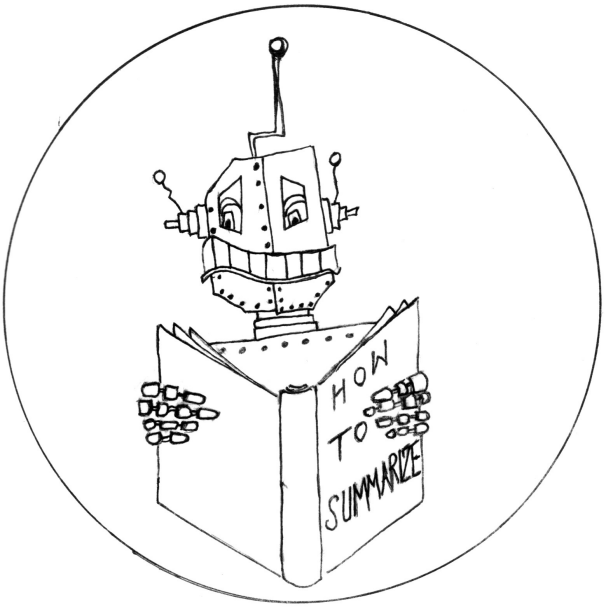
TOWARDS MORE HUMAN-LIKE TEXT SUMMARIZATION: STORY ABSTRACTION USING DISCOURSE STRUCTURE AND SEMANTIC INFORMATION

MAXIMILIAN DROOG-HAYES



Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

Department of Computer Science
School of Electronic Engineering and Computer Science
Queen Mary University of London

May 2019

Maximilian Droog-Hayes: *Towards more human-like text summarization: Story abstraction using discourse structure and semantic information,* Doctor of Philosophy (PhD), © May 2019

SUPERVISORS:
Geraint Wiggins
Matthew Purver

DECLARATION

I, Maximilian Droog-Hayes, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged within.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

London, May 2019

Maximilian Droog-Hayes

To all the characters of my story, this tale wouldn't have come together without you.

Any story worth its salt can handle a little shaking up.

— Haroun and the Sea of Stories

ABSTRACT

With the massive amount of textual data being produced every day, the ability to effectively summarise text documents is becoming increasingly important. Automatic text summarization entails the selection and generalisation of the most salient points of a text in order to produce a summary. Approaches to automatic text summarization can fall into one of two categories: abstractive or extractive approaches. Extractive approaches involve the selection and concatenation of spans of text from a given document. Research in automatic text summarization began with extractive approaches, scoring and selecting sentences based on the frequency and proximity of words. In contrast, abstractive approaches are based on a process of interpretation, semantic representation, and generalisation. This is closer to the processes that psycholinguistics tells us that humans perform when reading, remembering and summarizing. However in the sixty years since its inception, the field has largely remained focused on extractive approaches.

This thesis aims to answer the following questions. Does knowledge about the discourse structure of a text aid the recognition of summary-worthy content? If so, which specific aspects of discourse structure provide the greatest benefit? Can this structural information be used to produce abstractive summaries, and are these more informative than extractive summaries? To thoroughly examine these questions, they are each considered in isolation, and as a whole, on the basis of both manual and automatic annotations of texts. Manual annotations facilitate an investigation into the upper bounds of what can be achieved by the approach described in this thesis. Results based on automatic annotations show how this same approach is impacted by the current performance of imperfect preprocessing steps, and indicate its feasibility.

Extractive approaches to summarization are intrinsically limited by the surface text of the input document, in terms of both content selection and summary generation. Beginning with a motivation for moving away from these commonly used methods of producing summaries, I set out my methodology for a more human-like approach to automatic summarization which examines the benefits of using discourse-structural information. The potential benefit of this is twofold: moving away from a reliance on the wording of a text in order to detect important content, and generating concise summaries that are independent of the input text. The importance of discourse structure to signal key textual material has previously been recognised, however it has seen little applied use in the field of auto-

matic summarization. A consideration of evaluation metrics also features significantly in the proposed methodology. These play a role in both preprocessing steps and in the evaluation of the final summary product. I provide evidence which indicates a disparity between the performance of coreference resolution systems as indicated by their standard evaluation metrics, and their performance in extrinsic tasks. Additionally, I point out a range of problems for the most commonly used metric, ROUGE, and suggest that at present summary evaluation should not be automated.

To illustrate the general solutions proposed to the questions raised in this thesis, I use Russian Folk Tales as an example domain. This genre of text has been studied in depth and, most importantly, it has a rich narrative structure that has been recorded in detail. The rules of this formalism are suitable for the *narrative structure reasoning system* presented as part of this thesis. The specific discourse-structural elements considered cover the narrative structure of a text, coreference information, and the story-roles fulfilled by different characters.

The proposed narrative structure reasoning system produces high-level interpretations of a text according to the rules of a given formalism. For the example domain of Russian Folktales, a system is implemented which constructs such interpretations of a tale according to an existing set of rules and restrictions. I discuss how this process of detecting narrative structure can be transferred to other genres, and a key factor in the success of this process: how constrained are the rules of the formalism. The system enumerates all possible interpretations according to a set of constraints, meaning a less restricted rule set leads to a greater number of interpretations.

For the example domain, sentence level discourse-structural annotations are then used to predict summary-worthy content. The results of this study are analysed in three parts. First, I examine the relative utility of individual discourse features and provide a qualitative discussion of these results. Second, the predictive abilities of these features are compared when they are manually annotated to when they are annotated with varying degrees of automation. Third, these results are compared to the predictive capabilities of classic extractive algorithms. I show that discourse features can be used to more accurately predict summary-worthy content than classic extractive algorithms. This holds true for automatically obtained annotations, but with a much clearer difference when using manual annotations.

The classifiers learned in the prediction of summary-worthy sentences are subsequently used to inform the production of both extractive and abstractive summaries to a given length. A human-based evaluation is used to compare these summaries, as well as the outputs of a classic extractive summarizer. I analyse the impact of knowledge about discourse structure, obtained both manually and automatically, on summary production. This allows for some insight into the knock-

on effects on summary production that can occur from inaccurate discourse information (narrative structure and coreference information). My analyses show that even given inaccurate discourse information, the resulting abstractive summaries are considered more informative than their extractive counterparts. With human-level knowledge about discourse structure, these results are even clearer.

In conclusion, this research provides a framework which can be used to detect the narrative structure of a text, and shows its potential to provide a more human-like approach to automatic summarization. I show the limit of what is achievable with this approach both when manual annotations are obtainable, and when only automatic annotations are feasible. Nevertheless, this thesis supports the suggestion that the future of summarization lies with abstractive and not extractive techniques.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Droog-Hayes, Maximilian (2017). "The Effect of Poor Coreference Resolution on Document Understanding". In: *29th European Summer School in Logic, Language and Information (ESSLLI) 2017 Student Session*, pp. 209–220.

Droog-Hayes, Maximilian, Geraint Wiggins, and Matthew Purver (2018). "Automatic Detection of Narrative Structure for High-Level Story Representation". In: *The 5th Computational Creativity Symposium, AISB-2018*, pp. 26–33.

— (2019a). "Detecting Summary-Worthy Sentences: The Effect of Discourse Features". In: *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, pp. 381–384.

— (2019b). "A Creative, Cognitively Inspired Approach to Text Summarization". In: *The 6th Computational Creativity Symposium, AISB-2019*.

the reader is the very space in which are inscribed,
without any being lost, all the citations a writing consists of;
the unity of a text is not in its origin,
it is in its destination

— Roland Barthes (Barthes, 1967)

ACKNOWLEDGEMENTS

Acknowledgements are tricky things. Just like the texts discussed in this work, thesis acknowledgements have their own typical structure. Who gets thanked first? Who gets thanked last? Who gets left out, either intentionally or accidentally? The right people need to be thanked in the right way.

Sadly, I cannot thank all such people here. Suffice to say that if you are reading this, then I am endlessly appreciative of the support you have given me in all aspects of my life during this time. And there are also many deserving of thanks who will never read this, yet have deeply affected and aided me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

A summary is a brief statement. It gives an account of the main points of something in a succinct way. That something could be a text, image, video, or speech, and the summary itself could take any of those forms too. Aside from the modality, an ideal summary should also consider the many surrounding factors, such as length, purpose, genre, and intended audience.

This thesis contributes to the field of automatic text summarization: the automatic creation of textual summaries from textual documents. Since the beginnings of this field, there have been a variety of definitions as to what automatic summarization entails. Spärck Jones (1999) provides a broad definition of a summary as "a reductive transformation of source text to summary text through content reduction by selection and/or generalisation on what is important in the source" Spärck Jones (1999, p. 2). She goes on to say that the information gathered to produce a summary must be relevant to a particular subject or for a particular purpose (Spärck Jones, 2007). This definition implies a focused summary for a particular task and that many different summaries could be produced from a single document depending on the context. Radev, Hovy, and K. McKeown (2002) simply state that a summary should cover the main points of input text(s) and be no more than half their length. This definition covers the potential use of multiple documents as an input, however the restriction on summary length may be useful only as a general rule. As found by Harman and Over (2004), summary lengths can vary greatly due to differing granularities of the content, or simply the verbosity of the summarizer (human or otherwise). Other variations state that a summary is a subset of data which represents the information contained in an entire document. Definitions in this style imply that a binary decision can be made about the summary-worthy nature of each segment of a document, that there is entirely redundant information in the input document and that no abstraction is required.

Traditionally there have been two main branches of automatic text summarization: *Extractive* approaches and *Abstractive* approaches. Extractive approaches select and concatenate material directly from the input document, usually at the sentence level. In contrast, abstractive approaches involve understanding and representing the content of a document before generating a summary from this representation in a concise and original way. This thesis concerns only a small part of the overall summarization problem: can structural information be

automatically detected and used to produce informative, abstractive summaries?

1.1 MOTIVATION

Traditionally, a summary is produced by someone who reads a document and is required to provide a brief account of the most important points, possibly for a specific purpose. However, it is now the case that a vast and increasing amount of data is being produced every year. So much so that it is becoming difficult to identify the extent of what is available, let alone read it all. As such, it is becoming increasingly desirable to automate at least some of the summarization process.

Research into automatic text summarization began sixty years ago with the work of Luhn (1958). His work involved automatically creating literature abstracts of scientific articles by sentence extraction, where sentences were scored according to the proximity of frequently occurring words. This initial research focused on the creation of summaries in order to save a prospective reader time and effort in finding useful information. Luhn did not provide a statistical analysis of the generated summaries, but instead gave examples and discussed their potential. Aside from the removal of human effort, Luhn talked about the possibility of machine methods of summarization removing human bias. He suggested that the work of an abstracter is almost always influenced by their background, opinions, and interests, and that if the same person were to summarize the same document at two different times, the results would not be identical. Additionally, Luhn recognised the need for different types of document to be treated differently. He discusses how automatic summarizers must be based on properties of writing ascertained by analysis of specific types of literature.

Sixty years on, these important issues identified by Luhn (1958) are often ignored. It is not uncommon for current summarization systems to be trained on large corpora of documents and their corresponding human-written summaries. Luhn pointed out the diversity of summaries that can be produced by different people, or even the same person at different times. Nevertheless, summarization systems are being trained over collections of documents with only a single corresponding summary for each. Moreover, the most commonly used summarization evaluation metric (ROUGE) is based around lexical comparisons between a generated summary and a set of one or more supposedly ideal human written summaries. There are evident issues with this approach, given the range of ways in which a concept can be expressed.

Sixty years on, the field also remains dominated by extractive approaches to summarization. It is of course understandable that these

more limited approaches were appropriate when the field was in its infancy. However, they should no longer be the focus of the field. Abstractive approaches represent the natural human-like way of writing a summary, and are desirable in order to get closer to human levels of summarization performance. Extractive approaches have been called the low-hanging fruit of summarization, being technological rather than fundamental (Spärck Jones, 2007). In addition it has been shown that not only does human sentence extraction perform poorly in comparison to regular abstractive summarization, but that automatic extractive systems from several years ago were already approaching the ceiling of what can be achieved by human extractors (Genest and Lapalme, 2013).

1.1.1 *Motivating Examples*

In this section I discuss the summarization of two well known fairy tales. These are used for illustration purposes, as the issues that arise in the summarization of both tales elucidate the issues which motivate the direction of this thesis. Links to the full text versions of each fairy tale are provided in case the reader is unfamiliar with them. The examples clearly illustrate several issues and make the subject of this thesis more tangible.

1.1.1.1 *Below the Surface*

This first example is intended to demonstrate the limits of only considering shallow surface features of a text when determining summary-worthy content.

Consider *The Emperor's New Clothes*[1] (Andersen, 2008). In this story, an emperor who is excessively fond of clothes gets tricked by two swindlers posing as weavers. The swindlers tell the emperor that they can make a suit of clothes that is invisible to those who are unfit for their positions, or who are extraordinarily simple. In reality, they make no such suit and pretend to be working while the emperor continues to provide them with money and materials. The emperor, all of his court, and the general public all believe the clothes are invisible to them, but keep up the pretence that they can see the clothes for fear of being mocked. Finally, while the emperor is parading in his new 'clothes', a child cries out that the emperor is not wearing anything at all, and the crowd drop the pretence that they can see the emperor dressed in a fine suit of clothes.

The child plays an exceptionally important role here, his involvement changes the meaning of the story. However he has only two mentions in the whole text, both within the final five sentences. Even

1 The full text of this story can be read at http://www.gutenberg.org/cache/epub/1597/pg1597.txt

with coreference information, traditional surface-based metrics will not capture the importance of such a character. It is necessary to realise that if a character is only introduced so close to the end of the text, they must have a purpose.

1.1.1.2 *Beyond Extraction*

This second example illustrates the limits of using solely extractive techniques for summarization, even with knowledge of the important content. Below is an extractive summary of the widely known fairy tale *Little Red-Cap*[2], otherwise known as *Little Red Riding Hood* (Grimm, 2008). This is a summary I have created manually, selecting and concatenating the sentences which I believe best capture the salient points of the story in a concise manner.

> Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother, and there was nothing that she would not have given to the child. One day her mother said to her: "Come, *Little Red-Cap*, here is a piece of cake and a bottle of wine; take them to your grandmother, she is ill and weak, and they will do her good. The grandmother lived out in the wood, half a league from the village, and just as Little Red-Cap entered the wood, a wolf met her. The wolf thought to himself: "What a tender young creature!" Meanwhile the wolf ran straight to the grandmother's house and knocked at the door. The wolf lifted the latch, the door sprang open, and without saying a word he went straight to the grandmother's bed, and devoured her. Little Red-Cap, however, had been running about picking flowers, and when she had gathered so many that she could carry no more, she remembered her grandmother, and set out on the way to her. She was surprised to find the cottage-door standing open, and when she went into the room, she had such a strange feeling that she said to herself: "Oh dear! And scarcely had the wolf said *this*, than with one bound he was out of bed and swallowed up Red-Cap. The huntsman was just passing the house, and thought to himself: "How the old woman is snoring! Then just as he was going to fire at *him*, it occurred to him that the wolf might have devoured the grandmother, and that she might still be saved, so he did not fire, but took a pair of scissors, and began to cut open the stomach of the sleeping wolf. When he had made

2 This summary is based on the copyright-free version of the fairy tale available from Project Gutenberg. The full text can be read at https://www.gutenberg.org/files/2591/2591-h/2591-h.htm#link2H_4_0023

> two snips, he saw the little Red-Cap shining, and then he made two snips more, and the little girl sprang out, crying: "Ah, how frightened I have been! Red-Cap, however, quickly fetched great stones with which they filled the wolf's belly, and when he awoke, he wanted to run away, but the stones were so heavy that he collapsed at once, and fell dead.

The original text of this story is comprised of 84 sentences. My attempt at an extractive summary contains just 13 of those, or a little over 15%. While this summary reflects only my own opinions of the salient points of the story, it allows me to illustrate two important issues with extractive methods.

Firstly, there is a trade-off between ambiguity and concision. In the above summary I have emphasised three phrases that are particularly ambiguous without the context of the originally surrounding sentences. The first of these, 'Little Red-Cap' is included without the prior sentence elucidating that the little girl and the eponymous protagonist refer to the same character. The second, 'this', is referring to the preceding passage of the original text whereby the wolf attempts to trick Little Red-Cap. This passage constitutes a single narrative unit, made up of a dialogue between the wolf and Little Red-Cap. It only retains coherence if a contiguous span of 10-15 sentences are all included. However, doing so would double the length of this summary. Its inclusion would arguably result in a less cohesive summary too, expressing ideas at different levels of detail. The third point, 'him', is referring to the wolf. But the number of pronominal references in this sentence can lead to ambiguity.

The emphasised 'this' also highlights the second important issue. The preceding passage of the original text whereby the wolf attempts to trick Little Red-Cap can be expressed as just that. 'The wolf attempts to trick Little Red-Cap'. This is a more natural, human-like way of condensing related content and *abstracting* rather than *extracting*. A human reader is unlikely to recall this dialogue word-for-word, but they may well remember the narrative function it performs.

1.2 PROBLEM STATEMENT

The aim of this thesis then, is to investigate a more human-like approach to text summarization than is performed by current methods. I cover only a small part of the overall summarization problem: generic, single-document summarization. That is, the creation of general summaries of individual documents. My focus is on single-document rather than multi-document summarization as successful single-document summarization arguably demonstrates a greater level of document understanding. Multi-document summarizers can utilise the fact that certain concepts or spans of text may occur across

multiple documents in order to identify important content. Other dimensions of summarization such as purpose and intended audience are outside the scope of this current work.

Many current approaches to summarization consider only the surface features of a text in order to identify summary-worthy content. These approaches also rely on the surface of the input to generate summaries; spans of text, usually at the sentence level, are cut from the input document and joined together to form a summary. My particular focus is on the use of discourse-structural information to address these two issues. Psycholinguistic research into the cognitive models formed by humans as they read and recall texts provides us with information as to the textual elements that humans are most likely to recall, and thus include in a summary. It is my belief that discourse-structural information can be used to simulate some of these elements, and not only be used to better recognise summary-worthy content, but also to generate original summaries that are not dependent on the surface text of the input.

This thesis aims to answer the following questions:

1. Can discourse structural information be detected automatically?
2. Can this information be used to more accurately detect summary-worthy information than traditional approaches?
3. If so, which of these elements provides the greatest benefit?
4. Can this information be subsequently used to generated abstractive summaries?
5. If so, do these better convey the salient points of a text than summaries produced by existing algorithms?

To illustrate the general solutions proposed to the questions raised in this thesis, I use Russian Folk Tales as an example domain. This genre of text has been studied in depth and, most importantly, it has a rich narrative structure that has been recorded in detail. The rules of this formalism are suitable for the *narrative structure reasoning system* presented as part of this thesis. The specific discourse-structural elements considered cover the narrative structure of a text, coreference information, and the story-roles fulfilled by different characters.

In this work I detail a summarization system which allows these questions to be answered and evaluated. The summarization system itself is not the focus of this thesis, it only acts as a proof of concept and enables my research questions to be addressed. This is particularly important due to the imperfect processing stages involved in summarization. There are multiple required steps in summarization which complicate the success and evaluation of systems. Many

summarization systems rely on imperfect preprocessing steps such as Part of Speech tagging, dependency parsing and coreference resolution. As of yet none of these tasks can be automated perfectly, and it is unclear how errors from one step may propagate and be compounded. As such, some of the required annotations are carried out both manually and automatically. This allows a study of both the upper-bound potential of this approach, and what can be achieved with current technologies.

This thesis contributes to the field in the following ways. I present a narrative structure reasoning system, which provides a method by which the narrative structure of a text can be automatically determined for a suitably constrained domain of texts. This is evaluated for the example domain of Russian Folktales, and its utility in aiding the detection of summary-worthy content is shown. I present a method by which information about the structure of a text can be used to generate abstractive summaries, and I evaluate the summaries produced by this system. Methods of evaluation play an important role in this thesis. Throughout the studies of this work, I evaluate the effect of obtaining semantic annotations through varying degrees of automation. This allows for some insight into the knock on effects of inaccuracies in the preprocessing steps of summarization. Furthermore, various processes are evaluated with both automatic statistical evaluation metrics, and human-based studies. The results of this indicate indicate discrepancies between automatic statistical evaluation metrics, and human-based studies.

1.3 THESIS OUTLINE

This thesis presents a framework for abstractive summarization using discourse-structural information. It also includes a review of the relevant literature, and evaluations of multiple steps involved in the summarization process. The following outlines the contents of each chapter.

Chapter 2 is a discussion of related work broken down into three major sections. It begins with a review of some of the major approaches to summarization, enabling a discussion of their strengths and weaknesses, and the valuable concepts that they each employ. This is followed by a discussion of the necessary representations for summarization. This brings together research from psycholinguistics on the cognitive representations that humans form during reading, and research into discourse structure. I tie these together to suggest how discourse-structural information could be used to capture some of the elements of cognitive representations that psycholinguistics tells us human readers are more likely to store and recall, and thus include in a summary. Finally, I discuss existing work on the

evaluation of summarization systems, which occurs via the studying the summaries that they produce.

In Chapter 3 I describe two small-scale preliminary studies. These studies on coreference resolution and automatic summary evaluation were carried out in order to assess whether the use of these existing tools was appropriate to my work. The results of these studies influenced my methodological choices.

In Chapter 4 I provide an overview of my approach to summarization, motivated by existing literature. This facilitates the following three chapters which describe and evaluate the different stages of my approach.

Chapter 5 presents a narrative structure reasoning system. This uses knowledge about the narrative structure of instances of a genre in order to reason about the structure of a text from that domain. This system produces all valid interpretations of a text according to a set of encoded rules and constraints about the structure of its genre. I detail the application of this system to Russian Folktales, using the morphological analysis of Propp (2015) to obtain the structural constraints on this domain.

Chapter 6 describes the creation of statistical models used to predict summary-worthy content over the same example domain of Russian Folktales. The studies carried out for this chapter show the benefit of knowledge about discourse-structural and semantic features, including those obtained via the narrative structure reasoning systems. The predictive abilities of models using features obtained according to varying degrees of automation are compared to each other, as well as several existing extractive summarization systems. This analysis provides a starting point to compare these different approaches, as well as the knock-on effects of inaccuracies in automatic annotations.

In Chapter 7 I detail my approach to summary generation, which can occur to a desired length, and describe a method of comparative evaluation. I report the results of a human-based evaluation with nearly 2000 responses which compares four methods of summary production across two different summary lengths. These results show a clear indication of summary preference, and enable a continuation of the discussion about the knock-on effects of inaccuracies in automatic annotations.

I conclude this research in Chapter 8 with a summary of the results and contributions of this thesis. I then suggest some of the directions in which this could be continued in the future work of myself or others.

2 LITERATURE REVIEW

The primary focus of my research is an investigation into human-like approaches to text summarization. This chapter lays the groundwork to motivate the methods that I use, by bringing together strands of research from multiple fields.

I begin with a review of some of the major approaches to text summarization. This enables a discussion about the value in the methods employed by existing work, and highlights the differences to my own approach. Following this, I discuss the literature regarding the representations necessary for summarization. I summarize some of the psycholinguistic research which has proposed models for the cognitive representations of text that humans form while reading. The similarities between these models indicate the elements of a text that humans are most likely to recall. I consider research regarding the structure of discourse, particularly story structure and the work of Vladimir Propp, which is essential to this thesis. Finally, I provide a review of the methods that have been used to evaluate the success of summarization systems, which have also influenced my methodology.

2.1 APPROACHES TO SUMMARIZATION

Over the last sixty years an abundance of approaches to automatic text summarization have been investigated. Yet these cannot be placed neatly into a taxonomy of approaches due to the multiple dimensions of summarization. The summary construct, the *abstractive* and *extractive* distinction, appears to be the clearest divide between different approaches. However, there are several definitions of these terms. Additionally, some approaches may either merge or compress sentences. On the one hand, these techniques produce material not strictly in the original document, but they do not produce new lexical units or aggregate concepts across large spans of text. Spärck Jones (2007) proposes instead a division between *extractive* and *non-extractive*, which covers situations where summaries take different forms such as reviews.

Another complicating factor to classifying approaches is that many summarization systems are created to be domain specific. That is, they are either trained on documents from a certain domain or use specific features of a genre which make it inappropriate to use these systems in other contexts. For instance, the position of a sentence within a document is often considered as an important feature when

selecting the summary-worthy content of news articles. Some approaches to summarizing scientific articles have exploited the explicit structure of this domain. These methods are not always applicable or transferable to other domains of summarization.

Other dimensions of summarization, such as the purpose of a summary (such as answering a query (Bosma, 2005) or providing an indicative summary (Kazantseva and Szpakowicz, 2010)) further add to the difficulties of creating a taxonomy of summarization approaches. As such I focus on discussing a variety of different approaches to summarization, many of which have some degree of overlap. Some of these are domain specific; some are applicable only to a certain type of data; and some have more abstractive qualities than others.

2.1.1 *Heuristic Approaches*

Heuristic approaches to summarization rank and then extract sentences from a document based on a wide variety of potential features, such as position in the overall text or the usage of a word that appears in the title. These methods tend to learn which observable features of sentences best predict their inclusion or exclusion from a summary, without considering the meaning of the text.

The earliest work on automatic summarization involved the creation of extracts of scientific articles (Luhn, 1958). Sentences were ranked for selection based on the presence and proximity of frequently occurring words. Unlike the majority of later work on summarization, here the author admits the lack of sophistication in such an approach. Luhn does however suggest that uniformity of the summaries' derivation and the consistency of the outputs more than makes up for this.

During the early stages of research into automatic summarization, Edmundson (1969) used various heuristics based on the presence of cue phrases for the selection of sentences from scientific texts to create extracts. The goal was to replace the subjective notion of significance with more objective measures. The first of these heuristics was the use of a dictionary of cue words, indicating whether pragmatic words such as 'significant' counted as positively or negatively relevant. The selection of sentences including highly frequent content words was also used, as well as sentences containing words that appear in the document's title or headings. The final heuristic used the hypothesis that sentences which occur under certain headings are relevant, and that summary-worthy sentences tend to occur near the beginning or end of sections. The success of these heuristics was judged by calculating the sentence co-selection scores against manually created extractive summaries, where it was found that they performed noticeably better than a baseline of randomly selecting sentences.

Lal and Ruger (2002) describe a summarization system for news articles that scores sentences based on features such sentence length, position in the enclosing paragraph, and the position of that paragraph within the overall document. Their approach also examines whether a sentence includes frequently mentioned named entities, or if mentioned entities also appear in title elements of the text. This system was trained on a corpus of news articles, with sentences already annotated as extract-worthy, in order to learn the weights of features.

Sentence position has also been used by Nakao (2000) for the summarization of books. This was performed by first segmenting a text into a number of roughly equally sized topics and then producing a short two to three sentence summary on each of these, concatenated to create a one-page summary. In this work the author states that the lead sentence from each topic "probably indicates the contents of the subsequent parts in the same textual unit" (Nakao, 2000, p. 1). However it has been suggested that the use of sentence position works primarily for news articles and not for longer documents such as books (Ceylan and Mihalcea, 2007; Bamman and Smith, 2013). Ceylan and Mihalcea (2007) found that the use of sentence position as a selection heuristic led to worse results for their corpus of books, hypothesising that style and topic changes throughout books means that the lead sentences of sections do not necessarily cover essential aspects of the text.

2.1.2 *Sentence Manipulation*

Sentence manipulation techniques primarily involve the removal of redundant content from the sentences selected to form a summary. These rewriting techniques lead to summaries with a greater originality than purely extractive methods. However, the summary text is still heavily based upon the wording of the input text. Due to this, sentence manipulation is sometimes said to go beyond the abilities of extractive summarizers (Knight and Marcu, 2000), or categorised as a semi-extractive approach (Genest and Lapalme, 2013). There are two main forms of sentence manipulation; *sentence compression* and *sentence fusion*.

Sentence compression approaches aim to remove a redundant sequence of words from a selected sentence to produce a more concise output with the same meaning. Zajic, Dorr, and Schwartz (2004) use a set of linguistically-motivated heuristics to remove tokens from parse trees until they meet a constraint on their size. The aim of these heuristics is to remove the low-content aspects of the parses, such as determiners and relative clauses. This work was used to generate headlines from the leading sentence of a news article. This is an easier task than full document summarization, especially as a news headline does not need to be a full sentence. An alternative approach

by Knight and Marcu (2000) uses a probabilistic noisy channel model, trained on news articles paired with their human-written abstracts. In this model it is assumed that each sentence was generated from a shorter string, where the problem is to find the most probably short string that generated an observed sentence.

Sentence fusion methods aim to combine phrases conveying similar information from multiple sentences into a single generated sentence. The presence of such redundant information may be used in multi-document summarization to indicate its importance, and subsequently merged to remove redundancy and create new sentences not found in any of the input documents (Barzilay and K. McKeown, 2005). The Opinosis system of Ganesan, Zhai, and Han (2010) contains elements of both compression and fusion, extracting highly redundant content from a dataset of reviews of products and services. This system creates a directed graph structure with each unique word in the input corresponding to a single node. More frequently mentioned phrases appear as paths through the graph with a higher weighting. This captures repeatedly mentioned information and shows non essential elements that can be ignored. However, Carenini and Cheung (2008) have suggested that extractive approaches to multi-document summarization are not appropriate when there is corpus controversiality; when the corpus contains conflicting opinions.

2.1.3 *Information Extraction*

Information extraction approaches to summarization work by extracting structured information from unstructured natural language texts. Key information is both detected and subsequently presented in the form of a summary with the use of manually constructed templates, which move beyond the simple extraction and concatenation of material from input documents. Information extraction approaches to summarization began with the Fast Reading Understanding and Memory Program (FRUMP) system of DeJong (1982). This system would skim news articles with the aim of recognising and extracting key events in order to fill relevant templates. The idea behind such an approach was that pragmatic and semantic knowledge about certain events could be used to predict the information that should be reported.

M. White et al. (2001) proposed an information extraction approach to multi-document summarization for the domain of natural disasters. Their user-directed RIPTIDES system allowed users to optionally provide filters and preferences for the scenario templates used in the generation of summaries. Relevant information was then detected in input documents with the use of extraction patters. In addition to the template filling of elements such as named entities, dates, and

locations, sentence extraction was used to include information not covered by templates. Heuristic methods were used to prevent the inclusion of multiple, similar extracted sentences. This approach also considered the coherence of a generated summary by favouring the inclusion of fewer summary topics at a greater level of detail over the extraction of the highest n ranked sentences. Similar work has considered domain specific lists of aspects that should be covered in a summary, such as what/where/when information for the 'attack' domain (Genest and Lapalme, 2012).

2.1.4 *Lexical Chain Summarization*

Lexical chains are sequences of semantically related words. They capture the fact that a single concept may appear as multiple low-frequency words in a text which are, when taken individually, statistically unimportant. Realising that these occurrences all refer to the same idea gives a more accurate indication of the important concepts expressed in a text than simple word frequency. I use the methods described here to form lexical chains in my approach to summarization, as they are an easily identifiable type of cohesion in text. Identifying and linking cohesive elements of a document can help to create a more robust representation of a text, aiding the summarization process.

The work of Barzilay and Elhadad (1999) aims to go beyond brittle location based methods of summarization that rely on the structure of a text, to consider the content of a document. They considered the construction of lexical chains to be the most easily identifiable type of cohesion used to join different parts of text, also overcoming the limitations of word frequency information. To address the issue of some words having multiple senses, Barzilay and Elhadad (1999) explicitly create each possible lexical chain, and later select the most probable chains. In contrast, others have used WordNet (Miller, 1998) relations to implicitly choose the most likely chain and word sense for a candidate word based on its relatedness to members of each chain (Silber and McCoy, 2000a; Silber and McCoy, 2000b). A candidate word is compared to chain members according to hypernymy, hyponymy, synonymy and whether the candidate and a member element are sibling elements: sharing a hypernym. Important lexical chains are then selected based on the total frequency of all member elements. Barzilay and Elhadad (1999) present three potential heuristics for using this information to extract sentences and form a summary. The first of these involves selecting sentences which contain the first occurrence of each member of an important chain. The second leads to more concise summaries; for each chain, the first sentence containing the most frequently occurring member element is selected. The final heuristic examined finds the portion of the text with the most

mentions for a given lexical chain, and extracts the first sentence from this span which contains a member element. An alternate generation heuristic is proposed in Gonzàlez and Fort (2009) where sentences containing members of lexical chains are extracted until a predefined limit on the summary length is exceeded.

2.1.5 *Summarization by Sentiment Analysis*

Sentiment analysis is the process of identifying the emotive language in text and quantifying its affective state. This may be a simple classification of positive, negative, or neutral language. But some approaches assign an absolute value to emotive language on a scale between these extremes.

In summarization, attempts have been made to use sentiment analysis to separate subjective and objective information. This involves assigning sentiment scores to individual sentences and seeing how they compare to the rest of the document. Dabholkar, Patadia, and Dsilva (2016) examine the relative sentiment scores of sentences, the difference in sentiment between a sentence and the average of the document. They preferentially extract sentences with a relative sentiment score close to zero, claiming this indicates their neutrality in the context of their document. When too few sentences of this form are present, the sentences with the most polar sentiment values are additionally extracted. This method however makes no provision for situations where the entire document is highly emotive and subjective, where highly positive or negative sentences will appear as relatively neutral. In contrast, Reagan et al. (2016) use spans of highly emotive text to indicate major plot points of stories and classify their plot structure.

2.1.6 *Summarization Using Rhetorical Structure*

Rhetorical Structure Theory (RST) aims to explain the organisation of a text based on the rhetorical relations that hold between different spans, usually clauses, of text. The relations join the more critical *nucleus* statements to the less necessary *satellite* statements in one hierarchical tree structure. The primary difference between these two statement types is that the nuclei make sense independently of their satellites, but the converse is not true. Rhetorical Structure Theory will be explained in more detail in Section 2.2.2.1.

Many RST based approaches to summarization first create an RST parse tree of a document and then proceed to assign scores to each node in the tree. Ono, Sumita, and Miike (1994) describe a rule based approach to constructing RST parses, after which scores are assigned to nodes. Higher scores are given to nuclei and nodes closer to the root of the tree. The lowest ranked nodes are then removed from the

tree to leave a cut-down representation of the original document. A similar method is described by Marcu (1997) and O'Donnell (1997). However in O'Donnell (1997), nodes are further weighted according to the importance of the connecting RST relation, where the relative importance of each relation was decided empirically. After a method of scoring the content of a parse is applied, sentences or clauses of text are selected and concatenated until a given document compression ratio is reached.

Marcu (1998) suggests two ways to integrate the discourse structure of a document with RST summarization techniques. The first method involves scoring the importance of spans of text according to a combination of features. Both the position of a span in an RST parse, and more traditional heuristic methods such as sentence position or similarity to the document title are considered in the scoring process. However the author states that this still leads to the treatment of a document as a flat sequence of textual units, and that discourse structure should play a more central role in the summarization process. This is addressed in the second method, which focuses instead on obtaining the best discourse interpretation of a text. Discourse parsing is ambiguous, and multiple interpretations can be derived for a given text. As such, the authors propose a method which selects the interpretation which maximises the value of seven different heuristics inspired by other summarization techniques. These include a position-based metric, based on the assumption that important sentences occur near the start or end of a document, and so should occur near the root of a parse tree. The maker-based metric gives higher scores to interpretations which contain more rhetorical relations that explicitly occur in the input document. The spans of text closest to the root of this parse are then selected to form a summary.

Aside from RST, other approaches have aimed to discover and use information about the rhetorical status of spans of text for the purpose of summarization. The work of Teufel and Moens (2002) focuses on using rhetorical information to move beyond simple extractive approaches to summarization. They focus on the extraction of relevant sentences along with informative rhetorical tags. They recognise the desirability to go beyond the mere selection and concatenation of content in order to form a summary. However, the exent to which this is possible is dependent on contextual information about the selected content being preserved. Teufel and Moens (2002) define a scheme for the annotation of the rhetorical status of sentences specifically for the domain of scientific articles. Based on analyses of the particular rhetorical structure and requirements of summarization for this domain, they defined a set of seven rhetorical roles, designed to express the important discourse and argumentation aspects. Unlike RST, this scheme is non-hierarchical. While the authors agree that the overall structure of text is hierarchical, they believe that the determination of

the function of a span of text with regards to adjacent spans of text is not necessary to the understanding of the function of a span of text with regards to the overall message of a document. Using a corpus of scientific articles annotated with rhetorical role and relevance information, Teufel and Moens (2002) trained a classifier to provide a list of extracted sentences along with their rhetorical status for a given scientific article.

2.1.7 *Supervised Approaches*

These approaches to summarization use a training corpus of documents and their summaries in order to learn the combinations of features that predict the relevance of a sentence. For example, Lal and Ruger (2002) learn the importance of features such as the position of a sentence in a paragraph, the position of that in the overall text and the length of a sentence to predict summary-worthy sentences against a corpus annotated with extract-worthy sentences. However, approaches such as these depend entirely on the availability of a suitably large, annotated corpus. Supervised methods such as these also go against one of the ideas highlighted in the earliest work on automatic summarization: that the application of machine methods to summarization can remove bias from the abstracter's product, which can cause the quality of summaries to vary greatly.

Corpora may be used for a variety of purposes in connection with a summarization system. They have been used to discover cue phrases that indicate summary-worthy content (K. McKeown and Radev, 1995), as well as identify phrases that that humans are likely to omit when creating of summaries (Jing, 2000). Knowledge about the type of corpus that a document fits into can also be used to gain information about the relevant topics. Zajic, Dorr, and Schwartz (2004) use an unsupervised approach to discover topic models in a set of news articles in their work on summarization by headline generation. The topics associated with the target document are prepended to a compressed one-sentence summary to form a headline. Witbrock and Mittal (1999) also explores the use of a large corpus of news articles, in order to learn models of content selection between news articles and their accompanying headlines. Here the authors focus on the inability of extractive summarizers to produce an output shorter than a single sentence. They produce a statistical model of summarization by term selection and term ordering as an approximation, in place of what they call the ideal solution of document understanding and then producing an appropriate summary directly from that understanding. This may be as simple as the probability of a word in the source document being included in the headline-summary, or involve the relationships between words, or characteristics such as the Part of Speech (POS) tags of words. The probability of a surface realisation

of words is based on a bigram model of the probability of a sequence of words occurring. Summarization is then carried out by finding the combination of the words to be selected for the summary which gives the maximum probability.

2.1.8 *Hidden Markov Model approaches*

A Markov model describes a possible sequence of future states based only on the previous observable state or states. In a Hidden Markov Model (HMM), events are still observed and it assumed that the properties of a Markov model hold, but the states are hidden.

In some contexts large passages of text can be summarized very succinctly, meaning that individual phrases between source and summary may not align. HMM methods for summarization aim to find passages of text that align with a sequence of summary sentences (Conroy and O'Leary, 2001). Bamman and Smith (2013) propose two methods for HMM summarization. In the first, each HMM state corresponds to a passage of the input document according to the observation that summaries of long documents can condense entire passages into a single sentence. Their second method is a token model, where HMM states correspond to tokens in the input document.

2.1.9 *Term frequency-inverse document frequency approaches*

Term frequency-inverse document frequency (TF-IDF) is a term weighting scheme used in information retrieval to indicate the significance of the appearance of a word in a document relative to a corpus. It is the product of two statistics, the number of times a term occurs in a document and the proportion of documents in the corpus that contain the term. Common stop-words that appear in the majority of documents have a low importance, but a word that appears in relatively few documents across a corpus has a higher importance.

TF-IDF has been used in both single-document and multi-document summarization. In single-document summarization, TF-IDF has been used to look at the frequency of terms in a sentence relative to their appearance across a document (Term frequency-inverse sentence frequency). Scores are summed over each word in a sentence and sentences with scores above a certain threshold value are then concatenated to produce a summary (Neto et al., 2000). In multi-document summarization, TF-IDF has been used to identify clusters of similar documents, where a summary is produced per cluster. The words in the centroid of a cluster are identified and sentences containing these words are ranked. This ranking penalises sentences with a high word overlap to ensure a summary has broad coverage. Finally a summary is produced to a desired compression ratio by selecting the top scoring n sentences (Radev, Jing, et al., 2004).

TF-IDF has also been used in combination with other, mainly heuristic, approaches such as the position of a sentence within a document in order to select content for extractive summarization (Ceylan and Mihalcea, 2007; Saggion, 2008).

2.1.10 *Summarization by Latent Semantic Analysis*

Latent Semantic Analysis (LSA) is a technique used to find the relationships between terms and documents in a corpus based on the co-occurrence of words. It is used in information retrieval to find documents relevant to a query that don't necessarily contain the query terms. It has also been used in word sense disambiguation tasks, improving the Extended Lesk measure (Banerjee and Pedersen, 2003) by using the presence of words in a sense definition as well as the absence of words in other definitions to indicate the most likely word sense (Guo and Diab, 2012).

In summarization, LSA can be used to reduce a term-sentence matrix to a matrix of a lower rank. This is subsequently used to pick the highest ranked sentence about each of the n highest ranked topics (Gong and X. Liu, 2001). However, Steinberger and Jezek (2004) suggest two disadvantages to this method. First, prior knowledge about the number of topics that a document contains is unlikely, so the ideal size of the reduced matrix is not known. Furthermore, the larger this matrix, the less significant are the topics included in the summary. Second, sentences may be important to many topics, but unless they are the most important sentence to at least one topic, they will not be included in a summary. The authors propose an extension where the size of the reduced matrix is independent of the summary length and to weight scores towards sentences which are relevant to many topics. In evaluation, they find that this approach outperforms other LSA methods as well as simple heuristic and TF-IDF approaches.

The LSA summarization approach of Steinberger and Jezek (2004) is used as a comparison point in subsequent chapters of this thesis and as such it is worth expanding on the formula used. First of all, a term by sentence matrix A is created, where each column vector A_i represents the weighted term-frequency vector of sentence i. Singular Value Decomposition is then applied to the matrix A such that:

$A = U \sum V^T$

"where $U = [u_{ij}]$ is an mxn column-orthonormal matrix whose columns are called left singular vectors, $\sum = diag(\sigma_1, \sigma_2, ..., \sigma_n)$ is an nxn diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and $V = [v_{ij}]$ is an nxn orthonormal matrix, whose columns are called right singular vectors" (Steinberger and Jezek, 2004, p. 2).

In the approach of Gong and X. Liu (2001) the matrix V^T describes the importance of each topic in each sentence, and the single most im-

portant sentence for each topic is selected to form a summary. Steinberger and Jezek (2004) go further by computing the length of each sentence vector in the matrix V. This favours the index values in the matrix which correspond to the highest singular values. This is calculated as:

$$S_k = \sqrt{\sum_{i=1}^{r} v_{k,i}^2 \cdot \sigma_i^2}$$

where S_k represents the vector length for sentence k, and its importance score for summarization. r represents the level of dimensionality reduction, which is learned from the data.

Steinberger, Kabadjov, et al. (2005) investigated whether coreference information could be used to improve LSA based summarization. In this work, the authors found that simply substituting each referential mention for its semantic head did not lead to any significant improvement in results. However, adding separate information to indicate the entity that a mention refers to did lead to improved summarization results. This held true for both manually and automatically added coreference information, although better results were obtained with manual coreference resolution. In following work the authors showed that coreference resolution could also be used to improve generated summaries by correcting dangling anaphoric expressions (Steinberger, Poesio, et al., 2007).

2.1.11 *Neural Network Approaches*

In recent years there have been several neural network approaches to abstractive summarization, for example (Rush, Chopra, and Weston, 2015; Nallapati et al., 2016) and (Tan, Wan, and Xiao, 2017). Many of these approaches claim to be abstractive, as they generate original sentences not present in the input documents. However, they are typically unable to produce anything longer than single sentence outputs. I do not consider these approaches in detail, as they do not provide an interpretable intermediate representation. This prevents the summary generation process from occurring in a controllable yet creative way. My focus is on methods which include a symbolic representation, which can allow for control over the degree of summarization.

Highly relevant to these approaches is the field of machine translation, as the tasks of summarization and translation are somewhat analogous. Both the text of a document and a summary produced from it can be viewed as different languages, where the summary language expresses the same information more concisely. Both statistical (Banko, Mittal, and Witbrock, 2000), and more recently neural approaches to machine translation have been applied to summarization tasks. Initially developed for machine translation, the encoder-decoder neural network architecture has also been applied to summarization (Rush, Chopra, and Weston, 2015; Koupaee and Wang, 2018).

This architecture is useful for sequence prediction problems where the input and/or output may be of a variable length.

These approaches generally require vast amounts of training data. The approach of Koupaee and Wang (2018), for example, uses a dataset of over 200,000 articles, each of which is further subdivided into source paragraphs each with a target summary sentence. The majority of large scale summarization datasets consist solely of news articles; the scarcity of existing document-summary pairs for other types of text means that such approaches are not always applicable.

2.1.12 *Aspect Based Summarization*

Aspect based summarization involves generating a summary about the different facets of a dataset. As such it is usually most relevant in the context of product reviews, and has features in common with sentiment analysis and multi-document summarization.

Lu, Zhai, and Sundaresan (2009) study the summarization of product reviews by decomposing general reviews into summarized ratings for different aspects of a product and the quality of service provided in the transaction. This is desirable due to the vast quantity of product reviews available online, and the fact that customers may be willing to compromise on some aspects of a transaction but not on others. Reviews are parsed and the head words from each are clustered, where the most frequent clusters are taken as the aspects to be considered. Ratings for aspects are then assigned based on the average overall rating given to reviews which contain a given head word.

Knowledge about the discourse structure of reviews has also been exploited to infer the importance of aspects as well as the association between them. Gerani et al. (2014) use an approach involving an RST parser, removing all non aspect-based words and then merging the discourse trees. The subgraph containing the the most important aspects is identified, and this information is used along with sentence templates to generate summaries.

2.1.13 *Timeline Summarization*

Timeline summarization is a specific type of multi-document summarization used to present long-lasting news stories in a concise way (Tran, Alrifai, and Herder, 2015). Often constructed manually, these require a considerable amount of time and effort to create. Efforts at automatic summarization in this field have aimed to extract key sentences about events on a given date. However this style of summarization is generally confined to the news domain. News articles can be accompanied by meta-data, including the publication date, which reduces the need for temporal analysis.

Allan, Gupta, and Khandelwal (2001) discuss the need to produce an up-to-date timeline summary at regular intervals, stating that it does not make sense to wait for the topic to be finished before producing a summary. Their approach extracts sentences with the highest probability of being 'interesting'; the product of being both 'novel' and 'useful'. The idea of novelty is based on trying to select the first sentence about a topic, which is significantly different from preceding sentences. A 'useful' sentence is one which has a high probability of being generated from a language model of all sentences seen so far, indicating the importance of the material. Their model is compared against several baselines for sentence selection, including the random assignment of summary-worthiness scores to sentences. The authors note the relatively high performance of selecting sentences that have simply been scored in descending order of their position in a news article, supporting the belief that sentence position is a highly important feature in the domain of news.

Timeline summaries can alternatively be created with the use of headlines alone (Tran, Alrifai, and Herder, 2015). This was carried out with the aim of tackling the poor readability and low understandability of timeline summaries, in part by removing the issue of dangling anaphoric expressions. Their system was trained to only select self-contained headlines, those which explicitly describe an event, while also selecting those containing frequently mentioned events and having *influence*. That is, events that cause or have an impact on later events. This can be detected by subsequent mentions of the event or the date on which it occurred. While timeline summarization may be largely restricted to the domain of news, the idea of influential events is applicable more broadly. In other domains of text, such as stories, the desires of characters may prove important, as it is these which motivate the subsequent events.

2.1.14 *Graph-Based Summarization*

Graph-based approaches to summarization create a network-like intermediate representation of an input text. Summary-worthy content is typically obtained by using the structure of the graph to determine the most important vertices. Both the method for doing this, and the content represented by a node differs across approaches. Some graph-based approaches to summarization display far more *abstractive* qualities than methods discussed thus far.

TextRank (Mihalcea and Tarau, 2004) is an unsupervised graph-based ranking algorithm for graphs extracted from natural language texts. The importance of a node is determined by the number of edges connecting to it. This in turn influences the weight of the edges connecting to other nodes. In this way, re-ranking occurs until the values converge, using a similar idea to PageRank (Brin and Page, 1998). The

text units represented by vertices depends on the application, as does the relation used to connect vertices together. Mihalcea and Tarau (2004) describe two examples using this system, keyword extraction and sentence extraction. In keyword extraction relations are based on co-occurrence within a window, the closer two words occur, the stronger their link. The more words that co-occur with a given word, the more important that word is. For sentence extraction the vertices represent sentences, and the similarity function is the content overlap of a pair of sentences. Erkan and Radev (2004) describe a similar approach, LexRank, where edges are created instead based on the cosine similarity between a pair of TF-IDF vectors between any two given sentences.

Some graph based approaches to summarization have relied on the use of existing corpora of document-summary pairs. F. Liu et al. (2015) present a supervised learning approach to news summarization based on transformations of dependency parses. The parses of each sentence are merged, so that there are no duplicate nodes. This means, for example, that the graph contains only one instance of each verb which is connected to all of its arguments from every instance of the verb in the source text. As such it is unclear from the representation which actions are performed by whom. However, from this structure they learn to predict the subgraph which best represents a summary. The text from node labels is concatenated in order to form a bag of words summary. The generated words are output in no particular order, as this has no effect on ROUGE-1 evaluation metric being used. In a similar manner, Leskovec, Grobelnik, and Milic-Frayling (2004a) and Leskovec, Grobelnik, and Milic-Frayling (2004b) parse text to a subject-verb-object form and learn the mapping from a document graph to a summary graph. However, their approaches go further by annotating the graphs with more semantic information: synonyms are identified, and coreference information is added to create additional links on the graphs. Summaries are formed by selecting sentences from the input document containing at least one triple identified as summary-worthy. This leads to results which have a high recall but low precision, when evaluated against human extractive summaries in terms of sentence co-selection.

Graph approaches have also been applied to multi-document summarization, in the domain of news articles. W. Li (2015) define Basic Semantic Units (BSUs), which are action indicators with their associated actor and receiver arguments. These are used to extract semantic information from texts and form summaries. Constituency and dependency parses are used, along with coreference information and named entity recognition, to obtain the units that are used as nodes in the network. Nodes are linked based on semantic relatedness: a linear combination of the relatedness of arguments, the action verb, and the proximity of a pair of BSUs. Relatedness of the arguments

and action verb is measured by semantic analysis of Wikipedia data and WordNet similarity. BSUs with a high degree of similarity are clustered, and only the most salient unit is retained. Based on a training corpus of documents and their associated human summaries, if a BSU is extracted from both a source document and its associated summary then it is assigned a summary-worthy score. Unlike other work discussed so far, this approach includes a planing step for summary generation. An optimal path through the network is found which aims to traverse all nodes only once in a way that leads to adjacent generated sentences being semantically related and proceed in descending order of importance. BSUs contain enough semantic and syntactic information to then allow for sentence generation, but relying quite heavily on the original sentence structure. If two adjacent sentences contain the same subject, the latter mention is replaced by the relevant pronoun to improve linguistic quality. The authors state that this approach works well in domains containing facts and actions that can more easily be extracted in predicate form, but less so for texts expressing opinions.

There are many similar approaches to this, but which exhibit different, valuable features. Moawad and Aref (2012) describes a graph reduction process based on a set of heuristic rules, looking at the overlap of between subject-verb-object triplets rather than relying on a training corpus. This method also expressly generates sentences rather than the realisation of a bag of words. Lexical chains have also been used to provide additional semantic information. Balaji, Geetha, and Parthasarathi (2016) describe a bootstrapping approach to learning graph operations that transform the document semantic graph into a summary graph by a process of modifying, replacing, deleting, and adding nodes.

2.2 REPRESENTATIONS FOR SUMMARIZATION

In this section I discuss literature regarding the representation of narratives. An understanding of the cognitive representations formed by humans as they read is highly useful. This provides an indication of the material that readers are more likely to store and recall, and thus include in a summary, and why this is the case. I present research which has analysed the discourse structure of different genres of texts. I believe this can aid the computational modelling of some of the processes that human summarizers perform, and can potentially be detected automatically, leading to improved automatic summarization.

2.2.1 *Cognitive Representations*

Research in the field of psycholinguistics can provide us with some insight as to how we as humans understand and remember information from texts. My work aims to practically apply some of these findings to automatic summarization with the intention of producing summaries which are closer in quality to those that humans can write.

Existing work supports the view that humans construct a network-like mental representation of narrative during reading (Bower and Morrow, 1990; Goldman, Graesser, and Van den Broek, 1999; Graesser, Singer, and Trabasso, 1994; Tapiero, Van den Broek, and Quintana, 2002; Tzeng et al., 2005). Readers transform a text into a representation of its underlying conceptual propositions or events, which generally form the nodes of this mental network. The concepts are linked by the causal connections between events, and the identification of referents to the same entities. In the model of Goldman, Graesser, and Van den Broek (1999), information obtained while reading is used to update the model and change its state. Similarly, Tzeng et al. (2005) views reading comprehension as a cyclical process. Every cycle, a sentence or proposition is read and concepts fluctuate in their activation based on four information sources: current input, residual information from previous input, current episodic text representation and readers background knowledge. Each cycle, concepts are activated and added as nodes to the episodic memory, or strengthened if they already exist.

Fang and Teufel (2014) prototyped the feasibility of implementing the cognitive model of Kintsch and Teun A Van Dijk (1978) for text summarization. Here the authors focused in particular on the memory component, linking textual propositions parsed from a text to units already held in short and long term memory units. Summary generation could then proceed via the selection of text corresponding to the propositions that were most frequently retained in the short term memory component. Fang and Teufel (2014) demonstrated that memory-like components of cognitive models can indeed be simulated and used to generate summaries, albeit with modest results. In addition, the authors admit that they do nothing to account for the world knowledge component required to form 'macropropositions' and abstract away from the input text.

Some of the aforementioned cognitive theories suggest that such a mental network is hierarchical; at the layer above the surface text, readers form connections based on the overlapping arguments of predicates in the text. Above this propositional model, readers create a situational model by identifying the causes and motives which explain why the events and actions of a narrative occurred (Graesser, Singer, and Trabasso, 1994). Other theories suggest a flatter represen-

tation, with a distinct model for each separable element of a narrative. Bower and Morrow (1990) describes mental models as being comprised of two parts. The first describes the characters, their relations and goals, and the second describes the physical settings of the narrative. Goldman, Graesser, and Van den Broek (1999) suggests that internal representations of narratives include models of the settings, persons, objects and instruments that play a role in the drama.

Existing work has considered the classification of the people mentioned in a text for whether they should be included in a summary, and if so, how much background information is required for them. Working in the domain of news, Nenkova, Siddharthan, and K. McKeown (2005) trained classifiers with features such as number of referents, number of times a relative clause or apposition is used for a given person in order to determine whether they are a *major/minor* character (whether they should be mentioned in a summary), and whether they are *hearer old/new* (whether context is required for the character, or a reader is likely to have prior knowledge of them).

Across these different theories of how mental models are structured, the goals or motives of characters remain the most important feature. Character goals have been called the glue that links narrative events into a coherent causal network (Goldman, Graesser, and Van den Broek, 1999). Tapiero, Van den Broek, and Quintana (2002) state that readers use their naïve theories about causality to understand a text and to construct a coherent interpretation of it. The authors go on to discuss how causal connections do not have to be between events that are proximal in the surface text of a narrative, and discuss four different types of causal relation. Of these four types (*physical causality*, *motivation*, *psychological causation* and *enablement*), only the strength of a physical causality relationship is significantly affected by the distance between cause and effect. For example, the link between a person dropping a glass, and that glass smashing on the floor is greatly affected by the distance between those two statements in the text. Relations involving goals or the internal motivations of characters were strongly connected regardless of the distance between their elements.

Goals are the cause of character actions, and actions that more closely related to goals are more likely to be remembered. Studies have revealed that in these representations, readers consider the most significant parts of a story to be the events on the main causal chain (Bower and Morrow, 1990; Tapiero, Van den Broek, and Quintana, 2002). This is a path which connects the most important events of a narrative, starting with the initial setting and ending with the satisfaction of the primary goal. These events can be identified by observing nodes with a high degree, and by the main path that connects the events of a text. Black and Bower (1980) investigated what readers are likely to remember when the semantic content of a narrative is

changed, but its surface realisation and narrative structure are, as far as possible, untouched. They found that theories of story memory that focus on the causal chain of events could be used to better predict the statements that readers would recall than theories of story grammars. Schank and Abelson (1977) looked at the knowledge structures necessary to understand stories, stating that understanding occurs when each action can be seen as a step towards a goal. In this work, causal connections are important, and the more highly linked an event is the more important it is.

