Construction and Annotation of a French Folkstale Corpus

Anne Garcia-Fernandez*, **Anne-Laure Ligozat**** ***, **Anne Vilnat****
*Laboratoire d'Anthropologie Sociale - Collège de France - CNRS,
52 rue du Cardinal Lemoine, 75005 Paris, France
annegf [at] college-de-france [dot] fr
** LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
firstname.lastname [at] limsi [dot] fr
*** ENSIIE, 1 square de la résistance, 91000 Évry, France

Abstract

In this paper, we present the digitization and annotation of a tales corpus - which is to our knowledge the only French tales corpus available and classified according to the Aarne&Thompson classification - composed of historical texts (with old French parts). We first studied whether the pre-processing tools, namely OCR and PoS-tagging, have good enough accuracies to allow automatic analysis.
We also manually annotated this corpus according to several types of information which could prove useful for future work: character references, episodes, and motifs.
The contributions are the creation of an corpus of French tales from classical anthropology material, which will be made available to the community; the evaluation of OCR and NLP tools on this noisy corpus; and the annotation with anthropological information.
Keywords: corpus construction, digital humanities, corpus annotation

1. Introduction

Tales represent an important material in anthropology and ethnology. Classical tale corpora have been manually annotated and classified in these domains. Digitizing the corpora and annotation would enable both their wider dissemination and their use in automatic classification or information retrieval systems for example. Our work is a first step towards this objective.

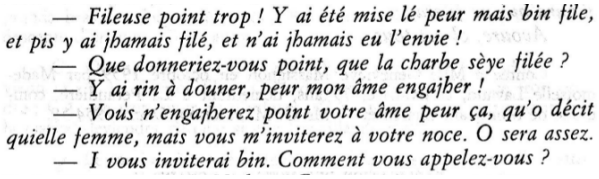
Many studies have tried to analyze tales either according to their types, their structures or the kind of information they contain. Several typologies of tales have been proposed. For example (Miller, 1893) proposed a typology containing three main kinds of tales: fairy tales, realistic tales and animal tales. This classification was extended by (Wundt, 1905) who added for example a class for mythological tales and moral tales. Other classifications are based on the content of tales. (Volkov, 1924) proposed 15 classes which correspond to topics addressed within the tale (The naive hero, The dragon-slayer...). This classification is based on different types of criteria since, for example, some classes are based on a characteristics of a character and some classes are based on events happening in the story. The most detailed classification was proposed by (Aarne and Thompson, 1973) and is described in section 2.2..

One of the drawbacks of these classifications is that a tale can correspond to several classes (for example, a tale with animals can also contain a moral), that the classification of a tale is subjective, and that classes are heterogeneous (Holbek, 1965; Propp, 1965).

The Aarne&Thompson tale type (A&T classification) is nevertheless the most detailed classification and a number of works used it to classify tales. For example, (Nikivorof, 1926) classified 125 Russian tales using A&T classification from which 20% are explicitly approximately classified. French tales were also classified in this way by (Delarue and Ténèze, 1997) who published several books citing more than 5 000 tale references and gathering more than 500 tales. A very large corpus of 40 000 Dutch tales (Meder, 2010) was also classified in this way.

These works classify tales without considering the content of the tales, but only analyzing its main category or the main event within it. Among works which studied the content of the tales itself, the main studies were proposed by (Propp, 1965). He proposed an analysis of the narrative structure of folktales, called the morphology of the folktales which concerns only fairy tales. This analysis considers tales as a sequence of segments (initial situation, departure of the hero... villain punishment, wedding). (Kwong, 2011) propose a structural annotation based on (Rumelhart, 1975) and (Mandler and Johnson, 1977). Even if this annotation remain simple and reduce the burden on the annotators, it has the default to mix several linguistic levels. Indeed, classes concern the semantic level (eg *Internal State* correspond to the emotion and state of mind of a protagonist), the discourse level (eg *Speech* correspond to direct speech among protagonists), the pragmatic level (eg *Episode* correspond to a self-contained description of a single incident).

