00	.00	.33
Style	E.g. punctuation, word choice, formal language	.08	.36	.64	.39	.03	.67	.33	.67
Number of details	Length or number of details	.02	.00	1.00	.10	.05	.00	1.00	1.00
Recount facts	E.g., '*narrative 1 could be true*'	.01	.63	.63	.08	.00	-	-	.00
Structure	E.g. repetition of events	.03	.60	.47	.16	.08	.56	.44	1.00
Story types	E.g., '*both are of the same story type*'	.00	-	-	.00	.46	.59	.43	1.00
Motifs	Elementary building blocks of narratives	.00	-	-	.00	.06	.43	.71	.67
Other	All remaining dimensions, such as the narrator, origin of the stories, etc.	.03	-	-	.23	.05	-	-	.67
None	E.g., '*they are not the same*'	.18	-	-	.51	.13	-	-	.67

Table 8: Dimensions of narrative similarity. For each group (crowd: 80 persons, 589 motivations, experts: 3 persons, 111 motivations), the table reports the fraction of motivations (M) or persons (P) mentioning a dimension, and for each dimension, the fractions that mentioned similarities (Sim) or differences (Diff).

	B	SE
Intercept	2.47***	0.12
Characters.sim	0.03	0.11
Characters.diff	0.04	0.18
Plot.sim	0.99***	0.12
Plot.diff	-0.44***	0.12
Genre.sim	0.27*	0.14
Genre.diff	-0.57*	0.27
Theme.sim	0.21·	0.12
Theme.diff	-0.63**	0.23
Settings.sim	0.46	0.36
Settings.diff	0.14	0.29
Style.sim	0.17	0.29
Style.diff	0.79***	0.21
Num details.diff	1.19**	0.36
Recount facts.sim	0.42	0.49
Recount facts.diff	-0.24	0.49
Structure.sim	-0.31	0.38
Structure.diff	-0.03	0.41
None	-0.90***	0.16
Adjusted $R^2 = 0.292$		

Table 10: OLS model (weights and standard errors). *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; · $p < 0.1$

For most dimensions, whether they are mentioned is influenced by the presented narrative pair. For example, the probability of a random motivation mentioning theme is 0.28. However, knowing that another worker has mentioned theme for the same pair, the probability goes up to 0.51.

To study the importance of these dimensions, we fitted an Ordinary Least Squares model (OLS) with the given score as the dependent variable (Table 10). We find that plot, genre and theme are the most important. Characters are not significant after including the other dimensions. Maybe surprisingly, mentioning differences between style and number of details receives a *positive* weight. From manual inspection, we find that when narratives are already very similar on other dimensions, workers tend to mention these more superficial differences.

We also analyzed the correlation between characteristics of workers (education, frequency of watching movies/reading books). For most of the dimensions, we did not observe a relation with the characteristics of workers. People who read more books more often mention the theme of a narrative (ρ = .223, $p < 0.05$). We also found that people who watch more movies more often pay attention to whether narratives differ in number of details or length ($\rho = .210$, $p < 0.1$).

Experts. Motivations given by the experts for the narrative pair in the introduction are shown in Table 11. Statistics based on manual annotation are shown in Table 8. The dimensions 'story types' and 'motifs' are used in folk narrative research. Motifs are small elementary building blocks of plots of narratives (e.g., 'disease caused by witchcraft'). As expected, motifs and story types were only mentioned by the experts. Story types were mentioned in many of the motivations (46%).

Both are the same: the narratives must demonstrate that witches are real. *Given score: 5. Dimensions: Theme, Characters*
Strong similarity in content, I doubt between box 4 and 5: 1 and 2 share the traditional element of a witch changing into a cat, getting hurt, and being recognized in her human form through the wound. Given score: 4. Dimensions: Characters, Plot
Clearly two narratives of the same type: Hexentier verwundet: Frau zeigt am folgenden Tag Malzeichen. Whether it is with multiple cats, or one, it doesn't matter. Moral: night cats are metamorphosed witches, and you don't want them near you. *Given score: 5. Dimensions: Story type, Theme, Characters*

Table 11: Translated motivations by experts

Other dimensions important to experts are the plot, characters and theme of the narratives. Style, whether true facts are recounted, and setting are not important to experts.

6. ESTIMATING NARRATIVE SIMILARITY

In this section we present preliminary experiments on how well unsupervised and supervised methods correspond with the crowd judgements. Studies on document similarity in other domains found low to moderate correlations between automatic measures and human judgements. For example, a correlation of less than 0.2 was observed using cosine similarity [33] and between 0.5-0.6 using different binary, count-based and LSA-based measures [18]. To our knowledge, we are the first to perform such experiments on narrative similarity.