Text memory is highly complex, involving many variables in terms of both a narrative and a reader. A knowledgeable reader can draw inferences and connect elements of a text based on prior knowledge, giving them expectations about what can, or is most likely to, occur in a given scenario. Research in psycholinguistics has highlighted the importance of a reader's world knowledge to achieve this. It has been stated that text comprehension improves when the reader has adequate background knowledge, and can not only draw relations between different parts of the text, but also between the text and the real world (Graesser, Singer, and Trabasso, 1994; Hutto, n.d. Tapiero, Van den Broek, and Quintana, 2002).

The integration of knowledge bases are outside the scope of this thesis, as my focus is on the use of discourse structure. However it is useful to be aware that such resources do exist and can be considered in future work. Knowledge bases store complex structured or unstructured information which could help to join together parts of a text that are not obviously related from surface text alone, but are, to human reader, clearly related. There is ongoing work involving the manual creation of knowledge bases, such as Cyc (Lenat, 1995), and ConceptNet (H. Liu and P. Singh, 2004). However, such approaches have been accused of lacking both breadth and depth, being made of ad-hoc relations. Several automatically generated knowledge bases have also been created. Some of these extract relations from natural language texts (Harrington and Clark, 2007), while others exploit existing stores of structured information (Dolan, Vanderwende, and Richardson, 1993; Informatics, 2014). YAGO (Informatics, 2014), for example, contains over 10 million entities and 120 million facts extracted and linked together from Wikipedia, WordNet and GeoNames data. Manual evaluation has shown that this has been performed with over 95% accuracy.

2.2.2 *Discourse Structure*

Research in psycholinguistics tells us about the kinds of cognitive representations that humans create while reading. These studies have indicated the importance of both the motivations of characters, and the causal links between events. Knowledge about the discourse struc-

ture of a given type of text can be used to go some way towards replicating these human processes.

Discourse structure is a term used to describe the way a text is organised. Knowledge about the structure of a text can aid a variety of natural language processing tasks. It can be used in the determination of important content and lead to formation of more coherent summaries. Humans tend to judge the quality of summaries not only in terms of the events they include, but also in terms of referential clarity and coherence (Webber and Joshi, 2012). Here, coherence refers to the ways in which the content of a text is semantically and logically connected in order to convey meaning to the reader. By paying attention to linguistic features beyond individual words and sentences, summarization systems can only improve.

Different types of text have a different structure, and discourse structural information is implicitly used in many approaches to news article summarization, both abstractive and extractive. Reasonable success can be achieved in this domain by focusing on the lead-in, the first few sentences of a news article, which often provides a summary of the content. News articles follow an *inverted pyramid* structure, where the text starts with the most important elements, later providing additional details and any background information or commentary (Pöttker, 2003).

Grosz and Sidner (1986) put forward a theory of discourse structure based on intentions and plans, and how spans of text aggregate into discourse segments to form the linguistic structure of a text. This is made of three interrelated components: the linguistic structure, intentional structure, and attentional state. The linguistic structure describes the sequences of clauses in text, which aggregate into discourse segments. The intentional structure is used to explain the discourse relevant purposes and their interrelations. The authors state here that cue phrases are the most distinguished linguistic means that a speaker has for both indicating the boundary of a discourse segment and to convey information about the purpose of a discourse segment. The third component of this model, the attentional state, represents the focus of a participant (a reader in the case of written text). This is a dynamic record of the properties and relations of objects currently in focus for a participant. The authors find that intentions are key to explaining the structure of a discourse and its coherence, but also that they are the most difficult aspect to identify. This is in part due to the fact that surface text alone may not provide enough indicators.

2.2.2.1 *Rhetorical Structure Theory*

Rhetorical Structure Theory (RST) gives an account of the structure of discourse based on a set of rhetorical relations. RST proposes that all propositional units of text in a coherent document must be connected

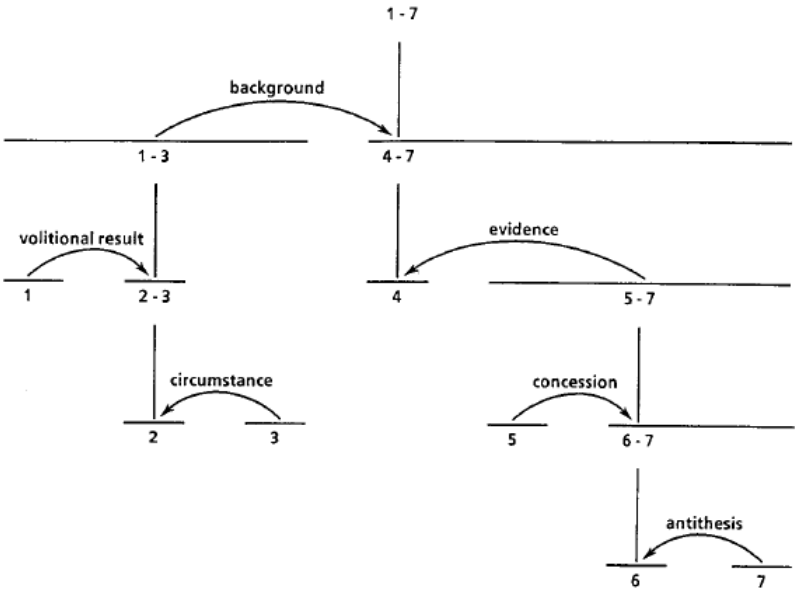


Figure 1: Diagrammatic example of an RST parse taken from Mann and S. A. Thompson (1987).

by some type of rhetorical or discourse relation. First developed by Mann and S. A. Thompson (1987), it is a theory of text parsing to show the discourse relations between spans of text and how they relate to each other in a hierarchical way. Mann and Thompson identify over 20 relations between spans of text such as *Evidence*, *Elaboration* and *Motivation*. However the authors do not claim this list is exhaustive, and it may require extension or modification.

Figure 1 shows an example RST parse for a short seven-sentence news story, demonstrating how spans of text are connected by various relations in a hierarchical structure. Each vertical line descends to a text span identified as a nucleus statement, and the labelled arrows are the relations joining satellite statements to their nuclei.

RST relations are mostly asymmetric: relations such as *Evidence* are defined such that if **A** is evidence for **B**, then **B** is not evidence for **A**. The result of this is that in a pair of text spans joined by a relation, one span often does not make sense without the other. Without the nucleus claim the satellite evidence carries no meaning, suggesting that some parts of the text are more crucial to conveying the text's meaning than others. It is easy to see the importance of nucleus statements in text, if such a statement is removed, the purpose of its associated satellite is unclear. Furthermore, a satellite statement providing *Evidence* can be substituted for alternate evidence and the text will still make sense. The same is not necessarily true of nucleus statements. Building from this, extracting only the nuclei statements in a text and joining them together can potentially give a coherent synopsis of the original text. Two of the relations identified by Mann and Thompson

are multi-nuclear: *Sequence* and *Contrast*, where included spans are equally important. In this work nuclei are arguably the most important spans of text, and using a multi-nuclear relation such as *Sequence* to join these spans together could prove to be a good way to write a summary, as a sequence of the key events of a text.

RST models of discourse capture relations such as *Motivation*. Research in psycholinguistics has highlighted the importance of such elements to story understanding and recall; an important consideration for tasks in natural language processing such as automatic summarization. Equally, applications of RST in summarization have indicated the relative low importance of other types of relation such as *Elaboration*, which convey non-essential information that can be omitted from a summary.

However, there are difficulties associated with creating RST parses of a text. The theory is not constrained enough that only a single valid interpretation can be created for any given text. Even a single human analyst can provide multiple RST parses for a single text. This subjectivity can come from differing opinions on the type of relation to use between spans, or where the boundaries between text spans lie and how they relate to each other on the more global scale of the text. There have been efforts towards automatically parsing texts to conform to Rhetorical Structure Theory (Surdeanu, Hicks, and Valenzuela-Escárcega, 2015), which do produce coherent results. However, it is hard to judge the accuracy of such systems due to the subjective nature of interpreting a text according to RST.

The work of Elson (2012) explored various methods of modelling discourse, with a particular focus on narratives and the ways in which narrative discourse relations can reveal insights about either a genre or a single text. He argued that existing work on the automatic analysis of discourse, such as RST, focuses on expository texts and does not deal with the additional relations that occur in narratives - primarily the plans and goals of characters. Elson (2012) provides a discussion of the benefits and drawbacks of three different descriptive representations for narrative discourse: causal networks, story grammars, and plot units. In a search for a formal representation that could unlock information about structure and content, Elson defined a set of discourse relations specific to narrative. He defined Story Intention Graphs (SIG), which represent three increasingly more abstract representations of the information contained in a text, in a similar manner to Graesser, Singer, and Trabasso (1994). The first of these, the textual layer, represents the utterances of the text in question. The timeline layer represents the propositions underlying these utterances, and the intepretive layer is a representation of the goals, plans, beliefs, and intentions of the characters.

Elson and K. R. McKeown (2007) and Elson (2012) present SCHEHERAZADE, a discourse modelling platform for experiments

and applications which build on symbolic representations of narratives. These narratives can express a diverse range of styles and come from a multitude of genres. Furthermore, the platform can be used for the encoding and analysis of both individual stories and corpora. The key objects stored by SCHEHERAZADE are the abstract structural elements of a narrative, world knowledge about the domain of the narrative, and the actual story content represented as the actions taken by particular characters. This structure allows for the modelling of semantics such as timelines, states, events, characters and goals, and can detect thematic patterns in the story. The aim of such a platform is to enable higher-level tools to perform different analyses on the narratives modelled and encoded by the system. Fundamentally however, these texts still need to be manually annotated under the SCHEHERAZADE platform.

2.2.2.2 *Story Structure*

Black and Bower (1980) describe a narrative as a selective rendering of a continuum of events, where the author omits non-essential or predictable elements under the assumption that the reader will have the requisite knowledge to fill in the full scenario. According to this basis, the difficulty of a text can be judged; how much specialised information does the reader require? This makes stories, such as folktales, fairytales, and myths of particular interest to my research. These genres are, for the most part, written to be accessible to anyone and require comparatively little world knowledge. This allows me to focus on the use of discourse structure to aid summarization without having too much concern for the lack of world knowledge.

As a genre stories have a specific structure, different from that of other domains, such as news articles. The events that comprise a story can be shuffled about and retold in a new order, giving a different narrative, while the story itself remains the same. Retellings may also omit details, or exchange elements or characters for their equivalents which fulfil the same purpose. Stories are robust to such changes, with people able to recognize the elements required to tell a story. As Butt the Hoopoe tells Haroun on the Ocean of the Streams of Story "Any story worth its salt can handle a little shaking up" (Rushdie, 1991, p. 78).

Story structure concerns the arrangement and relations between just the events in a narrative, not *how* it is told. This is the important distinction between story and plot. Consider[1]:

STORY: The king died and then the queen died.

PLOT: The king died, and then the queen died of grief.

1 Example taken from Forster (2005, p. 87)

A story is the sequence of events simply expressing what happened and the order it happened in. The plot belongs to the class of discourse elements, which cover *how* a narrative is told.

One approach to modelling the structure of stories has been story grammars. Story grammars, such as those proposed by Rumelhart (1975) and Thorndyke (1977), attempt to specify the units of stories and describe how they fit together. These grammars are comprised of rewrite rules to fit the structure of a given story. For example the first rule of Thorndyke's grammar is given below:

$$\text{Story} \rightarrow \text{Setting} + \text{Theme} + \text{Plot} + \text{Resolution}$$

Each of these elements have further rewrite rules. Most importantly, the plot can be comprised of any number of episodes, which can each be comprised of any number of attempts before the outcome of the episode is reached. With this availability of recursion, the grammars can fit to many different stories. However, these grammars remain too high level for the purpose of summarization. While generalisation is desirable, so too is some detail about the specifics of a story. Furthermore, Black and Bower (1980) found that story grammars could not be used to significantly predict the statements that people would recall from a short story. The main events on the causal chain, however, would.

Labov (2013) refined his earlier work detailing the structure of a narrative. This later version was comprised of six elements: the abstract, the orientation, the complicating action, the evaluation, the resolution, and the coda. Ouyang and K. McKeown (2014) demonstrated the possibility of automatically detecting information about discourse structure by predicting the complicating action element of Labov's theory. However, this was the only element they attempted to detect, and Labov's work just considers personal narratives.

Another approach to modelling story structure is *plot units*, proposed by Lehnert (1981) in order to aid summarization. Plot units are comprised of affect states and the tensions between characters. In turn, the affect states are made up of a single character, an event, and the affect state of the character (either negative, neutral, or positive). This approach required large amounts of knowledge engineering, however later work demonstrated the feasibility of constructing plot units automatically (Goyal, Riloff, and Iii, 2013).

Anthropological studies of stories have resulted in different detailed analyses. Lévi-Strauss (1955) performed a structural study of myths, to try and answer why myths from different cultures all over the world share so many similarities. He broke down myths into their constituent elements, or mythemes, so that different myths could be compared in terms of their structure rather than their content. Key to his structural analysis was the idea of binary oppositions, funda-

mental concepts that occur in all cultures. However, the choice of mythemes, how they are interpreted, and what the binary oppositions are make this process highly subjective. S. Thompson (1989) similarly analysed stories from across the world, resulting in a highly detailed index of the motifs that occur in each of them.

The structural analyses of stories discussed so far have been at an abstract level, such as story grammars, or very fine grain analyses which are only suitable for the classification and comparison of stories. My main interest is in structural information that places expectations on a story, and can be used computationally to get beneath the simple surface meaning of a text. To this end, the next section provides a detailed description of Vladimir Propp's *Morphology of the Folktale* (Propp, 2015).

2.2.2.3 *The Morphology of the Folktale*

Vladimir Propp's *Morphology of the Folktale* decomposes folktales into elementary components, and describes how these pieces fit together to form a given tale. Over a set of 100 folktales, a subset of Afanasyev's collection of Russian Folktales, Propp identified five categories of elements which he claims define a tale as a whole. A brief description of these categories is given below.

CHARACTER FUNCTIONS: These are a sequence of 31character-based functions, or narratemes. They are the narrative units of a tale, which each describe an abstracted event and are performed by the *dramatis personae* of the tale.

CONJUNCTIVE ELEMENTS: These elements describe how information is transferred from character to character. When successive functions are performed by different characters, the latter character must somehow be informed of everything that has occurred up until that point. This may for example occur when characters act *ex machina*, are all knowing, or overhear a dialogue between others.

CHARACTER MOTIVATIONS: The goals and aims of characters, which drive their actions.

CHARACTER APPEARANCE: This describes the method by which a character is first introduced to the story, rather than their physical appearance. For example a character may be encountered accidentally, or arrive suddenly.

ATTRIBUTIVE ELEMENTS: These describe the specific qualities, including physical appearance, of each character, for example their age or peculiarities of their appearance.

Key amongst these categories is the sequence of 31 character-based narrative units, that make up the actions of the story[2]. Each of these functions provide a generalised description of a key event in a tale, and the characters that it necessarily involves. Propp repeatedly stresses the importance of these character functions over other elements of the tale such as the characters who perform them. Table 1 gives the canonical ordering of these functions, with their designation and brief definition.

While there is a canonical order to these functions, any given function does not have to be present in a particular instance of a tale. In addition, many of these functions are logically paired, such as *struggle/victory*, *difficult task/solution*, and *pursuit/rescue*. Propp does however make one key restriction on what has to be be present within a tale; a tale necessarily has to include either an instance of *villainy* or *lack* which provides the motivation for the subsequent actions of the protagonist.

A single tale could hold multiple sequences of these character functions, either sequentially or embedded within one another. Propp defines a *move* as "any development from villainy (A) or a lack (a), through intermediary functions to marriage (W*), or to other functions employed as a denouement" [3] (Propp, 2015, p. 92). With this definition, every new instance of villainy or lack creates a new move.

Propp describes these functions with a series of short examples, often providing the indicator words for the presence of a function. That is, most of Propp's functions are detected via long and highly varied lists of cue words. Other functions, such as the violation of an interdiction, are highly dependent on the form that the preceding imperative command (the interdiction function) takes.

Each character function has multiple subtypes, detailing the specific forms that it may take. Consider the testing of the hero by a donor (*donor tests hero*). Propp describes 10 fine-grained ways in which this function might occur. These cover forms such as: *the testing of the hero*, *requesting mercy*, or *requesting the division of property*.

The final element of Propp's morphology requiring discussion here is that of the roles of the dramatis personae. Propp concluded that every character in a tale could be resolved into one of seven types according to their purpose. The Hero (who defeats the villain or resolves the lack), the Villain (the character who creates the main obstacle for the hero), the Donor (the character who may test the hero and gives them a magical agent), the Dispatcher (the character to send the hero on their quest), the Helper (an often magical agent who

2 Henceforth I shall refer to this set of character-based narrative units as *character functions* or *Propp functions*.

3 Propp's analysis of folktales implicitly refers to heroes as male, and being married off to a princess is a common form of the reward function that signals the end of a tale. These gender-specific roles do not reflect my own views, or necessarily Propp's own views. His morphology is a product of the corpus he analysed.

Table 1: The strict ordering of Propp's character based functions.

1. β Absentation	17. J Branding
2. γ Interdiction	18. I Victory
3. δ Violation	19. K Liquidation
4. ϵ Reconnaissance	20. ↓ Return
5. ζ Delivery	21. Pr Pursuit
6. η Trickery	22. Rs Rescue
7. θ Complicity	23. O Unrecognized arrival
8. A/a Villainy/Lack	24. L Unfounded claims
9. B Mediation	25. M Difficult task
10. C Beginning counteraction	26. N Solution
11. ↑ Departure	27. Q Recognition
12. D Donor tests hero	28. Ex Exposure
13. E Hero reacts	29. T Transfiguration
14. F Receipt of magical agent	30. U Punishment
15. G Transference	31. W Reward
16. H Struggle	

assists the hero), the Princess/Prize (the marriage to this character is often the goal of the hero), and the False Hero (who takes credit for the hero's actions and seeks the reward for themselves). Certain character functions are logically connected and grouped into *spheres of action*. Each of these spheres corresponds to one of the character types, and specifies the character functions that a character of a given type is involved in. However, one character in a tale may fulfil the role and the actions performed by several character types. For instance, it is not uncommon for the villain character to unwillingly fulfil the role of the donor too, accidentally leaving a magical agent behind only to be found by the hero.

APPLICATIONS OF PROPP'S MORPHOLOGY

Propp's description of a folktale as being comprised of a sequence of simple units, coupled with a description of how these units can be composed to create new tales, has led to its use in the field of computational story generation. Gervás (2014) gives an overview of Propp's semi-formal generation procedure as well as discussing a computational approach to generate instances of Russian folktales. Differing generation options are also considered here, as well as their evaluation by metrics inspired by Propp's morphology and his available annotations. In addition, there has been other work in story gener-

ation either inspired by Propp's work (Turner, 1993) or using it in combination with other methods such as case-based reasoning (Díaz-Agudo, Gervás, and Peinado, 2004).

Previous work (Valls-Vargas, Ontanón, and J. Zhu, 2013) has also attempted the identification of character roles according to Propp's morphology. This approach primarily considers the identification of characters fulfilling the roles of the hero and the villain, leaving the other roles unspecified. Using manually annotated data, Valls-Vargas et al. create role-action matrices, which specify the characters who are the subject and object for each verb in the text. These are used with a genetic algorithm to learn the actions that each type of character typically perform, as well as the actions that different types of character perform to each other. Only considering the assignment of the hero and the villain (the remaining character roles are grouped into a category for 'other'), this method achieves 78.99% classification accuracy over a small dataset of 8 Russian folktales.

Bod et al. (2012) present empirical work on the annotation of three single-move Russian folktales by a group of human annotators. The aim of this was to examine the objectivity and reproducibility of Propp's morphology. With limited training, participants of the first study were asked to assign character roles to the story characters as well as annotating three stories with character functions. Then in a subsequent study a different set of participants were given the roles of the dramatis personae and asked to annotate the same three tales. The authors found that providing the character role assignments had a large impact on assignment of character functions to a tale, but that there was low inter-annotator agreement in both studies. In addition to low agreement between participants, the authors found that none of the human annotations matched Propp's own. They claim that this is in part due to the vagueness of some of Propp's function descriptions. I would also consider the limited training that participants received to be an important factor, and that participants of the first study were trained on an example constructed by the authors, rather than an existing folktale. This research was continued by Fisseni, Kurji, and Löwe (2014) which showed that with significantly more training, inter-annotator agreement was much higher and that annotators could reproduce Propp's own function annotations.

Using a model merging algorithm created with merge rules derived from Propp's morphology, Finlayson (2016) demonstrates the feasibility of computationally learning a theory of narrative structure. Utilising a corpus of deeply annotated Russian folktales (Finlayson, 2015), the outlined method accurately learns to capture events corresponding to some of the key character functions in Propp's morphology, most notably *vilainy/lack*, *struggle/victory* and *reward*.

2.3 EVALUATING SUMMARIZATION

Approaches to summarization are evaluated through their product; the summaries that they generate. There are a range of metrics used to evaluate the quality of generated summaries, both automatic and manual. Automatic approaches to evaluation allow for relatively quick results, and on a large scale. However, these are largely restricted to analysing the surface text of a summary, rather than examining the meaning beneath it. Manual evaluations performed by human assessors typically require some form of ranking or judgement about the quality of a summary. These can be both time-consuming and subjective depending on the experimental setup. Nevertheless, currently only human-based approaches can truly judge if a summary is coherent and whether it covers the salient points of a text. Human evaluation is especially desirable for abstractive summarization, so that the semantic equivalence of summary content can be intelligently considered.

The types of evaluation measures can be broadly split into extrinsic task-based evaluations, and intrinsic measures which look at the content and quality of a produced text. Extrinsic evaluations consider factors such as the ability of a reader to use a generated summary in order to answer questions. Evaluating text quality requires the judgement of human assessors, while efforts to evaluate the content of a summary can be attempted automatically by using a reference summary as a point of comparison. In this section I describe the main intrinsic evaluation approaches in addition to several novel methods which consider different dimensions of evaluation.

2.3.1 *ROUGE*

Of all the methods used for the evaluation of automatically generated summaries, ROUGE, the *Recall-Oriented Understudy for Gisting Evaluation* is most common. ROUGE is an automatic evaluation metric which was first introduced in 2004; it evaluates a generated summary by comparing it a set of one or more 'ideal' reference summaries, usually written by a human (C.-Y. Lin, 2004b). This comparison is accomplished by looking at the percentage of units that occur in both the testing summary and the reference summary. These units are sequences of one or more words, possibly with a gap between them. There are multiple variants of ROUGE, which I shall briefly summarize, as to the best of my knowledge this is the most widely used summarization evaluation metric.

ROUGE-N: This considers the n-gram overlap between a generated summary and a set of one or more reference summaries, where n is the size of the n-gram to consider. This variant does not

account for the order in which the n-grams appear, so a small n leads only to a bag-of-words comparison.

ROUGE-L: This looks at the longest common subsequence of words between a pair of summaries at the sentence level. Each sentence in a reference-summary is compared to each sentence in the generated summary to find the longest sequence of words that they have in common. The intuition behind this is that longer common subsequences between two summary sentences indicate a higher degree of similarity between the two summaries. However, the words do not need to be strictly contiguous. The words of the sequence must follow the same order between both summary sentences, however the generated-summary sentence may contain arbitrary length gaps of other words.

ROUGE-W: This builds on ROUGE-L by weighting it in favour of word sequences that provide a match between consecutive words.

ROUGE-S: This is a skip-bigram co-occurrence measure; a skip-bigram is a pair of words in their sentence order with an arbitrary gap between them. ROUGE-S looks at the total number of these as a ratio of the number of possible combinations of skip-bigrams for a given reference sentence. This method will give low scores if the word order between two sentences being compared is substantially different.

ROUGE-SU: This accounts for the fact that ROUGE-S requires the word ordering to be similar between two sentences being compared by only looking at unigrams.

Despite the large number of ROUGE variants, often only a few are chosen to indicate the quality of a given summarization method. Commonly, ROUGE-N is used, where $n = 1$ and only a single reference summary is present. This may indicate a misuse of ROUGE, as it does not promote theoretically well grounded approaches to summarization. Instead it incentivises methods which are trained over large corpora of data to faithfully reproduce human summaries, or at least the key words contained in them with no regard to word ordering or grammatical structure.

Additionally, ROUGE is very rigid in that it cannot account for simple lexical differences between two summaries. By taking an arbitrary summary A, a semantically equivalent summary B can be created by simply substituting each word for a synonym where possible. ROUGE will score B poorly against the reference summary A.

C.-Y. Lin (2004a) suggests that inter-human agreement is usually very high on evaluating single-document summaries, arguing that

humans can normally agree on what a good summary contains and so ROUGE provides a good way of finding overlaps between automatically generated and reference summaries. Rankel, Conroy, and Schlesinger (2012) however, shows that the correlation with human judgements is not so high. Ng and Abrecht (2015) find two main issues with ROUGE. Firstly, it favours lexical similarities, making it unsuitable for the evaluation of highly paraphrased or abstractive summaries. Secondly, ROUGE neither rewards summaries with high fluency and readability, nor penalises those that do not display these qualities.

2.3.2 *Word Embeddings for ROUGE*

More recently, an extension to ROUGE has been proposed which aims to address one of its main flaws. Ng and Abrecht (2015) use word embeddings (Bengio et al., 2003) to compute the semantic similarity of words used in summaries, as opposed to only looking at the lexical overlaps as in ROUGE. This approach uses pre-trained word embeddings from word2vec (Mikolov, Yih, and Zweig, 2013), where the mappings of two words into a space are closer together the more semantically similar they are.

The authors found that this approach had a higher correlation with human judgement scores than the original ROUGE metric. However, it was largely evaluated on the summaries produced by extractive summarization systems. As such, the summaries being evaluated were each formed as a subset of the words in the input documents. This gives little scope for originality, or testing whether this method can accurately account for lexical differences between summaries. Additionally, this method still suffers from the other flaws of ROUGE. Primarily, it does not take the fluency or readability of generated summaries into account. A further point is that this specific work relies on word embeddings which were trained on a news dataset. This makes it less suitable for the evaluation of documents in other domains, where words may carry different meanings and so differ in their semantic similarity.

2.3.3 *Pyramid*

After ROUGE, Pyramid is one of the more common evaluation methods. Pyramid is based around the idea of identifying units of summary-worthy information from a pool of human summaries, and finding which of these occur in the generated summary (Nenkova and Passonneau, 2004).

To begin, each reference summary is marked with *Summarization Content Units* (SCUs), spans of text which are no larger than a clause. SCUs are weighted based on how many reference summaries they

occur in, and are subsequently partitioned by their weighting into a pyramid. The peak of the pyramid contains the few SCUs which occur in many reference summaries, while the base contains the many SCUs which occur in only a few reference summaries each. Then, according to the Pyramid metric, the optimal summary contains all of the SCUs from the top level of the pyramid, and additionally contain the SCUs from each successive tier depending on the length of the summary. A summary is then scored as the ratio of the sum of the weights given to its SCUs over the maximum sum of the weights of a summary containing the same number of SCUs. A similar method is described in Hovy, C.-Y. Lin, and L. Zhou (2005)

Creating a pyramid to use for the evaluation of new summaries requires an initial set of human summaries constructed from the same document set, as well as the annotation of SCUs which different annotators may have disagreements on. The authors admit that this is a laborious process which would ideally use some degree of automation. Furthermore, semantic content units are treated as independent from each other, so their ordering and relations between each other are not taken into account.

2.3.4 *SERA*

SERA, the Summarization Evaluation by Relevance Analysis metric was proposed as an evaluation metric for scientific articles after analyses found that ROUGE was not appropriate for this domain (Cohan and Goharian, 2016). It was found that correlations between ROUGE and Pyramid scores were weak for scientific articles, and that there was a large variation in correlation between different ROUGE variants.

SERA is based on the premise that concepts take meaning from the context that they are in, and that related concepts co-occur frequently. This metric is designed for multi-document summarization, treating the evaluation as an information retrieval task. Both a generated summary and a human reference summary are treated as search queries against the corpus of documents that was used to form the summaries. The ranking of retrieved documents are compared, where a more similar ranking leads to a higher evaluation score. The authors state more similar rankings indicate higher content quality, as this method is based on semantic relatedness rather than lexical overlap

While this method was intended for multi-document summarization, it could be adapted to single-document summarization, by treating each sentence or paragraph as a separate document.

2.3.5 *Human Ratings*

Human judges have been used to evaluate summaries according to a range of criteria. At the Text Analysis Conference, humans have

been asked to evaluate summaries in terms of their *readability* and *responsiveness*. Readability was assessed by judgements on the fluency, grammaticality, non-redundancy and overall coherence of a document. Assessors were asked to judge responsiveness according how well they thought a summary presented the requested information and how valuable the summary was to the reader.

Bhartiya and A. Singh (2014) discuss assessment criteria they devised to evaluate their own approach to summarization. Assessors are asked to rate generated summaries over a scale of 1-5 on five different aspects; Information Content, Grammatical Correctness, Abstraction, Expressiveness and Excess or Unnecessary Information. The ratings given to generated summaries are then compared against the ratings that participants gave to human-written summaries.

Evaluation according to human judgements does however have several flaws. The process can be time consuming and subjective, and the agreement between assessors can be low. Additionally, human ratings cannot easily distinguish between different types of text defect such as the fluency of a single sentence against the coherence of an entire text. This can result in judges disagreeing over ratings and inconsistencies from a single judge. However these methods do allow for a range of dimensions of a summary to be evaluated.

2.3.6 *Reading Times*

Zarrieß, Loth, and Schlangen (2015) investigated the use of reading times as a more objective human based evaluation metric. Longer reading times indicate that a reader has greater difficulty in reading a text and that the text may have issues such as complex grammatical features or inconsistencies. To test this, they developed Mouse Contingent Reading (MCR).

In this method, a covered text is displayed on screen to a user. Users can only view the text of a given sentence while the mouse is hovering over it. With this setup, it is possible to test the reading time for each sentence, as well as any transitions a reader makes back and forth through the text.

Different versions of the same text were then generated with defects in the word order, syntax or incorrect referring expressions. In their studies, participants were presented with a random set of texts, either with or without defects. Reading times and scan path under MCR were noted, and participants were also asked to rate the fluency and clarity of each text after reading it. The authors then built regression models using only human rankings, only sentence reading times, a combination of the two, and whole-text reading times to predict the quality of text. They found that a combination of the reading time metrics identified generated texts to a high degree of accuracy, compared to human ratings which could not distinguish different qual-

ities of generated text well. However the authors found that error free texts were not associated with faster reading times, and hypothesisised two reasons for this. First, participants knew they would have to rate a text after reading it. Second, readers may have rushed through clearly defective texts, skewing expected differences in reading times, as they were unable to clearly understand the meaning of the defective texts and so invested less time on them.

This approach demonstrates a more objective way of using human evaluation, which does not involve the opinions of individuals. In terms of time however, it is still a more costly approach than other automatic evaluation methods which can even be integrated into the development of summarization systems.

2.3.7 *Reading Comprehension*

Evaluating a summary based on how successfully it can be used to answer comprehension questions is a form of task-based extrinsic evaluation. This is again a more objective measure than human judgements of summaries, as it can be posed as a two-alternative forced choice of whether or not a summary can be used to answer a question.

Morris, Kasper, and Adams (1992) show how this approach can be used to compare summaries generated under different conditions from the same texts. Four different settings were tested: extractive summaries 20% of the length of the original text, extractive summaries 30% of the length of the original text, a manually written abstract, and no text at all. Participants were asked to answer multiple choice questions about the original text, having been given one of the four summary options. The setting without any summary text was used to see how many questions simply contained the answer within the question text.

This approach has the benefits of human involvement in the evaluation, while also removing the subjective nature by being a forced choice from predefined options. From the participants' point of view, this is also on the quicker end of human based evaluation methods.

2.4 SUMMARY

I began this chapter with a review of some of the many different approaches to summarization that have been explored since the initial work of Luhn (1958). However, regardless of the approach taken, the vast majority of work in summarization has focused on the domain of news articles. Many of these approaches to news article summarization implicitly use discourse structural information; extracting the first few sentences of an article, which already serve as a reasonable summary. This sets a high baseline for summarization of the news

domain, especially as it has been found that assessors take linguistic quality into account even when told not to (Conroy and Dang, 2008).

While the field of automatic summarization appears to be primarily focused on the summarization of news articles, there has been work that has considered other domains. These cover genres such as email threads (Zajic, Dorr, and Schwartz, 2004), biomedical literature (Ling et al., 2007), scientific articles (Luhn, 1958; Edmundson, 1969), including citation based summarization (Qazvinian et al., 2013), and patent documents (Tseng, C.-J. Lin, and Y.-I. Lin, 2007).

Steinberger and Jezek (2004) suggest that there are four categories of approaches to automatic text summarization: heuristic approaches that use techniques such as sentence position, corpus-based approaches using techniques such as TF-IDF, approaches which take discourse structure into account such as lexical chain methods, and knowledge rich approaches that make use of domain knowledge. However I believe that lexical chains barely scratch the surface of discourse structural information, and what they call knowledge-rich approaches actually use knowledge of a domain to anticipate the structure of a document and make best use of it.

Of most interest to me are the approaches which use discourse structure, as the knowledge gained through this structural information lets us progress towards understanding causal connections in the way that psycholinguistics tells us that humans do when reading. Graph-based approaches to summarization exhibit features closest to the mental models created by humans. They create a propositional text-base of concepts that move away from the surface text of a document, dropping elements such as tense and aspect to focus on the events of a text. Some of these, such as TextRank (Mihalcea and Tarau, 2004), use information from the entirety of a document to provide support for a given sentence.

While many of the discussed approaches to summarization have limitations, their methods all contain ideas of value. Variable compression rate in the generation of summaries, as seen in Radev, Jing, et al. (2004), is desirable to cater to a range of needs. Both the LSA approach of Steinberger and Jezek (2004) and lexical chain approaches capture the fact that some concepts that have little individual significance are part of an idea expressed lexically in multiple different ways. Tran, Alrifai, and Herder (2015) expresses the idea of influential events for timeline summarization, similar to the idea of events on the main causal chain. Coreference information has also been used in summarization, sometimes manually (Gerani et al., 2014), and should intuitively provide benefit. This is a computationally hard task, but something that humans perform automatically while reading in order to understand the actors involved in each event.

Generally, automatic evaluation methods can only look at the surface text of a summary, and so they have to be compared to a suppos-

edly ideal human summary. ROUGE is most commonly used, in part due to the fact that its speed and ready implementation make it desirable. Two of the most common methods, ROUGE and Pyramid do not take the ordering of summary content into account. Depending on the variant of ROUGE being used even the word order matters little. With Pyramid, the semantic content units are treated independently and so can appear in any order in a summary being evaluated. As a result, summarization systems can be trained to achieve their optimal score against these metrics, but produce summaries that are not particularly useful or cohesive for their intended human audience. Furthermore, automatic evaluation metrics are unable to accurately predict how well automatic summaries compare to human summaries. Rankel, Conroy, Slud, et al. (2011) find that the performance of summarization systems according to different evaluation metrics varies significantly across different sets of documents used at TAC, and that they can achieve unrealistically high results.

Human judgement evaluation methods are expensive to perform and there are issues with different judges having different opinions on a single summary, as well as the possibility of humans giving inconsistent results when evaluating a set of summaries. However human metrics do allow judges to examine the fluency of summaries. It is hard to automatically judge the fluency or readability of a summary, which humans can comment on during the evaluation process with little extra effort. This can even be done via reading times and eye tracking, to provide a more normative result. But it is possible that, depending on the purpose of a generated summary, fluency is not an important criteria. If computational summarization is just used to aid human summarizers and partially automate the summarization process, then issues of fluency and poor grammaticality can be corrected manually.

3

PRELIMINARY STUDIES

In this chapter, I provide a description of two two small-scale studies which have influenced my methodology. The first of these studies examines automatic coreference resolution systems and their evaluation metrics. This was performed in order to assess how suitable such systems are for tasks that require accurate knowledge about the entities of a text, and how reflective their evaluation metrics are of this performance. The second of these studies investigates the most widely used summarization evaluation metric, ROUGE, and how appropriate it is for the evaluation of this work.

3.1 THE EFFECT OF COREFERENCE RESOLUTION ON DOCUMENT UNDERSTANDING

Coreference resolution is the task of grouping together the expressions that refer to the same entities. This is far easier for human readers than for automatic systems, as humans can often determine the coreferentiality of mentions from just the mentions alone and some additional context when necessary (Morrow, 1985; Sachan, Hovy, and Xing, 2015). Consider:

Jon went to the shop. He bought a drink.

In the above, *He* unambiguously refers to *Jon* but the situation becomes more complex in longer texts as more characters are introduced and mentions may occur further away from their referent.

Coreference resolution is an important preprocessing step in tasks such as automatic summmarization, which require a level of document understanding. It aids a deeper level of comprehension by providing the ability to recognise which characters perform which actions. However, in practice automatic coreference resolution systems do not appear to perform as well as their evaluation metrics suggest.

End-to-end summarization contains multiple processing steps, each potentially introducing errors; and for abstractive summarization, high quality coreference information may be crucial. The exact contribution of coreference information required is unclear, as is the level of accuracy required. As the evaluation of summaries can be expensive and highly subjective, a more objective measure of evaluating the importance of coreference information is required. While intrinsic coreference metrics are available, it is unclear how well they relate to overall task performance.

In this section, I describe a small-scale study[1] which examines the effect of automatic coreference resolution on semantic understanding. I test the extent to which coreference information aids people in a closely related task: answering questions about short stories. These questions were designed to cover the critical aspects of the stories: those that must be understood in order to create a summary. The aim of this was not to find the best performing coreference resolution system, but to see whether such systems can be reliably used in automatic summarization without contributing significant inaccuracy. This study additionally enables an examination of the utility of intrinsic coreference evaluation metrics.

3.1.1 *Metrics*

There are three standard intrinsic evaluation metrics for coreference resolution systems: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and the entity-based CEAF variant, $CEAF_e$ (Luo, 2005). J. Cai and Strube (2010) provide a detailed description of all three metrics along with their relative strengths and weaknesses.

The MUC metric handles the scoring of coreference resolution systems by looking at the number of links that would need to be added or removed from a system response in order to replicate the set of gold standard entities and their mentions. As this metric scores systems in terms of links between mentions it should not be used on datasets with singleton entities, entities which only have a single mention in a text. By definition, singleton entities have no links to other mentions, and so MUC does not credit systems for separating singleton entities from other coreference chains.

B^3 takes singletons into account, by calculating precision and recall scores for each mention and then averaging over all mentions. Unlike MUC, this metric does not consider all types of errors to be equal, and some are penalised more harshly than others. B^3 does however assume that both the gold standard data and the system response are comprised of a matching sets of mentions (Stoyanov et al., 2009), which is not the case when the system is automatically identifying mentions as well as clustering them into co-referent entities.

CEAF uses the best alignment of system to gold responses in order to calculate precision and recall in terms of either mentions or entities. The reason for this is that Luo (2005) finds the results of both MUC and B^3 to be counter-intuitive as entities are used more than once in their calculation. However CEAF weights the alignment of all entities equally, regardless of the size of the coreference chain.

1 The contents of this section are an expanded version of work which has previously appeared in Droog-Hayes (2017).

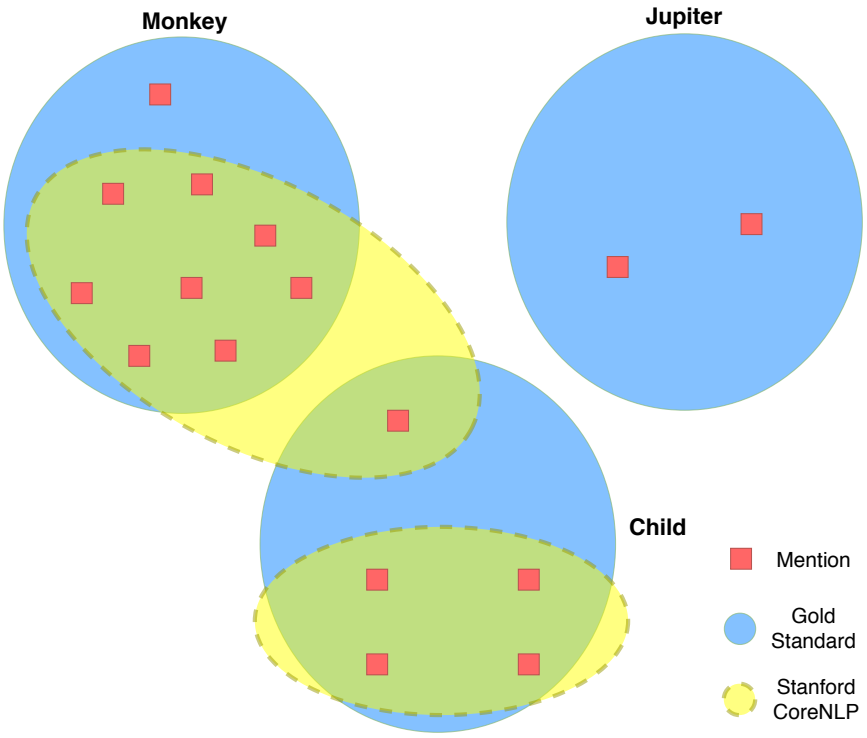


Figure 2: Coreferent mentions of entities for an example story according to gold standard annotations and CoreNLP annotations.

3.1.2 *Motivation*

It is desirable to use coreference resolution tools in automatic summarization. However, in reality, these systems do not seem to perform as well as their evaluation metrics would indicate.

Figure 2 shows the entities and their mentions from an example story used in this study. According to the gold standard manual annotations, the story contains three entities: two mentions referring to *Jupiter*, nine mentions referring to *Monkey*, and five mentions referring to *Child*. According to the output of the Stanford CoreNLP coreference resolution system (H. Lee et al., 2013; Manning et al., 2014), this story only contains two entities. One contains a subset of the mentions referring to *Child*, and the other refers to a combination of the remaining child element and a subset of the mentions referring to *Monkey*. No mentions for *Jupiter* are detected. This automatic grouping of entities could potentially be quite damaging to a human reader's understanding of the text. In particular, the merging of mentions from two distinct entities could lead to misunderstandings about the actions being performed by and to different characters.

While the output of CoreNLP may be damaging to document understanding, this is not necessarily reflected by coreference resolution evaluation metrics. MUC, B^3, and $CEAF_e$ evaluate according to different formulae. However, they all consider the mentions of an entity

as a chain, and evaluate in terms of the number of links that are different between the system output and the gold standard data. In Figure 2, one mention of *Monkey* is missing, and one incorrect link to a mention of *Child* should be removed and attached to the *Child* entity instead. Finally, while neither mention of *Jupiter* is detected, this only corresponds to a single missing link for this undetected entity.

This CoreNLP output results in the following F1 evaluation scores: MUC: 86.95%, B^3: 74.61%, and $CEAF_e$: 71.11%. These scores do not appear to be indicative of the effect that the inaccuracies of automatic coreference resolution systems have on document understanding, which motivated a further investigation into their efficacy for tasks involving semantic understanding.

3.1.3 *Approach*

To examine the impact of potentially inaccurate coreference information on document understanding, I examined the extent to which two different automatic coreference resolution systems could aid a question answering task. This method shares similarities with the reading comprehension methods of evaluating summaries discussed in Chapter 2.

3.1.3.1 *Materials*

Two automatic coreference resolution systems were used in this study: Stanford's CoreNLP system (H. Lee et al., 2013; Manning et al., 2014), and the Illinois coreference resolution system (Peng, Chang, and Roth, 2015). The CoreNLP coreference resolution system is a multi-pass sieve of deterministic coreference models, applied one by one from highest precision to lowest precision. The Illinois coreference resolution system is a pairwise classification model, deciding if mentions belong to the same equivalence class, trained on the Automatic Content Extraction datasets (Doddington et al., 2004).

A set of 10 stories were selected from a collection of Aesop's fables. These short self-contained stories were chosen as opposed to segmenting a longer text, as separating character mentions could impair the success of the coreference resolution systems. As many of Aesop's fables contain anthropomorphic animals, half of the selected stories specifically involve only human characters. This split was enforced to show that the automatic coreference resolution systems being tested are not disadvantaged by this type of data; they are often trained on corpora from sources such as news articles, magazines and web blogs (Peng, Chang, and Roth, 2015), which do not usually contain animals with human characteristics. The stories involving only humans contained 96 noun phrase mentions and those involving animals contained 118 noun phrase mentions.

A first set of participants were asked to produce questions and corresponding answers regarding the summary-worthy aspects of each story. The answers given in this step were taken as the ground truth values that the results of the next step should be compared to, where a second set of participants answered these questions with specific coreference information. This question answering task was performed on graphical representations of the text as opposed to the original plain text. By showing participants a fixed meaning representation of the text, this alleviates the possibility of participants giving a range of answers to the questions due to differing interpretations of a story's meaning. Abstract Meaning Representation (AMR) was used as the semantic representation language for this task. Tree-like visualisations of manual AMR parses of these stories were the main component of the graphical representation, along with PropBank (Palmer, Gildea, and Kingsbury, 2005) edge labels and coreference information.

3.1.3.2 *Method*

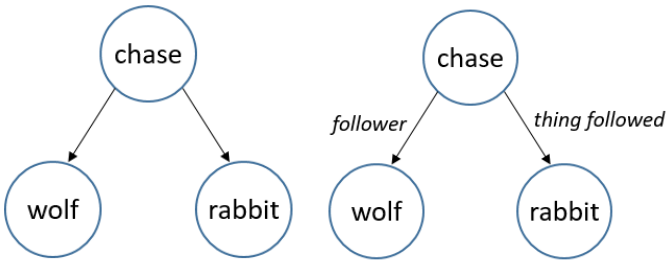
The questions and answers given by the first set of participants were combined for use in the rest of the study according to the following rules: Questions that could not be explicitly answered from the text, such as most *why* questions, were removed to avoid any ambiguity as to the correct answer. Repeating occurrences of the same question were removed, as different participants very often provided near identical questions. Finally, the wording of questions was modified to match the source text as closely as possible. As an example:

TEXT EXCERPT: The wolf chased the rabbit.

RESPONDENT QUESTION: Who ran after the rabbit?

For the above, the question text would be modified to *Who chased the rabbit?* This was carried out to make the questions as unambiguous as possible, given the graphical representation presented to the second set of participants. This process led to a final set of 42 questions. These questions and their answers under each coreference setting are given in Appendix A for reference.

Manual AMR parses for each text were then created for the construction of tree-like graphical representations of each story. These differed slightly from standard AMR parses in that intra-sentential mentions of a character were not resolved to the same entity, as this would act as manual coreference resolution. It was ensured that exactly the same number of character mentions occurred in the parses as in the original source text of each story, so that the outputs of automatic coreference resolution systems could be correctly annotated on the visualisations. For AMR parses of sentences containing coordinating conjunctions, the subject of the first clause was explicitly linked



(a) Without edge labels. (b) With edge labels.

Figure 3: Graphical representation for the text '*The wolf chases the rabbit*', illustrating the need for edge labels.

to the second clause. For humans reading plain text, it is self-evident that the subject is the same in both clauses of such sentences, but this may not be the case when reading AMR parses and so the link is added to ensure that related questions can be answered without difficulty. This was also performed as these conjunctions join two verb phrases with only a single subject, so there is no second noun phrase present for automatic coreference resolution systems to try and detect, let alone correctly resolve.

The graphical representation of a given story consisted of a force directed graph with each sentence's AMR parse. AMR parses do not show the PropBank frame roles that describe the links between two AMR entities. These were parsed onto the visualisations so that hovering on a link between two nodes displayed a description of the link type. Consider the graph fragment in Figure 3 for the text *'The wolf chased the rabbit.'* From this alone it is not possible to tell whether the wolf is chasing the rabbit or the rabbit is chasing the wolf. Adding on the relevant PropBank roles *follower* and *thing followed* makes this unambiguous.

A version of the graphical representation was produced to correspond to each coreference condition being tested: Manual coreference resolution; Literal - the literal text with no coreference resolution; CoreNLP; and Illinois. For each story under each condition, the graphical representation contained an indexed list of unique entities detected. This was displayed as a unique number and the head noun of a found entity. Coreferent mentions on the graph fragments were then annotated with the index of their corresponding entity, so the head noun could be read off of the list for question answering.

Two native English speakers were trained on how to read the graphical representations in order to answer the set of questions for each story under each coreference condition. More opinions were not necessary because answers to the set of questions were a matter of fact, does the graph contain an edge which represents an answer to the question and if so, what is that answer? Any disagreements between

the individual annotations of the two experts and myself were then discussed until a full consensus as to the correct answer to each question was reached. This can be justified by the principle of "what I tell you three times is true" (Carroll, 1979, p. 46).

This task was performed on graphical meaning representations, where the original text of the stories was not available to the two trained experts. The decision was made to perform the study in this way primarily to avoid readers compensating for coreference errors by mentally correcting them without thinking, in a similar manner to how readers can still read a text which contains simple spelling mistakes (Rayner et al., 2006). A screenshot of this question answering interface for the manual coreference condition can be seen in Droog-Hayes (2017).

3.1.4 *Intrinsic Coreference Evaluation*

Table 2 shows the performance scores for the two coreference resolution systems on this dataset according to MUC, B^3 and $CEAF_e$, as well as the mean of the F1 scores (CoNLL F1) as described in Pradhan et al. (2011). The results of all three metrics suggest that CoreNLP outperforms the Illinois system on this dataset.

For this dataset CoreNLP detected 177 noun phrase mentions, of which 14 were spurious and did not represent noun phrase mentions. By contrast, the Illinois system detected 142 mentions, only 4 of which were spurious. It is non-trivial to give a meaningful breakdown of the results, as these systems can produce a variety of *types* of mistake. This study produced twenty sets of automatic coreference resolution annotations, two sets for each story. Of these twenty, less than half contained the correct number of entities. In some cases the correct *number* of entities simply meant that gold standard entities were not detected, and additional spurious entities were introduced by the system. Furthermore, in some sets of annotations, a single gold standard entity was detected as several entities in a system output. Other types of error related to undetected or incorrectly resolved mentions.

Results are given for the separate splits of human-only stories and stories with anthropomorphic animals. This is to show that using stories containing animals speaking as human characters had no detrimental affect on the performance of the coreference resolution systems on this dataset. In fact, according to the MUC and B^3 metrics, both systems performed at least as well if not better on this split of the data. Table 2 shows that even on a small dataset, the differences in methods of calculation for these three metrics can noticeably affect the results. As such it is useful to compare the results under these metrics with the results of the semantic understanding task.

Table 2: Average F1 scores under commonly used evaluation metrics.

DATA SPLIT	MUC		B^3		$CEAF_E$		CONLL F1	
	CoreNLP	Illinois	CoreNLP	Illinois	CoreNLP	Illinois	CoreNLP	Illinois
Human	71.43	62.45	57.04	42.19	60.19	49.19	62.89	51.28
Animal	75.04	62.45	60.62	47.69	51.51	50.15	62.39	53.43
All	73.23	62.45	58.83	44.94	55.85	49.67	62.54	52.35

3.1.5 *Extrinsic Coreference Evaluation*

In judging the answers provided to the questions about the stories, a leniency rule was introduced. This acted to give a more positive reflection of the abilities of the automatic coreference resolution systems in situations where they did not detect mentions at all. In some cases, a mention of an entity was not detected, however the node text on the graphical representation provided a nominal expression. Coupled with the coreference annotations for the other mentions in the text, this could still lead participants to the correct answer to a question. This is not the same as the answer that would be obtained from the meaning representation with no coreference resolution performed, as other correctly identified mentions are additionally required. This rule was applied more frequently to the Stanford system than the Illinois system, without which the performance gap between the two would be more pronounced. The following example question and corresponding answer under each coreference condition illustrates this rule and the effect of the different conditions on the ability to answer the questions. Consider:

QUESTION: What did the father wish?

LITERAL: *Not Applicable* - as no pronouns are resolved, the question does not make sense in this particular instance.

CORENLP: *For his sons to give his farm the same attention as he had given to it* - where *his* and *he* correctly resolve to the father and *it* resolves to *the same attention*.

ILLINOIS: *For his sons to give the farm the same attention as he had given it* - where *his* and *he* correctly resolve to the father and *it* is not resolved at all.

In this example the answer under the Literal condition is incorrect; as none of the pronouns are resolved it cannot be said that any of them refer to the father in the question. Using CoreNLP coreference information, the answer is counted as incorrect as it reads *'For his sons to give his farm the same attention as he had given to the same attention'*, which is nonsensical. The answer using coreference information from the Illinois system is correct under the leniency rule, as *'it'* is neither correctly or incorrectly resolved, but reads true to a human reader.

Table 3 gives the results of this question answering task under the four different coreference conditions. Results are given for the splits of data both including and not including animal characters, as well as the overall results. These splits contained an unequal number of questions, hence the overall results are not a mean of the two averages. It is important to realize that given imperfect coreference resolution, a question itself may not make sense, in which case no answer can be

Table 3: Percentage of questions correctly answered under four different coreference conditions.

DATA SPLIT	LITERAL	MANUAL	CORENLP	ILLINOIS
Human	27.27%	100%	36.36%	54.55%
Animal	35.00%	100%	50.00%	45.00%
All	30.95%	100%	42.86%	50.00%

provided. *Literal* refers to the results of answering the set of questions where no coreference resolution has been performed and the answer, where possible, is given simply by the text label(s) on the relevant node(s). This score represents the questions that can be answered without character information or that concern spans of text where characters are explicitly named. The *Manual* column represents the questions that can be answered with ideal, independently verified, coreference resolution. As the results show, all questions could be answered correctly using this version of the graphical representations, as if being answered from the original text. The *CoreNLP* and *Illinois* columns give the percentage of questions that could be correctly answered using the outputs from each automatic coreference resolution system respectively.

These extrinsic results show that nearly 43% of the questions could be answered when using Stanford CoreNLP and 50% when using the Illinois system. Less than a third of the questions could be answered without any coreference resolution information; which highlights the importance of coreference information to document understanding. The set of questions that can be answered with no coreference resolution performed is not a proper subset of the questions that can be correctly answered with either automatic coreference resolution system. As some explicitly named character mentions are incorrectly resolved to other characters by the automatic systems, questions that can be answered without additional coreference information cannot necessarily be answered with it.

These results suggest that none of the intrinsic evaluation metrics discussed in this work are suitable for predicting a system's usefulness to this task. Not only do the intrinsic metrics fail to accurately predict performance at this task, they give a misleading idea of which coreference resolution system performs better. The intrinsic metrics suggest that CoreNLP is more accurate over this dataset, while for this task the converse appears true. Although the results indicate that CoreNLP correctly detected a greater number of the noun phrase mentions in the stories than the Illinois system, it appears to have performed noticeably worse in resolving the coreferentiality of these mentions.

A contributing factor to the lack of correlation between the results of the question answering study and the coreference resolution evaluation metrics could be the incorrect detection of head nouns. Regardless of the precision and recall of the automatic systems under any of the three established metrics examined here, assigning an incorrect head noun to an entity can result in lowered performance for question answering. A very pronounced example of this occurs in one of the stories examined here, where the *murderer* character is assigned the head noun *the man whom he murdered*, leading to nonsensical answers to all related questions. While only using head nouns may not give a true picture of the entities that they represent, they are required in practice to substitute co-referent mentions in order to answer questions and generate summaries. Steinberger, Poesio, et al. (2007) performs such substitutions in order to correct dangling anaphoric expressions in generated summaries. For this dataset, although multiple entities were assigned non-informative head words such as *he* by the automatic systems, it was never the case that a more specific co-referent mention in the same *detected* entity simply wasn't assigned. Having an anaphor assigned as an entity's head noun does little to aid the correct answering of questions related to this entity.

3.1.6 *Discussion*

This study was designed to give the automatic coreference resolution systems the best possible score they could achieve on the described task. While it is not possible to define an exact level at which the extrinsic performance of these systems would be considered acceptable, the results of this task suggest they offer little improvement over not being used at all. Even then, they do not offer a clear cut improvement and introduce different types of error. Judging by the intrinsic performance of the systems, one would expect the extrinsic results to be higher than what is reported here. In theory, the results of this task should have been significantly higher if the systems' primary loss of accuracy came from the incorrect resolution of minor characters. The questions used to evaluate these systems extrinsically concerned only the aspects of the stories deemed summary-worthy. As such, mistakes by the coreference resolution systems regarding minor characters are tolerable and would not affect the answers to the questions participants were asked. Conversely, coreference mistakes concerning major characters should have a very negative effect on their performance in this task. The results presented here then suggest that the areas where the automatic coreference resolution systems are failing concern the most critical characters.

The results of this study indicate that coreference information is needed, but that automatic coreference resolution in its current state provides too large a source of error for tasks requiring semantic un-

derstanding, and that manual coreference resolution should be performed where feasible. The mistakes made by such systems appear to cover important aspects of a text, which would seriously affect the automatic planning and generation of summaries. This is a small-scale study, but the results suggests that the typically used established evaluation metrics for coreference resolution cannot predict a system's usefulness to these tasks.