Tales were also studied as a distinctive type of corpus in Natural Language Processing. They were often used for emotion analysis (Alm et al., 2005; Kim et al., 2010; Volkova et al., 2010; Keshtkar and Inkpen, 2010) using in particular (Alm et al., 2005) corpus, for example for affective text-to-speech systems. (Doukhan et al., 2012) also have storytelling objectives, but they performed a comprehensive pre-processing and annotation of their corpus: normalization, annotation of episodic structure, speech acts, characters and linguistic information. (Malec, 2010) annotated a corpus of 20 tales with Propp's narrative functions in order to create a training corpus for the automatic annotation of these functions. (Bod et al., 2012) annotated four tales for which Propp had given a reference annotation, and tend to show that character type and narrative function annotation have low inter annotator agreement. Other works have performed various linguistic analyses on tales, such as discourse analysis (Kwong, 2011), syntactic



— *Fileuse point trop ! Y ai été mise lé peur mais bin file, et pis y ai jhamais filé, et n'ai jhamais eu l'envie !*
— *Que donneriez-vous point, que la charbe sèye filée ?*
— *Y ai rin à douner, peur mon âme engajher !*
— *Vous n'engajherez point votre âme peur ça, qu'o décit quielle femme, mais vous m'inviterez à votre noce. O sera assez.*
— *I vous inviterai bin. Comment vous appelez-vous ?*

Figure 1: Excerpt from a tale

and semantic annotation (El Maarouf and Villaneau, 2012), script recognition (Jans et al., 2012). Closer to our objectives is (Nguyen et al., 2013), who classified automatically Dutch folktales according to Aarne-Thompson-Uther, and Brunvand classifications. Yet, in our case, we have a much smaller corpus annotated according to a classification, which is the result of a digitization, and is in French.

Table 1 propose a summary of works processing tales corpora to develop models, annotations or tools.

In this paper, we present the digitization and annotation of a tales corpus - which is to our knowledge the only French tales corpus available and classified according to the Aarne&Thompson classification - composed of historical texts (with old French parts). We studied first whether the pre-processing tools, namely OCR and PoS-tagging, have good enough accuracies to allow automatic analysis.

We also annotated manually this corpus according to several types of information which could prove useful for future work: character references, episodes, and motifs.

Our contributions are the following:

- Creation of an corpus of French tales from classical anthropology material;
- Evaluation of OCR and NLP tools on this noisy corpus;
- Annotation with anthropological information: structural information, character mention and types, motifs.

2. Corpus Presentation

2.1. Books

Our corpus is composed of three volumes of "Le conte populaire français" (Delarue and Ténèze, 1997), which is a catalog of folkstales collected in France and French speaking countries. It was started by Paul Delarue, and finalized by Marie-Louise Ténèze, and classifies folktales according to Aarne&Thomspon classification. This book is no longer published, which makes its digitization and exploitation critical.

# tales	107
# pages	1,333
# words	85,600
mean #words per tale	800

Table 2: Corpus characteristics

Figure 1 presents an excerpt from a page.

2.2. Aarne & Thompson Typology

(Aarne and Thompson, 1973) proposed a classification of Folktales which is the first hierarchical classification proposed. It is organized in three levels: the first level contains broad categories corresponding to the tale's kind (like Fairy Tales, Religious Tales, Animal Tales, etc.), the second level indicates the type of the tale (like Supernatural Opponent, Supernatural Tasks...), and the third level corresponds to the variety of the tale. For example, the tale "*Hop-o'-My-Thumb*", also known as "*Little Thumbling*" [1], is a Fairy Tale, of the type Surpernatural Opponent and of the variety "The Dwarf and the Giant" while the story of "*Hansel and Gretel*" has the same kind and type, but of a different variety. This classification, called Aarne&Thompson tale type index, is composed of 10 kinds, 46 types and 2400 varieties.

Each tale type is described by the main motifs present in the story. A motif refers to a specific element of the story such as an event, a character, or an object. (Thompson, 1955) established a typology of thousands of motifs in which these classical elements of stories are hierarchically organized into 23 mains classes (such as magic, animals, or mythological motifs), and 143 sub-classes (such as mythical animals, magic animals, or animals with human traits).

