Folktale similarity based on ontological abstraction

Marijn Schraagen
Digital Humanities Lab, Utrecht University, The Netherlands
M.P.Schraagen@uu.nl

Abstract
This paper presents a method to compute similarity of folktales based on conceptual overlap at various levels of abstraction as defined in Dutch WordNet. The method is applied on a corpus of Dutch folktales and evaluated using a comparison to traditional folktale similarity analysis based on the Aarne–Thompson–Uther (ATU) classification system. Document similarity computed by the presented method is in agreement with traditional analysis for a certain amount of folktale pairs, but differs for other pairs. However, it can be argued that the current approach computes an alternative, data-driven type of similarity. Using WordNet instead of a domain-specific ontology or classification system ensures applicability of the method outside of the folktale domain.

1 Introduction
A folktale is a specific type of narrative that is particularly suitable for analysis of semantic structure. Although folktales may differ in various aspects, such as the characteristics of the main actors or the sequence of events, often similarities can be identified on a more general or more abstract level. In this paper similarity between folktales is investigated using an explicit abstraction of text according to the WordNet concept hierarchy. A comparison is provided to conventional folktale motif analysis. An example of folktale similarity on various levels of abstraction is provided by the folktales Sleeping Beauty and Snow White, which both feature a princess as specific character and a variable number of enchanted objects at a more general level. Similarities regarding events occur at various levels as well, for example the princess in Snow White is asked by the seven dwarfs to perform household tasks, whereas the girl protagonist from Hansel & Gretel is ordered by the witch to do housework. In this case both the actors and the actions are similar at various levels, depicted in Figure 1. This notion of abstraction-based semantic similarity can be computed automatically using a machine-readable concept hierarchy such as WordNet. The paper is structured as follows: Section 2 describes characteristics of folktales and provides an overview of the resources used in the analysis, Section 3 discusses related work, Section 4 provides details on similarity computation, Section 5 contains experimental results and a comparison to existing folktale analysis approaches, and Section 6 concludes.

2 Folktale similarity
The folktale texts used in the current research are extracted from the Dutch Folktale Database (Meder, 2010). This collection contains over 40,000 folktales (including jokes, urban legends, etc.) from written and oral sources, in Dutch, Frisian and several contemporary and historical Dutch dialects. The database is partially annotated using Aarne-Thompson-Uther (ATU) codes (see the remainder of this section for details). The database, maintained by the Meertens Institute, is available for research purposes upon request. For the current research a pilot set of 16 traditional fairy tales is used from a single original source (van Dongen and Grooten, 2009), with a total of

http://www.verhalenbank.nl, in Dutch

352
33,022 words. This set provides folktales in grammatically correct, modern Dutch, which increases applicability of natural language processing tools and methods. Several folktales in this set do not appear in the ATU catalog, illustrating the applicability of the current method on non-traditional folktale sources.

In the current research the Dutch Cornetto database is used (Vossen et al., 2013) to obtain term abstractions. Cornetto is modeled after the Princeton WordNet, which is a widely used ontology for English concepts (Fellbaum, 1998) containing a comprehensive set of terms and (hierarchical) relations for an extensive variety of domains. Concepts are organized in sets of (approximate) synonyms, called synsets, which are connected by relations such as hypernymy and meronymy. Cornetto contains over 92,000 lemmas and is available under academic license.

Traditionally, folktales are analyzed using the Thompson Motif Index (TMI). This index is a set of over 12,000 story elements (motifs), classified in semantic categories and subcategories (Thompson, 1960). Some examples are provided in Table 1. The motifs in this index are often specific to a single folktale (or folktale type, i.e., the set of variants of a story that are considered the same folktale), however more general motifs are used as well. The Aarne–Thompson–Uther (ATU) classification system (Uther, 2004) describes a folktale type as a list of motifs (typically two or three to about 20) from a subset of nearly 1,900 elements from the TMI, divided into thematic categories and subcategories. The ATU classification is centered around the type of protagonists and the general theme of the folktale, while the TMI is centered around events and relations. This may introduce semantic relatedness differences between the two systems, for example the classification of ATU 123 as Animal tales–Wild animals and domestic animals compared to ATU 333 which is classified as Fairy tale–Supernatural opponent, while two out of the total of four motifs of ATU 123 are also found in ATU 333 (see Table 1).