3.2 EVALUATING ROUGE AS AN EVALUATION METRIC

The evaluation of summaries is a difficult task even for human judges, being both time consuming and subjective. An automatic evaluation metric should remove the potential inconsistencies of human judges and address the time concern. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is the most widely used automatic summarization evaluation metric. The aim of the ROUGE metric is to rate a summary according to its lexical overlap with a set of one or more reference summaries. However, upon inspection there are several significant issues with ROUGE that make it unsuitable for the evaluation of generated summaries, especially abstractive summaries. Consider the example below:

TEXT: Jon went to the store. He bought a bottle of water at the shop.

SUMMARY 1: Jon bought a beverage from the store.

SUMMARY 2: Jon purchased a drink at a shop.

In this example, a human reader would consider the two summaries to be semantically equivalent. The primary difference between them is the use of synonyms for three of the words. However according to ROUGE, this pair of summaries have a 28.6% similarity at best[2]. This toy example illustrates how in practice a ROUGE score cannot be relied upon to indicate the quality of a summary.

In this section, I highlight some problems with ROUGE in order to justify why I do not consider it to be a suitable evaluation metric for my work. This consists of a discussion of issues with the formulation of the metric, its potential misuse, and an small-scale empirical study. The study tests ROUGE across three different domains of summarization, with the assumption that multiple human summaries of a given document should be of a similar high standard and score well under a summarization evaluation metric. The results indicate that ROUGE has a low upper bound score for human summaries, and cannot account for the variation in the wording between them.

2 Scored by ROUGE-N where $n = 1$, scoring the unordered unigram overlap. This value reflects both precision and recall scores as the summaries are of equal length.

3.2.1 *Criticisms of ROUGE*

Existing research has indicated the limitations of using n-gram overlap as a means of evaluating one summary against another. Work by Kwong (2010) has shown the disparity between the interpretations of three people over just three sentences of a story. Ceylan and Mihalcea (2007) have also shown that over a set of 50 documents, using ROUGE-1 to compare one human summary to another achieves a recall score of only 49.3%, setting a low upper bound for automatic summarization.

Harman and Over (2004) found a low unigram overlap between different human-written summaries of the same document, with the majority of summaries having no trigram overlap whatsoever. Their results showed not only a large variation in unigram overlap between summary pairs, but that none of them had an overlap above 30%. This demonstrates that even with 'ideal' human-written summaries, unigram overlap is not a clear indicator of document similarity. The authors provide examples and hypothesise that the unigram overlap between summaries can differ because they are written at different levels of granularity or include dissimilar material, leading to different lexical choices. They also find that a small unigram overlap between two human model-summaries does not predict a wide difference in human-assigned coverage scores (how well do the summaries cover the content of the documents) between summaries. This finding goes against the basis of ROUGE, that n-gram overlap correlates with summary quality. The authors suggest that this lack of correlation is because humans are able to compensate for different word choices between different summaries.

Research has also indicated the inability of ROUGE to distinguish between human and machine summaries (Rankel, Conroy, Slud, et al., 2011), and between systems claiming state-of-the-art performance. Hong et al. (2014) used ROUGE to compare the performance of various baseline summarizers and systems claiming to be state-of-the-art over the prior 10 year period. In doing so, the authors found that there had been no substantial improvement in summarization systems according to ROUGE. Additionally, the summaries generated by these systems had little content overlap despite the similarity in evaluation scores.

The ROUGE metric has many variants, Graham et al. (2015) highlighted that with all its parameters, there are 192 variants of ROUGE. Graham showed that contrary to what is commonly used, ROUGE-2 with stemming and stop word removal provides the best correlation with human judgements. She then replicated an evaluation of state-of-the-art summarization systems, and in doing so, she found the relative performance of the tested systems to be different from the existing rankings.

Regardless of the literature detailing alternative summarization evaluation methods, ROUGE still appears to be the most used. The two main ROUGE papers (C.-Y. Lin and Hovy, 2003; C.-Y. Lin, 2004b) have well over 4000 citations between them, while by comparison the paper detailing the Pyramid evaluation metric (Nenkova and Passonneau, 2004) has less than 500. As an easily available, fast and automatic evaluation metric, it is understandable how ROUGE can be preferable to a time consuming manual evaluation.

3.2.2 *Theoretical issues*

The ROUGE metric is used to determine the quality of a *system* summary by comparing it to one or more *reference* summaries, where the greater the lexical overlap between system and reference summaries, the better the system summary is (C.-Y. Lin and Hovy, 2003). This comparison is based on counting the overlapping units between the candidate and reference summaries. These units consist of words or word sequences, potentially with gaps in between them. A description of all ROUGE variants and their calculation is given by Lin (C.-Y. Lin, 2004b). The most frequently used is ROUGE-N. This metric looks at the n-gram overlap between a system summary and one or more reference summaries according to the following:

$$\text{ROUGE-N Precision} = \frac{\sum_{R \in \mathbb{R}} \sum_{g \in S} |\{g | g \in S \wedge g \in R\}|}{\sum_{R \in \mathbb{R}} \sum_{g \in S} |\{g | g \in S\}|}$$

$$\text{ROUGE-N Recall} = \frac{\sum_{R \in \mathbb{R}} \sum_{g \in S} |\{g | g \in S \wedge g \in R\}|}{\sum_{R \in \mathbb{R}} \sum_{g \in R} |\{g | g \in R\}|}$$

$\mathbb{R}$ is the set of reference summaries, R is the set representation of the n-grams it contains, and g is the n-gram of size N occurring in the *system* summary S. In other words, the recall is the ratio of overlapping n-grams between a system and a reference summary to the total number of n-grams in the reference summary. The precision can be expressed as the mean n-gram overlap between the system summary and a reference summary.

System summary: 'a b c d'
Reference summary 1: 'a b c d'
Reference summary 2: 'c d e f g'

P_1: 100% P_2: 50% $P_{\{1,2\}}$: 75%
R_1: 100% R_2: 40% $R_{\{1,2\}}$ 66.6%

Figure 4: An example of the effect of summary length on ROUGE precision and recall scores.

3.2.2.1 *Formulation issues*

There is a subtle difference between these formulae for precision and recall. The recall score will tend towards the recall score between the system and the longest of the reference summaries, whereas the precision is the mean of all of the single-reference precision values. This fact may not have been considered in the original evaluation of ROUGE, which was only performed on summaries of identical lengths (C.-Y. Lin and Hovy, 2003; C.-Y. Lin, 2004a; C.-Y. Lin, 2004b).

Figure 4 gives an example showing the effect that the length of multiple reference summaries has on the calculation of ROUGE precision and recall scores. The average precision using a single reference (P_1 or P_2) is the same as the precision for using both references together ($P_{\{1,2\}}$). However the average recall against a single reference (R_1 or R_2) is different to the recall of using both references ($R_{\{1,2\}}$), which tends towards the recall between the system and the longest reference summary.

Furthermore, without even considering the content of a system summary, its length alone affects the ROUGE score. When evaluating a system summary against an arbitrarily longer reference summary, it is self-evident that the count of overlapping n-grams will always be less than the count of n-grams in the reference summary, making a perfect recall score unattainable. Whereas to a human, it should not overly matter whether a summary is verbose or concise as long as it contains the same information. The opposite is true for ROUGE precision scores. With a short system summary and an arbitrarily long reference summary, the proportion of overlapping words relative to the length of the system summary will be high.

3.2.2.2 *Potential misuse*

It is tempting to place all the blame for the issues with ROUGE on its creators, however this would not be fair. Although it was originally intended, and evaluated, as a recall-based ranking metric, later versions included precision and now F1 scores for ROUGE are commonly reported. The various settings for multiple parameters additionally

mean that scores for 192 distinct ROUGE variants can be calculated, but in practice the value of these parameters are often omitted from system evaluations. As a result, even when systems are evaluated on the same dataset, direct comparisons are often not possible. Furthermore, even with a great number of variants, only a handful of them tend to be reported. In practice, ROUGE-1 using a single reference summary appears to be most common.

Moreover, ROUGE-1 was initially evaluated just as a ranking metric. The creators of ROUGE found a strong correlation between summary rankings according to their metric, and summary rankings according to human coverage judgements (C.-Y. Lin, 2004b). This correlation was stronger when multiple reference summaries were used. However, this only demonstrates the ability of ROUGE to rank a set of automatic summaries. Whereas it is often incorrectly applied as an absolute indicator of summary quality.

Another issue is that ROUGE was only evaluated on news data, which is intrinsically different to other domains of text. Cohan and Goharian (2016), for example, show that ROUGE is not reliable for evaluating scientific articles, as the ROUGE scores have a low correlation with Pyramid scores, and a large variance in the correlations depending on the ROUGE variant used.

There is also some confusion as to the scoring formula that should be used for ROUGE when multiple reference summaries are available. C.-Y. Lin (2004b) states that when multiple references are used, the pairwise ROUGE score between the candidate summary and every reference summary should be calculated, and only the maximum of these should be used. In direct contrast to this C.-Y. Lin (2004a) provides the averaging formula for recall shown above. The official ROUGE implementation[3] defaults to the latter, with the option to use the former. However when using a single reference summary there is no difference between the two, leading to the question of why the former scoring method was not set as the default. For this reason, and because relatively few papers report the ROUGE options used, I use the default scoring metric in my empirical study of ROUGE.

3.2.3 *Empirical issues*

With only the formula for ROUGE-N recall, it is evident that unless every reference summary consists of the same set of n-grams, it is not possible for any system to achieve a perfect ROUGE score against multiple references. As such, a greater number of reference summaries tends to reduce the absolute ROUGE score. This fact may help to explain the tendency to only use a single reference summary when using ROUGE to evaluate a system. However, C.-Y. Lin (2004a) indicates that multiple reference summaries make ROUGE evaluation results

3 http://berouge.com/ [Accessed 13 April 2017]

more stable. He cites the work of Nenkova and Passonneau (2004), which highlights the benefit of multiple references to avoid bias, as there is no single correct summary of a document (Jing et al., 1998). The key claim to this statement is more stable evaluation results, not the effect of multiple references on absolute ROUGE scores.

ROUGE has been used to try and optimise extractive summarization systems, (Narayan, Cohen, and Lapata, 2018) for example. However it has been shown that this is an NP-hard problem and that there is little understanding of how ROUGE reflects ideal human summaries (Schluter, 2017). Here I perform a small-scale pilot study to further motivate my decision against using ROUGE as an evaluation metric for my work. I use ROUGE to compare abstractive human summaries solely to other human summaries to gain a better understanding of the upper bound scores that can be achieved by this metric, as well as seeing whether a larger set of reference summaries can better account for the variability in a given summary.

3.2.3.1 *Method*

This empirical study evaluated ROUGE on three different domains of text: News Articles, Scientific Articles and Stories. The intention of this was to see whether ROUGE indicates a greater inter-human summary agreement and is thus more appropriate in one of these domains over another. Originally, ROUGE was only evaluated on summaries of the same length, however this does not indicate its performance on naturally written, abstractive summaries. In this study, participants were not told either how long a summary should be, or the level of detail that it should contain.

Six participants were asked to produce summaries for a set of ten stories and ten news articles. A seventh summary for each of these documents was created from existing data as a proxy for an additional participant. For each story the respective Wikipedia plot summary was used, and selected news articles were taken from a corpus already containing summaries for each document. Sentences in the Wikipedia plot summaries which discussed the context of a story, or differences between alternate versions of the story were removed. The inclusion of such sentences expressing information outside the content of the stories would have otherwise led to lower ROUGE scores. Due to the time investment required to summarize ten scientific articles, this was only performed by two human summarizers. The abstract of each article was used as a proxy for a third human summary. This study only considers single-document summary evaluation as that is most relevant to my own task.

In this study I examined results for ROUGE-N with values of n from 1 to 4, ROUGE-L and ROUGE-SU-4, as they are the most com-

monly reported variants. ROUGE scores were calculated[4] using leave-one-out cross-validation, scoring one summary against all possible combinations of references. For each of the three corpora, the ROUGE scores for each summary t of each document S were calculated in turn against every subset of references that could be created out of the remaining summaries of that document, as in Algorithm 1.

Algorithm 1: Calculation of ROUGE scores for a summary t against a set of M reference summaries for document S.

```
for C ∈ Corpora do
    for S ∈ C do
        for 1 ⩽ x < |S| do
            for t ∈ S do
                for M ∈ P_=x(S/{t}) do
                    ROUGE(t, M)
                end
            end
        end
    end
end
```

Scores were not averaged either within-domain or between domains as the considerable variation in absolute ROUGE scores would obscure the results. Instead I consider the ROUGE scores obtained when varying the number of references a document is evaluated against, and results are presented as an average on a per-document basis. Due to issues with ROUGE already discussed, I only consider recall scores in the following results.

ROUGE scores were averaged on a per-document basis for a given number of reference summaries. For instance, the ROUGE score obtained for every combination of using 2 reference summaries against a given test summary t was averaged. This method further motivates the use of the average, rather than the maximum ROUGE scoring formula. Against a single reference summary, the ROUGE score will be the same regardless of whether the average or maximum scoring method is used. But when every combination of reference summaries is tested and averaged, it is self-evident that as the number of reference summaries increased, so too will will the averaged maximum ROUGE score. For a given test summary t, as the size of the reference set increases, so too does the likelihood that the reference set will include the summary which has the greatest lexical overlap with t, thus giving the highest ROUGE score under the maximum scoring formula. As such it is not informative to examine here the effect of

4 The ROUGE options used in all the experiments described here are '-2 -4 -u -n 4 -f A -a -e relevant-data-folder' with the addition of -s for stop word removal and -m for stemming where applicable.

Table 4: Document and summary statistics averaged over each corpus.

FEATURE	STORIES	NEWS	SCIENCE
Single reference ROUGE-1	45.9	37.7	41.9
Document length	1297.8	482.1	5769
Summary length	173.6	44.2	133.8
S.D. Summary length	110.9	32.9	46.6
Compression ratio	7.8	11.5	44.9

multiple reference summaries on the maximum ROUGE score that can be achieved.

3.2.3.2 *Results*

Table 4 provides some statistics on the summary data used in this study, averaged over each corpus. The single reference ROUGE-1 score indicates the mean ROUGE-1 score for each corpus when only a single reference summary is used. As can be seen from the results, there is a positive correlation between this result and the average length of the summaries. This provides some empirical validation that the formulation of ROUGE means it is affected by summary length. This dataset further indicates that the length of summaries can vary greatly, contrary to the original evaluation of ROUGE. The *S.D. Summary length* row of Table 4 shows the mean of the standard deviations in the word counts of the summaries for each corpus. This represents the range of the differing lengths of summaries received from participants.

ABSOLUTE SCORES

Figure 5 shows how the average ROUGE-1 recall score for each document in the corpus of (a) Stories, (b) News Articles and (c) Scientific Articles changes with the number of reference summaries. The highest average ROUGE-1 scores are clearly obtained when only a single reference summary is used. These scores decrease with the number of references, but do start to level out. Across the three domains, the highest average ROUGE score obtained was a little over 50%, for the story corpus. Ideally, an automatic evaluation metric should be able to indicate that all of the summaries in this dataset are of a high quality.

This same inverse relationship between the number of references and ROUGE-1 score holds across all of these corpora regardless of whether ROUGE's stemming and/or stop word removal parameters are used. This inverse relationship also holds across ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-SU-4 for 50% of the Sto-

Table 5: ROUGE-1 scores for a given number of reference summaries averaged across the set of 10 stories.

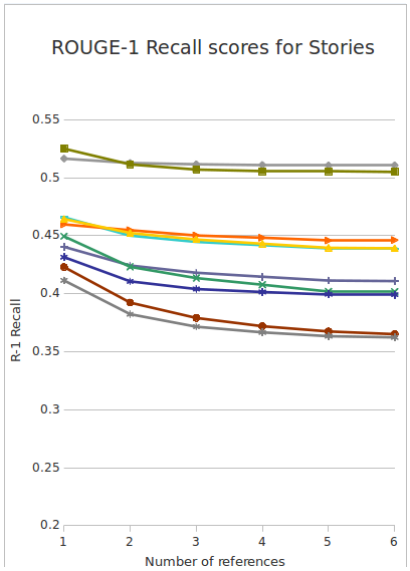
REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	45.9	34.6	49.0	39.6
2	44.2	33.3	47.1	38.2
3	43.5	32.8	46.4	37.6
4	43.1	32.5	46.0	37.3
5	42.9	32.3	45.7	37.1
6	42.8	32.3	45.7	37.0

Table 6: ROUGE-1 scores for a given number of reference summaries averaged across the set of 10 news articles.

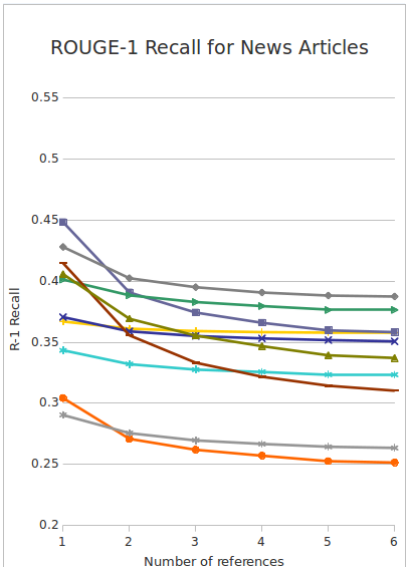
REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	37.7	33.4	40.7	37.9
2	35.1	30.8	37.8	34.9
3	34.1	29.9	36.8	33.9
4	33.7	29.4	36.2	33.3
5	33.3	29.0	35.8	32.9
6	33.2	28.9	35.7	32.8

ries, 90% of the News Articles and 70% of the Scientific articles. Furthermore, across all these variants of ROUGE, stop word removal led to a significant decrease in ROUGE scores, indicating that a sizeable component of these scores comes from function words. The results for these other ROUGE variants are reported in Appendix B.

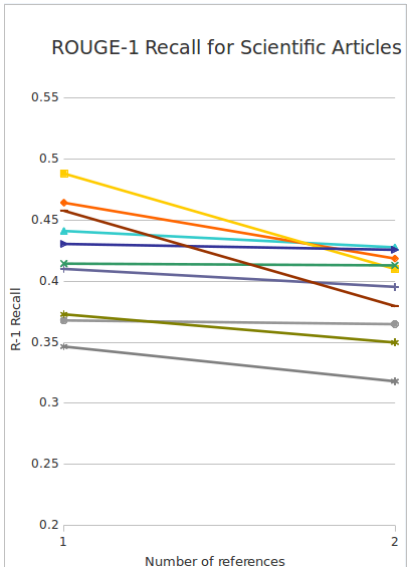
It is not trivial to present these trends in a succinct way. As already described, results are averaged on a per-document basis for a given number of reference summaries. However, to give an indication of these trends, the ROUGE-1 results are averaged over each dataset and presented in Table 5 for stories, Table 6 for news articles, and Table 7 for scientific articles. The figures in these tables show ROUGE-1 scores averaged for a given number of references as before, but here also across all 10 documents of a given dataset. As such, their purpose is to provide the reader with an indication of these trends only. These results show, intuitively, that the removal of stop words leads to lower ROUGE-1 scores due to a reduced lexical overlap. Stemming leads to higher ROUGE-1 scores as lexical differences arising from the conjugation of words are removed. The application of both stop word removal and stemming leads to overall lower ROUGE-1 scores.



(a) Stories



(b) News Articles



(c) Scientific Articles

Figure 5: Average ROUGE-1 recall scores plotted against number of reference summaries for Stories, News Articles, and Scientific Articles.

Table 7: ROUGE-1 scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	41.9	33.9	45.4	40.1
2	39.0	30.4	42.4	36.3

Table 8: The minimum and maximum differences in ROUGE-1 scores when using a single reference summary.

	STORIES	NEWS	SCIENCE
Min range	38.9	38.6	16.7
Max range	82.5	95.3	67.6

Table 9: The minimum and maximum differences in ROUGE-1 scores when using two reference summaries.

	STORIES	NEWS	SCIENCE
Min range	34.8	34.5	6.47
Max range	69.0	78.4	36.1

VARIABILITY

The results showed a large variation in the ROUGE scores that could be obtained based on the choice of reference summaries. Table 8 shows the effect of choosing one reference summary over another for evaluation with ROUGE-1 and a single reference summary. The minimum and maximum ranges indicate the smallest and largest disparity in absolute ROUGE-1 scores that could be obtained by changing the reference summary used. In other words, the *Min range* for stories indicates that there was a summary for which the choice of reference summary can make a nominal difference of 38.9% in the ROUGE-1 score. And the *Max range* for news articles indicates there was a summary for which the choice of reference summary can make a nominal difference of 95.3% in the ROUGE-1 score. This indicates that there were two news summaries that had a less than 5% unigram overlap between them. This illustrates just how significant the differences in ROUGE scores can be when just comparing two human written summaries of the same document.

Table 9 shows the best and worst case variation in ROUGE-1 scores when using two reference summaries. This shows the potential variation in ROUGE-1 scores is lower than when using a single reference summary. This variation continues to decrease as more reference summaries are used. For the story corpus, when using 6 reference summaries the minimum range in variation is 20.0% and the maximum range is 60.9%. When using 6 references for the news corpus, the minimum range in variation is 14.7% and the maximum range is 46.8%. While this does validate the claim of C.-Y. Lin (2004a) that multiple reference summaries make ROUGE evaluation results more stable, it also demonstrates that there can still be a large degree of variation depending on the references used.

3.2.4 *Discussion*

The ROUGE metric compares a test summary to a set of one or more model summaries, which act as the upper bound of what the test summary should contain. It works on the assumption that using a larger set of model summaries should therefore act as a better reference and account for the variability in summaries. This small study has shown that as more reference summaries are used, variation in ROUGE score does indeed drop, as expected; but also that lower absolute ROUGE scores are invariably obtained. This testing of ROUGE also exclusively covered human-written summaries, which has indicated a modest upper bound of what should then be reasonably achieved by automatic summarization.

Table 8 and Table 9 show that a larger set of reference summaries does indeed lead to a decrease in the variability of ROUGE scores. In other words, the lexical differences between the test summary and a reference summary have a smaller impact on the ROUGE score. However, even with up to 6 reference summaries, there is still a considerable amount of variation in scores. In practice, obtaining a large number of human-written reference summaries incurs a cost which may negate much of the benefit of an automatic evaluation metric.

These results constitute a significant caveat that, while the majority of work that uses ROUGE evaluation reports results with only a single reference summary, it is possible to achieve highly varied scores under this condition. I suggest that such varied results are not very informative, but highlight the importance of using reference summaries that are accessible to other researchers so that accurate comparisons of results can be performed.

As noted by (Rankel, Conroy, Slud, et al., 2011), human evaluators can clearly distinguish machine and human summaries, while metrics such as ROUGE cannot. Ideal evaluation of summaries requires not only an understanding of the document in question (as is the aim with summarization), but also its summary and whether the summary entails the original document. Instead, metrics such as ROUGE, which are computationally cheap, compare literal lexical similarity between a generated summary and a supposedly ideal summary. Two different summaries of the same text may both be perfectly acceptable, but simply summarize at different granularities of the text, or use different vocabulary.

I have also highlighted the potential misuse of ROUGE, providing a reminder that it was originally proposed as a recall-only ranking metric, although it appears to be rarely used this way in current summarization research. Additionally, it is necessary for users of ROUGE to report the exact options they have used, given that settings such as stop word removal or stemming can have a large impact on the absolute results across all variants of ROUGE. There are also many variants

of ROUGE, although I have found that unigram overlap achieves the highest scores when comparing human summaries to other human summaries. But ROUGE-N, especially where $n = 1$, does not promote good summarization. It promotes learning to copy the lexical choices of human summaries as faithfully as possible, but with no requirements for summaries to be meaningful or grammatically correct.

Evaluation of a summary by a human involves understanding whether the summary is representative of the original document, by first understanding both the document and the summary in their own right. Following this, I suggest that the task of automatically evaluating summaries is at least as hard as summarization itself, as document understanding is by no means a solved task. As such, I suggest that summary evaluation can only reliably be performed by human evaluators at present.

3.3 SUMMARY

This chapter has detailed two small scale studies which have influenced my methodology. The first of these studies suggests the limits of automatic coreference resolution tools, and that their evaluation metrics may not be entirely reflective of their performance. The results of this study motivated the manual annotation of coreference information in the subsequent work of this thesis. Although coreference information is additionally obtained from automatic coreference resolution tools for further for further analysis into the knock on effects of errors throughout the different stages of summary production described in subsequent chapters.

The second study of this chapter focuses on the most widely used summary evaluation metric, ROUGE. In addition to existing research suggesting the limits of ROUGE, this study led me to search for alternative methods of summary evaluation. In particular, the results of this study motivated the use of a human-based method of summary evaluation, allowing for summaries to be judged according to their meaning rather than purely by lexical choices.

The tools investigated in this chapter provide fast, fully automated approaches to some of the processes required for automatic summarization. However, the studies described here have highlighted some of their flaws. While performing these processes with human involvement can increase the time taken to both generate and evaluate summaries, it is important that these steps are performed accurately. As far as possible, this allows for the approach presented in this thesis to be evaluated independently of imperfect processing steps.

4

SYSTEM OVERVIEW

In the previous chapters I have discussed relevant literature on the generation and evaluation of summaries, and begun to draw together work from psycholinguistics and studies of narrative structure. I have also detailed two preliminary studies regarding important tasks in summarization, which have helped to shape the methodology that I follow.

In this chapter, I propose an approach to summarization which is motivated by methods that human readers use, and looks for meaning beyond the surface realisation of a text. I detail how existing literature motivated my abstractive framework for summarization, and provide an overview of my system. This supports the following three chapters, which describe the results of my studies, and answer my primary research questions.

4.1 APPROACH

My overall approach to summarization follows the process model put forward by Spärck Jones (1999). This model breaks down the process of summarization into three distinct stages. First is the interpretation stage; a source text must be interpreted to some form of text representation. Second is the transformation stage, whereby the representation of the source is transformed into a representation of the summary. Finally is the generation stage, where the surface text of the summary is generated from its intermediate representation.

This model provides an intuitive framework for summarization. The distinct processes make it easier to draw comparisons with other summarization systems, and understand the logic behind each of the stages. Each of these stages can be further broken down, for example the interpretation stage may first involve representing individual sentences before integrating them into a more global text representation. Furthermore, this model holds similarities with the processes that psycholinguistics tells us that humans perform when reading, understanding, and recalling a text.

Each of these stages are significant problems in their own right; my focus is on the interpretation and transformation of text, considering only simple generation options. Text generation is however a necessary component, as the evaluation of summarization systems tends to only consider the end product. Moreover, Conroy and Dang (2008) have found that assessors take the linguistic quality of summaries into account, biasing towards human-written summaries. An

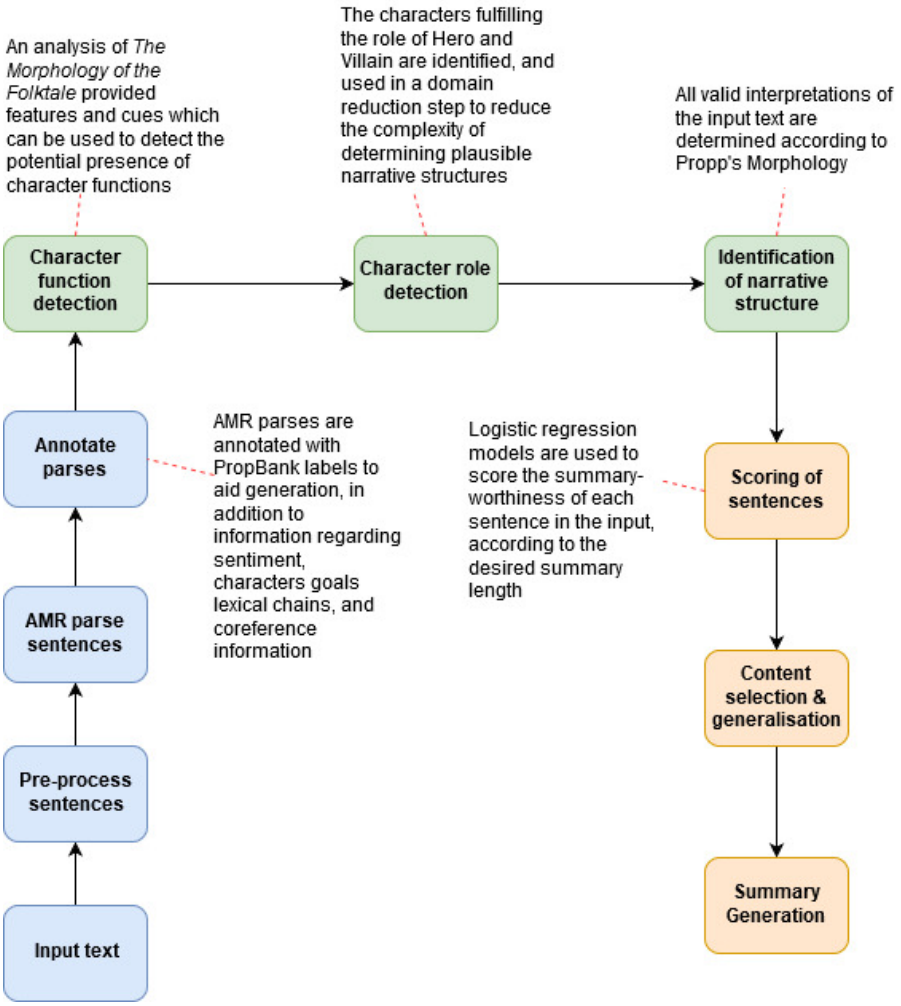


Figure 6: High-level diagram of the architecture of the proposed system.

overview as to the architecture of the proposed system can be seen in Figure 6.

4.1.1 *Interpretation*

In this first stage of processing, a text is transformed into a stripped down form of the surface text which preserves the meaning but without the syntax or features such as tense, aspect or determinacy of nouns. This semantic representation of the input text has been called a textbase by discourse psychologists (Teun Adrianus Van Dijk, Kintsch, and Teun Adrianus Van Dijk, 1983; Graesser, Millis, and Zwaan, 1997).

This representation stage is necessary at a computational level, in order to reason about the important content, and transform the text representation into a summary representation. Moreover, it enables a deeper semantic representation of the text to be performed. The

Once when I was six years old I saw a magnificent picture in a book, called True Stories from Nature, about the primeval forest.

```
(s / see-01
      :ARG0 (i / i)
      :ARG1 (p / picture
            :mod (m / magnificent)
            :location (b2 / book :wiki -
                  :name (n / name :op1 "True" :op2 "Stories" :op3
                  "from" :op4 "Nature")
                  :topic (f / forest
                        :mod (p2 / primeval))))
      :mod (o / once)
      :time (a / age-01
            :ARG1 i
            :ARG2 (t / temporal-quantity :quant 6
                  :unit (y / year))))
```

Figure 7: Example manually constructed AMR parse with its corresponding sentence. Obtained from the freely available *Little Prince* AMR Corpus (Knight, 2015)

ordering of sentences is preserved in the resulting graph-like textual representation, as the order of events is still important.

4.1.1.1 *Base Representation*

To form the basis of the text representation, I use the Abstract Meaning Representation (AMR) formalism. The AMR Bank is a manually constructed set of thousands of English sentences paired with their semantic representations (Knight, 2015). These representations are rooted, directed and labelled graph structures which each correspond to a single sentence. Nodes represent entities as opposed to words, which is the case in both dependency parsing and semantic role labelling. An example sentence parse according to the AMR formalism is shown in Figure 7.

This representation is not focused around the syntax of a sentence, it collapses morphological variations and focuses on the relations between objects in a sentence. This is beneficial to many areas of research, such as Event Detection, where AMR has been used to im-

prove traditional methods (X. Li, Huu Nguyen, and Cao, 2015). AMR uses PropBank framesets, so its main structure comes from verbal propositions and the arguments they take. Simplifying syntactical variations means that AMR does not consider tense or arity when representing the semantics of a sentence. The structure of AMR and its use of verbal propositions leads to sentences with similar meanings being represented in similar ways. This allows for the possibility of comparing sentences beyond simple lexical overlap. One way to perform this is with the AMR evaluation metric, SMATCH (S. Cai and Knight, 2013), which compares the semantic triples in one AMR structure against those in another.

Although AMR loses the exact representation of the surface text, some AMR parsers provide a word alignment function. This gives a mapping between the nodes of the parses and the original spans of text which they correspond to. To parse texts to this formalism, I use the open source tool JAMR[1] (Flanigan, Thomson, et al., 2014; Flanigan, Dyer, et al., 2016). Like all current methods of parsing text, AMR parsers are imperfect and can introduce mistakes. The only changes I manually perform are to ensure that the parses correctly contain nodes representing the co-referent entities of the original texts.

4.1.1.2 *Coreference Information*

Coreference information is desirable in order to better understand which actions are performed by which characters, and reason about the roles of these characters in the stories. For reasons discussed in Section 3.1, I carried out coreference resolution manually. I developed a simple annotation tool in order to perform these annotations directly on the AMR sentence parses of a text, rather than its surface form.

Automatic summarization contains multiple pre-processing steps, and it is of interest to observe the knock-on effects that inaccuracies in automatic coreference resolution can have on the final summary outputs. For this reason, automatically obtained coreference information was also stored. Coreference annotations according to both Stanford CoreNLP, and the neuralcoref system (Wolf, 2017) for spaCy [2] (Honnibal and Johnson, 2015), which claims state-of-the-art performance, were used [3]. The effects of using manual versus automatic coreference resolution are discussed in the results of Chapter 6 and Chapter 7.

1 While other AMR parsers have been produced, cf. (Artzi, K. Lee, and Zettlemoyer, 2015; Pust et al., 2015), their implementation was either inaccessible or did not produce word alignments.

2 Implementation available from `https://spacy.io/`

3 For implementation reasons, the Illinois coreference resolution system already discussed was not further used.

4.1.1.3 *Semantic Annotations*

In addition to coreference information, various semantic features are annotated in order to provide further inter-sentential links between concepts. I learn from the valuable concepts and methods employed by existing approaches to summarization in order to determine the important semantic information that should be present on a meaning representation of text. The aim of this is to create a richer text representation to reason about, and to more faithfully simulate the mental models of interconnected concepts that psycholinguistics tells us humans form during reading. Unlike coreference information, these features are automatically annotated.

Chains of semantically related nouns are identified and connected according to the lexical chaining algorithm described by Silber and McCoy (2000b). These lexical chains are built by finding nouns which are related in WordNet by either hypernymy, hyponymy, are instances of the same synset, or are siblings (nouns which share a hypernym). The ability of lexical chains to capture important content has been demonstrated by a number of attempts to create summarization systems that use them as the sole criteria for content selection (Barzilay and Elhadad, 1999; Brunn, Chali, and Pinchak, 2001).

Information about the emotive language in a text is also annotated. Sentiment analysis has previously been used in summarization to indicate important spans of text (Dabholkar, Patadia, and Dsilva, 2016). To identify emotive concepts I use the large connotation lexicon described by Feng et al. (2013). Feng et al. (2013) make the distinction that the connotative sentiments of words aids the interpretation of the subtle shades of sentiment beyond the surface meaning of a text considered by sentiment analysis. Sentiment analysis has also been used by Nalisnick and Baird (2013) to analyse the relations between pairs of characters in Shakespeare's plays, where sudden changes in character-to-character sentiment are indicative of key events. Reagan et al. (2016) describes how sentiment analysis can be used to classify the plot structure of stories. However, these methods work by examining the change in sentiment over the course of texts that are many times longer than the Russian Folktales considered in my work.

Research into the cognitive representations formed by humans as they read has indicated the importance of causality in the processes of storing and recalling information. For stories, this refers to the goals of the characters, and the main events on the narrative chain which lead to their fulfilment. The goals of characters provide important context for the reader. They motivate the actions of the story, and help to form a coherent text. To capture such information, I compiled a dictionary of cue words and phrases which are indicative of character goals, enabling the automatic marking of these indicators. This list is comprised of the lemmas for the classes of VerbNet (Schuler, 2005)

verbs representing desire, in addition to a list[4] of adverbial words and phrases of purpose.

While AMR uses PropBank framesets to structure the parses, the outputs do not contain PropBank edge labels. The outputs do contain argument numbers for the relevant PropBank verbal propositions, but these are not as informative. Automatically adding these natural language labels is beneficial as they can be used to aid text generation which is less dependent on the input surface text. Part of Speech tagging is also performed for the same reason.

4.1.1.4 *Discourse-Structural Information*

The annotated features described thus far are independent of the genre of text that is used as input. However, I consider only Russian Folktales so that discourse-structural features relating to Propp's Morphology of the Folktale (Propp, 2015) can also be annotated. These features, the constraints of Propp's Morphology, and a system to automatically determine this structural information are discussed in detail in Chapter 5.

Throughout this thesis I use Russian Folktales as an example domain of what is achievable with the approach that I describe, however the methods I present are general and can be applied to other domains. There are various reasons for choosing Russian Folktales as an example domain. Stories such as fairy tales and folktales are often told to a younger audience, and as such they tend to contain relatively simple language and sentence constructs. In theory, this should make the automatic parsing of such texts less error prone. Black and Bower (1980) describe narratives as a "selective rendering of a continuum of myriad events having different levels of description" (Black and Bower, 1980, p. 3). They say that it is according to this amount of implicit, specialised or common cultural information that educators distinguish the difficulty of a text. In the domain of Russian Folktales, the events of the texts tend to be explicit, and do not require much specialised or common cultural information. This is beneficial, as the integration of external world knowledge is outside of the scope of this thesis.

The primary reason for using the domain of Russian Folktales is due to the level of detail with which Propp (2015) describes it. His detailed analysis covers the narrative elements, the types of characters, and the rules and constraints which join these together to form valid instances of tales. This provides an excellent domain in which to study whether discourse-structural information can be detected automatically, and examine its utility for automatic summarization. Black and Bower (1980) state that while stories have a simple structure, it is difficult to capture significant generalisations. Propp's morphology can help to capture these generalisations.

4 http://www.thefreedictionary.com/Adverbs-of-Purpose.htm

Psycholinguistic research highlights the importance of causal connections, and the motivations of characters. While story grammars provide an alternative method of obtaining structural information, they cannot accurately predict the information that human readers will recall (Black and Bower, 1980). In contrast, the character functions of Propp's Morphology capture the motivations of characters, especially the initial instance of villainy or lack. Propp describes this event as the motivator for all subsequent events of the story.

Perhaps closest to Propp's analysis, but applicable to a larger domain of stories, is *The Hero's Journey* described by Campbell (2008). However the structure described by Campbell is far less constrained than Propp. Campbell (2008) describes the 17 stages of the monomyth, or hero's journey. These stages detail the events of a narrative, grouped into three major sections: departure, initiation, and return. However, the order of the stages is less rigid than Propp, and stories may focus on just one of the stages. Campbell gives examples and typical circumstances of each stage, however they are less well defined than those of Propp. They are expressed in terms of settings or events at a more abstracted human level. In contrast, Propp's work provides something that can be detected from the surface text of a tale. Propp's Morphology could potentially be applied to other types of folktale and fairy tale more generally, but that is outside of the scope of this thesis. I focus on the domain about which it created in order to try and best answer my primary research questions.

4.1.2 *Transformation*

The transformation step of the ITG summarization model involves determining the summary-worthy content of the intermediate text representation that should be transferred to the summary representation. This occurs via a process of selection and/or generalisation. Spärck Jones (1999) states that methods of capturing discourse structure are required in the interpretation of a document in order to support the condensation of material in the transformation stage. It is preferable to perform this stage over an intermediate representation so it is more robust against the specific wording of a text.

To aid this process I trained regression models in order to predict content marked as summary-worthy from the features that are annotated on the meaning representation of the input text. This resulted in a set of weights for the annotated features, which could then be used to indicate the elements of the meaning representation that should be included in the summary representation. The learning of these weights is described by Droog-Hayes, Wiggins, and Purver (2019) and in Chapter 6.

The presence of different annotated features results in a score for each element of the meaning representation, effectively giving a rank-

ing of the content in the meaning representation, rather than a binary split. As a result, summaries of different lengths can be generated as desired. The length of a summary directly affects the amount of generalisation and abstraction of content, rather than selection, that occurs in the transformation stage. This is discussed in detail in Chapter 7.

4.1.3 *Generation*

Once a representation of the summary content has been produced, it is necessary to generate the surface form of the output. This is required as it has been shown that human assessors take the linguistic quality of summaries into account even when instructed not to. Selecting important spans of text and concatenating them often does not result in coherent outputs, and is only half of what a summarization system needs to perform (Knight and Marcu, 2000).

Summary text is generated through a process of using the original surface text aligned with elements of the AMR parses, and template based sentence generation. Taking inspiration from existing approaches to summarization, sentence compression techniques are employed when spans of the original surface text are used. The processes involved in summary generation, as well as the evaluation of the final product of this thesis are presented in Chapter 7.

4.2 SUMMARY

This chapter has provided an overview of my approach, motivated by existing literature. I have taken valuable concepts from existing approaches to summarization, as well as research into the ways that human readers store and recall important information from a text. While this approach aims to replicate elements of the cognitive representations that psycholinguistics tells us that humans form while reading, I acknowledge that there are elements of these which are not considered at all. In particular, this approach does not consider the role of external world knowledge in summarization, the importance of which has been indicated by existing literature. However, the features used to create a deeper meaning representation of texts go some way towards accounting for this. Propp's character based functions each provide a generalised description of an event in a story. These each make explicit the information around a key event of a story, which may only be implicit in the text. They account for elements of the text that the author may have left implied, assuming that the reader will have adequate world knowledge.

This chapter has provided the necessary overview of my system to enable the results of this work to be presented in the following three chapters, and answer my primary research questions. Chapter 5 de-

scribes a narrative structure reasoning system, and in particular how I apply it to Propp's Morphology in order to obtain information about the discourse structure of a text. This begins to address whether discourse-structural information can be detected automatically, and answer my first research question, although an evaluation of this process is given extrinsically via the following chapters. Chapter 6 presents a study to learn how the annotated features of the intermediate text representation indicate summary-worthy content. This allows for a first evaluation of the narrative reasoning system, and a further evaluation of the effect of automatic coreference resolution. Most importantly, it shows the benefit of the annotated semantic and discourse features to the detection of summary-worthy information and answers my second and third research questions. Chapter 7 describes how these features are used to perform the transformation stage of the ITG model, and the processes used for the generation stage. I then present an evaluation of the summaries generated by this work, according to varying degrees of automation. This enables a final evaluation of the value of my work, this approach to summarization, and addresses my final two research questions.

5 A NARRATIVE STRUCTURE REASONING SYSTEM

'Narrative structure' refers to the underlying way in which a text is organised. Detecting this structural information can aid the recognition of textual elements that psycholinguistics tells us are important in the cognitive representations of human readers. Having a high level understanding of the structure and events of a text allows for a deeper understanding of the causal links between events.

This chapter addresses my research question of whether structural information can be detected automatically. I detail a system based on constraint logic programming which provides all valid interpretations of the narrative structure of a text according to a set of encoded rules and constraints. This is followed by a description of how the elements of Propp's Morphology can be detected, allowing for an application of this system to the domain of Russian Folktales. I give a qualitative discussion of the outputs, but leave the evaluation of this system, via an application to Propp's Morphology, to the subsequent two chapters.

5.1 SYSTEM DESCRIPTION

This system produces all valid interpretations of the narrative structure of a text according to a set of rules and constraints placed on its narrative elements. The aim of this system is to aid automatic summarization by providing a high-level, abstracted description of a text via knowledge about the structure of its genre. This can help the processes of both detecting summary-worthy content, and generating summaries which are less dependent on the surface form of the input.

The steps in this process are as follows. First, the rules and constraints about the structure of a given genre of texts must be formalised. Second, potential instances of the structural elements within a text of that genre must be detected. Third, a representation of the text with potential instances of its structural elements is passed as input to the system. Finally, the rules and constraints of the domain can be applied. This results in the production of all valid interpretations of the narrative structure of a text according to a given formalism.

There are several assumptions being made for the work presented in this thesis. Firstly, the narrative structure reasoning system is being implemented to specifically handle single-move Russian folktales. That is, Russian folktales with no sub-plots. Secondly, the corpus being used only contains tales with a single character fulfilling the

role of hero and, for the most part, a single character fulfilling the role of villain. The only domain-specific exception to this is when a tale contains multiple dragon characters, which are always villainous. Finally, the role of Hero and Villain cannot be fulfilled by the same character. These assumptions can of course be relaxed in order to handle a wider range of Russian folktales, or stories in general. However, they are being used here as they are applicable to the corpus used in these studies, and simplify the problem of reasoning about narrative structure. The absence of these assumptions simply leads to a less constrained system which is able to produce a greater number of valid story interpretations.

5.1.1 *Constraint Logic Programming*

I treat this task as one of constraint satisfaction. Constraint Logic Programming (Jaffar and Lassez, 1987) is a form of programming whereby the relations between variables are expressed as constraints. Such programs can then be queried about the provability of a goal. The constraints provide conditions which must be satisfied, either relating the value of one variable to another, or placing restrictions on the values which a variable may take. I consider in particular the Prolog implementation of Constraint Logic Programming over Finite Domains (CLPFD). CLPFD allows for reasoning about variables which have integer values and, among other things, provides a set of arithmetical and membership constraints.

Constraint Logic Programming is appropriate for tasks where there are multiple variables and a solution is required which fits all constraints placed over the set of variables. It is trivial to create a mapping between each element of a structural theory and an integer, which can be used to represent it. Constraints can then be implemented in terms of this integer representation. I then consider the task of determining the narrative structure of a text as one of sentence labelling. Sentences must be assigned numerical labels which represent structural elements. The resulting strings of integers must each conform to the provided rules and constraints of the genre.

5.1.2 *Constraints*

Common to different structural analyses such as those of Campbell (2008) and Propp (2015) are discrete structural elements that may be linked according to a variety of rules. By using a numerical representation for the structural elements of a given genre of texts, different types of logical constraint may be implemented. These represent rules that the over-arching structure of a text of that genre must follow.

The types of constraint implemented for this system are detailed below. This is not an exhaustive list of the rules that can be represented by this system, only a list of what was necessary for my task.

ORDER: Structural elements may require a strict ordering. This can be implemented by ensuring that sequential structural elements are mapped to integers of increasing value, and then restricting output strings to only contain integers of increasing value.

MUTUAL EXCLUSION: The presence of one narrative element may exclude the presence of another. This is implemented by ensuring that at most one of the two provided integer arguments may be present in the output.

MUTUAL NECESSITY: The presence of one narrative element may necessitate the presence of another. This can be implemented by ensuring either both or neither of the provided integer arguments are present in the output.

IMPLICATION: This requires that, for two given narrative elements, if the first is present, the second must also be present. However, the second element may be present regardless of the presence of the first.

ONE-OF-n: This requires that at least one of n provided elements must be present in the output. If, for example, the structural theory of a genre specifies that a text can end in multiple ways, this can be used to ensure that at least one of those elements must be present.

The above constraints can be further combined to restrict the structure of a text. For example ONE-OF-n can be combined with MUTUAL EXCLUSION to ensure that *exactly* one of the n elements is present.

5.1.3 *Text Representation*

The preprocessing of a text should result in the determination of the narrative elements that each sentence of the text could plausibly represent. Although some sentences may not be representative of any structural elements. Each sentence is treated as a variable, and the preprocessing of a text identifies the domain of each variable. This information is given as input to the narrative structure reasoning system in the format of a list of lists. Each entry in the outer list represents a sentence variable from the original text, and contains integer values representing the domain of narrative elements it may represent. The domain of each variable must include a null value, to indicate that it may not in fact represent any structural element at all.

5.1.4 *Output*

The application of the encoded constraints to a text results in the production of every valid interpretation of that text according to the provided structural information. This is a process of determining every possible configuration of value assignments that conforms to the given constraints.

These outputs are represented as strings of integers, where the position of each number in the string corresponds to a sentence of the input text. A zero indicates that the sentence represented by that position does not represent any narrative element. A genre of texts with a suitably constrained narrative structure is required in order to limit the number of interpretations.

5.2 DETECTING THE ELEMENTS OF PROPP'S MORPHOLOGY

I consider the application of this system to the domain of Russian Folktales. In particular, I consider what Propp refers to as *single-move* tales. Single-move tales contain a single sequence of events following from the initial motivating action to the conclusion of the tale. Propp's analysis states how these sequences of events can be embedded within each other to describe more complex tales. One way in which this can occur is by the presence of multiple protagonists, who each go their separate ways during the course of a tale. The tale may then follow the actions of each protagonist sequentially or embedded within each other. I made this decision as it further constrains the structure that a given tale can have by enforcing a strict ordering over Propp's character functions. This acts to reduce the number of plausible interpretations of a given tale.

I chose to investigate this genre of text as, in my research, I could not find another equally constrained and well-defined domain. Propp describes how "The storyteller is constrained" in several areas, including the "overall sequence of functions, the series of which develops according to the above indicated scheme" (his function scheme) Propp, 2015, p. 112. These constraints and restrictions between functions make this problem well-suited to Constraint Logic Programming. Here I begin with an explanation of the detection of instances of Propp's functions and the roles of the primary characters, before providing a description of the constraints he places on these elements.

5.2.1 *Character Function Detection*

Propp's Morphology describes 31 narrative units that comprise all possible events of a folktale. Although, a particular instance of a folktale will only contain a subset of these character functions. These each provide an abstracted description of an event from the initial

motivating action through to the conclusion of a tale. Additionally, they state the types of character that must necessarily be involved in each action.

I make two explicit additions to this set of functions which are left implicit in Propp's analysis. The first comes from prepending a function to account for Propp's description of an *Initial Situation* prior to his first function definition. The second arises from expressly splitting Propp's function VIII and VIIIa (*Villainy/Lack*) into two distinct functions. Propp describes how exactly one of these two actions must be present in a tale, as it motivates the subsequent events. This split ameliorates the implementation of constraints and enables a more informative representation of a tale.

As stated in Section 2.2.2.3, each character function has multiple subtypes which specify the exact forms an event may take. I only consider the presence of the overall function types, such as the defeat of the Villain (*Victory*), rather than distinguishing between its multiple subtypes, such as the Villain losing in a contest, being beaten in open combat, being killed, and so on. Propp's analysis details the pairing of some specific forms of functions; for example, a struggle in an open field (*Struggle*) is specifically paired with victory in an open field (*Victory*). However, these fine-grained pairings and their initial detection is highly dependent on the linguistic choices of the author, an aspect of the tale over which Propp says the storyteller has freedom. The surface text of a tale may state "they fought in an open field", but this does not guarantee a direct statement to the effect of "victory in an open field". While the act of victory may be explicit, a restatement of the location is superfluous and so may well be omitted from the text. Moreover, the subtypes of functions are highly specific, but the ability to generalise the actions into a more abstract event is beneficial to the task of summarization. As such I believe that it is not beneficial to attempt the recognition and assignment of these fine-grained function subtypes for this work. Aside from the plethora of conceivable ways by which these subtypes could be expressed in natural language, attempting to recognise them would further increase the number of valid interpretations of a tale.

The first stage in determining the narrative structure of a given folktale involves identifying the potential instances of each character function. That is, the domain of values for each variable must first be established. Some functions are each described as appearing in specific ways. The *Interdiction* function is described in terms of a command, an imperative. This can be recognised via the AMR parses as a sentence containing direct speech, where the main verb proposition is missing its first argument, the subject. Propp describes the majority of character functions in terms of the presence of cue words. While an instance of a character function, such as *Villainy* may in fact span several sentences of the tale, a single keyword such as 'attack' is often

enough to indicate this. Although on the surface the presence of cue words appear to be a simplistic approach, they have been recognised as the most prominent linguistic means for conveying the purpose of a discourse segment (Grosz and Sidner, 1986).

For character functions that Propp describes in this way, I obtain a seed-list of cue words based on the examples he discusses in *The Morphology of The Folktale* and the detailed annotations of Russian Folktales provided in the data of Finlayson et al. (2015). These lists of cue words are then expanded with the use of WordNet and FrameNet (Baker, Fillmore, and Lowe, 1998), with the intention of making them more applicable to a wider range of folktales. These lexical resources are comprised of *synsets*; sets of synonymous words. The synsets containing at least one cue word from a given seed-list of cue words are identified, and the members of these synsets used to expand the list of indicators for that character function. During this process verbs that Finlayson (2016) refers to as 'generic events' are omitted. Some of these verbs, like *go*, can be used to indicate many different character functions. While others, such as *say*, can be an indicator of every character function; in Propp's analysis, every character function can occur via an act of speech. This process results in a long and varied list of cue words that can indicate the presence of a character function. *Villainy*, for instance, may occur by acts ranging from 'exasperate' to 'immolate'.

The annotated parses for each sentence in the meaning representation of a folktale are used to determine potential instances of each character function. Each sentence is then labelled with integer values corresponding to the character functions that it may represent, or a null value to indicate that it does not represent any character function. Evidently, the majority of sentences in a tale will only be assigned a null value; the number of sentences in a tale often far exceeds the number of character functions. Moreover, Propp assigns at most 12 character functions to a single-move tale in his annotations. For most character functions, detecting their presence is a case of comparing the node-text of an AMR parse with the expanded list of cue words. This is performed with AMR parses rather than the surface text of a tale as the lists of cue words only provide the base form of each word, without all of their possible inflections.

As an example, consider Figure 8. This sentence is initially tagged as potentially being an instance of: Villainy, the Hero acquiring the use of a magical agent, a struggle between the Hero and Villain, a liquidation of lack, and the punishment of the Villain. Acquisition of a magical agent and the liquidation of a lack are indicated by *seize*, as these character functions can occur without the consent of other characters. However, with the additional information of which characters perform the roles of Hero and Villain, the majority of these

```
He seized her and he dragged her to his lair.
(a / and
      :op1 (s / seize-01
            :ARG0 (h / he)
            :ARG1 (s2 / she))
      :op2 (d / drag-01
            :ARG0 (h2 / he)
            :ARG1 (s3 / she)
            :ARG2 (l / lair
                  :poss (h3 / he))))
```

Figure 8: Example AMR parse of a sentence from one of the tales analysed by Propp.

options are subsequently ruled out and the domain of this variable is reduced.

5.2.2 *Character Role Detection*

Propp describes seven roles that the characters of a folktale can take: Hero, Villain, Donor, Dispatcher, Helper, Prize, and False Hero. The narratemes of Propp's Morphology are described in terms of the characters who are performing the actions. As such, it is desirable to know the mapping between the actors of a tale and the roles which they fulfil. This allows for a domain reduction of the character functions that each sentence could plausibly represent; if the Villain is not mentioned in a given sentence, then that sentence cannot represent an act of villainy.

Of primary interest is the knowledge of which characters take the roles of *Hero* and *Villain*. Table 10 shows the number of character functions that each type of character is necessarily involved in. In his enumeration of narrative units, Propp describes when a certain type of character must be involved in a given action. This reflects what Propp terms the *sphere of action*. Some types of character, such as the *Prize*, are not essential to any given character function. The Prize may be present in narrative units such as the rewarding of the Hero (*Reward*), but this reward may be monetary or take some other form which does not involve the Prize character. The majority of character functions require the involvement of at least the Hero or Villain, but do not necessarily require the other character roles. For this reason, there is little to gain from the detection of the other character roles.

Table 10: The number of character functions each type of character must necessarily be involved in.

ROLE	NUMBER OF FUNCTIONS
Hero	22
Villain	8
Donor	3
Dispatcher	0
Helper	0
Prize	0
False Hero	2

Moreover, existing research has shown that some of Propp's character roles are unclear, and that the roles of *Donor* and *Helper* are often performed by the same character (Valls-Vargas, Ontanón, and J. Zhu, 2013). The assignment of multiple character roles to a single character is another reason why I only consider the detection of the Hero and the Villain; intuitively (for this genre), and in the corpus of folktales studied by Propp, the roles of the Hero and the Villain are never performed by the same character.

Existing research has investigated the possibility of automatically determining the characters who fulfil the key roles in Propp's Morphology. Valls-Vargas, Ontanón, and J. Zhu (2013) attempt to assign one of Propp's character roles to each character identified in a text. In this work a story is parsed to determine the subject and object of each verb. This information is used to construct an $n + 1$ square matrix, where n is the number of characters identified. Each cell in this matrix then contains the actions (if any) that one character performs to another. Characters are then assigned roles by finding the best match according to a reference point of the actions that characters with each role are typically the subject or object of. When only classifying the role of each character as either *Hero*, *Villain*, or *other role* the authors achieve a classification accuracy of 78.99%. Other observable features such as the number of mentions a character has in a story are not considered in this task. This could potentially aid the classification process; in this domain it is typically the case that the Hero of a story has the greatest number of mentions in the text.