3. Evaluation of OCR and NLP tools

3.1. Corpus Digitization

The books were digitized previously to this work and consisted of a single pdf file. This file was split into 1,334 pdf pages, which were then digitized. The mean length of tales is of about 800 words.

We compared three OCR tools: gocr[2], ocrad [3] and tesseract [4]. The outputs of these tools were converted to utf-8 if necessary, and the files were normalized to unify some characters (for example different kinds of apostrophes were converted to the standard typewriter one).

In order to evaluate these tools, we manually corrected tesseract's output for 10 randomly chosen reference pages (chosen in the pages containing the tales themselves, and not comments or indexes). These pages represent a total of 3,429 words (proper excluding for example punctuation signs) and 18,124 characters.

Character precision was calculated with OCRopus [5] evaluation script `ocropus-econf`. Word precision was calculated with a script based on wdiff, and only takes into account the precision on words (excluding punctuation signs). The results are given in Table 3.

Tesseract obtained the best results with a 0.93 precision on words and 0.02 error rate on characters. The most frequent character confusion by far (44 occurrences in the reference file) is between 'h' and 'b', then the errors concern mostly confusions with punctuation such as between 'P' and '?', '.' and '!' or accents such as 'e" and ''. One particularity of the

[1] The original title of this tale by Charles Perault is "Le petit poucet".

[2] `http://jocr.sourceforge.net/`

[3] `http://www.gnu.org/software/ocrad/`

[4] `http://code.google.com/p/tesseract-ocr/`

[5] `http://code.google.com/p/ocropus/`

Language	# Tales	Source	
Russian	100	Russian fairy tales (Afanasev, 1945)	Model of narrative structure (Propp, 1965)
			Pre-processing tool for PFTML Annotations (Malec, 2010)
			Manual annotation of character types(Bod et al., 2012)
Dutch	$\sim$ 40 000	Dutch fairy tales (Meder, 2010)	Automatic classification (Nguyen et al., 2013)
			Narrative genre study (Nguyen et al., 2012)
English	185 (22 annotated)	Childrens fairy tales, including Grimms, H.C. Andersens and B. Potters stories (Alm et al., 2005)	Emotion annotation and classification (Alm et al., 2005)
French	89	GV-Lex corpus, coming mostly from a collaborative website (Doukhan et al., 2012)	Manual annotation of tale character references, manual and automatic annotation of episodes, speech turns, lexical information (Doukhan et al., 2012; Doukhan, 2013)
	139	Fairy Tales Corpus (copyrighted texts from a website) (El Maarouf and Villaneau, 2012)	Syntactic and semantic (referential, character types, semantic role) annotations (El Maarouf and Villaneau, 2012)

Table 1: Comparison with other works on tales corpora

Tool	character error rate	word accuracy
tesseract	.02	.93
gocr	.21	.21
ocrad	.31	.35

Table 3: OCR evaluation

TreeTagger	MElt	Stanford Parser
0.81	0.8	0.73

Table 4: Tagger accuracies on the reference files

corpus is that it contains old French or dialects of French words (such as "rin" or "jhamais" in Figure 1) which could have been a problem for OCR (for example tesseract has language dependent parameters), but there was no correlation between the number of non French words and tesseract accuracy.

3.2. Part-of-speech tagging

As our goal is to automatically analyze the tales, it is important to evaluate the performance of standard annotation tools on this kind of texts.
We conducted an experiment on part-of-speech tagging, and evaluated three tools:

- TreeTagger (Schmid, 1994)
- MElt (Denis and Sagot, 2009)
- Stanford Parser (Green et al., 2011)

All reference texts were automatically annotated by the three tools. The reference files were created from the corrected OCR files, since the objective is to evaluate the difficulty of this kind of texts (containing both modern and old French), independently of the OCR accuracy. The output of the TreeTagger was then corrected, in order to create the reference set (this tagger was chosen because it also lemmatizes the word, so actually both PoS and lemma annotation were corrected).
The results are given in Table 4.
The results are quite close for all taggers, yet the TreeTagger obtains the best results, which may be due to a small bias in the manual annotation, since the reference was created by correcting the TreeTagger output.
We also listed the most frequent PoS confusions for each tagger. For the Stanford Parser for example, one of the most frequent confusion is between proper and common nouns.