### 3 Related work

Folktale similarity using WordNet-based term matching has been previously investigated (McIntyre and Lapata, 2010; Lestari and Manurung, 2015) using the hierarchical similarity measure of Wu and Palmer (1994). In this approach a folktale is considered sequential, with similarity computation based on alignment of the sequence of actions and actors. In contrast, the current approach considers the (non-sequential) presence of terms and term abstractions, similar to a bag-of-words approach, while preserving event or situation similarity by comparing folktales on a sentence level. Abstraction based on Dutch WordNet for folktale similarity has been used by Nguyen et al. (2013), using abstractions of verbs as one of several features involved in similarity computation. The abstraction feature did not improve the results significantly, which is attributed to limited coverage of the abstraction lexicon and inaccuracy of the grammatical analysis.

Characterizing semantic relations between folktales using TMI motifs is discussed by Karsdorp et al. (2012), presenting the conclusion that motif-based methods suffer from the limited amount of motif overlap between folktales. A search tool for TMI motifs using WordNet based semantic abstraction is presented in (Karsdorp et al., 2015). A mapping of nominal phrases to folktale actors using a domain-specific ontology for term abstraction is described by Declerck et al. (2012).

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Category</th>
<th>ATU</th>
</tr>
</thead>
<tbody>
<tr>
<td>F451.5.1.2</td>
<td>Dwarfs adopt girl as sister</td>
<td>Marvels → Spirits and demons → Underground Spirits</td>
<td>709 Snow White</td>
</tr>
<tr>
<td>Q2</td>
<td>Kind &amp; Unkind</td>
<td>Rewards and punishments</td>
<td>440, 480, Frog King, Kind &amp; Unkind Girls, 513, 571, Wonderful Helpers, Golden Goose</td>
</tr>
<tr>
<td>F911.3</td>
<td>Animal swallows man (not fatally)</td>
<td>Marvels → Extraordinary occurrences → extraordinary swallowings</td>
<td>123, 333, 700 The Wolf &amp; the Seven Kids, Little Red Riding Hood, Tom Thumb</td>
</tr>
<tr>
<td>F913</td>
<td>Victims rescued from swallower’s belly</td>
<td>Marvels → Extraordinary occurrences → extraordinary swallowings</td>
<td>123, 333, The Wolf &amp; the Seven Kids, Little Red Riding Hood</td>
</tr>
<tr>
<td>J144</td>
<td>Well-trained kid does not open to wolf</td>
<td>The wise and the foolish → Wisdom acquisition → education</td>
<td>123 The Wolf &amp; the Seven Kids</td>
</tr>
<tr>
<td>K1832</td>
<td>Disguise by changing voice</td>
<td>Deceptions → Deception by disguise</td>
<td>123 The Wolf &amp; the Seven Kids</td>
</tr>
</tbody>
</table>

Table 1: TMI motif examples.

---

2 An open source database using the structure of Cornetto and translations of the content of English WordNet is available as an alternative (Postma and Vossen, 2014a).
An unsupervised exploration and visualization method for concept clustering in folktales has been proposed by Honkela (1997), using self-organizing maps trained on word trigrams. Natural computing approaches using (phylo)genetic algorithms are used to study variation within folk-tale types and between closely related types, using TMI motifs and other story elements as features (Ross et al., 2013; Tehrani, 2013). A vector-based method for semantic folktales clustering using Latent Semantic Mapping is described in (Vaz Lobo and Martins de Matos, 2010).

