FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

COMMUNICATING WITH CULTURE: HOW HUMANS AND MACHINES DETECT NARRATIVE ELEMENTS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Wolfgang Victor Hayden Yarlott

2022

To: John L. Volakis
College of Engineering and Computing

This dissertation, written by Wolfgang Victor Hayden Yarlott, and entitled Communicating with Culture: How Humans and Machines Detect Narrative Elements, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Phillip Carter

Giri Narasimhan

Monique Ross

Ning Xie

Mark A. Finlayson, Major Professor

Date of Defense: March 31, 2022

The dissertation of Wolfgang Victor Hayden Yarlott is approved.

John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

To my father, mother, and brother, for putting up with me for all these years.

ACKNOWLEDGMENTS

This work was made possible by FIU's Presidential Fellowship, FIU's KFSCIS Director's Fellowship, and DARPA via SBIR Phase II Prime contract FA8650-19-C-6017. This work was greatly aided by my collaborators at SIFT, the partner for the SBIR Phase II.

Many thanks to my advisor, Dr. Mark A. Finlayson, for encouraging me to explore avenues I otherwise would have missed and for his support and advice. I would also like to thank my committee members, Dr. Phillip Carter, Dr. Giri Narasimhan, Dr. Monique Ross, and Dr. Ning Xie, for their time and advice.

Additional thanks to all my fellow members of the Cognac Lab, including Josh, Deya, Labiba, Anurag (whose assistance on the ACUMEN project was invaluable), Mireya, Mohammed, Mustafa, and Samira, for their friendship, advice, and assistance over all of these years. I am also grateful to Diana, my undergraduate mentee, and Diego, another Cognac undergraduate mentee, who helped immensely with this work.

Thanks to annotators: David Roberson, Aeron Commins, and adjudicator Jacob Stulberg, who worked on Propp's morphology; and annotators: James Conlon, Orpaz Levy, Natasha Maldonado, Sivan Manoah, Robert McKendry, and Jean Mendez, who worked on motifs.

Special thanks to friends afar: Kyle and Tim, as well as internet friends, veelow, feenix, and seal. Thanks and apologies to friends and family unmentioned, either for brevity or from a slip of the mind.

To everyone else—go to hell.

ABSTRACT OF THE DISSERTATION

COMMUNICATING WITH CULTURE: HOW HUMANS AND MACHINES DETECT NARRATIVE ELEMENTS

by

Wolfgang Victor Hayden Yarlott

Florida International University, 2022

Miami, Florida

Professor Mark A. Finlayson, Major Professor

To understand how people communicate, we must understand how they leverage shared stories and all the knowledge, information, and associations contained within those stories. I examine four classes of narrative elements that convey a wealth of cultural knowledge: Propp's morphology, motifs, discourse structure, and motif-like elements.

My thesis has three aims: first, to demonstrate that people can reliably detect and identify all three of these narrative elements; second, to develop automatic detectors for discourse and motifs; third, to demonstrate the deep relation between these narrative elements and other theories of narrative structure and knowledge representation that I refer to as the continuum of cultural communication.

The first step of my work answers two key questions about Propp's morphology by demonstrating the reliability of annotators applying Propp's scheme across a variety of experiments, in a double-blind annotation study and demonstrates a shortcoming in Propp's scheme.

The second step of my work, showing that people familiar with motifs can reliably detect when they are being used to share information and associations, approaches this problem by performing a large-scale annotation study of 21,000 examples into four categories performed by three pairs of annotators over a period of 11 weeks.

The third step demonstrates the reliability of applying a theory of news discourse structure to news articles via a double-blind annotation study and, using the results of this annotation, demonstrate a preliminary detector of the news discourse function of paragraphs in news articles.

The fourth step of my work, detecting motific usage automatically, consists of a large-scale pipeline that achieves moderate performance. This pipeline is the first work towards automatically detecting motific usage of motifs and beats out simple baselines while comparing favorably too and generalizing better than a simple neural network baseline system.

Finally, I describe motif-like elements in a niche internet subculture and an exploration of the broader scope of narrative elements that communicate information between individuals who share a cultural or sub-cultural background. I describe these relations and provide a rough continuum of the landscape of narrative elements in cultural communication.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

Introduction

1.1 Motivation

To be truly intelligent and capable of communication with humans, many people believe that machines must be able to utilize *commonsense reasoning*, a longstanding problem at the heart of artificial intelligence. Commonsense reasoning has been recognized as "important in many AI tasks" [DM15], including natural language processing. Among many others, John McCarthy suggests that AI should "emphasize commonsense knowledge" [McC89, p.2]; Tandon et al. suggest that it is "important to endow machines with commonsense knowledge" [TVdM18]; and Cyc, a major artificial intelligence project, is "predicated on the idea that effective machine learning depends on having a core of knowledge that provides a context for novel learned information—what is known informally as "common sense."" [MWK+05].

Commonsense reasoning can be roughly defined as "what a typical seven year old knows about the world, including fundamental categories like time and space, and specific domains such as physical objects and substances; plants, animals, and other natural entities; humans, their psychology, and their interactions; and society at large" [Dav17]. I believe that especially critical is the latter part of this definition: humans, their psychology, their interactions, and society at large. This is closely related to what we might call "culture"—the "webs of significance" that man has spun [Gee73, p. 5] and the process by which humans "relate to one another so that we can understand, interpret, and, to a large extent, predict each other's behavior" [Mah13].

Crucial to this task, then, must be cultural knowledge: information that is naturally occurring within a cultural group and propagated throughout. In this work, I focus on narrative elements, how they are used in communication, and determine if humans are capable of reliably recognizing and identifying these narrative elements in text. In particular, I focus on a theory of plot structure derived from Vladimir Propp's Morphology of the Folktale [Pro68a], the discourse function of paragraphs in news as

formulated by Teun A. van Dijk [vD88], and motifs as described by Antti Aarne [Aar10] and Stith Thompson [Tho60]. I also, as a result of reflection on motifs, review a new class of narrative element which I refer to as motif-like (§7) and I propose a spectrum bridging culturally-relevant narrative elements. I give a brief introduction to each in the following sections.

1.1.1 Proppian Morphology

Vladimir Propp's Morphology of the Folktale [Pro68b] is one of the most formal treatments of narratology developed, and can be described as a *plot grammar*. Propp's theory focused on three parts of the tale: the role that a character may play at any given time (*dramatis personae*), the plot-advancing actions that characters may take part in as a role (*functions*), and the high-level organization of these actions in tales (*moves*).

Functions are the actions that a character may take during a story to drive the plot forward, for example a villain causing harm to a member of a family. Propp identified 31 such functions and noted that tales often begin with an initial situation. Functions are often linked: passing a test may result in the hero receiving a magical agent.

Compared to functions, moves are simple: they are comprised of instances of functions. "Single-move" tales start from a preparatory move that sets the stage, then a single move containing all the actions that drive the plot forward; there are also more complicated "multi-move" tales that interleave the sequences of functions in a variety of ways.

Propp's *Dramatis personae* are seven roles that characters fulfill in a story. Characters are not locked to a specific role for the entirety of the story, and characters often are not *dramatis personae* at all.

Propp's work has seen applications in a diverse range of fields, including textual story generation [FC04, GLM15, GB01], support and guidance for children during story creation activities [MPB01], and as a means of varying sign language for virtual characters based on conflict to enhance computer generated sign language [RB03].

1.1.2 News Discourse

Discourse structure is a key aspect of all forms of text, providing valuable information about the contents of a given span of text: what purpose they serve, what information they contain, and even where we might expect them to appear in the text. This is most obvious in academic, legal, (some) medical, and technical texts, which are often clearly delineated into sections containing, for example, introductory, background, or explanatory material—these types of texts are designed to make it easy to find specific information within them quickly. Understanding the discourse structure of a document has implications for text summarization, information extraction, question answering, and commonsense knowledge acquisition.

In this work, I focus on a theory of discourse in news developed by Van Dijk [vD88], which contains ten leaf elements describing the components of news grouped together under broader sections in a hierarchical manner; the details of this are described in Section 5.

1.1.3 Motifs

Motifs can be simply described as recurring cultural "memes" found in folklore and, more generally, cultural materials. Motifs are interesting because they provide a compact source of cultural information—they concisely communicate a constellation of related cultural ideas, associations, assumptions, and knowledge. Motifs are highly prominent and ubiquitous: because of this, the ability to automatically detect motifs grants access to a vast repository of important cultural information to computational analysis—an issue as of yet unaddressed [ATF21].

One common western motif that exemplifies the importance of motifs is the "troll under the bridge." One such story containing the motif, "The Three Billy Goats Gruff," goes as such: a troll, hiding under a bridge he claims ownership over, tries to devour the goats who pass over the bridge. To members of many western cultures, this combination entails a number of related ideas that are by no means directly

communicated by the surface meaning of the words: the bridge is along the critical path of our heroes, they must cross to achieve their goal; the troll lives under the bridge, crawling out to attempt to consume or otherwise extract some value from the passers-by; the troll is a squatter, not the 'officially' sanctioned master of the bridge; and the troll often meats his end at the hands of someone too strong to be eaten. The utility of the motif as a communicative device is clearly visible in the common term "patent troll"—which claim illegitimate ownership over *ideas*; attempt to consume the *companies* or otherwise extract value from them; and are often only gotten rid of once faced with a legal challenge from an entity too large for them to overcome.

Because of this density of information, motifs are often retained within a tale as it is passed between cultures and down generations, which has led folklorists to construct motif indices that identify motifs and their presence in specific tales). The most well-known motif index is the Thompson motif index (TMI) by [Tho60], which references tales from over 600 collections, indexed to more than 46,000 motifs and sub-motifs. Thompson's index designates each motif with a code; for example, "troll under a bridge" is referenced by the codes G304 and G475.2. In this case, "troll under a bridge" is represented by two motifs as Thompson generalizes trolls to ogres, a general class of monstrous beings; thus, the motifs are "troll as ogre" (G304) and "ogre attacks intruders on bridge" (G475.2).

Thompson informally defines a motif as items "worthy of note because of something out of the ordinary, something of sufficiently striking character to become a part of tradition, oral or literary. Commonplace experiences, such as eating and sleeping, are not traditional in this sense. But they may become so by having attached to them something remarkable or worthy of remembering" [Tho60, p. 19]. He notes that motifs generally fall into one of three subcategories: an event, a character, or a prop [Tho77, pp. 415–416]. Here we give an example of each (with their associated Thompson's motif code):

A **hero rescuing a princess** (B11.11.4) is perhaps one of the most well-known event motifs in western culture. Ask a westerner the following question: "A princess

has been kidnapped: who kidnapped her, who rescues her, and what does the rescuer need to do to effect a rescue?", and common answers will be "a dragon kidnapped her, the knight must rescue her, and he must kill the dragon." This motif may be the climax of the story, with a "happily ever after" ending just after the hero defeats the dragon, or it may just happen in the course of a story: in *Ivan Dogson and the White Polyanin*, a Russian folktale [Afa57a, Tale #139], Ivan slays three dragons, each with more heads than the last, rescuing a princess each time. The motif is prolific, found across the tales, literature, and movies of many cultures.

Old Man Coyote (A177.1) is a character motif: known in some Native American Indian tribes as Coyote, he is one of the most recognizable gods. In Native American Crow folklore, Old Man Coyote creates the earth and all the creatures on earth. He travels the world, teaching the animals how they should behave. Old Man Coyote, however, is far from a noble and elegant creator. He creates ridiculous costumes and tries to trick the Crow tribe into wearing them, only to be run off. He purposefully bungles rituals to produce food, such as transforming skin from his back to meat, in order to guilt his guests into performing the ritual correctly to get free food, later performing it correctly to discredit his former guests when they tell others he erred. Anywhere Old Man Coyote is referenced, he calls to mind someone who has done great things, but is lazy and often far too clever for their own good, falling pray to their own cunning.

A **magic carpet** (D1155) is a prop that allows the hero to fly through the sky, and is familiar to anyone who has watched Disney's *Aladdin*. In *One Thousand and One Nights*, Prince Husain encounters a merchant selling a carpet for an outrageous price; the merchant says: "O my lord, thinkest thou I price this carpet at too high a value? ... Whoever sitteth on this carpet and willeth in thought to be taken up and set down upon other site will, in the twinkling of an eye, be borne thither, be that place nearhand or distant many a day's journey and difficult to reach" [Bur09, p. 496]. Solomon, said to be the third king of Israel, was said to have a carpet 60 miles on each side that could transport him vast distances in a short amount of time. In Russian hero tales, magic carpets are common items that aid the hero in his quest.

Although the above examples are drawn from folklore, motifs have importance beyond folktales: they occur in modern tales, news stories, press releases, propaganda, novels, movies, plays, and anywhere that cultural materials are found. As discussed later, motif-like elements can even be found in internet subcultures, which lack a specific grounding in folklore. One powerful example is the use of the *Pharoah* motif in modern middle eastern discourse. The Pharaoh is an arrogant and obstinate tyrant who defies the will of God and is punished for it. In modern Islamist extremist narra-

tives, the Pharaoh is a symbol of struggles against anti-Islamic regimes and has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush, whom Osama bin Laden referred to as the "pharaoh of the century" as noted by [HCGJ11]. In calling George W. Bush the "pharaoh of the century," bin Laden intends to condemn Bush as the worst oppressor of his people that has been seen in the past one hundred years. Without this information, we would be unable to understand both the content of this message (that these leaders are considered oppressors) and the cultural group for whom this message was intended; instead, an uninformed human or automated tool may assume it is simply a way of referring to these individuals as leaders of their nations.

How Common Are Motifs?

It is important to address, as part of the motivation for why I find motifs an interesting and important topic, how frequently motifs actually occur in text. At first glance, our data suggests that motifs are relatively infrequent: roughly 7% of matches for lexical forms—the text of a motif (e.g., "Finn McCool")—are used to invoke the greater motific meaning, and that assumes that the lexical form occurs. However, within our lab and with our collaborators, we have estimated using several large corpora [Par11, Dav16] that this idea-invoking usage of motifs account for roughly 0.008% of all words in news articles. In comparison, the phrase "quid pro quo," oft-encountered to the point of fatigue, accounted for 0.01% of all words in news articles in the U.S. from August–November 2019. Thus, motifs are nearly as common as widely familiar phrases in news.

A More Refined Definition of Motif

One necessary step towards the computation of motifs is providing a more formal definition for them: recall Thompson's definition of motifs as *something remarkable or out of the ordinary.* While eating is not a motif, eating from a magical table is. However, Thompson described his analysis as selecting elements that he felt were of interest to

future scholars, suggesting a less principled and more intuition-driven approach. From Thompson's discussions on motifs, a more concise version of Thompson's definition might be: a motif is any remarkable or non-commonplace element in a story, where Thompson's definition of "element" are actors, items, and single incidents [Tho77, pp. 415–416]. Within this work, I refer to these elements as characters, props, and events, respectively.

This simple definition provides several problems, which some, such as Heda Jason, have attempted to address [Jas07]. I have found that further attempts to define motifs have their own problems; I address both the problems with Thompson's definition as well as further definitions in my early work on motifs [YF16a]. In this early work, I provide my own definition of motif based on Thompson's definition: "A motif is a set of closely-related variants of a non-commonplace, specific narrative element that is repeated across tales of the same type." Throughout this dissertation, this is the definition I intend when I used the term *motif*.

1.2 Interdisciplinary Connections

I have approached the work done in this dissertation from a computational perspective, as my training and work is primarily in computer science. However, this work touches (and is potentially relevant to) many fields outside of computer science, all of which are important to the whole of this work. A brief list of fields relevant to my work includes folkloristics, narratology, sociolinguistics, structuralism, cultural anthropology, linguistics, conversation analysis, stylistics, discourse analysis, literary theory, semiotics, and film theory, among others I have failed to list or am yet to be properly acquainted with. In particular, I draw heavily upon the works of Vladimir Propp, Teun A. van Dijk, Antti Aarne, and Stith Thompson. First, I will describe how my work relates to the work of those four, their fields, and some immediately related (predominately) non-computational work. After that, I will provide a brief description of

research that my work is not based upon, but that I feel is relevant to my work as a whole.

Vladimir Propp's Morphology of the Folktale [Pro68a] is what today we might refer to as a "plot grammar," which describes the moving parts that order how roles drive the plot forward and how these are arranged into broader moves that comprise story arcs; Lévi-Strauss refers to Propp's system of functions as "metastructure" [LS84]. There has been much other work on plot structure, such as Joseph Campbell's *The Hero with a Thousand Faces*, wherein he suggests that myths share a fundamental structure (the "monomyth") [Cam08] and Claude Lévi-Strauss argues that myths are composed of constituent units in higher-order structures [LS55]. In comparison to other approaches, Propp's work aims to capture the formulaic repetition in folktales in a formal way, placing content in a secondary role. Propp's focus on form over content in constructing his morphology has drawn criticism [LS84, p. 179], but his work has been deeply influential in performing structural analysis on tales even outside of Russian folklore [Dun62, Col73], and his work has been described as one of the "most valid Formalist contributions" as applied to fiction [Erl65]. Of course, as may be expected, Propp's work has also been deeply influential to computational approaches to narrative, with applications as early as 1965 [Dun65]. To those outside of the field of computation, the results presented in §3 almost certainly remain relevant as a whole: they demonstrate the reliably application of Propp's theory to text as well as identify a shortcoming in the form of potentially missing functions. The results strongly suggest that Propp's general approach was correct, with some oversights.

Teun A. van Dijk's theory of news discourse [vD88] is a hierarchical model, suggesting that topics are arranged according to ever-finer descriptions of the content contained within a given piece of discourse. While other work has focused on schemes describing inverted chronology [Bel98, Del00, VD86], I focus on the hierarchical model in part due to its presentation being easy to adapt for an annotation study, but also due to its nature lending itself to potential future computational applications such as information extraction or summarization. Van Dijk's theory has been discussed on the

basis of news comprehension [Bek06] and has been proposed for annotation of Dutch news articles [vdVBB+11]. Of potential interest to non-computational researches is, again, a demonstration of the reliability of van Dijk's scheme as applied by human annotators (§5). Additionally, my work describes a basic pipeline for automatic labeling of paragraphs according to van Dijk's theory, and further work by Banisakher, *et al.*, which I contributed to, improves significantly upon this model [BYA+20].

The work of Antti Aarne [Aar10], building on top of the idea of classifying folktales according to themes discussed by Roman Vólkov [V́24], describes a system of *tale types*, which incorporated motifs as a way to classify these tale types. Further revisions to tale type catalogs have been made by Stith Thompson [Tho60] and as recently as 2004 by Hans-Jörg Uther [Uth04]. Motifs are substantially more about the content, describing "unusual and striking" elements that "persist in tradition" [Tho77, p. 415–416], which has seen criticism: Lévi-Strauss notes that "no motif can be said to be indivisible" [LS84], as more complex motifs described within catalogs can be broken down into smaller elements; Dundes criticizes (among other aspects) the focus on character rather than plot, censorship, and "ghost entries" [Dun97]; and Thompson himself admits that the tale type index has entries that consist of a single motif [Tho77]. Propp, mentioned above, has a more biting critique of the approach: "If a division into categories is unsuccessful, the division according to theme leads to total chaos" [Pro68a, p. 7]. Even so, the idea of motifs is deeply interesting due to how central they are to communication within a cultural group: for example, the motif of the Pharoah has been used to decry leaders as great oppressors [HCGJ11]. This utility as a method of communication is what draws me to the concept of motif and drives my work towards an automatic motif detector (§6). I believe such a tool is of general interest to anyone who interacts with motifs (or, potentially, motif-like elements): as I express later in this work, and as those reading are no doubt aware, collecting and annotating motifs is a costly process, both in time and money. An automated tool could save countless hours of work, bootstrapping any effort related to motifs. I also

demonstrate the reliability of a motif annotation scheme through a large-scale study (§4).

The final piece of my dissertation (§7), inspired by a small case study into the niche internet subculture of involuntary celibates, or "incels" (the discussion on which may also be of interest to those investigating such subcultures), focuses on an expansion into motif-like elements and beyond, to a continuum of culturally-relevant communicative elements that spans many different disparate theories of narrative. The resemblance of these motif-like elements to other structures, including not only Propp's morphology and motifs, but also scripts [Tom87, SA75] and frames [Min74], while still generally being deployed in a motif-like fashion (hence the name) suggests deeper connections between these theories. I propose four axes on which we might express the relations between these elements: abstraction, identity, context, and purpose.

Work relevant to my work that did not form the basis for it includes Mikhail Bakhtin's ideas posited in *Discourse in the Novel*—that language is stratified into "socio-ideological" languages: "languages of social groups, 'professional' and 'generic' languages, languages of generations, and so forth" [Bak]. I feel this is a close analogue to the continuum of culturally-relevant communicative elements I propose (§7) and its deep relation to identity, both personal and cultural. The four axes I suggest may be viewed as axes on which language is stratified; because such stratification is indicative not just of who we are but, as Bakhtin suggests, who we are communicating with, these axes have deep implications for work related to both the source and audience of a piece of communication, such as text.

William Labov's work on oral narratives [LW97] suggests interesting applications of narrative theories in analyzing effectiveness and completeness of narrative structure in populations, as well as the potential of determining what makes someone a skilled storyteller; while these ideas are not explored in my work, they may be an avenue for an application in automatically performing such analyses. Although I do not do so in this work, I am deeply interested in performing similar work using the structural scheme defined by Labov [Lab97], especially exploring the idea of reportability and

viewpoint—I suspect that interesting information could be extracted by performing computational comparisons between many different viewpoints of the same story (for example, a noteworthy world event) to determine what each author considers to be reportable based on the structure of narrative.

Roland Barthes, in *Mythologies* [Bar72], demonstrates the ubiquity of domains to which my work might apply. The elements I examine as part of this dissertation: motifs, Proppian morphology, and news discourse hierarchy, may also be viewed as levels of description, noted as necessary to carry out strutural analysis by Roland Barthes [Bar66]. Discourse is the organization, Propp's morphology the form, and motifs the content. Barthes describes narrative as a "succession of tightly interlocking mediate and immediate elements" [Bar66, p. 270]—that a single unit may have correlates in different levels causes an irregular feeling to narrative. The axes I propose, in particular abstraction and narrative context, seem to me related to the idea of dimensions that creates narrative structure. There is also a relation between my work performing annotation studies and the idea of signification in myth, which Barthes discusses at length [Bar72]: on one level, the text itself is a signifier of the narrative elements I am interested in; on a second level, that the text signifies narrative elements is a signifier of the deeper meaning behind the theories that originated them. That there is some connection here is clear to me, although the exact nature of it is not.

I have referenced Lévi-Strauss multiple times where he refers to Propp, but his broader work is also relevant. Lévi-Strauss aligns closely with my work in his discussion on "mythemes," the constituent unit of myth—the elements I examine in my work may be viewed as forms of mythemes. Further afield from my work, I find particularly interesting his assertion that "[l]anguage is a social phenomenon," and that "much of linguistic behavior lies on the level of unconscious thought" [LS63, p. 56]—if the same thought can be applied to the higher-order narrative elements I explore, it makes it clearer that not just the words we use, but the manner in which we deploy elements of structure and meaning are deeply tied to our unconscious identity and how we express it. Lévi-Strauss explores this along many avenues, including as it applies to broader

social phenomenon such as taboo (like incest), marriage, and kinship. I believe that social structure, beliefs, and rules could be reflected, to a degree, in the narrative elements we use to communicate; one potential example comes from the "stereotypical situations" in the subculture of "incels" that I explore, the contents and structure of which deeply relate to their worldview.

Similarly related is the work by Algirdas Greimas [GPC89], whose work on narrative grammar has a clear connection to this work. Less obvious, and perhaps somewhat related to the concepts raised by Bakhtin on the social stratification of language and Lévi-Strauss on the unconscious, social nature of language, Greimas' discussion on sociolects, sociosemiotic roles, and semiotic groups is deeply related to the ideas of identity and narrative elements that I raise towards the end of my work.

As a whole, my work is both verification of the human ability to identify certain theories of narrative and a demonstration of computational tools that aim to do the same. I hope that those of you who have found this work from a far-flung (from my perspective) field are inspired to either try your hand at computational solutions or reach out to a colleague to collaborate and do so: interdisciplinary work is desperately needed to perform truly robust research on how humans interact and communicate, how we might aid, emulate, and understand these human qualities through computational means, and how the computational tools we build can further our ability to more deeply understand humans from the perspective of all fields.

1.3 Problem Statement and Research Components

I explore the problem of detecting communicative narrative elements in five ways: (1) demonstrating that humans can reliably annotate Propp's morphology; (2) demonstrating that humans are capable of detecting motifs via annotation; (3) developing a system for the automatic detection of the discourse function of news paragraphs; (4) developing a system for the detection of known motifs; and (5) exploring the range of motif-like elements that occur in cultures and subcultures that do not have a well-

grounded folklore from which motifs may be drawn and the spectrum of communicative elements.

1.3.1 Human Annotation of Narrative Elements: Proppian Morphology

I describe yet-to-be-published work by my advisor and I, building on previous work [YF16b], in which we demonstrate the reliability of Propp's morphology and point out shortcomings of Propp's theory. This work predates the motif annotation study and provides additional background for the annotation and analysis of theories of narrative structure.

1.3.2 Human Annotation of Narrative Elements: Motifs

Before attempting to automatically detect motifs, I demonstrate that humans can perform the same task reliably with a large-scale annotation study resulting in more than 21,000 examples across more than 9,000 news articles being annotated.

1.3.3 Preliminary Detection Work: Discourse Function in News

This work is a precursor to the motif detection work and demonstrates a simple method for automatically detecting the discourse function of news paragraphs using a well-known theory of news discourse.

1.3.4 Automatic Detection of Motifs

Can motifs be automatically detected? I develop a pipeline to test this question and discuss its shortcomings and strengths, as well as the potential for genaralizing the model beyond the set of data it was originally trained on.

1.3.5 Beyond Motifs

I ask three questions that address motifs outside of the domain of folklore: (1) does the four-element motif scheme generalize beyond folklore?; (2) If not, what modifications are necessary?; (3) Where do motifs lie on a continuum of narrative elements used for communication?

I address each of these questions, performing a small, internal annotation of forum posts from the internet subculture of involuntary celibates (incels), demonstrate the reliability of a complex annotation scheme of motif-like elements that shows similarities to not only motifs but to Proppian morphology and semantic frames. I discuss some trends in the incel data, as well as suggest a continuum of culturally-relevant communicative narrative elements and a path forward from motif-like elements.

1.4 Dissertation Contributions

Within this dissertation, I demonstrate that humans can reliably identify and annotate four different types of narrative elements: Propp's morphology, van Dijk's theory of news discourse, motifs, and motif-like elements found in an internet subculture; this demonstrates the viability of detecting them, their intrinsic relevance to human communication, and provides both a benchmark and a set of data from which to develop automated tools. I develop a preliminary pipeline for the automatic detection of news discourse and a pipeline for the automatic detection of motifs, providign a breakdown of the latter's performance by motif type, compare it to a neural net model, and provide the results of early generalizability results. Finally, I discuss a case study of a small internet subculture which lead to the development of a motif-like scheme and propose a broader spectrum of communicative narrative elements.

1.5 Outline

The dissertation is structured as follows: first, I summarize the related work relevant to each chapter of the thesis (§2). I then describe my work on annotating narrative elements, first starting with my work on Proppian morphology (§3), followed by a discussion of a large-scale annotation demonstrating the tractability of human detection of motifs (§4). I describe preliminary detection work on a system for automatically detecting the discourse function of paragraphs in news (§5), followed by the development and performance of an automatic motif detector (§6). I discuss motif-like elements and the broader spectrum of cultural narrative elements beyond that (§7). Finally, I describe anticipated future work (§8) and summarize my contributions (§9).

CHAPTER 2

Related Work

In this chapter, I work through the background of the three narrative elements I focus on in this work; first, I describe the body of work related to Proppian morphologies (§2.1), including annotation and its usage in the fields of computational creativity and story generation; second, I describe work related to motifs (§2.2), from its origins in folkloristics, explorations in computer science, and work in automatically parsing motif indices; finally, I describe the history of theories of news discourse (§2.3).

2.1 Proppian Morphology

One of the most notable attempts to annotate Proppian morphologies is the Proppian Fairy Tale Markup Language (PftML) [Mal01]. Malec discusses the method and difficulties encountered while creating his annotation scheme and applying it parts of 20 Russian magic tales, but the work is hard to assess as it does not include examples of annotation nor a description of the annotation scheme in the version available online [Mal01]. PftML, as described, does not appear to handle the annotation of implicit functions nor the annotation of dramatis personae and Proppian moves. Recent work on PftML looks towards the possibility of automatically classifying and annotating Russian folktales [Mal10].

Further work on PftML [DEL+10] brings in additional linguistic information, including annotating tokens, morpho-syntactic properties of the tokens, semantic relations, among others [LDDM10], and includes brief examples of annotation, but does not appear to support signals nor discontinuous regions representing functions. Continuing work on augmenting PftML has resulted in APftML [SD10] which has been used to encode the results of an information extraction approach to automated semantic annotation of folktales [DS10] and to support cultural heritage and digital humanities research by semi-automatically annotating fairy tales for dramatis personae and functions to be queried by both specialists in those domains and linguists in general

[DSL10]. APftML has also served as the inspiration for the development of NooJ [Sil15] lexicons and grammars aimed at collecting invariants of conceptual categories from Propp's morphology, such as all verbs that express INTERDICTION [LVDD10]

Explorations into the structure of discourse have also been inspired by Propp: George P. Lakoff [Lak72] used many of Propp's functions as part of a transformational model intended to have "sufficient formal power to describe accurately the complexities of the structure of fairy tales" (p. 129).

Work in story generation has resulted in ProppOnto, an OWL ontology based on Propp's morphology [PGDA04]. Peinado et al. explicitly state that their system is not intended to be complete (p. 6) and appear to use the system solely for story generation rather than annotation.

Recent work by Lendvai et al. [LDD+10, LDDM10] attempts to integrate PftML and ProppOnto together with linguistic information, demonstrating an approach to enrich both schemes. Currently, this work appears to be a proposal, with it being unclear whether or not work has proceeded in this integration. Further, this integration does not alleviate some of the issues that PftML and ProppOnto suffer from as annotation schemes.

Work by Bod et al. [BFKL12] is the work most suited to direct comparison with our results: they directly studied the reproducibility of Proppian narrative annotations. Bod's study was split into two trials. The first study consisted of nine students who were briefed for 45 minutes, given a small handout describing the functions and dramatis personae, and asked to annotated four single-move stories with functions, subfunctions, and dramatis personae. The second study was similar to the first, but omitted subfunctions, had dramatis personae already assigned, and only had six participants [BFKL12, pp. 18–20]. Bod *et al.* concluded that the dramatis personae had an important effect on the assignment of functions. Further, they conclude from the first study that dramatis personae cannot be reliably annotated and from the second study that even given the dramatis personae, human annotators cannot reliably annotate some functions (p. 20).