6.1 Goal and Evaluation

For each narrative pair, we take the mean of the received similarity judgements by the crowdworkers. We experiment with two different setups: 1) Classification, where the goal is to classify the pairs into *low* (<= 3) and *high* (>3) similarity. The performance is reported using the F-score. 2) Regression, where the goal is to predict the mean of the received judgements. We evaluate the performance using the Spearman correlation and Mean Squared Error (MSE).

6.2 Dataset Construction

We randomly divided the dataset into a training and test set. Feature development and parameter tuning was done using cross-validation on the training set. Like in the previous sections, we excluded the narrative pairs that received a low score for understandability. Statistics of the dataset are shown in Table 12. The documents were parsed using the Frog parser [30] and a stop word list of 76 frequent Dutch words was used.

Set	# Pairs	Mean	Low	High
Train	498	2.674	344 (69.08%)	154 (30.92%)
Test	400	2.683	271 (67.75%)	129 (32.25%)

Table 12: Statistics dataset

6.3 Method

We experiment with both unsupervised similarity metrics (e.g., cosine similarity) and supervised machine learning models. We use linear regression and logistic regression with Ridge (L2) regularization to prevent overfitting.

6.4 Features

We evaluate a variety of features, most of them based on the dimensions we identified in the previous section. In addition, we explore features based on manually annotated metadata.

First, we study the effectiveness of features that only measure lexical similarity. We experiment with different metrics (cosine similarity and Jaccard index) and representations (e.g., words versus character ngrams).

We also extract features from the narratives to approximate elements such as the plot, characters and theme in narratives. Plot elements are approximated by extracting subject-verb pairs. They are extracted by searching on subject (*'su'*) and verb complement (*'vc'*) relations from the Frog parser. Each 'plot element' is a character + root of a verb (e.g., 'lawyer_answer' or 'girl_disappear').

We extract the characters of a narrative by searching on subject (*'su'*) relations from the Frog parser. Only tokens classified as nouns, pronouns, or as 'special' are included. Unfortunately, the narratives are noisy because they come from a variety of sources, and therefore the Frog parser sometimes missed relations or incorrectly extracted them.

Themes are extracted using LDA [3]. We train a model on the training documents with 20 topics using the Gensim library [25]. We measure the similarity between the topic distributions using the Jensen-Shannon divergence.

Crowdworkers also pay attention to *style*. We therefore experiment with features that capture stylistic similarities based on statistics such as the length of words and sentences, and similarities in POS structures.

Our analyses also revealed that differences in *the amount of detail* in narratives play a role. We use the difference in length of the narratives to approximate this dimension.

We also study the usefulness of *manually annotated metadata*. They also capture dimensions identified in the previous section, such as whether the narratives have the same genre (mentioned by the crowd and experts), or story type (only mentioned by experts). In addition, we study whether manually annotated keywords and named entities are useful.

Below is an overview of the used features:

Lexical

1. Cosine similarity
2. Jaccard index

Story Elements

3. Plot
4. Theme (LDA)
5. Characters

Stylistic

6. Absolute difference between average word length
7. Absolute difference between average sentence length
8. 1-3 ngram POS patterns (Jaccard)

Other

9. Absolute length difference

Metadata (manual annotation)

10. Same story type (boolean)
11. Keywords (Jaccard)
12. Same genre (boolean)
13. Named Entities (Jaccard)

6.5 Results

We first study the individual features. Next, we study the performance achieved by combining them.

Individual features. We first evaluate the individual features in the regression setup. We report the Spearman correlations and MSEs (Table 13).

For the lexical features, we experimented with using the cosine similarity and Jaccard index. We also experimented with using word unigrams, word unigrams + bigrams, or character ngrams (of lengths 2-5). We find that using n-grams consistently achieves a better performance. In addition, the Jaccard index performs better than the cosine similarity.

We find that the stylistic features (POS patterns, word and sentence length) only obtain a low correlation. The features that aim to capture the story elements (e.g., theme) perform moderately. The features based on manually annotated metadata perform well, in particular the features based on story types and keywords.

Metric	ρ	MSE
Lexical		
Cosine - Unigrams	0.182	0.925
Jaccard - Unigrams	0.374	0.816
Cosine - Bigrams	0.206	0.918
Jaccard - Bigrams	0.383	0.865
Jaccard - Ngrams	0.418	0.817
Cosine - Ngrams	0.357	0.813
Story Elements		
Theme (LDA)	0.122	0.968
Characters	0.155	0.953
Plot	0.168	0.937
Stylistic		
Difference word length	0.076	0.981
Difference sentence length	0.073	0.980
POS ngrams	0.121	0.950
Other		
Length difference	0.079	0.975
Metadata		
Story type	0.336	0.873
Keywords	0.481	0.797
Genre	0.142	0.984
Named entities	0.184	0.946

Table 13: Individual features

Combination of features. We now combine the features using supervised machine learning models. We evaluate them in regression and classification tasks (Table 14).