Several semantic relatedness measures that use WordNet as knowledge base have been proposed, see, e.g., (Pedersen et al., 2004) for an overview. Considering hierarchy traversal, well-known approaches include the Wu-Palmer measure mentioned above, which defines similarity between two nodes as the path length from the first shared parent node to the root node of the hierarchy, and the Leacock-Chodorow measure, which finds the shortest path between two concepts (scaled for specificity of the hierarchy). Further graph-topological information is incorporated using PageRank (Agirre et al., 2009). Evaluation of graph-based semantic relatedness measures has been performed using comparison to human word-pair similarity ratings, e.g., (Postma and Vossen, 2014b). Recent approaches of similarity computation include path length weighting strategies (Gao et al., 2015) and domain-specific data (McInnes and Pedersen, 2015).

Using WordNet for similarity of documents has been investigated by, e.g., (Hotho et al., 2003; Sedding and Kazakov, 2004) for the task of document clustering. These approaches represent a document as a bag-of-words, consisting of terms in the document as well as term synonyms and hypernyms from WordNet. However, it is concluded that the investigated approach of adding WordNet relations does not improve clustering results significantly. Similar methods for document clustering do show improved results, e.g., (Wang and Taylor, 2007), suggesting a large impact of preprocessing and sense selection procedures. Further applications include information retrieval, matching a WordNet-expanded query to a set of (non-expanded) documents (Varelas et al., 2005).

Note that many approaches using WordNet for semantic similarity focus either on pairs of concepts (or synsets, words, lemmas, etc.), document clusters, or, in the folktale setting, variants of the same story. These tasks are generally motivated by the availability of evaluation resources, such as human concept similarity ratings, the Reuters categorized news corpus, or folktale corpora tagged by story type, respectively. In contrast, the current approach attempts to construct a network of documents based on semantic relatedness, by comparing document pairs on (non-sequential) sentence level. Evaluation of this approach is arguably less straightforward, however this task and the proposed WordNet-based method provide a shift in focus compared to traditional approaches.

4 Method

In the current approach a document collection is compared at sentence level. First, sentence boundaries, lemmas and part-of-speech tags are obtained using the Frog toolkit (van den Bosch et al., 2007). Lemmas tagged as noun (including proper names), adjective, or verb are selected (except for the common verbs be, have, can and will). The set of lemmas for a sentence is compared to the set of lemmas for all other sentences in all other folk-tales in the corpus. If a matching lemma is not found in the compared sentence then the WordNet hierarchy is consulted for a match at a higher level of abstraction, using the match level to adjust the similarity score. The similarity of two sentences is computed as the total of all match scores relative to the combined size of the lemma sets. Formally, the score \( s \in [0, 1] \) equals \( \frac{\sum_i \frac{1}{\text{level}(a_i)} + \sum_j \frac{1}{\text{level}(b_j)}}{|A| + |B|} \) for sentences \( A \) and \( B \) as sequences of WordNet lemmas and level(\( \ell \)) defined as the minimum level of the lemma \( \ell \) that matches a lemma (at any level) in the compared sentence. After computing similarity scores for all sentence pairs, for each (ordered) folktales combination \((f_A, f_B)\) the relative number of sentences in \( f_A \) is counted for which the most similar sentence in the corpus originated from \( f_B \). The procedure is described formally in Algorithm 1, an example is provided in Figure 2. For the mapping of sentence lemmas to WordNet the synset with the lowest WordNet sense number is selected, corresponding to some extent to the ‘default’ sense. Incorrect senses are assumed to be related or to have a minimal effect given the document size (cf. (Hotho et al., 2003)). During hierarchy traversal a random hypernym is selected for a given synset to limit the amount of branching. In the example the
Algorithm 1 F×F document pair similarity.