While Bod *et al.* state that their previous suggestion of a large-scale study to determine how reliably humans can apply Proppian morphology to narratives is "not worthwhile" (p. 21), they themselves admit that making Propp's "vague descriptions" (p. 20) understandable to annotators may require more time and training than they were given in the study. Bod *et al.* suggest that a property necessary for a formal framework to be the basis of an automated system is that sufficiently trained human annotators will annotate a narrative in the same way (p. 17). We believe that our study has sufficiently addressed the training shortcoming identified by Bod *et al.*, by giving annotators more than 30 hours training, as well as another 90 hours of time to produce annotations.

Although Propp noted that his particular list of 31 functions and subtypes, and their orders, was applicable only to this specific set of Russian tales, Alan Dundes [Dun62] and Benjamin Colby [Col73], have shown that Proppian-style analyses may profitably be performed for other cultures. Propp's work has been called "[o]ne of the most valid Formalist contributions to the theory of fiction" [Erl65, p. 249] and has seen extensions [Dun62, Gre83].

There have also been objections to Propp's theory. Propp's work has been criticized for its separation of form from content, treating content as "less important" than form [LS84, p. 179], making the mistake of trying to "characterize a tale without mentioning the motifs" [BVPR84, p. 194]. His work has been considered inconsistent, because even as he attempts to separate content from form, morphological criteria "reintroduce some aspects of content" [LS84, p. 179]. Additional criticisms target the reproducibility of Propp's work [BFKL12] and that Propp's work is not sufficient to account for the diversity of plots in fairy tales [BVPR84].

2.1.1 Use of Propp as a Computational Creativity Aid

Machado et al. use Propp's morphology as "the underlying theory of narrative" [MPB01, p. 2] for SAGA, their architecture for collaborative story creation applications, such as Teatrix [MPP01], which has seen use in classrooms by children between

ages 7 and 9. In SAGA, Propp's morphology enables the "Director Agent" [MBP04] component to support and guide the users through the creation of narrative.

Within the realm of computationally supported creativity, Propp's morphology has seen use as part of tools for authoring narratives. Hartmann, Hartmann, and Feustel [HHF05] have developed a tool that allows authors to use Propp's functions to construct the plot graphs for interactive dramas. Schneider, Braun, and Habinger [HHF05] map Propp's functions to scenes to aid authors in the development of general interactive narratives.

The GEIST project [SGBI02], aimed at developing interactive storytelling during city tours as an augmented reality game, also uses Propp's morphology extensively, as the underlying story model. Propp's functions are used in GEIST "as classes," with the scenes acting "as instances of those classes," (p.39). The next function in a sequence is selected based on the criteria for function progression that Propp gave in Morphology of the Folktale.

2.1.2 Use of Propp in Story Generation

Generation has a long history with Propp, with some very early work in story generation being based on Propp: Klein et al. used Propp's morphology to generate stories [Kle75, KAA+77], creating a system capable of generating folktale text based on Propp's morphology.

Grasbon and Braun [GB01] describe an implementation of an interactive storytelling system that used the dependencies and sequences inherent in Propp's functions to generate narratives from pre-written scenarios for each function.

Arinbjarnar [Ari05] describes the creation of a murder mystery game engine based on a Bayesian network designed with a morphology similar to Propp's morphology. Fairclough and Cunningham's work [FC02, FC03, FC04] also focuses on games, integrating Proppian characters and character functions into a game as part of a case-based planning and constraint satisfaction system designed to make agents react appropriately to player actions while following a plot.

Early work by Peinado and Gervás uses ProppOnto [PGDA04] as the cases of a case-based reasoning system intended to drive a multiplayer, directed interactive fiction engine mimicking tabletop (also called pen-and-paper) role-playing games [PG04]. Later work by Gervás used a similar approach to measure the semantic distance between situations and maintain "an independent story structure from the simulated world" [GDAP+05, p. 4].

Peinado and Gervás [PG05] also raise questions about the creativity of narratives generated by Proppian morphology, determining that it is possible to produce relatively novel narratives through generation. Further experiments have indicated that generated narratives do well in terms of narrative cohesion, but score less well than stories originally written by a person, although they still score better than a random baseline [PG06].

Newer work by Gervás describes the partial implementation of a story generation prototype based on Propp's morphology [Ger13] and work by Gervás, León, and Méndez [GLM15] attempts to reconcile existing schemes with Proppian character functions and extend a Propp-based generation system (Propper) to support schema-driven generation. Additional work on the Propper system aims to explore how dependency relations and constraints that occur at the end of a plot must be "taken into consideration when designing a constructive procedure for plots" [Ger16, p. 188].

Recent work by Gervás et al. has used Propp as the basis for developing an annotation scheme for musical theatre plots [GHLG16]. This work has resulted in the production of the "world's first computer-generated musical" [Bro15].

Thomas [Tho10] describes two methods for generating folktales using Propp's morphology. The first method expands existing analyses of tales (called *schemes* by Thomas) and filling in roles with random character. The second method attempts to generate fully-formed moves by considering the sequential order of Propp's functions and filling in subtypes for the functions in the move.

Imabuchi and Ogata [IO12b] describe a story generation system that uses Propp's morphology both as the basis of a story grammar and as part of a database. This

system has proposed applications as part of an integrated narrative generation and one experimental application, called KOSERUBE, has already been developed [IO12a]. Later work by Imabuchi and Ogata suggest revisions to the system to increase the flexibility by generalizing the generation mechanism [IO13].

2.2 Motifs

Motifs are retained throughout generations due to the density of information they contain. Folklorists have long observed that a tale's specific composition of motifs can be used to trace the tale's lineage [Tho77, Part 4,Chapter V]. This has led folklorists to construct motif indices that identify motifs and note their presence in specific tales (usually as represented in a particular folkloristic collection). The most well-known motif index is the Thompson motif index (TMI) [Tho60].

While Thompson's index is the best known, there are many other motif indices targeting specific cultures and periods, for example, early Irish literature [Cro52], traditional Polynesian narratives [Kir71], or Japanese folk-literature [Ike71]. In addition, the idea of motif was incorporated into another useful notion, the *tale type*, which seeks to classify whole tales based on their motifs. Antti Aarne constructed an index of tale types in 1910 [Aar10], with translations and revisions by Thompson [Tho60] and Uther [Uth04] (the last being known as the *ATU* catalog).

Thompson also has substantial discussion on motifs and the compilation of indices in his book *The Folktale* [Tho77]. While Thompson's motif index is perhaps the primary source of motif information used today, it has been criticized because of overlapping motif subcategories, censorship (primarily of obscenity), and missing motifs [Dun97]. These motif indices provide a substantial base for us to build upon and we draw heavily from both the Aarne-Thompson index as well as Tom Peete Cross' Motif-index of of Early Irish Literature [Cro52] and Dov Noy's Motif-index of Talmudic-Midrashic literature [Noy54] to select our group of motifs.

2.2.1 Computer Science

Darányi [Dar10] has called attention to the need for research into the automation of extraction and annotation of motifs in folklore. As a precursor to the AMICUS project, a proposed network for communication and collaboration on the task of automatic motif identification, Darányi suggests relevant existing classification schemes, including Proppian morphology, and proposes that automatic motif recognition has substantial potential to affect document processing and information extraction in the fields of folklore and scientific texts.

Further work by Darányi and Forró [DF12] have determined that motifs may not be the highest level of abstraction in narrative: their analysis of two segments of the Aarne-Thompson-Uther tale type catalog [Uth04] suggest that there exist motif strings, a higher level of abstraction, exemplified by triplets and quadruplets in the "supernatural adversaries" segment of the ATU.

Darányi et al. [DWF12] have also made headway towards using motifs as sequences of "narrative DNA", demonstrating examples of these motif sequences that are repeated and can be viewed as a type of "recombination" in the storytelling process; this analogy is extended to insertion, deletion, and duplication of motifs and motif strings in stories from the "tales of magic" tale type category, demonstrating how motifs and sequences of motifs are used, reused, and remixed as stories evolve and are retold through generations and cultures.

Work by Ofek et al. [ODR13] have demonstrated learning tale types based on this idea of sequences of motifs as "narrative DNA." his work demonstrated that for some classes of tale types, especially the domain of "tales of magic," the classifier obtained solid performance, suggesting that given enough data (as "tales of magic" was the most prominent class in the dataset), tale types can be learned and automatically detected based on sequences of motifs.

Declerck et al. [DLD12] have also done work on converting electronic representations of TMI and ATU to a format that enables multilingual, content-level indexing of folktale texts, building upon past work [DL11]. Currently, this work appears to be fo-

cused on the descriptions of motifs and tale types, without reference to the stories. In this work, we take the first steps towards automatically detecting motifs in folklore as well as verifying this system with gold-standard data created by in-culture annotators.

With regard to analyzing motif annotation schemes, Karsdorp et al. [KKM+12] present an analysis of the degree of abstraction present in the ATU catalog and the methods used to note what motifs belong to a given tale type. They find the ATU annotation insufficient for analyzing recurring motifs across types, in that it the ATU scheme fails to capture commonalities across closely related types while also failing to provide sufficient detail.

One important note is that motifs are not necessarily constitutive elements—that is, the presence of absence of a motif is not definitional for the identity of a particular tale. Motifs, rather, impose a "family resemblance" relationship between different version of the same tale. For example, in the well-known tale *Cinderella* [Cox93], found across many different cultures, several motifs commonly recur across retellings: three evil step-sisters, a fairy godmother, a glass slipper, and so forth. But the story will continue to be recognizable as *Cinderella* if the pumpkin carriage (F861.4.3—*Carriage from pumpkin.*) is replaced by another means of transportation or does not appear at all. A story having all the motifs of other tales of the same type is sufficient, but not necessary, for it to be recognized as a member of that tale type. Fisseni and Lawrence [FL13] have shown results where, in some cases, modifying the motifs involved may result in a story very similar to the original in what they refer to as a "simple solution to the problem of integrating the proposed change" (p. 103). Ignoring these non-constitutive motifs smooths over details that may potentially contain cultural information and, thus, is not in our interests.

Jason [Jas07] makes an effort to more clearly define motifs, leveling similar complaints on the clarity of Thompson's definition of motifs to those in this paper. Jason provides a definition of motifs as narrative elements that meet the following criteria: they must be (1) the simplest unit of content that fill a primary formal slot of literary structure (a character or deed) and (2) context-free (not belonging to a certain plot).

There are issues with this definition. Jason does not appear to define what simplest means beyond filling a slot of literary structure. Restricting motifs to characters or deeds ignores the importance of props within a story, such as magic carpets (D1155). And context-free motifs ignore the vast wealth of cultural knowledge that motifs contain: to encapsulate cultural knowledge, motifs necessarily arise from related tales (a tale type) within a culture.

To address the concerns raised by other definitions, my past work [YF16a] proposed my own definition based on Thompson's original definition: "A motif is a set of closely-related variants of a non-commonplace, specific narrative element that is repeated across tales of the same type." This work also proposed a general framework for a potential motif extraction and detection system; the system presented in this work builds upon this past work.

2.2.2 Automatically Parsing Motif Indices

In addition to work on defining and using motifs, my past work [YF16a] also attempted to automate the parsing of Thompson's motif index. However, I uncovered numerous challenges in this apparently simple task: first, there is no high-quality digitized version of Thompson's motif index. One commonly cited online source, hosted at `Ruthenia.ru` [Rut], a joint effort between Moscow-based publisher OGI and the Department of Russian Literature at Tartu University to provide sources for Russian language research, has inconsistent HTML and numerous OCR errors that makes parsing of the index difficult. The MOMFER effort to parse the motif index with the intention of creating a search engine [KvdMMvdB15], provides code for parsing the HTML motif index hosted at Ruthenia.ru, but is incomplete and does not accurately parse large parts of the index.

Through this effort, other issues with Thompson's motif index have come to light: many of his references to "tales" are simply cross-references to other collections (such as Cross' index of Irish literature [Cro52] or Boberg's index of Icelandic literature [Bob66], among many others). Thus the index does not provide in many cases a direct

link between motifs and tales: many stories are cited for only a single motif, despite containing more. Many of the cited stories and collections are hard to find or may no longer be accessible. Due to these issues, the motif index will likely not provide a solution to the initial problem we identified: the need for a corpus with many related motifs.

2.3 News Discourse

There has been a substantial work describing how the structure of news operates with regards to the chronology of real-world events. Much news follows an inverted chronology—called the inverted pyramid [Bel98, Del00] or relevance ordering [VD86]—where the most important and typically the most recent events come first. Bell claims that *"news stories... are seldom if ever told in chronological order"* [Bel94, p. 105], which is demonstrated by Rafiee *et al.* for both Western (Dutch) and non-Western (Iranian) news [RSS17]. Rafiee *et al.* also show that many stories follow a hybrid structure, which combines characteristics from both inverted and chronological structures.

In this work, we focus on van Dijk's structural approach to the structure of news discourse [vD88], which is organized as a tree. We choose this work as our focus due to the presentation and description of the schemata, which facilitated the quick development of an annotation guide. A more in-depth description of van Dijk's theory is presented in Section 5.2.

Discussing van Dijk's theory of news discourse, Bekalu states that analysis of "the processes involved in the production of news discourses and their structures will ultimately derive their relevance from our insights into the consequences, effects, or functions for readers in different social contexts, which obviously leads us to a consideration of news comprehension" [Bek06, p. 150]. The theory proposed by van Dijk has also been proposed for use in annotating the global structure of elementary discourse units in Dutch news articles [vdVBB+11].

Pan and Kosicki [PK93], in a similar analysis, present a framing-based approach that provides four structural dimensions for the analysis of news discourse: syntactic structure, script structure, thematic structure, and rhetorical structure. Of these, the syntactic structure is most closely aligned with van Dijk's theory. In this paper, we chose to focus on van Dijk's theory as Pan and Kosicki do not provide a list or description of the structure that could be readily translated into an annotation scheme.

White [Whi98] treats the structure of news as being centered around the headline and lead. White suggests that the headline and lead, which act as a combination of both synopsis and abstract for the news story, serve as the nucleus for the rest of the text: *"the body which follows the headline/lead nucleus—acts to specify the meanings presented in the opening headline/lead nucleus through elaboration, contextualisation, explanation, and appraisal"* [Whi98, p. 275]. We focus on van Dijk's theory for this paper as we find it to provide a higher degree of specificity: White's specification modes serve roughly the same purpose as higher-level groupings in van Dijk's theory.

Finally, building on top of the work described in this thesis (§5)), further work by Banisakher et al. [BYA+20] has improved the state of the art in automatically detecting discourse function of paragraphs in news.

CHAPTER 3

Human Detection of Narrative Elements: Proppian Morphology

3.1 Motivation

Vladimir Propp's Morphology of the Folktale is an approach to plot structure born out of Russian formalism. Propp describes it as an attempt to "make an examination of the forms of the tale which will be as exact as the morphology of organic formations" [Pro68b, p. xxv]. Propp's insight was that rather than looking at the characters, a structural analysis should look instead at what actions they perform to advance the plot forward: what he called "functions."

This is in contrast to motifs, which originate from "something of sufficiently striking character" that "become a part of tradition, oral or literary" [Tho60]. While motifs characterize the most important single elements—characters, props, or events—that comprise a body of a cultural group's folklore, Propp's morphology characterizes the manner in which stories progress and how the roles characters fill drive the action forward. In a way, some motifs and parts of Propp's morphology can be considered complementary: whereas Propp's work describes the form, the motifs describe the content.

This work was done alongside my advisor, Mark Finlayson, as an extension of prior work [Fin16, YF16b] on stories from folklore and what narrative elements exist there that encapsulate cultural knowledge[1]. As an extension of prior work, we continued looking at Propp's morphology and the 46 Russian folktales that he examined as part of his work. While this section of work focuses on folktales, it is applicable beyond them: prior work has focused on story generation and computational creativity. Further, Propp's approach is generally applicable to stories, which are ubiquitous—they appear across culture and across domains, in allegories, plays, modern tales, and news. Stories often contain a wealth of information, such as moral imperatives, the

[1]It should be noted that much of this work is drawn from work from an in-progress journal article. Owing to the substantial contributions of my advisor, Dr. Finlayson, to this work, I use plural pronouns for the majority of this section.

thoughts and actions of people in love, and cultural beliefs, norms, and traditions. Critical to understanding and extracting this information is the way the information is presented: the plot structure of the narrative.

Plot structure, broadly speaking, is the manner in which characters and events relate to each other over the course of action of a story. Much research on plot structure has focused on what relations occur and how to capture and express these relations in a way that can be used to describe a wide set of stories. In Joseph Campbell's The Hero with a Thousand Faces [Cam08], Campbell suggests that important myths all share a fundamental structure (what he terms a monomyth) describing how the hero interacts with the world. Claude Lévi-Strauss, similarly, argued that myths are composed of constituent units arranged in higher-order structures [LS55].

Previous work on folktales by Roman Volkov [V24] and Antti Aarne [Aar10] attempted to create a system for expressing and describing the components of the folktale, focusing on motifs (repeating plot elements) and tale types (classes of similar folktales). Propp dismissed these approaches for being unscientific and suggesting the notion that there is always a clear-cut division of folktales into types. As the first example of its kind, Propp's work aimed to capture the formulaic repetition that is present in folktales in a precise and relatively formal way.

With today's mathematical and conceptual machinery, Propp's theory can be described as a *plot grammar* (we describe Propp's theory in greater detail in §3.3). To derive and provide evidence for his theory he analyzed approximately 100 Russian hero tales drawn from the collection of Aleksander Afanas'ev [Afa57b], providing for each tale a list of functions that appear and the order of their combination.

Within the fields of computational linguistics and natural language processing, Propp's morphology has a substantial number of potential applications for three reasons. First, Propp's morphology is one of the most formal narratological treatments developed so far, having a relatively clear method for determining and extracting theory components from text. Second, Propp's work separates content from form, allowing a description and analysis of a plot without requiring its instantiation directly into lan-

guage. These two points combine to make Propp's morphology readily applicable to the creation of computational models. Third, properties of Propp's morphology—that functions always occur in the same sequence—makes them a powerful tool for story generation and computation creativity. Propp's work has been applied in systems as diverse as textual story generation [FC04, GLM15, GB01], support and guidance for children during story creation activities [MPB01], and as a means of varying sign language for virtual characters based on conflict to enhance computer generated sign language [RB03].

3.2 Three Questions and Description of Work

Surprisingly, despite the deep interest in Propp's work in computational circles, there are fundamental questions regarding Propp's morphology that need to be answered. We have raised, as part of this and previous work, three questions that are critical to determining the *reliability* of Propp's theory—that is, whether independent people will agree when applying it—and test them with the aid of annotators who have been given substantial training with Propp's morphology. First, will annotators given a list of Propp's functions and the functions identified in a tale agree with each other one where and whether those functions appear? Second, will the annotators, given only a list of functions, agree with each other when asked to find the functions in a tale? Third, will annotators, armed only with their training and experience, agree with each other and with Propp about which functions are indicated by a set of tales? This third question has been addressed to a degree by prior work and the questions are described in more detail in §3.4.

My advisor performed a large-scale annotation study for which I provide the analysis to answer two of these questions: as a result, we provide a fully-annotated corpus of all of Propp's work, drawn from Aleksander Afanas'ev's corpus, for which Propp

provided annotations[2]. These annotations use our previously developed annotation scheme for Propp's morphology, ProppML [YF16b].

Annotation was done in three batches: first, annotators worked on nine batches, consisting of 28 stories, with full access to the material. Second, annotators worked on another 11 batches, consisting of 18 stories, using a redacted version of Propp's Morphology of the Folktale. Finally, they returned to the second batch of stories, with full access to the material. Afterwards, the annotations were adjudicated and merged: the merge of the annotations with full access form a gold-standard set, while the restricted access annotations form a best-effort set. We use both the individual and the merged annotation sets are used to answer our questions about Propp's theory and determine that it is *reliable*: agreement measures calculated over this set of data shows that Propp's system can be reliably applied by human annotators to all of the tales that Propp annotated. This holds true even in the absence of the functions identified by Propp in specific tales. While prior work has explored this (e.g., [BFKL12]), the authors suggest a lack of adequate training for their annotators may contribute to their results. Our work addresses this shortcoming and robustly demonstrates the reliability of applying Propp's system. Open material human annotation has a high inter-annotator agreement ($F_1 > 0.7$ and $\kappa > 0.9$ for functions, $F_1 > 0.55$ and $B^3 > 0.65$ for moves, and $F_1 > 0.75$ for *dramatis personae*). In general, these results are very good, with the κ for functions being "near perfect."

3.3 Overview of Propp's Morphology

Before describing the experiments and results in more detail, it is necessary to first provide a description of Propp's morphology. Vladimir Propp's Morphology of the Folktale [Pro68b], as described in the motivation, is one of the most formal treatments of narratology developed. Propp's theory focused on three parts of the tale: the role

[2]In English editions of *Morphology of the Folktale*, Propp provides only 45 annotations. Other translations, such as the 1972 German translation [PEW72], include Propp's annotation for a 46th tale.

that a character may play at any given time (*dramatis personae*), the plot-advancing actions that characters may take part in as a role (*functions*), and the high-level organization of these actions in tales (*moves*). In this section, we will describe them in the order of their complexity: first functions, then moves, then *dramatis personae*.

3.3.1 Functions

Functions are, generally speaking, the actions that a character may take during a story to drive the plot forward. For example, a villain may cause harm or injury to some member of a family: Propp defines this as "*villainy*" and assigns it the symbol A. Propp identifies 31 such functions which can be divided into two groups:

1. Seven functions which take part in what we call the "preparatory move": these functions serve to set up the start of the story. For example, a preparatory movement might have an interdiction ("Don't leave your brother alone outside," symbol: γ) followed by a violation of the interdiction ("The older sister went inside to play with her dolls," symbol: δ). These functions leave the situation primed for the start of the story.
2. 24 functions, such as *villainy*, which drive the action forward. These functions are driven forward by the specific roles characters play, identified by Propp and described in 3.3.3.

In addition to these 31 functions, Propp also noted that tales usually begin with an initial situation, which he denoted with the symbol α. Many of these functions are inextricably linked: a violation of an interdiction(δ) implies an interdiction (γ), and if after a test (D) the hero is found worthy, they will acquire a magical agent (F).

Propp's functions also have subtypes (for example, in A, the harm may be caused by abduction, theft of a magical agent, or other villainy), inversions (in δ, the hero might not violate the interdiction) [Pro68b, p. 116, note 4], or repetition (what Propp called *trebling*, where a function is repeated three times for emphasis).

3.3.2 Moves

In comparison to Propp's functions, moves are relatively simple: a move is comprised of instances of functions and simple, "single-move" tales have a preparatory move, which sets the stage, and then a single move in which plot-forwarding actions take place.

For more complex "multi-move" tales, there are multiple ways in which they can take place, including (but not limited to):

- A move follows the resolution of a previous move.
- A move takes place during a previously-started move as an "episode."
- Moves may be interrupted and resumed in arbitrary fashions.
- Moves may have a common ending.

For multi-move tales, each lack or act of villainy creates a new move [Pro68b, p. 92].

3.3.3 Dramatis Personae

Dramatis personae are the seven roles identified by Propp that each character fulfills in the story. A character does not need to be the same *dramatis personae* for the entire story (a donor in one part of the tale may later become a villain) and a character need not be a *dramatis personae* at all.

The seven roles, identified by Propp, are: Hero, Villain, Donor, Helper, Princess, Dispatcher, and False Hero.

3.3.4 Propp's Original Data

Propp's original analysis of the folktales contains three substantial pieces of information: first, the non-preparatory functions that occur, including subtype and inversion; second, the move in which they occur; third, how parts of the tale or different tales

Table 3.1: Propp's analysis of tale #98: Daughter and Stepdaughter.

98	I	A^9	B^5		$\uparrow$	$\left\{ \right.$	D^7	E^7	f^9	$\left. \right\}$	$\downarrow$
							D^1	E^1	f^1		
	II	a^9	B_2^5	C	$\uparrow$	$\left\{ \right.$	D^7	$E\_^7$	$f\_^9$	$\left. \right\}$	$\downarrow$
							D^1	$E\_^1$	$f\_^1$		

combine. All three of these are present in Figure 3.1. Notably missing is information about exactly where in the text a function occurs. The information in 3.1 is taken from the 1968 translation of *Morphology of the Folktale* [Pro68b, pp. 136–137]. The individual moves are represented as the row headers I and II in the second-most left column. Surrounded in brackets towards the center of the page is an example of functions that occur simultaneously within the respective moves: passing a donor's test and providing a service in the first, and failing in the second.

Not all of the parses are present in the 1968 edition of *Morphology of the Folktale* [Pro68b]: the 1972 German translation, *Morphologie des Marchens*, [PEW72] contains the parse for a 46th tale: Tereshechka (#112). In total, we analyzed the differences between our annotations and six different translations: the 1966 Italian translation [Pro66], the 1969 Russian edition [Pro69], the 1970 French edition [Pro70], and the 1972 German edition [PEW72].

3.4 Experimental Design

The question at hand is whether Propp's theory is *reliable*. By *reliable*, we mean something quite specific: will independent people agree with each other when applying Propp's theory? In the introduction, we raise three questions that we feel are key to answering this broader question. We define them here in greater detail—assuming annotators who have extensive training and experience with annotating Propp's morphology:

Q1. Given a list of Propp's functions and the identified functions in specific tales, will these independent annotators agree with each other as to where and whether those functions appear in those tales?

Q2. Given a list of Propp's functions, will these independent annotators agree with each other, and also agree with Propp (as appropriate), when asked to find his functions in tales?

Q3. Without the resources from the prior two questions, will these independent annotators agree with each other, and with Propp (as appropriate), as to the set of functions that are indicated by a particular set of tales?

In past work, we have examined whether or not annotators, given a list of Propp's functions and the identified functions in specific tales (Q1), agree with each other with regard to a subset of Propp's work, consisting of simpler, "single-move" tales [YF16b]. In this work, we address this question for all of Propp's tales, including the more complex "multi-move" tales. Additionally, we address the second question (Q2) for the entire corpus, examining whether or not annotators, given a list of Propp's functions, will agree with each other and with Propp's original annotations. The final question (Q3) has been addressed manually [Dun64, Col73] and we have begun to address it computationally [Fin16].

What experimental design is appropriate to answer Q1? As pointed out by Bod *et al.* ([BFKL12]), a double-annotation paradigm is one appropriate approach to addressing the reliability of a textual marking scheme, and is the approach we follow here. In a double-annotation experiment, two people are trained in the operation of the scheme (these are called the *annotators*), and are asked to independently mark up texts with the scheme. The agreement between the two sets of markings created by the annotators is measured using appropriate statistical agreement measures such as the F-measure [FBY92, vR79] or Fleiss' kappa [Fle71]. High agreements indicate a positive answer to the question. Further, a gold standard marking of the texts (suitable for machine learning) can be generated by having the two annotators confer to resolve disagreements, sometimes assisted by a third party (called the *adjudicator*).

Conducting such a double-annotation experiment entails the following steps:

1. Define an appropriate and complete annotation scheme. (§3.5)
2. Select the texts to be annotated. (§3.6.1)
3. Assemble or build the tools required to do the annotation, for which we use Story Workbench. [Fin11b]
4. Train the annotators and adjudicator in both the scheme and the tools. (§3.6.2)
5. The annotators perform the annotation. (§3.6.3)
6. *Optional:* The adjudicator eliminates disagreements to generate gold standard data. (§3.6.3)
7. Measure agreement between the annotators. (§3.7)

In reality, there is often a loop between steps 7 and 1, as noted elsewhere [PS13, FE16], because analysis of the data reveals flaws in the scheme, which requires revision and a repeat of the experiment. We had already progressed through such a loop several times during previous attempts at annotation of Propp's scheme, and we discuss the lessons we learned in those loops, and how they were integrated into this final scheme, in the next section.

3.5 Design of ProppML

In prior work [YF16b], we have provided a full description of our annotation scheme, including a full formal BNF specification. In this section, we provide a brief overview of the scheme, including motivations, moving parts, and an example of the annotation, to acquaint the reader with the scheme for our discussion of the annotation procedure and the results.

In ProppML, we split the task into three separate schemes: functions, moves, and *dramatis personae*. Collectively, we refer to these as *ProppML*. Although the schemes are annotated separately, they do cross-reference each other in specific places (i.e.,

the move scheme refers to the function scheme), as well as reference other related annotation schemes described below.

ProppML allow association of Propp's theoretical constructs of functions, moves, and *dramatis personae* with the text under consideration. This association is implemented by reference to character offsets and anchored by identified token boundaries. As described in [Fin15], the texts are first run through a tokenizer (in this case, the Stanford CoreNLP suite [MSB+14a]). These tokens are indexed to character offsets and the ProppML schemes refer to the tokens by unique ID numbers.

3.5.1 ProppML: Function Scheme

The first of the three schemes allows an annotator to notate the presence of functions in text. Functions are split into two portions: the **function tag**, which constitutes the description of the function itself (as described in §3.3), and one or more **instances**, which declare where the function occurs in the text.

In the **function tag**, annotators mark the type (the *Initial* situation, a *Preparatory* function, or a *Normal* function). This separation is necessary because Propp's annotations do not specifically note where the Preparatory and Initial functions occur and they do not participate in the normal move structure of the tale [Pro68b, pp. 108–109]. Annotators also mark the symbol, the subtype (denoted by sub- and superscripts), and inversion of the function (denoted by an underscore).

One of the innovations of ProppML is the notion of a **function instance**. In our early approaches to marking Propp [Fin15], we noted there often was not one unambiguous occurrence of the function, which in some cases was indicated by Propp as *trebling*. ProppML allows all of these occurrences to be marked as separate "instances" of the same function, allowing for interleaving during repeated sequences of functions.

Additionally, annotators are required to mark the **type** for each function: either a function is explicit in the text, which the annotators mark as the *Actual* type, or it is implicit, in which case they mark the closest logically related instance and give it the

type *Antecedent* if the instance occurs before the function or *Subsequent* if it occurs before.

Finally, annotators can also mark text spans corresponding to *signals*, which allow the identification of a key word/verb that most strongly indicate the presence of the function or inversion of a function.

3.5.2 ProppML: Move Scheme

The second scheme allows for the annotation of the move structure in tales. Moves are numbered, with move zero corresponding to the "preparatory" move. Each move is represented as a sequence of function instances, rather than functions: this is important, as different instances of a repeated function be spread across different moves (see, for example, Tale 93 in [Pro68b, pp. 136–137]). Within a move, function instances are ordered by their appearance in the text.

3.5.3 ProppML: Dramatis Personae Scheme

The final annotation scheme in ProppML is the *dramatis personae* scheme. The data was previously annotated with "coreference groups" corresponding to bundles of co-refering referential expressions [Fin15]. Annotators assign any number of the seven roles Propp identified to these groups, including none: as described in §3.3.3, characters may fulfill different roles at different times, or may fulfill no roles at all.