In order to determine the Hero of the tale, I define a metric of character importance. This metric considers the number of times a character is mentioned, over the space of the tale in which they perform actions. That is, the density of coreferent mentions to a character over the span of text in which they are mentioned. This is calculated as the product of the number of times a character is mentioned and the distance (number of sentences) between their first

and final mentions, divided by the distance of the first mention from the end of the text. This metric ideally requires accurate coreference resolution information, and so mistakes in automatic coreference resolution can impact the determination of the characters who fulfil the key roles of a tale.

Character Importance =

$$\frac{\text{count(mentions)} * (\text{pos(last\_mention)} - \text{pos(first\_mention)})}{\text{text\_length} - \text{pos(first\_mention)}}$$

This metric is motivated by the example of the child character at the end of the *The Emperor's New Clothes* in Section 1.1.1.1. It aims to capture the importance of characters who are only mentioned over a short span of text, or are only introduced near the end of a tale, and so cannot have a high total number mentions. It is evident that a character introduced near the denouement of a tale must serve a purpose, however this importance is not captured by simply counting the number of times a character is mentioned.

The role of Hero is assigned to the character which has the highest score according to this metric. Consider the tale *Nikita the Tanner*[1] as an illustration of this. The hero, Nikita, is only introduced a third of the way through the tale, while the dragon (the Villain) is introduced in the very first sentence. In the text, both of the aforementioned characters have the same number of referents, however Nikita only has the opportunity to perform actions over a shorter span of text, and so obtains a higher score.

The role of the Villain is subsequently assigned to the character who has at least one direct interaction with the Hero and is involved in the greatest number of verb predicates with negative connotation. I use the connotation lexicon created by Feng et al. (2013) in order to identify verb structures with negative connotations. Looking at verb connotations helps to fill in for the common sense knowledge of a reader, to recognise that certain actions may indicate villainous behaviour. The condition of direct interaction is enforced by only considering characters who are involved in at least one AMR verb predicate which also has an argument referring to the Hero. The role assignments are performed in this order as the Hero of the story is often also an argument of a large number of verbs with negative connotations; being involved in struggles against the Villain.

While this approach generally works for this genre, I do make one concession. In some instances of these Russian folktales there are multiple characters who share the role of Villain. For example, in

1 In *Nikita the Tanner*, a dragon has been devouring maidens from the city of Kiev. After the dragon kidnaps a princess and weds her, the king and queen approach the eponymous hero for aid. He reluctantly agrees to help and eventually defeats the dragon, subsequently returning to his work of tanning hides. The full text of this story is given in Section C.1

various tales studied by Propp, the Hero will first face a three-headed dragon, then a six-headed and finally twelve-headed dragon. In such instances there is no single antagonist, and multiple characters should share the role of Villain. To account for this, domain knowledge can be used in addition to Propp's work to recognise that dragons always play a villainous role in these folktales.

In the assignment of characters to the roles of Hero and Villain, I make the assumption that these two roles must be filled by two distinct entities. While it is possible to envisage the situation where a single character acts as both the Hero and the Villain in other genres, I believe this to be highly unlikely in folktales. Propp discusses the roles of several dramatis personae being filled by one story character, however his examples cover the possibilities of combinations such as a Donor-Helper, Donor-Villain, or Donor-Dispatcher character. With the assumptions already discussed, this method of character role detection leads to the correct assignation of the role of Hero in all of the folktales examined. Using domain-specific knowledge regarding multiple dragon characters fulfilling the single role of a Villain, the correct characters are also assigned the role of Villain in every tale examined.

5.3 DETERMINING THE STRUCTURE OF RUSSIAN FOLKTALES

In this section, I report each of the rules and constraints that I have implemented in order to determine the structure of Russian Folktales. These are based on descriptions from Propp's own work. However, I take a somewhat strict interpretation of his work. This is necessary in order to limit the number of interpretations that can be created for a given folktale. Propp does not always strictly follow the constraints he lays out in *The Morphology of the Folktale*, as can be seen by inconsistencies in the appendices of the same book.

5.3.1 *Domain Reduction*

Prior to the application of narrative constraints, a domain reduction step is performed. Without even considering the structural rules laid out by Propp, certain assignments of functions to sentences can be trivially ruled out. This domain reduction reduces the complexity of the constraint satisfaction problem and reduces the number of outputs.

The character functions of Propp's Morphology necessarily involve characters. Each of them describes an event performed by, or to, a character of the story. However, there are of course sentences in many stories which do not contain referents to any characters. By this rule, and with the use of coreference information, any sentence that does not contain any references to a character cannot be an instance of

one of Propp's functions. In other words, the domain of such variables is reduced to the null value. This rule can be extended to cover character functions that describe the involvement of the Hero and/or Villain. Character functions requiring the involvement of these characters cannot occur in sentences which do not contain referents to the required characters. The relevant values are removed from the domains of variables which do not meet this criteria.

Considering again the tale of *Nikita the Tanner*, the hero of this tale is only mentioned for the first time in the eleventh sentence. However the fifth sentence is a valid candidate for *Departure*, the character function representing the departure of the Hero, as this is often indicated by verbs of motion. With coreference information and character role assignment it can be automatically detected that the fifth sentence makes no mention of the Hero and so in fact cannot represent an instance of *Departure*. In practice, this process greatly reduces the number of plausible function assignments available to each sentence.

I further make the intuitive decision that the first thirteen functions, up to and including the Hero's departure, must occur in the first half of the text. Propp terms the initial functions, those leading up to the lack or act of villainy, as the preparatory functions. The character functions from *Lack/Villainy* to *Departure* represent what Propp calls the complication. I create the restriction that the values representing these functions cannot be present in the domains of sentences occurring in the latter half of a tale. As the functions have a sequential ordering, this removes assignments of character functions whereby all of the values are bunched-up at the end of a tale, and and represent a tale whose first half carries no meaning. This again greatly reduces the size of the search space and the number of valid outputs.

5.3.2 *Narrative Constraints*

Below is a description of each of the constraints on the structure of Russian Folktales that I have implemented in the narrative structure reasoning system. I follow what Propp describes in theory, as opposed to what is present in his available annotations of stories. I observe that what Propp describes is more tightly constrained that what his annotations show in practice. In particular, Propp described pairs of functions which must either be both present or both absent from the annotation of a tale. However not all of his annotations strictly conform to these rules. As far as possible, I describe these constraints in the same sequential ordering that the character functions of a tale follow.

SEQUENTIAL ORDERING I place the constraint that there is a strictly uniform order in which character functions can occur. "The sequence of functions is always identical" (Propp, 2015, p. 22). Although Propp also states that there is some freedom within very

narrow and restricted limits. Existing work using Propp's analysis for the purpose of story generation has not enforced such a strict constraint over the ordering (Gervás, 2014). However, I do not consider that possibility here, as it would significantly increase the size of the search space.

INTERDICTION ⇒ VIOLATION The statement of a command to the Hero and its subsequent violation follow the rules of a logical implication: "Functions II and III form a paired element. The second half can sometimes exist without the first" (Propp, 2015, p. 27).

RECONNAISSANCE ⇒ DELIVERY The Villain making an attempt at reconnaissance and the result of this action form a similar pairing: "As in other similar instances, the second half of the paired function can exist without the first" (Propp, 2015, p. 29).

TRICKERY ⇒ COMPLICITY This pairing describes the relation Propp implies that if the Villain makes an attempt to deceive a character, that character will submit to this deception and unwittingly help the Villain.

VILLAINY ⊕ LACK Propp describes that each tale must include either one of these elements, and that they are in effect mutually exclusive. "it is necessary to choose a single element which is obligatory for all tales and to make the division according to its varieties. A (villainy), or a (lack) are the only such obligatory elements" (Propp, 2015, p. 102).

VILLAINY ⇔ LIQUIDATION The *Liquidation* function refers to the liquidation of the initial misfortune or lack. Propp states that "this function, together with villainy (A), constitutes a pair".

LACK ⇒ BEGINNING COUNTERACTION A tale motivated by the Hero lacking and seeking something implies the subsequent event whereby the seeker-hero agrees to or decides upon counteraction.

DONOR TESTS HERO ⇔ HERO REACTS In the sequence of character functions after the departure of the Hero (*Departure*), there are three functions which express the testing of the Hero by the Donor character. These cover: the testing of the Hero by the Donor, the Hero's reaction to this test, and finally the receipt (or lack thereof) of a magical agent by the Hero. The testing of the Hero, and the Hero's reaction to this test form paired elements.

HERO REACTS ⇒ RECEIPT The reaction of the Hero to the Donor's test implies the presence of the subsequent receipt event. I do

not enforce the constraint that the three donor functions necessitate each other as Propp does not expressly state this. Moreover, his appendix of annotations contains situations whereby a folktale may contain the receipt of a magical agent without the two prior events.

TRANSFERENCE $\Rightarrow$ RECEIPT Following the departure of the Hero (*Departure*), the sequence of character functions specifies the three Donor functions and then the *Transference* function. The *Transference* function specifies the spacial transference of the Hero from one place to another. It is implicit in Propp's analysis, and logical to a reader, that the transference of the Hero cannot directly follow the departure of the Hero. Thus the act of transference implies that at least one of the Donor functions must occur prior to this point.

STRUGGLE $\Rightarrow$ VICTORY A struggle between the Hero and Villain must be followed by the eventual victory of the Hero. However, these two functions do not mutually necessitate each other, as Propp describes ways in which instances of *Victory* can occur without a prior struggle.

PURSUIT $\Leftrightarrow$ RESCUE These two functions describe the pursuit of the Hero by the Villain, and the subsequent rescue of the Hero. It is implied, intuitively, that this pair of functions mutually necessitate each other.

RETURN $\oplus$ PURSUIT Propp implies that the return home of the Hero cannot co-occur with the pursuit and rescue functions. He states that the return function 'implies a surmounting of space' and does not need to be followed by another function indicating travel. He further adds that return can sometimes take the form of fleeing, and so it would not logically be followed by the pursuit and rescue functions.

UNRECOGNISED ARRIVAL $\Leftrightarrow$ RECOGNITION Functions 23 - 28[2] in the sequence reported by Propp describe the events pertaining to the False Hero character. These follow the unrecognised arrival of the Hero, through the presenting of unfounded claims by the False Hero, the proposal of a difficult task to the real Hero, the solution of this task by the Hero, and subsequent exposure of the False Hero. Propp states that the recognition of the Hero is "almost always preceded by an unrecognised arrival" (Propp, 2015, p. 62).

BRANDING $\Rightarrow$ RECOGNITION During the fight with the Villain, the Hero may be branded. Propp states that "recognition serves as a

2 These refer to the function numbers in Propp's original analysis, prior to the two additions I make.

function corresponding to branding and marking" (Propp, 2015, p. 62). However, as there are occurrences whereby the Recognition function occurs without prior branding, these functions do not necessitate each other.

DIFFICULT TASK ⇔ SOLUTION Propps analysis implies logically that the difficult task posed to the Hero and the Hero's solution of must occur together.

STRUGGLE ⊕ DIFFICULT TASK When discussing the classification of tales, Propp remarks that the struggle-victory pairing, and difficult task and its resolution are mutually exclusive, and these developments of events cannot occur within the same tale.

RECEIPT + TRANSFERENCE + STRUGGLE + PURSUIT This constraint specifies that at least one of the four following functions must be present in a tale: the receipt of a magical agent by the Hero, the transference of the Hero to a new location, the struggle between Hero and Villain, or the pursuit of the Hero. Along with the constraints already discussed, this ensures that a tale has some form of middle element and cannot simply be comprised of just one or two actions which do not describe a coherent story. This constraint is not based on Propp's description of character functions, however an analysis of his own annotations reveals that folktales must contain at least one of these elements.

LIQUIDATION + RESCUE + REWARD I place the constraint that at least one of the following must be present in a tale: liquidation of the initial lack (for folktales that follow a story about a lack), the rescue of the Hero from pursuit, or the rewarding of the Hero. This enforces the requirement that a tale must have an ending which completes the preceding sequence of events. "Terminal functions are at times a reward (F), a gain or in general the liquidation of misfortune (K), an escape from pursuit (Rs)" (Propp, 2015, p. 92).

Although not expressly stated by Propp, I place constraints to ensure that a representation of a tale expresses a beginning, middle and end. The motivation for this is to allow the constraint solver to reject strings of character functions which do not represent complete tales. This decision is reinforced by the findings of Gervás (2014) that in the generation of tales according to Propp's morphology, a relatively low percentage of tales had satisfactory endings. Requiring an instance of either *Villainy* or *Lack* provides a beginning element to the tale. To represent the end of a tale, I constrain the assignments of functions to contain an instance of either a rescue from pursuit, the Hero's return or the rewarding of the Hero. While the return or rewarding of the

Hero provides a natural end to the tale, Propp states that "A great many tales end on the note of rescue from pursuit" (Propp, 2015, p. 58). The forms that the midsection of a tale can take are, as is to be expected, exceptionally varied. I aim to keep this variability, while ensuring that some form of action is represented, by placing a loose restriction over the function assignments to make certain that one of the following four events is present: the testing of the Hero by the donor, the spatial transference of the Hero to the object of a search, the struggle between Hero and Villain, the pursuit of the Hero.

While I have discussed several diversions from a strict adherence to Propp's morphology, these modifications limit the number of outputs from what would otherwise be an infeasible search space. The intuitive restrictions I have imposed merely act to omit the nonsensical outputs, which would not represent a complete tale in any case.

5.3.3 *Application of Constraints*

For the task posed here, I consider the labelling of sentences in a tale with Propp's character functions. Here, each sentence in a tale represents a distinct variable which must take a value from the domain of all character functions, or a zero-value to indicate that the given sentence does not represent a character function. Character functions are given a unique integer identifier, meaning that each sentence must be labelled with an integer value from 0 to 33.

The narrative constraints discussed above are implemented in the CLPFD narrative structure reasoning system and a numerical representation of the text is passed as input to the system. A folktale is represented as a list of lists, where each element in the outer list is a variable corresponding to a sentence of the input text. The value of each element is a list of integers representing the domain of the variable, the character functions that a sentence may represent. With the application of the constraints, the domains of many variables are greatly reduced. Value assignments which would break the constraints and lead to nonsensical outputs are removed from their domains. All sequences of character functions which could represent a given tale while satisfying the imposed constraints are then enumerated by the program.

5.4 ANALYSIS

A formal evaluation of the narrative structure reasoning system and the value of Propp's structural analysis of folktales is performed via the studies of the following two chapters. However, here I provide a qualitative analysis of the results, and a case study to compare the outputs of the narrative structure reasoning system to the gold standard annotations of Propp.

5.4.1 *Qualitative Analysis*

With the constraints I have employed, it is not possible to reproduce Propp's own function assignment for every tale he analyses. Propp's description of narrative constraints is stricter than his own annotations contained in the appendices of *The Morphology of the Folktale*. The placement of some character functions is unclear, and some tales appear to be far different from what Propp's annotations would suggest (Finlayson, 2015). However it is necessary to enforce a strict interpretation of Propp's work in order to maintain a manageable number of results.

This process of determining structural information results in a large number of interpretations even for short tales which are only in the order of tens of sentences long. Below is the shorthand[3] for the ideal sequence of character functions for the tale *Nikita the Tanner*, as annotated by Propp[4]. The exact correspondence between these functions and the text of the tale can be seen in Section C.1.

$$\mathrm{A\ B\ C} \uparrow \mathrm{H\ I\ K} \downarrow \mathrm{W}$$

This system does correctly identify the above sequence of functions as one potential representation of *Nikita the Tanner*. But it also identifies a further 231 other possible function sequences for this tale which are also valid according to Propp's morphology. These range in length, representing the tale with between four and eleven functions. With the constraints described, no less than four character functions can represent a tale, as this would not represent a complete tale. Propp makes no mention of a minimum number of functions that should be used to represent a tale, however in his annotations he assigns no less than six functions to a single-move tale. Bod et al. (2012) make some observations in this manner about the number of functions that human annotators typically assigned to a story in comparison to Propp's own annotations.

While this process produces a significant number of interpretations for such a short tale, it represents only a tiny portion of the original search space, where each of 34 sentences could be labelled with a 0 or an increasing value between 1 and 33. Such a high number of interpretations may help to explain the low inter-annotator agreement that has been found in studies on the replication of Propp's annotations. It also indicates that Propp's morphology is under-constrained for the unambiguous annotation of stories in this way. Furthermore, this number of interpretations includes the fact that many outputs represent the same narrative structure with the same meaning, but the sentences tagged as representing each function differ. That is, once

3 These shorthand function names correspond to the function designations given in Table 1.

4 Propp's annotations do not include any of the Preliminary functions.

the zero-padding of the outputs is removed, the number of actual interpreted meanings of a tale is narrowed down further.

I observe that this approach can give rise to significantly different, but still meaningful interpretations of a single tale. One tale analysed by Propp, *The Witch*[5], tells the story of a boy who is kidnapped by a witch while he is out fishing, but eventually manages to escape with the aid of some geese and return home to his parents. Propp's annotations mark this as a tale about an act of villainy, with the subsequent pursuit and rescue of the Hero. My approach is able to detect this interpretation of the story, however it produces significantly different, but arguably correct interpretations in addition to this. One of these interpretations marks this as a tale about the boy's lack of fish and his wish to go fishing, which is granted by his parents. Although this may not be the intended interpretation of the story, the generated sequence of character functions which represent this does conform to Propp's Morphology and is a credible interpretation.

5.4.2 *Case Study*

In this section I compare the outputs of the narrative structure reasoning system to the ideal annotations of Propp for a single Russian Folktale. I consider the tale *Nikita the Tanner* as an illustrative example, and use the exact alignments between Propp's narrative functions and the texts as in Finlayson et al. (2015). In this folktale, the gold standard annotations for each function correspond to a single sentence. However, this is not the case for all tales. This folktale was chosen as an example as it is the shortest in the dataset being considered.

The narrative structure reasoning system produces 232 valid interpretations for this tale, with the given constraints. These interpretations were converted into a probability distribution in order to compare them to the gold standard annotations. The total probability for each Propp function is divided proportionally between the sentences which represent it, according to the produced interpretations. For example, if the Villainy function were to occur in sentence 1 for 174 interpretations of this tale, and sentence 2 for 58 of the interpretations. Then sentence 1 would be assigned a probability of 0.75 of representing the Villainy function, and sentence 2 would be assigned a probability of 0.25.

As already discussed, the gold standard annotations for *Nikita the Tanner* are comprised of 9 of Propp's narrative functions. Figure 8 compares the probability distributions for each of these, obtained via both gold standard annotations and the output of the narrative structure reasoning system. Each of these pairs of probability distributions is accompanied by a value for the cross entropy between the two dis-

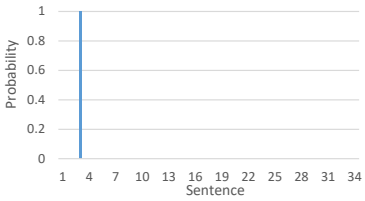
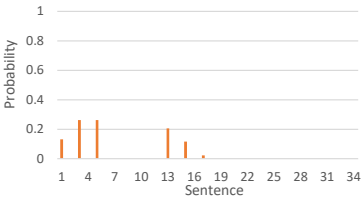
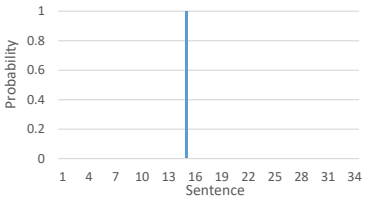
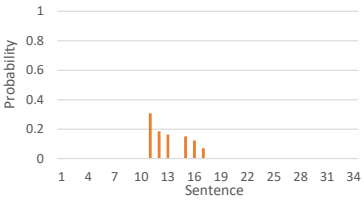
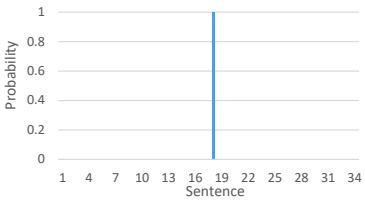
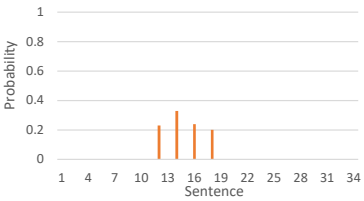
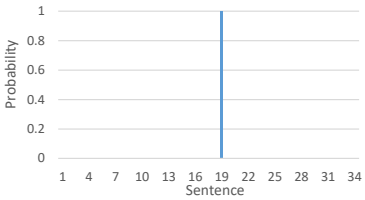
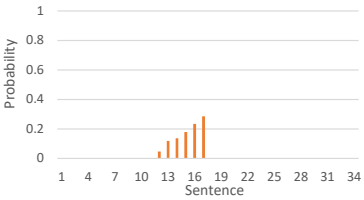
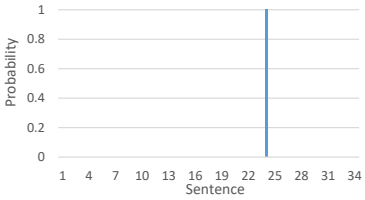
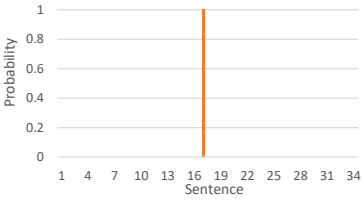
5 The full text of this tale can be found in Section C.9

tributions. In order to make this calculation possible, a small constant of 1×10^{-7} is added to all zero-probability events, and the probability of all other events is reduced in scale accordingly. It is worth noting for comparison purposes that the cross entropy value between any of these gold standard distributions and a uniform probability distribution is 5.08 for this 34 sentence story.

Figure 8 shows that for this tale, the narrative structure reasoning system most accurately predicts the occurrence of the Victory and Reward functions. The least accurately predicted functions are the Departure, Struggle, and Return functions. In these cases the system fails to create any interpretations whereby these narrative functions are assigned to the same sentences as in the gold standard annotations. The gold standard annotations tag both the Struggle and Victory as occurring in the same span of text. The narrative structure reasoning system perfectly recognises the Victory function. However the sentence which it tags as an instance of the Struggle function instead describes the hero being implored to fight, rather than the fight itself. The poor prediction of the Departure and Return functions is not surprising as they convey acts of movement, and the associated cue words are common throughout the stories. As such, it is a difficult task to accurately replicate the location of these functions as in the gold standard data using only the methods described. In addition to the outputs of the narrative structure reasoning system shown here, the system additionally detected the presence of four functions of this story which do not occur in the gold standard annotations. These are the three functions involving a Donor character testing the Hero, and the Punishment function.

One reason for the low predictive performance shown here is that the single set of gold standard function assignments is being compared to a probability distribution over 232 interpretations of the tale. Any pair of these interpretations may be mutually exclusive, either due to the narrative functions which they contain, or the relative points in the story at which they place these narrative events. This also highlights that Propp's Morphology is underconstrained for an exact replication of his annotations.

This case study provides some quantitative insight into capabilities of the narrative structure reasoning system. While the results may indicate that the system cannot predict the gold standard annotations with a high degree of accuracy, Figure 8 demonstrates that the system is able to identify the more general locations of important content in a text. Nevertheless, the utility of the narrative structure reasoning system will be shown via the analyses of the following two chapters. As these studies will demonstrate, while the narrative structure reasoning system may not accurately replicate the gold standard annotations, it does detect useful, summary-worthy elements of a text.

	GOLD STANDARD	SYSTEM OUTPUT	CROSS ENTROPY
VILLAINY			1.92
	(a) Villainy: gold standard assignment	(b) Villainy: narrative structure reasoning system assignment	
MEDIATION			2.73
	(c) Mediation: gold standard assignment	(d) Mediation: narrative structure reasoning system assignment	
COUNTERACTION			2.32
	(e) Beginning Counteraction: gold standard assignment	(f) Beginning Counteraction: narrative structure reasoning system assignment	
DEPARTURE			23.3
	(g) Departure: gold standard assignment	(h) Departure: narrative structure reasoning system assignment	
STRUGGLE			23.3
	(i) Struggle: gold standard assignment	(j) Struggle: narrative structure reasoning system assignment	

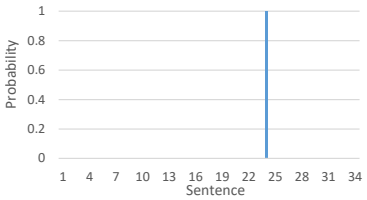
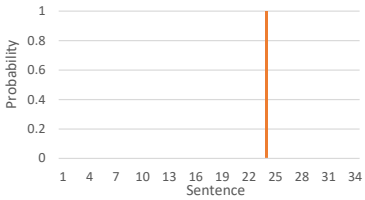
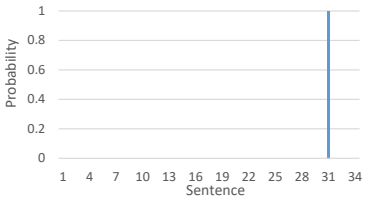
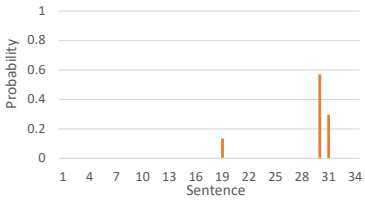
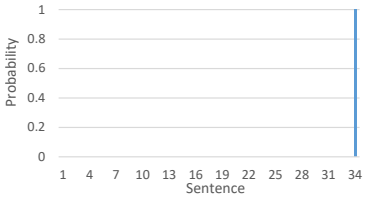
	GOLD STANDARD	SYSTEM OUTPUT	CROSS ENTROPY
VICTORY			0
	(k) Victory: gold standard assignment	(l) Victory: narrative structure reasoning system assignment	
LIQUIDATION			1.76
	(m) Liquidation: gold standard assignment	(n) Liquidation: narrative structure reasoning system assignment	
RETURN			23.3
	(o) Return: gold standard assignment	(p) Return: narrative structure reasoning system assignment	
REWARD			0
	(q) Reward: gold standard assignment	(r) Reward: narrative structure reasoning system assignment	

Figure 8: Probability distributions of Propp function assignments for the folktale *Nikita the Tanner* according to both gold standard annotations, and the output of the narrative structure reasoning system. Cross entropy values are given for each pair of probability distributions.

5.5 SUMMARY

In this chapter I have described a narrative structure reasoning system. This system produces every valid interpretation of a text according to a set of predefined constraints about the structure of that genre. The aim of this is to obtain high-level interpretations of the meaning of a text. I believe that this structural information can be used to simulate some of the decisions that psycholinguistics tells us human readers make about the summary-worthy nature of the content of a text.

I have reported how this system can be applied to the example genre of Russian Folktales, including methods for detecting the narrative elements and roles of the key characters. It is desirable to perform these annotations automatically due to the difficulties associated with obtaining accurate human annotations of Propp's morphology have been observed in empirical studies (Bod et al., 2012; Fisseni, Kurji, and Löwe, 2014). I have detailed the constraints placed on the structure of instances of this genre, based on my interpretation of the work of Propp (2015).

This genre was chosen as Propp's analysis provides the most constrained definition of the narrative structure of a genre that I found. Moreover, stories require comparatively little world knowledge and contain relatively simpler text, ameliorating their automatic parsing. Structural information also goes some way to accounting for world knowledge. World knowledge acts to provide information about what a given event entails, as well as likely outcomes. Propp's analysis provides this information via the descriptions of narrative functions, in addition to their interrelations and constraints.

The evaluation of this system in its own right is non trivial. Here I have only provided a qualitative analysis of the results, and a case study for a single folktale. In the following chapter I demonstrate how structural information relating to Propp's Morphology aids the detection of summary-worthy content. The manual annotation of these features provides the largest benefit, but the utility of the automatic annotations via the methods discussed in this chapter is still clear. I further show the value of Propp's constraints; that Propp's analysis has use beyond simply providing a list of useful cue words. Then in Chapter 7 I analyse how this structural information affects the summaries generated as the final product of this thesis.

6

THE DETECTION OF SUMMARY-WORTHY CONTENT

In this chapter, I discuss the creation of statistical models which aim to recognise the summary-worthy content of a text. I train models on a binary classification task: the prediction of sentences containing summary-worthy information. These models are trained on a set of discourse-structural and semantic features to predict sentences marked as summary-worthy by human annotators. The purpose of this is two fold. First, the statistical models learned by this work are used to inform the selection of content used to produce summaries. This provides an objective way of determining the relative importance of the various annotated features to the indication of summary-worthy content. Second, this provides a means by which to evaluate the utility of the annotated features. This forms the basis of an answer to one of my primary research questions: whether knowledge about the discourse structure of a text aids the recognition of summary-worthy content.

Summarization evaluation is both difficult and subjective. The work carried out for the creation of these models also allows multiple comparisons to take place. I compare the predictive capabilities of the models created when the same set of training features are generated via varying degrees of automation. This enables an examination of the impact of automatic coreference resolution, as well as provide an extrinsic evaluation of the narrative structure reasoning system over the domain of Russian Folktales. Additionally, this chapter compares the predictive abilities of these features to the abilities of existing extractive summarization systems. The validity of such a comparison is presented prior to these results.

6.1 APPROACH

The features used in this study are the combination of annotated semantic features described in Chapter 4, and the structural annotations of Propp's Morphology. I begin this section with a brief report of each of these features, followed by a description of the dataset used and how it was annotated with these features. Multiple sets of training data were created in order to compare the automatic detection of certain features with their manual annotation. I then explain how the ground-truth values for which sentences contain summary-worthy information were obtained. Finally, I detail the method employed for determining the utility of the discussed features.

6.1.1 *Annotated Features*

My overall approach to summarization is a semantic, abstractive approach: building structured semantic representations of the content and using these to generate text summaries. Given this, my interest here is to look at the efficacy of observable features that give a deeper, more human-like, insight into the meaning of a document. Of the 43 features examined, 35 were based around the annotation of character roles and narrative units according to Propp's Morphology. I consider many of these features to carry more semantic information than those commonly used by existing summarization systems. It is worth emphasising however, that this analysis is not focused on the merits of a particular feature, or the work of Propp. Instead, it focuses on the usage of semantic and structural features in general, and comparing them to the capabilities of more traditional approaches. The features, and the reasoning behind their inclusion, are given below.

SENTENCE POSITION The relative position of a sentence within a story. This feature encodes the percentage distance of a given sentence through the story, where the first sentence has a score of 0 and the final has a score of 1. This is used rather than absolute sentence number so that the feature can be studied independently of story length. Sentence position is a commonly used feature in extractive approaches to summarization. It proves especially useful for certain types of document, such as news articles, where the beginning of a document essentially provides a summary of the content.

NUMBER OF CHARACTER MENTIONS The number of noun-phrases in a given sentence which co-refer with the characters of the story. I include this feature to examine whether a high density of character mentions can be used to indicate important events and thus potentially summary-worthy content.

NUMBER OF UNIQUE CHARACTERS MENTIONED The number of unique story-characters mentioned in a given sentence. This feature performs a subtly different role from the number of character mentions. When creating very short summaries only the key characters will be mentioned, and so a high number of distinct characters may act as an indicator of non summary-worthy content.

NUMBER OF LEXICAL CHAINS The number of unique lexical chains that have items present in a given sentence. Lexical chains capture parts of the cohesive structure of a text, helping identify the importance of the salient concepts mentioned throughout a story.

SPEECH This is a binary flag indicating whether a sentence contains direct speech or not. Stories, and especially summaries, tend to be written as reported facts; the presence of direct speech may therefore be an indicator against a sentence containing summary-worthy content.

HERO A binary flag indicating whether a sentence contains a reference to the hero of the story. This is based on the character role identification methods discussed in Chapter 5. A summary is more likely to include events involving the main characters of a story, and so the presence of the hero may be a strong indicator of summary-worthy information.

VILLAIN A binary flag indicating whether a sentence contains a reference to the villain of the story. This is based on the character role identification methods discussed in Chapter 5. A summary is more likely to include events involving the main characters of a story, and so the presence of the villain may be a strong indicator of summary-worthy information.

PROPP FUNCTION WEIGHTS Sentences are annotated with 33 distinct features, each corresponding to one of Propp's narrative functions.[1] The total score for each Propp function feature is 1 if the story contains the function, else 0. A score of 1 is shared proportionally between the sentences which represent that function as a probability distribution. For example if the *Villainy* function is expressed over two sentences in a story, each of these sentences will receive a score of 0.5 for this feature, and all other sentences will have a 0 score for this feature. I initially treated the presence of each Propp function as a binary feature, but this led to worse results.

NUMBER OF PROPP FUNCTIONS The total number of Propp functions that a given sentence is a part of. As Propp functions span multiple sentences it is not uncommon for there to be some overlap. This means a sentence may be part of the expression of multiple Propp functions simultaneously. Equally, some sentences are simply required to form a cohesive story and will not represent any Propp functions whatsoever. Such sentences are unlikely to be summary-worthy; conversely, sentences at the point of overlap of multiple Propp functions may be important. I initially treated this as a binary feature rather than a count, however this led to worse results.

DESIRING VERB This is a binary feature indicating whether or not a given sentence contains a verb indicating desire. The list of such

1 As discussed in Chapter 5, I make two additions to the 31 character functions defined by Propp. The first represents his description of an 'Initial Situation'. The second comes from explicitly splitting of *Villainy* and *Lack* into two distinct functions.

words is compiled from the VerbNet synsets for desirous verbs such as 'want' and 'long'. Statements of desire by the hero or villain of a tale often act as the motivating event for the subsequent story events, and psycholinguistic research indicates the importance of such causal actions

GOAL PHRASE This is a binary feature indicating whether or not a given sentence contains a goal based phrase such as 'in order ...' or 'so that ...' These phrases stem from the adverbial phrases of purpose discussed in the semantic annotations of my system. In a similar manner to desiring verbs, these phrases indicate purpose and thus may signify summary-worthy sentences.

6.1.2 *Training and Testing Data*

I annotated every sentence in 10 of the Russian folktales[2][3] studied by Propp (Propp, 2015) with the semantic and discourse features discussed. Given the annotations of Propp's narrative functions and coreference information, all other features can subsequently be annotated automatically. Stories were manually annotated with Propp's narrative functions according to his own labels (Propp, 2015), and the exact sentence boundaries provided in the dataset of Finlayson (2015) and Finlayson et al. (2015). Coreference resolution was carried out manually as described in Chapter 4. Coreference information is used in the determination of the hero and villain, the number of noun-phrase mentions in a given sentence, and the number of unique characters mentioned in a sentence.

In addition to this manual annotation of the dataset, I separately stored automatically obtained annotations for both Propp's narrative functions and coreference information. This allowed for the creation of multiple versions of the dataset, annotated with the same set of features but according to a varying degree of automation. Performing this step enabled a study into the impact of obtaining these features automatically: what knock-on effects do errors in the automatic approaches introduce? Given the time consuming nature of manual annotations, the benefits to obtaining this information automatically are evident. Each of these datasets is described below.

2 The full text of all Russian folktales used in this thesis can be found in Appendix C. The 10 tales used for training were: *Nikita the Tanner*, *The Magic Swan Geese*, *The Crystal Mountain*, *Shabarsha the Laborer*, *Ivanko the Bear's Son*, *Frolka Stay at Home*, *The Witch*, *The Seven Simeons*, *Prince Danila Govorila*, and *The Merchant's Daughter and the Maidservant*

3 The stories used as training and testing data for this thesis differ from those used in my published work (Droog-Hayes, Wiggins, and Purver, 2019). As such the values of results presented here will differ slightly, however the trends are the same. This change occurred after publication as I discovered that some stories held out of this study for testing purposes contained Propp function features unseen during the training stage.

MANUAL Both coreference resolution and the assignment of Propp functions were carried out manually.

AUTO-CORENLP Stanford's CoreNLP system (H. Lee et al., 2013; Manning et al., 2014) was used for automatic coreference resolution; Propp function assignment carried out manually. Coreference information impacts the determination of the following features: NUMBER OF CHARACTER MENTIONS, NUMBER OF UNIQUE CHARACTERS MENTIONED, HERO, and VILLAIN.

AUTO-SPACY As Auto-CoreNLP above, but using the neuralcoref (Wolf, 2017) extension to spaCy for coreference resolution.

AUTO-PROPP Coreference resolution was performed manually, but Propp's character functions were assigned automatically using the methods described in Chapter 5. This produces every valid interpretation of a given tale which conforms to Propp's Morphology. These interpretations are used to create a probabilistic distribution of character function assignments for each sentence of a story, and obtain the scores for the PROPP FUNCTION WEIGHTS features. The method used to produce this distribution is very similar to the calculation of Propp function weights from manual annotations. The only difference is that a score of 1 is not divided equally among the sentences which represent a function, it is weighted based on the number of interpretations in which a given sentence represents a given function. As the interpretations created by the narrative structure reasoning system can be conflicting, I did not expect this approach to be mature enough to be a substitute for manual Propp function assignments; I assess that expectation further on.

PROPP-CORENLP Entirely automatic feature extraction, using automatic character function assignments, and automatic coreference resolution performed by Stanford's CoreNLP. Automatic coreference resolution directly affects the determination of the hero and villain. This subsequently impacts the determination of valid interpretations of the story by the narrative structure reasoning system.

PROPP-SPACY Entirely automatic feature extraction, using automatic character function assignments as above, but with automatic coreference resolution performed by neuralcoref for spaCy.

6.1.3 *Ground-Truth Data*

Two different ground-truth datasets were created, one of short summaries and the second of long summaries, as follows. Annotators were first given each story to read in its entirety before beginning

the annotation of its sentences. For the short summary ground-truth data, the assessors were asked to mark sentences as summary-worthy only if they were essential to convey the main events of a story. For the long summary ground-truth data, assessors were asked to additionally mark sentences as summary-worthy if they contained information about noteworthy events in the narrative chain of the story. Sentences were marked as summary-worthy or not in this way by three independent assessors working in isolation. Following this, the three assessors were brought together for a session in which they discussed the labelling of any sentences where there was a disparity between their annotations until a full consensus had been reached.

In total, 947 sentences were annotated. Of these, 126 sentences were marked as summary-worthy for short summaries, and 225 were marked as summary-worthy for long summaries. The summary-worthy sentences for short summaries are a proper subset of those selected for the long summaries.

6.1.4 *Method*

To analyse the utility of these features, the data was set up as a classification task: I try to predict which sentences are marked as summary-worthy by the annotators. I used a logistic regression classifier, as implemented in the Weka toolkit (Hall et al., 2009), and evaluated performance via 10-fold cross-validation to prevent over-fitting. These folds are class-balanced, created from the dataset in its entirety as opposed to leave-one-story-out.

For this task of detecting summary-worthy sentences, both false positive and false negative predictions are damaging. In other words, while it is not desirable to misclassify a summary-worthy sentence as unimportant, the impact of misclassifying a non-summary-worthy sentence should not be understated. A summary not only loses some of its utility with a lack of concision but may also lose cohesion. For this reason, recall metrics are not used to evaluate this study. Instead, results are analysed based on scores obtained by Cohen's Kappa coefficient. This statistic measures the agreement between two or more raters, and takes into account the possibility of the agreement occurring by chance. In this study, agreement is measured between the ground-truth values and the values predicted by logistic regression models. A Kappa score of 0 indicates the performance expected by chance, not the incorrect classification of all items. I use Cohen's Kappa rather than classification accuracy as this is a class-imbalanced binary classification task. For the short-summary ground truth values, the target class corresponds to only 13% of the overall dataset. For the long-summary ground truth values, this only increases to 24%. With this in mind, it would be trivial to achieve a classification

Table 11: Kappa scores for the prediction of short and long summary-worthy content under different algorithms.

ALGORITHM	SHORT	LONG
ZeroR	0.00	0.00
Naive Bayes	0.16	0.19
Sequential Minimal Optimisation	0.12	0.17
K-Nearest Neighbours	0.16	0.19
PART	0.25	0.28
RepTree	0.16	0.08
J48	0.28	0.09
Logistic Regression	0.28	0.27

accuracy of 87% for the short-summary dataset, and 76% for the long-summary dataset with nothing more than a majority-class classifier.

The choice of a logistic regression classifier was made after first comparing the baseline performance of a variety of different algorithms. By this I mean the performance of the algorithms with their default parameters and no consideration of any cost-sensitive parameters which I examined later. Table 11 shows the Kappa scores for predicting short and long summary-worthy sentences for 8 different classification algorithms. These are: ZeroR, Naive Bayes, Sequential Minimal Optimisation, K-Nearest Neighbours, the PART rule based algorithm, the decision tree algorithms REPTree and J48, and Logistic Regression. As the results show the best overall performance was obtained by the logistic regression algorithm. In addition to giving the best results, the logistic regression models have interpretable results; a qualitative discussion on the coefficients of the learned models shall be given in the results section of this chapter.

It is interesting to consider how well the predictive abilities of the learnt logistic regression models compare to the performance of existing extractive algorithms. Four different extractive summarization algorithms were implemented to act as comparison points for this study. To begin, it is important to elucidate why this is a fair comparison to make. Extractive summarization systems use some criteria to rank and select sentences which are then concatenated to form a summary. In other words, such systems make a judgement about whether or not each sentence in a document is summary-worthy. While my overall approach to summarization is abstractive, this chapter focuses on the prerequisite step of determining whether a sentence *contains* summary-worthy content. Once these sentences have been identified, the corresponding summary-worthy content in the representation of the input text can be transformed into a representation of the summary. It is from this summary representation that the subse-

quent summary generation process occurs. As such, my prerequisite step to summary generation also involves making a decision about the summary-worthy nature of each sentence in a document. Given the ground-truth dataset, both my approach and existing extractive approaches can be compared for their ability to predict summary-worthy sentences.

I compare my method against the following extractive approaches: Luhn's original summarization algorithm (Luhn, 1958), TextRank (Mihalcea and Tarau, 2004), LexRank (Erkan and Radev, 2004), and a Latent Semantic Analysis (LSA) approach (Steinberger and Jezek, 2004). Luhn's summarization algorithm is based on word frequency and the proximity of significant words within a sentence. Both TextRank and LexRank are graph-based approaches to summarization where a graph is constructed by creating a node for each sentence in the document and the edges are based on the semantic similarity of sentences. The TextRank algorithm creates edges between sentences based on the number of words that they have in common. LexRank creates edges based on the cosine similarity of TF-IDF vectors of two sentences. The LSA approach to summarization analyses the relationships between the sentences of a document and the terms that they contain. The sentences which best represent each important concept are then identified and joined together to form a summary.

6.2 RESULTS

Here I provide a comparative analysis of my results in three sections. First, I discuss the utility of each discourse feature for predicting summary-worthy content. This comparison is based on manual feature extraction (i.e. the *Manual* dataset in Section 6.1.2). Then I compare the results obtained by this manual feature extraction with the automatic feature extraction. Finally, I compare my approaches to predicting summary-worthy sentences with the performance of several different extractive summarization algorithms. It is important to understand that the focus of this analysis is *comparative*, and not about the absolute values of the results.

6.2.1 *Utility of Features*

6.2.1.1 *Feature Values*

Table 12 shows the coefficients obtained by logistic regression models for predicting short and long summary-worthy sentences. These coefficients correspond to two separate regression models, one for the prediction of short summaries and the other for the prediction of long summaries. I am aware of the shortcomings of bald comparisons of such numerical weights. I use them here only to aid a qualitative dis-

cussion regarding the utility of each of these features. Table entries prepended with 'F-' correspond to individual Propp functions. Five of Propp's functions are not displayed in this table as they do not occur at all in the 10 stories used in this training dataset, and as such each of them has a coefficient of 0.

From Table 12 it can be seen that the coefficients with the greatest positive or negative values correspond to Propp's narrative functions. In the prediction of both long and short summary-worthy sentences, the highest positive-valued coefficient corresponds to the *Violation* function. This function describes a character disobeying a previous command that has been given to them (the *Interdiction* function), such as going to a forbidden location. Additionally, Propp describes this as the point of a tale at which the villain is introduced. The importance of such an element of a tale is clear, as it is this violation of a command that provides the opportunity for subsequent events to occur. This function has an even higher valued coefficient for the long summary data.

As previously stated, many Propp functions form pairs, and the *Interdiction* and *Violation* functions are one such example. It can be observed that for paired functions the coefficient for the later function always has the higher value. The first function of a pair may even have a negative coefficient; its presence is an indicator against a sentence being summary-worthy. In fact, the coefficients corresponding to the *Interdiction* function have the greatest negative values. The *Pursuit* and *Rescue* functions are another clear instance of this. I attribute this trend to the fact that the meaning of the first function is implicit even when only the second half is present. For example, a rescue implies a prior chase just as a violation of a command implies the prior stating of the command. This is reflected in the ground-truth data, where sentences representing the second half of each paired function are more often marked as summary-worthy than those representing the first half. This is especially true for the short summary data, where the first half of each pair can be seen as somewhat superfluous.

Relatively high valued coefficients can also be observed for the *Villainy* and *Lack* functions. Propp describes these as very important aspects of the tale, being the means by which the actual movement of the tale is created. The act of villainy or a character's lack (this is typically expressed as a desire for wealth or marriage) is the only one of Propp's elements that he explicitly says must be present in each tale. The values for *Lack* are comparatively low, but this can be explained by observing that 8 out of the 10 tales are about acts of villainy rather than a lack. In the ground truth data, at least one sentence marked as summary-worthy in each story is annotated as being part of either the *Villainy* or *Lack* function.

Table 12: Values of coefficients for each feature used to train logistic regression models for the prediction of summary-worthy content over both the short and long summary datasets.

FEATURE	SHORT	LONG
Sentence Position	-1.00	-0.01
No. of Noun-phrase Mentions	0.22	0.27
No. of Unique characters	-0.10	0.08
No. of Lexical chains	0.57	0.41
Speech	-0.62	-0.51
Verb Goal	1.19	0.80
Phrase Goal	0.89	2.66
Hero?	0.72	0.23
Villain?	0.93	0.82
Propp function	-0.10	-0.25
F-Initial situation	1.80	2.28
F-Absentation	-16.57	1.52
F-Interdiction	-147.27	-41.42
F-Violation	48.48	56.0
F-Delivery	19.45	14.81
F-Trickery	-0.63	-0.29
F-Complicity	-50.05	14.76
F-Villainy	21.12	14.83
F-Lack	2.63	1.36
F-Mediation	1.46	1.68
F-Beginning counteraction	1.53	1.08
F-Departure	1.67	0.96
F-Donor tests Hero	1.79	4.16
F-Hero reacts	0.97	2.54
F-Receipt of magical agent	1.63	2.20
F-Transference	2.51	2.82
F-Struggle	1.02	0.33
F-Victory	1.90	1.81
F-Liquidation	4.09	3.86
F-Return	3.50	3.13
F-Pursuit	-12.98	6.41
F-Rescue	23.61	15.78
F-Unrecognised arrival	2.03	14.03
F-Recognition	20.05	14.43
F-Exposure	17.32	12.29
F-Transfiguration	0.19	-0.43
F-Punishment	23.18	14.48
F-Reward	2.76	2.13

I observe that certain coefficients with low, or negative, values for the short summary data have higher values for the longer summary data. Propp functions such as *Initial situation* and *Absentation* are background elements to a tale, with little importance to a short summary. However the coefficients for features such as these rise for the prediction of long summary data. I attribute this to the inclusion of additional events in longer summaries, which are unnecessary for summarizing the main points. Other features such as *Phrase Goal* and the number of noun-phrase mentions also receive a higher weight. These increases could be attributed to the inclusion of more explanatory sentences, and the inclusion of a greater number of characters in a summary. It is intuitive that a short summary will focus primarily on the protagonist and antagonist, while a longer summary will introduce more characters. This is further supported by the increase in the importance of the following functions for the creation of long summaries: *Donor tests Hero*, *Hero reacts*, and *Receipt of magical agent*. These three functions are closely linked, as they are the only elements which express the involvement of the Donor character, who tests and subsequently rewards the hero. A character fulfilling such a role is only mentioned in sentences marked as summary-worthy for the ground truth data for one of the ten stories for short summaries.

The *Propp function* feature, which indicates the number of Propp functions that a given sentence is a part of, has a low negative value. Overlapping narrative functions tend to each span multiple sentences. Thus the features corresponding to those functions have low values, and individual sentences that are part of these events have low importance. Additionally, by far the most common overlapping functions correspond to the donor functions. Sentences containing these functions do not tend to be marked for inclusion in summaries, and have low valued coefficients. As a result, it is not surprising that this feature is a slight indicator against summary-worthy content.

Aside from Propp's functions, the benefit of the other discourse features is also evident. The importance of phrase goals is particularly evinced by its coefficient for the prediction of long summary data. In addition, my hypothesis about the low importance of speech-containing sentences is supported by the low-negative speech coefficient for both short and long summaries. However the importance of the number of unique characters mentioned in a sentence, and the total number of noun-phrase mentions in a sentence appears to be relatively low. Although it is understandable that the value of these coefficients would be higher for long summary data, where more characters would commonly be included in the resulting summary text. Similarly, the relative position of a sentence within the overall text appears to provide little information, unlike in other domains such as news summarization.

Table 13: Kappa scores for the prediction of summary-worthy content with the removal of groups of annotated features.

FEATURE REMOVED	SHORT	LONG
None	0.280	0.273
Sentence Position	0.286	0.290
No. of Noun-phrase Mentions	0.267	0.253
No. of Unique characters	0.283	0.284
No. of Lexical chains	0.267	0.217
Speech	0.286	0.262
Hero?	0.286	0.273
Villain?	0.289	0.255
Verb Goal	0.268	0.250
Phrase Goal	0.274	0.264
Propp Features	0.063	0.177

6.2.1.2 *Feature Ablation*

In order to further study the utility of these features to the prediction of summary-worthy content, a feature ablation was performed. This followed a leave-one-out method whereby the logistic regression models were retrained with one feature held out at a time. The 33 features for indicating the presence of individual Propp functions, and the feature to indicate the number of Propp functions in a sentence were grouped together and counted as a single feature for this process. All other features were treated individually. While other logical groupings of features exist, such as those indicating the presence of the hero and villain, these groupings become increasingly subjective and so were not considered. Furthermore, while it is evident that some of these features interact with each other, it was not feasible to retrain models with every possible subset of features. Even when grouping the features related to Propp's narrative functions into a single unit, there remain 2^{10} possible subsets of these features.

Table 13 shows the kappa scores obtained for the prediction of both short and long summary-worthy content when individual features are removed. As a comparison point, the *None* row indicates the baseline performance of this model with all features present. It is important to note that the utility of each of these features was originally examined before resplitting the set of training and test data used for these studies. In the original set of training data, every feature had a positive impact on the kappa score obtained by these models. This study was performed as a post hoc analysis after the studies of the subsequent chapter had been completed.

In this split of training data, it appears that the features *Sentence Position*, and *No. of Unique characters* have a slight negative impact on the predictive capabilities of the logistic regression models. For the prediction of short summary-worthy content, the *Hero* and *Villain* features also have a slight negative impact on the kappa scores obtained. However, these results alone may not give a true indication of the utility of the features analysed. While the individual removal of the *Hero* and *Villain* features lead to higher kappa scores, their joint removal leads to a decrease in kappa scores. Removing these two features together results in a kappa scores of 0.274 and 0.253 for the prediction of short and long summary-worthy content respectively. These figures both reflect decreases on the baseline performance of the models. I suggest that the increased kappa scores obtained from the individual removal of these features may arise from the information shared with the Propp function features. For example, an indication of a fight between the hero and villain implicitly holds the same information as these features. However, there will be references to these characters throughout the texts which are not also tagged as being an instance of any narrative function.

These results indicate that the features related to the narrative structure of the text provide the greatest predictive capabilities. As with the previous study discussing the value of the coefficients of these features, this is more true for the prediction of short summary content than for long summary content. The presence of lexical chains appears to be a far more important feature for the prediction of long summary-worthy content than for short. The goal related features additionally appear to carry more weight for the prediction of content for the longer summaries.

6.2.2 *Automatic Performance*

Here I compare the results of predicting summary-worthy content based on manual annotations with varying degrees of automatic annotation. As already stated, given coreference information and assignments of Propp's narrative functions to the sentences of a tale, all other features can be subsequently annotated automatically.

Coreference information impacts the determination of the following sentence features: number of noun-phrase mentions, number of unique characters, Hero, and Villain. As well as the incorrect resolution of referents, not all noun-phrase mentions will necessarily be picked up by automatic coreference resolution. Incorrect referencing will impact the indicators of whether or not a given sentence mentions the hero and villain of a story, as well as affecting the determination of which characters fulfil those roles. Automatic Propp function assignment impacts the 33 features each corresponding to a

Table 14: Kappa scores for the prediction of summary-worthy content with varying degrees of automation.

SETTING	SHORT	LONG
Manual	0.280	0.273
Auto-CoreNLP	0.270	0.235
Auto-spaCy	0.270	0.210
Auto-Propp	0.057	0.169
Propp-CoreNLP	0.033	0.162
Propp-spaCy	0.035	0.147

Propp function, as well as the feature indicating the number of Propp functions present in a given sentence.

For the analysis of automatic annotations, I examine the Kappa scores that are obtained by logistic regressions performed over the various versions of the dataset. These are shown in Table 14. The Short and Long columns each correspond to the prediction of sentences marked as summary-worthy for short summaries and long summaries respectively.

Table 14 shows that the best Kappa scores for both short and long summary data are achieved when the dataset has been annotated manually. The 'Auto-spaCy' and 'Auto-CoreNLP' rows show how the use of automatic coreference resolution information causes a decrease in these scores, with CoreNLP performing marginally better than spaCy. The 'Auto-Propp' row shows the effect of performing coreference resolution manually, but automatically obtaining the assignments of Propp functions. From this it is clear that the effect of inaccurate information about the discourse structure of a story has a greater impact than poor coreference information. Finally, 'Propp-spaCy' and 'Propp-CoreNLP' show the results of the fully automatic prediction of summary-worthy content with spaCy and CoreNLP coreference resolution systems respectively. As can be seen by the poor performance of fully automatic approaches, some manual annotation is still desirable for this task until the accuracy of the requisite systems has improved.

It is evident from these results that the automatic assignment of Propp's functions is more damaging to the formation of short summaries than long summaries. I believe this is because knowledge about the discourse structure of a story is more critical in the creation of shorter summaries. As shorter summaries are necessarily more condensed and must only cover the most key aspects of a text, it is more important to know where in the text the key narrative events occur. As Table 12 shows, the other annotated features hold a greater

Table 15: Optimised Kappa scores for the best predictive performance of the system according to manual and fully automatic annotations.

	SHORT	LONG
Manual	0.344	0.316
Propp-CoreNLP	0.099	0.239
Propp-spaCy	0.135	0.258

importance in the generation of longer summaries, and so there is less reliance on the accuracy of Propp function annotations.

When considering these results, there are several factors which help to explain the poor predictive abilities of the automated annotations. While Propp provides a detailed list of rules and constraints for the structure of Russian folktales, they are not strict enough that only a handful of valid interpretations can be produced. For some tales it is even arguable whether the key event is an act of villainy, or some lack of the hero (Propp expressly states that either one of these events must occur in a tale). As such, many interpretations of a tale can be produced which are in contradiction to one another, rather than simply a shifting of the exact sentences in which a function occurs. As a result, this can lead to probabilistic distributions of Propp functions where either the score for a function is split across a large number of sentences, or the distribution includes the presence of functions which are mutually exclusive according to Propp. In addition, errors stemming from the automatic coreference resolution systems can propagate and affect multiple annotated features. Coreference information is used to determine the characters which fulfil the roles of the hero and villain, as well as information about the number of characters mentioned in each sentence of a story.

As I am most interested in the correct prediction of the small class of positive instances of summary-worthy sentences in this imbalanced dataset, I further examined the use of a weighted cost function in the logistic regression classifier. This was implemented via a cost-sensitive logistic regression classifier in Weka, which penalises the incorrect classification of summary-worthy sentences more harshly in training. I initially used a cost parameter which was inversely proportional to the size of each class (summary-worthy vs. not summary-worthy) for both the short and long summary datasets separately. However, this did not result in the optimal Kappa score so I optimised the cost weight parameter separately for both the short and long summary datasets, to give the best Kappa score over cross-validation.

Table 15 shows the best Kappa scores that can be achieved by this system with cost-optimisation of both the manual and completely

Table 16: Optimised Kappa scores for the best prediction of summary-worthy content by existing extractive algorithms.

ALGORITHM	SHORT	LONG
Luhn	0.095 (90%)	0.095 (50%)
TextRank	0.061 (40%)	0.123 (40%)
LexRank	0.067 (70%)	0.050 (50%)
LSA	0.118 (50%)	0.223 (50%)

automatic annotation of data[4]. When using a weighted cost function, spaCy outperformed CoreNLP for the prediction of summary-worthy sentences over both short and long datasets. Comparing this with the results of Table 14, it can be seen that this optimisation process especially benefits the prediction of short summary content by manually annotated data and the prediction of long summary content from automatically annotated data.

6.2.3 *Extractive Comparisons*

Here I examine the predictive abilities of four classic extractive summarization algorithms (Luhn, TextRank, LexRank, and LSA) over the same dataset of Russian Folktales. These algorithms were chosen for ease of implementation, and to compare approaches that have been developed many years apart. Table 16 shows the performance of these algorithms for the prediction of sentences marked as summary-worthy for the formation of both short and long summaries.