1: function WORDNETLOOKUP (sentence S)
2: set of tuples R ← ∅
3: for all (term, position, level = 0) in S do
4: synset syn ← WORDNET(term)
5: while syn ≠ undefined do
6: R ← R ∪ {syn, position, level}
7: level ← level + 1
8: syn ← WORDNETHYPERNYM(syn)
9: return R

10: function MAIN (document set F)
11: for all folktales fa ∈ F do
12: for all sentences A ∈ fa do
13: score_max ← 0, fm ← undefined
14: SYN_A ← WORDNETLOOKUP(A)
15: for all folktales fb ∈ F – {fa} do
16: for all sentences B ∈ fb do
17: s ← 0, ma ← [∞], mb ← [∞]
18: SYN_B ← WORDNETLOOKUP(B)
19: for all combinations ((ta, pa, la) ∈ SYN_A, (tb, pb, lb) ∈ SYN_B) do
20: if ta = tb and la < ma[pa] then
21: m_a[pa] ← la
22: if ta = tb and lb < mb[pb] then
23: m_b[pb] ← lb
24: for all matches m in ma, mb do
25: s ← s + 1/m
26: if s > score_max then
27: score_max ← s
28: fm ← fb
29: scores[fa][fm] ← scores[fa][fm] + 1/|A|
30: return scores[ ]

Figure 2: Sentence similarity example showing a score of \((\frac{1}{3} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{2}) + (0 + \frac{1}{3} + \frac{1}{3} + \frac{1}{2})\)/(5 + 4) = 0.47. English word translations in italics, grey nodes represent terms not listed in WordNet.

Overall values are somewhat lower than the correlation for the hierarchy traversal measures reported by Postma and Vossen (2014b) on Dutch WordNet, which might be caused by the lack of hierarchy depth awareness of the current method. However, the current measure is not intended as a stand-alone word pair similarity computation, instead it is part of an asymmetrical sentence matching procedure intentionally designed for matching on any level of the hierarchy.

<table>
<thead>
<tr>
<th>scored term</th>
<th>McNo</th>
<th>McRel</th>
<th>McSim</th>
<th>RgNo</th>
<th>RgRel</th>
<th>RgSim</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>0.64</td>
<td>0.60</td>
<td>0.64</td>
<td>0.54</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>target</td>
<td>0.44</td>
<td>0.39</td>
<td>0.49</td>
<td>0.53</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>lowest</td>
<td>0.59</td>
<td>0.54</td>
<td>0.63</td>
<td>0.53</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>average</td>
<td>0.62</td>
<td>0.56</td>
<td>0.65</td>
<td>0.58</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>highest</td>
<td>0.58</td>
<td>0.53</td>
<td>0.61</td>
<td>0.58</td>
<td>0.54</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 2: Spearman’s ρ correlation between the abstraction measure and human similarity ratings.
5 Results and evaluation

Application of the current method on the folktale test corpus results in a matrix of pair-wise directed similarity scores, shown in Table 3. The graph of scores above a threshold of 10.0 (i.e., the most similar sentence for at least 10% of the sentences in folktale A was found in folktale B) is provided in Figure 3. The graph contains a number of central nodes, most notably Snow White, Hansel and Gretel, and The Wonderful Helpers. These nodes can be interpreted as representing a prototypical folktale, more specifically the fairy tale subgenre.

Increasing the similarity threshold to 13.0 (solid lines in Figure 3) reveals two clusters in the graph. The left cluster contains folktales featuring civilian protagonists, who find themselves in potentially harmful circumstances. The right cluster contains royal protagonists dealing with issues of moral values. The exception is Snow White, which has a royal protagonist, who is however banned from the royal court, living as a civilian house guest annex maid, and subject of murder attempts.

For comparison, the same method is applied without using WordNet abstractions, i.e., counting overlap in (lemmatized) terms as found in the text. This comparison (see Figure 4) shows that plain term overlap is less structured or partitioned in general and pair-wise relations display less topic overlap as compared to the abstraction method.

To provide an evaluation of the proposed similarity measure, a comparison is performed to the traditional ATU classification and associated TMI motif sets. To address the problem of limited motif overlap between folktales, the hierarchical TMI numbering system can be used for partial or abstract motif overlap using an approach similar to the term similarity computation described in Sec-

Table 3: Pair-wise WordNet-based similarity scores for the test corpus. Thresholds indicated in italics (10.0) and bold (13.0).
is used, which considers the number of edges of a node and the direct neighbors of this node and compares these numbers to the overall degree distribution of the network. In Figure 6 the assortativity values for both types of similarity computation are shown as a correlation graph. The figure shows correspondences for peripheral nodes and the central Snow White node, as well as differences for nodes which are central in one of the two graphs only. Note that assortativity is not a measure of centrality as such. The definition takes into account the difference in degree between neighboring nodes, i.e., a larger part of the network is measured as compared to single degree count.