3.5.4 Example Annotations

Figure 3.1 shows excerpts from an actual annotated file of The Magic Swan Geese [Gut73, p.349], provided for illustrative purposes, with the function, move, and *dramatis personae* annotations. These files are in Story Workbench annotation format, which is described elsewhere [Fin08, Fin11b]. Ellipses indicate removal of data to improve readability. We have included **six** annotation layers, enclosed in `<rep>` tags: the text itself, tokens, Proppian function annotation, coreference annotation, move annotation,

and archetype annotation. For the function, coref, move, and archetype layer, we only include the data applicable to the selection of text in the text layer.

```
<?xml version="1.0" encoding="UTF-8"?>
<story>
  <rep id="edu.mit.story.char">
    <desc id="0" len="4064" off="0">
    ...
    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 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 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 wings.
    ...
    </desc>
  </rep>
  ...
  <rep id="edu.mit.parsing.token">
    <desc id="3" len="2" off="353">An</desc>
    <desc id="4" len="3" off="356">old</desc>
    <desc id="5" len="3" off="360">man</desc>
    ...
    <desc id="125" len="2" off="885">on</desc>
    <desc id="126" len="5" off="889">their</desc>
    <desc id="127" len="5" off="895">wings</desc>
    <desc id="128" len="1" off="900">.</desc>
    ...
  </rep>
  <rep id="edu.mit.discourse.rep.refexp">
    <desc id="2158" len="10" off="353">3~4~5</desc>
    <desc id="2159" len="3" off="375">8</desc>
    <desc id="2160" len="12" off="375">8~9~10</desc>
    <desc id="2161" len="4" off="389">12</desc>
    <desc id="2162" len="10" off="398">14~15</desc>
    <desc id="2163" len="27" off="398">14~15~16~17~18~19</desc>
    <desc id="2164" len="12" off="413">17~18~19</desc>
    <desc id="2165" len="8" off="429">22</desc>
    <desc id="2166" len="8" off="439">24</desc>
    <desc id="2167" len="10" off="455">28~29</desc>
    <desc id="2168" len="165" off="467">
      31~32~33~34~35~36~37~38~39~40~41~42~43~44~45~46~47~48~49~50~51~52~53~54~55~56~57~58~59~60~61~62~63~64~65~66~67~68~69~70~71~72
    </desc>
    ...
  </rep>
  ...
  <rep id="edu.mit.discourse.rep.coref">
    <desc id="2456" len="3701" off="353">father|2158,2159,...</desc>
    <desc id="2457" len="3668" off="375">mother|2160,2167,...</desc>
    <desc id="2458" len="3665" off="389">parents|2161,2169,2170,2179,2182,...</desc>
    <desc id="2459" len="3538" off="398">the children|2163,...</desc>
    <desc id="2460" len="3638" off="398">daughter|2162,2165,2166,2171,2174,2176,2180,2183,2184,2185,...</desc>
    <desc id="2461" len="3211" off="413">son|2164,2177,2186,2192,2193,...</desc>
    <desc id="2462" len="235" off="467">what they had told her|2168,2181</desc>
    <desc id="2463" len="5" off="515">a bun|2172</desc>
    <desc id="2464" len="7" off="530">a dress|2173</desc>
    <desc id="2465" len="10" off="551">a kerchief|2175</desc>
    <desc id="2466" len="3338" off="621">the house|2178,...</desc>
    <desc id="2467" len="29" off="727">the grass beneath the window|2187</desc>
    <desc id="2468" len="10" off="746">the window|2188</desc>
    <desc id="2469" len="10" off="771">the street|2189</desc>
    <desc id="2470" len="5" off="806">games|2190</desc>
    <desc id="2471" len="3117" off="814">Swan Geese|2191,2194,...</desc>
    <desc id="2472" len="2691" off="889">Geese wings|2195,...</desc>
    ...
  </rep>
  <rep id="edu.mit.semantics.rep.function" ver="0.5.0">
    <desc id="2555" len="73" off="353">
      INITIAL::alpha:false::|ACTUAL:::3~4~5~6~7~8~9~10~11~12~13~14~15~16~17~18~19~20
    </desc>
    <desc id="2558" len="182" off="450">
      PREPARATORY::gamma:false:1:2|ACTUAL:27::27~28~29~30~31,57~58~59~60~61~62~63~64~65~66~67~68~69~70~71~72
    </desc>
    <desc id="2560" len="186" off="468">
      PREPARATORY::beta:false::1|ACTUAL:34,75::32~33~34~35~36,73~74~75~76
    </desc>
    <desc id="2559" len="107" off="704">
      PREPARATORY::delta:false::|ACTUAL:88,98,105::87~88~89~90~91~92~93~94~95~96~97~98~99~100~101~102~103~104~105~106~107~108
    </desc>
    <desc id="2514" len="87" off="814">
      NORMAL::A:false::1|ACTUAL:116,122::110~111~112~113~114~115~116~117~118~119~120~121~122~123~124~125~126~127~128
    </desc>
    ...
  </rep>
  <rep id="edu.mit.semantics.rep.move" ver="0.1.0">
    <desc id="2556" len="458" off="353">0|2555,2558,2560,2559</desc>
    <desc id="2549" len="3015" off="814">1|2514,...</desc>
  </rep>
  <rep id="edu.mit.semantics.rep.archetype" ver="0.1.0">
    <desc id="2529" len="3638" off="398">HERO|2460</desc>
    <desc id="2544" len="3211" off="413">PRINCESS|2461</desc>
    <desc id="2530" len="3117" off="814">VILLAIN|2471</desc>
    ...
  </rep>
</story>
```

Figure 3.1: Selection of the annotation of The Magic Swan Geese.

3.6 Data Production

3.6.1 Selection of Texts

Because answering Q1 (from §3.4) requires us to know Propp's list of functions for a tale, our raw text is necessarily drawn from Propp's original corpus, which he selected from Aleksandr Afanas'ev's collection of Russian folktales [Afa57b]. Propp analyzed 100 of these tales, publishing a subset of his analyses (46 tales) in a table at the end of his monograph [Pro68b].

Table 3.2: Corpus-wide statistics of the moving pieces of Propp's morphology.

	# Words	# Tokens	# Functions	# Functions Trebeled	# Dramatis Personae	# Moves
Total	89,094	109,120	920	130	407	85
Average	1,936.8	2,372.2	20.0	2.8	8.8	1.8
Standard Deviation	1,112.9	1,361.5	9.4	2.6	4.6	0.9

For this study we analyze the full set of 46 tales for which he published his analyses, building on our prior work restricted to single-move tales [YF16b]. Table 3.2 shows the corpus-wide statistics for the words, tokens, and moving pieces of Propp's morphology.

While Propp's work was performed on the original Russian text, all of our work is performed on English translations. Some of the tales did not have a readily-available English translation. For these tales, we commissioned a translation of them by expert Russian translators. Work on translations of texts is generally considered acceptable for first-order structural analysis [Fis63].

3.6.2 Annotator Training

Annotation was done in a double-blind manner by two highly-trained annotators. Both annotators were students at Harvard University in Cambridge, MA[3] To begin we trained the annotators for three weeks where the annotators were first asked to read Propp's book from start to finish, and then asked to annotate the Magic Swan

[3]The study was begun while my advisor, Dr. Finlayson, was a researcher at MIT, also in Cambridge.

Geese tale, the analysis of which is explained in detail in Propp's book [Pro68b, pp. 96–99]. Reading the book and the initial annotation of the Magic Swan Geese took approximately 20 hours total. The annotators were then brought together then with the adjudicator for a three-hour meeting to discuss any questions and compare their annotations. The adjudicator was already highly trained in Propp's system: he was one of the original annotators who helped produce the first set of Propp annotations [Fin12, Fin15], and was also a Ph.D. student in English literature at Harvard University. After the annotators received feedback on their annotations, they re-did their annotations on the Magic Swan Geese and had another meeting (another 10 hours total). At this point agreement was very good, and annotation of the remainder of the data began.

The annotation of the 28 tales that were not part of the closed annotation study acted as further training for the annotators, ensuring that when performing the closed annotation, they had extensive experience in annotating narrative texts with Propp's morphology. Annotators were also trained separately in the operation of the Story Workbench, the tool we used for this work, which took approximately one hour.

3.6.3 Annotation Procedure

Annotation of the texts after the Magic Swan Geese was performed at a rate of approximately 2,000 words/week, with annotators spending approximately 7 hours a week annotating, and 3 hours/week in an adjudication meeting. Therefore, after the initial training period of 30 hours, the annotators spent approximately 9 weeks annotating, and each annotator spent a total of approximately 120 hours on the project. This is substantially more training than had been given in prior work, which included annotation by the scheme developer [Mal01] and a 45-minute briefing including a handout [BFKL12].

The adjudicator spent approximately 24 hours on the project, not counting the annotation that he performed in previous years (which constituted, at a minimum, approximately 100 hours of work). During each adjudication meeting disagreements between the annotators were discussed, and additional discussion of subtleties of Propp's

system was held as needed. Further, a gold standard marking was produced by the team. Thus, the project produced three sets of marked texts: one marked by annotator 1, another marked by annotator 2, and a gold-standard set corrected by the adjudicator. If the team had a disagreement that could not be resolved, they consulted Dr. Finlayson, who ran the annotation.

For the closed annotation, annotators were not allowed to access the full version of Propp's Morphology of the Folktale nor allowed to use the notes they had taken as part of their prior annotation. Instead, the annotators were provided with a redacted version of Propp's morphology that removed any references to which functions he found in which tales. This includes examples at the end of descriptions of functions, Propp's in-depth analysis of The Magic Swan Geese, and the table in the appendix with Propp's annotations.

After the closed annotation study was complete, the annotators then created a second annotation of the tales with full access to their notes and other materials. This second, open-material annotation serves as the baseline against which the closed annotations are compared in our analysis.

3.7 Results

The goal of this annotation study was to address a high-level concern regarding Propp's theory of narrative structure: is Propp's theory *reliably*—that is, will independent people agree with each other when asked to apply his morphology to text? To this end, in designing the experiment, we raised three questions (§3.4) that we felt addressed this concern: will trained annotators, given a list of functions and the functions in the tales (essentially, a function oracle), agree with each other on where they occur? Will trained annotators, given just the list of functions, agree with each other and with Propp when asked to find those functions? Will trained annotators, armed only with their wit and experience, agree with each other and Propp as to the set of functions that occur in a set of tales?

In this work, we build upon prior work [YF16b] to fully address the first question for all of Propp's annotations and address the second question; as stated in §3.4, other work [Dun64, Col73] has manually addressed the third question.

We start by describing the agreement metrics that we use for analyzing these results, the ways in which they deviate from standard agreement metrics (if they do so), and the reason they do so (§3.7.1). We follow up by examining the results to the first question (§3.7.2), noting an interesting observation for Propp's moves that applies to all of our results (§3.7.2). We address the second question in two parts: first, the agreement between independent annotators (§3.7.3), and second, the agreement between those annotators and a gold-standard representing Propp's annotations (§3.7.4). We conclude with a brief discussion and interpretation of the results (§3.7.5).

3.7.1 Definition of Agreement Metrics

We measure agreement between annotators for functions, moves, and *dramatis personae* in several ways.

For **functions**, we report two F_1 measures: the first is a 'strict' F_1 that requires annotator markings to be identical, and the second is an 'identification' F_1 (F_1^i) that marks agreement if there is any overlap between the annotators' marked instances. The identification metric first performs a "best effort" alignment of functions, where exact agreements are paired off, and then function markings are paired off in descending order of degree of overlap.

We also report the Fleiss kappa (κ_{sym}) for identification of the primary function symbols. Although annotators were given the list of functions to mark, they were allowed to change the symbol identity or subtype if they felt it was necessary. In most cases, however, they did not make any changes: we detail these changes in our comparison between editions of Propp's annotations.

For **moves**, we report a 'strict' F_1 and a 'grouping' (F_1^g) measure: the strict measure operates on moves as a whole, while the grouping measure is calculated between the function instances involved in moves.

In addition to the F_1 metrics, we report a modified F_1 calculated using a slight modification of the B-CUBED [BB98b, BB98a] metric developed for scoring cross-document coreference, which produces precision and recall. This metric, in particular, was selected after deliberation, as characteristics of the moves made many standard agreement metrics unsuitable: we briefly discuss this at the end of this section. For the B-CUBED F_1 metrics, we report both 'strict' (B^3) and a 'grouping' ($B^{3,g}$) measures, where the grouping metric is first run through a "best effort" alignment.

For ***dramatis personae***, we report a strict F_1 and a identification (F_1^i) measures, as well as Fleiss kappa (κ_{dp}) for assignment of *dramatis personae* labels. The identification measure marks whether the annotators agreed on whether a character was a *dramatis personae*, not necessarily on the label(s).

Selection of Move Agreement Metrics

We analyzed several agreement metrics for their suitability in calculating inter-annotator agreement for moves. The F_1 [vR79] and Fleiss' kappa [Fle71, SF79] are often-used and well-suited metrics—we note how we calculate these from the data in the previous subsection. While the F_1 and Fleiss' kappa results roughly reflected what we expected to see for functions and dramatis personae, we had expected higher agreement for moves based on our experience while adjudicating the annotations. We explored four alternative metrics: the Adjusted Rand Index [Ran71, HA85], the Jaccard Index [Jac01, Jac12], the MUC-6 scoring metric [VBA+95], and the B-CUBED precision and recall metrics [BB98b, BB98a].

Calculating the agreement between moves shares many similarities to the classes of problems that these agreement metrics are intended for and work well on: functions my occur in the same move or in different moves. However, functions may also be annotated in only one tale if the annotators disagree as to *which* function occurs or *whether* a function occurs at all. Internally, our system attempts to resolve the first issue by doing a best-effort alignment of functions, including those that cover the same span but are not marked by the same symbol as a last resort. When annotators

disagree on *whether* a function occurs—that is, one annotator marks some function covering a portion of text and the other does not—no alignment resolves this. There are ways to resolve this: missing functions can either be created as singleton moves in the story missing them, added to a move containing all of the missing functions, or ignored. Ignoring missing functions loses valuable information, so we needed an agreement metric that took into account the fact that the annotators had disagreed and showed a reasonable, but not outrageous, penalty.

Many of the evaluated metrics performed poorly on moves due to missing functions. The Adjusted Rand Index, with singleton moves, performed poorly because the chance adjustment highly penalizes the one-off moves. With monolithic moves containing all the missing functions, the ARI trended upwards as more functions were missing—in cases of extreme mismatch (completely disjoint function identification), the ARI trended towards perfect agreement. The Jaccard Index, like the F_1 score that is calculated in a similar way, trends low and was not cross-move, but was calculated from an average of the move scores. The model theoretic approach from MUC-6 depends on the notion of equivalence classes—because instances of a function can theoretically appear across multiple moves, it was not possible to construct a reasonable equivalence class.

Of the evaluated metrics, B-CUBED performed well, possibly due to the close analogy between coreference chains in moves: functions can be viewed as roughly equivalent to relevant documents. Again, however, when functions were missing from a move, B-CUBED had unfavorable characteristics—while the function behaved roughly as expected, even in extreme cases, we found that it was non-symmetric: reversing the "truth" and "test" examples would cause non-symmetric changes in the precision and recall. This issue was resolved by using a slightly modified B-CUBED metric: both the precision and recall are calculated twice and averaged, which produces symmetric precision and recall. With this modification in place, B-CUBED was used alongside F_1 to measure inter-annotator agreement for moves.

Table 3.3: Microaveraged annotation agreement measures for the training, open, and combined set.

	Functions			Moves				*Dramatis Personae*		
Comparison	F_1	F_1^i	κ_{sym}	F_1	F_1^g	B^3	$B^{3,g}$	F_1	F_1^i	κ_{dp}
Training	0.266	0.738	0.928	0.093	0.660	0.556	0.679	0.729	0.800	0.490
Open	0.251	0.746	0.929	0.054	0.490	0.521	0.629	0.681	0.783	0.408
Combined	0.258	0.742	**0.929**	0.075	0.562	0.537	**0.652**	0.703	**0.791**	0.448

3.7.2 Do annotators, given a function oracle, agree when asked to find Propp's functions?

We begin by addressing whether or not, given a *function oracle*—that is, a list of Propp's functions and the list of functions Propp identified in each tale—two independent annotators will agree. In prior work [YF16b], we have examined this question for a small subset of Propp's full work: 15 single-move tales. In this section, we present the results of the annotation of all 46 stories Propp provided annotations for.

Table 3.3 shows the results for the training, open, and combined sets of data. The training set is the first nine batches, comprised of 28 tales, that were used to finalize annotator training. The open set is the final 11 batches, comprised of 18 stories, annotated with open access to materials. In bold are the best agreement measures for each aspect of Propp's theory: functions have near-perfect agreement, moves have good agreement, and *dramatis personae* have very good agreement. For our analysis, we focus primarily on the combined set (noted in **bold** in the table), which represents the entire corpus. We include the training and open sets for completeness and note that there is a slight drop in between the training and open data sets: one possible explanation for this is the difference in the level of feedback that annotators received during training and during the open material annotation.

For functions, the strict F_1 result in Table 3.3 is low, at 0.258. This result is expected—as described in §3.7.1, any disagreement at all in the annotation of functions, including disagreement over the exact extent of a function (which might span several hundred tokens) gives a penalty. In contrast, the grouping F_1 is very good,

especially considering the complexity of Propp's system: 0.742. Further, the Fleiss kappa is a *near-perfect* 0.929.

Annotators had less agreement on moves: while the strict F_1 is low, as expected, the grouping F_1 and strict B^3 are acceptable, but not outstanding. The grouping B^3 result is good, but not great: 0.652. The overall lower scores for moves is a trend that occurs throughout all of the experiments, indicating that identifying moves is a harder task than identifying functions.

Table 3.4: Confusion matrix between annotators on the open material experiment.

	HERO	VILLAIN	DONOR	HELPER	PRINCESS	DISPATCHER	FALSE HERO	UNSPECIFIED
HERO	**61**	0	1	0	1	0	0	2
VILLAIN	0	**90**	0	0	0	0	2	34
DONOR	0	0	**19**	2	0	0	0	8
HELPER	0	1	8	**40**	0	1	0	30
PRINCESS	2	0	0	0	**31**	0	0	16
DISPATCHER	0	1	0	0	1	**15**	0	15
FALSE HERO	0	0	0	0	0	0	**3**	3
UNSPECIFIED	4	20	5	23	10	6	1	**224**

Dramatis personae had very interesting inter-annotator agreement results: both the strict F_1 and the identification F_1 had very good scores: 0.703 and 0.791, respectively. In contrast to the high F_1 scores, the Fleiss' kappa is relatively low (0.448), indicating moderate agreement: this pattern occurs throughout the entire corpus.

Annotator Confusion on Dramatis Personae

We will briefly expand upon the results of the *dramatis personae* agreement: we found it unusual that there would be such a substantial drop in agreement between the F_1 and Fleiss kappa, especially as we saw the result repeated across all of our results. Reexamining our results and agreement metrics gave a possible cause: the annotators agreed *which characters* played important roles in the story, but not *what role* they played.

The differences in the agreement calculations support this: the Fleiss kappa measures the annotator agreement when assigning categorical ratings and is thus dependent on the exact *dramatis personae*. This is also true for the strict F_1, but it is *not* true for the identification F_1, which marked agreement when annotators agreed that a character was a *dramatis personae*, but ignored labels.

Table 3.4 shows the confusion matrix for dramatis personae. This confusion matrix is a sum of the confusion matrices for all of the annotated tales. The numbers represent

the number of coreference bundles that were identified as that particular *dramatis personae* by the annotators. In bold is the diagonal, showing where annotators agreed on the identity of the characters. Unspecified means that a coreference bundle was not identified as a *dramatis personae*. Bolded on the diagonal are the number of coreference bundles that annotators agreed on for the entire corpus; anything off the diagonal is disagreement. Out of the 680 coreference bundles in the annotation, annotators disagreed on 197 of them—this means that annotators disagreed on roughly 29% of the assigned dramatis personae. This high level of disagreement supports what we see with the F_1 and Fleiss' kappa scores for dramatis personae: while they can identify which characters are and are not important, they have substantially more trouble agreeing on what role these characters play.

3.7.3 Do annotators agree with each other on where and what functions occur in tales?

We now address whether or not, given a list of Propp's functions, annotators agree with each other when asked to find the functions in tales. This is the first half of our second question (§3.4) and, as described in §3.6, annotators were instructed to use only a redacted copy of Propp's Morphology of the Folktale. While this copy retained Propp's description of the functions, any reference to which functions were found in which tales was removed.

Table 3.5: Microaveraged annotator agreement for the open and closed annotations.

	Functions			Moves				*Dramatis Personae*		
Comparison	F_1	F_1^i	κ_{sym}	F_1	F_1^g	B^3	$B^{3,g}$	F_1	F_1^i	κ_{dp}
IAA, open	0.251	0.746	0.929	0.054	0.490	0.521	0.629	0.681	0.783	0.408
IAA, closed	0.124	0.616	**0.787**	0.078	0.203	0.539	**0.545**	0.589	**0.707**	0.289

Table 3.5 shows the results of our closed annotation experiment, as well as the results for the open annotation from the previous experiment. The closed annotation was intended to determine the degree to which annotators agree on which functions occur in the tale. In bold are the best results from the closed annotation for each aspect

of Propp's theory: functions and *dramatis personae* both have very good agreement and moves have an acceptable level of agreement. For almost all of the categories, the results for the closed study are worse than the open study. The one exceptions to this are the strict F_1 and B^3 for moves; however, the differences are so small that this is almost certainly nothing more than chance, especially given that the increase is not supported by the more lenient measures.

The large difference in agreement, overall, was to be expected: even with the extensive training the annotators had, the larger search space for annotation made the task substantially more difficult.

Focusing just on the results of the closed annotation (noted in the table in **bold**), we see that the results are still relatively good.

For functions, the identification F_1 is good, at 0.616, and the Fleiss kappa is very good, at 0.787. This result, especially the high Fleiss kappa, seems to indicate that the annotators were able to agree a substantial amount of the time on where functions occurred (F_1) and what function was occurring (Fleiss kappa).

As briefly noted above, for moves, the strict F_1 and B^3 metrics show a small increase over the open annotator agreement. Despite the increase, the strict F_1 remains terrible and the strict B^3 remains acceptable. Interesting, however, is that the grouping F_1 experiences a substantial drop down to 0.203 (from 0.49) and the grouping B^3 is nearly the same as the strict B^3 (0.545, down from 0.629).

One possible explanation is the degree to which moves depend on functions: as the agreement on the functions decreases, the agreement on the moves must also necessarily decrease. Beyond that, however, if annotators disagree as to which functions are occurring, it is probable that they disagree as to which moves the functions belong to.

For *dramatis personae*, The identification F_1, which focuses only on identifying which characters are *dramatis personae*, experiences a drop, but remains very good at 0.707. However, the strict F_1 and Fleiss kappa both experience a larger drop, indicating that the annotators had substantially more difficulty identifying the role that characters played.

This is backed up by looking at annotator confusion for *dramatis personae*: annotators disagreed on 141 of the 363 coreference bundles in the annotation. This indicates a disagreement on roughly 39% of the characters, which is a substantial increase from the 29% disagreement for the open annotation agreement. As with moves, we suspect that the reason might be the disagreement in functions. *Dramatis personae* are intentionally linked to the functions in Propp's theory: each function specifies what roles take the action—*hero* agrees to *villain*'s persuasions, *donor* interrogates *hero*, and so on. As with moves, a disagreement on what functions occur propagates to a disagreement on what roles the characters play.

3.7.4 Do annotators agree with Propp on where and what functions occur in tales?

Finally, we examine the second part of our second question (§3.4): will annotators, given only a list of Propp's functions, agree with Propp when asked to find the functions in tales. Both of the annotation sets compared in this experiment are merged from previous annotation studies.

The "gold standard" set is an adjudicated merge of the two annotations done with the list of what functions Propp identified in each tale. As such, it is as close as possible to an annotation done by Propp himself—extensively trained annotators identified, for each function Propp listed as being in a tale, the precise location of that function in the text.

The second set of data, referred to as the "closed" set in Table 3.6, is an adjudicated merge of the "closed" annotations, where annotators were only given access to a redacted copy of the primary source material, as described in the previous section. This merged version represents the best-effort of human annotators attempting to recreate Propp's original analysis: that is, determining not just *where* the functions are, but also *what* functions occur.

Table 3.6: Microaveraged annotation agreement for the open, closed, and gold-standard test.

	Functions			Moves				*Dramatis Personae*		
Comparison	F_1	F_1^i	κ_{sym}	F_1	F_1^g	B^3	$B^{3,g}$	F_1	F_1^i	κ_{dp}
IAA, open	0.251	0.746	0.929	0.054	0.490	0.521	0.629	0.681	0.783	0.408
IAA, closed	0.124	0.616	0.787	0.078	0.203	0.539	0.545	0.589	0.707	0.289
A1 closed vs. Gold	0.208	0.636	0.833	0.131	0.276	0.523	0.546	0.788	0.862	0.571
A2 closed vs. Gold	0.157	0.571	0.768	0.112	0.274	0.543	0.572	0.590	0.720	0.288
Closed vs. Gold	0.260	0.662	**0.879**	0.138	0.397	0.533	**0.620**	0.772	**0.877**	0.548

Table 3.6 shows the results for the comparison between annotators and the gold standard, as well as the results for the open and closed studies, described in the previous two sections, for comparison. Also shown are the results from comparing each individual annotator (A1 and A2) with the gold standard. In bold are the best results for annotator agreement with Propp on each aspect of Propp's theory: agreement for functions and *dramatis personae* is very good and agreement for moves is good. Of note is the fact that, from the inter-annotator agreement in the closed study, agreement between annotators and the gold standard is substantially higher. The contributor to this may be adjudication: as described in §3.6, during the merge process annotators discuss with a highly-trained adjudicator to resolve conflicts in their annotations. During any conflict, the annotators and adjudicators must agree on the resolution, and so the final annotation is a combination of the best annotations from both annotators.

Briefly, we will touch on the individual annotator results in comparison with the gold standard: generally speaking, the results are good to very good, and are generally comparable with the results of the merged closed annotation. However, there are some interesting minor characteristics that get smoothed out as part of the merge: annotator one had substantially higher agreement with the gold standard for functions and *dramatis personae* than the second annotator.

Additionally, the final closed merge for functions, moves, and identification of important characters (identification F_1 of *dramatis personae*) outperforms both of the annotators when compared to the gold standard. This indicates that, although the first annotator's annotation was generally closer to the gold standard, annotations

from both annotators were used in resolving conflicts, leading to a higher score almost entirely across the board.

For the remainder of this section, we will focus primarily on the results comparing the merged closed set and the gold-standard annotation (shown in **bold** in Table 3.6). Note that, by the design of the experiment, these agreement measures are only calculated on the set of 18 tales that were part of the closed annotation experiment.

Functions demonstrate the expected characteristics from the two previous studies: low strict F_1 (0.260) and high identification F_1 (0.662) and Fleiss kappa (0.879). The identification F_1 is good and the Fleiss kappa is very good: this indicates that, in comparison with Propp, annotators perform well at identifying and locating functions in tales.

Move results are at the expected levels: low strict F_1 performance, relatively low grouping F_1, acceptable B^3 and good grouping B^3. As in all of our analyses, these results are to be expected, given the apparently greater difficulty of determining what functions belong to which moves.

Agreement on *dramatis personae* is very high: the strict F_1 is good (0.772), the identification F_1 is very good (0.877), and agreement between annotators and the gold standard on which role characters play is higher than for either of the previous experiments. This substantially improved performance on the task of assigning roles is almost certainly due to adjudication. As expected, confusion matrix analysis shows a decrease in total confusion: annotators disagreed on 86 out of 315, or about 21% of coreference bundles. This is lower than the disagreement for both the open study (29%) and the closed study (39%).

3.7.5 Overall Analysis

Looking solely at functions, agreement between annotators was good across all three experiments. With unrestricted access to Propp's Morphology of the Folktale and their own notes, annotators demonstrated outstanding agreement ($F_1^i = 0.746$, $\kappa_{sym} = 0.929$), with near-perfect identification of functions, as expected. With only restricted

access to the source material, annotators agree with each other to a substantial degree ($F_1^i = 0.616$, $\kappa_{sym} = 0.787$), and annotators agree with the gold standard to nearly the same degree as they agree with each other in the open study ($F_1^i = 0.662$, $\kappa_{sym} = 0.879$).

These results show that independent annotators can reliably apply Propp's analysis to the tales (Q1, §3.4). Additionally, the results from the closed studies show that annotators agree with each other and with Propp as to what functions occur and where they occur (Q2, §3.4).

Outside of functions, the results have very interesting trends. As discussed in the individual sections, *dramatis personae* demonstrate a very high degree of agreement across the experiments. Annotators perform very well at identifying which characters play an important role in the story but not what role they play—that is, the character is a *dramatis personae*, but the annotators did not agree as to which ones they acted as.

This analysis is backed up both by the definition of the agreement metrics we used and the confusion matrices for the annotation. The highest-scoring metric, the identification F_1, only measures whether or not the character is a *dramatis personae*, not which one, while the other two metrics, which always perform worse, take into account the role. The confusion matrices showed that the size of the disagreement coincides with the degree to which annotators disagreed on which coreference bundles had a given role.

For moves, the scores were overall lower than the functions, but were acceptable to good, depending on the metric and the study. One possible explanation for this drop is that identifying moves is a more difficult task than identifying functions. This is, in ways, analogous to the difficulty annotators had in identifying *dramatis personae*: annotators agree on *where* and *which* functions occur, but they have trouble agreeing on how to group these functions together into the high-level structure of moves.

One thing shown was the effect that identifying functions had on both moves and *dramatis personae*: in the closed study (§3.7.3), there was a substantial drop in

agreement on moves. This is somewhat expected: if annotators don't agree on the functions, how can they agree on what moves they belong to? More interesting is the effect functions had on *dramatis personae*: while all the metrics dropped, the two that required identifying the specific role a character played dropped by more. One potential reason is this: the functions are, indeed, functions of the *dramatis personae* and, as such, knowing what characters are involved in which functions narrows down the possible roles they play by a substantial degree.

Additionally, the comparison to the gold standard (§3.7.4) showed interesting effects of the adjudication and merge process. Agreement between the merge and gold standard was better for almost all metrics than the individual annotations: this may be because the adjudication resolutions are the best of both annotations. The agreement on *dramatis personae* is also highest between the merge and the gold standard than for any other experiment: this is likely also a result of the merge and adjudication process.

As a result of these experiments, we have answered the two remaining core questions that we raised at the beginning of this paper. This, combined with prior work, demonstrates that Propp's theory can be reliably applied to text by independent annotators, given sufficient training.