Metric	ρ	MSE	F-score
Categories			
Lexical	0.431	0.759	0.590
Story elements	0.181	0.922	0.455
Stylistic	0.124	0.949	0.408
Metadata	0.494	0.746	0.614
Lexical + Category			
Lexical + story elements	0.435	0.761	0.590
Lexical + stylistic	0.491	0.715	0.611
Lexical + metadata	0.569	0.614	0.652
All			
Automatic (Lexical + story elem. + stylistic + other)	0.494	0.715	0.600
Automatic + metadata (Lexical + story elem. + stylistic + other + metadata)	**0.592**	**0.598**	**0.657**

Table 14: Feature combinations

We find that a reasonable performance is obtained using only the lexical features. Although the story elements features alone (plot, characters, theme) obtained a moderate performance, they do not help improve on the performance using the lexical features. We suspect this has several reasons. First, the story elements features are directly derived from the text as well and therefore highly correlated with the lexical features. For example, we find a Spearman correlation of .468 between the characters feature and the Jaccard n-grams feature. In addition, manual inspection shows that the extracted story elements are noisy, and thus the extraction of the features itself can be improved.

The metadata alone are already very effective. However, one should keep in mind that for new narratives no metadata will be available.

Using only lexical + stylistic features a good performance is achieved. Adding the remaining features does not lead to improvements. However, the best performance is obtained using both the automatically extracted features and the metadata. While the obtained correlation is moderate (.592), we should keep in mind that it is a difficult task. For example, when we randomly selected a judgement for each narrative pair and correlated that with the average of the remaining judgements, a Spearman correlation of .556 was obtained (see the section on agreement analysis).

7. DISCUSSION AND IMPLICATIONS

We analyzed the relationship between story types and human perception of similarity. While most narrative pairs from different story types are indeed perceived as not similar, within a story type there may be much variation. Dimensions such as genre and style that do not play a role in the definition of story types, do play a role in perception of similarity. This suggests that a more nuanced view of narrative similarity is desired.

Our results highlighted that non-experts and experts differ in how they judge narrative similarity. Therefore, how similarity between narratives is estimated should depend on the intended users and goal of the application.

We also found that non-experts vary in which dimensions they consider. Therefore, efforts to personalize systems that deal with narrative similarity could be an interesting direction of research. In addition, to help users understand the output of an automatic system, explicit explanations of how narratives are related would be useful as well.

Our study has limitations. First, free-text motivations were used to study the importance of dimensions. Users only mentioned dimensions they considered relevant, but (unconsciously) they may have also been influenced by other dimensions. Second, the mentioned dimensions and provided ratings may also have been influenced by the previous pairs a user has seen. We randomized pairs within a HIT to reduce possible effects of displaying order. However, further research is needed to study the influence of sampling and displaying order on the user judgements. Third, to enable a large-scale experiment, we included a large number of narratives from the Dutch Folktale Database. While we posed several restrictions to the final set to improve readability and also asked workers to indicate whether they understood the narratives, unclear or noisy narratives may have led to noise in the obtained judgements and mistakes in the automatic extraction of the features in the prediction experiments. Fourth, our experiments were performed on one specific dataset. Although we expect that our experimental setup can be used in other domains as well, other datasets (for example, movie reviews) should be used to verify this.

8. CONCLUSION

This paper presented a study on how humans perceive narrative similarity. A better understanding of narrative similarity is a first step towards better clustering and retrieval systems dealing with narrative collections. Data was collected by asking crowdworkers and folktale experts to rate the similarity between narrative pairs. We analyzed the provided similarity scores as well as their provided motivations. Our results showed that non-experts pay attention to more dimensions than experts, and that story types only give a limited view of narrative similarity.

Many of the identified dimensions can currently only be approximated in a shallow way using automatic methods. Further work is needed on automatically extracting dimensions such as style, structure, plot etc. of narratives to improve the automatic estimation of narrative similarity. Based on the findings in this paper, we plan to develop better clustering systems for narratives.

While this paper focused on a particular domain (narratives), we expect that the setup of the experiment and the types of data analyses performed can also be used to shed light on how similarity is perceived in other domains.