4. Manual tale-specific annotations

4.1. Annotations

Our annotations (see Table 5) are partly based on those of (Doukhan et al., 2012), with the addition of Thompson motifs. The following elements were annotated:

- Structural elements: episode boundaries and types;
- Named entities: character references and types; Thompson motifs.

4.2. Annotation Protocol

We used brat [6] to annotate the tales. Though brat is slower when processing large files, its user-friendly interface makes it easy to annotate entities and relations.
Two annotators (through the authors of this article) each processed three tales: for each reference page (with corrected OCR output), we gathered the pages corresponding to the whole page from tesseract's outputs.
A first tale was annotated to test the annotation guidelines, then, after a consensus, two more tales were annotated to get first annotation agreements.

[6] http://brat.nlplab.org/

Type	Element	Annotations	Annotator 1	Annotator 2
structural	episodes	boundaries and types	34	37
named entity	characters	references and types	17	14
	motifs	text mention	55	34

Table 5: Annotations

4.2.1. Episodes

Following (Doukhan et al., 2012), episodes are defined as linear non overlapping segments which contain one or several sentences, with episode boundaries being generally associated to temporal, spatial or thematic changes, or to the emergence of new characters.
For episodes, the tale itself was annotated to indicate its beginning and end, and then the beginning and the end of each episode was marked, with a label attributed to each episode: title, exposition, triggering event, epilogue, and refrain. Definitions of these labels are the following (Doukhan et al., 2012) :

- Exposition: introduction, initial situation and character presentation.
- Triggering-event: event that modifies the initial situation and provokes the departure of the hero.
- Scene: atomic unit, part of the tale that moves the story forword.
- Refrain: passage of the tale that recurs with a nearly identical manifestation.
- Epilogue: ending of the tale with a conclusion and/or moral; includes hero recognition, exposure, transfiguration, punishment of the villain and wedding.

The objective with this annotation was to check if the texts could be segmented into several episodes with a high enough annotator agreement, as well as to assess the difficulty of label attribution.
As episode boundaries annotation corresponds to a text segmentation task, we used text segmentation evaluation metrics, provided by the SegEval package,[7] to compare the annotations made by each annotator. These measures are Pk and WindowDiff.
Equation 1 represents the window size (k), where N is the total number of sentences. Equation 2 is the traditional definition of WindowDiff, where R is the number of reference boundaries in the window from i to $i + k$, and C is the number of computed boundaries in the same window. The comparison (> 0) is sometimes forgotten, which produces strange values not bound between 0 and 1; thus we prefer equation 3 to represent WindowDiff, as it emphasizes the comparison (Scaiano and Inkpen, 2012).

(1) $k = \frac{N}{2 * number\ of\ segments}$

(2) $PK(r,h) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|\delta(r_i, r_{i+k}) - \delta(h_i, h_{i+k})|)$

(3) $WD(r,h) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(r_i, r_{i+k}) - b(h_i, h_{i+k})| > 0)$

[7] http://segeval.readthedocs.org/, (Fournier, 2013)

These metrics were used in this work to evaluate an inter-annotator agreement, and not to measure the performance of a system.

4.2.2. Characters

Character annotation corresponds to named entity annotation. In this work, we decided to annotate only the first mention of a character (since our objective is to assess the presence of a particular character type), with its character type (Propp, 1965):

- Hero: main character, nice, with whom the reader will normally associate most strongly, and who is the key person around which the story is told.
- Villain: struggles directly against the hero, morally bad, and against whom the Hero will typically fight at the end of the story.
- False hero: who appears to act heroically and may even be initially mistaken for the real Hero. Will try to steal the Hero's thunder, grabbing the credit, and will not pass the intermediate trial.
- Donor: met by chance by the hero who will help him, gives the Hero something special, such as a magical weapon or some particular wisdom.
- (magical) Helper: supports the Hero in his or her quest. Can appear along the way as friends or random people who act pro-socially to support the Hero, at critical moments to provide support, or may be found in a support role.
- Princess: goal of the quest. The princess may herself be the object of the quest, or the reward.
- Princess father: constrains the Princess or may dispatch the Hero on his mission.
- Dispatcher: sends the Hero on a quest or a set of quests to be completed before he gains the reward. This may be a family member such as a mother or father.