3.8 New Insights

One class of criticisms against Propp's analysis is whether or not it represents the ground truth: does Propp's analysis account for those pieces that actually occur in narrative? Our annotation provides some insight into this question: while it appears that Propp got the broad morphology correct, he missed some details—in particular, a close analysis of the coverage of text by functions reveals that some function-like elements appear to be missing.

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. 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. 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. 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. 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." 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. Nikita, having done his heroic deed, would not accept any reward, but returned to currying hides.

Figure 3.2: Document showing the text covered by functions in Nikita the Tanner (Tale No. 148).

This analysis is enabled by our approach to verifying the reliability of Propp's morphology: by performing an in-depth text annotation including the spans of the functions, we are able to see what is and is not covered by Propp's functions.

Figure 3.2 shows a close view of the coverage of one tale, Nikita the Tanner (Tale No. 148) [Gut73, pp.168–170], provided for illustrative purposes. Text with a grey background is covered by some function. For the purposes of analysis, the coverage of a function is extended to bridge small gaps (less than three tokens, where a token is a word or punctuation mark) and from the start of the extent of the function to the end of the sentence it occurs in. Looking at just one tale, it is remarkable how much text remains uncovered by Propp's functions. However, one must keep in mind that uncovered text is not necessarily indicative of important elements being overlooked. Propp's original definitions of functions is as follows: "Function is understood as an act of a character, defined from the point of view of its significance for the course of action" [Pro68b, p. 21]. In short, functions are what drive the action of the story forward: incidental dialogue, failed actions, and actions unrelated to the overall course

of action do not qualify as functions. We term these sections of the text *unproductive narrative*.

Looking at the gaps in Figure 3.2 and keeping in mind Propp's definition of functions, we note that there are plenty of text that is relatively unproductive: the story does not start before the tsar begs Nikita to take action, but everything between the starting villainy and the tsar's request does nothing to advance the story. However, towards the end of the story, there's an interesting element: Nikita deceives his opponent, the dragon, into crawling into the sea, which enables Nikita to kill his opponent. The death of the dragon is an essential part of the action, but Propp does not assign any function covering this event, nor does he provide a function that reasonably covers it.

We performed a brief, exploratory analysis of all 46 tales covered by Propp's analysis, looking for similar function-like elements. This analysis was performed by recording and counting function-like elements in tales, establishing reportability requirements for what we call "missing functions", and an additional pilot annotation on top of the gold standard corpus to establish the rough extent of missing functions in the narrative text.

Recording and counting was performed with a simple read-through and a recording of any elements that looked like functions, along with the tale that it was found in. Similar elements were given a temporary designation and grouped together in a manner similar to Propp's subtypes for functions where it made sense. This first pass of the stories found 13 function-like elements. Some of these were existing functions that had not been listed by Propp as part of his analysis. These were disqualified for consideration.

To establish reportability requirements for missing functions, we needed to understand how frequent Propp's functions occurred in his analysis. Figure 3.3 shows the number of tales covered for each function, with subtypes grouped together under a single symbol, sorted in descending order of number of occurrences. Also included is α, the initial situation. This graph shows that a substantial number of Propp's functions have ten or fewer occurrences: we use this to establish a cutoff criteria of ten

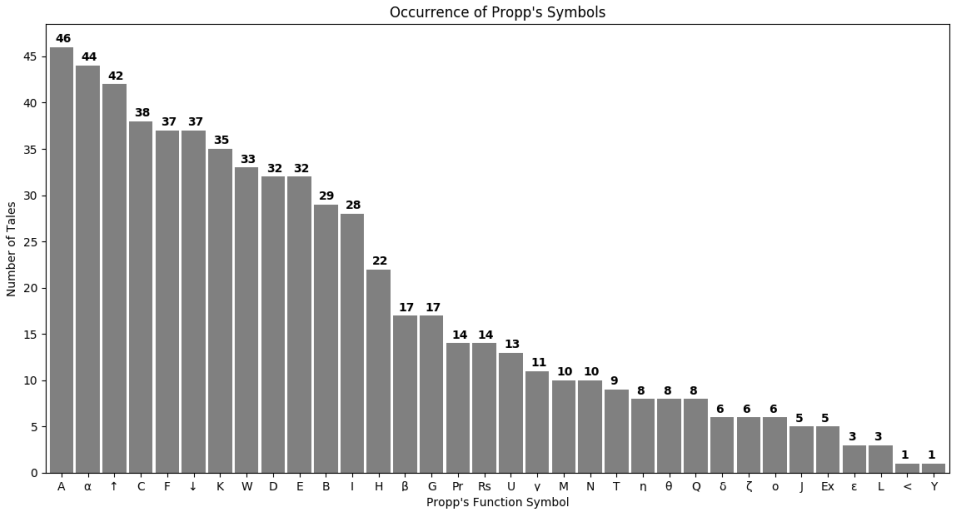

Figure 3.3: The number of tales covered by each of Propp's functions in our corpus.

occurrences for missing functions. While many functions occur frequently, many are relatively rare: five preparatory and eight normal functions occur ten or less times, and of those, one preparatory and three normal functions occur five or less times. With more than 40% (13) of Propp's 31 functions occurring ten or less times, a reasonable requirement for missing functions was that they must occur in at least ten tales.

Table 3.7: Description of the missing functions and the number of tales they occur in.

Designation	# of Occurrences	Description
Tmp5	17	Deceit by the Clever Hero
Tmp5^1	9	The clever hero uses deceit to obtain what they desire.
Tmp5^2	8	The clever hero uses deceit to defeat an opponent.
Tmp7	12	Carelessness by the hero results in a plight.
Tmp7^1	5	Carelessness by the hero places the hero or others in danger.
Tmp7^2	2	Carelessness by the hero acts as a violated interdiction.
Tmp7^3	5	Carelessness by the hero results in their expulsion.

Establishing this reportability requirement reduced the list of function-like elements we had identified to just two elements, which we call missing functions. Table 3.7 shows the identified missing functions and their subtypes. Tmp5 and Tmp7 are the parent functions, with the subtypes denoted by superscript. The numbering in the function designation is an artifact of the process by which candidate elements were generated. These missing functions have a substantial presence in the corpus: combined, they

occur in 25 of the 46 tales in Propp's corpus, with the individual counts of each missing function and subtype being listed in the table.

After finalizing the two missing functions, a pilot annotation was performed with Story Workbench in order to understand the extent of these functions in the text. Given this purpose, the annotation ignored some finer details of the full annotation process: inversions and signals were ignored and the annotation was done to cover the full extent of the function in the text.

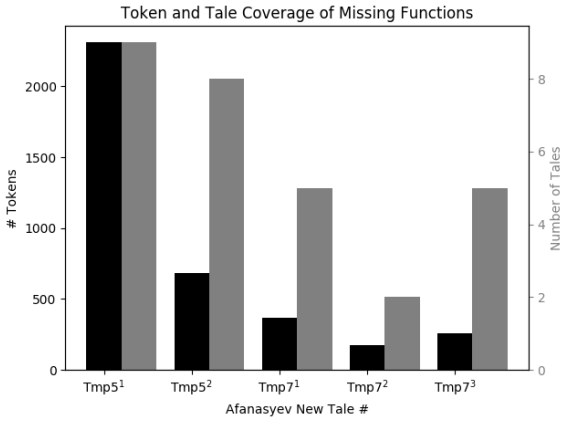

Figure 3.4: The coverage of each missing function subtype, shown in tokens (black) and number of tales (gray).

Figure 3.4 shows the number of tokens covered for each subtype of the missing functions, as well as the number of tales that each subtype occurs in. Missing functions cover a substantial amount of tales but do not cover a substantial amount of tokens. Notable is the coverage of Tmp5[1], one of the earliest candidates for missing functions, identified informally in our past work. At a glance, there appears to be a mild correlation between the number of tales and the number of tokens (as expected), but the sample size is too low for this result to be statistically significant. More interesting, however, is the large amount of tokens covered by Tmp5[1]. One potential reason for this is it may be the most common missing function that occurs in Propp's analysis: previous work identified it as one potential example of a missing function [Fin12, p. 80].

Looking at the corpus, we found that, on average, 32.5% of the tokens in a story were covered by functions, with 41 of the tales having less than 50% of the tokens

covered. The new functions increase the coverage by a relatively small amount: with missing functions, 37% of the tokens are covered, with 39 tales having less than 50% of the tokens covered.

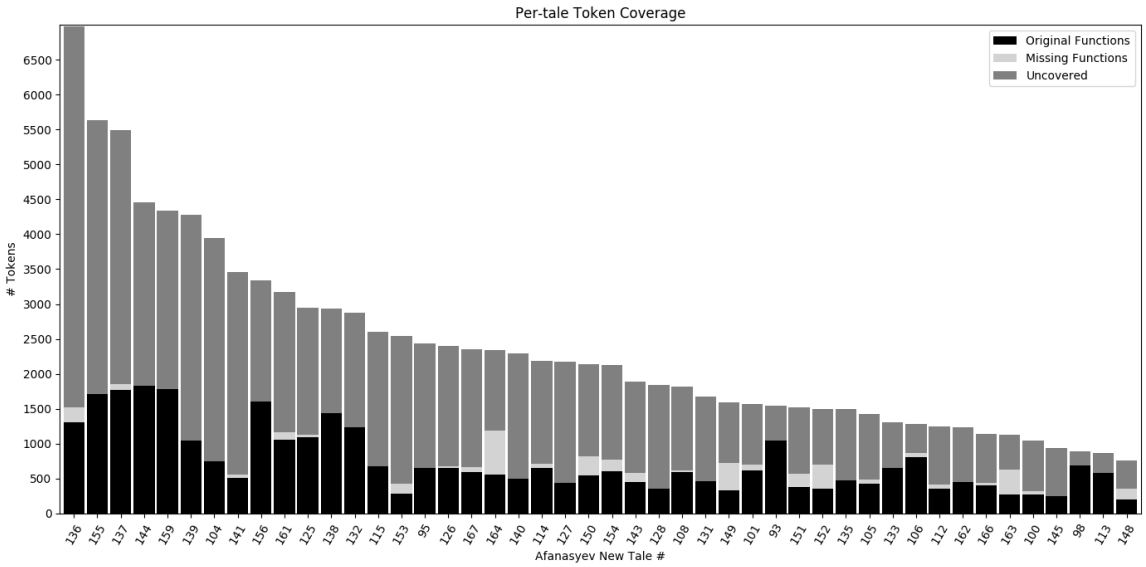

Figure 3.5: Bar graph showing the number of tokens covered by Propp's original functions vs. missing functions.

Despite the number of tales they occur in, it is obvious that these tokens do not cover much of the text. Figure 3.5 shows the total token coverage for the original functions (in black), the number of tokens covered by the missing functions (in light gray), and the number of uncovered tokens (in dark gray). Even after annotating the missing functions, substantial amounts of text remain uncovered: this may be unproductive text (e.g., dialogue, actions that produce no movement of the story) or text covered by unidentified functions.. There are many tales that have a substantial number of tokens uncovered: one example is tale #137 - Ivan the Bull's Son. In many tales, including Ivan the Bull's Son, much of the story is dialog, including banter. As discussed at the beginning of the section, unproductive narrative is to be expected, although the magnitude of it was not.

Figure 3.6 shows the same information as Figure 3.5 as a percent chart, for the reader to get a better idea of the relative amount of the stories that are covered by functions. The figure shows Propp's original set of functions (in black) and the percent of tokens covered by the missing functions (in light gray). The order, as in Figure 3.5,

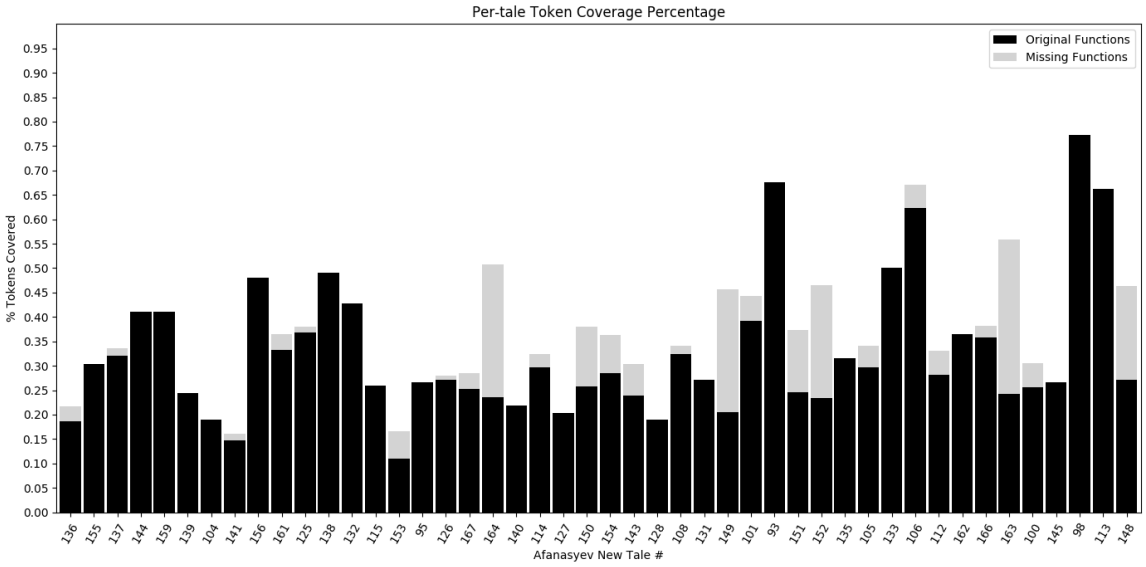

Figure 3.6: Bar graph showing the percent of tokens covered by Propp's original functions vs. missing functions.

is in descending size of the tale, to provide a consistent point of reference. Five of the 25 tales with missing functions gain at least 25% token coverage (#163, #164, #149, #152, and #148. Of these, two of the tales gain at least 50% token coverage: #163 and #164. The largest gains in coverage, percentage-wise, are in Bukhtan Bukhtanovich (#163) and Kozma Quickrich (#164): in part, this is to be expected, as both stories focus on a "trickster" archetype, which was identified in past work as a substantial missing element [Fin12, p. 80].

Our annotation and analysis of the missing functions present in the corpus that Propp did not describe in his original morphology is far from exhaustive. However, this surface analysis and identification demonstrates the utility of our approach to verifying the reliability of Propp's corpus via text annotation, as well as the utility of the corpus we have produced. Without such fine-grained annotations, it would be substantially more difficult to extract precisely which text sections were uncovered, and thus more difficult to examine what interesting elements might have been overlooked by Propp. Further, this analysis demonstrates that while Propp's morphology is a reliably, powerful tool for describing the structure of narratives, it is by no means complete or global.

3.9 Summary of Contributions

Propp's morphology has had a deep impact in narratology, narrative understanding, and computational approaches to narrative, especially in story generation. Propp's work has inspired and influenced a substantial body of work on narrative structure since its translation into English. ProppML, the scheme we present here, is a complete annotation scheme for Proppian morphologies, succeeding previous work on annotation schemes targeting Propp's work.

Our experiments show that Propp's functions can be applied and important characters can be identified with a very high degree of reliability by highly-trained human annotators. In addition, these highly-trained annotators can assign functions to moves and identify what role a character plays with an acceptable degree of reliability.

We have fully answered two of the core questions that we raised in this and previous work: first, we have shown, for both single-move and multi-move tales, that given Propp's general approach, list of functions, and functions identified in specific tales, independent annotators agree with each other as to where and whether those functions appear in the tales Propp suggests: $F_1^i = 0.742$, $\kappa_{sym} = 0.929$ for functions, $F_1^g = 0.562$, $B^{3,g} = 0.652$ for moves, and $F_1^i = 0.791$, $\kappa_{dp} = 0.448$ for *dramatis personae*.

Second, we have shown that given Propp's general approach and list of functions, independent annotators agree not only with each other, but also agree with Propp, when asked to find his functions in tales. For inter-annotator agreement, our results are: $F_1^i = 0.616$, $\kappa_{sym} = 0.787$ for functions, $F_1^g = 0.203$, $B^3 = 0.545$ for moves, and $F_1^i = 0.707$, $\kappa_{dp} = 0.289$ for *dramatis personae*. Per-annotator agreement with the gold standard annotation are good, with the best results (across annotators) being: $F_1^i = 0.636$, $\kappa_{sym} = 0.833$ for functions, $F_1^g = 0.276$, $B^3 = 0.572$ for moves, and $F_1^i = 0.862$, $\kappa_{dp} = 0.571$ for *dramatis personae*.

Comparing the adjudicated, best-effort merge of the closed annotations to the gold standard results in very good agreement measures: $F_1^i = 0.662$, $\kappa_{sym} = 0.879$ for functions, $F_1^g = 0.397$, $B^3 = 0.620$ for moves, and $F_1^i = 0.877$, $\kappa_{sym} = 0.548$ for *dramatis personae*. These results show that annotation of new texts, for which a list

of pre-identified functions will not exist, can be done with high reliability by highly-trained annotators.

Finally, we suggest that Propp's theory, as it stands, is broadly correct, but does not capture all elements present in the corpus of tales he annotated. We perform a preliminary analysis of the corpus, using the annotations produced from this study to analyze the uncovered areas, and hypothesize and present the elements that Propp may have missed.

CHAPTER 4

Human Detection of Narrative Elements: Motifs

4.1 Motivation

Motifs are, as I have belabored in the introduction, remarkably interesting and potentially useful sources of information and, in particular, *cultural* information that are used frequently within narratives. However, despite there being massive repositories of motifs in the form of the motif indices, this does not provide a guarantee that these motifs are meaningful to the groups from which they originate, merely that they occur in their literature. Neither does this guarantee that even when meaningful to in-group persons that they can reliably identify when a motif is being used in a *motific* manner: that is, being used to invoke the constellation of ideas that is expected to be associated with a motif. To that end, it is necessary to determine whether or not members of a cultural group can reliably identify motific usage of motif phrases with a high degree of agreement before creating a tool to automatically perform the same task. After all, if humans are unable to identify motifs they are deeply familiar with, the results of a tool doing so amounts to not much more than noise.

Further, for the development of a tool for the automatic detection of motifs, a relatively large collection of data is required: to this end, the annotation process can be combined with adjudication to produce a set of gold standard data, as none existed at the time of this work.

Thus, in short: to build a tool, both verification of the phenomenon and data are needed, thus necessitating annotation.

4.2 Approach

There were several necessary steps from start to completion of the annotation process:

- selecting the cultural groups from which to draw motifs,
- selecting the specific motifs,

- selecting and acquiring data,
- developing of an annotation guide and scheme,
- selecting an annotation tool and developing an annotation pipeline,
- sanity checking the annotation and testing the annotation pipeline,
- hiring annotators,
- annotator training,
- and adjudication.

I describe each of these in detail in the relevant sections below, explaining at each step the considerations that I and my collaborators kept in mind while making decisions.

4.2.1 Time and Money: the Cost of Annotation

This portion of work started March 2019 and lasted until September 2020, from the start of cultural group selection to the completion of annotation and adjudication. The annotation and adjudication itself took a total of 11 weeks and cost roughly $15,000.

4.3 Selection of Cultural Groups

Before selecting motifs, it was important to identify which cultural groups from which to draw motifs. As part of a larger team using this data, we had two necessary criteria: first, there needed to be a strong authoratative source of motifs (a motif index, folklore collection, or something similar)—this made it substantially easier to identify motif candidates that may be interesting. This task was split among myself, my colleague Anurag Acharya, and my mentee Diana Gomez. Our starting seed was Thompson's Motif Index [Tho60], which provided the names of many other motif indices, allowing us to quickly expand our search.

The second criteria was that cultural groups needed large populations near either FIU or our collaborators, SIFT. This restriction was deemed necessary for an ongoing

survey study that is part of the same effort this work falls under, to ensure that we were easily able to find participants (although world events quickly made these concerns obsolete).

The three cultural groups selected were Irish, Puerto Rican, and Jewish, each of which had a strong source to provide potential motifs. For Irish, I used T.P. Cross' "Motif-Index of Early Irish Literature" [Cro52] as a main source. For Puerto Rican, Diana Gomez drew motifs from S.R. Lamarche's "The Mythology and Religion of the Tainos," [HdAL21], R.E. Alegría's "The Three Wishes: A Collection of Puerto Rican Folktales" [AAHC69], and J. Ramírez-Rivera's "Puerto Rican Tales: Legends of Spanish Colonial Times" [RRKS77]. For Jewish motifs, Anurag Acharya referenced D.N. Noy's "Motif-index of Talmudic-Midrashic literature" [Noy54]. Additionally, each of these cultural groups had a strong presence in either Miami or Boston (where a collaborator was located), fulfilling the second criteria.

4.4 Selection of Motifs

Once these groups were selected, individual motifs needed to be identified from within these groups. This was necessary as there are a substantial number of potential motifs in existence: Thompson's motif index lists over 46,000, drawn from many different cultures. To draft a tractable list of motifs for annotation, I developed three criteria for how to select these motifs:

1. **Motifs must be commonly used**: the simplest test of this was to do simple searches to see if the motif was used either on social media, such as Twitter, or in the news. This criteria was intended to simplify the process of looking for motif usage—if we couldn't find it, there was no point in including
2. **Motifs must have a source within the cultural group**: by *source*, I mean an associated, well-known story within the same body of folklore as other motifs in the cultural group. This criteria is intended to provide proof of relevance for the motif to the cultural group in question. If a motif had no definitive source

within the folklore of the cultural group we drew it from, there was a possibility that it was a motif that was now common, but drawn from other cultural sources.

3. **The motifs needed to have a high potential strength**: this was a subjective measurement based on *how* the motifs were used when found in social media or news. If a motif seemed to be used in a way to allude to a greater idea rather than a simple reference or usage as a name, this suggested a high potential strength.

These criteria were used by the group working on this project to draft a selection of motifs: I handled Irish motifs, and again Anurag and Diana handled Jewish and Puerto Rican motifs, respectively. During the selection of motifs, we aimed for a total of 30 motifs, roughly 10 from each culture.

4.4.1 List of Motifs

From the initial selection phase, the following motifs were chosen (with definitions for each motif provided in the appendices):

Irish The Salmon of Wisdom, Finn McCool, leprechaun, King Conchobar, aos si, banshee, Cu Chulainn, the wren, the magic harp, tir na nog, shamrock, fairy fort, and the children of lir.

Jewish Haman, golem, Amalek, babel, leviathan/behemoth, 70 languages, name in vain, milk with meat, the ark of the covenant, and kiddush.

Puerto Rican Reyes Magos/Three Kings, Agueybana, Atabey, Roberto Cofresi, Divina Providencia, Guanina, Juan Bobo, Yocahu, the coqui, and Hormigueros.

This is a total of 33 motifs: 13 Irish, 10 Jewish, and 10 Puerto Rican. There were many motifs that simply didn't make the cut for a variety of reasons. For example: the Jewish motif "sukkot" was eliminated for having no apparent motific use, only direct references; the Irish motif "king of cats" was rejected for having no clearly identifiable source; and the Puerto Rican motif "three camels come for grass on January 5th" was removed for being too difficult to find examples of.

Eventually, given the dataset that was being used, this list of motifs was insufficient for the completion of the Puerto Rican annotation, with too few articles to continue annotating. As such, the annotators were asked to provide additional motifs that they felt were relevant, resulting in the addition of five additional motifs:

Puerto Rican jibarito, guaraguao, pitirre, chupacabra, and pava.

These five motifs were used for the final two batches, for a total of roughly 1500 examples. More detailed information on these motifs can be found in Appenix **??**.

4.5 Selection and Acquiring of Data

Selection of data was done through NexisUni, a university-focused version of LexisNexis, a tool for searching through articles. This tool was chosen due to the group's previous familiarity with it from identifying the common presence and strength of motifs from the motif selection phase. The process involved searching for motif-related terms using NexisUni and batch downloading these articles. These articles were then further processed by my own Lucene-based lexical matcher, with fuzzy rules for a variety of lexical forms for each motif, to verify the presence of motifs and produce initial tags in the brat standoff annotation format for use by the annotators. In total, myself, Anurag, and Diana downloaded over 100,000 articles, of which 7,946 were annotated. All of the data was collected in English, as there are substantially more NLP tools available for English than other relevant languages.

4.6 Aside: What is Text Annotation?

Before continuing to describe the steps in demonstrating human detection of motifs, it is necessary to define what I mean when I mention annotation. By *annotation*, I refer to linguistic annotation, which "covers any descriptive or analytic notations applied to raw language data" [BL01]. Consider the following simple example wherein the bolded text is the motif candidate I am interested in:

Figure 4.1: What an annotator sees while annotating using brat.

> As the report put it, the state has too many "little box" governments: 66 counties, 1,018 cities and boroughs, and 1,546 townships. A political **Tower of Babel**.

In this particular example, annotators must decide what label to assign to the phrase "Tower of Babel," and in this case, both annotators for this example have determined it to be *motific*. From the annotator's point of view, using the brat rapid annotation tool [SPT+12], they see the screen in Figure 4.1.

Annotations are pre-tagged with their motif label; to simplify the annotation process, annotators were instructed to leave the tag if the example was motific and to select the correct tag otherwise.

4.7 Annotation Guide and Scheme

The annotation guide, provided in Appenix **??**, describes what annotation is, the purpose, the idea of what a motif is, the procedure, the tool itself, and provides a basic catalog of the motifs.

Additionally, the guide contains a section that was heavily revised as the annotation proceeded: "Special Cases and Specific Considerations," which was used to list any decisions or information from the adjudication sessions that I felt needed to be cached. Discussions that were not a simple resolution of a disagreement or correction of a mistake, but resulted in a decision about how specific *classes* of annotations should be handled were noted in this section.

The annotation guide itself, in particular the structure and description of annotation, is heavily based off of annotation guides developed by my advisor, Mark A. Finlayson, during his Ph.D. work.

The annotation scheme was tested over the course of many pilot annotations done with members of both FIU and SIFT: among those that helped are Anurag Acharya, Diana Gomez, Diego Castro, and Laurel Bobrow. Initially, I suggested three categories for motif candidates: unrelated, referential, and motific. However, the pilot annotations quickly demonstrated an edge case: for example, the metal band "Behemoth" is named after a beast from Jewish folklore known to be monstrous. Imagine I say the following:

> I'm excited to go to the **Behemoth** concert later this month!

How should the usage of "behemoth" in this context be annotated? It's clearly not unrelated—it draws its name from the source folklore. It's clearly not motific—there is no attempt to invoke any of the ideas associated with the motif. It's also not referential—there is no discussion of the source nor the definition of the motif. For motifs that are impactful within a culture or have spread beyond it, there is an additional class of *things named after motifs*. This necessitated the addition of a fourth category: eponyms. I define each of the four categories as follows:

Unrelated A usage that is unrelated to the cultural group or cannot be established as directly related (e.g., "behemoth" as a monster in a game).

Referential A usage that directly refers to the folklore origin of the motif or its definition (e.g., discussing the "behemoth" origin).

Eponymic A usage that references the motif in a name—this distinction is made because while it is highly similar to motific usage, it is typically not used as such (e.g., the band "behemoth" may be referred to with no additional meaning beyond the band).

Motific A usage that intends to invoke the cultural associations of a motif (e.g. referring to something large and monstrous as a "behemoth").

The annotation scheme is remarkably simple, thanks to brat, my selected annotation tool: from the user side, they see the interface shown in Figure 4.2. From my

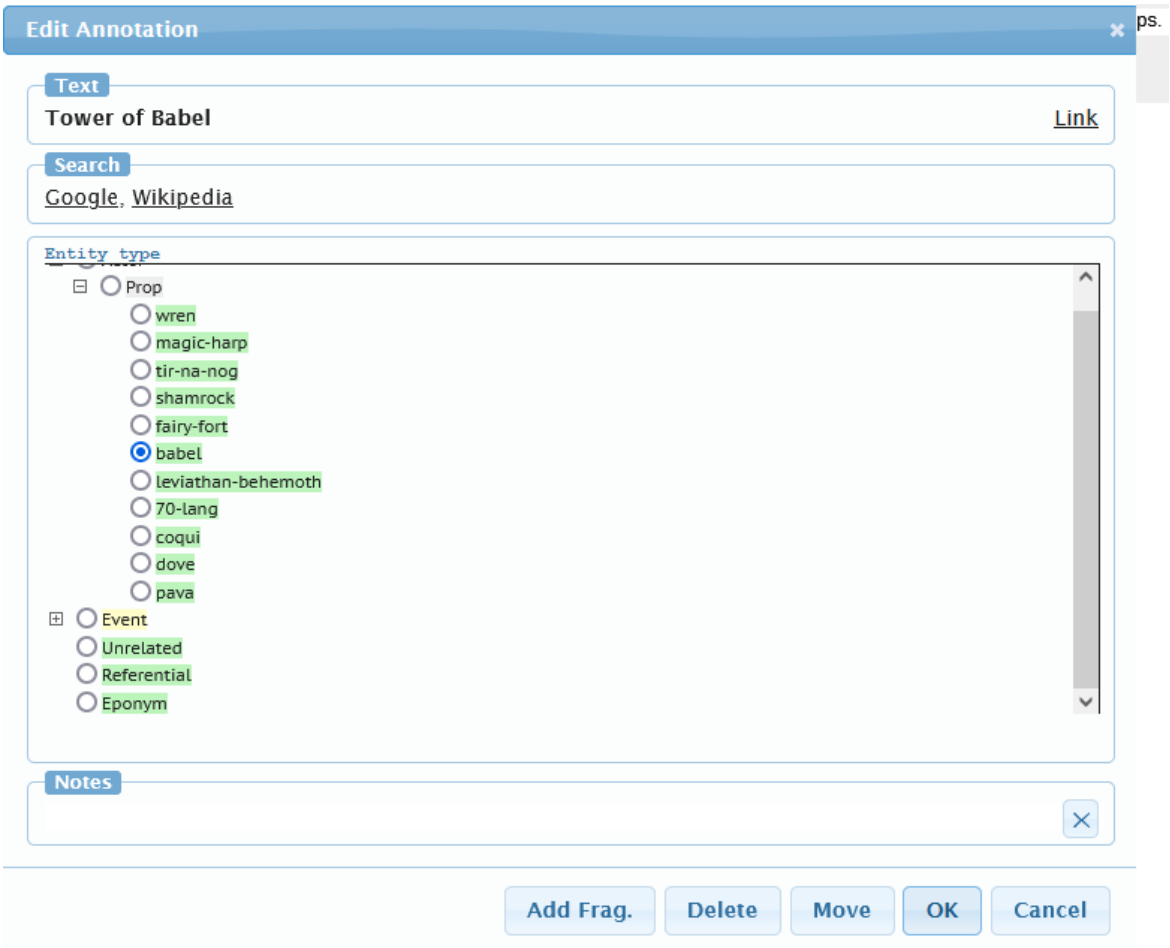

Figure 4.2: What an annotator sees while modifying an annotation using brat.

side, I get the brat standoff annotation format, which is relatively simple:

`ID<tab>label starting-token ending-token<tab>raw-text`

An example of what a brat file looks like can be seen in Figure 4.3.