Acknowledgements

This research was supported by the Folktales as Classifiable Texts (FACT) project, part of the CATCH programme funded by the Netherlands Organisation for Scientific Research (NWO). We would also like to thank the folktale experts for participating in the experiment.

9. REFERENCES

[1] J. Abello, P. Broadwell, and T. R. Tangherlini. Computational folkloristics. *Communications of the ACM*, 55(7):60–70, July 2012.

[2] D. Bär, T. Zesch, and I. Gurevych. A reflective view on text similarity. In *Proceedings of RANLP 2011*, 2011.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] M. Cherubini, R. de Oliveira, and N. Oliver. Understanding near-duplicate videos: a user-centric approach. In *Proceedings of ACM Multimedia*, 2009.

[5] A. Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, 34(3):195–202, 1997.

[6] K. Eckert, M. Niepert, C. Niemann, C. Buckner, C. Allen, and H. Stuckenschmidt. Crowdsourcing the assembly of concept hierarchies. In *Proceedings of JCDL 2010*, 2010.

[7] D. K. Elson. Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, 2012.

[8] M. Fay. Story comparison via simultaneous matching and alignment. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, 2012.

[9] B. Fisseni and B. Löwe. Which dimensions of narrative are relevant for human judgments of story equivalence? In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, 2012.

[10] L. Friedland and J. Allan. Joke retrieval: recognizing the same joke told differently. In *Proceedings of CIKM 2008*, 2008.

[11] R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Proceedings of NIPS 2011*, 2011.

[12] R. Grundkiewicz and F. Gralinski. How to distinguish a kidney theft from a death car? Experiments in clustering urban-legend texts. In *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition*, 2011.

[13] J. Kim, G. Kazai, and I. Zitouni. Relevance dimensions in preference-based IR evaluation. In *Proceedings of SIGIR 2013*, 2013.

[14] A. Kovashka and M. Lease. Human and machine detection of stylistic similarity in art. In *Proceedings of CrowdConf 2010*, 2010.

[15] E. Kypridemou and L. Michael. Narrative similarity as common summary. In *Proceedings of the Workshop on Computational Models of Narrative 2013*, 2013.

[16] K. A. La Barre and C. L. Tilley. The elusive tale: leveraging the study of information seeking and knowledge organization to improve access to and discovery of folktales. *Journal of the American Society for Information Science and Technology*, 63(4):687–701, 2012.

[17] J. H. Lee. Crowdsourcing music similarity judgments using mechanical turk. In *Proceedings of ISMIR 2010*, 2010.

[18] M. D. Lee, B. Pincombe, and M. B. Welsh. An empirical evaluation of models of text document similarity. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2005.

[19] T. Meder. From a Dutch Folktale Database towards an International Folktale Database. *Fabula*, 51(1-2):6–22, 2010.

[20] L. Michael. Similarity of narratives. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, 2012.

[21] G. Nathalie, L. B. Hervé, H. Jeanny, and G.-D. Anne. Towards the introduction of human perception in a natural scene classification system. In *NNSP 2002*, 2002.

[22] D. Nguyen, D. Trieschnigg, T. Meder, and M. Theune. Automatic classification of folk narrative genres. In *Proceedings of the Workshop on Language Technology for Historical Text(s) at KONVENS 2012*, 2012.

[23] D. Nguyen, D. Trieschnigg, and M. Theune. Folktale classification using learning to rank. In *Proceedings of ECIR 2013*, 2013.

[24] E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(Feb):79–92, 2014.

[25] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *New Challenges for NLP Frameworks*, 2010.

[26] J. J. Tehrani. The phylogeny of Little Red Riding Hood. *PloS one*, 8(11):e78871, 2013.

[27] S. Thompson. *The folktale*. Dryden Press, 1951.

[28] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder. An exploration of language identification techniques for the Dutch folktale database. In *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage, LREC 2012*, 2012.

[29] J. Urbano, J. Morato, M. Marrero, and D. Martín. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR workshop on crowdsourcing for search evaluation*, 2010.

[30] A. van den Bosch, B. Busser, S. Canisius, and W. Daelemans. An efficient memory-based morphosyntactic tagger and parser for Dutch. *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, 2007.

[31] R. Vliegendhart, M. Larson, and J. A. Pouwelse. Discovering user perceptions of semantic similarity in near-duplicate multimedia files. In *First International Workshop on Crowdsourcing Web Search*, 2012.

[32] J. Yi, R. Jin, A. K. Jain, S. Jain, and T. Yang. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *Proceedings of NIPS 2012*, 2012.

[33] M. Zengin and B. Carterette. User judgements of document similarity. *Proceedings of the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)*, 2013.