Roles can be combined together. For example, the dispatcher may also be combined with the false hero who then trails along behind (perhaps disguised as a helper). The presence of the roles is optional, and most tales include only a subpart of them.
As (Bod et al., 2012) had low inter-annotator agreement for the annotation of the character types, we tried to precise as much as possible their characterizations in the guidelines.
To evaluate the inter-annotator agreement on characters, we computed the strict F-measure, F_S (Doukhan et al., 2012).

(4) $F_S = \frac{2sm}{|A1|+|A2|}$

with $|A1|$ and $|A2|$ for the number of annotations of annotator 1 and 2, sm for the number of strictly common annotations.
We used the brat evaluation script BRATEval [8] which computes the F-measure between two annotation sets. This tool also enables to compute the F-measure including non exact matches, i.e. two entities will be considered as common if they overlap, but as the results show, there was no disagreement on character mention boundaries.

4.2.3. Motifs

Finally, we annotated Aarne & Thompson motifs. We chose to use the second level of this hierarchy, which seemed precise enough to enable tale classification for example, but contains a restricted number of motifs (142) so that their annotation be possible. Concerning brat configuration, we created entities for the first level of the hierarchy, and the second level was treated as an attribute corresponding the the first level selected.
This information can be considered as an event annotation. A motif is annotated only if it corresponds to an event or an action. For example the motif "*Royalty and nobility*" is not annotated at each occurrence of the king character.
We compared the mere presence of a motif in both annotations, in order to have an evaluation independent of the chosen text spans, as well as the F-measure.

4.3. Results of the annotation

Annotation of the first tale enabled us to verify our hypothesis about the difficulty of annotating each kind of information, and to clarify the annotation guidelines.
For example, for character references, we decided to annotate complete noun phrases (such as "*les trois messieurs*" - "*the three gentlemen*"), and to consider as characters only those who make at least one action in the tale or who have an importance in the development of the study. For the segmentation of tales, we decided that each part of the story repeated at least one time, even if it is small (one sentence only), have to be annotated as a "*refrain*" episodes. The annotation of the second and third tale gave us preliminary results.

4.3.1. Episodes

The segmentation of tales into episodes presents a WindowDiff of 0.41 and Pk of 0.38 on the second tale, and of 0.15 and 0.15 for the third tale, which is the same order of magnitude as the values found by (Doukhan et al., 2012). Disagreement are mostly due to difference in setting boundaries. For example, for the second tale, both annotators identified 13 episodes but only 4 boundaries are in common.
Concerning the identification of episodes "*exposition*", "*triggering event*", and "*epilogue*", annotators are always in agreement. Indeed, "*exposition*" is the first episode, "*triggering event*" is the second, and "*epilogue*" is the last one. On the contrary, the distinction between "*refrain*" and "*scene*" cause more disagreement.

4.3.2. Characters

For character mentions and types, the strict F-measure is 0.8750. For some characters, only one annotator indicated them, but the types and boundaries given to all other characters are identical (for a total of about 20 characters in these tales). When only one annotator identified a character, it is never an "*hero*", a "*villain*", neither a "*dispatcher*" but a "*donor*", or an "*helper*".

4.3.3. Motifs

For motif presence, the strict F-measure is of 0.65.
The result analysis shows that some types of motifs present a higher inter-annotator agreement: the motifs which are the most explicit in the texts, such as the Recognition of person transformed to animal. These are also the motifs for which the annotation corresponds to the motifs listed by Aarne Thompson for the corresponding tale type. For other motifs (such as Selling oneself and escaping), the annotations are much more divergent.

5. Conclusion

In this paper, we presented the creation of the digitized version of a French folkstale corpus. OCR and NLP tools were evaluated on this corpus, and the results make it possible to build automatic systems for this corpus.
We also annotated the corpus with structural, character and motif information. Results on the preliminary annotations show that this annotation can be made. Our objective is to create automatic annotations for retrieval and classification purposes, which will enable this corpus, as well as similar corpora, to be disseminated and examined.