4.7.1 Aside: What is the difference between a motif, a motif candidate, and motific usage?

When I use the term **motif**, I am referring to the idea of a motif: for example, a *trickster god* or a *magic sword*. When I use the term **motif candidate**, I am referring to a span of text (such as the italicized portions of the previous sentence) that matches the lexical form of a motif. When I use the term **motific usage**, I am referring to a

```
T1      Eponym 334 348  Tower of Babel
T2      Eponym 884 898  Tower of Babel
T3      Referential 1413 1427   Tower of Babel
T4      Referential 1942 1956   Tower of Babel
T5      Eponym 2376 2390        Tower of Babel
T6      babel 2921 2935 Tower of Babel
T7      Referential 3477 3491   Tower of Babel
T8      Referential 4033 4047   Tower of Babel
T9      babel 4567 4581 Tower of Babel
T10     Referential 5101 5115   Tower of Babyl
T11     Unrelated 5646 5657     0 languages
T12     babel 6188 6202 Tower of Babel
T13     Referential 6739 6753   Tower of Babel
T14     Referential 7290 7305   Towers of Babel
T15     leviathan-behemoth 7786 7794    behemoth
T16     leviathan-behemoth 8384 8392    behemoth
T17     Referential 8917 8931   tower of Babel
T18     Referential 9456 9470   tower of Babel
T19     babel 10022 10036       Tower of Babel
T20     babel 10588 10602       Tower of Babel
T21     babel 11147 11161       tower of Babel
T22     Referential 11706 11720 tower of Babel
T23     Referential 12267 12281 tower of Babel
T24     babel 12686 12700       tower of babbl
T25     Eponym 13048 13062      Tower of Babel
T26     babel 13608 13622       tower of Babel
T27     babel 14174 14189       towers of Babel
T28     babel 14741 14756       towers of Babel
T29     leviathan-behemoth 15264 15272  behemoth
T30     leviathan-behemoth 15841 15849  behemoth
```

Figure 4.3: An example of a brat file.

motif candidate that is not eponymic, referential, or unrelated—simply put, a motif candidate whose usage invokes the cultural associations of the motif.

4.8 Selecting an Annotation Tool and Developing an Annotation Pipeline

Selecting an annotation tool was a relatively simple matter: while I explored several tools, including my lab's internal tool Story Workbench [Fin11a], I eventually settled on brat [SPT+12] as the simplest and easiest to deploy tool for the annotation.

However, displaying full articles was inefficient—some articles may contain a single motif, or have motif candidates spread far apart. I decided to instead show snippets of articles for context and display multiple sections with motif candidates per annotation file. To enable this, it was necessary to develop tools to combine annotation files, split them apart, and calculate the agreement. I worked with my colleague, Anurag, to develop these tools in Python.

4.9 Sanity Checks: Annotation and Pipeline

As previously described, simple sanity check annotations were performed in the group to ensure that the pipeline could go from individual raw text files and lexical matcher annotation files to the combined format, back, and calculate agreement. These annotations took place using small batches of around 100 samples each and also helped to refine the annotation guide and scheme (for example, as stated earlier, the addition of the eponym class).

4.10 Hiring Annotators

For the purposes of this annotation study, it was necessary to hire annotators, a process which was done with the approval of and close advisement of the PIs in charge of the ACUMEN project, which this work was a part of. We decided to hire annotators who were considered cultural experts. We went through a temp agency, Robert Half, and were billed $20.80 per hour. We interviewed each candidate provided by the agency and required that they identify as belonging to one of the three cultural groups from

which we drew motifs, possessed or were in the process of completing a college degree, and were fluent in English. Annotators were asked about their technical abilities, their access to a computer and internet connection, their qualifications, language abilities, and their connection to the cultural group. We also assessed their apparent interest in the project itself (e.g., did they ask questions about the project or seem excited to participate), under the belief that interested annotators would perform better than disinterested annotators. In total, we hired three sets of two annotators to perform the double-blind annotations.

4.11 Annotation

4.11.1 Annotator Training

I gave annotators a single two-hour session of training that included reviewing the annotation guide, covering any questions or concerns, and running through a small sample annotation together as a team. Further, I held a two-hour adjudication session with each pair of annotators every week to cover the week's annotations: these served to help reinforce the annotator's skills.

4.11.2 Annotation Procedure

Annotation was done in a double-blind manner, as annotators were asked to perform their annotations independently of each other with no contact outside of the weekly adjudication session. Annotators were allowed full access to the annotation guide during their annotating and were free to annotate at their leisure so long as the week's task was completed. Annotators were, however, limited to a total of 10 hours, at my advisor's suggestion: as annotation can, at times, be a tedious task, we believe that more than 10 hours a week of annotation would reduce the amount of focus the annotators were able to muster and lower the quality of the annotations.

Annotation batches started at 300 examples for the first week and ramped up to 1,000+ by the final weeks; the exact numbers varied depending on the articles selected for the week, as articles were not split apart.

Annotation took a total of 11 weeks, although not all groups participated for the full 11 weeks: the Irish group reached a substantial level of agreement (Fleiss' $\kappa > 0.55$) on the fifth week of annotation, with the Puerto Rican team reached this on the third week and the Jewish team reaching it on the second week; the Jewish team participated for 9 weeks and the Puerto Rican team participated for 10 weeks.

4.11.3 Adjudication

Adjudication was a relatively simple process: I sat down with both annotators in a pair to discuss the previous batch of annotations done. We focused solely on disagreements and I allowed the annotators to come to a decision on the correct annotation except for times when I was asked for my input. These sessions could last as little as 30 minutes or up to the allotted two hours in cases where there was a high degree of disagreement. Any issues with the data were addressed at these meetings and any substantial decisions about annotations (e.g., special cases) were recorded in the annotation guide and distributed to all six annotators.

4.12 Annotation Results

To calculate the agreement between annotators, I use Fleiss' kappa [Fle71], which generalizes Cohen's kappa [Coh60] (a statistic to measure agreement between a pair of raters) to measure agreement to any number of raters. Fleiss' kappa (F_k) was intentionally chosen for use in the pilot runs, which had up to five participants, and maintained throughout to provide a point of comparison for sanity check purposes.

The Jewish and Puerto Rican teams participated produced 7,000 annotations with an average agreement of $F_\kappa > 0.7$ while the Irish team produced 7,000 annotations with an average agreement of $F_\kappa > 0.55$. In Table 4.1 I provide a full listing of the

Table 4.1: The week-by-week agreement in Fleiss' kappa of the annotation process.

	Irish		Jewish		Puerto Rican	
Week	F_k	#	F_k	#	F_k	#
1	-0.18	379	0.07	363	0.19	326
2	-0.05	536	0.579	554	0.518	440
3	0.00	881	0.638	912	0.552	864
4	-0.006	861	0.739	895	0.68	838
5	0.559	863	0.802	904	0.731	887
6	0.699	978	0.821	992	0.725	977
7	0.61	970	0.822	1349	0.798	1013
8	0.633	988	0.652	984	0.817	992
9	0.557	1047	0.779	971	0.738	922
10	0.477	1013	-	-	0.748	1454
11	0.429	1174	-	-	-	-

per-batch results. The Irish group had the most difficulties in achieving the baseline of $F_k > 0.55$, the cutoff selected for the purposes of data collection. While at various points in time the agreement dips for all three groups (especially for the Irish team, who fell under 0.55 for the last two batches), the overall average of the data never fell under 0.55 for Irish team nor under 0.7 for the Jewish and Puerto Rican teams.

4.13 Discussion

I found that this study demonstrated, first and foremost, that humans can reliably identify motific usage of motifs in text and distinguish these from other usage. Further, I believe this study demonstrates the necessity for identifying strong sets of motifs and the difficulty in ensuring that the strong presence of a motif is coherent with its cultural source—this is likely the cause for the higher agreement between the Jewish group than the Irish group.

4.13.1 Jewish Team Agreement

Jewish team had consistently high agreement throughout. Many of the motifs used, such as Amalek or Haman, had very specific and distinct meanings that are not present in a broader cultural audience, which I believe is responsible for this high agreement.

4.13.2 Irish Team Agreement

One note is that the Irish team agreement dipped in the final two annotation batches. There are a few potential causes of this: (1) some of the less common, but more distinctive motifs began to disappear (e.g., there were no more articles containing the motif "Children of Lir" after a certain point), which left motif candidates that had spread to a broader audience and thus were less clear in their usage; (2) the Irish annotation lasted the longest by far, which could have had an impact on annotator morale. I believe that both of the reasons likely contributed: many Irish motifs, when they become more common, have a diluted meaning—for example, the "leprechaun" is viewed in a more negative light in Ireland than in its usage in North America; however, the lengthy annotation process no doubt reduced annotator performance, and dips in performance can be seen in all three groups as they reached the end of the annotation period.

4.13.3 Puerto Rican Team Agreement

The Puerto Rican team also experienced a dip in the last two batches; however, one note is that the Puerto Rican annotators had new motifs, suggested by them, introduced in these batches, which is a likely contributor to this.

4.14 Motifs as a Function of Genre

An additional result made possible as a result of the annotation was measuring the frequency of motifs in editorial or op-ed articles when compared to other articles.

The broad usage of motifs suggests their cognitive important as touchstones of cultural knowledge and their cultural relevance hints at their importance for pieces intended to represent an opinion or convince other of an opinion: for example, editorial articles. Thus, I expect that in editorial articles, as compared to non-editorial articles, motifs would occur more frequently.

Table 4.2: Comparison of motif frequency per article, sentence, and token between editorial and non-editorial articles.

	Editorial/Opinion	Non-Editorial/Opinion	Ratio of Ed. vs. Non-ed
Motif/Article	0.7607	0.2026	3.75
Motif/Sentence	0.0185	0.0061	3.03
Motif/Token	0.0008	0.0003	2.67

Of the 7,946 articles that were annotated, 5,109 had either editorial tags or other genre tags; the remaining 2,678 articles did not. Using a sentence-level opinion classifier [YH03] modified by a colleague, Joan Zheng, to operate at the document level that performed well on the already-categorized data ($F_1 > 0.90$), we re-categorized these articles as either editorial or not, resulting in a total of 117 editorial and 7,829 non-editorial pieces.

Calculating the rate of motifs per article, sentence, and token, I found that motifs were roughly three times as frequent (3.75x, 2.95x, and 2.93x, respectively) in editorial articles than in non-editorial articles. The detailed results of this experiment are present in Table 4.2.

I hypothesize this difference in frequency is due one of several potential factors: (1) editorial articles take a more casual form of discourse in comparison to articles written to report on an event or topic; (2) editorial articles are crafted to appeal to a certain audience; (3) editorial articles are more likely to rely on emotional appeal; or (4) editorial articles are arguing from a specific stance and more likely to use powerful rhetoric devices. I believe that these results strongly suggest the important of motifs for understanding human communication.

4.15 Summary of Contributions

One of the major contributions of this work is a large-scale annotation study which serves to demonstrate exactly that—that human annotators can reliably agree with a second, independent annotator on how to annotate candidate motif phrases with a high degree of agreement ($F_k > 0.55$). This suggests the importance of motifs generally

and confirms that they have meaning that can be reliably identified and extracted by members of a specific group.

Additionally, this annotation study resulted in 21,123 annotations after agreement had reached a substantial level, which were used for the development, training, and testing of the system described in §**??**.

Finally, this annotation study allowed for the identification of a trend in editorial vs. non-editorial articles with respect to motifs, hinting at their importance especially in the field of understanding articles based around human-to-human communication and opinion discourse than factual reporting.

CHAPTER 5

Preliminary Detection Work: Discourse Function in News

5.1 Motivation

Similarly to how Propp's morphology describes the form of narrative and motifs describe the content, discourse structure describes the purpose and content of given portions of a narrative. This work focuses primarily on non-fiction, by leveraging Teun A. van Dijk's theory of news discourse [vD88]. However, discourse structure can also be applied to fictional narratives; further, while the annotation study I conducted as part of this work was done at the paragraph level, discourse structure can be analyzed at all levels of text.

As I described in the introduction, discourse structure is a vital part of text and communication at large: purpose, topic, location, and more may be captured through a description of the discourse structure of a given piece of text. Discourse structure is important across a diverse set of domains and is deeply relevant for computational applications such as summarization, information extraction, question answering, and more.

In this work, previously reported in the Workshop on Events and Stories in the News [YCGF18], I apply an established hierarchical theory of news discourse [vD88] to model how paragraphs operate as units of discourse within news articles to capture the importance of events within a story. Using this model, I tested two hypotheses: first, that humans can reliably annotate news articles with van Dijk's theory; second, that these discourse labels can be predicted by machine learning.

5.2 Van Dijk's Theory of News Discourse

Van Dijk's hierarchical theory of news discourse [vD88] is shown in Figure 5.1, which I apply to a subset of the news articles of the ACE Phase 2 corpus. The model is composed of ten leaves; here, I briefly describe each of these categories as well as their

parent categories as appropriate. I provide further annotation detail for discourse types where I believe van Dijk's description was underspecified, as done in the guide provided to my annotators.

Summary elements express the major subject of the article, with the *headline* being a special construct that introduces a topic, and the *lead* summarizing the topic introduced by the headline. While annotators were initially instructed to annotate the headline, I do not include it in the annotations, as the ACE Phase 2 corpus has the headline separate as part of its annotation scheme.

Situation elements are the actual events that comprise the major subject of the article. *Episodes* concern *main events*, which are those events that directly relate to the major subject of the article, and the *consequences* of those events. The *background* consists of the *context*, which are any *circumstances* that contribute to understanding the subject as well as any *previous events*. Where circumstances may be non-specific, previous events refer to a specific event that has occurred recently. *History* elements are those events that have not occurred recently, typically referenced in terms of years prior, rather than months, weeks, or days. These elements of the discourse structure provide important information about the relation of each paragraph with respect to the central events of a news story.

Conclusions are those *comments* made by the journalistic entity (the newspaper, reporter, etc.) regarding the subject. These can be *expectations* about the resolution or consequences of an event, or *evaluations* of the current situation. In contrast, *verbal reactions* are *comments* solicited from an external source, such as a person involved in the events of the article, an expert, etc. These elements of the discourse provide further supporting context for the central events of an article.

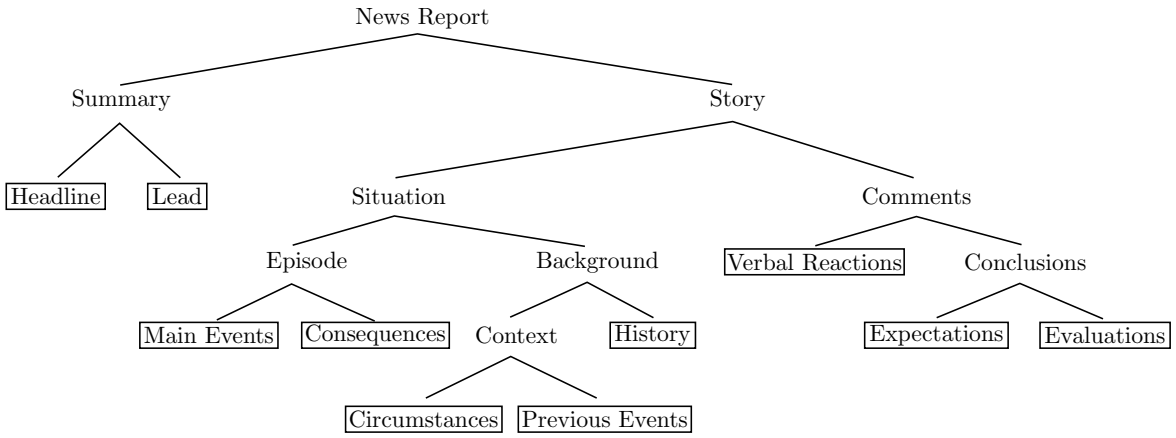

Figure 5.1: Van Dijk's hierarchical discourse structure of news.

5.3 Data & Annotation

One of the major contributions of this work, as a result of the annotation study, was the production of a gold-standard corpus of paragraph-level discourse structure based on the ACE Phase 2 corpus. This dataset comprises 50 documents containing 28,236 words divided in 644 paragraphs. This annotation is, to the best of my knowledge, the first of this kind, and provides additional information about the corpus that, until now, is not considered in any knowledge extraction method.

5.3.1 Selection of Texts

I selected the ACE Phase 2 corpus because it is a major standard corpora of news articles that satisfied three criteria: it is widely-used, has relevance to other tasks, and was readily available to me. For annotation, 50 articles were randomly selected from the development set, divided into ten sets of five documents each. Within these sets, documents were swapped or replaced to obtain a uniform length for each batch of articles. Table 5.1 shows the corpus-wide statistics for the number of words and paragraphs, where each paragraph is given a single type in accordance to van Dijk's theory. Paragraphs in the majority of documents were either already clear due to lexical markers such as empty lines or indentation. The remaining texts were divided by the adjudicator based on either contextual or structural clues, such as abrupt change in topic or unnatural line breaks.

	Words	Paragraphs
2-3 Total	28,236	644
Average	564.7	12.9
Standard Deviation	322.1	4.9

Table 5.1: Corpus-wide statistics on the relevant lexical features for annotating the news articles.

5.3.2 Annotation

Annotation was done in a double-blind manner by three annotators, one of whom also acted as the adjudicator. All three annotators were Ph.D. students in computer science with a focus on natural language processing, with experience in both annotating and running annotation studies.

Annotator Training

In contrast to other studies done as part of this dissertation, annotators taking aprt in this project were given minimal training. A single adjudication meeting was held after annotation for the first two sets of documents was completed. The primary purpose of this adjudication meeting was to resolve any questions the annotators had, discover any uncertainty in the annotation guide, and revise the annotation guide to address these questions. The annotation guide contains descriptions of each discourse label in addition to an example of a fully-annotated news article, shown in Figure 5.2, parts of which are omitted for brevity. Annotation was primarily performed using the comment system present in the popular word processor, Microsoft Word. While this process is highly non-standard, it was a program all three annotators had and sidestepped the need for selection of or creation of a tool for this annotation task.

Annotation Procedure

Annotation was performed over the course of a month, as the annotator's time and schedule allowed. However, each set took roughly 45 minutes to an hour, resulting in roughly ten hours of work for the adjudicator (who performed annotation on all ten sets of documents) and six hours of work for the other two annotators. The division of

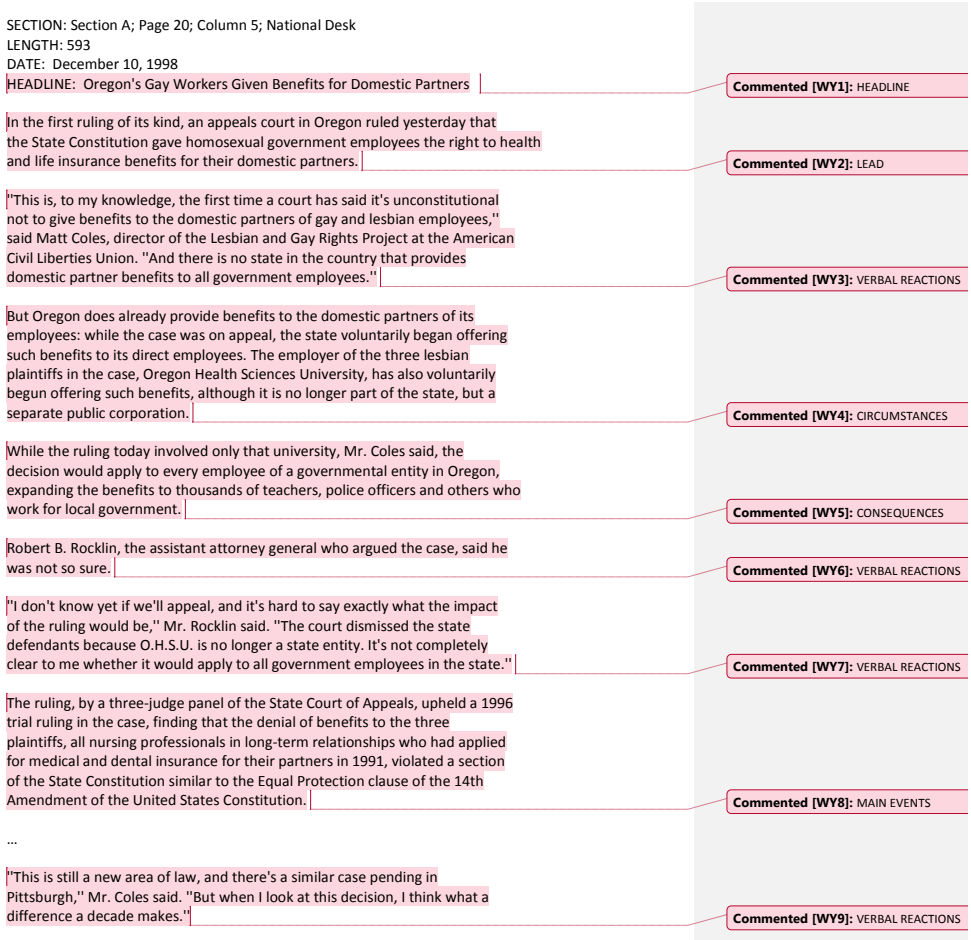

Figure 5.2: Example annotation included in the annotation guide.

the annotation batches is shown in Figure 5.3. When confronted with multiple labels that seemed to fit, annotators were instructed to choose the label that seemed the most applicable.

Figure 5.3: Division of work for the annotation study.

The adjudication procedure took a further hour for each set of documents, resulting in another ten hours of work for the adjudicator and another two hours for the other two annotators, who were only required to participate in adjudication of the first two sets of documents. The purpose of this group adjudication meeting was to resolve any outstanding questions or confusions regarding the annotation procedure. The annota-

tion resulted in triple annotation for the first ten documents, and double annotation for the remaining forty documents. The multiple annotations were merged into a gold standard for every document. Additionally, although annotators were instructed to annotate the headline for each document, these labels are not included as part of the gold standard because within the ACE Phase 2 dataset, the headlines themselves are clearly annotated.

5.3.3 Annotation Results

This annotation study had two goals: first, to evaluate whether or not humans can reliably apply van Dijk's theory of discourse to real news articles; second, to produce a benchmark dataset of document-level discourse annotations for machine learning applications. As a reminder from the previous sections on annotation, by *reliable*, I mean that annotators have a high degree of agreement with respect to each other. To measure agreement, I use the standard F_1 score [vR79], treating one of the annotators as the correct labels, as well as Cohen's kappa coefficient for inter-rater agreement [Coh68].

The results of the annotation study are shown in Table 5.2. Inter-annotator agreement between annotators A1 and A2 was measured over ten documents; inter-annotator agreement between the annotators and the adjudicator, as well as the annotators and the gold standard, was measured over 30 documents. The comparison between the adjudicator and the gold standard was measured over the entire collection of 50 documents.

The results of the annotation study are shown in Table 5.2, which contains the microaveraged agreement measures between the annotators (A1, A2), adjudicator (Adj.), and the merged gold standard (Gold). These results include precision (P), recall (R), balanced F-measure (F_1), relative observed agreement among raters (p_0), probability of chance agreement (p_e), and Cohen's kappa (κ, derived from p_0 and p_e).

Comparison	# Docs	P	R	F_1	p_0	p_e	κ
A1 vs. A2	10	0.76	0.79	0.77	0.63	0.18	0.55
Adj. vs. A1	30	0.81	0.85	0.83	0.71	0.19	0.64
Adj. vs. A2	30	0.80	0.83	0.82	0.69	0.18	0.62
A1 vs. Gold	30	0.93	0.92	0.92	0.86	0.19	0.83
A2 vs. Gold	30	0.92	0.90	0.91	0.83	0.19	0.80
Adj. vs. Gold	50	0.93	0.87	0.90	0.81	0.18	0.77

Table 5.2: Agreement for news discourse annotation.

Table 5.3 provides the distribution of van Dijk's labels (sans headlines, of which there are 50: one for each document, annotated within the ACE Phase 2 corpus). The majority of paragraphs fall under the categories of verbal reactions or circumstances.

Label	Count
Lead	42
Main	60
Consequences	19
Circumstances	103
Previous Events	64
History	27
Verbal Reactions	252
Expectations	21
Evaluations	56
Total	644

Table 5.3: Distribution of the labels within the annotated corpus.

Discussion

I observed that the inter-annotator agreement between the adjudicator and the individual annotators is high ($F_1 = 0.8$, $\kappa = 0.6$). Moreover, the results in Table 5.2 indicate that annotators, even with minimal training, can reliably apply van Dijk's theory.

Inter-annotator agreement between the two annotators is also high, although lower than agreement with the adjudicator ($F_1 = 0.75$, $\kappa = 0.55$). One possible reason is that the adjudicator was also responsible for the annotation guide: since the adjudicator is the source of the initial examples and instructions for annotation, is reasonable that the annotators would agree more strongly with the adjudicator than with each other.

Comparisons with the gold-standard are included for completeness: the all-around high agreement with the gold standard ($F_1 = 0.85$, $\kappa = 0.75$) demonstrate that the gold-standard is not dominated by a single annotator.

Although the distribution of labels is highly skewed, I find that this is roughly in-line with the style of reporting featured in the ACE Phase 2 corpus, which seeks comments and analysis from experts within the field as well as explaining the immediate context that has an effect on the main event.

5.4 Discourse Label Prediction

I built on top of my annotation study to demonstrate the automated learning of document-level discourse on a per-paragraph basis. I used machine learning algorithms included in Scikit-learn [PVG+11] for my classifiers: in particular, I use the SVM, decision tree, and random forest models. I include decision tree and random forest results despite their lower performance because they are particularly interesting for this experiment, as the theory itself is hierarchical. In addition to features from Scikit-learn, I also use the paragraph vectors [MSC+13] implementation in Gensim [ŘS10].

5.4.1 Feature Selection

In this section, I briefly describe the features I use and explain my rationale behind them.

Bag of Words I use Scikit-learn's `text.CountVectorizer` class with the standard English stopwords to provide a count of the tokens in each paragraph. This feature was selected based on the idea that paragraphs from different types of discourse would use different language.

TF-IDF I use Scikit-learn's `text.TfidfTransformer` class with standard parameters. TF-IDF was selected as one method to approximate topics within a given paragraph.

Paragraph Vectors I use Gensim's `models.doc2vec.Doc2Vec` class using the distributed bag-of-words model, with a minimum α of 0.01, a minimum word occurrence of five, and 50 steps (`dm=0, min_alpha=0.01, min_count=5, steps=50`). I use this as a second method of approximating the topic of a given paragraph.

Previous Paragraph's Label I also include the label from the previous paragraph. This feature is based on the idea that there is, to some degree, some sequential ordering or restriction in discourse type. One simple example is that a lead paragraph is never followed by another lead paragraph.

The bag of words, TF-IDF, and paragraph vector models are built across the entire training corpus and roughly measure what topics and words correspond to specific label types.

5.4.2 Results

My best experimental results were obtained using grid search to maximize the micro-averaged performance of the classifier, as measured across five folds. The SVM classifier uses a linear kernel with $C = 10$ and the class weights balanced based on the training data; the decision tree classifier uses the default parameters with the class weights balanced; the random forest uses 50 estimators with balanced class weights.

Feature Groups	P	R	F_1
Baseline #1 (Most Freq. Class)	0.39	0.39	0.39
Baseline #2 (SVM + Bag of Words)	0.46	0.46	0.46
Decision Tree	0.41	0.41	0.41
Random Forest	0.43	0.43	0.43
SVM	0.54	0.54	**0.54**

Table 5.4: Results from label prediction using SVM. All results are micro-averaged across instances, including precision (P), recall (R), and balanced F-measure (F_1). For the final three classifiers, all four features are described in §5.4.1.

Table 5.4 presents the results from my experiments, showing that my classifier is a substantial improvement over the most-frequent-class and bag-of-words baselines. I note that because these features are fairly generic, and do not include potentially

more informative and semantically and syntactically rich features (such as, e.g., event-, coreference-, dialog-, or discourse-specific features), these results give me hope of much better performance with further experimentation.

Label Type	F_1
Lead	0.87
Main	0.23
Consequences	0.13
Circumstances	0.46
Previous Events	0.18
History	0.05
Verbal Reactions	0.76
Expectations	0.08
Evaluations	0.19

Table 5.5: Per-label F_1 results. Best performance occurs for the lead, circumstances, and verbal reactions.

Table 5.5 presents the per-label results from my experiments. The relatively strong performance on *circumstances* and *verbal reactions* is not surprising, given their predominance. Similarly it is not surprising that I have low performance on labels that occur, on average, about once a document. I observe an unexpected level of performance on *lead* paragraphs, given their relative scarceness in the dataset. Within the data, I find that leads, with a single exception, occur at the start of the document: this accounts for the high performance, given that the first paragraph's previous paragraph is represented as -1, allowing the classifier to take advantage of their strong positional tendency.

5.5 Discussion

I find that using my SVM classifier, I achieve reasonable performance (65% of human performance). I suspect that an increase in performance can be gained by additional feature engineering. Moreover, I expect that including high-precision rule-based prediction will further improve the performance of the system: this is based on comments during adjudication from annotators, who stated that they relied heavily on lexical

clues such as quotation marks and specific words ("said," "commented," etc.) to select certain categories (in this case, *verbal reactions*).

While I expected the tree-oriented methods—decision trees and random forests—to outperform the SVM classifier, this was not the case in practice and they were outperformed by one of the baselines. I believe that this is because the features currently used fail to capture the higher-level semantic ideas that van Dijk used to group together the discourse types. While people understand that verbal reactions from experts, expectations, and evaluations are all types of comments, my current features do not capture these relations.

5.6 Additional Work

Further work on this was done by Banisakher, *et al.* [BYA+20], on which I significantly assisted; however, as Banisakher was the primary technical contributor to this work, it will not be described in full within this dissertation. As a brief summary, the work was an appliction of a technique using conditional random fields developed as part of Banisakher's other work in medical information extraction. This model, along with the addition of lexical, further positional, syntactical, and semantic features resulted in a substantial increase in performance. The model alone resulted in an increase from $F_1 = 0.54$ to $F_1 = 0.59$, while the additional features further raised performance to $F_1 = 0.71$.

5.7 Summary of Contributions

This portion of work has several key contributions. First, I have demonstrated that humans can reliably learn and annotate news articles with van Dijk's theory of news discourse with a high degree of agreement. Second, I have developed a system that can predict the document-level discourse labels for paragraphs within a news article with reasonable performance (65% of human performance). Third, I have generated a

gold-standard corpus of these labels, along with an annotation guide, to support future work, and which has already been applied to further improvements of the model.

CHAPTER 6

Automatic Detection of Motifs

6.1 Motivation

The main contribution of this work is a pipeline for the automatic detection of motifs. As demonstrated in the previous section, categorizing motifs as motific, referential, eponymic, or unrelated requires a substantial amount of time, effort, and money to complete manually and produces relatively little data. Given the centrality of motifs to communication in culture, the rarity of annotated motif data, and the difficulty of gaining access to cultural experts to manually annotate data, I believe automatic motif detection to be a highly important task.