In the interest of comparing against the best possible performance of these algorithms, they were individually optimised, both for the prediction of the long and short ground truth data. Each of the four algorithms examined produces a ranking of sentences in the input document, rather than classifying each sentence as summary-worthy or not. The bracketed percentages in Table 16 indicate that the highest-ranking n percentage of sentences were selected from each story in order to form a summary. These values were optimised over the entire dataset in order to obtain the highest Kappa scores. This means, for example, that the highest-ranked 40% of sentences were selected from each story in order to obtain the best Kappa score of 0.061 for the prediction of short summary data by the TextRank algorithm.

The results show that of the four algorithms tested, the most semantically driven approach, Latent Semantic Analysis, performs best for the prediction of both short and long summaries. It is inter-

4 The following numbers indicate the cost that was applied to penalise the misclassification of summary-worthy sentences: Manual+Short 2.6, Manual+Long 2.0, Propp-CoreNLP+Short 2.6, Propp-CoreNLP+Long 1.7, Propp-spaCy+Short 2.3, Propp-spaCy+Long 2.6.

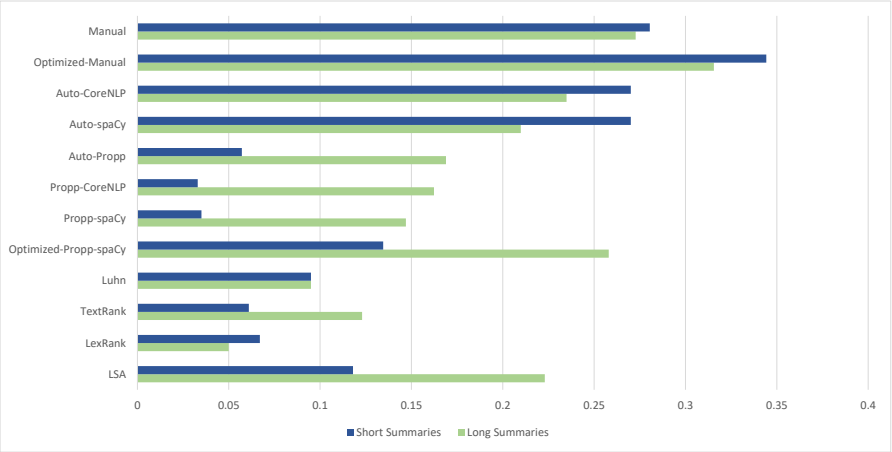


Figure 9: A comparison of the Kappa scores obtained by all experiments, across both short and long summary data.

esting to note, however, that the relatively simple approach of Luhn compares well against these far more complex and computationally intensive approaches devised many decades later. This provides some additional support to the belief that there is little further to be gained by research in extractive summarization.

Figure 9 provides a clear comparison between the predictive performance of the approaches examined for this task. Moving down the bars of the graph, it can be seen that an increased level of automation leads to a reduction in performance. The largest impact on performance comes from the automation of assigning Propp's narrative functions to the stories, and the impact is most pronounced for the prediction of short summary-worthy content. This is intuitive, as shorter summaries will primarily focus on the key events of a story, and so inaccuracies in the information regarding narrative structure will have a greater impact. In contrast, longer summaries will contain additional information and will have comparatively less reliance on this structural information.

The differences in performance between the logistic regression models learnt from semantic and discourse-structural features, and the tested extractive systems can be clearly seen in Figure 9. From this, the benefit of annotating sentences with semantic and discourse features is evident. When using an optimisation parameter even the automatically annotated data (Optimised Propp-spaCy) clearly outperforms every extractive method examined for the prediction of both short and long summaries. This particular result is surprising to me, given my awareness of the errors that can propagate through these annotations via incorrect coreference information and assignment of Propp's narrative functions. The benefit of these discourse features is even more evident when comparing the results of the optimised manual annotations with the scores obtained by the extractive algorithms.

6.2.4 *An evaluation of the Narrative Structure Reasoning System*

In the previous chapter, I described a narrative structure reasoning system, which automatically determines possible interpretations of a text that conform to a given set of rules. While Propp provides a lot of rules and constraints on the structure of Russian Folktales, these are still not enough to build an unambiguous interpretation of a tale. On the one hand this freedom can allow for creativity.

Due to the multitude of outputs that can be produced by the narrative structure reasoning system, it is somewhat unclear how an intrinsic evaluation of it can be carried out. As such, I perform an extrinsic evaluation of the performance of this system on the domain of Russian Folktales in two parts. The first is to assess that, for the example domain of Russian Folktales, the application of rules in the narrative structure reasoning system provides benefit. The second is to assess how it affects the generation of summaries, which shall be discussed in the next chapter.

The preceding sections have shown that the largest impact to the automatic performance of the system is automatic Propp function assignment. The purpose of this evaluation is to show that the rules and constraints of Propp's Morphology have value, and thus that the narrative structure reasoning system has value. Propp describes the majority of the narrative functions he identifies in terms of cue words. As such it is worth evaluating the value of Propp's work beyond just providing identifiers for these narrative functions. As an example domain, this shows the value of the narrative structure reasoning system to successfully reduce the set of possible interpretations of a text and can be used to more accurately detect summary-worthy sentences.

The narrative structure reasoning system takes as input every potential instance of each of Propp's functions and applies a set of constraints. Earlier in this chapter I have shown the results of using the output of this process in a probabilistic distribution as features to aid the detection of summary-worthy content. Here then, I evaluate the use of the *input* to this process, effectively skipping this step in my approach to summarization.

Table 17 shows the Kappa scores for the prediction of summary-worthy content based on a probabilistic distribution over the potential presence of Propp functions without the application of any constraints. The *Auto-Propp* results are given again for convenience. The table shows the results both with and without the use of a weighted cost function. Table 17 shows that both with and without optimisation, the output of the narrative structure reasoning system is better at the prediction of summary-worthy content than the input.

Table 17: Kappa scores for the prediction of summary-worthy content, based on fully automatic annotations, before and after applying Propp's constraints.

	SHORT	LONG
Cue-Propp	0.029	0.158
Auto-Propp	0.057	0.169
Optimised-Cue-Propp	0.119	0.253
Optimised-Auto-Propp	0.147	0.270

6.3 SUMMARY

This chapter provides an analysis of the benefit of using discourse-structural and semantic features in the detection of summary-worthy content for summarization. I test a variety of machine learning algorithms and find logistic regression models to perform best in this task. These models are created in order to facilitate the detection of summary-worthy content in unseen story data for summary generation, however this work also enabled an analysis of multiple elements.

I provide a comparative analysis of results in three sections; the value of individual features, the knock-on effects of inaccurate automated annotations, and a comparison between the performance of this system and four different extractive approaches to summarization. These results show not only the value of using discourse features in summarization, but also the improvements they can give over standard extractive approaches. I have shown that with optimisation even an imperfect, but fully automatic, approach to annotation can outperform these extractive methods. In particular, I have shown the value of knowledge about the discourse structure of a text for selecting content to form summaries.

In addition, I have provided a first analysis of the narrative structure reasoning system. This has shown that the narrative structure reasoning system provides benefit, and that Propp's work has value beyond simply providing lists of indicators for narrative units. The evaluation of generated summaries in the following chapter will enable a further analysis of this work.

For this study, it is important to consider the risk of overfitting. Overfitting a regression model occurs when too many parameters are estimated from the data. Here 43 semantic and discourse-structural features were used to train logistic regression models over what is a relatively small dataset. In this work the sample size could not be increased due to the scarcity of gold standard Propp function annotations for single-move tales. In addition, the reason for such a large number of parameters comes from the fact that 33 of them each cor-

responded to Propp's narrative functions and so had to be included in their entirety or not at all in order to maintain relevance. However, 5 of these narrative functions never occurred in the dataset used and so were never annotated.

While the results presented in this chapter show promise, I believe that there is one crucial way in which they could be improved. In this work, the value of each sentence is primarily examined in isolation. In other words, the decision about the summary-worthiness of a sentence does not consider the context provided by the surrounding sentences. This is somewhat mitigated by the annotation of Propp's narrative functions, which relate to a story as a whole. However some Propp functions such as *Pursuit* and *Rescue* tend to span multiple sentences, and the individual value of a sentence annotated as such is very low without the surrounding context. In addition the lexical chain feature helps consider elements of the story as a whole by identifying the salient concepts. However it is not natural for a human, either as a reader or summarizer, to judge the value of a sentence in a story without considering its context. I leave this open as an area of future research to improve upon these content selection results. The results also indicate the negative impact of poor automatic coreference resolution and Propp function detection. Due to the time required to perform these annotations manually, there is a clear benefit to investigating improvements that can be made to the automation steps.

7

THE GENERATION AND EVALUATION OF SUMMARIES

In this chapter I present my approach to the generation of summaries, and detail their evaluation. I begin by describing my approach to the Transformation and Generation stages of the ITG model of summarization. I investigate the generation of summaries to a desired length, summarizing hierarchically and abstracting to suitably fit this length constraint. As such, these two stages of summarization are highly intertwined. The desired summary length directly impacts the selection and transformation of story content into summary content. Building on the detection of summary-worthy content reported in the previous chapter, I describe the application of these models to the transformation of a story representation into a summary representation. Following this, I detail my procedure for generating natural language summaries from this intermediate summary representation.

In the second part of this chapter I describe my procedure for evaluating the produced summaries. These abstractive summaries can be generated according to either the manual or automatic annotation of coreference information and narrative structural information. Both of these types of abstractive summary are compared to extractive summaries whose generation is informed by the same set of features, and extractive summaries produced by the LSA summarizer discussed in the previous chapter. Summaries generated according to all four of these generation procedures are evaluated at two different target lengths.

7.1 SUMMARY GENERATION

In the previous chapter I showed the value of discourse features to the detection of sentences *containing* summary-worthy content. In extractive summarization, these features are simply used to inform sentence selection. In abstractive summarization, sentences expressing related concepts are condensed and expressed in a new way. In this section I describe the processes involved in both the transformation of a text representation into a summary representation, and the subsequent generation of a natural language summary from this. The process of content selection is informed by the logistic regression models trained to predict summary-worthy content described in the previous chapter. The output summaries are generated to a desired length, as a compression ratio of the length of the input text. The shorter the desired output, the more the content *abstraction* occurs in compari-

son to *extraction*. Here the compression ratio is measured in terms of sentences, relative to the number of sentences in the input text. This is in contrast to some existing work which generates summaries up to a given word limit. This decision was made as it is more natural to consider generation in terms of well-formed sentences given that it has been shown that human assessors take linguistic quality into account. Additionally the focus of this work is on story understanding and so only simple generation methods, at the sentence level, are considered.

7.1.1 *Content Scoring*

The intermediate representation of an input text is comprised of AMR fragments: parses of individual sentences. These are each annotated either manually or automatically with coreference information and Propps narrative functions. All other semantic and discourse-structural features previously discussed can then be automatically annotated on the basis of this information. A summary representation is then created based on the reductive transformation of this space. Abstraction introduces new elements into the summary representation, in the form of AMR fragments merged into a new single concept.

The first step in the transformation process is to score the content of each AMR fragment in the meaning representation of the input text. These scores are indicative of summary-worthy content that should be included in the summary representation. To obtain these scores, I use the coefficients of the kappa-optimised logistic regression models trained to predict summary-worthy content from Chapter 6. Each fragment is assigned a score of summary-worthiness based on the logistic regression probability function given below.

$$P(\text{summary-worthy}) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 * x_1 + \ldots + \theta_{43} * x_{43})}}$$

Here, θ_0 corresponds to the y-intercept. The subsequent θ values correspond to the coefficients learned by the logistic regression models for each of the 43 annotated structural and semantic features. The x values paired with each θ correspond to the observed value for the given feature in a given AMR parse.

For this process, either the set of coefficients for short summary-worthy data or long summary-worthy data are used based on the desired summary length. On average, sentences marked as summary-worthy for the formation of short summaries in Chapter 6 corresponded to 13% of the original text length, and 24% for long summaries. When it is desired to produce summaries of less that 19% of the length of the input, the coefficients for short summary data are used, else the coefficients for long summary data are used. The

coefficients based on manual or automatic annotations are used as appropriate. In other words, one of four different sets of coefficients are used to score content depending on the annotations used and desired summary length.

While this formula gives probability values, they are treated as absolute scores in this process. Content which is compressed and abstracted always results in a single sentence. The summary-worthiness of such a sentence is calculated as the sum of the summary-worthy scores for each of its constituent AMR fragments. As such, it is possible that the score for a compressed sentence will not represent a valid probability. This is a valid procedure, as my approach only uses these scores to determine the relative importance of each fragment, and never treats them as probabilities.

7.1.2 *Content Abstraction*

In this approach to summarization, abstraction occurs via knowledge about the structure of a given domain of texts. In this thesis I use Russian Folktales as an example domain, using the structural information provided by Propp's *Morphology of the Folktale* (Propp, 2015). Any given narrateme may be represented across multiple sentences of the input story. These instances translate to the representation of narratemes across multiple AMR fragments in the meaning representation built from the interpreting the input text. Propp's narrative functions each describe an abstracted event of a story. This information can be used to abstract away from the text of the input story, and compress content into a more concise form.

Just as with individual AMR fragments, the importance of these compressed structures for inclusion in a summary must be determined. Abstracted content is not necessarily more summary-worthy than individual AMR fragments. To obtain summary-worthy scores for compressed content, further logistic regression models were created in a similar manner to those described in Chapter 6. Over the same set of 10 folktales used as training data, every occurrence of a narrative function covering multiple sentences was replaced with a single sentence representing the abstract event. These sentences were based on Propp's brief description of the corresponding narrative functions, and the characters they necessarily include. Over this set of compressed stories, three annotators were again asked to mark the sentences which contain information necessary to form a short or a long summary[1]. Logistic regression models were then trained to predict summary-worthy sentences over the same set of 43 annotated features. For reference, the performance of these models is shown in

1 This set of compressed training data contains 605 sentences. 112 sentences were marked as summary-worthy for the formation of short summaries, and 203 sentences were marked as summary-worthy for the formation of long summaries.

Table 18: Kappa scores for the prediction of summary-worthy content over compressed sentences.

	SHORT	LONG
Manual	0.401	0.365
Optimised-Manual	0.490	0.439

Table 18. The coefficients of these models could then be used to score the summary-worthiness of compressed content[2].

These new logistic regression models were only created for the dataset manually annotated with Propp's narrative functions. The output of the narrative structure reasoning system does not result in the unambiguous automatic annotation of sentences, only a probabilistic distribution of Propp function assignments to each sentence. As such, there is no clear method for determining which sentences should be compressed into a single narrative function. Furthermore, there is no clear distinction between the narrative functions that a given sentence is most indicative of. For these reasons, the coefficients from manually annotated compressed sentences were used to score the summary-worthiness of compressed sentences for automatic approaches too.

7.1.3 *Transformation Procedure*

An operation procedure is performed on the lowest scoring content until the the size of the summary representation falls to the desired length threshold. That is, the summary representation is created via a reductive transformation: removing and compressing content from the meaning representation of the input text. This is performed as opposed to the constructive process of transferal and compression of the highest scoring content until the desired length is reached for reasons that shall become evident.

The operation procedure involves either the removal of an AMR fragment from the meaning representation, or the abstraction of multiple related fragments into a new single fragment. This process is performed repeatedly until the size of the meaning representation is reduced to meet the desired length threshold. The cost of operating on a given AMR fragment is calculated as the sum of the scores of every fragment that represents the same Propp function. In cases where an AMR parse does not represent any Propp functions, or it is the only parse to represent a particular function, the cost of operation is equal to the score of that fragment. In cases where multiple

2 For the scoring of both individual AMR fragments and abstracted content, the coefficients from the kappa-optimised variants of the respective logistic regression models were used.

AMR parses represent a single Propp function, the cost of operating on any of them is set as the sum of the summary-worthy scores of the individual parses. The cost of operating on a fragment created from a prior abstraction operation is calculated with the use of the coefficients from the logistic regression models discussed in Section 7.1.2.

Summaries are created in this way, via the reduction of content rather than the transferal of the most important content, in order to produce summaries which are more cohesive. The content marked as least summary-worthy is operated on until the summary length falls below the desired length threshold. Content abstraction occurs, but only when it is necessary. The aim of this is to retain as much of the original information as possible and maintain textual cohesion. In contrast, constructing a summary representation from the ground up would involve selecting the most important content in descending order. While this results in a greater level of abstraction, summaries produced in this way will be comprised primarily of highly compressed content, scattered with sentences regarding very particular elements of the input text. This produces summaries which contain content described at varying levels of granularities, reducing textual cohesion.

7.1.4 *Generation Procedure*

Prior to sentence realisation, a linear ordering is placed on the content of the summary representation. Both AMR fragments and indicators of abstracted content are placed in sequence to follow the order of the events of the input text which they represent. Abstracted events, which may correspond to multiple sentences of the input text, are placed in order according to the position of their first occurrence (using information from either manual or automatic annotations as appropriate).

The desired length of a summary affects the size of the summary representation which is produced. From this, a single sentence is generated for each item in the summary representation. A sentence is created for each AMR fragment in the meaning representation according to a simple generation and compression procedure. Sentence templates are used to generate original sentences for content which is abstracted in the transformation process.

7.1.4.1 *AMR Sentence Generation*

Sentences are generated from the AMR fragments in the summary representation with the use of word alignments, and sentence compression techniques. The automatic AMR parser used in this work, JAMR (Flanigan, Thomson, et al., 2014; Flanigan, Dyer, et al., 2016), provides alignments between the nodes of an AMR parse and spans of text from the sentence it corresponds to. The tree structure of each

AMR parse is flattened into a sentence, using the word alignments to insert the necessary particles to produce grammatical sentences.

Before flattening an AMR parse, any non-essential sub-trees are removed. Existing research has investigated the removal of non-essential information from the Universal Networking Language (UNL) representation. The Universal Networking Language (Uchida, M. Zhu, and Della Senta, 1999) is a formal language designed to represent the semantic data of documents. Previous research on the summarization of UNL documents has proposed the removal of non-essential elements from the UNL representation, and defined heuristics for pruning their content (Sornlertlamvanich, Potipiti, and Charoenporn, 2001; Martins and Rino, 2002). In a similar manner, I identify and remove non-essential elements of AMR parses. This is based on the list of the non-core roles of parses provided in the AMR Guidelines (Banarescu et al., 2018). However, the generation process only removes the child nodes involved in a subset of these relations. These are 'concession', 'location', 'mod', 'manner', 'time'. I find the other 'non-core' roles such as 'name' and 'purpose' to be essential for the generation of meaningful and grammatically correct sentences.

Coreference information, either manual or automatic, is used to further improve the generated sentences, and make them less dependent on the surface text of the input story. Pronouns are substituted with the head-word of the entity that they refer to in order to ensure the resulting summary is coherent.

7.1.4.2 *Template Sentence Generation*

Natural Language Generation is a significant field of research in its own right. As it is not the focus of this thesis, templates are used to guide the generation of sentences from abstracted content. Abstraction occurs via knowledge about the structure of a genre of texts; related content is compressed into an abstracted description of a key narrative event. Using the example domain of Russian Folktales, sentence templates are created to represent each of Propp's narrative functions.

Below is a listing of the templates prepared for each of Propp's narrative functions. To reiterate, a given folktale will only include a subset of these functions. Each of these templates is based on the short description Propp provides at the beginning of his explanation as to the purpose of each narrative element. The terms in italics denote template items which are filled in via the manual or automatic annotation of the input text. Theses templates were designed such that they would be meaningful in the event that any given template filler cannot be identified. This enables the desirable abstraction and generalisation of content, while still using story-specific information.

INITIAL SITUATION *The initial characters* are introduced.

β *The hero* leaves home.

γ *The hero* is told *a command*.

δ *The hero* takes no notice of the command.

ϵ *The villain* makes an attempt at reconnaissance.

ζ *The villain* receives information about *the victim*.

η *The villain* deceives *the victim*.

θ *The victim* submits to the deception.

VILLAINY *The villain* performs a *villainous act*.

LACK *The hero* lacks *something important*.

B *The hero* is approached with a request.

C *The hero* decides upon counteraction.

↑ *The hero* departs on the quest.

D *The hero* is tested by a magical donor, *donor*.

E *The hero* reacts to the test by *the donor*.

F As a result, *the hero* acquires the use of a magical agent.

G *The hero* is transported to a new location.

H *The hero* fights the villainous *villain*.

J In the fight, *the hero* is branded by *the villain*.

I *The hero* defeats *the villain*.

K *The hero* overcomes the problem.

↓ *The hero* then returns home.

PR *The villain* pursues *the hero*.

RS *The hero* is rescued from the pursuit.

O Unrecognised, *the hero* arrives.

L The false hero presents unfounded claims.

M A difficult task is proposed to *the hero*.

N *The hero* resolves the difficult task.

Q At last, *the hero* is recognised.

EX The false hero is exposed.

T *The hero* is given a new appearance.

U *The villain* is punished.

W Finally, *the hero* is rewarded.

In addition to the detection of the *hero* and *villain* of the tale described in Chapter 5, the character fulfilling the role of *donor* is identified for the generation of content. The donor character cannot easily be identified at the same time as the hero and villain; it is not uncommon for the donor character to fulfil an addition role, such as that of villain. Furthermore, the detection of the donor relies on information about the presence of any of the three donor narrative functions (D, E, F), rather than informing their detection. An instance of a Russian Folktale does not have to contain these functions, and thus does not have to contain a donor character, but it does have to include a hero and a villain. If the determined narrative structure of a given tale includes reference to at least one of the donor functions, an attempt is made to recognise the character who fulfils this role. The role of donor is fulfilled by the character who has the most direct interactions with the hero over the spans of text representing the donor functions. This is detected via coreference information and the AMR parses of the sentences representing the donor functions.

The character fulfilling the role of the *false hero* is not detected. The functions involving this character (L, EX) occur rarely in the folktales. However, the structure of the text representing these functions can vary greatly. For example, it is non-trivial to distinguish between the character presenting unfounded claims and the character these are being presented to, due to the wide variety of sentence constructs which can express this event.

Aside from the detection of characters fulfilling particular roles, the semantic annotations present in the meaning representation of the input text are used to fill the other templates. Here I provide a brief description of how each of these are identified.

INITIAL CHARACTERS The characters enumerated in the initial setting of the story are detected with coreference information, either manual or automatic.

A COMMAND The template for *a command* refers to the imperative given to the hero in the Interdiction function. With the knowledge of the AMR fragment representing this function, the command can be detected in the same manner as the Interdiction function itself: the AMR parse representing this function is a sentence containing direct speech, where the main verb proposition is missing its first argument, the subject. The word alignments between the parse and its corresponding sentence are used to identify the verb proposition and its child nodes.

THE VICTIM This is the same as the hero character. This distinction is only maintained so that the template makes sense in its own right.

VILLAINOUS ACT After the detection of Propp's functions (either manually or automatically), the villainous act can be identified as the main verb proposition of the corresponding AMR parse.

SOMETHING IMPORTANT This is detected as the second argument, the object, of the main verb in the AMR parse tagged as an instance of the Lack function.

7.2 SUMMARY EVALUATION

In this section, I provide an evaluation of the summaries generated by the summarization approach laid out in this thesis. I start with an explanation of the types of summary that are compared, and provide an example of each. Following this I describe the set of summaries used in evaluation, and the evaluation procedure. Finally, I report the outcomes of this study and discuss the results.

7.2.1 *Approach*

7.2.1.1 *Summary Types*

In this evaluation, four different methods of summary generation are compared. The approach to summarization laid out in this thesis is compared on the basis of both manual and automatic annotations. This allows for an evaluation of the approach to summarization using discourse-structural and semantic features independently of the impact of any errors stemming from automatic annotations. A third, extractive summarizer is also compared. This uses the same set of discourse-structural and semantic features for the sole purpose of scoring and ranking the sentences of the input text. The top ranking n sentences are concatenated until the desired length threshold is reached. In addition to these approaches, the best performing extractive summarizer tested in Chapter 6, the LSA summarizer of Steinberger and Jezek (2004), is used as a baseline comparison point. These methods, and the shorthand that shall be used to refer to them are given below.

MANUAL The approach to abstractive summarization presented in this thesis, using manual annotations for coreference resolution and the assignment of Propp functions.

AUTOMATIC The approach to abstractive summarization presented in this thesis, using automatic annotations for coreference res-

olution (spaCy), and the automatic assignment of Propp functions via the narrative structure reasoning system.

EXTRACTIVE An extractive approach to summarization, using automatic annotations for coreference resolution (spaCy), and the automatic assignment of Propp functions via the narrative structure reasoning system to rank and select the most summary-worthy sentences.

LSA The Latent Semantic Analysis summarizer used as a comparison point in the detection of summary-worthy content in Chapter 6 is used here as a baseline.

7.2.2 *Examples*

An example summary generated according to each of these four procedures is given below. These are summaries of the folktale *Nikita the Tanner*, the full text of which can be found in Section C.1. The summaries presented here are generated with a length threshold of 10%. That is, each summary is generated to be no longer than 10% of the length of the input text. The full listing of summaries used in this study can be found in Appendix D.

MANUAL

The princess, the Tsar, and the dragon are introduced.
The dragon seized the princess and dragged her to his lair but did not devour her, because the princess was a beauty.
Nikita, having done his heroic deed, would not accept reward, but returned to currying hides.

AUTOMATIC

A dragon performs a villainous act.
Nikita defeats A dragon.
Nikita overcomes the problem.

EXTRACTIVE

Finally, it was the fate of the tsar's daughter to go to the dragon.
The princess had a little dog that had followed her to the dragon's lair.
She would attach her letter to the dog's neck, and the dog would take it to them and even bring back the answer.
At that moment Nikita was currying hides and held twelve hides in his hands; when he saw that the tsar in person had come to see him,

he began to tremble with fear, his hands shook, and he tore the twelve hides.

LSA

Instead, he took her to wife.
At that moment Nikita was currying hides and held twelve hides in his hands; when he saw that the tsar in person had come to see him, he began to tremble with fear, his hands shook, and he tore the twelve hides.
But no matter how much the tsar and tsarina entreated him, he refused to go forth against the dragon.
Nikita, having done his heroic deed, would not accept any reward, but returned to currying hides.

The first sentence of the *Manual* summary and all three sentences of the *Automatic* summary demonstrate the use of templates. In the automatic coreference resolution of this story, the literal string detected for the villain character is *A dragon*, which is reflected in the associated summary. The *Extractive* summary is created by concatenating the sentences ranked as having the highest summary-worthy scores. The difference between this summary and the *Manual* and *Automatic* summaries demonstrates the effect of operating on the lowest scoring content first. Important content that forms part of narrative events is recognised, abstracted, and given a higher summary-worthy score.

7.2.2.1 *Materials*

A set of 15 Russian folktales were considered in the evaluation of this system. Of these, 10 were used in the training of logistic regression models, and 5 were held out[3] purely for testing purposes. As stated in the previous chapter, the training set was constructed so that there were no unseen Propp functions in the test set. A larger test set was not possible, due to the availability of gold standard annotations from Propp (2015) and Finlayson et al. (2015). Furthermore, this thesis only considers what Propp terms *single-move* tales in order to limit the output of the narrative structure reasoning system.

For each of these 15 tales, summaries of two different lengths were generated according to all four of the described generation methods. Summaries were generated at both 5% and 10% of the length of the input. These lengths were chosen in order to limit the time needed for participants to complete the evaluation task. The 15 tales ranged in length from 34 sentences to 156 sentences, with an average length of

3 The full text of all Russian folktales used throughout this thesis can be found in Appendix C. The tales used in training are referenced in Chapter 6. The tales held out for testing were: *Bukhtan Buktanovich*, *The Runaway Solider and the Devil*, *Ivan Popyalov*, *The Serpent and the Gypsy*, and *Dawn, Evening, and Midnight*.

101 sentences. Each study consisted of one tale and two summaries, giving an average reading time of approximately five minutes.

7.2.2.2 *Method*

In this thesis I have highlighted the issues with the most common summarization evaluation metric, ROUGE, and argued against the automatic evaluation of summaries at the present time. To evaluate the method of abstractive summarization presented in this thesis, I set up a two-alternative forced choice study. This removes some of the subjectivity of human-based evaluations, although it is still likely that participants take linguistic quality into account when judging the summaries.

Each participant was presented with the full text of a Russian folktale followed by the statement:

> Read the following two summaries of this story carefully. Then select the checkbox for the summary that you think better conveys the salient points of the story.

Below this, two summaries were presented side by side. These summaries corresponded to the given tale, and were generated according to the same length parameter. This setup resulted in 180 variants of the study. There are 6 combinations of summary pairs for each of 2 different summary lengths, giving 12 summary pairs for each of 15 folktales.

This study was hosted online, and each participant could take part in the study as many times as they liked. The variant of the study shown to a participant was loaded from a queue. The stories were queued in a repeating sequence in order of their numerical index, but the order of the 12 summary pairs for each story was randomised. The left-right ordering of each summary pair was further randomised. This ensures that any habits of participants to always click left/right are lost as noise in the results. As such, a participant would have to take part 15 times in order to see the same folktale, and 180 times to see the same folktale with the same summary pairing.

7.2.3 *Results*

Here I present the results of two separate studies. In the first study, a total of 540 responses were recorded from unpaid participants via a standalone website, providing 3 results for each variant of the study. The second study was carried out by paid participants via Amazon's Mechanical Turk (MTurk) crowdsourcing platform. In this study, 1440 responses were recorded, providing an additional 8 results for each variant of the study. While the content of these two studies is identical, a separation is maintained between the sets of results as the

reliability of Mechanical Turk has previously been called into question.

Although these studies provide a small number of responses for each specific comparison of folktale, summary pair, and summary length, they can be aggregated in various ways. Of primary interest is the preference participants exhibited for each summary type. As such, the results for individual folktales can be combined. However, in both studies, a separation of results is maintained between the folktales used for training logistic regression models and those held out for testing in this study. While the summarization system presented in this thesis consists of many more steps than the trained logistic regression models, this split allows for any differences in results to be observed with complete transparency. Furthermore, an evaluation is performed to examine whether there is any significant difference in the preferences of participants across the two summary lengths that were tested.

7.2.3.1 *Web Results*

Table 19 and Table 20 show the percentage of participants who selected one summary type over another, according to each summary pairing. Table 19 aggregates these results across the 10 folktales used in the training of the logistic regression models, giving 30 data points per comparison. Table 20 aggregates these results across the 5 folktales held out for this study, giving 15 data points per comparison. The combined results across both the training and test splits are given in Table 21. The columns *Preference-Short* and *Preference-Long* refer to the results for each of the two lengths of summary used in this study. Each result refers to the percentage of participants who preferred the first summary type in the *Comparison* column to the second. This ordering of comparisons does not reflect the study, where the order of summary types shown to participants was randomised. The values in these tables appended with an asterisk indicate a statistically significant preference at p=0.01, according to both the binomial and χ^2 tests. That is, a rejection of the null hypothesis that there is no significant difference in preference for one summary type over another.

The order of preference for summary types over the split of folktales used in training logistic regression models and used solely for this study are given below.

TRAINING & SHORT: Manual > LSA > Extractive = Automatic

TRAINING & LONG: Manual > LSA > Extractive > Automatic

TESTING & SHORT: Manual > LSA > Automatic > Extractive

TESTING & LONG: Manual > Automatic > LSA > Extractive

Table 19: Percentage preference for the first summary type in each pair, over the 10 folktales used in training logistic regression models. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	97%*	97%*
Manual	Extractive	97%*	100%*
Manual	LSA	80%*	97%*
Automatic	Extractive	50%	43%
Automatic	LSA	27%	33%
Extractive	LSA	33%	40%

Table 20: Percentage preference for the first summary type in each pair, over the 5 folktales held out for this study. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	100%*	93%*
Manual	Extractive	87%*	100%*
Manual	LSA	100%*	80%
Automatic	Extractive	87%*	87%*
Automatic	LSA	47%	60%
Extractive	LSA	40%	40%

Table 21: Percentage preference for the first summary type in each pair, over all 15 single-move folktales. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	98%*	96%*
Manual	Extractive	93%*	100%*
Manual	LSA	87%*	91%*
Automatic	Extractive	62%	58%
Automatic	LSA	33%	42%
Extractive	LSA	36%	40%

Across all datasets, the approach of this thesis with manual annotations was always preferred. Following this, the LSA baseline approach is ranked in second place for all but long summary setting over the 5 folktales held out for testing. Across the training set, extractive summaries formed from my approach were preferred at least as often as summaries generated by my abstractive approach with automatic annotations. In contrast, the abstractive summaries generated with the use of automatic annotations were preferential to extractive summaries formed from automatic annotations across the 5 folktales held out for testing.

Additional χ^2 tests were performed to discover if there was a significant difference in summary preference between the two possible summary lengths shown to participants. These tests did not reveal any statistically significant difference in the preferred summary type across the two different summary lengths at either p=0.05 or p=0.10. In this setup, a higher p-value provides greater confidence of the similarity between the two sets of results. This result can be used to indicate that summary length did not factor into the preferences of participants.

7.2.3.2 *Mechanical Turk Results*

Amazon's Mechanical Turk is a crowdsourcing platform where Requesters upload Human Intelligence Tasks (HITs) to be performed by Workers for a monetary reward. As of 2010, Amazon reported half a million registered Workers from over 190 countries (AWS, 2011). The vast number of workers on the platform allows for the fast collection of results at a smaller cost than traditional surveys (Bentley, Daskalova, and B. White, 2017). However, the reliability of Mechanical Turk has been previously questioned, due to the use of automation and bots by MTurk workers (Dreyfuss, 2018). Nevertheless, studies have shown that differences in results obtained via MTurk as opposed to more traditional methods are so small as to have no practical consequences (Bartneck et al., 2015; Bentley, Daskalova, and B. White, 2017).

Several measures were employed in order to reduce the risk of low quality results due to bots, and the random selection of answers by workers. MTurk allows requesters to place restrictions on the workers who can access a task, such as their location and prior HIT experience. I required that workers must have completed at least 1000 prior tasks, and had a 97% acceptance rate of their work. I further required that the workers must be registered in a primarily English speaking country: UK, Ireland, USA, or Canada. Furthermore, the study page shown to MTurk workers included a CAPTCHA to prevent bots. Lastly, the order of every summary pair was swapped for half the tasks. That is, for each summary pair in the 180 variants of the study, four participants were shown A-B, and four were shown

Table 22: Percentage preference for the first summary type in each pair, over the 10 folktales used in training logistic regression models. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	71%*	79%*
Manual	Extractive	86%*	90%*
Manual	LSA	80%*	76%*
Automatic	Extractive	70%*	69%*
Automatic	LSA	54%	64%
Extractive	LSA	43%	33%*

Table 23: Percentage preference for the first summary type in each pair, over the 5 folktales held out for this study. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	83%*	73%*
Manual	Extractive	83%*	85%*
Manual	LSA	70%	88%*
Automatic	Extractive	78%*	85%*
Automatic	LSA	55%	65%
Extractive	LSA	15%*	30%

B-A. This means the results are balanced against any bias of participants to only click the left or only click the right option.

Table 22 and Table 23 show the percentage of participants who selected one summary type over another, according to each summary pairing. Table 22 aggregates these results across the 10 folktales used in the training of the logistic regression models, giving 80 data points per comparison. Table 23 aggregates these results across the 5 folktales held out for this study, giving 40 data points per comparison. The combined results across both the training and test splits are given in Table 24. The columns *Preference-Short* and *Preference-Long* refer to the results for each of the two lengths of summary used in this study. The values in these tables appended with an asterisk indicate a statistically significant preference at p=0.01, according to both the binomial and χ^2 tests.

Table 24: Percentage preference for the first summary type in each pair, over all 15 single-move folktales. Results appended with an asterisk indicates a statistically significant preference at p=0.01.

COMPARISON		PREFERENCE	
		SHORT	LONG
Manual	Automatic	75%*	75%*
Manual	Extractive	85%*	87%*
Manual	LSA	77%*	80%*
Automatic	Extractive	73%*	74%*
Automatic	LSA	54%	64%*
Extractive	LSA	33%*	32%*

Across both the training and testing splits of folktales, and across both summary lengths that were tested, the order of preference for summary types is as follows: Manual > Automatic > LSA > Extractive. In this second study, the abstractive summaries generated with manual annotations were again selected in preference to all three other summary types. The abstractive summaries generated on the basis of automatic annotations were selected in preference to both the LSA baseline summaries and extractive summaries formed on the basis of the discourse-structural and semantic features used in the generation of the abstractive summaries. However, the baseline Latent Semantic Analysis extractive summaries were preferred to the extractive summaries generated in this work.

Again, χ^2 tests were performed to discover if there was a significant difference in summary preference between the two possible summary lengths shown to participants. These tests only revealed a significant difference for one case: the strength of summary type preference for the Extractive-LSA comparison over the set of 5 folktales used only for this study at p=0.05. There was no significant difference in preference for this comparison when considering all 15 folktales together. For all other comparisons, there was no significant difference in the strength of summary preference with regards to the length of summary shown.

7.2.3.3 *Discussion*

While the results of the first study indicated a lack of preference for abstractive summaries generated based on automatic annotations, this is not the case in the second study performed on Mechanical Turk. There is no clear explanation for this difference in the preference ranking of summary types. A lack of preference for summaries generated with the use of automatic annotations (the Automatic and Extractive summary types) would be unsurprising given how errors in these

processes have knock-on-effects in summary generation. Inaccuracies in automatic coreference resolution propagate to impact the determination of the hero and villain, and subsequently the determination of structural information. Furthermore, the head-words assigned to coreferential mentions are not always informative. In some cases pronouns are assigned as the head-words of entities, which impacts the linguistic quality of the Automatic type summaries when mentions are substituted for their head-word. It has been previously shown that humans will judge summaries on the basis of linguistic quality even when instructed not to do so (Conroy and Dang, 2008). Furthermore, without stricter constraints, the narrative structure reasoning system is unable to determine a single, unambiguous interpretation for the instances of Russian Folktales used in this work.

Across both studies, little preference was shown for the Extractive-type summaries. These summaries are formed by selecting the highest scoring sentences according to the set of discourse-structural and semantic features. As shown by the example summaries of Section 7.2.2, the selection of the highest scoring content leads to summaries which differ greatly from those formed via content abstraction. The Manual and Extractive summaries both exhibit a similar level of linguistic quality, but the Manual (and Automatic) summaries demonstrate the value of recognising the importance of related content and expressing it concisely.

The results of Chapter 6 show that the method used to select the content of Extractive summaries better detects the summary-worthy content of Folktales than the baseline LSA approach. Especially as the optimal performance of the LSA summarizer occurred when it was used to generate summaries to 50% of the length of the input, rather than 5% or 10% as in this study. However, both studies presented in this chapter show that in practice participants always preferred the baseline LSA summaries to the extractive summaries. This result fits with the findings of my study on automatic coreference resolution metrics in 3.1. It is not always the case that automatic metrics can accurately predict the results obtained via human-based studies.

7.3 SUMMARY

In this chapter I have described my approach to the Transformation and Generation stages of the ITG model of summarization put forward by Spärck Jones (1999). The identification of summary-worthy content is informed by the logistic regression models discussed in the previous chapter, which demonstrated the value of discourse-structural and semantic features. Summaries are generated to a desired length, which results in a variable level of abstraction. I have presented simple methods for sentence generation, which ensure summaries are produced with reasonable linguistic quality. Finally, I

have reported the results of an evaluation comparing summaries generated according to four different methods at two different lengths.

The comparative evaluation of these four summary types has shown that the abstractive approach to summarization proposed in this thesis, with manual annotations, best conveys the salient points of the folktales examined. The fully automatic abstractive approach to summarization was ranked less favourably, although it was preferential to the baseline LSA approach in the results of the MTurk study. In the first web study, the LSA summaries were preferred to the Automatic summaries more often than not. The preference of the Manual summaries shows the overall value of my approach, while the lower preference for the Automatic summaries gives a more realistic picture of what can be achieved with currently available tools.

This chapter has provided answers to my final research questions. It has demonstrated a method by which discourse-structural information can be used to produce abstractive summaries, and shown that these are more informative than extractive summaries. A comparison of summaries generated on the basis of both manual and automatic annotations has shown the value of my approach, but also highlighted that improvements are still needed in the required, but imperfect, preprocessing steps.

8

CONCLUSIONS

Summarization is a vast topic. Aside from text, many other forms of media can also be summarized in a variety of ways; ways that consider the audience, role, purpose and genre of a text. This thesis has focused on only a small part of the problem, can structural information about a given genre of text be used to better inform its summarization? I have described a generalised approach as to how this can be achieved, and applied this approach to the domain of Russian Folktales.

In this final chapter, I conclude with a summary of the major contributions and findings of this thesis, and suggest possible directions of future work.

8.1 SUMMARY OF CONTRIBUTIONS

The primary contribution of this thesis has been to present a framework for abstractive summarization that makes use of structural information about a given domain of texts. It is my belief that such information can be used to go some way towards modelling the cognitive representations that psycholinguistics tells us humans form while reading. I have provided a overview of the many approaches to summarization, which can be more or less broken down into two categories: extractive and abstractive approaches. Although, there is clearly some overlap between the methods, and disagreement as to where the exact boundary between these two categories lies. The most evident disagreement covers the techniques of sentence compression and sentence fusion. While my approach to summary generation involves sentence compression techniques, I do not believe this alone qualifies a method as abstractive. The approach I describe moves away from the surface text of the input in multiple ways. Key to this is the abstraction of ideas: recognising that a passage of text represents a single coherent concept that can be expressed concisely, and in a way that is different from that of the original text. The text of the input is also used, but adapted to be more suitable for the purposes of a summary. Sentence compression occurs to remove superfluous content, and pronouns are substituted for their referents. This substitution improves the readability and understanding of summaries, without which it would be necessary to include more of the original text to give context to the pronominal references

Throughout this thesis I have used Russian Folktales as an example domain for both illustration purposes and to evaluate my work,

and several assumptions have been made in this context. These assumptions do not hinder the generalisation of the findings of this thesis. They were formalised only as they were found to be applicable to the corpus used, and reduced the complexity of reasoning about narrative structure. Specifically, this work does not consider stories that contain sub-plots, where there are multiple heroes/villains, or where a single character is both heroic and villainous. Nevertheless, the solutions presented to my research questions have remained general, because the approach does not in any way rely on the nature of the summarized text. Using the detailed analysis of the structure of Russian Folktales performed by Vladimir Propp, I have presented a method by which discourse structural information about a text can be determined and applied to abstractive summarization.

In answer to my first research question I have detailed a narrative structure reasoning system, showing how structural information can be detected automatically given sufficient knowledge about a domain of texts. I have subsequently shown, in answer to the second, that such structural information can be used to predict the summary-worthy content of a story. I have also compared regression models built using the detected structural information to the predictive capabilities of existing extractive algorithms. These studies indicated that, over a modestly sized dataset, the detected information can indeed be used to more accurately detect summary-worthy information than traditional approaches. To answer my third research question, the analyses of Chapter 6 have indicated that the elements which provide the greatest benefit to the detection of summary-worthy content relate specifically to the key narrative events of a given domain. In Chapter 7 I aimed to answer my final two research questions. This chapter demonstrated the application and benefit of discourse structural information to the generation of abstractive summaries. The generation process proceeds according to a desired summary length, where a shorter summary necessitates a greater amount of abstraction. An evaluation of the produced summaries show that this method can be used to produce abstractive summaries that better capture the salient points of a text than existing methods. This is a significant finding as, to the best of my knowledge, no other summarization research has explicitly addressed the detection of structural information and also used it for text abstraction.

In addition to the evaluation of the end product of this research, I have aimed to evaluate each step of this work in isolation where possible. I had initially intended to rely more heavily on existing tools, namely automatic coreference resolution systems and the most commonly used evaluation metric, ROUGE. However, an early evaluation of these tools motivated my comparison of both manual and automatic annotations throughout, as well a human based final evaluation. The choice to manually annotate data necessarily limited the

scope of the evaluations that could be performed, but it allowed for alternate, valuable comparisons. The final evaluation of summaries has allowed me to demonstrate the clear benefit and upper bound of my overall approach, but also shown what can be realistically achieved with automatic annotations. It is interesting to note that the results of Chapter 6 suggest a better performance of my automatic approach against the baseline LSA approach than is reflected by the human based evaluation of Chapter 7. This however, is not surprising. This thesis has shown that automatic evaluation metrics do not necessarily align with human judgements. This is true of both literature discussing ROUGE and the evaluation of coreference resolution systems in Section 3.1.

8.2 FUTURE WORK

As previously argued, summarization is a large, multidimensional problem. It would not be possible to cover all aspects of the field within a single thesis. There are multiple directions which are outside of the scope of this work, such as the consideration of context factors like summary purpose and intended audience. Additionally, there are significant sub-problems involved in automatic summarization. This research has indicated that further improvements ideally need to occur in preprocessing steps such as coreference resolution. Natural language generation is also a major field in its own right, and only simple generation techniques are considered in this work. Further to this, the evaluation of summaries is non-trivial, and, I believe, should only be carried out by humans at present. A more robust automatic evaluation metric that could better indicate the utility of generated summaries would enable the quicker evaluation of new methods and greatly benefit the field.

There are of course limitations to the work of this thesis, and general improvements that can be made. I aim to capture salient concepts in the text by using WordNet to identify lexical chains. However, this is a man-made taxonomy. Some parts of WordNet have a very deep hierarchy while others are shallow, making it difficult to define a degree of separation at which concepts may still be considered to be related. My use of both lexical chains and discourse-structural information acted as a stand-in for world knowledge, which can often be necessary to provide additional context. There are many existing knowledge bases both manually defined (Lenat, 1995), and automatically created (Ouyang and K. McKeown, 2014) which store complex structured or unstructured information. The integration of such resources is an interesting future direction to consider.

There are also improvements that can be made to this work with specific regard to the domain of Russian Folktales. Propp describes how the concept of *trebling* applies to instances of these stories. This

may happen within individual functions (e.g. a hero asks for help from three different characters and only the third agrees), across multiple functions (e.g. a hero must fight and defeat three different dragons), or across entire moves (e.g. the tale sequentially follows the paths of three different brothers). Detecting trebling would allow for further abstraction of content in the way a human would. A potential method for this is provided by the AMR evaluation metric SMATCH. Preliminary tests indicate that while these events may be lexically quite dissimilar, the parses of these events appear to exhibit a clear degree of structural similarity. In fact, it is evident that trebling occurs in wider domains of fairy tales too, such as *Goldilocks and the three bears*, or *The three little pigs*. Further abstraction could also occur via a hierarchy with multiple levels of compression. Some of Propp's functions are logically connected such as the three donor functions, or the acts of Pursuit and Rescue. Beyond the abstraction of text into a single narrateme, multiple narratemes such as these could be compressed together. This possibility was not considered here as it represents a departure from a strict interpretation of Propp's work. Moreover, this thesis has only considered single-move Russian Folktales, although this could be extended to cover all instances of this genre. However, this should ideally occur with further rules constricting the form that valid interpretations may take.

Another future direction should consider the application of the narrative structure reasoning system, and even Propp's Morphology, to other domains. While Propp defines his Morphology specifically for the domain of Russian Folktales, this genre contains many elements common to other folktales and fairy tales. However, Propp states how the tales of Brothers Grimm present the same scheme in general, but in a less pure and stable manner. He states that "All kinds of foreign influences alter and sometimes even corrupt a tale. Complications begin as soon as we leave the boundary of the absolutely authentic tale" (Propp, 2015, p. 100). Perhaps then the Morphology could be adapted to make it more widely applicable to a range of tales instead. Propp's constraints require a folktale to follow the events of either an act of villainy, or a hero's lack of something. These two paths are clearly analogous to two of the seven basic plots described by Booker (2004). However, by modifying the rules of Propp's Morphology, or introducing new narrative units, the additional types of plot described by Booker (2004) could also be represented. Aside from Propp's Morphology, the narrative structure reasoning system could be applied to other, if less restricted, genres. For example in *The Hero with a Thousand Faces*, Campbell (2008) describes the typical structure of the journey which a hero undergoes in myths. This analysis describes a more linear structure of stories, but still contains detectable elements and restrictions. Subsequent work by Vogler (2007) has also applied Campbell's analysis to describe the common structure of film scripts.

If nothing else, an investigation along these lines could shed light on to just how constrained a formalism must be to be practically usable with the narrative structure reasoning system.

Finally, each output of the narrative structure reasoning system provides a valid interpretation of the input text. In the generation of summaries I have used a probability distribution across all interpretations in order to inform the selection of content. Selecting only a single one of these interpretations could lead to interesting and computationally creative results. Many of the outputs of the system put an interesting twist on more standard interpretations of these tales, and this could have applications in the field of story generation. This could be further expanded by reducing or removing the constraints on the characters who fulfil the roles of Hero and Villain.

8.3 FINAL REMARKS

In sixty years of research on automatic text summarization, very many papers and books have been published, and yearly workshops and conferences are held on the topic. Research on summarization has covered a multitude of topics and techniques. Nevertheless, the field is still very young; it is hard to identify a clear direction of progress.

It is evident that further research and improvements need to occur in a range of related tasks. This thesis has demonstrated that required preproceesing steps such as coreference resolution are not yet reliable enough. Formal, large-scale evaluations of summarization have occurred (DUC, TAC), but these have primarily considered only a single domain: News. The focus of this thesis has been on the structure of text, and while new articles certainly have their own structure, the very nature of this structure makes it difficult to judge the value of new approaches to summarization. The lead-in sentences of a news article provides a more than adequate summary, and so extractive summarizers which work to this end produce very high baseline summaries. Furthermore, the most widely used summarization evaluation metric is fundamentally flawed. The premise of ROUGE is that summaries can be compared on the basis of lexical similarity alone. They cannot. If a comparison between only the surface text of two documents could indicate quality, automatic summarization would have progressed further and faster since its inception. By only considering the similarity of word choices, ROUGE implicitly promotes extractive summarization. This method provides the greatest chance of lexical overlap with a reference summary. It does not encourage research into more original, theoretically grounded, and human-like approaches to summarization.

Extractive approaches have a limit, which I believe has already been achieved. My aim has been to transcend that limit and move summarization technology into a new space of possibilities.

BIBLIOGRAPHY

Allan, James, Rahul Gupta, and Vikas Khandelwal (2001). "Temporal summaries of new topics". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 10–18.

Andersen, Hans Christian (2008). *Andersen's Fairy Tales*. Project Gutenberg. P.O. Box 2782, Champaign, IL 61825-2782, USA: Urbana, Illinois: Project Gutenberg. URL: http://www.gutenberg.org/cache/epub/1597/pg1597.txt.

Artzi, Yoav, Kenton Lee, and Luke Zettlemoyer (Sept. 2015). "Broad-coverage CCG Semantic Parsing with AMR". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1699–1710.

AWS (Jan. 2011). *AWS Developer Forums: MTurk CENSUS: About how many workers were on Mechanical Turk in 2010?* Accessed March 30, 2019. URL: https://forums.aws.amazon.com/thread.jspa?threadID=58891.

Bagga, Amit and Breck Baldwin (1998). "Algorithms for scoring coreference chains". In: *The first international conference on language resources and evaluation workshop on linguistics coreference*. Vol. 1. Citeseer, pp. 563–566.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The berkeley framenet project". In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 86–90.

Balaji, J, TV Geetha, and Ranjani Parthasarathi (2016). "Abstractive summarization: A hybrid approach for the compression of semantic graphs". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 12.2, pp. 76–99.

Bamman, David and Noah A Smith (2013). "New alignment methods for discriminative book summarization". In: *arXiv preprint arXiv:1305.1319*.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (2018). "Abstract Meaning Representation (AMR) 1.2.5 Specification". In: URL: https://github.com/amrisi/amr-guidelines/blob/master/amr.md.

Banerjee, Satanjeev and Ted Pedersen (2003). "Extended gloss overlaps as a measure of semantic relatedness". In: *Ijcai*. Vol. 3, pp. 805–810.

Banko, Michele, Vibhu O Mittal, and Michael J Witbrock (2000). "Headline generation based on statistical translation". In: *Proceed-*

ings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 318–325.

Barthes, Roland (1967). "The Death of the Author". Trans. by Richard Howard. In: *Aspen* 5–6.

Bartneck, Christoph, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska (2015). "Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment". In: *PloS one* 10.4, e0121595.

Barzilay, Regina and Michael Elhadad (1999). "Using lexical chains for text summarization". In: *Advances in automatic text summarization*, pp. 111–121.

Barzilay, Regina and Kathleen McKeown (2005). "Sentence fusion for multidocument news summarization". In: *Computational Linguistics* 31.3, pp. 297–328.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (Feb. 2003). *A neural probabilistic language model. 3: 1137–1155*.

Bentley, Frank R, Nediyana Daskalova, and Brooke White (2017). "Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 1092–1099.

Bhartiya, Divyanshu and Ashudeep Singh (2014). "A Semantic Approach to Summarization". In: *arXiv preprint arXiv:1406.1203*.

Black, John B and Gordon H Bower (1980). "Story understanding as problem-solving". In: *Poetics* 9.1-3, pp. 223–250.

Bod, Rens, Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe (2012). "Objectivity and reproducibility of Proppian narrative annotations". In: *Proceedings of the Third Workshop on Computational Models of Narrative. Ed. by Mark Alan Finlayson*, pp. 17–21.

Booker, Christopher (2004). *The seven basic plots: Why we tell stories*. A&C Black.

Bosma, Wauter (2005). "Query-based summarization using rhetorical structure theory". In: *LOT Occasional Series* 4, pp. 29–44.

Bower, Gordon H and Daniel G Morrow (1990). "Mental models in narrative comprehension". In: *Science* 247.4938, pp. 44–48.

Brin, Sergey and Lawrence Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems*.

Brunn, Meru, Yllias Chali, and Christopher J Pinchak (2001). "Text summarization using lexical chains". In: *Proc. of Document Understanding Conference*.

Cai, Jie and Michael Strube (2010). "Evaluation metrics for end-to-end coreference resolution systems". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL, pp. 28–36.

Cai, Shu and Kevin Knight (2013). "Smatch: an Evaluation Metric for Semantic Feature Structures". In: *ACL (2)*, pp. 748–752.

Campbell, Joseph (2008). *The hero with a thousand faces*. Vol. 17. New World Library.

Carenini, Giuseppe and Jackie Chi Kit Cheung (2008). "Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality". In: *Proceedings of the Fifth International Natural Language Generation Conference*. Association for Computational Linguistics, pp. 33–41.

Carroll, Lewis (1979). *The Annotated Snark by Lewis Carroll*. Ed. by Martin Gardner. Penguin Books.

Ceylan, Hakan and Rada Mihalcea (2007). "Explorations in Automatic Book Summarization". In: Association for Computational Linguistics.

Cohan, Arman and Nazli Goharian (2016). "Revisiting Summarization Evaluation for Scientific Articles". In: *arXiv preprint arXiv:1604.00400*.

Conroy, John M and Hoa Trang Dang (2008). "Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality". In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 145–152.

Conroy, John M and Dianne P O'Leary (2001). "Text summarization via hidden markov models". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 406–407.

Dabholkar, Salil, Yuvraj Patadia, and Prajyoti Dsilva (2016). "Automatic Document Summarization using Sentiment Analysis". In: *Proceedings of the International Conference on Informatics and Analytics*. ACM, p. 49.

DeJong, Gerald (1982). "An overview of the FRUMP system". In: *Strategies for natural language processing* 113, pp. 149–176.

Díaz-Agudo, Belén, Pablo Gervás, and Federico Peinado (2004). "A case based reasoning approach to story plot generation". In: *Advances in Case-Based Reasoning*, pp. 142–156.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel (2004). "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation". In: *LREC*. Vol. 2, p. 1.

Dolan, William, Lucy Vanderwende, and Stephen D Richardson (1993). "Automatically deriving structured knowledge bases from on-line dictionaries". In: *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, pp. 5–14.

Dreyfuss, Emily (Nov. 2018). *A Bot Panic Hits Amazon's Mechanical Turk*. Accessed March 30, 2019. URL: https://www.wired.com/story/amazon-mechanical-turk-bot-panic/.

Droog-Hayes, Maximilian (2017). "The Effect of Poor Coreference Resolution on Document Understanding". In: *29th European Summer School in Logic, Language and Information (ESSLLI) 2017 Student Session*, pp. 209–220.

Droog-Hayes, Maximilian, Geraint A Wiggins, and Matthew Purver (2019). "Detecting Summary-Worthy Sentences: The Effect of Discourse Features". In: *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, pp. 381–384.

Edmundson, Harold P (1969). "New methods in automatic extracting". In: *Journal of the ACM (JACM)* 16.2, pp. 264–285.

Elson, David K (2012). "Modeling narrative discourse". PhD thesis. Columbia University.