6 Discussion and future work

The current WordNet-based similarity measure divides the example folktale set into two clusters corresponding to civilian protagonists in threatening circumstances and royal protagonists presented with moral choices, respectively. This result shows that the method is able to differentiate general topics in folktales based on overlap in terms and term abstractions. Using term ab-

<table>
<thead>
<tr>
<th>ATU</th>
<th>Title</th>
<th>Motif description</th>
<th>Motif code</th>
<th>match level</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>The Wolf &amp; the Seven Kids</td>
<td>Disguise by changing voice</td>
<td>K1832</td>
<td></td>
</tr>
<tr>
<td>333</td>
<td>Little Red Riding Hood</td>
<td>Wolf puts flour on his paw to disguise himself</td>
<td>K1839.1</td>
<td>4</td>
</tr>
<tr>
<td>533</td>
<td>The Speaking Horsehead</td>
<td>Disguise as goose-girl (turkey-girl)</td>
<td>K1816.5</td>
<td>3</td>
</tr>
<tr>
<td>709</td>
<td>Snow White</td>
<td>Imposter forces oath of secrecy</td>
<td>K1933</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compassionate executioner: substituted heart</td>
<td>K0512.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Example motif matches for The Wolf & the Seven Kids.

Figure 5: Graph of directed pairwise folktale similarity scores using the Thompson Motif Index. The number of motif pairs with the highest overlap is shown for pairs of documents (relative to the number of source document motifs), restricted to the top two highest ranked target nodes for each document.
Stratification increases the level of clustering. The comparison with traditional folktales shows corresponding similarity relations and centrality for a number of folktales, but deviating results for others. However, even though both analysis methods measure folktales similarity, the WordNet similarity measure considers the full text of a document, involving both syntax and semantics, while motif analysis is based on a small set of key events or themes, resulting in a highly specific semantic comparison on a considerably reduced and condensed representation of the document. The difference in approach leads to different results of similarity computation as well.

Rather than an alternative approach of computing TMI similarity, the WordNet method should be considered an alternative text-oriented measure of similarity of folktales. The current approach has the advantage that a domain-specific classification system is no longer required. Within the folktales domain this addresses the issue of selective motif attribution and differences in motif granularity for folktales featured in the existing catalogs, as well as the possibility to include folktales outside of the catalog, as demonstrated in the current test corpus. This advantage extends to potential use outside of the folktales domain, e.g., using general literary works or non-fictional narratives.

The current method is ranking-based, therefore a strong match between two documents (e.g., two variants of the same narrative) may cause less pronounced similarities to remain undetected. This behaviour can be exploited for incremental clustering, by leaving out the comparison of highly similar document pairs in subsequent iterations.

In future work, the granularity of the WordNet hierarchy and the relative position in the concept tree can be used to adjust term matching weights. Word sense disambiguation can be taken into account. The method can be applied on larger or more heterogeneous corpora, e.g., folktales documents lacking standardized spelling or grammatical sentences could be used to test the robustness of disambiguation-based approaches. The approach could be extended towards discourse analysis to accommodate story element matching across sentence boundaries. Scalability issues resulting from the current method of comparing every pair of sentences for every pair of documents could be addressed using various pruning, ranking, and selection mechanisms. Finally, the development of an informative baseline (e.g., using existing clustering toolkits) and an automatic evaluation procedure tailored towards the current notion of narrative similarity (e.g., using story variants as in (Nguyen et al., 2013)) is desired to increase understanding and interpretation of current results.

References


Figure 6: Correlation of assortativity of similarity graphs.