6. References

A. A. Aarne and S. Thompson. 1973. *The types of the folktale: a classification and bibliography*. Suomalainen tiedeakatemia-Academia scientiarum Fennica, Helsinki, 2nd revised and augmented edition.

A. Afanasev. 1945. *Russian fairy tales*. Pantheon Books: New York.

C. O. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics.

R. Bod, B. Fisseni, A. Kurji, and B. Löwe. 2012. Objectivity and Reproducibility of Proppian Narrative Annotations. In *Proceedings of the workshop on Computational Models of Narratives, Istanbul*, pages 26–27.

P. Delarue and M.-L. Ténèze. 1997. *Le Conte populaire français. Catalogue raisonné des versions de France et des pays de langue française d'outre-mer*. Maisonneuve & Larose, new single volume edition.

P. Denis and B. Sagot. 2009. Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort. In *PACLIC*, pages 110–119.

[8] https://bitbucket.org/nicta_biomed/brateval

D. Doukhan, S. Rosset, A. Rilliard, C. d'Alessandro, and M. Adda-Decker. 2012. Designing French Tale Corpora for Entertaining Text To Speech Synthesis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

D. Doukhan. 2013. *Synthèse de parole expressive au-delà du niveau de la phrase : le cas du conte pour enfant*. Ph.D. thesis, Université Paris-Sud.

I. El Maarouf and J. Villaneau. 2012. A French Fairy Tale Corpus syntactically and semantically annotated. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

C. Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, page to appear, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Green, M.-C. de Marneffe, J. Bauer, and C. D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735. Association for Computational Linguistics.

B. Holbek. 1965. On the Classification of Folktales. In *International Congress for Folk Narrative-Research*, pages 159–161.

B. Jans, S. Bethard, I. Vulić, and M. F. Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344. Association for Computational Linguistics.

F. Keshtkar and D. Inkpen. 2010. A corpus-based method for extracting paraphrases of emotion terms. In *Proceedings of the NAACL HLT 2010 Workshop on Computational approaches to Analysis and Generation of emotion in Text*, pages 35–44. Association for Computational Linguistics.

S. M. Kim, A. Valitutti, and R. A. Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

O. Y. Kwong. 2011. Annotating the Structure and Semantics of Fables. In *PACLIC*, pages 275–282.

S. Malec. 2010. AutoPropp: Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales. In *First International AMICUS Workshop*.

J. M. Mandler and N. S. Johnson. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, 9(1):111–151.

T. Meder. 2010. From a Dutch folktale database towards an international folktale database. *Fabula*, 51:6–22.

V. F. Miller. 1893. *O Sbornikie materīalov dlia opisanīia miestnosteĭ i plemen Kavkaza, izdavaemom Upravlenīem Kavkazskago uchebnago okruga*. Tip. kantseliarīi Glavnonachalstvuiushchago grazhdanskoĭ chastiu na Kavkazie.

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

D.-P. Nguyen, D. Trieschnigg, and M. Theune. 2013. Folktale classification using learning to rank. In *35th European Conference on IR Research, ECIR*.

A. I. Nikivorof. 1926. Skazochi materialy Zaonezja, sobrannye v 1926 godou. *Obzor Rabot Skazossno Komissi*.

V. Propp. 1965. *Morphologie du conte*. Seuil, Original Ed. 1928.

D. E. Rumelhart. 1975. Notes on a schema for stories.

M. Scaiano and D. Inkpen. 2012. Getting more from segmentation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 362–366. Association for Computational Linguistics.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

S. Thompson. 1955. *Motif-index of Folk-literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-books, and Local Legends*. Indiana University Press.

R. Volkov. 1924. Skazka. *Razyskanija po sjužetusloženiju narodnoj skazki (Le conte. Recherches sur la composition des sujets au conte populaire), Odessa*, page 6.

E. P. Volkova, B. J. Mohler, D. Meurers, D. Gerdemann, and H. H. Bülthoff. 2010. Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106. Association for Computational Linguistics.

W. Wundt. 1905. *Volkerpsychologie. Mythus und religion*, volume 2. Wilhelm Engelmann, Leipzig.