Elson, David K and Kathleen R McKeown (2007). "A Platform for Symbolically Encoding Human Narratives." In: *AAAI Fall Symposium: Intelligent Narrative Technologies*, pp. 29–36.

Erkan, Günes and Dragomir R Radev (2004). "Lexrank: Graph-based lexical centrality as salience in text summarization". In: *Journal of Artificial Intelligence Research* 22, pp. 457–479.

Fang, Yimai and Simone Teufel (2014). "A summariser based on human memory limitations and lexical competition". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 732–741.

Feng, Song, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi (2013). "Connotation lexicon: A dash of sentiment beneath the surface meaning". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1774–1784.

Finlayson, Mark A (2015). "ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory". In: *Digital Scholarship in the Humanities* 32.2, pp. 284–300.

— (2016). "Inferring Propp's functions from semantically annotated text". In: *Journal of American Folklore* 129.511, pp. 55–77.

Finlayson, Mark A et al. (2015). "Supplementary materials for "ProppLearner: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory"". In:

Fisseni, Bernhard, Aadil Kurji, and Benedikt Löwe (2014). "Annotating with Propp's Morphology of the Folktale: reproducibility and trainability". In: *Literary and Linguistic Computing* 29.4, pp. 488–510.

Flanigan, Jeffrey, Chris Dyer, Noah A Smith, and Jaime G Carbonell (2016). "CMU at SemEval-2016 Task 8: Graph-based AMR Parsing with Infinite Ramp Loss". In: *SemEval@ NAACL-HLT*, pp. 1202–1206.

Flanigan, Jeffrey, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith (2014). "A discriminative graph-based parser for the abstract meaning representation". In:

Forster, Edward Morgan (2005). *Aspects of the Novel. 1927*. Penguin Books.

Ganesan, Kavita, ChengXiang Zhai, and Jiawei Han (2010). "Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions". In: *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, pp. 340–348.

Genest, Pierre-Etienne and Guy Lapalme (2012). "Fully abstractive approach to guided summarization". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 354–358.

— (2013). "Absum: a knowledge-based abstractive summarizer". In: *Génération de résumés par abstraction* 25.

Gerani, Shima, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat (2014). "Abstractive summarization of product reviews using discourse structure". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1602–1613.

Gervás, Pablo (2014). "Reviewing Propp's story generation procedure in the light of computational creativity". In: *AISB Symposium on Computational Creativity, AISB-2014*.

Goldman, Susan R, Arthur C Graesser, and Paul Van den Broek (1999). *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso*. Routledge.

Gong, Yihong and Xin Liu (2001). "Generic text summarization using relevance measure and latent semantic analysis". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 19–25.

Gonzàlez, Edgar and Marıa Fuentes Fort (2009). "A New Lexical Chain Algorithm Used for Automatic Summarization". In: *CCIA*, pp. 329–338.

Goyal, Amit, Ellen Riloff, and Hal Daumé Iii (2013). "A computational model for plot units". In: *Computational Intelligence* 29.3, pp. 466–488.

Graesser, Arthur C, Keith K Millis, and Rolf A Zwaan (1997). "Discourse comprehension". In: *Annual review of psychology* 48.1, pp. 163–189.

Graesser, Arthur C, Murray Singer, and Tom Trabasso (1994). "Constructing inferences during narrative text comprehension". In: *Psychological review* 101.3, p. 371.

Graham, Yvette et al. (2015). "Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE". In: *EMNLP*, pp. 128–137.

Grimm, The Brothers (2008). *Grimms' Fairy Tales*. Project Gutenberg. P.O. Box 2782, Champaign, IL 61825-2782, USA: Urbana, Illinois: Project Gutenberg. URL: https://www.gutenberg.org/files/2591/2591-h/2591-h.htm.

Grosz, Barbara J and Candace L Sidner (1986). "Attention, intentions, and the structure of discourse". In: *Computational linguistics* 12.3, pp. 175–204.

Guo, Weiwei and Mona Diab (2012). "Learning the latent semantics of a concept from its definition". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pp. 140–144.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). "The WEKA data mining software: an update". In: *SIGKDD Explorations* 11.1, pp. 10–18.

Harman, Donna and Paul Over (2004). "The effects of human variation in duc summarization evaluation". In: *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*. Citeseer, pp. 10–17.

Harrington, Brian and Stephen Clark (July 2007). "ASKNet: Automated Semantic Knowledge Network". In: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Hong, Kai, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova (2014). "A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization". In: *LREC*, pp. 1608–1616.

Honnibal, Matthew and Mark Johnson (Sept. 2015). "An Improved Non-monotonic Transition System for Dependency Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1373–1378. URL: https://aclweb.org/anthology/D/D15/D15-1162.

Hovy, Eduard, Chin-Yew Lin, and Liang Zhou (2005). "Evaluating duc 2005 using basic elements". In: *Proceedings of DUC*. Vol. 2005. Citeseer.

Hutto, Daniel D (n.d.). "Narrative Understanding". In:

Informatics, Max Planck Institute for (2014). *YAGO: A High-Quality Knowledge Base*. Accessed June 18, 2016. URL: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

Jaffar, Joxan and J-L Lassez (1987). "Constraint logic programming". In: *Proceedings of the 14th ACM SIGACT-SIGPLAN symposium on Principles of programming languages*. ACM, pp. 111–119.

Jing, Hongyan (2000). "Sentence reduction for automatic text summarization". In: *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, pp. 310–315.

Jing, Hongyan, Regina Barzilay, Kathleen McKeown, and Michael Elhadad (1998). "Summarization evaluation methods: Experiments and analysis". In: *AAAI symposium on intelligent summarization*, pp. 51–59.

Kazantseva, Anna and Stan Szpakowicz (2010). "Summarizing short stories". In: *Computational Linguistics* 36.1, pp. 71–109.

Kintsch, Walter and Teun A Van Dijk (1978). "Toward a model of text comprehension and production." In: *Psychological review* 85.5, p. 363.

Knight, Kevin (2015). *Abstract Meaning Representation (AMR)*. Accessed November 10, 2015. URL: http://amr.isi.edu/.

Knight, Kevin and Daniel Marcu (2000). "Statistics-based summarization-step one: Sentence compression". In: *AAAI/IAAI* 2000, pp. 703–710.

Koupaee, Mahnaz and William Yang Wang (2018). "WikiHow: A Large Scale Text Summarization Dataset". In: *arXiv preprint arXiv:1810.09305*.

Kwong, Oi Yee (2010). "Constructing an Annotated Story Corpus: Some Observations and Issues". In: *LREC*. Citeseer.

Labov, William (2013). *The language of life and death: The transformation of experience in oral narrative*. Cambridge University Press.

Lal, Partha and Stefan Ruger (2002). "Extract-based summarization with simplification". In: *Proceedings of the ACL*.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2013). "Deterministic coreference resolution based on entity-centric, precision-ranked rules". In: *Computational Linguistics* 39.4, pp. 885–916.

Lehnert, Wendy G (1981). "Plot units and narrative summarization". In: *Cognitive Science* 5.4, pp. 293–331.

Lenat, Douglas B (1995). "CYC: A large-scale investment in knowledge infrastructure". In: *Communications of the ACM* 38.11, pp. 33–38.

Leskovec, Jure, Marko Grobelnik, and Natasa Milic-Frayling (2004a). "Learning semantic graph mapping for document summarization". In: *Proceedings of ECML/PKDD-2004 Workshop on Knowledge Discovery and Ontologies*.

— (2004b). "Learning sub-structures of document semantic graphs for document summarization". In:

Lévi-Strauss, Claude (1955). "The structural study of myth". In: *The journal of American folklore* 68.270, pp. 428–444.

Li, Wei (2015). "Abstractive multi-document summarization with semantic information extraction". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1908–1913.

Li, Xiang, Thien Huu Nguyen, and Ralph Cao Kaiand Grishman (2015). "Improving event detection with abstract meaning representation". In: *ACL-IJCNLP 2015*, p. 11.

Lin, Chin-Yew (2004a). "Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?" In: *NTCIR*.

— (2004b). "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out: Proceedings of the ACL-04 workshop*. Vol. 8.

Lin, Chin-Yew and Eduard Hovy (2003). "Automatic evaluation of summaries using n-gram co-occurrence statistics". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pp. 71–78.

Ling, Xu, Jing Jiang, Xin He, Qiaozhu Mei, Chengxiang Zhai, and Bruce Schatz (2007). "Generating gene summaries from biomedical literature: A study of semi-structured summarization". In: *Information Processing & Management* 43.6, pp. 1777–1791.

Liu, Fei, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith (2015). "Toward abstractive summarization using semantic representations". In:

Liu, Hugo and Push Singh (2004). "ConceptNet - a practical commonsense reasoning tool-kit". In: *BT technology journal* 22.4, pp. 211–226.

Lu, Yue, ChengXiang Zhai, and Neel Sundaresan (2009). "Rated aspect summarization of short comments". In: *Proceedings of the 18th international conference on World wide web*. ACM, pp. 131–140.

Luhn, Hans Peter (1958). "The automatic creation of literature abstracts". In: *IBM Journal of research and development* 2.2, pp. 159–165.

Luo, Xiaoqiang (2005). "On coreference resolution performance metrics". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, pp. 25–32.

Mann, William C and Sandra A Thompson (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *ACL (System Demonstrations)*, pp. 55–60.

Marcu, Daniel (1997). "From discourse structures to text summaries". In: *Proceedings of the ACL*. Vol. 97. Citeseer, pp. 82–88.

— (1998). "Improving summarization through rhetorical parsing tuning". In: *The 6th Workshop on Very Large Corpora*, pp. 206–215.

Martins, Camilla Brandel and Lucia Helena Machado Rino (2002). "Revisiting UNLSumm: Improvement through a case study". In: *the Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing*. Vol. 1. Citeseer, pp. 71–79.

McKeown, Kathleen and Dragomir R Radev (1995). "Generating summaries of multiple news articles". In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 74–82.

Mihalcea, Rada and Paul Tarau (2004). "TextRank: Bringing order into texts". In: Association for Computational Linguistics.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *HLT-NAACL*, pp. 746–751.

Miller, George (1998). *WordNet: An electronic lexical database*. MIT press.

Moawad, Ibrahim F and Mostafa Aref (2012). "Semantic graph reduction approach for abstractive Text Summarization". In: *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*. IEEE, pp. 132–138.

Morris, Andrew H, George M Kasper, and Dennis A Adams (1992). "The effects and limitations of automated text condensing on reading comprehension performance". In: *Information Systems Research* 3.1, pp. 17–35.

Morrow, Daniel G (1985). "Prominent characters and events organize narrative understanding". In: *Journal of Memory and Language* 24.3, pp. 304–319.

Nakao, Yoshio (2000). "An algorithm for one-page summarization of a long text based on thematic hierarchy detection". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 302–309.

Nalisnick, Eric T and Henry S Baird (2013). "Character-to-character sentiment analysis in shakespeare's plays". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 479–483.

Nallapati, Ramesh, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. (2016). "Abstractive text summarization using sequence-to-sequence rnns and beyond". In: *arXiv preprint arXiv:1602.06023*.

Narayan, Shashi, Shay B Cohen, and Mirella Lapata (2018). "Ranking sentences for extractive summarization with reinforcement learning". In: *arXiv preprint arXiv:1802.08636*.

Nenkova, Ani and Rebecca J Passonneau (2004). "Evaluating Content Selection in Summarization: The Pyramid Method". In: *HLT-NAACL*. Vol. 4, pp. 145–152.

Nenkova, Ani, Advaith Siddharthan, and Kathleen McKeown (Oct. 2005). "Automatically Learning Cognitive Status for Multi-Document Summarization of Newswire". In: *Proceedings of Hu-*

man Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 241–248. URL: https://www.aclweb.org/anthology/H05-1031.

Neto, Joel Larocca, Alexandre D Santos, Celso AA Kaestner, Neto Alexandre, D Santos, et al. (2000). "Document clustering and text summarization". In:

Ng, Jun-Ping and Viktoria Abrecht (2015). "Better Summarization Evaluation with Word Embeddings for ROUGE". In: *arXiv preprint arXiv:1508.06034*.

O'Donnell, Mick (1997). "Variable-length on-line document generation". In: *Proceedings of the 6th European Workshop on Natural Language Generation*, pp. 82–91.

Ono, Kenji, Kazuo Sumita, and Seiji Miike (1994). "Abstract generation based on rhetorical structure extraction". In: *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 344–348.

Ouyang, Jessica and Kathleen McKeown (2014). "Towards Automatic Detection of Narrative Structure". In: *LREC*, pp. 4624–4631.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The proposition bank: An annotated corpus of semantic roles". In: *Computational linguistics* 31.1, pp. 71–106.

Peng, H., K. Chang, and D. Roth (July 2015). "A Joint Framework for Coreference Resolution and Mention Head Detection". In: *CoNLL*. ACL, p. 10. URL: http://cogcomp.cs.illinois.edu/papers/MentionDetection.pdf.

Pöttker, Horst (2003). "News and its communicative quality: The inverted pyramid-when and why did it appear?" In: *Journalism Studies* 4.4, pp. 501–511.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (2011). "Conll-2011 shared task: Modeling unrestricted coreference in ontonotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 1–27.

Propp, Vladimir (2015). "Morphology of the Folk Tale. 1968". In: First Edition Translated by Laurence Scott.

Pust, Michael, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May (2015). "Parsing english into abstract meaning representation using syntax-based machine translation". In: *Training* 10, pp. 218–021.

Qazvinian, Vahed, Dragomir R Radev, Saif M Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon (2013). "Generating extractive summaries of scientific paradigms". In: *Journal of Artificial Intelligence Research* 46, pp. 165–201.

Radev, Dragomir R, Eduard Hovy, and Kathleen McKeown (2002). "Introduction to the special issue on summarization". In: *Computational linguistics* 28.4, pp. 399–408.

Radev, Dragomir R, Hongyan Jing, Małgorzata Styś, and Daniel Tam (2004). "Centroid-based summarization of multiple documents". In: *Information Processing & Management* 40.6, pp. 919–938.

Rankel, Peter, John M Conroy, and Judith Schlesinger (2012). "Better Metrics to Automatically Predict the Quality of a Text Summary". In:

Rankel, Peter, John M Conroy, Eric V Slud, and Dianne P O'Leary (2011). "Ranking human and machine summarization systems". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 467–473.

Rayner, Keith, Sarah J White, Rebecca L Johnson, and Simon P Liversedge (2006). "Raeding wrods with jubmled lettres there is a cost". In: *Psychological science* 17.3, pp. 192–193.

Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds (2016). "The emotional arcs of stories are dominated by six basic shapes". In: *EPJ Data Science* 5.1, p. 31.

Rumelhart, David E (1975). "Notes on a schema for stories". In: *Representation and understanding*. Elsevier, pp. 211–236.

Rush, Alexander M, Sumit Chopra, and Jason Weston (2015). "A neural attention model for abstractive sentence summarization". In: *arXiv preprint arXiv:1509.00685*.

Rushdie, Salman (1991). *Haroun and the Sea of Stories. 1990*. Granta Books.

Sachan, Mrinmaya, Eduard Hovy, and Eric P Xing (2015). "An active learning approach to coreference resolution". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1312–1318.

Saggion, Horacio (2008). "A robust and adaptable summarization tool". In: *Traitement Automatique des Langues* 49.2.

Schank, Roger C and Robert P Abelson (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Lawrence Erlbaum.

Schluter, Natalie (2017). "The limits of automatic summarisation according to ROUGE". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 41–45.

Schuler, Karin Kipper (2005). "VerbNet: A broad-coverage, comprehensive verb lexicon". In:

Silber, H Gregory and Kathleen F McCoy (2000a). "An efficient text summarizer using lexical chains". In: *Proceedings of the first inter-*

national conference on Natural language generation-Volume 14. Association for Computational Linguistics, pp. 268–271.

Silber, H Gregory and Kathleen F McCoy (2000b). "Efficient text summarization using lexical chains". In: *Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, pp. 252–255.

Sornlertlamvanich, Virach, Tanapong Potipiti, and Thatsanee Charoenporn (2001). "UNL document summarization". In: *Proceedings of the First International Workshop on Multimedia Annotation*.

Spärck Jones, Karen (1999). "Automatic summarizing: factors and directions". In: *Advances in automatic text summarization*, pp. 1–12.

— (2007). "Automatic summarising: The state of the art". In: *Information Processing & Management* 43.6, pp. 1449–1481.

Steinberger, Josef and Karel Jezek (2004). "Using latent semantic analysis in text summarization and summary evaluation". In: *Proc. ISIM'04*, pp. 93–100.

Steinberger, Josef, Mijail A Kabadjov, Massimo Poesio, and Olivia Sanchez-Graillet (2005). "Improving LSA-based summarization with anaphora resolution". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1–8.

Steinberger, Josef, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek (2007). "Two uses of anaphora resolution in summarization". In: *Information Processing & Management* 43.6, pp. 1663–1680.

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie, and Ellen Riloff (2009). "Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pp. 656–664.

Surdeanu, Mihai, Thomas Hicks, and Marco A. Valenzuela-Escárcega (2015). "Two Practical Rhetorical Structure Theory Parsers". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT): Software Demonstrations*.

Tan, Jiwei, Xiaojun Wan, and Jianguo Xiao (2017). "Abstractive document summarization with a graph-based attentional neural model". In: *Proceedings of the 55th Annual Meeting of the ACL*. Vol. 1, pp. 1171–1181.

Tapiero, Isabelle, Paul Van den Broek, and Marie-Pilar Quintana (2002). "The mental representation of narrative texts as networks: The role of necessity and sufficiency in the detection of different types of causal relations". In: *Discourse Processes* 34.3, pp. 237–258.

Teufel, Simone and Marc Moens (2002). "Summarizing scientific articles: experiments with relevance and rhetorical status". In: *Computational linguistics* 28.4, pp. 409–445.

Thompson, Stith (1989). *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*. Vol. 4. Indiana University Press.

Thorndyke, Perry W (1977). "Cognitive structures in comprehension and memory of narrative discourse". In: *Cognitive psychology* 9.1, pp. 77–110.

Tran, Giang Binh, Mohammad Alrifai, and Eelco Herder (2015). "Timeline Summarization from Relevant Headlines". In: *ECIR*. Vol. 9022, pp. 245–256.

Tseng, Yuen-Hsien, Chi-Jen Lin, and Yu-I Lin (2007). "Text mining techniques for patent analysis". In: *Information Processing & Management* 43.5, pp. 1216–1247.

Turner, Scott R (1993). "Minstrel: a computer model of creativity and storytelling". In:

Tzeng, Yuhtsuen, Paul Van Den Broek, Panayiota Kendeou, and Chengyuan Lee (2005). "The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension". In: *Behavior research methods* 37.2, pp. 277–286.

Uchida, Hiroshi, Meiying Zhu, and T Della Senta (1999). "Universal Networking Language: A gift for a millennium". In: *The United Nations University, Tokyo, Japan*.

Valls-Vargas, Josep, Santiago Ontanón, and Jichen Zhu (2013). "Toward character role assignment for natural language stories". In: *Proceedings of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, pp. 101–104.

Van Dijk, Teun Adrianus, Walter Kintsch, and Teun Adrianus Van Dijk (1983). "Strategies of discourse comprehension". In:

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A model-theoretic coreference scoring scheme". In: *Proceedings of the 6th conference on Message understanding*. ACL, pp. 45–52.

Vogler, Christopher (2007). *The Writer's journey: Mythic Structure for Writers*. Michael Wiese Productions Studio City, CA.

Webber, Bonnie and Aravind Joshi (2012). "Discourse structure and computation: past, present and future". In: *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, pp. 42–54.

White, Michael, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff (2001). "Multidocument summarization via information extraction". In: *Proceedings of the first international conference on Human language technology research*.

Witbrock, Michael J and Vibhu O Mittal (1999). "Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 315–316.

Wolf, Thomas (July 2017). *State-of-the-art neural coreference resolution for chatbots*. URL: https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30.

Zajic, David, Bonnie Dorr, and Richard Schwartz (2004). "BBN/UMD at DUC-2004: Topiary". In: *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pp. 112–119.

Zarrieß, Sina, Sebastian Loth, and David Schlangen (2015). "Reading Times Predict the Quality of Generated Text Above and Beyond Human Ratings". In: *ENLG 2015*, p. 38.

A

COREFERENCE STUDY MATERIALS

This appendix includes the questions asked in the study on the effect of automatic coreference resolution, and the associated answers under each of the four coreference conditions. The answers to the questions under the manual coreference resolution setting represent the ground truth answers. These stories come from a collection of Aesop's Fables, the text for which can easily be found online.

A.1 JUPITER AND THE MONKEY

1. *What did Jupiter promise?*

   MANUAL A royal reward.

   NONE A royal reward.

   CORENLP A royal reward.

   ILLINOIS A royal reward.

2. *Who did the monkey present?*

   MANUAL A young monkey (her son).

   NONE A young monkey.

   CORENLP The 'young monkey' presented 'young monkey'. Regardless, the question is not coherent in this case.

   ILLINOIS 'monkey' presented 'monkey'. These are both different singleton entities.

3. *Did the crowd laugh at the monkey?*

   MANUAL Yes.

   NONE In the surface text 'she' was laughed at, so the question itself cannot be understood.

   CORENLP They laughed at the 'young monkey', a different character.

   ILLINOIS They laughed at 'she' (surface text) which refers to 'a mother'.

A.2 THE EAGLE AND THE JACKDAW

1. *What did the eagle carry off?*

   MANUAL A lamb.

NONE 'he' - the surface text doesn't say expressly say 'lamb' at this point.

CORENLP Jackdaw.

ILLINOIS 'his' - the headword of the entity referred to.

2. *What did the jackdaw become tangled in?*

MANUAL The ram's fleece.

NONE 'he' - but the surface text doesn't name the jackdaw at a point at which the question would make sense.

CORENLP The ram's fleece.

ILLINOIS The ram's fleece.

3. *What happened to the jackdaw's wings?*

MANUAL The shepherd clipped them.

NONE 'he' clipped them, but the act of clipping is enough for a correct answer.

CORENLP The jackdaw clipped the jackdaw's wings. However, saying the wings were clipped does answer the question correctly without needing to observe the mistake in coreference resolution.

ILLINOIS The shepherd clipped the jackdaw's wings.

4. *Who did the Shepherd give the jackdaw to?*

MANUAL His children.

NONE 'he' gave 'him' to 'his children' - no mention of either jackdaw or shepherd in the surface text so this question doesn't make sense.

CORENLP The question doesn't make sense. The jackdaw gave the jackdaw to the jackdaw's child - child is not detected as a mention.

ILLINOIS The shepherd gave the shepherd to his child. This is incorrect, 'his children' is detected as a separate entity which does not refer to this mention.

A.3 THE FARMER AND HIS SONS

1. *What did the father wish?*

MANUAL For his sons to attend his farm as he did (give the same attention as he did).

NONE For 'his' sons to give 'his' farm the same attention as 'he' had given 'it'. Without coreference information, none of these mentions are coreferent, and so the answer is wrong.

CORENLP For his sons to give his farm the same attention as he had given to 'it', where it incorrectly refers to 'the same attention' rather than the farm.

ILLINOIS For his sons to give the farm the same attention as he had given 'it'. But 'it' does not link to the farm entity.

2. *Who did the father call to his bedside?*

MANUAL His sons.

NONE 'he' called 'they' to the bedside. No mention of father so question doesn't make sense.

CORENLP 'he' called 'his sons' to the father's bedside. Incorrect as the father didn't call anyone - the mention of 'he' doing the calling was not detected.

ILLINOIS 'they' (singleton mention).

3. *What did the father say was buried in the vineyards?*

MANUAL A (great) treasure.

NONE Father didn't say, 'he' said, so the question cannot be answered.

CORENLP 'he' said that a great treasure was buried. But the 'he' mention was not detected and does not link to 'father' - so the question does not make sense.

ILLINOIS A great treasure.

4. *Did the sons find treasure?*

MANUAL No.

NONE 'they' didn't find treasure. No mention of sons, so this is not the correct answer.

CORENLP No.

ILLINOIS 'they' (singleton mention) found no treasure. No referent to the sons, so the question does not make sense.

5. *What was the sons' reward?*

MANUAL An extraordinary, superabundant crop.

NONE 'they' received the reward, not the sons, so the question does not make sense.

CORENLP An extraordinary, superabundant crop.

ILLINOIS An extraordinary, superabundant crop.

A.4 THE FATHER AND HIS TWO DAUGHTERS

1. *Who did the first daughter/one marry?*

MANUAL A gardener.

NONE 'one' married a gardener.

CORENLP 'one' married a gardener.

ILLINOIS 'one' married a gardener.

2. *Who did the other daughter marry?*

MANUAL A tile-maker.

NONE 'other' married a tile maker.

CORENLP 'other' married 'person-make-tile', no indication of daughter though so this is incorrect.

ILLINOIS 'other' married person-make-tiles.

3. *What did the first daughter/one wish?*

MANUAL Heavy rain (to water plants well).

NONE 'i' wished - the question cant be answered.

CORENLP Both daughters refer to the entity 'the daughter who had married the gardener' According to CoreNLP the same entity wished for both rain and sun.

ILLINOIS The entity referred to as 'me' wished for rain. It is not clear that this refers to the first daughter, so the question does not make sense with this coreference information.

4. *What did the second daughter/other wish?*

MANUAL For dry weather to continue (to dry the bricks).

NONE 'i' wished - the question cant be answered.

CORENLP Both daughters refer to the entity 'the daughter who had married the gardener' According to CoreNLP the same entity wished for both rain and sun.

ILLINOIS 'the daughter who had married the gardener' wished for dry weather, the system is mixing references to both daughters.

5. *Did the man join the wishes of either daughter?*

MANUAL No.

NONE The question requires coreference information to make sense.

CORENLP The man didnt join either wish.

ILLINOIS The question does not make sense with the given coreference information.

A.5 THE FOX AND THE CROW

1. *What was the crow carrying in its beak?*

   MANUAL Cheese (piece of cheese).

   NONE Cheese.

   CORENLP Cheese.

   ILLINOIS A piece of cheese.

2. *Who greeted the crow?*

   MANUAL The fox.

   NONE 'he'

   CORENLP 'Master Reynard'.

   ILLINOIS 'he' greeted 'mistress crow'.

3. *What caused the crow to drop the item it was carrying in its beak?*

   MANUAL The crow opened her mouth (to caw her best).

   NONE The question requires coreference information to make sense.

   CORENLP 'her' opening 'her' mouth, is the closest text, but there is no reference to the crow.

   ILLINOIS 'her' opening 'her' mouth, is the closest text, but there is no reference to the crow.

4. *Did the fox grab the cheese?*

   MANUAL Yes.

   NONE 'master fox' snapped up the cheese.

   CORENLP 'a fox' snapped up the cheese.

   ILLINOIS Yes.

5. *What was Master Reynard's advice?*

   MANUAL Don't trust flatterers

   NONE The question requires coreference information to make sense.

   CORENLP Master Reynard does not refer to an entity who gives advice with this coreference information.

   ILLINOIS Master Reynard does not refer to an entity who gives advice with this coreference information.

A.6 THE FOX AND THE GOAT

1. *What did the fox fall into?*

   MANUAL A (deep) well.

   NONE A deep well.

   CORENLP A deep well.

   ILLINOIS A deep well.

2. *Who asked the fox what he was doing?*

   MANUAL A goat.

   NONE The question requires coreference information to make sense.

   CORENLP A goat. This mention is not detected, but the answer is valid with the leniency rule.

   ILLINOIS A goat.

3. *Why did the fox claim to have jumped down the well?*

   MANUAL To have water nearby/because of a great drought.

   NONE The question requires coreference information to make sense.

   CORENLP To have water nearby.

   ILLINOIS The fox makes no claims according to this coreference information.

4. *Who was convinced to jump into the well?*

   MANUAL The goat.

   NONE The goat.

   CORENLP The goat. This mention is not detected, but the answer is valid with the leniency rule.

   ILLINOIS The goat.

5. *What did the fox jump on in order to get out of the well?*

   MANUAL The goat's back.

   NONE 'her' back, no mention of the goat so this is incorrect.

   CORENLP Using the headword of the referring entity, 'her' back. This is not enough for a correct answer.

   ILLINOIS The relevant 'fox' mention is not detected and so the question does not make sense.

A.7 THE LAMB AND THE WOLF

1. *Who was pursuing the lamb?*

   MANUAL The wolf.

   NONE The wolf.

   CORENLP The wolf.

   ILLINOIS The wolf.

2. *Who would slay the lamb if caught?*

   MANUAL The priest.

   NONE The question requires coreference information to make sense.

   CORENLP The wolf.

   ILLINOIS The question does not make sense with the given coreference information.

3. *What did the lamb think was better than being eaten?*

   MANUAL Being sacrificed at the temple.

   NONE for 'I' to be sacrificed at the temple.

   CORENLP The wolf being sacrificed in the temple, not the lamb.

   ILLINOIS For 'I' to be sacrificed in the temple, but this is not a reference to the lamb.

A.8 THE MANSLAYER

1. *Who was chasing the murderer?*

   MANUAL A relative of the man he murdered.

   NONE A person related to the man 'he' murdered.

   CORENLP The murderer is the same entity as 'the man who he murdered'. According to this coreference information the murderer was chasing himself.

   ILLINOIS A relative of the man he murdered.

2. *Why did the murderer climb a tree?*

   MANUAL He was (fearfully) afraid.

   NONE The question requires coreference information to make sense.

   CORENLP 'the man whom he murdered' climbed a tree , not the murderer. The question does not make sense with the given coreference information.

   ILLINOIS The question does not make sense with the given coreference information.

3. *What did the murderer find in the tree?*

   MANUAL A serpent.

   NONE The question requires coreference information to make sense.

   CORENLP The question does not make sense with the given coreference information.

   ILLINOIS 'he' found a serpent, but this does not refer to the murderer.

4. *Did a crocodile eat the murderer?*

   MANUAL Yes.

   NONE The question requires coreference information to make sense.

   CORENLP The question does not make sense with the given coreference information.

   ILLINOIS Yes.

5. *What did the earth, the air and the water refuse the murderer?*

   MANUAL Shelter.

   NONE Shelter.

   CORENLP They refused shelter to 'person', which does not refer to the murderer.

   ILLINOIS Shelter.

A.9 THE PROPHET

1. *What was the wizard doing in the marketplace?*

   MANUAL Sitting (telling fortunes of passers-by).

   NONE Sitting (telling fortunes of passers-by).

   CORENLP The relevant mention is not detected.

   ILLINOIS Sitting (telling fortunes of passers-by).

2. *What happened to the wizard's house?*

   MANUAL The door was broken open (and his goods were being stolen).

   NONE The question requires coreference information to make sense.

   CORENLP The question does not make sense with the given coreference information.

   ILLINOIS The question does not make sense with the given coreference information.

3. *Who should have foreseen their own fortunes?*

   MANUAL The wizard.

   NONE 'you'.

   CORENLP 'you'.

   ILLINOIS 'you'.

A.10 THE TWO MEN WHO WERE ENEMIES

1. *Were the two deadly enemies on the same ship?*

   MANUAL Yes.

   NONE Yes.

   CORENLP Yes.

   ILLINOIS Yes.

2. *Why did they sit at opposite ends of the ship?*

   MANUAL They were determined to keep far apart (because they were enemies).

   NONE The question requires coreference information to make sense.

   CORENLP They were determined to keep far apart (valid only by the leniency rule).

   ILLINOIS They were determined to keep far apart.

3. *What put the ship in danger of sinking?*

   MANUAL A violent storm.

   NONE A violent storm.

   CORENLP A violent storm.

   ILLINOIS A violent storm.

4. *Who asked the pilot which end of the ship would go down first?*

   MANUAL The one in the stern.

   NONE The one in the stern.

   CORENLP The one in the stern.

   ILLINOIS The one in the stern.

B

ADDITIONAL ROUGE RESULTS

This appendix includes further results from the preliminary study of the ROUGE evaluation metric presented in Chapter 3. Specifically results are given here for the variants ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-SU-4 across all three of the corpora examined.

B.1 ROUGE-2

Table 25: ROUGE-2 scores for a given number of reference summaries averaged across the set of 10 stories.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	14.8	7.7	16.3	9.5
2	14.5	7.4	16.0	9.2
3	14.4	7.3	15.8	9.0
4	14.3	7.3	15.7	9.0
5	14.2	7.3	15.7	8.9
6	14.2	7.2	15.7	8.9

Table 26: ROUGE-2 scores for a given number of reference summaries averaged across the set of 10 news articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	13.6	13.5	14.8	15.6
2	11.7	11.8	12.8	13.6
3	11.2	11.2	12.2	13.0
4	10.9	10.9	11.9	12.7
5	10.8	10.8	11.7	12.5
6	10.7	10.7	11.7	12.4

Table 27: ROUGE-2 scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	12.8	11.3	14.0	13.5
2	11.5	9.5	12.7	11.5

B.2 ROUGE-3

Table 28: ROUGE-3 scores for a given number of reference summaries averaged across the set of 10 stories.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	4.7	2.2	5.4	2.7
2	4.7	2.2	5.4	2.7
3	4.7	2.2	5.4	2.7
4	4.7	2.2	5.4	2.7
5	4.7	2.2	5.4	2.7
6	4.7	2.2	5.4	2.7

Table 29: ROUGE-3 scores for a given number of reference summaries averaged across the set of 10 news articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	6.9	6.2	7.5	7.4
2	5.7	5.1	6.2	6.1
3	5.4	4.8	5.8	5.8
4	5.2	4.7	5.7	5.6
5	5.1	4.6	5.5	5.5
6	5.0	4.5	5.5	5.4

Table 30: ROUGE-3 scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	6.6	5.7	6.9	6.2
2	5.7	4.5	6.0	5.1

B.3 ROUGE-4

Table 31: ROUGE-4 scores for a given number of reference summaries averaged across the set of 10 stories.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	2.1	0.9	2.4	1.0
2	2.2	1.0	2.5	1.1
3	2.2	1.0	2.5	1.2
4	2.2	1.1	2.5	1.2
5	2.3	1.1	2.6	1.2
6	2.3	1.1	2.6	1.2

Table 32: ROUGE-4 scores for a given number of reference summaries averaged across the set of 10 news articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	4.2	3.4	4.5	3.9
2	3.3	2.7	3.6	3.1
3	3.0	2.5	3.3	2.9
4	2.9	2.4	3.2	2.8
5	2.9	2.4	3.1	2.7
6	2.8	2.3	3.1	2.7

Table 33: ROUGE-4 scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	4.2	3.5	4.3	3.6
2	3.5	2.7	3.7	2.8

B.4 ROUGE-L

Table 34: ROUGE-L scores for a given number of reference summaries averaged across the set of 10 stories.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	31.9	27.3	33.5	30.8
2	30.7	26.3	32.3	29.8
3	30.4	26.0	31.9	29.4
4	30.3	25.8	31.8	29.2
5	30.2	25.7	31.7	29.1
6	30.2	25.6	31.7	29.1

Table 35: ROUGE-L scores for a given number of reference summaries averaged across the set of 10 news articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	28.6	28.1	30.7	31.6
2	26.6	26.0	28.5	29.3
3	25.9	25.3	27.8	28.6
4	25.6	25.0	27.4	28.1
5	25.4	24.7	27.2	27.8
6	25.3	24.6	27.1	27.7

Table 36: ROUGE-L scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	25.8	23.7	27.3	27.4
2	23.9	20.7	25.4	24.1

B.5 ROUGE-SU-4

Table 37: ROUGE-SU-4 scores for a given number of reference summaries averaged across the set of 10 stories.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	24.0	14.4	27.2	18.2
2	21.8	12.8	24.7	16.4
3	20.8	12.2	23.5	15.6
4	20.3	11.8	22.8	15.1
5	19.9	11.6	22.4	14.8
6	19.7	11.6	22.2	14.7

Table 38: ROUGE-SU-4 scores for a given number of reference summaries averaged across the set of 10 news articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	15.5	13.0	17.6	15.9
2	12.4	10.0	14.4	12.4
3	11.4	9.0	12.9	11.1
4	10.8	8.4	12.3	10.5
5	10.4	8.1	11.8	10.0
6	10.2	7.9	11.6	9.8

Table 39: ROUGE-SU-4 scores for a given number of reference summaries averaged across the set of 10 scientific articles.

REFERENCES	ORIGINAL	STOP WORDS	STEMMING	BOTH
1	18.7	13.0	21.9	17.7
2	14.9	8.6	17.7	12.6

C

STORY MATERIALS

This appendix includes the full text of the 15 Russian Folktales used for studies within this thesis. The first of these, Nikita the Tanner, is marked with the gold standard character function annotations according to Propp (2015) and Finlayson et al. (2015).

C.1 NIKITA THE TANNER

α [*A dragon appeared near Kiev; he took heavy tribute from the people - a lovely maiden from every house, whom he then devoured. Finally, it was the fate of the tsar's daughter to go to the dragon.*] A [*He seized her and dragged her to his lair but did not devour her, because she was a beauty.*] Instead, he took her to wife. Whenever he went out, he boarded up his house to prevent the princess from escaping. The princess had a little dog that had followed her to the dragon's lair. The princess often wrote to her father and mother. She would attach her letter to the dog's neck, and the dog would take it to them and even bring back the answer. One day the tsar and tsarina wrote to their daughter, asking her to find out who in this world was stronger than the dragon. The princess became kindlier toward the dragon and began to question him. For a long time he did not answer, but one day he said inadvertently that a tanner in the city of Kiev was stronger than he. When the princess heard this, she wrote her father to find Nikita the Tanner in Kiev and to send him to deliver her from captivity. Upon receiving this letter, the tsar went in person to beg Nikita the Tanner to free his land from the wicked dragon and rescue the princess. At that moment Nikita was currying hides and held twelve hides in his hands; when he saw that the tsar in person had come to see him, he began to tremble with fear, his hands shook, and he tore the twelve hides. B [*But no matter how much the tsar and tsarina entreated him, he refused to go forth against the dragon.*] So they gathered together five thousand little children and sent them to implore him, hoping that their tears would move him to pity. The little children came to Nikita and begged him with tears to go fight the dragon. C [*Nikita himself began to shed tears when he saw theirs.*] ↑ [*He took twelve thousand pounds of hemp, tarred it with pitch, and wound it around himself so that the dragon could not devour him, then went forth to give him battle.*] Nikita came to the dragon's lair but the dragon locked himself in. "Better come out into the open field," said Nikita, "or I will destroy your lair together with you!" And he began to break down the door. The dragon, seeing that he could not avoid trouble, went out to fight in the open field. H, I [*Nikita fought*

him for a long time or a short time; in any event, he defeated him.] Then the dragon began to implore Nikita: "Do not put me to death, Nikita the Tanner; no one in the world is stronger than you and I. Let us divide all the earth, all the world, into equal parts; you shall live in one half, I in the other." "Very well," said Nikita, "let us draw a boundary line." He made a plow that weighed twelve thousand pounds, harnessed the dragon to it, and the dragon began to plow a boundary from Kiev; he plowed a furrow from Kiev to the Caspian Sea. "Now," said the dragon, "we have divided the whole earth." "We have divided the earth," said Nikita, "now let us divide the sea; else you will say that your water has been taken." K [*The dragon crawled to the middle of the sea; Nikita killed him and drowned him in the sea.*] That furrow can be seen to this very day; it is fourteen feet high. Around it the fields are plowed, but the furrow is intact; and those who do not know what it is, call it the rampart. ↓,*W* [*Nikita, having done his heroic deed, would not accept any reward, but returned to currying hides.*]

C.2 THE MAGIC SWAN GEESE

An old man lived with his old wife; they had a daughter and a little son. "Daughter, daughter," said the mother, "we are going to work; we shall bring you back a bun, sew you a dress, and buy you a kerchief. Be careful, watch over your little brother, do not leave the house." The parents went away and the daughter forgot what they had told her; she put her brother on the grass beneath the window, ran out into the street, and became absorbed in games. Some magic swan geese came, seized the little boy, and carried him off on their wings. The girl came back and found her brother gone. She gasped, and rushed to look in every corner, but could not find him. She called him, wept, and lamented that her father and mother would scold her severely; still her little brother did not answer. She ran into the open field; the swan geese flashed in the distance and vanished behind a dark forest. The swan geese had long had a bad reputation; they had done a great deal of damage and stolen many little children. The girl guessed that they had carried off her brother, and rushed after them. She ran and ran and saw a stove. "Stove, stove, tell me, whither have the geese flown?" "If you eat my cake of rye I will tell you." "Oh, in my father's house we don't even eat cakes of wheat!" The stove did not tell her. She ran farther and saw an apple tree. "Apple tree, apple tree, tell me, whither have the geese flown?" "If you eat one of my wild apples, I will tell you." "Oh, in my father's house we don't even eat sweet apples." She ran farther and saw a river of milk with shores of pudding. "River of milk, shores of pudding, whither have the geese flown ?" "If you eat of my simple pudding with milk, I will tell you." "Oh, in my father's house we don't even eat cream." She would have run in the fields and wandered in the woods for a long time, if she

had not luckily met a hedgehog. She wanted to nudge him, but was afraid that he would prick her, and she asked: "Hedgehog, hedgehog, have you not seen whither the geese have flown?" "Thither," he said, and showed her. She ran and saw a little hut that stood on chicken legs and turned round and round. In the little hut lay Baba Yaga with veined snout and clay legs, and the little brother was sitting on a bench, playing with golden apples. His sister saw him, crept near him, seized him, and carried him away. But the geese flew after her: if the robbers overtook her, where would she hide? There flowed the river of milk with shores of pudding. "Little mother river, hide me!" she begged. "If you eat my pudding." There was nothing to be done; she ate it, and the river hid her beneath the shore, and the geese flew by. She went out, said "Thank you," and ran on, carrying her brother; and the geese turned back and flew toward her. What could she do in this trouble? There was the apple tree. "Apple tree, apple tree, little mother, hide me !" she begged. "If you eat my wild apple." She ate it quickly. The apple tree covered her with branches and leaves; the geese flew by. She went out again and ran on with her brother. The geese saw her and flew after her. They came quite close, they began to strike her with their wings; at any moment they would tear her brother from her hands. Luckily there was the stove on her path. "Madam Stove, hide me!" she begged. "If you eat my cake of rye." The girl quickly stuck the cake in her mouth, went into the stove, and sat there. The geese whirred and whirred, quacked and quacked, and finally flew away without recovering their prey. And the girl ran home, and it was a good thing that she came when she did, for soon afterward her mother and father arrived.

C.3 BUKHTAN BUKHTANOVICH

In a certain kingdom in a certain land there lived one Bukhtan Bukhtanovich, who had a stove built on pillars in the middle of a field. He lay on the stove in cockroach milk up to his elbows. A fox came to him and said: "Bukhtan Bukhtanovich, would you like me to marry you to the tsar's daughter?" "What's that you're saying, little fox?" "Do you have any money?" "I have one five kopek piece." "Hand it over!" The fox took the coin, exchanged it for smaller coins - kopeks, pennies, and halfpennies. He went to the tsar, and said: "Tsar, give me a quart measure to measure Bukhtan Bukhtanovich's money." The tsar said: "Take one!" The fox took it home, stuck one kopek behind the hoop around the measure, brought it back to the tsar, and said: "Tsar, a quart measure is not big enough; give me a peck measure measure Bukhtan Bukhtanovich's money." "Take one!" The fox took it home, stuck a kopek behind the hoop of the measure, and brought it back to the tsar. "Tsar, a peck measure is not big enough; give me a bushel measure." "Take one!" The Fox took it home, stuck what

remained of his coins behind the hoop, and brought it back to the tsar. He said: "Have you measured all his money, little fox?" The fox answered: "All of it. Now, tsar, I have come for a good purpose: give your daughter in marriage to Bukhtan Bukhtanovich." "Very well; show me the suitor." The fox ran home. "Bukhtan Bukhtanovich, have you any clothes? Put them on." He dressed and, accompanied by the fox, went to the tsar. They walked along the market place and had to cross on a board over a muddy ditch. The fox gave Bukhtan a push and he fell into the mud. The fox ran to him. "What is the matter with you, Bukhtan Bukhtanovich?" Saying this, the fox smeared him with mud all over. "Wait here, Bukhtan Bukhtanovich, I shall run to the tsar." The fox came to the tsar and said: "Tsar, I was walking with Bukhtan Bukhtanovich on a board over a ditch - it was a wretched little board. We were not careful enough and somehow fell into the mud. Bukhtan Bukhtanovich is all dirty and unfit to come to town; have you some clothes you could lend him?" "Here, take these." The fox took the clothes and came to Bukhtan Bukhtanovich. "Here, change your clothes, Bukhtan, and let us go." They came to the tsar, and at the tsar's palace the table was already set. Bukhtan did not look at anything except himself: he had never seen such clothes in his life. The tsar winked to the fox: "Little fox, why does this Bukhtan Bukhtanovich look only at himself?" "Tsar, I think he is ashamed to be wearing such clothes; never in his life has he worn such mean garments. Tsar, give him the garment that you yourself wear on Easter Sunday." And to Bukhtan the fox whispered: "Don't look at yourself!" Bukhtan Bukhtanovich stared at a chair - it was a gilded one. The tsar again whispered to the fox: "Little fox, why does Bukhtan Bukhtanovich look only at that chair?" "Tsar, in his house such chairs stand only in the bathhouse." The tsar flung the chair out of the room. The fox whispered to Bukhtan: "Do not look at one thing; look here a bit and there a bit." They began to talk about the purpose of their visit, the match. And then they celebrated the wedding -does a wedding take long in a tsar's palace? There no beer need be brewed, no wine distilled - everything is ready. Three ships were loaded for Bukhtan Bukhtanovich and they traveled homeward. Bukhtan Bukhtanovich and his wife were on one of the ships, and the fox ran along the shore. Bukhtan saw his stove and cried: "Little fox, little fox, there is my stove." "Be quiet, Bukhtan Bukhtanovich, that stove is a disgrace." Bukhtan Bukhtanovich sailed on, and the fox ran ahead of him on the shore. He came to a hill and climbed it. On the hill stood a huge stone house, and around it was an enormous kingdom. The fox went into the house and at first saw no one; then he ran into a chamber, and there in the best bed lay Dragon, Son of the Dragon, stretching himself. Raven, Son of the Raven, was perched on the chimney, and Cat, Son of the Cat, sat on a throne. The fox said: "Why are you sitting here? The tsar is coming with fire and the tsarina with light-

ning, they will scorch and burn you." "Little fox, whither shall we go?" "Cat, Son of the Cat, go into the barrel." And the fox sealed him up in the barrel. "Raven, Son of the Raven, go into the mortar!" And the fox sealed him up in the mortar; then he wrapped Dragon, Son of the Dragon, in straw and took him out into the street. The ships arrived. The fox ordered all the beasts to be thrown into the water; the Cossacks threw them in at once. Bukhtan Bukhtanovich moved all his possessions into that house; there he lived happily and prospered, ruled and governed, and there he ended his life.

C.4 THE CRYSTAL MOUNTAIN

In a certain kingdom in a certain land there lived a king who had three sons. One day they said to him: "Father, our gracious sovereign, give us your blessing; we wish to go hunting." The father gave them his blessing and they set out in different directions. The youngest son rode and rode and lost his way; he came to a clearing, and there lay a dead horse, around which were gathered beasts of many kinds, birds, and reptiles. A falcon rose, flew up to the prince, perched on his shoulder, and said: "Prince Ivan, divide that horse among us. It has lain here for thirty years, and we have been quarreling ever since, unable to find a way of sharing it." The prince climbed down from his good steed and divided the carcass: he gave the bones to the beasts, the flesh to the birds, the skin to the reptiles, and the head to the ants. "Thank you, Prince Ivan," said the falcon. "For your kindness you shall be able to turn into a bright falcon or an ant whenever you wish." Prince Ivan struck the damp earth, turned into a bright falcon, soared up into the air, and flew to the thrice tenth kingdom. More than half of that kingdom had been swallowed into a crystal mountain. The prince flew straight into the royal palace, turned into a goodly youth, and asked the palace guards: "Will your king take me into his service?" "Why should he not take such a goodly youth?" they answered. Thus he entered the service of that king and lived in his palace for one week, then a second, then a third. The king's daughter asked her father: "Father, my sovereign, give me leave to take a ride with Prince Ivan to the crystal mountain." The king gave her leave. They mounted good steeds and set out. When they approached the crystal mountain, a golden goat jumped suddenly out from nowhere. The prince chased it; he galloped and galloped, but could not catch the goat, and when he returned the princess had vanished. What was he to do? How could he dare to appear before the king? He disguised himself as a very old man, so that he would be unrecognizable, came to the palace, and said to the king: "Your Majesty, hire me as your herdsman." "Very well," said the king, "be my herdsman. When the three-headed dragon comes to your herd, give him three cows; when the six-headed dragon comes, give him six cows;

and when the twelve-headed dragon comes, count off twelve cows." Prince Ivan drove his herd over mountains and valleys. Suddenly the three-headed dragon came flying from a lake and said: "Ah, Prince Ivan, what kind of work are you engaged in? A goodly youth like you should be vying in combat, not tending cattle. Well, let me have three cows!" "Won't that be too much?" asked the prince. "I myself eat only one duck a day, and you want three cows. But you won't get any!" The dragon flew into a rage, and instead of three cows, seized six. Prince Ivan straightway turned into a bright falcon, cut off all the three heads of the dragon, and drove the cattle home. "Well, grandfather," asked the king, "has the three-headed dragon come? Did you give him three cows?" "No, Your Majesty," replied Prince Ivan, "I did not give him any." Next day the prince drove his herd over mountains and valleys, and the six-headed dragon came from the lake and demanded six cows. "Ah, you gluttonous monster," said the prince, "I myself eat only one duck a day, and see what you demand! I won't give you any!" The dragon flew into a rage and instead of six, seized twelve cows; but the prince turned into a bright falcon, fell upon the dragon, and cut off his six heads. He drove the herd home and the king asked him: "Well, grandfather, has the six-headed dragon come? Has my herd grown much smaller?" "Come he did, but he took nothing," answered the prince. Late at night Prince Ivan turned into an ant and crawled into the crystal mountain through a little crack. Lo and behold, the princess was in the crystal mountain. "Good evening!" said Prince Ivan. "How did you get here?" "The twelve-headed dragon carried me off," said the princess. "He lives in father's lake and has a coffer in his side. In this coffer is a hare, in this hare is a duck, in this duck is an egg, and in this egg is a seed. If you slay the dragon and get that seed, it will be possible to destroy the crystal mountain and rescue me." Prince Ivan crawled out of the mountain, turned again into a herdsman, and drove his herd. Suddenly the twelve-headed dragon flew up to him and said: "Ah, Prince Ivan, you are not doing what you should; a goodly youth like you should be vying in combat, not tending a herd. Well, count off twelve cows for me!" "That will be too much for you!" said the prince. "I myself eat only one duck a day, and see what you demand!" They began to fight, and after a long struggle or a short struggle, Prince Ivan defeated the twelve-headed dragon, slashed open his trunk, and found the coffer in his right side. In the coffer he found a hare, in the hare a duck, in the duck an egg, in the egg a seed. He took the seed, set it alight, and brought it to the crystal mountain, which soon melted away. Prince Ivan led the princess to her father, who was overjoyed and said to the prince: "Be my son-in-law!" The wedding was held at once.

C.5 SHABARSHA THE LABORER

Shabarsha set off to work as a laborer, and times were bad. There was no grain at all, and the vegetables didn't grow. But one owner thought, a deep thought. How could he chase away his misery, what could he live on, where could he get some money? "Don't worry about it," said Shabarsha to him. "There'll come a day and there will be both grain and money!" And Shabarsha set off for the millpond. "At any rate I can catch some fish. I'll sell them and I'll have some money. But I've no line and no hook. Wait a minute and I'll make one." He asked the miller for a handful of hemp, sat down on the shore, and started weaving his tackle. He wove and wove, and then a boy in a black shirt and red hat jumped out of the water. "Grandfather! What are you doing here?" he asked. "I'm weaving a line." "Why?" "I intend to cleanse this pond of all you devils by pulling you out of the water." "Oh, no! Wait a little. I'll go and tell grandfather." The imp dove deep down and Shabarsha went back to his work. "Wait a minute," he thought, "I'll play a trick on you, you cursed ones. You'll bring me some gold and silver." And Shabarsha began digging a pit and when he had dug it, he placed his hat with a cutout hole over it. "Shabarsha, oh Shabarsha! Grandfather says that you and I should trade. What will you take not to drag us out of the water?" "Fill that hat full of gold and silver." The imp dove back into the water. He came back and said, "Grandfather says that you and I should first wrestle." "What do you mean, milksop; why should I wrestle with the likes of you? You couldn't even deal with my middle brother, Misha." "And where is your Misha?" "Look over there. He's resting in that ravine under a bush." "How can I call him out?" "Go up to him and hit him in the side. Then he'll get up of his own accord." The imp went into the ravine, found the bear, and whacked him in the side with his club. Misha rose up on his hind legs and grabbed the imp so that all his bones cracked. He forced himself out of the bear's claws and ran to the old man in the water. "Well, grandfather." He said in his fright, "Shabarsha has a younger brother, Misha, and he was about to wrestle with me when he cracked all my bones. What would it have been like if I had started to wrestle Shabarsha?" "Hmm, go and try to have a foot race with Shabarsha; we'll see who can win that." So the boy in the red hat once more came up to Shabarsha. He repeated his grandfather's words and Shabarsha replied, "Why should I race with the likes of you? My little brother, Bunny, will leave you far behind." "And where is your little brother Bunny?" "He's over there, lying in the grass. He wanted a rest. Go up closer to him and touch him on the ear. Then he'll race with you." The imp ran up to Bunny, touched him on the ear, and the hare jumped up with the imp right behind him. "Wait, wait for me, Bunny. Let me come up even with you. Oh, you've got away!" "Well, grandfather," he said to the water spirit, "I

tore after him running. But how? He wouldn't let me catch up to him and he wasn't even Shabarsha - just his youngest brother." "Hmm," the old man muttered, screwing up his eyebrows. "Go to Shabarsha and tell him to see who can whistle loudest." "Shabarsha, Shabarsha! Grandfather orders me to see which of us can whistle loudest." "You whistle first." The imp whistled so loudly that Shabarsha could hardly stand on his own two feet and the leaves started falling off the trees. "You whistle well," said Shabarsha, "but that's not how I do it. When I whistle, you won't stand on your feet and your ears won't be able to stand it. Lie down on your face and cover your ears with your fingers." The imp lay face down on the ground and stuck his fingers in his ears. Shabarsha took his club and with all his might he whacked him on the neck and whistled. He whistled on and on. "Oh, grandfather, grandfather! You wouldn't believe how Shabarsha whistles. Sparks fell from my eyes and I could hardly get up from the ground, and all the bones in my neck and spine were broken." "Oho! You see, you're not so strong, imp. Go and take my iron club that's in the rushes and try this: see who can throw it higher into the air." The imp took the club, put it on his shoulder, and set off for Shabarsha. "Well, Shabarsha, grandfather has ordered us to try one last time. Who can throw this club higher into the air?" "You throw it first and I'll watch." The devil threw the club and it flew higher and higher until it was just a black dot in the sky. They waited impatiently for it to come back to earth. Then Shabarsha took the club - it was heavy! He stood it on the toes of one foot, leaned on it with his palm, and began gazing at the sky. "Why don't you throw it? What are you waiting for?" asked the imp. "I'm waiting until that little cloud comes up, and then I'll throw this club at it. My brother, the blacksmith, is sitting up there and he can use this iron in his business." "Oh, no, Shabarsha! Don't throw that club into the cloud or else grandfather will be angry." The devil grabbed the club and dove back to his grandfather. The grandfather heard from his grandson that Shabarsha had nearly thrown his club out of sight and he got seriously afraid and ordered them to drag the money out of the pool and buy him off. The imp dragged up more and more money; he dragged out a whole lot, but the hat still wasn't full. "Well, grandfather, that hat of Shabarsha is a marvel. I've put all your money in it, but it's still empty. There's only that one chest of yours full of money." "Carry it off to him quickly. Is he still weaving that line?" "He is." Then there was nothing else to do so the imp started with the secret chest and began filling Shabarsha's hat. He poured and poured it in, until finally he had filled it. And since then, since that time, the laborer has lived in glory.