Automatic motif detection opens up the possibility of exploring the utility of motifs for a broad variety of tasks, including creating culturally-aware adaptions of information extraction and question answering systems. The presence of motifs in editorial pieces also points to the potential use of motif detection in the field of opinion mining. Further, I believe that motif detection has uses in the detection of disinformation and bad actors.

This work is an extension of preliminary results which I have published prior [YOA+21].

6.2 Approach

My approach uses a Java based feature extraction pipeline that caches the results of many different NLP tools developed both externally and within the Cognac Lab and SIFT. The output of these tools are parsed into features related to the motif candidates themselves (as many are generalized, sentence- or document-level features) by Python scripts and cached. These features are then fed into a decision tree model.

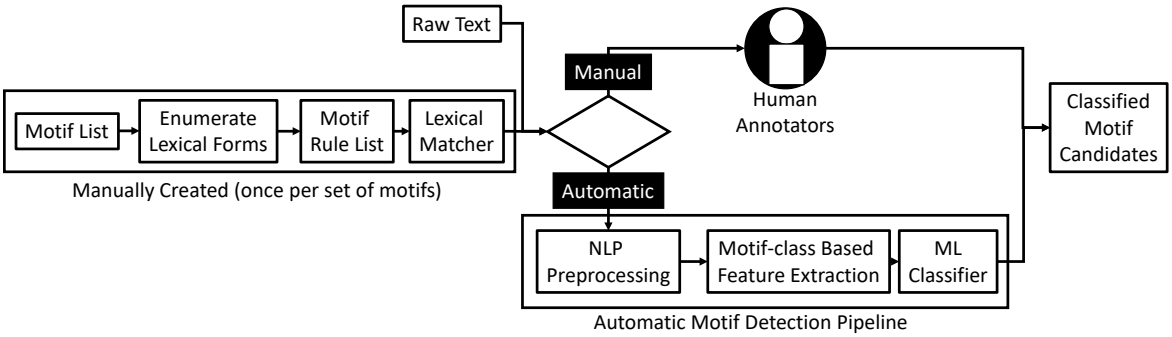

Figure 6.1: End-to-end Motif Detection Pipeline.

6.3 System Architecture

A simplified form of the system architecture is present in Figure 6.1. Displayed are necessary manual preprocessing steps. My automatic system, which I validate with annotated data, performs the task assigned to the human annotators: identifying which of the candidates from my high-recall lexical matcher are motific, eponymic, referential, or unrelated. This system is designed, as suggested in the motivation, to substitute for a human annotator in the classification of motif candidates into either motific, eponymic, referential, or unrelated.

Figure 6.2 shows the components of the pipeline as well as the flow of data through the pipeline. The pipeline is roughly divided into three steps: manual prepocessing, preprocessing, and motif detection. Although the majority of the components displayed are in use, some are not used due to their minimal impact on performance for the cost of running the tool. Many of these tools take a substantial amount of time to run over larger texts: because of this, the caching step situated between preprocessing and motif detection is essential for running the system in a timely manner for experiments. However, as the majority of these components are not trained for use specifically on the motif dataset, they can easily be run ad-hoc on new data without needing to reprocess the entire dataset.

The **manual preprocessing** step is the only step in the process that requires human input; this input requires the enumeration of known lexical forms of a motif. While it may be possible to automate this step in the future, I did not implement this

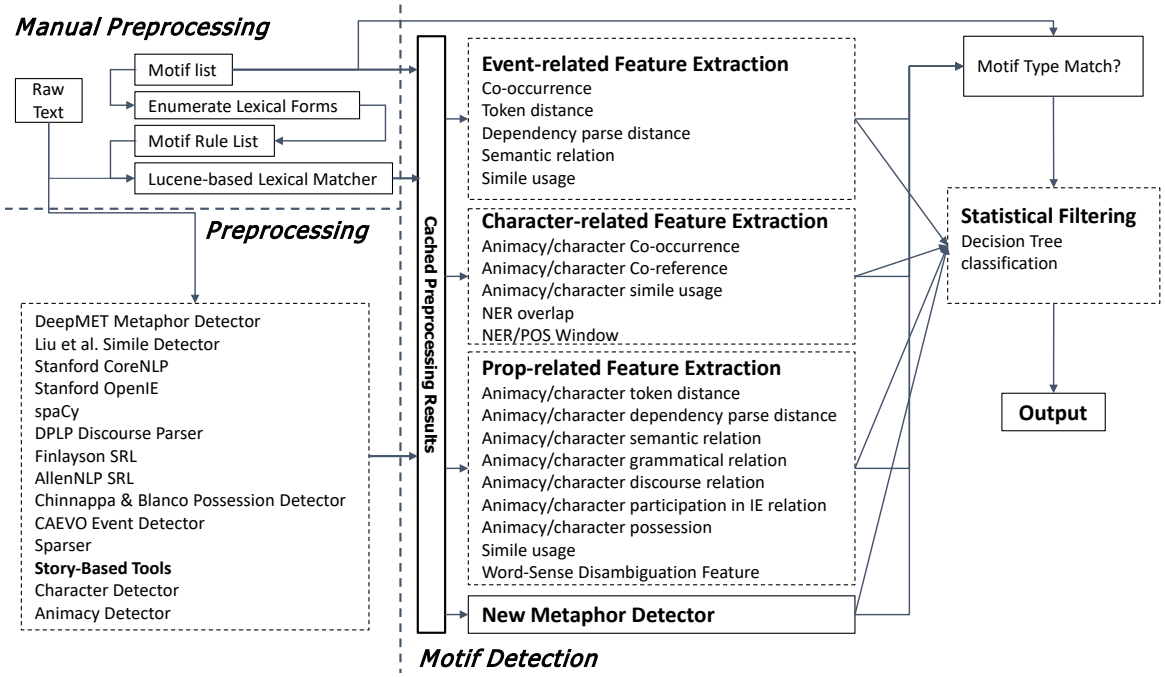

Figure 6.2: A detailed view of the automatic motif detection pipeline.

due to the difficulty in automatically processing motif indices, as I discuss in my short paper proposing the automatic detection of motifs [YF16a].

Of the components in this pipeline, I was directly involved in the refactoring of the animacy [JYRF21] and character [JMYF20] detectors that enabled generalizability experiments as well as all of the code in the motif detection portion of the pipeline.

6.4 Description of Features

Features for the pipeline are somewhat divided into three feature types: those intended to target event motifs, those intended to target character motifs, and those intended to target prop motifs. In addition to that, there are more general features that come out of the preprocessing pipeline. In the pipeline, there are 23 features available for use, which are as follows:

6.4.1 Actor-related Features

Co-occurrence with animate entities A boolean representing whether or not a candidate co-occurs with an animate entity, as detected by Jahan et al.'s animacy detector [JYRF21].

Coreference with animate entities A boolean representing whether or not a candidate participates in an animate coreference chain, using the same animacy detector.

Coreference chain length The length of the coreference chain the candidate is part of.

Distance to nearest animate coreference chain The token distance to the nearest animate coreference chain.

Distance to nearest animate referring expression The token distance to the nearest animate referring expression.

6.4.2 Event-related Features

Dependency tree distance from an event The distance from an event as detected by the CAEVO event detector [CCMB14].

Strict co-occurrence with an event The co-occurrence with an event as detected by CAEVO.

Token distance from an event The token distance from an event detected by CAEVO.

Semantic relation with an event The semantic relation, as detected by the SRL in Finlayson's Story Workbench [Fin11a] with an event detected by CAEVO.

6.4.3 Prop-related Features

Grammatical relation to the nearest inanimate entity The grammatical relation, as detected by Stanford CoreNLP [MSB+14b] between the candidate and the nearest inanimate entity.

Token distance to the nearest inanimate entity The token distance between the candidate and the nearest inanimate entity.

6.4.4 General Features

Windowed NER tags A 1-, 2-, or 3-wide window surrounding a candidate of named entity tags, as detected by CoreNLP.

Windowed POS tags A 1-, 2-, or 3-wide window surrounding a candidate of part-of-speech tags, as detected by CoreNLP.

Direct NER tags The majority vote for the entire span of a candidate for the named entity tags.

Direct POS tags The majority vote for the entire span of a candidate for the part-of-speech tags.

Simplified semantic role A simplified semantic role (in the form of ARG-1, ARG-2, ARG-M, etc.) as produced by the Finlayson SRL.

Word sense ranking The index of a candidate's word sense in WordNet [Uni10].

Presence of a comparative The presence of a comparative term (e.g., "greater," "better," "more," "less," etc.) in the same sentence as a candidate.

Semantic perplecity The embeddings representing the "surprisal" of BERT, a reimplementation of Li, et al. [LZT+21] by Armando Ochoa.

Document genre The genre of the document, as partially obtained from article metadata and partially computed using a high-performance opinion classifier [YH03] reimplemented by Joan Zheng.

Expected motif type The expected motif type, taken from the original motif list, of either actor, prop, or event.

Metaphoricity A boolean representing whether a candidate is metaphoric or not, from DeepMET [SFH+20].

Participation in simile A boolean representing whether a candidate participates in a simile or not, from Liu et al. [LHS+18].

6.5 Machine Learning Classifier

For this pipeline, I use the decision tree and random forest classifiers implemented as part of the Scikit-learn Python package [PVG+11]; for the decision tree model, the class weights are balanced by the implementation and it uses the entropy classification criteria; for the random forest model, the class weights are balanced by the implementation for every tree grown.

6.6 Results

The best results from the pipeline are shown in Table 6.1, which use a subset of the features extracted from the pipeline. This comparison shows the results for metaphor alone, simile alone, and the decision tree and random forest classifiers using all of the previously mentioned features with the exception of semantic role, word sense ranking, comparatives, and simile. The random forest model performs the best. These results are the average results of 10-fold cross validation on the dataset, split at runtime, performed to prevent overfitting to a test or validation set.

These results demonstrate high performance on eponymic and referential candidates, with lower performance on unrelated candidates and motif candidates being the most difficult to classify. A substantial amount of the performance comes from metaphoricity of motifs, but the additional features in the pipeline provide a boost of roughly 25% in performance over metaphor alone.

6.6.1 Discussion

I believe that metaphoricity provides the boost it does due to the overlap between metaphor, which represents non-literal language, and motif, which uses a term to refer

Table 6.1: Brief summary of the macro-average results.

Features	Motif F_1	Eponym F_1	Referential F_1	Unrelated F_1	Macro F_1
Metaphor	0.35	0.00	0.59	0.00	0.21
Simile	0.00	0.00	0.59	0.00	0.16
Decision Tree	0.41	0.74	0.80	0.61	0.56
Random Forest	0.44	0.79	0.85	0.68	0.59

to concepts not present in the text itself. Further, many motifs may be considered examples of metaphoric langauge: a "sword in the stone" or a "silver bullet" are both common metaphors that stem from motific origins. The pipeline does substantial work over metaphor in identifying eponymic and unrelated usage of motifs—this is least surprising in the case of eponyms, which represent an edge case of motifs in which an entity is referred to using a motif. Surprisingly, despite the expectation of high performance as another form of figurative language, the simile feature performs relatively poorly and does not change performance substantially even when used in combination with other features. This suggests that motifs are very rarely used as direct comparisons.

6.7 Error Analysis

I go over the per-motif type (actor, prop, event) and per-tag breakdowns of the performance of both the decision tree and random forest models, which reveal substantial weakness in specific tags and types.

6.7.1 Per-Type

Table 6.2 has the per-type breakdown of the macro averaged results for both the decision tree (DT) and random forest (RF) models used in the pipeline. With the current set of features, the pipeline has substantial difficulty with the actor class.

Interestingly, despite the inclusion of features intended to target actor features, the pipeline has substantial difficulty on them. This may be due to the similarity between

Table 6.2: Motif detection results per type.

Model	Category	Motif F_1	Eponym F_1	Ref. F_1	Unrelated F_1	Macro Avg.
DT	actor	0.16	0.79	0.77	0.32	0.51
DT	prop	0.54	0.69	0.81	0.66	0.68
DT	event	0.80	0.69	0.88	0.20	0.64
RF	actor	0.10	0.85	0.84	0.36	0.54
RF	prop	0.60	0.76	0.85	0.72	0.73
RF	event	0.79	0.57	0.88	0.16	0.60

actor motifs and ordinary named entities: there isn't much difference between "Finn McCool" as a name or the usage of the name to refer to the mythical figure.

Prior to the implementation of semantic perplexity, props performed as poorly as actors. However, semantic perplexity seems to have substantially benefited props, likely due to the fact that I expect prop motifs to be used in substantially different semantic contexts than expected: for example, the chaos of the stock exchange described as a "Tower of Babel" is distant from any contexts in which you may expect to see a tower or a reference to babel.

The high performance of the detector on events is likely due to the inclusion of a dedicated event detector, CAEVO, as part of the preprocessing step, which allows features that directly assess participation and distance from events within the article.

6.7.2 Per-Motif

Table 6.3 has the per motif breakdown of the random forest model per motif tag type. N/A indicates that a category did not return a valid result as there were no samples for that category. Some of the highest performance, in terms of motifs, comes from Jewish motifs: kiddush, leviathan/behemoth, and babel. Additionally, this breakdown shows that there are many motifs that are simply missing one of the categories: these are separated at the bottom of the table. This suggests that for banshee, magic harp, and fairy fort, despite being present in some form in the dataset, they do not have motific examples. Notable is the presence of eight motifs on which the classifier fails to classify any motific examples correctly: of these, six are considered actors, providing further evidence of the difficulty in classifying actor motifs.

Table 6.3: Motif detection results per motif tag type.

Motif Label	Motif F_1	Eponym F_1	Ref. F_1	Unrelated F_1	Macro Avg.
kiddush	0.81	0.58	0.86	0.00	0.56
leviathan-behemoth	0.78	0.89	0.07	0.00	0.44
babel	0.61	0.59	0.76	0.00	0.49
finn-mccool	0.35	0.85	0.84	0.00	0.51
shamrock	0.32	0.80	0.80	0.78	0.67
reyes	0.24	0.78	0.25	0.42	0.42
coqui	0.14	0.68	0.91	0.47	0.55
amalek	0.09	0.00	0.80	0.14	0.26
leprechaun	0.00	0.86	0.88	0.00	0.44
cu-chulainn	0.00	0.31	0.94	0.00	0.31
aos-si	0.00	0.00	0.00	0.73	0.18
70-languages	0.00	0.00	0.38	0.93	0.33
haman	0.00	0.62	0.84	0.21	0.42
hormigueros	0.00	0.08	0.94	0.31	0.33
cofresi	0.00	0.89	0.71	0.00	0.40
jibarito	0.00	0.43	0.03	0.00	0.12
banshee	N/A	0.32	0.84	0.86	0.67
magic-harp	N/A	0.00	0.00	0.79	0.26
fairy-fort	N/A	0.80	0.00	N/A	0.40
ark-of-the-covenant	0.29	N/A	0.91	0.00	0.40
golem	0.08	0.96	0.83	N/A	0.62
tir-na-nog	0.00	0.78	0.53	N/A	0.44
salmon	0.00	N/A	0.95	N/A	0.48
juan-bobo	0.00	0.74	0.61	N/A	0.45
chupacabra	0.00	0.63	0.86	N/A	0.50

Table 6.4: Results of an experiment classifying the actor class separately from the prop and event classes.

Model	Motif F_1	Eponym F_1	Ref. F_1	Unrelated F_1	Macro Avg.
DT, Actor Only	0.14	0.74	0.74	0.22	0.54
DT, No Actor	0.50	0.65	0.76	0.63	0.60

6.7.3 Separating the Actor Category

Table 6.4 contains the results of an experiment separating the actor class out to be classified separately using a decision tree classifier to determine if the actor class required different features for detection. Unfortunately, the actor class performs worse in the absence of the remaining data, and while the prop and event shows better performance on motifs ($F_1 = 0.5$), the performance degrades for the eponym and referential categories.

Table 6.5: Comparison of best performing pipeline model vs. a neural net model.

Model	Category	Motif F_1	Eponym F_1	Ref. F_1	Unrelated F_1	Macro Avg.
RF	All	0.44	0.79	0.85	0.68	0.59
NN #1	All	0.61	0.91	0.91	0.91	0.84
RF	actor	0.10	0.85	0.84	0.36	0.54
NN #1	Actor	0.17	0.92	0.91	0.57	0.64
RF	prop	0.60	0.76	0.85	0.72	0.73
NN #1	Prop	0.71	0.88	0.91	0.87	0.84
RF	event	0.79	0.57	0.88	0.16	0.60
NN #1	Event	0.94	0.88	0.97	NaN	0.93
NN #2	All	0.08	0.46	0.50	0.11	0.29

6.8 Comparison to Neural Net Results

My colleage, Armando Ochoa, helpfully implemented a neural net using RoBERTA [LOG+19] for the classification task. I provide his results here as a comparison to the pipeline model I have developed, which was designed to be a more generalizable model. Table 6.5 contains this comparison. In the table, NN #1 refers to the standard experiment of training on all motifs and testing on all motifs, with no motifs excluded. NN #2 refers to training on a subset of motif forms and testing on a disjoint set of motifs, which can be used as a test of generalizability.

6.8.1 Discussion

Disappointingly, the pipeline fails to achieve the same performance as the neural net model. Interestingly, the neural model also experiences substantial difficulty with the actor category, and mirrors the strengths of the pipeline in prop and events. This suggests that actors are the most difficult type of motif to categorize regardless of approach. However, one area in which the neural net model has substantial difficulty is in generalizability; one expectation of the pipeline is that it should be more generalizable due to the selection of linguistic features.

Table 6.6: Results of generalizability tests across cultural subsets of motifs.

Train	Test	Motif F_1	Eponym F_1	Ref. F_1	Unrelated F_1	Macro F_1
IR/JW	PR	0.044	0.382	0.604	0.110	0.285
IR/PR	JW	0.133	0.451	0.492	0.073	0.287
JW/PR	IR	0.012	0.422	0.376	0.245	0.264
IR	JW/PR	0.141	0.421	0.361	0.103	0.257
JW	IR/PR	0.063	0.380	0.279	0.242	0.241
PR	IR/JW	**0.249**	0.443	0.488	0.119	0.325

6.9 Generalizability

Table 6.6 contains the results for early testing of the generalizability of the decision tree model over the different combinations of cultural groups for motifs. While the random forest model was tested, it performed substantially worse at generalizing.

Interestingly, on sets that do not contain Jewish motifs, the model generalizes to a small degree, with the best performance being training on Puerto Rican motifs and testing on the overall set of motifs. While these tests are not strictly comparable to the generalizability experiment performed for the neural net model, they operate on similar sized splits of data: the neural net model was tested on 75:25 splits, cross-validated, while these experiments are done on 66:33 and 33:66 splits, as noted above.

These experiments suggest that the pipeline may generalize somewhat decently with a solid set of data. It is not entirely clear why the Jewish motifs perform so poorly compared to the Irish and Puerto Rican motifs at generalizing: one possibility is that the Jewish motif's greatest strength in human annotation, their distinctness, is a detriment in attempting to generalize beyond the subset. Another possibility is that the Jewish data contains the highest number of events, which the system performs the best on; however, that doesn't account for abysmal performance with Jewish training sets and Puerto Rican test sets, as the Puerto Rican data contains the second highest number of events and a substantial number of props. Additionally, there could be something about the motifs themselves: while the Irish motifs have become rather diluted due to their propagation into pop culture, the Jewish motifs are substantially obscure and nuanced. During the annotation sessions, annotators had a lengthy debate over the exact meaning of *Kiddush hashem* before backing off to a

more general cultural definition of it. This could indicate that the Irish motifs are too general to provide substantial information on motifs while the Jewish motifs are far too specific to learn anything about a broader range of motifs, or require substantially more cultural knowledge. One final potential reason for this difference could be the fact that the Puerto Rican team was asked to provide additional motifs from their own experience towards the end of annotations: these final motifs could be of higher quality, being originally selected by in-culture individuals. All of these are avenues for future work exploring generalizability.

6.10 Summary of Contributions

The main contribution of this work is the development of a motif detection pipeline and a characterization of its performance on different classes of motif, as well as a characterization of its potential for generalizability. The selection of and evaluation of features provides a path forward for the refinement of future motif detection models and suggests areas in which substantial improvements to the model can be made to increase performance.

CHAPTER 7

Exploring Beyond Motifs

7.1 Motivation

With the completion of the human annotation, I had determined that motifs could be reliably annotated by humans. However, I was also aware that motifs were not the only interesting narrative element from my prior work assisting my advisor, Mark Finlayson, with continuing his work on Vladimir Propp's morphology [Pro68a]. Further, while the annotation scheme I had developed worked well for the domain of well-grounded motifs with a folklore background, it wasn't clear that it would transfer to other domains cleanly. Thus, I was left with three questions, which I list in the order I will present them:

1. Does the motif annotation scheme generalize to other domains outside of folklore?
2. If not, what modifications are necessary to achieve substantial agreement?
3. Where do motifs lie on the continuum of narrative elements that communicate substantial amounts of information?

While motifs are an important part of communication, I believe that they paint an incomplete picture of how humans communicate with one another and express their thoughts and identities via language. Motifs are not the sum of cultural identity in narrative. One need look no further than the biblical story from which the modern idea of the shibboleth comes from: a word whose pronunciation itself is a signifier of identity, with lethal consequences.

Motifs themselves act as a form of shibboleth: there are subtle differences in the way that in-culture people use a motif vs. how someone from outside of a cultural group perceive or use the same motif. From the extremes of how Irish motifs are treated in American pop culture (a cereal mascot!) in comparison to their original form (a mischievous fairy), which was the source of endless complaints from my Irish annotators, to subtle differences that may catch someone off guard even if they are

familiar with a culture (despite being the land of youth, tír na nÓg is associated with the tragedy of overreaching one's self).

One such community in which motifs and motif-like elements serve as a shibboleth is the online community of "involuntary celibates," whose plight is immediately apparent from their chosen demonym. Commonly referred to as "incels," and sometimes viewed as an internet-based extremist hate group (due to members of the subculture participating in or claiming responsibility for heinous acts of violence in the real world), they are a niche internet subculture of primarily males who discuss their views on relationships and society. This subcommunity was suggested as one of potential interest to the project this work falls under by our collaborators at SIFT, who were using similar data in a different project, and had noticed motif-like elements within the data.

Notable about the subculture is the highly adversarial relationship that incels have with the rest of the internet: this strong focus on in-group/out-group suggested to me that they likely had a strongly defined culture that was readily apparent in their posts, making them an interesting community to study the generalizability of the motif annotation scheme to other domains.

7.2 Approach

The start of this part of my work proceeded very similarly to my approach to the main annotation phase:

- selecting the specific motifs,
- selecting and acquiring data,
- developing of an annotation guide and scheme,
- annotator training,
- adjudication,
- and iteration on the annotation guide and scheme.

The primary differences between this and the work on the folklore motif annotations is there was no need to select cultures (as this focused on a specific subculture), no need to build and test a pipeline (as it was still in place from the main annotation), and no hiring of annotators. This last point is worth lingering on: as a group, the people involved with the project decided that the subject matter was not suitable to subject people to without allowing them to back out, and needing to retrain annotators (and pay to train them) should an annotator choose to opt out of continuing made it an unappetizing option.

Annotations were, therefore, performed within the group, and were given the opportunity to opt out of the study at any point in time. These annotations were done in smaller batches and adjudicated as time allowed, with adjudication sessions primarily being for the sake of refining the annotation guide.

7.3 Selection of Motif-likes

The domain of incels is radically different from the folkloric domain used for the primary study: there are no stories in which motifs are grounded and the subculture itself is substantially newer than centuries of folklore: some reports suggest the earliest usage of the term "involuntary celibate" date to the mid-90s [LMMF18]. This means that there are no existing motif indices for the incel subculture. There is, however, a substitute: wiki articles that list terms of interest within the subculture, used to introduce new members to the subculture to the concepts.

From this wiki, I drew 80 terms of potential interest, which I worked to narrow down to a more sizable 30 motif-like elements, excluding terms that seemed too generic (e.g., "high T," a term referring to high testosterone) or referred to communities separate from the primary incel community (e.g., "MGTOW," short for "men going their own way," a separate community with some overlap).

7.3.1 Aside: What does motif-like mean?

I use the term motif-like to distinguish the elements that I found in my investigation of the incel subculture from motifs proper as they occur in folklore. One major reason is that, as I discovered over the course of this work, they are expressed different and are often used in a different way than standard motifs. However, many usages still share similarities to motifs: hence, "motif-likes."

7.3.2 List of Motif-likes

The list of motif-likes identified and used for this study of incel subculture are: alpha, beta, betabux, Chad, cope, cuck, femoid, hypergamy, incel, IT/Incel Tears, inhibition, KHHV, landwhale, NT/neurotypical, normie, numale, omega, oneitis, -pill, beta provider, roastie, sperg, Stacy, thirstie, volcel, and wizard.

Full descriptions of these motif-likes can be found in the appendices.

7.4 Selection of Data

The data chosen was from an archive of Reddit (a news aggregator) posts from incel-related subforums on the site. There were a total of 1,075,506 posts from 35,295 users, including 217,015 messages from users whose accounts were deleted. These posts spanned a total of 10 months. Reddit was chosen as, prior to the deletion of the subforum, it hosted one of the largest incel communities on the internet.

These posts were anonymized by Anurag Acharya using a script designed for the anonymization of forum posts. After running our lexical matcher over the dataset, I found that 21.4% of the posts contained motif-like candidates. Of the 30 chosen motif-likes, 27 appeared as candidates in our dataset. The breakdown of the top ten candidate instances is in Figure 7.1.

Unsurprisingly, the term "incel" is the most common, given that it is also the name of the subforum and the demonym for the subculture. Interestingly, "chad" and "normie" are both extremely common as well—I believe this is a result of the strong

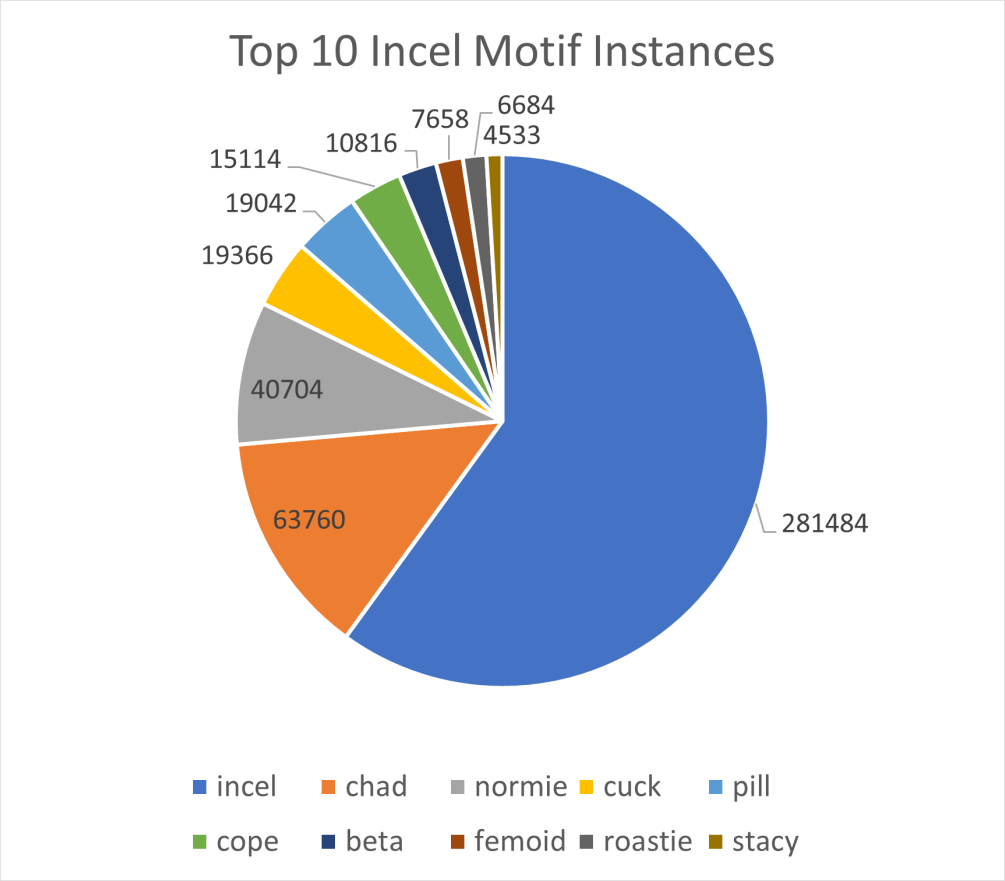

Figure 7.1: Breakdown of top ten candidates in terms of number of instances.

in-group/out-group dynamic, as many members of the subreddit consider themselves to be the polar opposite of either "chads" or "normies" in society.

7.4.1 Topic, Users, Targets, Sentiment, and Themes in Incel Data

As part of exploring the incel data, I annotated 100 adjudicated posts that were not eponymic, unrelated, group references, proper terms, or duplicates for speaker, target, sentiment, and general theme. Within these 100 posts, the distribution of motif-likes, shown in Figure 7.2, is similar to the distribution of the corpus overall (Figure 7.1), with incel, chad, and normie filling the top three spots in terms of frequency.

The vast majority of posters do not clearly identify themselves, as demonstrated in Figure 7.3. This is not particularly surprising, given the psuedo-anonymous nature

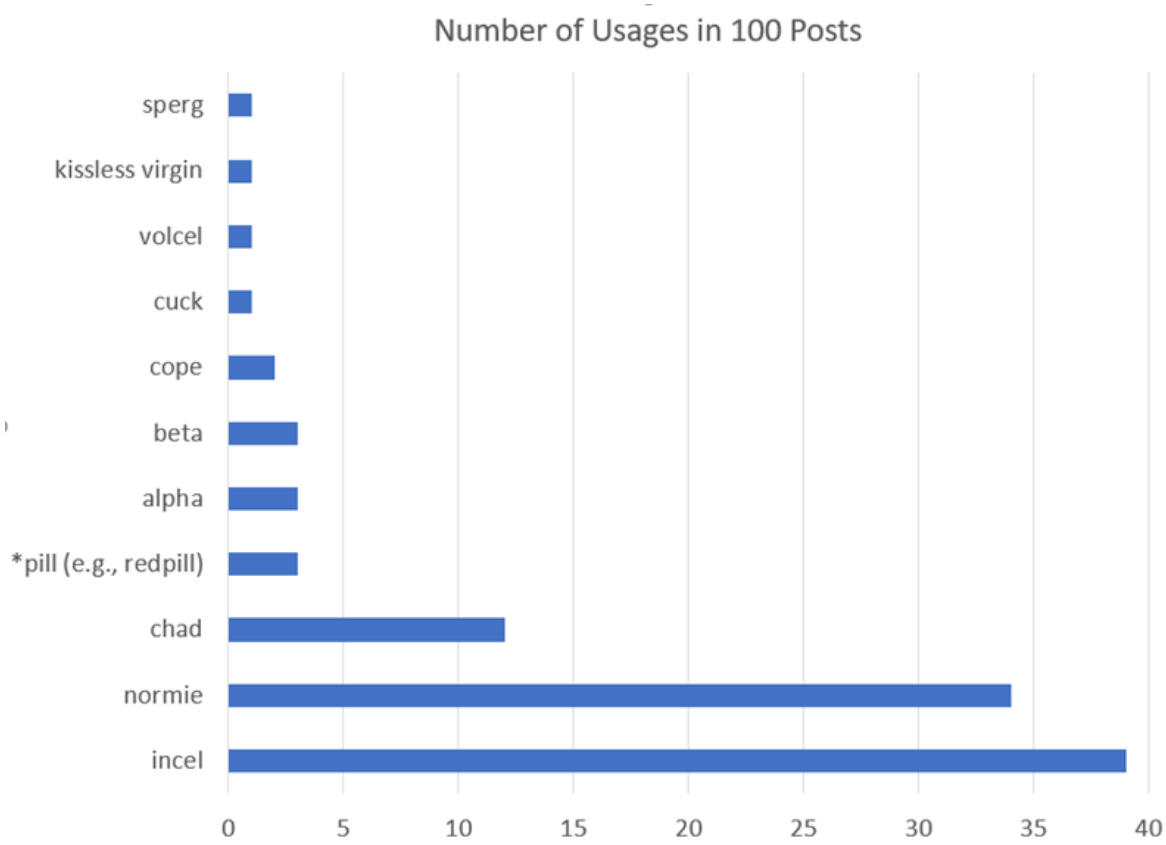

Figure 7.2: Number of instances of the categories present in the batch of 100 motif-likes annotated.

of Reddit subforums. Of the posters that do provide a clear identification within their post, nearly 7x as many posters are self-identifying incels vs. self-identifying "normies," considered to be a sort of opposing faction within the subreddit.

Shown in Figure 7.4 are the targets of posts. Overwhelmingly, the majority of posts target a specific group (e.g. incels or normies), regardless of sentiment. Following this are posts targeting a specific person with no mention of gender, followed by women, and individuals who use specific motifs (e.g., "people who use the term chad"). Of the nine posts that targeted women, the majority were negative (six negative, two neutral, one positive). Of the six posts targeting motif users, all instances took a mocking tone.

Figure 7.5 contains a breakdown of the overall sentiment of the 100 posts taken from the incel subreddit. The majority of posts have a negative sentiment, with only seven posts taking a positive tone. The sentiment category "mocking" was introduced to denote posts that mock a user or category of people, often using a strawman.

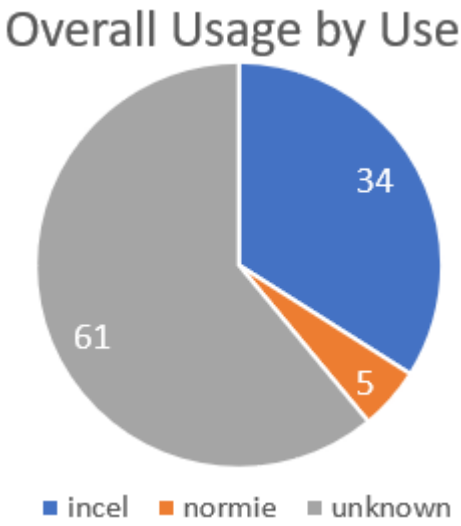

Figure 7.3: Breakdown of the 100 posts by identity of the user posting.

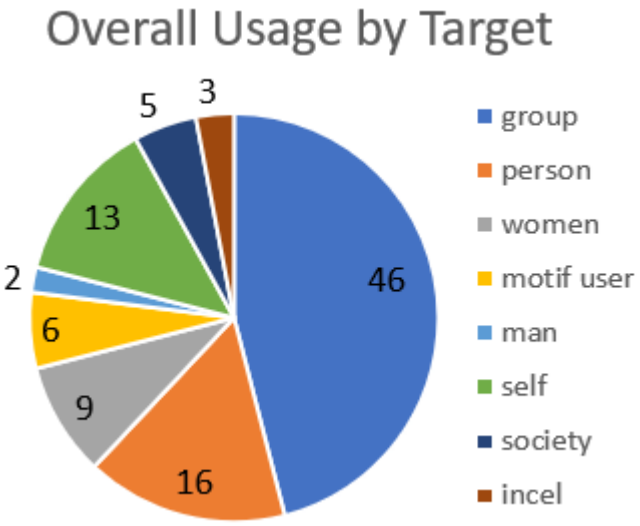

Figure 7.4: Breakdown of the 100 posts by the identity of the target of the post.

Finally, I annotated the usage by theme, introducing new categories as they seemed fit. The breakdown of this annotation is shown in Figure 7.6. The most present category is "persecution"—the belief by members of the incel subreddit that they are persecuted, either by specific groups ("chads" or "normies") or by society at large, and expressing a feeling that members of the incel subculture are treated unfairly or dealt an unfair hand in life.

Interestingly, the second most common theme was "counter motif"—posts that deny a motif-like's meaning, either positively or negatively. In some cases, this is to

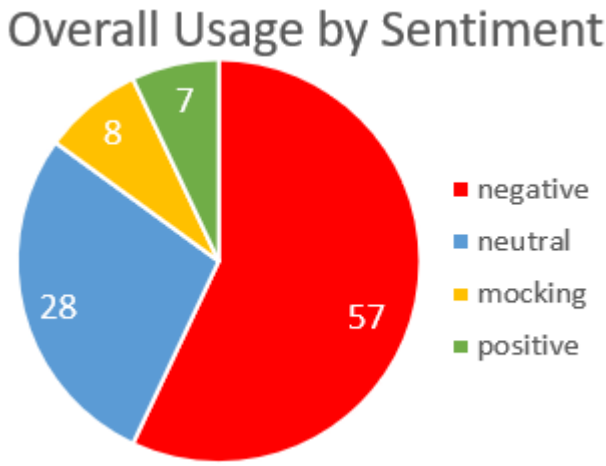

Figure 7.5: Breakdown of the 100 posts in terms of sentiment.

decry incel beliefs and insinuate that members of the incel subreddit are mistaken in their worldview. However, in rare cases, this is intended to be positive reinforcement for members of the community: for example, suggesting that things will get better and that not all of society aims to persecute them.

Many posts fall under the explanatory category, which neutrally addresses situations. Another set of posts are "belief affirming," the opposite of "counter motif" posts in that they reinforce incel motif-likes and the narrative they produce. Many categories ("anti-group," "insult," "don't understand," "superior," "no help," "membership denying," "dismissive," or "immature") are essentially us-vs-them rhetoric, being highly oppositional posts suggesting an inability to communicate with the other group or a fault in the other group.

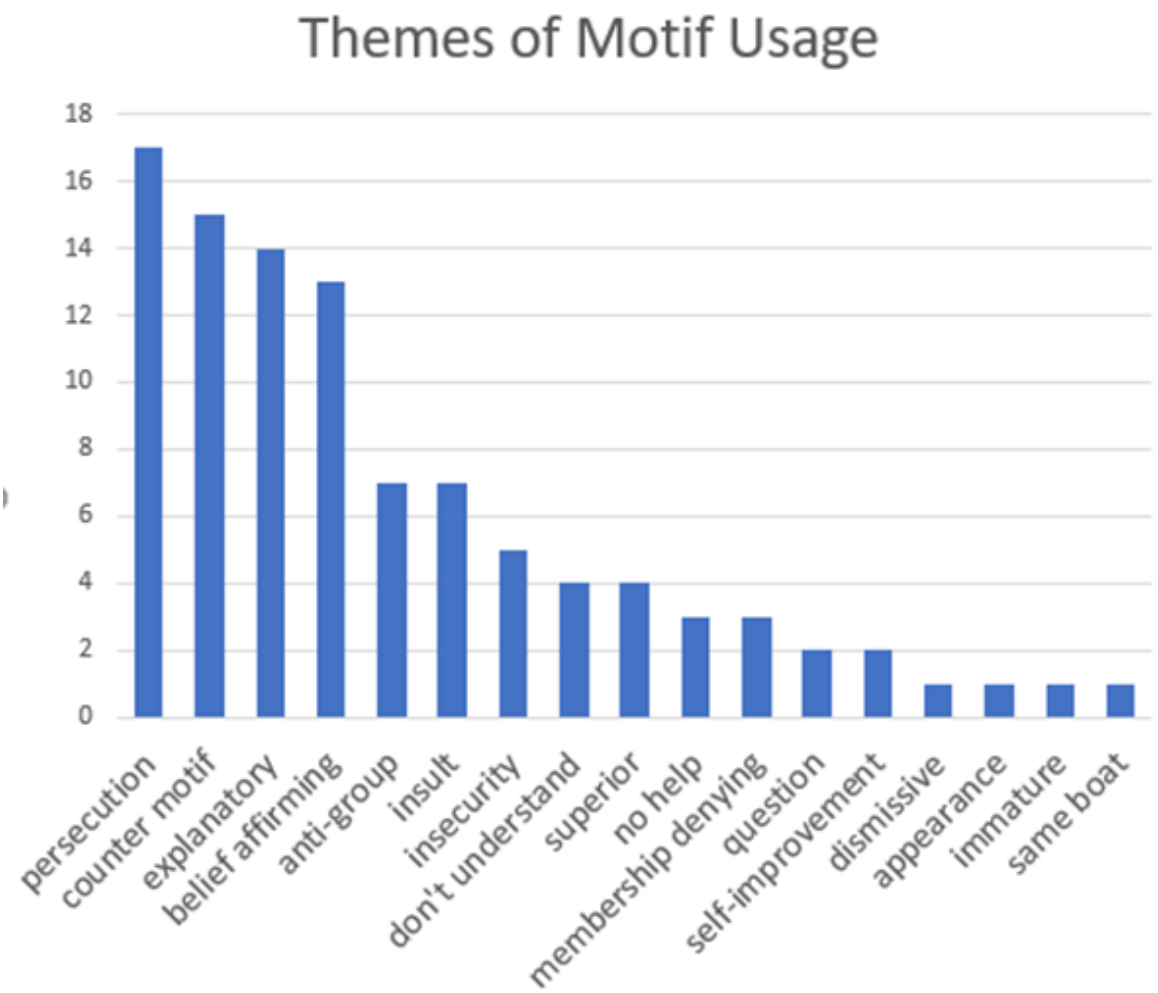

Figure 7.6: Overall themes present in the 100 annotated examples of incel data.

7.5 Annotation Guide and Scheme

For the initial batch of annotations, the original motif annotation guide was used. Past that, for the most part, the annotation guide remains the same as the guide for motifs. A partial version of the guide, with elements that are shared between the guides removed, can be found in the appendices.

The annotation scheme evolved substantially over the course of nine internal annotations, starting from the original four category scheme, expanding to five, the fourteen, before settling on thirteen elements. The process by which the scheme evolved is described as part of the following section (§7.6) as are the intermediary annotation schemes and the final annotation scheme is described in §7.7.

7.6 Annotation

7.6.1 Annotator Training

Annotations were performed by myself and two of my colleagues, Anurag and Diana. Annotators were not given any special training other than being provided with the motif guide and annotation scheme and being informed that the nature of content contained within some posts on the incel subforum could be substantially disturbing. Annotators were instructed to cease annotating at any point in time that they felt they could no longer continue with the annotation work.

7.6.2 Annotation Procedure

Annotation procedure remained the same as in the main annotation: annotations were done in a double-blind manner with an adjudication meeting held afterwards to discuss any issues and disagreements. These annotations were not done on a strict schedule, but rather completed as time allowed.

7.6.3 Batch 1: The Mystery of the Many Motifs

For the first batch of annotations, we retained the four element annotation scheme from the original annotation: motific, referential, eponymic, and unrelated. While the annotations had relatively high agreement ($F_k = 0.538$), I noticed that the incidence of valid motif-likes (those elements tagged as "motific") was substantially higher than in the main annotation: 337/394 candidates were marked as motific.

After the adjudication meeting, I assumed that there were many times that a term had become part of the subcultural group's vocabulary without having specific figurative usage. To address this, I introduced the "common usage" category with the intent of capturing motif-like usage that had degraded to common vocabulary terms. For example, in the sentence "rot in front of my computer posting on **incel**

sites," incel generally refers to the group, but doesn't immediately seem to invoke any broader connotations.

7.6.4 Batch 2: Not-so-common Usage

The result of the introduction of the "common usage" category was a substantial decline in performance, down to $F_k = 0.151$. As part of the adjudication meeting, I decided to revise the category from "common usage" to "grounded instance," intended to clarify the primary usage of motif-like terms that had previously been considered common usage—referring to either a specific group ("**incels** on this forum") or a specific person ("okay **chad**").

7.6.5 Batch 3 and 4: Ungrounded

The 3rd and 4th batches of annotations remained disappointing ($F_k = 0.18$ and $F_k = 0.36$ respectively). After discussing things with the other annotators and with my advisor, I decided to undertake a substantial revision of the annotation scheme by starting from the data and identifying categories from the samples there. The result of this process was a substantially different annotation scheme featuring a massive fourteen elements:

Descriptive Role Assigned to a specific, real person, with the intent of making it obvious what their character traits are or aren't.

Mythic Role A reference to the role itself, rather than any specific individual possessing the role.

Group Role The usage of a role to discuss a specific, real group and either assign traits to it, suggest stereotypes, or draw upon associations.

Active Event An event featuring action that the subject initiates; i.e. something they intentionally do.

Passive Event An event featuring action that the subject does not initiate; i.e. something forced upon them.

Stereotypical Situation A description of a stereotypical event. These describe in part the action/consequences implied by the motifs.

Event as Analogy The use of a motific event to describe or reframe something that is happening.

Standard Prop The usage of a motif as a prop.

Motif-referencing Character Trait A character trait either expressed as an adjective or as a direct trait that is specifically named after a motif.

Standard Eponym The usage of a motif as a name.

Group Reference The usage of a role to discuss a group in a manner that does not imply anything about the group nor require any knowledge about the group to understand.

Definitional Reference The usage of a role either to define it or in it's definitional sense.

Proper Usage The dictionary definition usage of a term rather than usage that has any motific associations, for terms that originate from outside the subculture.

Standard Unrelated An unrelated (and likely erroneous) tagging.

7.6.6 Batch 5: Hope and Minor Revision

This new annotation scheme, despite being substantially more complex, resulted in an increase in agreement between annotators, from $F_k = 0.36$ in batch 4 to $F_k = 0.40$ in this batch. Despite the increase in performance being somewhat small, I felt that this annotation scheme had substantial potential given that batch 4 was the second annotation with the previous scheme and this was the first annotation with a substantially more complicated scheme.

At the adjudication meeting, after discussion, I decided to remove the group role category, as it was causing the most confusion due to overlapping substantially with both descriptive and mythic roles—both of these categories were revised to allow for either groups or individuals that fit the description. This final revision resulted in an annotation scheme with thirteen categories.

With the removal of the group role category, ten of the categories also had their definitions revised:

Descriptive Role A role assigned to a specific, real person, with the intent of making it obvious what their character traits are or aren't based on the associations of the role. This can be a group.

Mythic Role A reference to the role itself, rather than any specific individual possessing the role, usually to illustrate a general idea or as part of a stereotypical event. This can be a group.

Active Event An event featuring action that the subject initiates; i.e. something they intentionally do.

Passive Event An event featuring action that the subject does not initiate; i.e. something forced upon them.

Stereotypical Situation A description of a stereotypical event, usually involving the usage of props or actors. These describe in part the action/consequences implied by the motifs. Note that when these may include a mythic role or descriptive role; you should use this tag rather than one of those.

Event as Analogy The use of a motific event to describe or reframe something that is happening. Non-incel example: "cutting the gordian knot" as an analogy for problem solving.

Standard Prop The usage of a motif as a figurative prop: e.g., an object that can be used (silver bullet), given to someone (panacea), etc.

Motif-referencing Trait A trait either expressed as an adjective or as a direct trait (e.g. inceldom; the nature of being an incel) that is specifically named after a

motif in the space and thereby depends upon an understanding of the original motif to understand the trait. item[Standard Eponym] The usage of a motif as a name; note that this does NOT include inferred usages that require specific knowledge, e.g. calling something "redpill lite" should not be inferred to refer to an eponym due to lack of expertise.

Definitional Reference The usage of a motif to define it, provide information about its meaning, or as reference to its original source.

7.6.7 Batch 6, 7, and 8: Success

Batch 6 saw a substantial improvement in agreement: $F_k = 0.61$. With batch 7, we performed a three-person annotation as part of training, as we had added an additional annotator, which resulted in an agreement of $F_k = 0.6$, with the pairwise agreements being between $F_k = 0.55$ and $F_k = 0.68$. This was substantially promising, as even the worst performance with a complicated annotation scheme achieved the expected performance from the main motif annotation phase.

For batch 8, I performed two sets of annotation with the other two annotators performing a single set of annotation. This resulted in agreements of $F_k = 0.54$ and $F_k = 0.65$, demonstrating sustained high agreement. This strongly suggests that this annotation scheme can be reliably applied to the domain of incel subculture motif-like elements.

7.7 Final Annotation Scheme and Close Analogues

Table 7.1 shows the final thirteen categories used for the motif, as previously described, along with close analogues to their role in narrative. *Char.* is short for character and *DP* stands for *dramatis personae*. For examples, the motif-like is in **bold**. Their definitions can be found earlier in this chapter. In addition to the categories, I also list close analogues that I either recognized as similar during the annotation process or were direct inspirations for the categories themselves.

Table 7.1: The thirteen final motif-like categories.

Analogue	Category	Example
Char.	Descriptive Role	Bullshit I'm 5'5 and not a **Chad** by any means but have gotten a few girls.
Char., DP	Mythic Role	Now that women can pick and choose who to mate with **incels** are gonna go the way of the dodo bird.
Event	Active Event	Until all women are **red pilled** and stop spouting shit like LOOKS DON'T MATTER/AREN'T THE ONLY THING THAT MATTERS
Event	Passive Event	Reading this back I think it's becoming clearly where my deep-seated and intense hatred of cucks/**being cucked** and exclusivity and shit comes from.
Function, Script, Frame	Stereotypical Situation	Oh college aged **woman** *like sex* alright. Just *not sex with* types like **Incels**.
Event, Frame	Event as Analogy	Seriously, nobody is actually planning for world domination here, the fantasy is just a **cope**, a harmless daydream.
Prop	Standard Prop	The **red pill** is truth.
Char., Prop, Eponym, Frame	Motif-referencing Trait	...your abrasiveness and distorted viewpoints are to blame for your **inceldom**, and you should take responsibility for that?
Eponym	Standard Eponym	Is being an annoying bitch on **/r/incels** doing anything for you?
Char., Referential	Group Reference	Entitled' is a fairly meaningless word that does not accurately describe most of the **incels** on this sub or elsewhere on the Internet.
Referential	Definitional Reference	most young white males are **involuntary celibate**.
Unrelated	Proper Usage	I've been around people with even mild autism, as someone who was once thought to have **Asperger's** (though that was not a correct diagnosis).
Unrelated	Standard Unrelated	Overall they're all fairly average, some slightly above average, with or without their **cel**ebrity status.

These analogues serve to demonstrate the complex relationship the motif-like elements in the incel subculture have with established theories of narrative structure. I define the analogues as follows:

Character, Event, Prop, and Eponym As previously discussed, these are the original motific categories. Some categories, such as "Motif-referencing Trait" share elements of multiple motif types (in this case, a trait is named after an actor or prop motif, which is then applied as an adjective to suggest the role of the target of the trait is identical to the original actor or prop).

Dramatis Personae The dramatis personae are seven roles identified by Propp [Pro68a] that a character fulfills in a story. This is almost identical to the mythic role in many cases, which refers to the idea of a character and the role they fulfill in society.

Function Also drawn from Propp, functions can generally be described as the actions that a character may take to drive the plot forward. These differ from event motifs, which describe *singular incidents* in that they describe how dramatis personae interact with the plot and other dramatis personae. Thus, they are substantially more generalized in comparison to event motifs.

Semantic Frame The idea of frames comes from Marvin Minsky's article, "A framework for representing knowledge" [Min74]. Here, I draw analogy with frames in their usage as storage structures for either stereotyped situations (e.g., what roles fit into specific situations and their implicit attitudes, such as women hating incels), by situations being drawn into an analogy with a known event motif-like (e.g., that taking up a hobby might be a form of *coping*, a known passive event), or by attaching motif-likes as a quality (e.g., the quality of being an incel might be called inceldom).

Script Scripts come from Silvan Tomkins' development upon affect theory [Tom87] and were extended for early AI work by Roger Schank and Robert P. Abelson [SA75], which use a frame-like structure to represent "mundane situations," such as dining at a restaurant. This fits closely to the stereotypical situations, which often fulfill the role of a more-grounded script.

7.7.1 Discussion

The final annotation scheme was sufficiently complicated that one of my collaborators was intensely skeptical when I first suggested it—despite concerns, however, it has been reliably applied by human annotators to the incel dataset I have collected. The similarities of motif-likes to other theories of narrative structure tie into one of my

original questions: where do motifs lie on the continuum of narrative elements? Where are these new motif-like elements situated?

7.8 The Continuum of Cultural Communication

There exists a spectrum of culturally-relevant communicative elements used in narrative that express a greater constellation of knowledge. From frames, to scripts, to Proppian morphology, to motif-likes, to motifs themselves, they form a continuum of expressions of cultural knowledge.

I propose that there are four axes of interest in the cultural communication continuum:

1. Abstraction
2. Identity
3. Context
4. Purpose

I attempt to place the various narrative elements described in the whole of this dissertation in the following sections, using examples, both synthetic and real, where appropriate to illuminate my reasoning as to their positioning in communication space.

7.8.1 Abstraction

By abstraction, I mean in the sense that a physical apple is less abstract than the idea of an apple, which is less abstract than the idea of food, and so on. I refer to the ends of this scale as *abstract* and *grounded*. For narrative elements, a very grounded example is a real-world occurrence or a specific story, while a very abstract element might be a narrative itself, the top-level element in narrative, which contains all others.

Below, I provide a rough ranking of narrative elements, from grounded to abstract, alongside my reasoning. Removed from consideration are the motif-like categories of proper usage, standard unrelated, group reference, and definitional reference, in

addition to the motif categories of unrelated and referential, as they are artifacts of the annotation scheme to describe meta-narrative examples of the elements or instances that aren't examples at all. At times, I use the term "cultural narrative" to refer to the portion of a narrative which is related to the source cultural group: for example, "Finn McCool" is part of the cultural narrative, but "George, the man down the street" may not be, even if both are mentioned in the same text.

Grounded

Descriptive Role The most grounded possible example, where a motif is exemplified in a single, real entity.

Motifs, active event, passive event, and props These examples are all motifs or motif-like elements that are grounded in a specific story, either real or fictional.

Eponym, eponym motifs, and motif-referencing traits To a degree, these all take a specific, real entity with no particular relation to the narrative and glues them to an idea from within the narrative, either by name or trait. I believe these to be less abstract as they originate from a cultural narratively uninvolved entity, providing metainformation about the subject.

Event as Analogy This is taking a cultural narratively uninvolved entity and contextualizing it as a more abstract event: in essence, stating, "this event is an example of this motif-like."

Mythic Role Mythic roles represent an "idealized" version of a motif and are often referred to as such: for example, the mythic "chad" who is always successful with women or the mythic "beta" who is always second-best.

Dramatis Personae I consider dramatis personae to be a slight abstraction of mythic roles, in that they are less culturally-relevant: a hero may occur across many cultural groups, but a mythic role generally originates from a specific cultural group.

Stereotypical Situation Stereotypical situations represent a further degree of abstraction in that they often represent not just mythic roles, but the expected interactions between mythic roles.

Functions, Moves As with dramatis personae being less grounded than mythic roles, so too are Propp's functions and moves less grounded than stereotypical situations.

Scripts, Frames, Discourse Function As schemes designed strictly for knowledge representation, these elements have almost no grounding within a story.

Abstract

Notable is that several of the narrative elements that I consider more abstract operate using what I consider a form of dereification: removing real, grounded entities that have no involvement in the cultural narrative and abstracting them, via indirect reference (e.g., analogy or naming) to bring them into the cultural narrative.

7.8.2 Identity

I propose that there is also an axis that covers identity: in particular, the degree to which a narrative element can be used as a signifier of belonging to a specific cultural group. Within this axis, I consider three groups: identity-agnostic, identity-theoretical, and identity-representational.

Identity-Agnostic These narrative elements have no relation to identity within a group: for example, understanding that members of group A may hate members of group B.

Identity-Theoretical These narrative elements are reasoning about the beliefs, characters, and expectations within a cultural group; many of these may be observable from outside of the group itself. For example, noticing that incels hate women.

Identity-Representational These narrative elements are deeply related to the identity of a group. For example, the wealth of ideas embodied in the "incel" mythic role: that they are persecuted, that society is against them, and that they are unfairly maligned and deprived of relationships that they view as a right; that "femoid" mythic roles are intentionally perpetuating a societal status quo and naturally gravitate towards the oppressing "chads," thereby making it natural for incels to consider femoids a group deleterious to their well-being and belonging in society.

7.8.3 Context

By context, I refer to the context in which we, as outside observers to a narrative, understand that narrative elements can either represent reasoning *about* the context of a narrative or reasoning *within* the context a narrative.

Context About a Narrative These are narrative elements that occur and may experience differences across cultures, such as Proppian moves or functions, but describe what is happening from an outside perspective.

Context Within a Narrative These narrative elements are those that directly drive the story forward using characters, events, and the reasoning and relations of those.

Hybrid Context Hybrid contexts were noticed among the incel motif-likes, especially stereotypical situations, which are often a meta-commentary on the outside perspective, similar to scripts, but drive character action within a narrative.

To illustrate the differences between these contexts and how they might all refer to the same narrative, consider the following examples:

Context About: F_{I1} `hero courts princess` F_{I2} `princess rebuffs hero` F_{I3} `villain courts princess` F_{I4} `villain succeeds and hero fails.`

Context Within: I asked some *Stacy* out and she claimed to not be looking for a relationship, but I later saw her dating some *Chad*.

Hybrid Context: *Femoids* are only interested in *Chad*.

7.8.4 Purpose

Purpose is simply the role that a narrative element serves as a unit for computation: I suggest that there is a axis where one end has knowledge-poor, context-rich elements that represent the connection between other elements; on the other end are the knowledge-rich, context-poor elements that encapsulate a constellation of ideas in a rich manner.

This differs from the axis of abstraction in that grounded elements, such as event motifs, still have a higher degree of context-richness in comparison to character and prop motifs, but occupy roughly the same level of abstraction.

7.8.5 Discussion

I believe that these axes suggest a rich relationship and intermixing of narrative elements used for communication. These represent ways to distinguish between them depending on what the most pressing concerns for a task are: to understand grounding, abstraction; to understand cultural group membership, identity; to understand narrative interaction, context; and to understand connections between pieces of knowledge, purpose.

I believe that there are substantial revisions that could be made to these axes and that they are likely not the only axes upon which communicative elements can be measured and compared. However, I believe this is an important step towards understanding how different levels and theories of narrative are related.

7.9 Culture, Identity, and the Path Forward

My work is deeply tied to the ideas of culture and identity. Motifs themselves are deeply rooted in cultural folklore. Motif-like elements are deeply tied to the ideas of

identity within the subcultural group of incels, and strongly relate to a variety of other theories of narrative structure.

I think there is great potential in being able to computationally detect these motif-like elements, now that I have demonstrated that they can be reliably detected by humans: anomalies in the manner in which actors present themselves and how they communicate may enable a path towards robust disinformation identification. Additionally, the range of communicative elements I have discussed in this chapter lend themselves readily towards culturally aware knowledge extraction, question answering, and, potentially, cultural commonsense.

I also have reservations about where these techniques might be taken: the subculture of involuntary celibates on Reddit only developed in the psuedo-anonymity of the internet forum. Can automatic detection of communicative narrative elements be used to de-anonymize individuals online? To what degree can their identity be ascertained based on the elements a person uses in communication? Certainly, humans can do so—in the simplest case, a friend mentioning a meme related to a TV show you're familiar with may clue you in to their belonging as part of the fandom for that show. To do so at scale, however, may be subject to abuse.

However, a robust understanding of how these narrative elements relate to one's personal and cultural identity, as well as the ability to automatically detect them, may lead to the very opposite: methods for automatically disguising one's identity and affiliations. One potential method of doing so might be to look for analogous elements between cultures and exchange near-matches in meaning to "scramble" the identity of a person based on how they express their meaning. Such a tool may be a great boon to individuals communicating in situations which may pose a risk to themselves—for example, in-culture informants reporting on human rights violations within an autocratic regime.

7.10 Summary of Contributions

I have demonstrated that, beyond the motifs used for the majority of this work, there exists a category of narrative elements that I refer to as motif-likes, which share similarities to other theories of narrative structure. Further, I demonstrate that these motif-likes can be reliably identified by human annotators ($F_k > 0.55$).

Finally, I propose that there is a continuum of communicative narrative elements which serve to transmit and express ideas related to culture and identity that individuals who share in that identity are able to understand to a greater degree. I describe, in depth, four axes I have identified, which strongly suggest deep interconnections between theories of narrative, culture, and knowledge representation.

CHAPTER 8

Future Work

One substantial area of future work is the continued development and improvement of automatic motif detection systems: the ultimate goal is a fully-generalizable system that does not depend on human input. To achieve this goal, not only does the performance of the current system need to be improved and made more generally applicable, but there is a necessity to automate the curation of lists of potential motif candidates. While one possibility is the automatic parsing of motif indices, which has had limited success in the past, I believe it preferable to develop a system to automatically cluster anomalous narrative elements that seem to operate in the same information-rich manner that motifs do. Some of my early work [YF16a] suggests one potential procedure by which this might be achieved.

I believe that what I refer to as the "continuum of cultural communication" can be further formalized and explored through subsequent annotation studies and analyses of different fields; in particular, one avenue which I believe worthy of exploration is the evolution of communicative narrative elements over time. Does the complexity of the motif-like scheme, which fills a hybrid gap in between motifs and several other theories of narrative structure, lessen over time, as a culture's age and grounding in a substantial body of work smoothen out odd narrative structures? Further, is there something substantially different between cultural group that are grounded in specific stories (e.g. Irish folklore) and cultural groups with no specific grounding (e.g. internet subcultures such as the incel community)? Understanding these relations and deepening the theory of how communicative narrative elements relate has implications for story understanding, information extraction, and narrative generation, and no doubt any other field that depends on textual understanding.

There is, of course, much that can build on top of the work done here, such as the inclusion of a deeper level of cultural knowledge into artificially intelligent systems. This has implications for information extraction, question answering, and commonsense reasoning, enabling such systems to operate in a more culturally aware manner

by being able to utilize motifs and other narrative structures that embody cultural commonsense.