C.6 IVANKO THE BEAR'S SON

In a certain village there lived a wealthy peasant and his wife. One day the wife went to the forest for mushrooms, lost her way, and stumbled into a bear's den. The bear kept her with him, and after some time, a long time or a short time, she had a son by him. This son was a man down to the waist and a bear below the waist; his mother called him Ivanko the Bearlet. Years went by, and when Ivan grew up he wanted to go away with his mother and live with the peasants in the village; they waited until the bear went to a beehive, made ready, and ran away. They ran and ran and finally came to their own village. The peasant saw his wife and was overjoyed - he had given up hope that she would ever return. Then he beheld her son and asked: "And who is this freak?" His wife told him all that had happened. How she had lived in the bear's den and had a son by him and that this son was human to the waist and a bear from the waist down. "Well, Bearlet," said the peasant. "Go to the back yard and slaughter a sheep; we must make dinner for you." "And which one shall I slaughter?" "Whichever one stares at you." Ivanko the Bearlet took a knife, went out to the yard, and called the sheep; all of them began to stare at him. He forthwith slaughtered them all, skinned them, and went to ask the peasant where he should store the skins and the meat. "What's this?" yelled the peasant. "I told you to slaughter one sheep, and you have slaughtered them all!" "No, father, you told me to slaughter whichever one stared at me; but when I came out into the yard all of them, without exception, began to stare at me." "You certainly are a clever fellow. Take the meat and skins into the barn, and at night guard the door against thieves and dogs." "Very well, I will guard it." It so happened that on that night a storm broke and the rain fell in buckets. Ivanko the Bearlet broke the door off the barn, took it into the bath house, and spent the night there. Thieves took advantage of the darkness; they found the barn open and without a guard, so they took whatever they pleased. Next morning the peasant arose, went to see whether everything was in order, and found that nothing was left: what the thieves had not taken, the dogs had eaten up. He looked for the guard, found him in the bath house, and began to chide him even more severely than the first time. "But, father, it is not my fault," said Ivanko. "You yourself told me to guard the door, and I did guard it. Here it is; the thieves did not steal it, nor did the dogs eat it up." "What can I do with this fool?" thought the peasant to himself. "If this goes on for a month or two, he will ruin me completely. I wonder how I can get rid of him." Then he hit upon an idea; the next day he sent Ivanko to the lake and told him to wind ropes of sand. In that lake dwelt many devils, and the peasant hoped that they would drag him into the water. Ivanko the Bearlet went to the lake, sat on the shore, and began to wind ropes

of sand. Suddenly a little devil jumped out of the water and asked: "What are you doing here, Bearlet?" "Can't you see? I'm winding ropes; I want to thrash the lake and torment you devils, because you live in our lake but do not pay any rent." "Wait a while, Bearlet, I'll run and tell my grandfather," said the little devil, and - flop! He jumped into the water. Five minutes later he was out again and said: "Grandfather said that if you can run faster than I, we'll pay the rent; if not, he told me to drag you down into the lake." "Aren't you a nimble fellow!" said Ivanko. "But you cannot hope to run faster than I. Why, I have a grandson who was born only yesterday, and even he can outrun you. Do you want to race with him?" "What grandson?" "He is lying there behind a bush," said the Bearlet, and cried to a hare. "Hey, hare, do not fail me!" The hare darted off into the open field like mad and in a trice vanished from sight; the devil rushed after him, but it was of no use; he was half a verst behind. "Now, if you wish," said Ivanko, "race with me. But on one condition - if you lag behind, I will kill you." "O no!" said the devil, and once more flopped into the water. After a while, he jumped out, carrying his grandfather's iron crutch, and said: "Grandfather said that if you can throw this crutch higher than I can, he will pay the rent." "Well, you throw first!" The devil threw the crutch so high that it was hardly visible; it fell back with a terrible rumble and thrust half its length into the ground. "Now you throw it," said the devil. The Bearlet took the crutch in his hand and could not even move it. "Wait a while," he said, "a cloud is coming near, I shall throw the crutch on it." "O no, that won't do, grandfather needs his crutch!" said the little devil. He snatched the crutch and rushed into the water. After a while, he jumped out again, saying: "Grandfather said that if you can carry this horse around the lake at least one more time than I can, he will pay the rent; if not, you will have to go into the lake." "Is that supposed to be hard? All right, begin!" answered Ivanko. The devil heaved the horse on to his back and dragged it around the lake; he carried it ten times, till he was exhausted and sweat streamed down his snout. "Well, now it's my turn," said Ivanko. He mounted the horse and began to ride around the lake; he rode so long that finally the horse collapsed under him. "Well, brother, how was that?" he asked the little devil. "I must admit," said the devil, "that you carried it more times than I, and in what a strange fashion! Between your legs! That way I couldn't have carried it even once! How much rent must we pay?" "Just fill my hat with gold, and work for a year as my laborer - that's all I want." The little devil ran to fetch the gold; Ivanko cut the bottom out of his hat and placed it above a deep pit; the devil kept bringing gold and pouring it into the hat. He worked at this for a whole day and only by evening was the hat filled. Ivanko the Bearlet got a cart, loaded it with gold coins, had the devil drag it home, and

said to the peasant: "Now be happy, father! Here is a laborer for you, and gold too."

C.7 THE RUNAWAY SOLIDER AND THE DEVIL

This soldier got permission to go on leave. He got ready and set off along the way. He walked and walked, but nowhere did he see any water and he wanted to wet his hardtack and eat a little along the way and road. His belly had long since been empty. There was nothing to be done and so he dragged himself on a bit further. Then he looked and saw a creek running. He went up to this creek and got three biscuits out of his pack and put them in the water. Besides the hardtack the soldier had this fiddle. In his free time he played various songs to chase away his boredom. So the soldier sat down next to the stream and began playing. Suddenly out of nowhere the Unclean One with the appearance of an old man came up to him with a book in his hands. "Greetings, mister soldier!" "Good health to you, my good man!" The devil wrinkled up his face when the soldier addressed him as "My good man!" "Listen, my friend: let's do a trade. I'll give you my book and you give me your fiddle." "Oh, old man, why do I need your book? I've served his majesty for ten years but I've never been able to read or write. I didn't know how before and now it's too late to learn." "Never mind, soldier, this is the sort of book that whoever looks at it will be able to read." "Alright, give it to me and I'll try." The soldier opened up the book and started reading, just as if he had done so from his early years. He was delighted and immediately handed over his fiddle. The Unclean One took it and began working at it with the bow, but no way did things go right! He couldn't make it play anything at all. "Listen, brother," he said to the soldier. "Let me stay as your guest for a couple of days and you teach me how to play this fiddle. I'll be very grateful." "No, old man," the soldier answered, "I have to go on home and in three days I'll be far, far away." "Please, soldier, if you'll stay and teach me to play this fiddle, I'll get you home in a single day; by a postal troika I'll drive you right there." The soldier sat and thought: should he stay or not? He took the hardtack out of the stream as he wanted to eat something. "Oh, brother soldier," said the Unclean One. "That's really awful stuff you've got there. Eat with me!" He opened his sack and got some white bread, some roast beef and vodka, and all sorts of other things. "Eat it, I don't want to." The soldier ate and drank his fill and then he agreed to stay on and teach this unknown old man how to play the fiddle. He stayed with him for three days and then he asked to go home. The devil led him outside and there before the porch stood a troika of fine horses. "Get in, soldier! I'll get you there in an instant!" The soldier got in the cart and they started the horses and so he went home - and how many versts flashed by his eyes! He got there in

nothing flat. "But will you recognize the village?" asked the Unclean One. "How could I not know it!" the soldier replied. "I was born in this village and I grew up here." "Well, farewell." The soldier got down from the cart, went in to his relatives, and they all exchanged greetings and asked about each other. For how long and when he had been given leave, It seemed to him that he had been with the Unclean One for no more than three days but in fact it had been three years. His leave had long since finished and he was considered a runaway in his regiment. The soldier became timid. He didn't know what to do. His stupidity simply wouldn't leave him. He went outside and thought, "Now where am I to go? If I go back to the regiment, they'll run me through the gauntlet. Oh, Unclean One, you've played a nasty trick on me." He had no sooner spoken these words than the Unclean One was there. "Don't be so sad, soldier! Stay with me; your service in the regiment wasn't much to brag about. They gave you dry biscuits to eat and they beat you with sticks. I'll make you happy. Do you want me to make you a merchant?" "Alright, that would be fine. Merchants live well and I'd like to try happiness." The Unclean One made him a merchant and gave him a large shop in the capital city with all sorts of valuable goods and he said. "Well, goodbye, brother. I'll be going beyond the thrice-nine land into the thrice-ten land. The king there has a splendid daughter, Princess Maria . I'm going to torment her." So our merchant lived and worried about nothing. Happiness just tumbled onto his household. And in his trading he had just one task; to become even more prosperous. The other merchants became envious, however. "Let's ask him what sort of person he is, where he came from, and does he know how to trade here? Why, he's taken all our trade away - let him have nothing for a while!" So they went to him and questioned him, and he replied to them: "My brothers, now I've got so much business that I've no time to discuss this with you. Come tomorrow and you'll find out everything." The merchants went their separate ways to their own homes. The soldier thought: what should he do? How should he answer them? He thought and thought and then he decided to leave his shop and go out of the town that very night. So he took all his money that was to hand and set out for the thrice-ten kingdom. He walked and walked until he came to a barrier. "Who are you?" asked the sentry. "I am a doctor and I'm coming to your tsardom because the daughter of your king is ill and I want to cure her," he answered. The sentry reported this to the courtiers and the courtiers reported this to the king himself. The king called in the soldier. "If you cure my daughter, I'll marry you to her." "Your majesty, just give me three decks of cards, three bottles of sweet wine, three bottles of strong spirits, three pounds of nuts, three pounds of bullets, and three bundles of strong wax candles." "Good, all will be ready." The soldier waited until evening, bought himself a fiddle, and set off to

the princess's. In her rooms he lit the candles and began drinking and carousing, and playing on his fiddle. At midnight the Unclean One came, heard the music, and rushed to the soldier. "Greetings, brother." "Greetings." "What are you drinking?" "I'm sipping a little kvass." "Give me some." "Certainly." But he brought him a full glass of the strong spirits. The devil drank it all and his eyes popped right out on his forehead. "Oh, that's strong stuff you've got. Let me have a bite to eat." "Here are some nuts, take them and eat," the soldier said, but he gave him some bullets. The devil chewed and chewed, but he just broke his teeth. Then they started playing cards. Meanwhile, time passed, the cocks crowed, and the Unclean One disappeared. The king asked the princess. "How did you sleep last night?" "Peacefully, thanks be to God!" The next night passed the same way. But on the third night the soldier asked the king: "Your majesty, order them to make some pincers weighing fifty pounds and three brass rods, three iron ones, and three lead ones." "Good, all will be ready." At deepest midnight the Unclean One appeared. "Greetings, soldier. I've come to carouse with you again." "Greetings! Who isn't happy with a cheerful companion!" They started drinking and carousing. The Unclean One saw the pincers and asked, "What is that?" "Well, you see, the king has taken me into his service and ordered me to teach some musicians to play the fiddle. But all their fingers are crooked, no better than yours, and so I've got to straighten them out in these pincers." "Oh, brother," asked the Unclean One, "Couldn't you straighten out my fingers? I still don't know how to play the fiddle." "And why not? Put your fingers just here." The devil put both hands in the pincers and the soldier squeezed them. Then he picked up the rods and let the devil have it! He beat him and repeated, "This is for your merchant's work!" The devil pleaded, the devil begged, "Let me go, please! I'll never come within thirty versts of the palace again!" But he kept on flogging him. The devil hopped and hopped. He twisted and twisted. With all his might he broke loose and said to the soldier "Even if you marry the princess, you won't get out of my hands. Once you go beyond the thirty versts, I'll grab you!" When he had said this, he disappeared. So the soldier married the princess and lived with her in love and harmony, and then after several years the king died and he began ruling the entire tsardom. Once the new king and his wife went into the garden to walk. "Oh, what a lovely garden," he said. "This isn't much of a garden," said the princess. "Just outside the town, about thirty versts from here, there is one to fall in love with." The king got ready and drove there with the queen. When he had stepped down from the carriage, the Unclean One came up and said. "Why have you come? Have you perhaps forgotten what you were told? Well, brother, you are at fault yourself. Now you'll never get out of my claws." What was he to do? "That is clearly my fate. Let me say farewell to my young wife." "Say goodbye then, but hurry!"

C.8 FROLKA STAY AT HOME

There was once a king who had three daughters, and such beauties they were as no tongue can tell of nor pen describe. Their garden was big and beautiful and they liked to walk there at night. A dragon from the Black Sea took to visiting this garden. One night the king's daughters tarried in the garden, for they could not tear their eyes away from the flowers; suddenly the dragon appeared and carried them off on his fiery wings. The king waited and waited but his daughters did not come back. He sent his maidservants to look for them in the garden, but all in vain; the maidservants could not find the princesses. The next morning the king proclaimed a state of emergency and a great multitude of people gathered. The king said: "Whoever finds my daughters, to him I shall give as much money as he wants." Three men agreed to undertake this task; a soldier who was a drunkard, Frolka Stay-at-Home, and Erema; and they set out to look for the princesses. They walked and walked till they came to a deep forest. As soon as they entered it they were overwhelmed by drowsiness. Frolka Stay-at-Home drew a snuffbox out of his pocket, tapped on it, opened it, shoved a pinch of tobacco into his nose, and cried: "Eh, brothers, let us not sleep, let us not rest, let us keep going." So they went on; they walked and walked and finally came to an enormous house, and in that house was a five-headed dragon. For a long time they knocked at the gate, but no one answered. Then Frolka Stay-at-Home pushed the soldier and Erema away and said: "Let me try, brothers!" He snuffed up some tobacco and gave such a knock at the gate that he smashed it. They entered the yard, sat in a circle, and were about to eat whatever they had. Then a maiden of great beauty came out of the house and said: "Little doves, why have you come here? A very wicked dragon lives here, who will devour you. You are lucky that he happens to be away." Frolka answered her: "It is we who shall devour him." He had no sooner said these words than the dragon came flying and roared: "Who has ruined my kingdom? Do I have enemies in the world? I have only one enemy, but his bones won't even be brought here by a raven." "A raven won't bring me," said Frolka, "but a good horse did." The dragon upon hearing this said: "Have you come for peaceful purposes or to fight?" "I have not come for peaceful purposes," said Frolka, "but to fight." They moved apart, faced each other, and clashed, and in one stroke Frolka cut off all the five heads of the dragon. Then he put them under a stone and buried the body in the ground. The maiden was overjoyed and said to the three brave men: "My little doves, take me with you." "But who are you?" they asked. She said that she was the king's eldest daughter; Frolka told her what task he had undertaken, and they were both glad. The princess invited them into the house, gave them meat and drink, and begged them to rescue her sisters. Frolka said: "We were

sent for them too!" The princess told them where her sisters were. "My next sister is even worse off than I was," she said. "She is living with a seven-headed dragon." "Never mind," said Frolka, "we shall get the better of him too; it may be somewhat harder to deal with a twelve-headed dragon." They said farewell and went on. Finally they came to the abode of the second sister. The house where she was locked up was enormous and all around it there was a high iron fence. They approached it and looked for the gate; finally they found it. Frolka banged upon the gate with all his strength and it opened; they entered the yard and, as they had done before, sat down to eat. Suddenly the seven-headed dragon came flying. "I smell Russian breath here," he said. "Bah, it is you, Frolka, who have come here! What for?" "I know what for," answered Frolka. He began to fight with the dragon and in one stroke cut off all seven of his heads, put them under a stone, and buried the body in the ground. Then they entered the house; they passed through one room, a second, and a third, and in the fourth they found the king's second daughter sitting on a sofa. When they told her why and how they had come there, she brightened, offered them food and drink, and begged them to rescue her youngest sister from the twelve-headed dragon. Frolka said: "Of course, that is what we were sent for. But there is fear in my heart. Well, perhaps God will help me! Give us each another cup!" They drank and left; they walked and walked till they came to a very steep ravine. On the other side of the ravine there stood enormous pillars instead of a gate, and on the pillars were chained two ferocious lions that roared so loudly that only Frolka remained standing on his feet; his two companions fell to the ground from fear. Frolka said to them: "I have seen worse terrors, and even then I was not frightened. Come with me!" And they went on. Suddenly an old man, who looked to be about seventy, came out of the castle; he saw them, came to meet them, and said: "Whither are you going, my friends?" "To this castle," answered Frolka. "Ah, my friends," said the old man, "you are going to an evil place; the twelve-headed dragon lives in this castle. He is not at home now, else he would have devoured you at once." "But he is the very one we have come to see," said Frolka. "If so," said the old man, "come with me, I will help you get to him." The old man went up to the lions and began to stroke them, and Frolka with his companions got through to the courtyard. They entered the castle; the old man brought them to the room where the princess lived. Upon seeing them she quickly jumped off her bed and began to question them as to who they were and why they had come. They told her. The princess offered them food and drink and began to make ready to go. As they were preparing to leave the house, they suddenly saw the dragon flying at a verst's distance from them. The king's daughter rushed back into the house and Frolka and his companions went out to meet and fight the dragon. At first the dragon attacked them

with great force, but Frolka, a clever fellow, managed to defeat him, cut off all of his twelve heads, and cast them into the ravine. Then they returned to the house and in their joy reveled even more than before. Following this feast they set out on their way, stopping only for the other princesses. Thus they all came back to their native land. The king was overjoyed, opened his royal treasury to them, and said: "Now, my faithful servants, take as much money as you want for a reward." Frolka was generous: he brought his big three-flapped cap, the soldier brought his knapsack, and Erema brought a basket. Frolka began to fill his cap first. He poured and poured, the cap broke, and the silver fell into the mud. Frolka began to pour again; he poured, and the money dropped from the cap! "There is nothing to be done," said Frolka. "Probably all of the royal treasury will fall to me." "And what will be left for us?" asked his companions. "The king has enough money for you too," said Frolka. While there was still money, Erema began to fill his basket, and the soldier his knapsack; having done this, they went home. But Frolka remained near the royal treasury with his cap and to this very day he is still sitting there, pouring out money for himself.

C.9 THE WITCH

There once lived an old couple who had one son called Ivashko; one can tell how fond they were of him! Well, one day, Ivashko said to his father and mother: "I'll go out fishing if you'll let me." "What are you thinking about! you're still very small; suppose you get drowned, what good will there be in that?" "No, no, I shan't get drowned. I'll catch you some fish; do let me go!" So his mother put a white shirt on him, tied a red girdle round him, and let him go. Out in a boat he sat and said: Canoe, canoe, float a little farther, Canoe, canoe, float a little farther! Then the canoe floated on farther and farther, and Ivashko began to fish. When some little time had passed by, the old woman hobbled down to the river side and called to her son: Ivashechko, Ivashechko, my boy,Float up, float up unto the waterside; I bring thee food and drink. And Ivashko said: Canoe, canoe, float to the waterside; That is my mother calling me. The boat floated to the shore: the woman took the fish, gave her boy food and drink, changed his shirt for him and his girdle, and sent him back to his fishing. Again he sat in his boat and said: Canoe, canoe, float a little farther, Canoe, canoe, float a little farther. Then the canoe floated on farther and farther, and Ivashko began to fish. After a little time had passed by, the old man also hobbled down to the bank and called to his son: Ivashechko, Ivashechko, my boy, Float up, float up, onto the waterside; I bring thee food and drink. And Ivashko replied: Canoe, canoe, float to the waterside; That is my father calling me. The canoe floated to the shore. The old man took the fish, gave his boy food

and drink, changed his shirt for him and his girdle, and sent him back to his fishing. Now a certain witch had heard what Ivashko's parents had cried aloud to him, and she longed to get hold of the boy. So she went down to the bank and cried with a hoarse voice: Ivashechko, Ivashechko, my boy,Float up, float up, onto the waterside I bring thee food and drink. Ivashko perceived that the voice was not his mother's, but was that of a witch, and he sang: Canoe, canoe, float a little farther, Canoe, canoe, float a little farther; That is not my mother, but a witch who calls me. The witch saw that she must call Ivashko with just such a voice as his mother had. So she hastened to a smith and said to him: "Smith, smith make me just such a thin little voice as Ivashko's mother has: if you don't, I'll eat you." So the smith forged her a little voice just like Ivashko's mother's. Then the witch went down by night to the shore and sang: Ivashechko, Ivashechko, my boy, Float up, float up, unto the waterside; I bring thee food and drink. Ivashko came, and she took the fish, and seized the boy and carried him home with her. When she arrived she said to her daughter Alenka, "Heat the stove as hot as you can, and bake Ivashko well, while I go and collect my friends for the feast." So Alenka heated the stove hot, ever so hot, and said to Ivashko, "Come here and sit on this shovel!" "I'm still very young and foolish," answered Ivashko: "I haven't yet quite got my wits about me. Please teach me how one ought to sit on a shovel." "Very good," said Alenka; "it won't take long to teach you." But the moment she sat down on the shovel, Ivashko instantly pitched her into the oven, slammed to the iron plate in front of it, ran out of the hut, shut the door, and hurriedly climbed up ever so high an oak-tree (which stood close by). Presently the witch arrived with her guests and knocked at the door of the hut. But nobody opened it for her. "Ah! That cursed Alenka!" she cried. "No doubt she's gone off somewhere to amuse herself." Then she slipped in through the window, opened the door, and let in her guests. They all sat down to table, and the witch opened the oven, took out Alenka's baked body, and served it up. They all ate their fill and drank their fill, and then they went out into the courtyard and began rolling about on the grass. "I turn about, I roll about, having fed on Ivashko's flesh!" cried the witch. "I turn about, I roll about, having fed on Ivashko's flesh." But Ivashko called out to her from the top of the oak: "Turn about, roll about, having fed on Alenka's flesh!" "Did I hear something?" said the witch. "No it was only the noise of the leaves." Again the witch began: "I turn about, I roll about, having fed on Ivashko's flesh!" And Ivashko repeated: "Turn about, roll about, having fed on Alenka's flesh!" Then the witch looked up and saw Ivashko, and immediately rushed at the oak on which Ivashko was seated, and began to gnaw away at it. And she gnawed, and gnawed, and gnawed, until at last she smashed two front teeth. Then she ran to a forge, and when she reached it she cried, "Smith, smith!

make me some iron teeth; if you don't I'll eat you!" So the smith forged her two iron teeth. The witch returned and began gnawing the oak again. She gnawed, and gnawed, and was just on the point of gnawing it through, when Ivashko jumped out of it into another tree which stood beside it. The oak that the witch had gnawed through fell down to the ground; but then she saw that Ivashko was sitting up in another tree, so she gnashed her teeth with spite and set to work afresh, to gnaw that tree also. She gnawed, and gnawed, and gnawed - broke two lower teeth, and ran off to the forge. Smith, smith!" she cried when she got there, "make me some iron teeth; if you don't I'll eat you!" The smith forged two more iron teeth for her. She went back again, and once more began to gnaw the oak. Ivashko didn't know what he was to do now. He looked out, and saw that swans and geese were flying by, so he called to them imploringly: Oh, my swans and geese, Take me on your pinions, Bear me to my father and my mother, To the cottage of my father and my mother. There to eat, and drink, and live in comfort. "Let those in the centre carry you," said the birds. Ivashko waited; a second flock flew past, and he again cried imploringly: Oh, my swans and geese! Take me on your pinions, Bear me to my father and my mother, To the cottage of my father and my mother. There to eat, and drink, and live in comfort. "Let those in the rear carry you!" said the birds. Again Ivashko waited. A third flock came flying up, and he cried: Oh, my swans and geese! Take me on your pinions, Bear me to my father and my mother, To the cottage of my father and my mother. There to eat, and drink, and live in comfort. And those swans and geese took hold of him and carried him back, flew up to the cottage, and dropped him in the upper room. Early the next morning his mother set to work to bake pancakes, baked them, and all of a sudden fell to thinking about her boy. "Where is my Ivashko?" she cried; "would that I could see him, were it only in a dream!" Then his father said, "I dreamed that swans and geese had brought our Ivashko home on their wings." And when she had finished baking the pancakes, she said, "Now, then, old man, let's divide the cakes: there's for you, father! There's for me! There's for you, father! There's for me." "And none for me?" called out Ivashko. "There's for you, father!" went on the old woman, "there's for me." "And none for me!" repeated the boy. "Why, old man," said the wife, "go and see whatever that is up there." The father climbed into the upper room and there he found Ivashko. The old people were delighted, and asked their boy about everything that had happened. And after that he and they lived on happily together.

C.10 THE SEVEN SIMEONS

Once, in a far away land, there was a man who had seven sons - all by the name of Simeon. They were seven lazy, good-for-nothing loafers. The laziest sons anyone ever had anywhere! They never did a single useful thing. This was unbearable for their father. Eventually, it got to the point that he simply brought all seven Simeons to the czar and enlisted them in his service. The czar was grateful to the father for bringing him so many fine, able men. He asked the father about their skills, "What do they know how to do?" "Ask them yourself, your royal majesty," their father said. Accordingly, the czar summoned the oldest Simeon and asked him, "What is your trade?" "Thievery, your royal majesty." "Hmm, you will probably be valuable to me at some point." He called in the second, "And what about you?" "I'm an artisan; I can make anything valuable." "You will also be useful." He called in the third Simeon, asking: "And what can you do?" "I can shoot any bird - even in flight, your royal majesty." "Great!" said the czar. He asks the fourth, "And you?" "If a marksman shoots a bird, I will retrieve it, better than any dog." "Fine!" "And what are you proficient at?" he asked the fifth. "From a vantage point, I can see and then relate all that is happening in any part of the kingdom." "Sounds great!" After the fifth Simeon, he questioned the sixth. "I'm very good at building boats; for me it's 'slam-bam,' and I've built a boat." "Good." Then asking the seventh, "And what do you do?" "I am a healer." "Sounds good." said the czar. Then he dismissed them all. After some time passed, the czar remembering the seven Simeons, decided to put one of their skills to use. The king asked the fifth Simeon, "Okay, Simeon, will you find out what is going on in various places?" Simeon climbed to a high place, looked around and related, "Here such-and-such is happening, and there, such-and-such." They checked his statements with the newspapers and found he was exactly right! After another long period passed, the czar decided to marry a certain princess. But, how was he to abduct her? He just didn't know; there was no one suitable for the job. Then he remembered the seven Simeons. Summoning them, he commissioned them as soldiers, with the assignment to bring him that princess. The Simeons quickly convened. As expert craftsmen they, slam-bam, built a boat, boarded and sailed off to the kingdom of the bride-to-be-princess. One, looking from the high mast, said the princess was currently alone and so vulnerable. The maker of valuables went with his brother, the thief, to sell his wares in the palace. They barely arrived when the thief stole the princess. They immediately cut anchor and set sail. When the princess realized they were carrying her away, she turned into a white swan and flew from the boat. The marksman, without delay, grabbed his gun and fired a shot into her left wing. Simeon number four darted into the water,

retrieved the swan, as any good dog would, bringing her back to the boat. The swan turned back into a princess, but her left arm was still injured. Healer Simeon immediately cured the princess' arm. With their healthy and successful return to their own kingdom, the Simeons fired a round from the cannon. The czar heard it, but had already forgotten about the Simeons. He thought to himself, "What's going on with that boat out there?" "Please go," he said. "Run and find out what's going on." Whether somebody ran or rode, they quickly returned with news concerning the seven Simeons and the czar's future bride. The czar rejoiced in the Simeons' success and commanded they be met with honor, cannon-fire, and the beating of drums. Only the princess didn't want marry the czar. He was already too old! Therefore, he asked her who she wanted to marry. The princess answered, "I want to marry the man who kidnapped me!" Afterall, the thief was a dashing fellow whose nobility appealed to the princess. The czar, without another word, ordered them to be married. Years later, he wanted some peace and quiet, so he placed Simeon the Thief on the throne, and made all of his brothers great noblemen.

C.11 IVAN POPYALOV

Once upon a time there was an old couple, and they had three Sons. Two of these had their wits about them, but the third was a simpleton, Ivan by name, surnamed Popyalof. For twelve whole years Ivan lay among the ashes from the stove; but then he arose, and shook himself, so that six poods of ashes fell off from him. Now in the land in which Ivan lived there was never any day, but always night. That was a Snake's doing. Well, Ivan undertook to kill that Snake, so he said to his father, "Father make me a mace five poods in weight." And when he had got the mace, he went out into the fields, and flung it straight up in the air, and then he went home. The next day he went out into the fields to the spot from which he had flung the mace on high, and stood there with his head thrown back. So when the mace fell down again it hit him on the forehead. And the mace broke in two. Ivan went home and said to his father, "Father, make me another mace, a ten pood one." And when he had got it he went out into the fields, and flung it aloft. And the mace went flying through the air for three days and three nights. On the fourth day Ivan went out to the same spot, and when the mace came tumbling down, he put his knee in the way, and the mace broke over it into three pieces. Ivan went home and told his father to make him a third mace, one of fifteen poods weight. And when he had got it, he went out into the fields and flung it aloft. And the mace was up in the air six days. On the seventh Ivan went to the same spot as before. Down fell the mace, and when it struck Ivan's forehead, the forehead bowed under it. Thereupon he said, "This mace will do for the Snake!" So when he had got everything

ready, he went forth with his brothers to fight the Snake. He rode and rode, and presently there stood before him a hut on fowl's legs, and in that hut lived the Snake. There all the party came to a standstill. Then Ivan hung up his gloves, and said to his brothers, "Should blood drop from my gloves, make haste to help me." When he had said this he went into the hut and sat down under the boarding. Presently there rode up a Snake with three heads. His steed stumbled, his hound howled, his falcon clamored. Then cried the Snake: "Wherefore hast thou stumbled, O Steed! hast thou howled, O Hound! hast thou clamored, O Falcon?" "How can I but stumble," replied the Steed. "When under the boarding sits Ivan Popyalof?" Then said the Snake, "Come forth, Ivanushka! Let us try our strength together." Ivan came forth, and they began to fight. And Ivan killed the Snake, and then sat down again beneath the boarding. Presently there came another Snake, a six-headed one, and him, too, Ivan killed. And then there came a third, which had twelve heads. Well, Ivan began to fight with him, and lopped off nine of his heads. The Snake had no strength left in him. Just then a raven came flying by, and it croaked: "Krof? Krof!" Then the Snake cried to the Raven, "Fly, and tell my wife to come and devour Ivan Popyalof." But Ivan cried: "Fly, and tell my brothers to come, and then we will kill this Snake, and give his flesh to thee." And the Raven gave ear to what Ivan said, and flew to his brothers and began to croak above their heads. The brothers awoke, and when they heard the cry of the Raven, they hastened to their brother's aid. And they killed the Snake, and then, having taken his heads, they went into his hut and destroyed them. And immediately there was bright light throughout the whole land. After killing the Snake, Ivan Popyalof and his brothers set off on their way home. But he had forgotten to take away his gloves, so he went back to fetch them, telling his brothers to wait for him meanwhile. Now when he had reached the hut and was going to take away his gloves, he heard the voices of the Snake's wife and daughters, who were talking with each other. So he turned himself into a cat, and began to mew outside the door. They let him in, and he listened to everything they said. Then he got his gloves and hastened away. As soon as he came to where his brothers were, he mounted his horse, and they all started afresh. They rode and rode; presently they saw before them a green meadow, and on that meadow lay silken cushions. Then the elder brothers said, "Let's turn out our horses to graze here, while we rest ourselves a little." But Ivan said, "Wait a minute, brothers!" and he seized his mace, and struck the cushions with it. And out of those cushions there streamed blood. So they all went on further. They rode and rode; presently there stood before them an apple-tree, and upon it were gold and silver apples. Then the elder brothers said. "Let's eat an apple apiece." But Ivan said, "Wait a minute, brothers; I'll try them first," and he took his mace, and struck the apple-tree

with it. And out of the tree streamed blood. So they went on further. They rode and rode, and by and by they saw a spring in front of them. And the elder brothers cried, "Let's have a drink of water." But Ivan Popyalof cried: "Stop, brothers!" and he raised his mace and struck the spring, and its waters became blood. For the meadow, the silken cushions, the apple-tree, and the spring, were all of them daughters of the Snake. After killing the Snake's daughters, Ivan and his brothers went on homewards. Presently came the Snake's Wife flying after them, and she opened her jaws from the sky to the earth, and tried to swallow up Ivan. But Ivan and his brothers threw three poods of salt into her mouth. She swallowed the salt, thinking it was Ivan Popyalof, but afterwards - when she had tasted the salt, and found out it was not Ivan - she flew after him again. Then he perceived that danger was at hand, and so he let his horse go free, and hid himself behind twelve doors in the forge of Kuzma and Demian. The Snake's Wife came flying up, and said to Kuzma and Demian, "Give me up Ivan Popyalof." But they replied: "Send your tongue through the twelve doors and take him." So the Snake's Wife began licking the doors. But meanwhile they all heated iron pincers, and as soon as she had sent her tongue through into the smithy, they caught tight hold of her by the tongue, and began thumping her with hammers. And when the Snake's Wife was dead they consumed her with fire, and scattered her ashes to the winds. And then they went home, and there they lived and enjoyed themselves, feasting and revelling, and drinking mead and wine.

C.12 THE SERPENT AND THE GYPSY

In old times there was a village where a serpent frequently flew. He had devoured all but one of the villagers. At that time a gypsy came to the village. He came at night and no matter where he looked, he couldn't find anyone! He finally entered the last hut where the last man was sitting and crying. "Greetings, good man!" "Why are you here, gypsy? Surely, you must be tired of life." "Why do you say that?" "You see, the serpent has been coming here devouring people. He has eaten everyone. He only left me alone until tomorrow morning when he will fly back and devour me too. In fact, he won't spare you either! He'll eat us both in one sitting!" "Or maybe he'll choke! Why don't you let me sleep here tonight? Tomorrow I'll see what kind of serpent this is." So, they spent the night. The next morning, a strong windstorm suddenly arose. The hut shook when the serpent landed. "Aha," said the serpent, "I've already turned a profit! I left only one man, but now I have two. Now I'll have something for lunch as well!" "Do you really think you can eat us?" asked the gypsy. "Yes, indeed, I will eat you." "You lie, devilish fiend. You will choke!" "So then, you think you are stronger than me?" "I should say so! I think even

you realize I have greater strength than you." "Well then, let's just see who's stronger between the two of us." "Let's!" The serpent grabbed a millstone. "Look at this, gypsy! I will crush this stone with only one hand." "Fine! I'll watch." The serpent grabbed the stone with the palm of his hand and squeezed it so tightly that it turned into fine sand - a cloud of dust arose from his fist. "What a wonder," said the gypsy, "but try gripping a rock so hard that water comes out of it. Watch how I squeeze!" There was a chunk of cheese curd on the table, the gypsy grabbed it and really gave it a squeeze. Whey dripped out onto the ground. "Well, did you see that? So who's stronger?" "True, your hand is stronger than my own. Let's see who can whistle the loudest." "Well, whistle!" The serpent whistled so loudly that leaves fell from all the trees. "You whistle very well, brother, but still not as well as I," said the gypsy. "First, you had better protect your eyeballs, or else they'll pop right out of your head!" The serpent believed it and covered his eyes with a scarf. "Okay, go ahead and whistle." The gypsy grabbed a club and as he whistled he thumped the serpent, who yelled at the top of his voice, "That's enough! That's enough, gypsy! Don't whistle anymore! In one go, my eyes have almost come right out." "Alright, however, I am quite prepared to whistle some more." "No, no, you don't need to do that. I don't want to argue with you anymore. Better yet, let's become brothers. You be the older brother, and I'll be the younger." "Agreed!" "Well, brother," said the serpent, "if you please, out there in the meadow a herd of oxen is grazing. Pick out the fattest, grab it by the tail and drag it back for lunch." There was nothing to be done - the gypsy went to the meadow. He saw a large herd of oxen grazing. He started to tie them together by their tails. The serpent waited and waited and couldn't wait any longer. He ran out to the meadow to see what was going on. "What is taking so long?" "Just wait a little longer. I'm tying together around fifty of these. Then in one trip I can drag them all back home so that we'll have enough for a whole month!" "Oh, you! Why do we need to spend our whole lives here? One is enough." Then the serpent grabbed the biggest oxen by the tale, ripped off the hide, shouldered the meat and dragged it home. "Brother, what about all the animals I tied together? We're not going to just leave them?" "Forget them." They returned to the hut, and set up two cauldrons for the beef, but there was no water. "Here, take this ox-hide," said the serpent to the gypsy, "if you please, fill this up with water and carry it back here. We'll start boiling our lunch." The gypsy grabbed the hide and dragged it to the well. Completely empty he could barely drag it, to say nothing of it when filled with water. He got to the well and started to dig a deep trench around it. The serpent again waited and waited and couldn't wait any longer. He ran out to see for himself what was taking so long. "What are you doing, brother?" "I want to dig all the way around the well and then drag the whole well back

to the hut, so that you won't have to go out for water anymore." "Oh, you! You're trying to do too much. To dig out the well would take a long time." The serpent dropped the hide into the well, filled it with water, dragged it out and carried it home. "Brother," he said to the gypsy, "go into the forest. Choose a dry oak and pull it to the hut. It's time to start a fire!" The gypsy went into the forest and pulled out bark fiber to twist into a rope. He braided an extremely long rope and set to wrapping up the oak trees. The serpent waited and waited and couldn't wait any longer. He ran out to see for himself what was taking so long. "Why are you dilly-dallying?" "Well, I want to bring in twenty oaks at once by tying them together and then dragging them in, roots and all, so that you will have firewood for a long time!" "Oh, you! You do everything your own way," said the serpent, then he ripped out the widest oak with its roots and dragged it to the hut. The gypsy pretended like he was extremely angry. Pouting, he sat silently. The serpent cooked the beef, and called the gypsy to lunch. But the gypsy fervently replied, "I don't want to!" And so the serpent scarfed a whole ox, drank a whole oxen-skin of water and then questioned the gypsy, "Tell me, brother, why are you upset?" "Because whatever I do, it's just not right. It's not how you like it done!" "Well, don't be angry. Let's make peace!" "If you want to make peace with me, then visit me as my guest." "Please, I'm ready now, brother!" Right away, the serpent got his wagon, harnessed three of the best stallions, and they left together for the gypsy camp. They were just approaching, when the little gypsy children saw their father. Barely dressed, they ran to meet him. At the top of their voices they exclaimed, "Father has arrived! He brought a serpent!" The serpent became frightened and asked the gypsy, "Who are they?" "Those are my children! I suppose they are hungry right now. Look how greedily they are coming for you." The serpent jumped out of the wagon and ran away. The gypsy sold the trio of horses with the wagon and began to live the good life.

C.13 PRINCE DANILA GOVORILA

There was once an old princess; she had a son and a daughter, both well built, both handsome. A wicked witch disliked them; she pondered and pondered as to how she could lead them into evil ways and destroy them. In the end she conceived a plan. Like a cunning fox she came to their mother and said: "My little dove, my dear friend, here is a ring for you; put it on your son's finger. With its help he will be healthy and wealthy, but he must never take it off, and he must marry only that maiden whom the ring fits." The old woman believed her and was overjoyed; before her death she enjoined upon her son that he take to wife a woman whom the ring would be found to fit. Time went by and the little son grew up. He grew up and began to seek a bride; he would like one girl, then another, but upon trying the ring

he always found it to be too big or too small; it did not fit either the one or the other. He traveled and traveled through villages and cities, tried the ring on all the lovely maidens, but could not find one whom he could take as his betrothed; he returned home and was pensive and sad. "Little brother, why are you grieving?" his sister asked him. He told her his trouble. "Why is the ring so troublesome?" said the sister. "Let me try it." She put it on her finger and the ring clasped it, and began to gleam; it fitted her as though made to her size. "Ah, my sister," said the brother "you have been chosen for me by fate, you shall be my wife." "What are you saying, my brother? Think of God, think of the sin; one does not marry one's own sister." But her brother did not heed her; he danced for joy and ordered that preparations be made for the wedding. The sister burst into bitter tears, went out of her room, sat on the threshold, and wept and wept. Some old women passed by; she invited them in and offered them food and drink. They asked her what her grief was, why she was sad. It was of no use to hide it; she told them everything. "Weep not, grieve not," said the old women. "But listen to us. Make four little dolls, seat them in the four corners; when your brother calls you to your wedding, go; when he asks you to come to the bridal chamber, do not hurry. Put your hope in God. Farewell!". The old women left. The brother wed his sister, went to the room, and said: "Sister Catherine, come to the featherbed." She answered: "I will come in a minute, only let me remove my earrings." And the dolls in the four corners cried like cuckoos: Cuckoo, Prince Danila, Cuckoo, Govorila, Cuckoo, he takes his sister, Cuckoo, for a wife, Cuckoo, earth open wide, Cuckoo, sister, fall inside! The earth began to open, the sister began to fall in. Her brother cried: "Sister Catherine, come to the featherbed!" "Just a minute, my brother, let me unclasp my girdle." The dolls cuckooed: Cuckoo, Prince Danila, Cuckoo, Govorila, Cuckoo, he takes his sister, Cuckoo, for a wife, Cuckoo, earth open wide, Cuckoo, sister, fall inside! Only the sister's head was still above ground. The brother called her again: "Sister Catherine, come to the featherbed!" "Just a minute, my brother, I must remove my slippers." The dolls cuckooed, and she vanished into the earth. The brother called her, he called her again in a louder voice, but she did not come. He ran to her room, banged at the door, and the door broke. He looked everywhere, but his sister was gone. Only the dolls were sitting in the corners and crying: "Earth, open wide! Sister, fall inside!" He seized an axe, cut off their heads, and threw them into the stove. The sister walked and walked underground and saw a little hut on chicken legs, turning round and round. "Little hut, little hut," she said, "stand the old way with your back to the woods and your front to me." The little hut stood still and the door opened. Inside sat a lovely maiden embroidering a towel with silver and gold. She received her guest with kindness, then sighed and said: "My little dove, my heart is glad to

see you, I will welcome you and fondle you while my mother is out. But when she comes back there will be trouble for both of us, for she is a witch." The guest was frightened by these words, but she had nowhere to go, so she sat with her hostess at the embroidery frame; they embroidered the towel and talked together. After a long time or a short time, when the hostess knew that her mother was about to come. She turned her guest into a needle, thrust the needle into a birch broom, and put the broom in a corner. She had no sooner done all this than the witch appeared at the door. "My good daughter, my comely daughter, I smell a Russian bone!" the witch said. "Madam mother, passers-by came in to drink some water." "Why did you not keep them here?" "They were old people, my mother, they would not have been to your liking." "Henceforth, mind you, invite all into the house, do not let anyone go; I will leave now to get some booty." She left; the maidens sat at the frame, embroidered the towel, talked and laughed together. The witch came flying home; she sniffed about in the house. "My good daughter, my comely daughter, I smell a Russian bone!" she said. "Some little old men stopped in to warm their hands; I tried to keep them but they did not want to stay." The witch was hungry; she chided her daughter, and flew away again. The guest had been sitting in the broom. They hastened to finish embroidering the towel; while working thus hurriedly they planned how to escape from their trouble and run away from the wicked witch. They had hardly had time to exchange a few whispers, when the witch (talk of the devil and he will appear) stood in the doorway, catching them by surprise. "My good daughter, my comely daughter, I smell a Russian bone!" she cried. "There, my mother, a lovely maiden is awaiting you," said her daughter. The maiden looked at the witch and her heart failed her. Before her stood Baba Yaga the Bony-legged, her nose hitting the ceiling. "My good daughter, my comely daughter, make a good hot fire in the stove," said the witch. They brought wood, oak and maple, and made a fire; the flame blazed forth from the stove. The witch took a broad shovel and began to urge her guest: "Now, my beauty, sit on the shovel." The beauty sat on it. The witch shoved her toward the mouth of the stove, but the maiden put one leg into the stove and the other on top of it. "You do not know how to sit, maiden. Now sit the right way," said the witch. The maiden changed her posture, sat the right way; the witch tried to shove her in, but she put one leg into the stove and the other under it. The witch grew angry and pulled her out again. "You are playing tricks, young woman!" she said. "Sit quietly, this way - just see how I do it." She plumped herself on the shovel and stretched out her legs, and the two maidens quickly shoved her into the stove, locked her in, covered her up with logs. They plastered and tarred the opening, and then ran away, taking the embroidered towel and a brush and comb with them. They ran and ran, and looking back beheld the wicked witch;

she had wrenched herself free, caught sight of them, and was hissing: "Hey, hey, hey, are you there?" What could they do? They threw down the brush and there appeared a marsh thickly overgrown with reeds. The witch could not crawl through it, but she opened her claws, plucked out a path, and again came close. Where could they go? They threw down the comb, and there appeared a dark, thick forest: not even a fly could fly through it. The witch sharpened her teeth and set to work: each time she clamped her teeth she bit off a tree by its roots. She hurled the trees to one side, cleared a path, and again came close - very close. The maidens ran and ran till they could run no longer; they had lost all their strength. They threw down the gold-embroidered towel, and there spread before them a sea, wide and deep, a sea of fire. The witch soared high; she wanted to fly across the sea, but fell into the fire and was burned. The maidens remained alone, little doves without a home; they did not know where to go. They sat down to rest. A servant came to them and asked them who they were, then reported to his master that in his domain sat not two little birds of passage but two marvelous beauties, one exactly like the other; they had the same brows, the same eyes. "One of them," said the servant, "must be your sister, but which of the two that is, it is impossible to guess." The master went to see them and invited them to his home. He saw that his sister was there, but which of the two she was he could not guess; his servant had told the truth. She was angry and would not tell him herself. What could be done? "This is what can be done, master," said the servant. "I will fill a sheep's bladder with blood, you will put it under your arm, and while you speak to your guests, I will come near you and strike you with a knife in your side; blood will flow and your sister will reveal herself." "Very well !" They did what they had planned; the servant struck his master in the side and blood gushed forth. The brother fell, the sister rushed to embrace him, and she cried and lamented: "My beloved, my dearest!" The brother jumped up, safe and sound, embraced his sister, and married her to a good man; and he himself married her friend, on whose finger the ring fitted, and all of them lived happily forever after.

C.14 THE MERCHANT'S DAUGHTER AND THE MAIDSERVANT

There was once a very wealthy merchant who had a marvelously beautiful daughter. This merchant carried goods to various provinces, and one day he came to a certain kingdom and brought precious cloths to the king as a gift. The king said to him: "Why can I not find a bride for myself?" The merchant answered: "I have a beautiful daughter, and she is so clever that no matter what a man is thinking, she can guess it." The king immediately wrote a letter and called his guards. "Go to that merchant's house," he told them, "and deliver

this letter to the merchant's daughter." The letter said: "Make ready to get married." The merchant's daughter took the letter, burst into tears, and prepared to go, taking also her maidservant; and no one could distinguish this maid from the merchant's daughter, they were so like each other. They dressed in dresses that were alike and went to the king for the marriage. The maid was full of spite, and said: "Let us take a walk on the island." They went to the island; there the maidservant gave the merchant's daughter sleeping potions, cut out her eyes, and put them in her pocket. Then she came to the guards and said: "Gentlemen of the guard, my maid servant has gone to sea." They answered: "We need only you, we have no use for that peasant girl." They went to the king; he married the maid at once, and they began to live together. The king thought to himself: "The merchant must have cheated me; she cannot be a merchant's daughter. Why is she so ignorant? She does not know how to do anything." Meanwhile the merchant's daughter recovered from the illness that her maid had brought upon her. She could not see; she could only hear and she heard an old man tending cattle. She said to him: "Where are you, grandfather?" "I live in a little hut." "Please take me into your hut." The old man took her in. She said to the old man: "Little grandfather, drive out the cattle." He heeded her and drove away the cattle. She sent the old man to a shop, saying: "Get velvet and silk on credit." The old man went; none of the wealthy merchants would give him goods on credit but a poor shopkeeper gave him some. He brought the velvet and silk to the blind maiden. She said to him: "Little grandfather, lie down to sleep. As for me, day and night are one and the same." And she began to sew a royal crown of velvet and silk; she embroidered such a beautiful crown that it was a pleasure to behold it. Next morning the blind maiden roused the old man and said: "Go and take this to the king, and for payment accept only an eye. And fear not, no matter what they do to you." The old man went to the palace with the crown. Everyone admired it and wanted to buy it from him, but the old man asked for an eye in payment. Straightway the king was told that he was asking for an eye. The king came out, was delighted with the crown, and began to bargain for it, but the old man still asked for an eye. The king began to curse and threatened to put him in prison; but no matter what the king said, the old man held his ground. Then the king cried to his guards: "Go and cut out an eye from a captive soldier." Just then his wife, the queen, rushed out, took an eye from her pocket, and gave it to the king. The king was overjoyed. "Ah, you have helped me out, little queen !" he said, and gave the eye to the old man, who took it, left the palace, and returned to his hut. The blind maiden asked him: "Did you get my eye, little grandfather?" He said: "I did." She took it from him, went outside at twilight, spat upon it, put it into its socket, and was able to see once more. Then again she sent the old man to the shops, giving him

money, and asked him to pay what he owed for the silk and velvet and to bring more velvet and gold thread. He got what he needed from the poor shopkeeper and brought these things to the merchant's daughter. She sat down to sew another crown, finished it, and sent the old man to the same king. "Do not take anything," she told him, "but an eye. And if you are asked where you got this crown, answer only: 'God gave it to me.'" The old man came to the palace. There everyone was amazed, for although the first crown was beautiful, the second was even lovelier. The king said: "I will buy it from you at any price." "Give me an eye," said the old man. The king at once ordered a guard to cut away an eye from a prisoner; but his wife again gave him an eye. The king was overjoyed and thanked her, saying: "Ah, little mother, you have been a great help to me!" The king asked the old man: "Where do you get these crowns?" He answered: "God gives them to me." And he left the palace. He came to his hut and gave the eye to the blind maiden. She again went outside at twilight, spat on the eye, put it in its socket, and could see with both eyes. She lay down to sleep in the hut, but upon awakening she suddenly found herself in a glass house and began to live in magnificent style. The king went to see this marvel, wondering who had built such a fine house. He drove into the yard and the merchant's daughter was delighted. She received him hospitably and bade him sit down at table. He feasted there and upon leaving asked the maiden to come to see him. He returned to his palace and said to his queen: "Ah, little mother, what a house there is in such and such a place! And what a maiden there is in it! No matter what one is thinking, she knows it." The queen guessed who it was and thought to herself: "It must be the same maiden whose eyes I cut out." The king again went to visit the maiden, and the queen was full of spite. The king came, feasted, and invited her to his palace. She began to make ready and said to the old man: "Farewell! Here is a chest of money; you will never reach its bottom, it will always be full. You will go to sleep in this glass house, but you will awaken in your old hut. Now I am going to make a visit. I shall not be alive tomorrow; I shall be killed and cut into little pieces. Arise in the morning, make a coffin, gather my remains, and bury them." The old man wept for her. Soon the guards came, seated her in a carriage, and drove away. They brought her to the king's palace, and the queen did not even look at her; she wanted to shoot her on the spot. She went out into the courtyard and said to the guards: "When you bring this maiden home, cut her into little pieces at once, take out her heart, and bring it to me." They took the merchant's daughter home and talked to her glibly, but she knew what they wanted to do and said to them: "Cut me up quickly." They cut her in pieces, took out her heart, buried her in the ground, and returned to the palace. The queen came out to meet them, took the heart, rolled it up into an egg, and put it in her pocket. The old man

went to sleep in a glass house but awoke in a hut and burst into tears. He wept and wept, then he set about his appointed task. He made a coffin and went to seek the maiden; he found her in the earth, dug her up, gathered all the pieces, put them in the coffin, and buried them in his own land. The king did not know of all this, so he went again to visit the merchant's daughter. When he arrived at the place, there was no house, no maiden; but at the spot where she was buried a garden had grown. He returned to the palace and told the queen: "I drove and drove, but I found neither house nor maiden, only a garden." When the queen heard this she went into the courtyard and said to the guards: "Go and cut that garden down." They came to the garden and began to cut it but it turned into stone. The king longed to see the garden again and went to that place. When he came he beheld a boy there - and what a handsome boy he was! "Surely some lord went for a drive and lost him here," he thought. He took the boy to his palace and said to his queen: "Mind you, little mother, do not maltreat him." Meanwhile the boy began to cry and there was no way of appeasing him: no matter what they gave him, he kept on crying. Then the queen took from her pocket the egg she had made from the maiden's heart and gave it to the boy; he ceased crying and began to skip around the rooms. "Ah, little mother," said the king to the queen, "you have made him happy." The boy ran to the yard and the king ran after him; the boy ran into the street and the king ran into the street; the boy ran to the field and the king ran to the field; the boy ran to the garden and the king ran to the garden. There the king saw the maiden and was overjoyed. She said to him: "I am your bride, the merchant's daughter, and your queen is my maidservant." They went to the palace. The queen fell at her feet. "Forgive me," she said. "You have never forgiven me," said the merchant's daughter. "Once you cut out my eyes, and then you ordered me cut in little pieces." The king said: "Guards, cut out her eyes and let her be dragged by a horse over the field." The maidservant's eyes were cut out, she was tied to a horse, and dragged to her death over the open field. And the king began to live happily with the young queen and to prosper. The king always delighted in her and dressed her in gold.

C.15 DAWN, EVENING, AND MIDNIGHT

In a certain kingdom there was a king who had three daughters of surpassing beauty. The king guarded them more carefully than his most precious treasure; he built underground chambers and kept his daughters there like birds in a cage, so that rough winds could not blow upon them nor the red sun scorch them with his rays. One day the princesses read in a certain book that there was a marvelous bright world; and when the king came to visit them, they straightway began to implore him with tears in their eyes, saying: "Sovereign, our

father, let us out to see the bright world and walk in the green garden." The king tried to dissuade them but to no avail. They would not even listen to him; the more he refused, the more urgently they besought him. There was nothing to be done, so the king granted their insistent prayer. And so the beautiful princesses went out to walk in the garden. They beheld the red sun, the trees, and the flowers, and were overjoyed that they had the freedom of the bright world. They ran about in the garden and enjoyed themselves when a sudden whirlwind seized them and carried them off far and high, no one knew whither. The alarmed nurses and governesses ran to report this to the king; the king straightway sent his faithful servants in all directions, promising a great reward to him who should find traces of the princesses. The servants traveled and traveled but did not discover anything and came back no wiser than they had set out. The king called his grand council together and asked his councilors and boyars whether anyone among them would undertake to search for his daughters. To any man who might find them, he said, he would give the princess of his choice in marriage, and a rich dowry. The king asked once and the boyars were silent; he asked a second time and they still did not answer; he asked a third time and no one made a sound! The king burst into tears. "Apparently I have no friends or helpers here," he said, and ordered that a call be issued throughout the kingdom. He hoped that someone from among the common people would undertake the heavy task. At that time there lived in one village a poor widow who had three sons; they were mighty champions. All of them were born in one night; the eldest in the evening, the second at midnight, and the youngest in the early dawn, and therefore they were called Evening, Midnight, and Dawn. When the king's call reached them, they straightway asked for their mother's blessing, made ready for their journey, and rode to the capital city. They came to the king, bowed low, and said: "Rule for many years, sovereign! We have come to you not to celebrate a feast, but to perform a task. Give us leave to go in search of your daughters." "Hail, good youths! What are your names?" "We are three brothers - Evening, Midnight, and Dawn." "What shall I give you for your voyage?" "We do not need anything, sire; only do not forget our mother, care for her in her poverty and old age." The king took the old woman into his palace, and ordered that she be given food and drink from his table and clothes and shoes from his stores. The good youths set out on their way. They rode one month, a second, and a third; then they came to a wide desert steppe. Beyond that steppe was a thick forest, and close to the forest stood a little hut. They knocked at the window and there was no answer; they entered and no one was in the hut. "Well, brothers," said one of the three, "let us stop here for a time and rest from our travels." They undressed, prayed to God, and went to sleep. Next morning Dawn, the youngest brother, said

to Evening, his eldest brother: "We two shall go hunting, and you stay at home and prepare our dinner." The eldest brother consented. Near the hut there was a shed full of sheep; without thinking much he took the best ram, slaughtered and cleaned it, and put it onto roast for dinner. He prepared everything and lay down to rest on a bench. Suddenly there was a rumbling noise, the door opened, and there entered a little man as big as a thumb, with a beard a cubit long. He cast an angry look around and cried to Evening: "How dared you make yourself at home in my house, how dared you slaughter my ram?" Evening answered: "First grow up - otherwise you cannot be seen from the ground! I shall take a spoonful of cabbage soup and a crumb of bread and throw them in your eyes!" The old man as big as a thumb grew more furious: "I am small but strong!" He snatched up a crust of bread and began to beat Evening on the head with it; he beat him till he was half dead and threw him under the bench. Then the little old man ate the roasted ram and went into the woods. Evening tied a rag around his head and lay moaning. The brothers returned and asked him: "What is the matter with you?" "Eh, brothers, I made a fire in the stove, but because of the great heat I got a headache; I lay all day like one dazed, I could neither cook nor roast!" Next day Dawn and Evening went hunting, and Midnight was left at home to prepare the dinner. Midnight made a fire, chose the fattest ram, slaughtered it, and put it in the oven; then he lay on the bench. Suddenly there was a rumbling noise, and the old man as big as a thumb, with a beard a cubit long, came in and began to beat and thrash him; he almost beat him to death. Then he ate the roasted ram and went into the woods. Midnight tied up his head with a handkerchief and lay under the bench and moaned. The brothers returned. "What is the matter with you?" Dawn asked him. "I have a headache from the fumes of the stove, brothers, and I have not prepared your dinner." On the third day the two elder brothers went hunting and Dawn stayed at home; he chose the best ram, slaughtered and cleaned it, and put it on to roast. Then he lay on the bench. Suddenly there was a rumbling noise - and he saw the old man as big as a thumb, with a beard a cubit long, carrying a whole hayrick on his head and holding a huge cask of water in his hand. The little old man put down the cask of water, spread the hay over the yard, and began to count his sheep. He saw that another ram was missing, grew angry, ran to the house, jumped at Dawn, and hit him on the head with all his strength; Dawn jumped up, grabbed the little old man by his long beard, and began to drag him around, repeating: "Look before you leap, look before you leap!" The old man as big as a thumb, with a beard a cubit long, began to implore him: "Have pity on me, mighty champion, do not put me to death, let my soul repent!" Dawn dragged him out into the yard, led him to an oaken pillar, and fastened his beard to the pillar with a big iron spike; then he returned to the house and sat down to wait for

his brothers. The brothers came back from their hunting and were amazed to find him safe and sound. Dawn smiled and said: "Come with me, brothers, I have caught your fumes and fastened them to a pillar." They went into the yard, they looked - but the old man as big as a thumb had long since run away. But half of his beard dangled from the pillar, and blood was spattered over his tracks. Following this clue, the brothers came to a deep hole in the ground. Dawn went to the woods, gathered lime bast, wound a rope, and told his brothers to drop him underground. Evening and Midnight dropped him into the hole. He found himself in the other world, released himself from the rope, and walked straight ahead. He walked and walked, and saw a copper castle. He entered the castle, and the youngest princess, rosier than a pink rose, whiter than white snow, came out to meet him and asked him kindly: "How have you come here, good youth - of your own will or by compulsion?" "Your father has sent me in search of you, princess." She straightway seated him at the table, gave him meat and drink, and then handed him a phial with the water of strength. "Drink of this water," she said, "and you will have added strength." Dawn drank the phial of water and felt great power in himself. "Now," he thought, "I can get the better of anyone." At this moment a wild wind arose, and the princess was frightened. "Presently," she said, "my dragon will come." And she took Dawn by his hand and hid him in the adjoining room. A three-headed dragon came flying, struck the damp earth, turned into a youth, and cried: "Oh, there is a Russian smell in here! Who is visiting you?" "Who could be here? You have been flying over Russia and you have the Russian smell in your nostrils - that is why you fancy it is here." The dragon asked for food and drink; the princess brought him a variety of meats and drink and poured a sleeping potion into the wine. The dragon ate and drank his fill and was soon overwhelmed by drowsiness; he made the princess pick the lice from his hair, lay on her knees, and fell sound asleep. The princess called Dawn. He came forth, swung his sword, and cut off all of the dragon's three heads; then he made a bonfire, burned the foul dragon and scattered his ashes in the open field. "Now farewell, princess! I am going to seek your sisters; and when I have found them I shall come back for you," said Dawn, and set out. He walked and walked, and came to a silver castle; in that castle lived the second princess. Dawn killed a six-headed dragon there and went on. After a long time or a short time, he reached a golden castle, and in that castle lived the eldest princess; Dawn killed a twelve-headed dragon and freed that princess from captivity. The princess was overjoyed, made ready to return home, went out into the wide courtyard, waved a red handkerchief, and the golden kingdom rolled up into an egg; she took the egg, put it in her pocket, and went with Dawn to seek her sisters. These princesses did the same thing: they rolled up their kingdoms into eggs, took them,

and all of them went to the hole. Evening and Midnight pulled their brother and the three princesses out into the bright world. They all came together to their own land; the princesses rolled their eggs into the open field, and straightway three kingdoms appeared, a copper, a silver and a golden one. The king was more overjoyed than any tongue can tell; he immediately married Dawn, Evening, and Midnight to his daughters, and at his death made Dawn his heir.

D

SUMMARY MATERIALS

This appendix includes the full set of summaries used in the final study of this thesis. Eight summaries were generated for each of the 15 Russian Folktales examined as described in Chapter 7. Four different generation procedures were used to create summaries of two different lengths relative to the length of the input story. Short summaries correspond to no more than 5% of the length of the input text, and long summaries correspond to no more than 10%.

D.1 NIKITA THE TANNER

D.1.1 *Short Summaries*

D.1.1.1 *Manual*

The dragon seized the princess and dragged her to his lair but did not devour her, because the princess was a beauty. Nikita, having done his heroic deed, would not accept reward, but returned to currying hides.

D.1.1.2 *Automatic*

Nikita is tested by a magical donor, a. Nikita then returns home.

D.1.1.3 *Extractive*

That furrow can be seen to this very day; it is fourteen feet high. Around it the fields are plowed, but the furrow is intact; and those who do not know what it is, call it the rampart.

D.1.1.4 *LSA*

Instead, he took her to wife. At that moment Nikita was currying hides and held twelve hides in his hands; when he saw that the tsar in person had come to see him, he began to tremble with fear, his hands shook, and he tore the twelve hides.

D.1.2 *Long Summaries*

D.1.2.1 *Manual*

The dragon seized the princess and dragged her to his lair but did not devour her, because the princess was a beauty. Nikita fought him

for a time; in any event, Nikita defeated him. Nikita, having done his heroic deed, would not accept reward, but returned to currying hides.

D.1.2.2 *Automatic*

A dragon performs a villainous act. Nikita is tested by a magical donor, a. Nikita then returns home.

D.1.2.3 *Extractive*

"Better come out into the open field," said Nikita, "or I will destroy your lair together with you!" And he began to break down the door. That furrow can be seen to this very day; it is fourteen feet high. Around it the fields are plowed, but the furrow is intact; and those who do not know what it is, call it the rampart.

D.1.2.4 *LSA*

Instead, he took her to wife. At that moment Nikita was currying hides and held twelve hides in his hands; when he saw that the tsar in person had come to see him, he began to tremble with fear, his hands shook, and he tore the twelve hides. But no matter how much the tsar and tsarina entreated him, he refused to go forth against the dragon. Nikita, having done his heroic deed, would not accept any reward, but returned to currying hides.

D.2 THE MAGIC SWAN GEESE

D.2.1 *Short Summaries*

D.2.1.1 *Manual*

Some magic geese came, seized the boy, and carried him off on their wings. His sister saw the son, crept near him, seized him, and carried him away.

D.2.1.2 *Automatic*

His sister then returns home. At last, His sister is recognized.

D.2.1.3 *Extractive*

Be careful, watch over your little brother, do not leave the house." The parents went away and the daughter forgot what they had told her she put her brother on the grass beneath the window, ran out into the street, and became absorbed in games. "Oh, in my father's house we don't even eat cream."

D.2.1.4 *LSA*

The stove did not tell her. "Apple tree, apple tree, tell me, whither have the geese flown?" "If you eat of my simple pudding with milk, I will tell you."

D.2.2 *Long Summaries*

D.2.2.1 *Manual*

Mother, the son, the daughter, and Father are introduced. Some magic geese came, seized the boy, and carried him off on their wings. The girl guessed that the magic swan geese had carried off her brother, and rushed after them. His sister saw the son, crept near him, seized him, and carried him away. The geese whirred and whirred, quacked and quacked, and finally flew away without recovering their prey.

D.2.2.2 *Automatic*

the daughter performs a villainous act. His sister lacks something important. As a result, His sister acquires the use of a magical agent. His sister then returns home. At last, His sister is recognized.

D.2.2.3 *Extractive*

Be careful, watch over your little brother, do not leave the house." The parents went away and the daughter forgot what they had told her she put her brother on the grass beneath the window, ran out into the street, and became absorbed in games. She gasped, and rushed to look in every corner, but could not find him. She ran and ran and saw a stove. "Oh, in my father's house we don't even eat cream." She ran and saw a little hut that stood on chicken legs and turned round and round.

D.2.2.4 *LSA*

"Daughter, daughter," said the mother, "we are going to work we shall bring you back a bun, sew you a dress, and buy you a kerchief. The girl guessed that they had carried off her brother, and rushed after them. The stove did not tell her. "Apple tree, apple tree, tell me, whither have the geese flown?" "If you eat of my simple pudding with milk, I will tell you." "Oh, in my father's house we don't even eat cream."

D.3 BUKHTAN BUKHTANOVICH

D.3.1 *Short Summaries*

D.3.1.1 *Manual*

Bukhtan defeats the dragon.

D.3.1.2 *Automatic*

The fox took it home, stuck one kopek behind the hoop around the measure, brought it back to the tsar, and said: "Tsar, a quart measure is not big enough; give me a peck measure measure Bukhtan Bukhtanovich's money." The fox took it home, stuck a kopek behind the hoop of the measure, and brought it back to the tsar. He said: "Have you measured all his money, fox?"

D.3.1.3 *Extractive*

"Take one!" Put them on." "Here, take these." On the hill stood a huge stone house, and around it was an enormous kingdom.

D.3.1.4 *LSA*

He lay on the stove in cockroach milk up to his elbows. "Tsar, a peck measure is not big enough; give me a bushel measure." He dressed and, accompanied by the fox, went to the tsar. "Little fox, whither shall we go?" "Cat, Son of the Cat, go into the barrel."

D.3.2 *Long Summaries*

D.3.2.1 *Manual*

A fox came to him and said: "Bukhtan Bukhtanovich, would you like me to marry you to the tsar's daughter?" And the fox sealed Cat up in the barrel. "Raven, Son of the Raven, go into the mortar!" And the fox sealed him up in the mortar; then the Fox wrapped Dragon, Son of the Dragon, in straw and took him out into the street. The ships arrived. The fox ordered the beasts to be thrown into the water; the Cossacks threw them in at once.

D.3.2.2 *Automatic*

The fox took it home, stuck one kopek behind the hoop around the measure, brought it back to the tsar, and said: "Tsar, a quart measure is not big enough; give me a peck measure measure Bukhtan Bukhtanovich's money." The fox took it home, stuck a kopek behind the hoop of the measure, and brought it back to the tsar. He said:

"Have you measured all his money, fox?" He dressed and, accompanied by the fox, went to the tsar. The fox took the clothes and came to Bukhtan Bukhtanovich. "I think Tsar is ashamed to be wearing such clothes; never in his life has he worn mean garments.

D.3.2.3 *Extractive*

"Take one!" "Take one!" Put them on." "Here, take these." "Tsar, in his house such chairs stand only in the bathhouse.' There no beer need be brewed, no wine distilled - everything is ready. On the hill stood a huge stone house, and around it was an enormous kingdom.

D.3.2.4 *LSA*

He lay on the stove in cockroach milk up to his elbows. "Tsar, a peck measure is not big enough; give me a bushel measure." "Bukhtan Bukhtanovich, have you any clothes? Bukhtan Bukhtanovich and his wife were on one of the ships, and the fox ran along the shore. The tsar is coming with fire and the tsarina with lightning, they will scorch and burn you." "Little fox, whither shall we go?' "Cat, Son of the Cat, go into the barrel." And the fox sealed him up in the barrel.

D.4 THE CRYSTAL MOUNTAIN

D.4.1 *Short Summaries*

D.4.1.1 *Manual*

The twelve-headed dragon performs a villainous act. Ivan fights the villainous the twelve-headed dragon. Ivan defeats the twelve-headed dragon.

D.4.1.2 *Automatic*

Prince Ivan is tested by a magical donor, a. When the three headed dragon comes to your herd, give him three cows when the six headed dragon comes, give him six cows and when the twelve headed dragon comes, count off twelve cows." Prince Ivan then returns home.

D.4.1.3 *Extractive*

But you won't get any!" The dragon flew into a rage, and instead of three cows, seized six. "He lives in father's lake and has a coffer in his side. Well, count off twelve cows for me!"

D.4.1.4 *LSA*

More than half of that kingdom had been swallowed into a crystal mountain. He disguised himself as a very old man, so that he would

be unrecognizable, came to the palace, and said to the king: "Your Majesty, hire me as your herdsman." "Won't that be too much" asked the prince. The dragon flew into a rage, and instead of three cows, seized six.

D.4.2 *Long Summaries*

D.4.2.1 *Manual*

The dead horse, the other sons, Ivan's father, the falcon, the birds, the reptiles, Ivan, and the beasts are introduced. More than half of kingdom had been swallowed into a mountain. The prince chased the goat Ivan and galloped, but could not catch the goat, and when he returned the princess had vanished. Ivan fights the villainous the twelve-headed dragon. Ivan defeats the twelve-headed dragon. Ivan then returns home.

D.4.2.2 *Automatic*

As a result, Prince Ivan acquires the use of a magical agent. Prince Ivan lacks something important. Prince Ivan is tested by a magical donor, a. When the three headed dragon comes to your herd, give him three cows when the six headed dragon comes, give him six cows and when the twelve headed dragon comes, count off twelve cows." Prince Ivan then returns home. At last, Prince Ivan is recognized.

D.4.2.3 *Extractive*

But you won't get any!" The dragon flew into a rage, and instead of three cows, seized six. "Good evening" said Prince Ivan. "How did you get here?" "He lives in father's lake and has a coffer in his side. In this coffer is a hare, in this hare is a duck, in this duck is an egg, and in this egg is a seed. Well, count off twelve cows for me!"

D.4.2.4 *LSA*

In a certain kingdom in a certain land there lived a king who had three sons. Thus he entered the service of that king and lived in his palace for one week, then a second, then a third. The king gave her leave. He disguised himself as a very old man, so that he would be unrecognizable, came to the palace, and said to the king: "Your Majesty, hire me as your herdsman." A goodly youth like you should be vying in combat, not tending cattle. The dragon flew into a rage, and instead of three cows, seized six. "Well, grandfather," asked the king, "has the three-headed dragon come?

D.5 SHABARSHA THE LABORER

D.5.1 *Short Summaries*

D.5.1.1 *Manual*

Shabarsha fights the villainous the imp. Shabarsha defeats the imp. there was nothing to do so the imp started with the chest and began filling Shabarsha's hat. The imp poured and poured it in, until the imp had filled it. And since then, since time, the laborer has lived in glory.

D.5.1.2 *Automatic*

grandfather lacks something important. Shabarsha performs a villainous act. As a result, grandfather acquires the use of a magical agent. grandfather overcomes the problem. grandfather then returns home.

D.5.1.3 *Extractive*

"Look over there. He's resting in that ravine under a bush." "Go up to him and hit him in the side. Then he'll get up of his own accord." The imp went into the ravine, found the bear, and whacked him in the side with his club. There's only that one chest of yours full of money."

D.5.1.4 *LSA*

What are you doing here?" he asked. "Look over there. "And where is your little brother Bunny?" The imp ran up to Bunny, touched him on the ear, and the hare jumped up with the imp right behind him. "Hmm," the old man muttered, screwing up his eyebrows. "Oho!

D.5.2 *Long Summaries*

D.5.2.1 *Manual*

Shabarsha set off to work as a laborer, and times were bad. There was no grain at all, and the vegetables didn't grow. How could One owner chase away his misery, what could he live on, where could he get money? Shabarsha departs on the quest. Shabarsha fights the villainous the imp. Shabarsha defeats the imp. The grandfather heard from his grandson that Shabarsha had thrown his club out of sight and Shabarsha got afraid and ordered them to drag the money out of the pool and buy him off. there was nothing to do so the imp started with the chest and began filling Shabarsha's hat. He poured and poured it in, until the imp had filled it. And since then, since time, the laborer has lived in glory.

D.5.2.2 *Automatic*

grandfather lacks something important. grandfather departs on the quest. Shabarsha performs a villainous act. As a result, grandfather acquires the use of a magical agent. grandfather overcomes the problem. The imp ran up to Bunny, touched him on the ear, and the hare jumped up with the imp behind him. grandfather then returns home. Shabarsha took his club and with all his might Shabarsha - just his youngest brother whacked him on the neck and whistled. The imp took the club, put it on his shoulder, and set off for Shabarsha. The devil grabbed the club and dove back to his grandfather.

D.5.2.3 *Extractive*

"And where is your Misha?" "Look over there. He's resting in that ravine under a bush." "Go up to him and hit him in the side. Then he'll get up of his own accord." The imp went into the ravine, found the bear, and whacked him in the side with his club. I've put all your money in it, but it's still empty. There's only that one chest of yours full of money." 'Carry it off to him quickly. Is he still weaving that line?" "He is."

D.5.2.4 *LSA*

What are you doing here?" he asked. He's resting in that ravine under a bush." "Hmm, go and try to have a foot race with Shabarsha; "He's over there, lying in the grass. The imp ran up to Bunny, touched him on the ear, and the hare jumped up with the imp right behind him. "Wait, wait for me, Bunny. Oh, you've got away!" "Hmm," the old man muttered, screwing up his eyebrows. "Go to Shabarsha and tell him to see who can whistle loudest." The imp lay face down on the ground and stuck his fingers in his ears. "Oho!

D.6 IVANKO THE BEAR'S SON

D.6.1 *Short Summaries*

D.6.1.1 *Manual*

the next day Father sent Ivanko to the lake and told him to wind ropes of sand. Ivanko fights the villainous the little devil. Ivanko defeats the little devil. Ivanko then returns home.

D.6.1.2 *Automatic*

a grandson who was born only yesterday lacks something important. The bear kept his wife with him, and after some time, a long time or a time, his wife had a son by him. a grandson who was born only yesterday is tested by a magical donor, a. the peasant performs a

villainous act. a grandson who was born only yesterday then returns home.

D.6.1.3 *Extractive*

Do you want to race with him?" "What grandson?" "Hey, hare, do not fail me!" The hare darted off into the open field like mad and in a trice vanished from sight; the devil rushed after him, but it was of no use; he was half a verst behind.

D.6.1.4 *LSA*

In a certain village there lived a wealthy peasant and his wife. "Go to the back yard and slaughter a sheep; "No, father, you told me to slaughter whichever one stared at me; "If this goes on for a month or two, he will ruin me completely. But on one condition - if you lag behind, I will kill you." "Grandfather said that if you can throw this crutch higher than I can, he will pay the rent."

D.6.2 *Long Summaries*

D.6.2.1 *Manual*

The bear kept Mother with him, and after some time, a long time or a time, Mother had a son by him. the next day Father sent Ivanko to the lake and told him to wind ropes of sand. In lake dwelt many devils, and the peasant hoped that the grandfather and the little devil would drag him into the water. Ivanko fights the villainous the little devil. Ivanko defeats the little devil. Ivanko overcomes the problem. Ivanko the Bearlet got a cart, loaded it with gold coins, had the devil drag it home, and said to the peasant: "Now be happy, father! Here is a laborer for you, and gold too."

D.6.2.2 *Automatic*

a grandson who was born only yesterday lacks something important. The bear kept his wife with him, and after some time, a long time or a time, his wife had a son by him. his wife had lived in the bear's den and had a son by him and that son was human to the waist and a bear from the waist down. a grandson who was born only yesterday departs on the quest. As a result, a grandson who was born only yesterday acquires the use of a magical agent. a grandson who was born only yesterday is tested by a magical donor, a. the peasant performs a villainous act. a grandson who was born only yesterday then returns home. A false hero presents unfounded claims. a grandson who was born only yesterday overcomes the problem.

D.6.2.3 *Extractive*

Do you want to race with him?" "What grandson?" "He is lying there behind a bush," said the Bearlet, and cried to a hare. "Hey, hare, do not fail me!" The hare darted off into the open field like mad and in a trice vanished from sight; the devil rushed after him, but it was of no use; he was half a verst behind. "Now, if you wish," said Ivanko, "race with me. But on one condition - if you lag behind, I will kill you." Between your legs! That way I couldn't have carried it even once!

D.6.2.4 *LSA*

In a certain village there lived a wealthy peasant and his wife. "Go to the back yard and slaughter a sheep; "No, father, you told me to slaughter whichever one stared at me; Ivanko the Bearlet broke the door off the barn, took it into the bath house, and spent the night there. Next morning the peasant arose, went to see whether everything was in order, and found that nothing was left: "You yourself told me to guard the door, and I did guard it. "If this goes on for a month or two, he will ruin me completely. Then he hit upon an idea; "Wait a while, Bearlet, I'll run and tell my grandfather," said the little devil, and - flop! But on one condition - if you lag behind, I will kill you." "Grandfather said that if you can throw this crutch higher than I can, he will pay the rent."

D.7 THE RUNAWAY SOLIDER AND THE DEVIL

D.7.1 *Short Summaries*

D.7.1.1 *Manual*

The soldier got in the cart and the devil and the soldier started the horses and so the soldier went home and how many versts flashed by his eyes! The Unclean One made the solider a merchant and gave the soldier a shop in the city with of valuable goods and the devil said. I'm going to torment the princess." The soldier thought and thought and then the soldier decided to leave his shop and go out of the town night. Meanwhile, time passed, the cocks crowed, and the Unclean One disappeared. When the Unclean One had said this, he disappeared. So the soldier married the princess and lived with the princess in love and harmony, and then after several years the king died and the soldier began ruling the tsardom.

D.7.1.2 *Automatic*

the king lacks something important. The devil performs a villainous act. the king is tested by a magical donor, a. As a result, the king

acquires the use of a magical agent. the king overcomes the problem. the king then returns home. At last, the king is recognized.

D.7.1.3 *Extractive*

Then they started playing cards. "Good, all will be ready." "Greetings! Who isn't happy with a cheerful companion!" "And why not? "This isn't much of a garden," said the princess. "Just outside the town, about thirty versts from here, there is one to fall in love with." "Say goodbye then, but hurry!"

D.7.1.4 *LSA*

He stayed with him for three days and then he asked to go home. He had no sooner spoken these words than the Unclean One was there. The other merchants became envious, however. The soldier thought: what should he do? He thought and thought and then he decided to leave his shop and go out of the town that very night. So he took all his money that was to hand and set out for the thrice-ten kingdom. He walked and walked until he came to a barrier. Who isn't happy with a cheerful companion!"

D.7.2 *Long Summaries*

D.7.2.1 *Manual*

The soldier got ready and set off along the way. "Listen, my friend: let's do a trade. I'll give you my book and you give me your fiddle." "No, man," the soldier answered, "I have to go on home and in three days I'll be, away." "Please, soldier, if you'll stay and teach me to play fiddle, I'll get you home in a single day; The soldier ate and drank his fill and then the soldier agreed to stay on and teach man how to play the fiddle. The soldier got in the cart and the devil and the soldier started the horses and so the soldier went home and how many versts flashed by his eyes! The soldier got down from the cart, went in to his relatives, and His relatives and the soldier all exchanged greetings and asked other. The Unclean One made him a merchant and gave him a shop in the city with of valuable goods and the devil said. I'm going to torment the princess." He thought and thought and then the soldier decided to leave his shop and go out of the town night. So the soldier took all his money that was to hand and set out for the ten kingdom. "I am a doctor and I'm coming to your tsardom because the daughter of your king is ill and I want to cure the princess," the soldier answered. Meanwhile, time passed, the cocks crowed, and the Unclean One disappeared. When the devil had said this, he disappeared. So the soldier married the princess and lived with the princess in love and harmony, and then after several years the king died and the soldier began ruling the tsardom.

D.7.2.2 *Automatic*

He walked and walked, but nowhere did This soldier see water and he wanted to wet his hardtack and eat a little along the way and road. Besides the hardtack the soldier had this fiddle. the king lacks something important. The devil performs a villainous act. the king is tested by a magical donor, a. the king departs on the quest. As a result, the king acquires the use of a magical agent. the king overcomes the problem. His leave had long since finished and "Oh, brother soldier was considered a runaway in his regiment. the king then returns home. The sentry reported this to the courtiers and the courtiers reported this to the king himself. At last, the king is recognized. Then the devil picked up the rods and let the devil have it! But the devil kept on flogging the devil. So the soldier married the princess and lived with the princess in love and harmony, and then after several years the king died and the king began ruling the tsardom.

D.7.2.3 *Extractive*

"Good, all will be ready." "Greetings." "Certainly." Then they started playing cards. "How did you sleep last night?" "Peacefully, thanks be to God!" "Your majesty, order them to make some pincers weighing fifty pounds and three brass rods, three iron ones, and three lead ones." "Good, all will be ready." "Greetings! Who isn't happy with a cheerful companion!" "And why not? I'll never come within thirty versts of the palace again!" "This isn't much of a garden," said the princess. "Just outside the town, about thirty versts from here, there is one to fall in love with." "Now you'll never get out of my claws." "Say goodbye then, but hurry!"

D.7.2.4 *LSA*

"Oh, old man, why do I need your book? "No, old man," the soldier answered, "I have to go on home and in three days I'll be far, far away." "Please, soldier, if you'll stay and teach me to play this fiddle, I'll get you home in a single day; The soldier ate and drank his fill and then he agreed to stay on and teach this unknown old man how to play the fiddle. He stayed with him for three days and then he asked to go home. He had no sooner spoken these words than the Unclean One was there. The other merchants became envious, however. The soldier thought: what should he do? How should he answer them? He thought and thought and then he decided to leave his shop and go out of the town that very night. So he took all his money that was to hand and set out for the thrice-ten kingdom. He walked and walked until he came to a barrier. Who isn't happy with a cheerful companion!" They started drinking and carousing. "Well, you see, the king has taken me into his service and ordered me to teach some musicians to play the fiddle. But all their fingers are crooked, no

better than yours, and so I've got to straighten them out in these pincers."

D.8 FROLKA STAY AT HOME

D.8.1 *Short Summaries*

D.8.1.1 *Manual*

The twelve-headed dragon, the first daughter, the second daughter, the third daughter, and the king are introduced. suddenly the dragon appeared and carried the first daughter, the second daughter and the third daughter off on his fiery wings. Frolka fights the villainous the twelve-headed dragon. Frolka defeats the twelve-headed dragon. Erema, the soldier, the first daughter, the second daughter, the third daughter and Frolka all came back to their land.

D.8.1.2 *Automatic*

suddenly the dragon appeared and carried them off on his fiery wings. the lions is tested by a magical donor, a. the lions lacks something important. the lions then returns home. At last, the lions is recognized.

D.8.1.3 *Extractive*

Do I have enemies in the world? I have only one enemy, but his bones won't even be brought here by a raven." "There is nothing to be done," said Frolka. "Probably all of the royal treasury will fall to me." "And what will be left for us?" asked his companions. having done this, they went home.

D.8.1.4 *LSA*

Their garden was big and beautiful and they liked to walk there at night. "Little doves, why have you come here? He had no sooner said these words than the dragon came flying and roared: Do I have enemies in the world? "I smell Russian breath here," he said. Then they entered the house;

D.8.2 *Long Summaries*

D.8.2.1 *Manual*

The twelve-headed dragon, the first daughter, the second daughter, the third daughter, and the king are introduced. suddenly the dragon appeared and carried the first daughter, the second daughter and the third daughter off on his fiery wings. The king said: "Whoever finds

my daughters, to him I shall give as much money as he wants." a soldier, Frolka Stay at Home, and Erema; and they set to look for the princesses. Frolka fights the villainous twelve-headed dragon. Frolka defeats the twelve-headed dragon. Erema, the soldier, the first daughter, the second daughter, the third daughter and Frolka all came back to their land. The king was overjoyed, opened his royal treasury to them, and said:

D.8.2.2 *Automatic*

suddenly the dragon appeared and carried them off on his fiery wings. the lions departs on the quest. the lions is tested by a magical donor, a. the lions lacks something important. the lions overcomes the problem. She said that She was the king's daughter; the lions then returns home. A false hero presents unfounded claims. At last, the lions is recognized. The man went up to the lions and began to stroke them, and Frolka with his companions got through to the courtyard. the man brought them to the room where the princess lived.

D.8.2.3 *Extractive*

He sent his maidservants to look for them in the garden, but all in vain; "Who has ruined my kingdom? Do I have enemies in the world? I have only one enemy, but his bones won't even be brought here by a raven." "A raven won't bring me," said Frolka, "but a good horse did." "Have you come for peaceful purposes or to fight?" Frolka was generous: "There is nothing to be done," said Frolka. "Probably all of the royal treasury will fall to me." "And what will be left for us?" asked his companions. While there was still money, Erema began to fill his basket, and the soldier his knapsack; having done this, they went home.

D.8.2.4 *LSA*

There was once a king who had three daughters, and such beauties they were as no tongue can tell of nor pen describe. Their garden was big and beautiful and they liked to walk there at night. suddenly the dragon appeared and carried them off on his fiery wings. The king said: "Whoever finds my daughters, to him I shall give as much money as he wants." "Little doves, why have you come here? He had no sooner said these words than the dragon came flying and roared: "Who has ruined my kingdom? Do I have enemies in the world? finally they found it. "I smell Russian breath here," he said. Then they entered the house; they passed through one room, a second, and a third, and in the fourth they found the king's second daughter sitting on a sofa.

D.9 THE WITCH

D.9.1 *Short Summaries*

D.9.1.1 *Manual*

Mother, Ivashko, and Father are introduced. Now a witch had heard what Ivashko's parents had cried to him, and the witch longed to get hold of the boy. The witch deceives the victim. But the moment Alenka sat down on the shovel, Ivashko pitched Alenka into the oven, slammed to the plate in front of it, ran out of the hut, shut the door, and climbed up ever so high an oak tree (which stood close).

D.9.1.2 *Automatic*

As a result, Ivashko acquires the use of a magical agent. Ivashko is tested by a magical donor, a. Ivashko lacks something important. she performs a villainous act. Ivashko then returns home. At last, Ivashko is recognized.

D.9.1.3 *Extractive*

"What are you thinking about! you're still very small; suppose you get drowned, what good will there be in that?" So his mother put a white shirt on him, tied a red girdle round him, and let him go. Ivashko came, and she took the fish, and seized the boy and carried him home with her. "Ah! There's for me."

D.9.1.4 *LSA*

I'll catch you some fish; I bring thee food and drink. Canoe, canoe, float a little farther. Ivashko perceived that the voice was not his mother's, but was that of a witch, and he sang: They all sat down to table, and the witch opened the oven, took out Alenka's baked body, and served it up. "No it was only the noise of the leaves." The witch returned and began gnawing the oak again.

D.9.2 *Long Summaries*

D.9.2.1 *Manual*

Mother, Ivashko, and Father are introduced. Now a witch had heard what Ivashko's parents had cried to Ivashko, and the witch longed to get hold of the boy. Ivashechko, Ivashechko, my boy, Float up, float up, onto the waterside I bring thee food and drink. The witch saw that the witch must call Ivashko with a voice as his mother had. So the witch hastened to a smith and said to him: "Smith, make me such a thin little voice as Ivashko's mother has: if you don't, I'll eat you." So the smith forged her a little voice like Ivashko's mother's.

Then the witch went down by to the shore and sang: Ivashechko, Ivashechko, my boy, Float up, float up, unto the waterside; I bring food and drink. But the moment Alenka sat down on the shovel, Ivashko pitched Alenka into the oven, slammed to the plate in front of it, ran out of the hut, shut the door, and climbed up ever so high an oak tree(which stood close). Oh, my swans and geese, Take me on your pinions, Bear me to my father and my mother, To the cottage of my father and my mother.

D.9.2.2 *Automatic*

As a result, Ivashko acquires the use of a magical agent. Ivashko is tested by a magical donor, a. Canoe, canoe, float to the waterside; Ivashko lacks something important. she performs a villainous act. Canoe, canoe, float a farther, Canoe, canoe, float a little farther; Ivashko overcomes the problem. They all ate their fill and drank their fill, and then they all went out into the courtyard and began rolling about on the grass. Ivashko then returns home. she pursues Ivashko. At last, Ivashko is recognized.

D.9.2.3 *Extractive*

"What are you thinking about! you're still very small; suppose you get drowned, what good will there be in that?" So his mother put a white shirt on him, tied a red girdle round him, and let him go. Ivashko came, and she took the fish, and seized the boy and carried him home with her. "Very good," said Alenka; "Ah! "No doubt she's gone off somewhere to amuse herself." There to eat, and drink, and live in comfort. There to eat, and drink, and live in comfort. Oh, my swans and geese! There to eat, and drink, and live in comfort. There's for me."

D.9.2.4 *LSA*

And after that he and they lived on happily together. I'll catch you some fish; do let me go!" I bring thee food and drink. the woman took the fish, gave her boy food and drink, changed his shirt for him and his girdle, and sent him back to his fishing. Canoe, canoe, float a little farther. Ivashko perceived that the voice was not his mother's, but was that of a witch, and he sang: "Very good," said Alenka; "No doubt she's gone off somewhere to amuse herself." Then she slipped in through the window, opened the door, and let in her guests. They all sat down to table, and the witch opened the oven, took out Alenka's baked body, and served it up. Then she ran to a forge, and when she reached it she cried, "Smith, smith! if you don't I'll eat you!"

D.10 THE SEVEN SIMEONS

D.10.1 *Short Summaries*

D.10.1.1 *Manual*

Lookout Simeon, Retriever Simeon, Thief Simeon, Artisan Simeon, the Czar, Healer Simeon, Marksman Simeon, the father, and Builder Simeon are introduced. As Healer Simeon, Marksman Simeon, Artisan Simeon, Builder Simeon, Lookout Simeon, Thief Simeon, Retriever Simeon, slam bam, built a boat, boarded and sailed off to the kingdom of the bride to be princess. With their healthy and successful return to their own kingdom, the Simeons fired a round from the cannon.

D.10.1.2 *Automatic*

After long period passed, the czar decided to marry a princess. The maker of valuables went with his brother, the thief, to sell his wares in the palace. When the princess realized the maker of valuables were carrying that princess away, that princess turned into a swan and flew from the boat.

D.10.1.3 *Extractive*

He thought to himself, "What's going on with that boat out there?" "Please go," he said. "Run and find out what's going on." He was already too old!

D.10.1.4 *LSA*

"I'm very good at building boats; When the princess realized they were carrying her away, she turned into a white swan and flew from the boat. The marksman, without delay, grabbed his gun and fired a shot into her left wing. The swan turned back into a princess, but her left arm was still injured.

D.10.2 *Long Summaries*

D.10.2.1 *Manual*

Lookout Simeon, Retriever Simeon, Thief Simeon, Artisan Simeon, the Czar, Healer Simeon, Marksman Simeon, the father, and Builder Simeon are introduced. After long period passed, the czar decided to marry a princess. Summoning them, the Czar commissioned them as soldiers, with the assignment to bring him princess. As Healer Simeon, Marksman Simeon, Artisan Simeon, Builder Simeon, Lookout Simeon, Thief Simeon, Retriever Simeon, slam bam, built a boat, boarded and sailed off to the kingdom of the bride to be princess.

They barely arrived when the stole the princess. With their healthy and successful return to their own kingdom, the Simeons fired a round from the cannon.

D.10.2.2 *Automatic*

After long period passed, the czar decided to marry a princess. As they, slam bam, built a boat, boarded and sailed off to the kingdom of the bride to be princess. The maker of valuables went with his brother, the thief, to sell his wares in the palace. When the princess realized the maker of valuables were carrying that princess away, that princess turned into a swan and flew from the boat. the princess didn't want marry the czar. Therefore, the czar asked Only the princess who Only the princess wanted to marry.

D.10.2.3 *Extractive*

But, how was he to abduct her? He just didn't know; there was no one suitable for the job. The marksman, without delay, grabbed his gun and fired a shot into her left wing. He thought to himself, "What's going on with that boat out there?" "Please go," he said. "Run and find out what's going on." He was already too old!

D.10.2.4 *LSA*

The czar was grateful to the father for bringing him so many fine, able men. "I can shoot any bird - even in flight, your royal majesty." "I'm very good at building boats; The Simeons quickly convened. The marksman, without delay, grabbed his gun and fired a shot into her left wing. The swan turned back into a princess, but her left arm was still injured. Healer Simeon immediately cured the princess' arm.

D.11 IVAN POPYALOV

D.11.1 *Short Summaries*

D.11.1.1 *Manual*

Now in the land in which Ivan lived there was never day, but night. Ivan Popyalov fights the villainous the Snake. Ivan Popyalov defeats the Snake. After killing the Snake, Ivan Popyalof and his brothers set off on their home.

D.11.1.2 *Automatic*

Once upon a time there was an couple, and an old couple had. Now in the land in which Ivan lived there was never day, but night. they then returns home. Then the brothers said, "Let's turn out our horses to graze here, while we rest ourselves a little."

D.11.1.3 *Extractive*

"Wherefore hast thou stumbled, O Steed! hast thou howled, O Hound! hast thou clamored, O Falcon?" Ivan came forth, and they began to fight. And the elder brothers cried, "Let's have a drink of water." But they replied:

D.11.1.4 *LSA*

Then Ivan hung up his gloves, and said to his brothers, "Should blood drop from my gloves, make haste to help me." His steed stumbled, his hound howled, his falcon clamored. And Ivan killed the Snake, and then sat down again beneath the boarding. presently they saw before them a green meadow, and on that meadow lay silken cushions. Then he perceived that danger was at hand, and so he let his horse go free, and hid himself behind twelve doors in the forge of Kuzma and Demian.

D.11.2 *Long Summaries*

D.11.2.1 *Manual*

Ivan Popyalov, Mother, the brothers, and Father are introduced. Now in the land in which Ivan lived there was never day, but night. Ivan undertook to kill that Snake, so Ivan Popyalov said to his father, "Father make me a mace five poods in weight." Ivan went home and told his father to make him a third mace, one of weight. Ivan Popyalov fights the villainous the Snake. Ivan Popyalov defeats the Snake. After killing the Snake, Ivan Popyalof and his brothers set off on their home. Then Ivan Popyalov perceived that danger was at hand, and so he let his horse go free, and hid himself behind twelve doors in the forge of Kuzma and Demian.

D.11.2.2 *Automatic*

Once upon a time there was an couple, and an old couple had three sons. Now in the land in which Ivan lived there was never day, but night. And when Ivan had got the mace, he went out into the fields, and flung it straight up in the air, and then he went Ivan. As a result, they acquires the use of a magical agent. they then returns home. As soon as Ivan Popyalof came to where his brothers were, he mounted his horse, and they all started afresh. Then the brothers said, "Let's turn out our horses to graze here, while we rest ourselves a little." They rode and rode;

D.11.2.3 *Extractive*

Then cried the Snake: "Wherefore hast thou stumbled, O Steed! hast thou howled, O Hound! hast thou clamored, O Falcon?" "How can I

but stumble," replied the Steed. Ivan came forth, and they began to fight. And the elder brothers cried, "Let's have a drink of water." But they replied: "Send your tongue through the twelve doors and take him." So the Snake's Wife began licking the doors.

D.11.2.4 *LSA*

Two of these had their wits about them, but the third was a simpleton, Ivan by name, surnamed Popyalof. On the fourth day Ivan went out to the same spot, and when the mace came tumbling down, he put his knee in the way, and the mace broke over it into three pieces. Then Ivan hung up his gloves, and said to his brothers, "Should blood drop from my gloves, make haste to help me." His steed stumbled, his hound howled, his falcon clamored. And Ivan killed the Snake, and then sat down again beneath the boarding. The Snake had no strength left in him. But Ivan cried: "Fly, and tell my brothers to come, and then we will kill this Snake, and give his flesh to thee." presently they saw before them a green meadow, and on that meadow lay silken cushions. Then he perceived that danger was at hand, and so he let his horse go free, and hid himself behind twelve doors in the forge of Kuzma and Demian.

D.12 THE SERPENT AND THE GYPSY

D.12.1 *Short Summaries*

D.12.1.1 *Manual*

The serpent had devoured all but one of the villagers. So, the last man and the Gypsy spent the night. The Gypsy fights the villainous the serpent. the serpent got his wagon, harnessed three of the stallions, and the Gypsy and the serpent left for the camp. The serpent jumped out of the wagon and ran away. The Gypsy sold the trio of horses with the wagon and began to live the life.

D.12.1.2 *Automatic*

He had devoured but one of the villagers. As a result, they acquires the use of a magical agent. There was a chunk of curd on the table, the gypsy grabbed it and gave it a squeeze. they then returns home. "Brother, what about the animals I tied? The serpent jumped out of the wagon and ran away.

D.12.1.3 *Extractive*

Don't whistle anymore! "Agreed!" Pick out the fattest, grab it by the tail and drag it back for lunch." One is enough." "Forget them." Choose a dry oak and pull it to the hut. It's time to start a fire!"

D.12.1.4 *LSA*

Now I'll have something for lunch as well!" I think even you realize I have greater strength than you." Let's see who can whistle the loudest." He started to tie them together by their tails. The serpent dropped the hide into the well, filled it with water, dragged it out and carried it home. "Brother," he said to the gypsy, "go into the forest. "If you want to make peace with me, then visit me as my guest."

D.12.2 *Long Summaries*

D.12.2.1 *Manual*

The serpent had devoured all but one of the villagers. "You see, the serpent has been coming here devouring people. So, the last man and the Gypsy spent the night. "Do you think you can eat us" asked the gypsy. "So, you think you are stronger than me?" I think even you realize have strength than you." "Well then, let's see who's stronger between the two of us." The Gypsy fights the villainous the serpent. And so the serpent scarfed a ox, drank a oxen skin of water and then questioned the gypsy," Tell me, brother, why are you upset?" away, the serpent got his wagon, harnessed three of the stallions, and the Gypsy and the serpent left for the camp. The serpent jumped out of the wagon and ran away. The gypsy sold the trio of horses with the wagon and began to live the life.

D.12.2.2 *Automatic*

He had devoured but one of the villagers. they departs on the quest. As a result, they acquires the use of a magical agent. There was a chunk of curd on the table, the gypsy grabbed it and gave it a squeeze. The serpent whistled so loudly that leaves fell from the trees. they then returns home. the gypsy is punished. "Brother, what about the animals I tied? "Here, take ox hide," said the serpent to the gypsy, "if you please, fill this up with water and carry it back here. The gypsy grabbed the hide and dragged it to the well. The serpent cooked the beef, and called the gypsy to lunch. The serpent jumped out of the wagon and ran away.

D.12.2.3 *Extractive*

So who's stronger?" Let's see who can whistle the loudest." "Okay, go ahead and whistle." Don't whistle anymore! "Agreed!" Pick out the fattest, grab it by the tail and drag it back for lunch." "What is taking so long?" "Just wait a little longer. One is enough." "Forget them." To dig out the well would take a long time." Choose a dry oak and pull it to the hut. It's time to start a fire!"

D.12.2.4 *LSA*

So, they spent the night. Now I'll have something for lunch as well!" I think even you realize I have greater strength than you." "Fine I'll watch." Let's see who can whistle the loudest." Pick out the fattest, grab it by the tail and drag it back for lunch." He saw a large herd of oxen grazing. He started to tie them together by their tails. The serpent dropped the hide into the well, filled it with water, dragged it out and carried it home. "Brother," he said to the gypsy, "go into the forest. He braided an extremely long rope and set to wrapping up the oak trees. "If you want to make peace with me, then visit me as my guest." They were just approaching, when the little gypsy children saw their father.

D.13 PRINCE DANILA GOVORILA

D.13.1 *Short Summaries*

D.13.1.1 *Manual*

The witch pondered and pondered as to how she could lead Prince Danila and Princess Catherine into evil ways and destroy them. Princess Catherine is rescued from the pursuit. the witch tried to shove Princess Catherine in, but Princess Catherine put one leg into the stove and the other under it. They plastered and tarred the opening, and then ran away, taking the embroidered towel and a brush and comb with them. The brother fell, the sister rushed to embrace him, and Princess Catherine cried and lamented: The brother jumped up, safe and sound, embraced his sister, and married Princess Catherine to a man; and Prince Danila himself married her friend, on whose finger the ring fitted, and all of them lived forever after.

D.13.1.2 *Automatic*

The woman believed an old princess and was overjoyed; before her death an old princess enjoined upon her son that your son take to wife a woman whom the ring would be found to fit. Sister is tested by a magical donor, a. young woman performs a villainous act. Sister then returns home. "One of them," said the servant, "must be your sister, but which of the two that is, it is impossible to guess." The master went to see them and invited them to his home. the servant struck his master in the side and blood gushed.

D.13.1.3 *Extractive*

"But listen to us. Farewell!". "Earth, open wide! "Why did you not keep them here?" What could be done? "This is what can be done,

master," said the servant. "Very well !" the servant struck his master in the side and blood gushed forth.

D.13.1.4 *LSA*

In the end she conceived a plan. Like a cunning fox she came to their mother and said: "What are you saying, my brother? Farewell!". "Sister Catherine, come to the featherbed." The brother called her again: "My little dove, my heart is glad to see you, I will welcome you and fondle you while my mother is out. But when she comes back there will be trouble for both of us, for she is a witch."

D.13.2 *Long Summaries*

D.13.2.1 *Manual*

There was once an princess; she had a son and a daughter, both well built, handsome. The witch pondered and pondered as to how she could lead Prince Danila and Princess Catherine into evil ways and destroy them. The woman believed the Witch and was overjoyed; before her death the old Princess enjoined upon her son that Prince Danila take to wife a woman whom the ring would be found to fit. Baba Yaga pursues Princess Catherine. Princess Catherine is rescued from the pursuit. The sister walked and walked underground and saw a hut on legs, turning round and round. the witch tried to shove Princess Catherine in, but Princess Catherine put one leg into the stove and the other under it. They plastered and tarred the opening, and then ran away, taking the embroidered towel and a brush and comb with them. The brother fell, the sister rushed to embrace him, and Princess Catherine cried and lamented: The brother jumped up, safe and sound, embraced his sister, and married Princess Catherine to a man; and Prince Danila himself married her friend, on whose finger the ring fitted, and all of them lived forever after.

D.13.2.2 *Automatic*

There was once an princess; an old princess had a son and a daughter, both well built, handsome. The woman believed an old princess and was overjoyed; before her death an old princess enjoined upon her son that your son take to wife a woman whom the ring would be found to fit. He traveled and traveled through villages cities, tried the ring on all the maidens, but could not find one whom your son could take as his betrothed; As a result, Sister acquires the use of a magical agent. Sister is tested by a magical donor, a. young woman performs a villainous act. In the fight, Sister is branded by young woman. They brought wood, oak and maple, and made a fire; the flame blazed from the stove. The witch shoved the witch toward the mouth of the stove, but the maiden put one leg into the stove and the

on top of it. the witch tried to shove the witch in, but the witch put one leg into the stove and the other under it. Sister then returns home. Where could the two maidens go? "One of them," said the servant, "must be your sister, but which of the two that is, it is impossible to guess." The master went to see them and invited them to his home. the servant struck his master in the side and blood gushed.

D.13.2.3 *Extractive*

one does not marry one's own sister." "But listen to us. Farewell!". Only the dolls were sitting in the corners and crying: "Earth, open wide! "Madam mother, passers-by came in to drink some water." "Why did you not keep them here?" "Henceforth, mind you, invite all into the house, do not let anyone go; "Now, my beauty, sit on the shovel." "Sit quietly, this way - just see how I do it." his servant had told the truth. What could be done? "This is what can be done, master," said the servant. "Very well !" the servant struck his master in the side and blood gushed forth. "My beloved, my dearest!"

D.13.2.4 *LSA*

In the end she conceived a plan. Like a cunning fox she came to their mother and said: He grew up and began to seek a bride; "What are you saying, my brother? "Weep not, grieve not," said the old women. Farewell!". "Sister Catherine, come to the featherbed." The brother called her again: "My little dove, my heart is glad to see you, I will welcome you and fondle you while my mother is out. But when she comes back there will be trouble for both of us, for she is a witch." I will leave now to get some booty." "My good daughter, my comely daughter, I smell a Russian bone!" she said. The guest had been sitting in the broom. They had hardly had time to exchange a few whispers, when the witch (talk of the devil and he will appear) stood in the doorway, catching them by surprise. The witch took a broad shovel and began to urge her guest: "Now, my beauty, sit on the shovel."

D.14 THE MERCHANT'S DAUGHTER AND THE MAIDSERVANT

D.14.1 *Short Summaries*

D.14.1.1 *Manual*

there the maidservant gave the merchant's daughter sleeping potions, cut out her eyes, and put them in her pocket. The old man took the daughter in. She lay down to sleep in the hut, but upon awakening the daughter found herself in a glass house and began to live in style. There the king saw the maiden and was overjoyed. She said to the king: "I am your bride, the merchant's daughter, and your queen is

my maidservant." The maidservant's eyes were cut out, the maidservant was tied to a horse, and dragged to the maidservant death over the field.

D.14.1.2 *Automatic*

There was once a very merchant who had a beautiful daughter. his queen is tested by a magical donor, a. At last, his queen is recognized. his queen then returns home.

D.14.1.3 *Extractive*

The letter said: "Make ready to get married." Here is a chest of money; you will never reach its bottom, it will always be full. Arise in the morning, make a coffin, gather my remains, and bury them." "Surely some lord went for a drive and lost him here," he thought. he ceased crying and began to skip around the rooms. "Forgive me," she said.

D.14.1.4 *LSA*

The merchant answered: "I have a beautiful daughter, and she is so clever that no matter what a man is thinking, she can guess it." "Go to that merchant's house," he told them, "and deliver this letter to the merchant's daughter." She sent the old man to a shop, saying: "Get velvet and silk on credit." And fear not, no matter what they do to you." "Give me an eye," said the old man. The queen guessed who it was and thought to herself: "It must be the same maiden whose eyes I cut out." Now I am going to make a visit.

D.14.2 *Long Summaries*

D.14.2.1 *Manual*

There was once a very merchant who had a beautiful daughter. The merchant answered: "I have a beautiful daughter, and the daughter is so clever that no matter what a man is thinking, she can guess it." and no one could distinguish maid from the merchant's daughter, the maidservant and the daughter were so like other. there the maidservant gave the merchant's daughter sleeping potions, cut out her eyes, and put them in her pocket. The old man took the daughter in. The daughter overcomes the problem. She lay down to sleep in the hut, but upon awakening the daughter found herself in a glass house and began to live in style. They cut the daughter in pieces, took out her heart, buried her in the ground, and returned to the palace. They came to the garden and began to cut it but it turned into stone. There the king saw the maiden and was overjoyed. She said to him: "I am

your bride, the merchant's daughter, and your queen is my maidservant." The maidservant's eyes were cut out, the maidservant was tied to a horse, and dragged to the maidservant death over the field. And the king began to live with the and to prosper.

D.14.2.2 *Automatic*

There was once a very merchant who had a beautiful daughter. his queen is tested by a magical donor, a. her performs a villainous act. his queen lacks something important. She could not see; As a result, his queen acquires the use of a magical agent. She took it from him, went at twilight, spat upon it, put it into its socket, and was able to see more. She went outside at twilight, spat on the eye, put it in its socket, and could see with eyes. In the fight, his queen is branded by her. He feasted there and upon leaving asked the maiden to come to see him. his queen then returns home. They cut the merchant's daughter in pieces, took out her heart, buried her in the ground, and returned to the palace. There the king saw the maiden and was overjoyed.

D.14.2.3 *Extractive*

The letter said: "Make ready to get married." Why is she so ignorant? She does not know how to do anything." "I live in a little hut." Here is a chest of money; you will never reach its bottom, it will always be full. You will go to sleep in this glass house, but you will awaken in your old hut. Now I am going to make a visit. I shall not be alive tomorrow; Arise in the morning, make a coffin, gather my remains, and bury them." "Surely some lord went for a drive and lost him here," he thought. he ceased crying and began to skip around the rooms. "Forgive me," she said.

D.14.2.4 *LSA*

The king said to him: "Why can I not find a bride for myself?" The merchant answered: "I have a beautiful daughter, and she is so clever that no matter what a man is thinking, she can guess it." "Go to that merchant's house," he told them, "and deliver this letter to the merchant's daughter." The old man took her in. She said to the old man: "Little grandfather, drive out the cattle." Next morning the blind maiden roused the old man and said: "Go and take this to the king, and for payment accept only an eye. And fear not, no matter what they do to you." The king said: "I will buy it from you at any price." The king was overjoyed and thanked her, saying: "Ah, little mother, you have been a great help to me!" He returned to his palace and said to his queen: "Ah, little mother, what a house there is in such and such a place! Now I am going to make a visit. The old man wept for her. she wanted to shoot her on the spot.

D.15 DAWN, EVENING, AND MIDNIGHT

D.15.1 *Short Summaries*

D.15.1.1 *Manual*

Dawn defeats the Three-Headed dragon. Dawn overcomes the problem.

D.15.1.2 *Automatic*

In a kingdom there was a king who had three daughters of surpassing beauty. when a whirlwind seized them and carried them off far and high, no one knew whither. They knocked at the window and there was no answer; He snatched up a crust of bread and began to beat Evening on the head with it; She seated him at the table, gave him meat and drink, and then handed him a phial with the water of. They all came to their own land;

D.15.1.3 *Extractive*

The king tried to dissuade them but to no avail. And so the beautiful princesses went out to walk in the garden. I shall take a spoonful of cabbage soup and a crumb of bread and throw them in your eyes!" "Presently," she said, "my dragon will come." Who is visiting you?" "Who could be here? The king was more overjoyed than any tongue can tell;

D.15.1.4 *LSA*

In a certain kingdom there was a king who had three daughters of surpassing beauty. The king burst into tears. Evening answered: "First grow up - otherwise you cannot be seen from the ground! I shall take a spoonful of cabbage soup and a crumb of bread and throw them in your eyes!" Next day Dawn and Evening went hunting, and Midnight was left at home to prepare the dinner. Midnight made a fire, chose the fattest ram, slaughtered it, and put it in the oven; then he lay on the bench. The little old man put down the cask of water, spread the hay over the yard, and began to count his sheep.

D.15.2 *Long Summaries*

D.15.2.1 *Manual*

The eldest Princess, the youngest Princess, the King, and the middle Princess are introduced. when a whirlwind seized them and carried them off far and high, no one knew whither. Dawn came, swung his sword, and cut off all of the dragon's three heads; then Dawn made a bonfire, burned the foul dragon and scattered his ashes in the open

field. “Now farewell, princess! I am going to seek your sisters; and when I have found them I shall come back for you,” said Dawn, and set out. Dawn walked and walked, and came to a castle; in that castle lived the second princess. Dawn killed a six headed dragon there went on. After a long time or a time, Dawn reached a castle, and in that castle lived the eldest princess; Dawn killed a twelve headed dragon and freed that from captivity. They all came to their own land;

D.15.2.2 *Automatic*

In a kingdom there was a king who had three daughters of surpassing beauty. when a whirlwind seized them and carried them off far and high, no one knew whither. They knocked at the window and there was no answer; morning Dawn, the brother, said to Evening, his eldest brother: “We two shall go hunting, and you stay at home and prepare our dinner.” He snatched up a crust of bread and began to beat Evening on the head with it; Suddenly there was a rumbling noise, and the man as as a thumb, with a beard a cubit, came in and began to beat and thrash him; Suddenly there was a rumbling noise and he the old man and he the old as big as a thumb, with a beard a cubit long, carrying a whole on his head and holding a huge of water in his hand. The man as as a thumb, with a beard a cubit long, began to implore him: “Have pity on, champion, do not put me to death, let my soul repent!” They went into the yard, your fumes looked but the old as big a thumb had long since run away. She seated him at the table, gave him meat and drink, and then handed him a phial with the water of. After a long time or a time, A three-headed dragon reached a castle, and in that castle lived the eldest princess; Dawn killed a twelve headed dragon and freed that from captivity. They all came to their own land;

D.15.2.3 *Extractive*

The king tried to dissuade them but to no avail. And so the beautiful princesses went out to walk in the garden. Beyond that steppe was a thick forest, and close to the forest stood a little hut. Near the hut there was a shed full of sheep; He prepared everything and lay down to rest on a bench. I shall take a spoonful of cabbage soup and a crumb of bread and throw them in your eyes!” At this moment a wild wind arose, and the princess was frightened. “Presently,” she said, “my dragon will come.” A three-headed dragon came flying, struck the damp earth, turned into a youth, and cried: “Oh, there is a Russian smell in here! Who is visiting you?” “Who could be here? You have been flying over Russia and you have the Russian smell in your nostrils - that is why you fancy it is here.” The king was more overjoyed than any tongue can tell;

D.15.2.4 *LSA*

In a certain kingdom there was a king who had three daughters of surpassing beauty. The king called his grand council together and asked his councilors and boyars whether anyone among them would undertake to search for his daughters. The king asked once and the boyars were silent; he asked a third time and no one made a sound! The king burst into tears. He hoped that someone from among the common people would undertake the heavy task. The king took the old woman into his palace, and ordered that she be given food and drink from his table and clothes and shoes from his stores. Evening answered: "First grow up - otherwise you cannot be seen from the ground! I shall take a spoonful of cabbage soup and a crumb of bread and throw them in your eyes!" Next day Dawn and Evening went hunting, and Midnight was left at home to prepare the dinner. Midnight made a fire, chose the fattest ram, slaughtered it, and put it in the oven; then he lay on the bench. The little old man put down the cask of water, spread the hay over the yard, and began to count his sheep. But half of his beard dangled from the pillar, and blood was spattered over his tracks.