Finally, I believe that, once such systems mature, there is much work that can be done in using the way we, as humans, use narrative elements to communicate our cultural identity to identify things such as disinformation via anomalous identity signals and automated spam via a lack of cohesion in identity signals. Beyond that, the deep relation that each person has to their cultural identity has implications for automatic anonymization and de-anonymization.

CHAPTER 9

Contributions

Some of my early work, in close collaboration with my advisor, Mark Finlayson, answered two key questions regarding Propp's morphology: first, that for both single-move and multi-move tales, that given Propp's general approach, list of functions, and functions identified in specific tales, independent annotators agree with each other as to where and whether those functions appear in the tales Propp suggests; second, that given Propp's general approach and list of functions, independent annotators agree not only with each other, but also agree with Propp, when asked to find his functions in tales.

Further, we suggested that Propp's theory, as it stands, is broadly correct, but does not capture all elements present in the corpus of tales he annotated. We perform a preliminary analysis of the corpus, using the annotations produced from this study to analyze the uncovered areas, and hypothesize and present the elements that Propp may have missed.

One of the major contributions of this work is a large-scale annotation study which serves to demonstrate exactly that—that human annotators can reliably agree with a second, independent annotator on how to annotate candidate motif phrases with a high degree of agreement ($F_k > 0.55$). This suggests the importance of motifs generally and confirms that they have meaning that can be reliably identified and extracted by members of a specific group.

Additionally, this annotation study resulted in 21,123 annotations after agreement had reached a substantial level, which were used for the development, training, and testing of the system described in §**??**.

Finally, this annotation study allowed for the identification of a trend in editorial vs. non-editorial articles with respect to motifs, hinting at their importance especially in the field of understanding articles based around human-to-human communication and opinion discourse than factual reporting.

I have also demonstrated that humans can reliably learn and annotate news articles with van Dijk's theory of news discourse with a high degree of agreement, developed a system that can predict the document-level discourse labels for paragraphs within a news article with reasonable performance (65% of human performance), and generated a gold-standard corpus of these labels, along with an annotation guide, to support future work.

The main contribution of this work is the development of a motif detection pipeline and a characterization of its performance on different classes of motif, as well as a characterization of its potential for generalizability. The selection of and evaluation of features provides a path forward for the refinement of future motif detection models and suggests areas in which substantial improvements to the model can be made to increase performance.

I have demonstrated that, beyond the motifs used for the majority of this work, there exists a category of narrative elements that I refer to as motif-likes, which share similarities to other theories of narrative structure. Further, I demonstrate that these motif-likes can be reliably identified by human annotators ($F_k > 0.55$).

Additionally, I propose that there is a continuum of communicative narrative elements which serve to transmit and express ideas related to culture and identity that individuals who share in that identity are able to understand to a greater degree. I further describe four axes that I believe exist in this continuum and the place of various narrative elements upon those axes.

BIBLIOGRAPHY

[AAHC69] R.E. Alegria, R.E. Alegría, L. Homar, and E. Culbert. *The Three Wishes: A Collection of Puerto Rican Folktales*. Harcourt, Brace & World, 1969.

[Aar10] Antti Amatus Aarne. *Verzeichnis der Märchentypen*. Suomalainen tiedeakatemia, 1910.

[Afa57a] Aleksandr Nikolaevich Afanas'ev. *Narodnye Russkie Skazki*. Moscow: Gos. Izd-vo Khudozh Lit-ry., 1957.

[Afa57b] Aleksandr Nikolaevich Afanas'ev. *Narodnye Russkie Skazki*. Moscow: Gos. Izd-vo Khudozh Lit-ry., 1957.

[Ari05] María Arinbjarnar. *Murder She Programmed: Dynamic Plot Generating Engine for Murder Mystery Games*. Thesis, Reykavik University, 2005.

[ATF21] Anurag Acharya, Kartik Talamadupula, and Mark A. Finlayson. Toward an atlas of cultural commonsense for machine reasoning. In *Proceedings of the Workshop on Common Sense Knowledge Graphs (CSKGs)*, Online, February 2021. Held in conjunction with the 35th AAAI Conference on Artificial Intelligence (AAAI 2021).

[Bak] Mikhail Mikhaĭ Bakhtin. *The dialogic imagination: Four essays*.

[Bar66] Roland Barthes. Structural analysis of narratives. *Image, music, text*, 1966.

[Bar72] Roland Barthes. Mythologies, trans. *Annette Lavers (New York: Hill and Wang, 1972)*, 151:82, 1972.

[BB98a] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer, 1998.

[BB98b] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics, 1998.

[Bek06] Mesfin Awoke Bekalu. Presupposition in news discourse. *Discourse & Society*, 17(2):147–172, 2006.

[Bel94] Allan Bell. Telling stories. *Media texts: Authors and readers*, pages 100–118, 1994.

[Bel98] Allan Bell. The discourse structure of news stories. *Approaches to media discourse*, pages 64–104, 1998.

[BFKL12] Rens Bod, Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe. Objectivity and Reproducibility of Proppian Narrative Annotations. In Mark Alan Finlayson, editor, *Third Workshop on Computational Models of Narrative (CMN)*, pages 17–21, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[BL01] Steven Bird and Mark Liberman. A formal framework for linguistic annotation. *Speech communication*, 33(1-2):23–60, 2001.

[Bob66] Inger Margrethe Boberg. *Motif-index of early Icelandic literature*. Munksgaard, 1966.

[Bro15] Mark Brown. World's first computer-generated musical to debut in London. *The Guardian*, 2015.

[Bur09] Richard Francis Burton. *The Arabian nights*. Barnes & Noble, 2009.

[BVPR84] Claude Bremond, Jean Verrier, Thomas G Pavel, and Marylin Randall. Afanasiev and propp. *Style*, pages 177–195, 1984.

[BYA+20] Deya Banisakher, W Victor Yarlott, Mohammed Aldawsari, Naphtali Rishe, and Mark Finlayson. Improving the identification of the discourse function of news article paragraphs. In *1st Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020)*, 2020.

[Cam08] Joseph Campbell. *The hero with a thousand faces*, volume 17. New World Library, 2008.

[CCMB14] Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284, 2014.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[Coh68] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[Col73] Benjamin N Colby. A Partial Grammar of Eskimo Folktales. *American Anthropologist*, 75:645–662, 1973.

[Cox93] Marian Roalfe Cox. *Cinderella: Three hundred and Forty-Five Variants of Cinderella, Catskin, and Cap o'Rushes*, volume 31. Folklore Society, 1893.

[Cro52] Tom Peete Cross. *Motif-index of early Irish literature*. Indiana University, 1952.

[Dar10] Sándor Darányi. Examples of Formulaity in Narratives and Scientific Communication. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 29–35, 2010.

[Dav16] Davies, Mark. Corpus of news on the web (now). 2016.

[Dav17] Ernest Davis. Logical formalizations of commonsense reasoning: a survey. *Journal of Artificial Intelligence Research*, 59:651–723, 2017.

[Del00] Judy Delin. *The language of everyday life: An introduction*. Sage, 2000.

[DEL+10] Thierry Declerck, Kerstin Eckart, Piroska Lendvai, Laurent Romary, and Thomas Zastrow. Towards a Standardized linguistic annotation of fairy tales. In *Workshop on Language Resource and Language Technology Standards*, pages 60–63, 2010.

[DF12] Sándor Darányi and László Forró. Detecting Multiple Motif Co-occurrences in the Aarne-Thompson-Uther Tale Type Catalog: A Preliminary Survey. *Anales de Documentación*, 15(1), 2012.

[DL11] Thierry Declerck and Piroska Lendvai. Linguistic and semantic representation of the thompson's motif-index of folk-literature. In *Research and Advanced Technology for Digital Libraries*, pages 151–158. Springer, 2011.

[DLD12] Thierry Declerck, Piroska Lendvai, and Sándor Darányi. Multilingual and Semantic Extension of Folk Tale Categories. In *Proceedings of the 2012 Digital Humanities Conference (DH 2012)*, 2012.

[DM15] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[DS10] Thierry Declerck and Antonia Scheidel. An information extraction approach to the semantic annotation of folktales. In *First Interna-*

tional AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, Vienna, Austria. University of Szeged, Hungary, 2010.

[DSL10] Thierry Declerck, Antonia Scheidel, and Piroska Lendvai. Proppian content descriptors in an augmented annotation schema for fairy tales. *ECAI 2010*, page 35, 2010.

[Dun62] Alan Dundes. From etic to emic units in the structural study of folktales. *The Journal of American Folklore*, 75(296):95–105, 1962.

[Dun64] Alan Dundes. *The Morphology of North American Indian Folktales*. Folklore Fellows Communications, 1964.

[Dun65] Alan Dundes. On computers and folk tales. *Western Folklore*, 24(3):185–189, 1965.

[Dun97] Alan Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, pages 195–202, 1997.

[DWF12] Sándor Darányi, Peter Wittek, and László Forró. Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways. In Mark A. Finlayson, editor, *Third Workshop on Computational Models of Narrative (CMN)*, pages 2–10, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[Erl65] Victor Erlich. *Russian formalism: history-doctrine*. Morton & Co., second edition, 1965.

[FBY92] William B Frakes and Ricardo Baeza-Yates. *Information retrieval: data structures and algorithms*. Prentice Hall PTR, 1992.

[FC02] Chris R Fairclough and Pádraig Cunningham. An Interactive Story Engine. In *Proceedings of the 13th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2002)*, pages 171–176, 2002.

[FC03] Chris R Fairclough and Padraig Cunningham. A Multiplayer Case Based Story Engine. In *Proceedings of the 4th International Conference on Intelligent Games and Simulation (GAME-ON 2003)*, pages 41–47. EUROSIS, 2003.

[FC04] Chris R Fairclough and Pádraig Cunningham. AI Structuralist Storytelling In Computer Games. In *Proceedings of the 5th International Conference on Computer Games: Artificial Intelligence, Design and Education (CGAIDE 2004)*. University of Wolverhampton, 2004.

[FE16] Mark Alan Finlayson and Tomaz Erjavec. Overview of Annotation Creation: Processes & Tools. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer, 2016.

[Fin08] Mark Alan Finlayson. Collecting Semantics in the Wild: The Story Workbench. In Jacob Beal, Paul Bello, Nick Cassimatis, Michael Coen, and Patrick Winston, editors, *Proceedings of the AAAI Fall Symposium on Naturally Inspired Artificial Intelligence (published as Technical Report FS-08-06, Papers from the AAAI Fall Symposium)*, volume 1, pages 46–53, Arlington, VA, 2008. AAAI Press, Menlo Park, CA.

[Fin11a] Mark A Finlayson. The story workbench: An extensible semi-automatic text annotation tool. In *Intelligent Narrative Technologies*, 2011.

[Fin11b] Mark Alan Finlayson. The Story Workbench: An Extensible Semi-Automatic Text Annotation Tool. In Emmett Tomai, Jonathan P. Rowe, and David K. Elson, editors, *Proceedings of the 4th Workshop on Intelligent Narrative Technologies (INT4)*, pages 21–24, Stanford, CA, 2011. AAAI Press, Menlo Park, CA.

[Fin12] Mark Mark Alan Finlayson. *Learning narrative structure from annotated folktales*. PhD thesis, Massachusetts Institute of Technology, 2012.

[Fin15] Mark A Finlayson. Propplearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory. *Digital Scholarship in the Humanities*, page fqv067, 2015.

[Fin16] Mark Alan Finlayson. Inferring Propp's Functions from Semantically-Annotated Text. *Journal of American Folklore, Special Issue on Computational Folkloristics*, 129(511):53–57, 2016.

[Fis63] J L L B Fischer. The Sociopsychological Analysis of Folktales. *Current Anthropology*, 4(3):235–295, 1963.

[FL13] Bernhard Fisseni and Faith Lawrence. A Paradigm for Eliciting Story Variation. In *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 100–105. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[GB01] Dieter Grasbon and Norbert Braun. a morphological approach to interactive storytelling. In *Proceedings of cast01, Living in Mixed Re-*

alities, pages 337–340. FhG - Institut Medienkommunikation (IMK), German Federal Ministry of Education and Research, 2001.

[GDAP+05] Pablo Gervás, Beln Daz-Agudo, Federico Peinado, Raquel Hervás, and BelÃ©n DÃaz-Agudo. Story plot generation based on CBR. *Knowledge-Based Systems*, 18(4-5):235–242, 2005.

[Gee73] Clifford Geertz. *The interpretation of cultures*, volume 5043. Basic books, 1973.

[Ger13] Pablo Gervás. Propp's Morphology of the Folk Tale as a Grammar for Generation. In Mark A Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister, editors, *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 106–122. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.

[Ger16] Pablo Gervás. Computational drafting of plot structures for russian folk tales. *Cognitive computation*, 8(2):187–203, 2016.

[GHLG16] Pablo Gervás, Raquel Hervás, Carlos León, and Catherine V. Gale. Annotating Musical Theatre Plots on Narrative Structure and Emotional Content. In Ben Miller, Antonio Lieto, Rémi Ronfard, Stephen G. Ware, and Mark A. Finlayson, editors, *7th Workshop on Computational Models of Narrative (CMN 2016)*, volume 53 of *OpenAccess Series in Informatics (OASIcs)*, pages 1–16, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[GLM15] Pablo Gervás, Carlos León, and Gonzalo Méndez. Schemas for Narrative Generation Mined from Existing Descriptions of Plot. In *Proceedings of the 6th Workshop on Computational Models of Narrative (CMN'15)*, pages 54–70, 2015.

[GPC89] Algirdas Julien Greimas, Paul Perron, and Frank Collins. On meaning. *New Literary History*, 20(3):539–550, 1989.

[Gre83] A J Greimas. *Structual Semantics: An Attempt at a Method*. University of Nebraska Press, Lincoln, Nebraska, 1983.

[Gut73] Norbert Guterman. *Russian Fairy Tales*. Pantheon Books, 1973.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[HCGJ11] Jeffry R Halverson, Steven R Corman, and HL Goodall Jr. *Master narratives of Islamist extremism*. Palgrave Macmillan, 2011.

[HdAL21] A. Hurley, C.R.R. de Arellano, and S.R. Lamarche. *The Mythology and Religion of the Tainos*. Independently Published, 2021.

[HHF05] Knut Hartmann, Sandra Hartmann, and Matthias Feustel. Motif definition and classification to structure non-linear plots and to control the narrative flow in interactive dramas. In *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, pages 158–167. Springer, 2005.

[Ike71] Hiroko Ikeda. *A type and motif index of Japanese folk-literature*. Orient Cultural Service, 1971.

[IO12a] Shohei Imabuchi and Takashi Ogata. A story generation system based on propp theory: As a mechanism in an integrated narrative generation system. In *Advances in Natural Language Processing*, pages 312–321. Springer, 2012.

[IO12b] Shohei Imabuchi and Takashi Ogata. Story generation system based on propp theory as a mechanism in narrative generation system. In *Digital Game and Intelligent Toy Enhanced Learning (DIGITEL), 2012 IEEE Fourth International Conference on*, pages 165–167. IEEE, 2012.

[IO13] Shohei Imabuchi and Takashi Ogata. Methods for generalizing the propp-based story generation mechanism. In *International Conference on Active Media Technology*, pages 333–344. Springer, 2013.

[Jac01] Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[Jac12] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50, 1912.

[Jas07] Heda Jason. About 'motifs','motives','motuses','-etic/s','-emic/s', and 'allo/s-', and how they fit together. an experiment in definitions and in terminology. *Fabula*, 48(1-2):85–99, 2007.

[JMYF20] Labiba Jahan, Rahul Mittal, W Victor Yarlott, and Mark Finlayson. A straightforward approach to narratologically grounded character identification. In *28th International Conference on Computational Linguistics (COLING 2020)*, 2020.

[JYRF21] Labiba Jahan, W Victor Yarlott, Mittal Rahul, and Mark A Finlayson. Confirming the generalizability of a chain-based animacy detector. In *1st Workshop on Artificial Intelligence for Narratives (AI4N 2020)*, 2021.

[KAA+77] Sheldon Klein, JOHN F Aeschlimann, MATTHEW A Appelbaum, DF Blasiger, ELIZABETH J Curtis, MARK Foster, SD Kalish, SJ Kamin, YD Lee, LA Price, et al. Modeling propp and lévi-strauss in a metasymbolic simulation system. *Patterns in Oral Literature*, pages 141–222, 1977.

[Kir71] Bacil F Kirtley. *A motif-index of traditional Polynesian narratives*. University of Hawai'i Press, 1971.

[KKM+12] FB Karsdorp, P Kranenburg, Theo Meder, Dolf Trieschnigg, and A Bosch. In search of an appropriate abstraction level for motif annotations. In *Proceedings of the 2012 Workshop on Computational Models of Narrative*, 2012.

[Kle75] Sheldon Klein. Meta-compiling text grammars as a model for human behavior. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 84–88. Association for Computational Linguistics, 1975.

[KvdMMvdB15] Folgert Karsdorp, Marten van der Meulen, Theo Meder, and Antal van den Bosch. Momfer: A search engine of thompson's motif-index of folk literature. *Folklore*, 126(1):37–52, 2015.

[Lab97] William Labov. Some further steps in narrative analysis. 1997.

[Lak72] George Lakoff. Structural complexity in fairy tales. *The study of man*, 1972.

[LDD+10] Piroska Lendvai, Thierry Declerck, Sándor Darányi, Pablo Gervás, Raquel Hervás, Scott Malec, and Frederico Peinado. Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1996–2001. European Language Resources Association (ELRA), 2010.

[LDDM10] Piroska Lendvai, Thierry Declerck, Sándor Darányi, and Scott Malec. Propp revisited: Integration of linguistic markup into structured content descriptors of tales. In *Proceedings of the Conference for Digital Humanities 2010*, 2010.

[LHS+18] Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, 2018.

[LMMF18] Justin Ling, Jill Mahoney, Patrick McGuire, and Colin Freeze. The 'incel'community and the dark side of the internet. *The Globe and Mail*, 24, 2018.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LS55] Claude Lévi-Strauss. The structural study of myth. *The Journal of American Folklore*, 68(270):428–444, 1955.

[LS63] Claude Lévi-Strauss. Structural analysis in linguistics and in anthropology. *Structural anthropology*, 1:31–54, 1963.

[LS84] Claude Levi-Strauss. Structure and Form: Reflections on a Work by Vladimir Propp. In Vladimir Propp, editor, *Theory and History of Folklore*, chapter 11, pages 167–210. University of Minnesota Press, Minneapolis, MN, 1984.

[LVDD10] Piroska Lendvai, Tamás Váradi, Sándor Darányi, and Thierry Declerck. Assignment of character and action types in folk tales. *Selected Papers from the NooJ*, pages 102–111, 2010.

[LW97] William Labov and Joshua Waletzky. Narrative analysis: Oral versions of personal experience. 1997.

[LZT+21] Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. How is bert surprised? layerwise detection of linguistic anomalies, 2021.

[Mah13] Sarah J Mahler. *Culture as Comfort: Many things you know about culture (but might not realize)*. Pearson Education, 2013.

[Mal01] Scott A Malec. Proppian structural analysis and xml modeling. *Proc. of Computers, Literature and Philology (CLiP 2001)*, 2001.

[Mal10] Scott Malec. Autopropp: Toward the automatic markup, classification, and annotation of russian magic tales. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 112–115, 2010.

[MBP04] Isabel Machado, Paul Brna, and Ana Paiva. 1, 2, 3.... action! directing real actors and virtual characters. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 36–41. Springer, 2004.

[McC89] John McCarthy. Artificial intelligence, logic and formalizing common sense. In *Philosophical logic and artificial intelligence*, pages 161–190. Springer, 1989.

[Min74] Marvin Minsky. A framework for representing knowledge, 1974.

[MPB01] Isabel Machado, Ana Paiva, and Paul Brna. Real characters in virtual stories. In *Virtual Storytelling Using Virtual Reality Technologies for Storytelling*, pages 127–134. Springer, 2001.

[MPP01] Isabel Machado, Ana Paiva, and Rui Prada. Is the wolf angry or... just hungry? In *Proceedings of the fifth international conference on Autonomous agents*, pages 370–376. ACM, 2001.

[MSB+14a] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[MSB+14b] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[MWK+05] Cynthia Matuszek, Michael Witbrock, Robert C Kahlert, John Cabral, Dave Schneider, Purvesh Shah, and Doug Lenat. Searching for common sense: Populating cyc from the web. *UMBC Computer Science and Electrical Engineering Department Collection*, 2005.

[Noy54] Dov Neuman Noy. *Motif-index of Talmudic-Midrashic literature*. Indiana University, 1954.

[ODR13] Nir Ofek, Sándor Darányi, and Lior Rokach. Linking Motif Sequences with Tale Types by Machine Learning. In Mark A Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister, editors, *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 166–182, Hamburg, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Par11] Parker, Robert, Graff, David, Kong, Junbo, Chen, Ke, and Maeda, Kazuaki. English gigaword fifth edition. 2011.

[PEW72] Vladimir Jakovlevič Propp, Karl Eimermacher, and Christel Wendt. *Morphologie des Märchens*. Carl Hanser Verlag, Munich, Germany, 1972.

[PG04] Federico Peinado and Pablo Gervás. Transferring game mastering laws to interactive digital storytelling. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 48–54. Springer, 2004.

[PG05] Federico Peinado and Pablo Gervás. Creativity Issues in Plot Generation. In *Working Notes on Workshop on Computational Creativity, at 19th International Joint Conference on Artificial Intelligence (2nd IJWCC '05)*, pages 45–52. Departamento de Ingeniera del Software e Inteligencia Artificial, Universidad Complutense de Madrid, 2005.

[PG06] Federico Peinado and Pablo Gervás. Evaluation of automatic generation of basic stories. *New Generation Computing*, 24(3):289–302, 2006.

[PGDA04] Federico Peinado, Pablo Gervás, and Belén Díaz-Agudo. A description logic ontology for fairy tale generation. In *Language Resources for Linguistic Creativity Workshop, 4th LREC Conference*, pages 56–61. Citeseer, 2004.

[PK93] Zhongdang Pan and Gerald M Kosicki. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75, 1993.

[Pro66] Vladimir Propp. *Morfologia della Fiaba*. Guilio Einaudi Editore, Turin, Italy, 1966.

[Pro68a] Vladimir Propp. *Morphology of the Folktale*, volume 9. University of Texas Press, 1968.

[Pro68b] Vladimir Propp. *The Morphology of the Folktale (2nd ed.)*. University of Texas Press, Austin, TX, 1968.

[Pro69] Vladimir Propp. *Mofologija skazki*. Nauka Verlag, Moscow, Russia, 1969.

[Pro70] Vladimir Propp. *Morphologie du conte*. Editions Gallimard, 1970.

[PS13] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly, Sebastopol, CA, 2013.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[RB03] Thomas Rieger and Norbert Braun. Narrative use of sign language by a virtual character for the hearing impaired. *Computer Graphics Forum*, 22(3):651–660, 2003.

[RRKS77] J. Ramírez-Rivera, B. Klein, and J. Slemko. *Puerto Rican Tales: Legends of Spanish Colonial Times*. Ediciones Libero, 1977.

[ŘS10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[RSS17] Afrooz Rafiee, Wilbert Spooren, and José Sanders. Culture and discourse structure: A comparative study of dutch and iranian news texts. *Discourse & Communication*, page 1750481317735626, 2017.

[Rut] Ruthenia. S. Thompson. Motif-index of folk-literature. `http://www.ruthenia.ru/folklore/thompson/`. Accessed: 2016-03-09.

[SA75] Roger C Schank and Robert P Abelson. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157, 1975.

[SD10] Antonia Scheidel and Thierry Declerck. Apftml-augmented proppian fairy tale markup language. In *Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, page 95, 2010.

[SF79] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[SFH+20] Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, 2020.

[SGBI02] Ulrike Spierling, Dieter Grasbon, Norbert Braun, and Ido Iurgel. Setting the scene: playing digital director in interactive storytelling and creation. *Computers & Graphics*, 26(1):31–44, 2002.

[Sil15] Max Silberztein. *La formalisation des langues: l'approche NooJ*. ISTE éd., 2015.

[SPT+12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, 2012.

[Tho60] Stith Thompson. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*, volume 4. Indiana University Press, 1960.

[Tho77] Stith Thompson. *The folktale*. Univ of California Press, 1977.

[Tho10] Craig Michael Thomas. *The Algorithmic Expansion of Stories*. Thesis, Queen's University, 2010.

[Tom87] Silvan Tomkins. Script theory. the emergence of personality. eds. joel arnoff, ai rabin, and robert a. zucker, 1987.

[TVdM18] Niket Tandon, Aparna S Varde, and Gerard de Melo. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record*, 46(4):49–52, 2018.

[Uni10] Princeton University. About wordnet, 2010. Retrieved on May 9, 2016 from: http://wordnet.princeton.edu.

[Uth04] Hans-Jörg Uther. *The types of international folktales: a classification and bibliography, based on the system of Antti Aarne and Stith Thompson*. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, 2004.

[V́24] Roman M Vólkov. *Shazka. Rozyskanija po sjužetosloženiju narodnoj skazki.*, volume 1. Ukrainian State Publishing House, 1924.

[VBA+95] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics, 1995.

[VD86] Teun A Van Dijk. *Studying Writing: Linguistic Approaches. Written Communication Annual: An International Survey of Research and*

Theory Series, Volume 1., chapter News Schemata, pages 155–185. ERIC, 1986.

[vD88] Teun A van Dijk. *News as Discourse*, chapter Structure of News, pages 52–57. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1988.

[vdVBB+11] Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171, 2011.

[vR79] Cornelis J. van Rijsbergen. *Information retrieval*. Butterworths, London Boston, 1979.

[Whi98] Peter R White. *Telling media tales: The news story as rhetoric*. Department of Linguistics, Faculty of Arts, University of Sydney, 1998.

[YCGF18] W Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33, 2018.

[YF16a] W Victor H Yarlott and Mark A Finlayson. Learning a better motif index: Toward automated motif extraction. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[YF16b] W. Victor H. Yarlott and Mark A. Finlayson. ProppML: A Complete Annotation Scheme for Proppian Morphologies. In Ben Miller, Antonio Lieto, Rémi Ronfard, Stephen G. Ware, and Mark A. Finlayson, editors, *7th Workshop on Computational Models of Narrative (CMN 2016)*, volume 53 of *OpenAccess Series in Informatics (OASIcs)*, pages 1–19, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[YH03] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, 2003.

[YOA+21] W. Victor H.. Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrado, Diana Gomez, David Mcdonald, Chris Miller, and Mark Alan Finlayson. Finding trolls under bridges: Preliminary work on a motif detector. *Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems*, 2021.

APPENDIX: MOTIF ANNOTATION GUIDE V2.2

Purpose of the Project

In 2002, Osama bin Laden called George W. Bush the "pharaoh of the century." However, bin Laden did not intend to refer to Bush as a great leader nor associate him with the impressive works of the sphinx or pyramid. This begs the question: what did bin Laden mean when referring to Bush as a pharaoh?

When he made this statement, Osama bin Laden was making a reference to a story from the Qur'an describing the oppression of the Israelites by the Egyptians. In context, the pharaoh is an oppressive ruler who enslaves and dominates others and is considered a great enemy.

In calling George W. Bush the "pharaoh of the century," bin Laden intends to condemn Bush as the worst oppressor of his people that has been seen in the past one hundred years. Note that without this information, we would be unable to understand both the content of this message (that Bush is considered an oppressor) and the cultural group for whom this message was intended.

This cultural information is captured in what we call motifs: repeated, cultural "memes" grounded in a well-known story. Other examples of motifs that you may see in modern usage may be a "knight in shining armor" or a "troll under a bridge."

The high-level scientific question of this project revolves around this: how do we use computers to track this cultural information automatically? To answer this question, we need to build systems that attempt to track this information. To build these

systems, we require data to train and test models of how cultural information occurs in narratives.

Specifically, we require data with the cultural information already identified by a human. The process of generating this data is called **annotation**.

What is Annotation?

Annotation is the process of making implicit information explicit by applying a descriptive or analytic notation to raw language data. In this project, we are performing *semantic text annotation*—this means we are explicitly noting the meaning of the language rather than the form. In this case, we are annotating : actors, events, props, etc., that occur within a written text. In this experiment, our annotations will identify the occurrence and location of motifs in narratives, mapping these motifs to Thompson's Motif-index of folk-literature when possible, and noting when there is no match.

What are you going to do?

For this project, you will be asked to correct annotations generated by a high-recall lexical matcher. As the lexical matcher creates annotations based only on the text, it generates many potential annotations that are incorrect. Your task is to look at these annotations and determine, in-context, if the annotation is correct.

The Concept of a Motif

Stith Thompson is a well-known scholar of folklore who most prominently revised and added to a system of folktale classification known as the tale type index (originally created by Antti Aarne) and developed an indexing system for narrative elements known as motifs. Thompson described motifs as follows:

> Stith Thompson, *The Folktale*, pp.415–416 A *motif* is the smallest element in a tale having a power to persist in tradition. In order to have this power

> it must have something unusual and striking about it. Most motifs fall into three classes. First are the actors in a tale—gods, or unusual animals, or marvelous creatures like witches, ogres, or fairies, or even conventionalized human characters like the favorite youngest child or the cruel stepmother. Second come certain items in the background of the action—magic objects, unusual customs, strange beliefs, and the like. In the third place there are single incidents—and these comprise the great majority of motifs. It is this last class that can have an independent existence and that may therefore serve as true tale-types. By far the largest number of traditional types consist of these single motifs.

Motifs concisely embody a large amount of culturally-relevant information: for example, if we say that a dragon arrives at a castle and seeks the princess, what might we assume is going to happen? People from Western cultures will no doubt call to mind some or all of the following scenes:

1. An evil dragon (B11.9—*Dragon as power of evil.*)
2. captures a princess (R11.1—*Princess (maiden) abducted by monster (ogre).*)
3. and a hero fights to rescue the princess (B11.11.4—*Dragon fight in order to free princess.*).

This story is almost the canonical example of a hero tale in Western folklore, so it may seem only natural that it springs to mind when we begin to talk about dragons. However, people from Asian cultures might have a different story in mind:

1. The king of all dragons (B11.12.5—*The dragon-king.*)
2. arrives at a castle and transforms himself into a man (B11.5.1—*Dragon's power of self-transformation.*)
3. in order to marry the princess (B11.12.7—*Human-dragon marriage.*).

Any of the motifs listed above are somewhat entangled with each other and with many other motifs. While the above examples are folklore-themed, motifs are present in news, literature, press releases, and many other sources of narrative. This is only a single example of the wealth of culturally-relevant knowledge that is contained within motifs. The ability to automatically identify these motifs within a story opens up that wealth of information for analysis, leading to the ability to have machines that can reason about culture and its relationship with narrative.

The purpose of this annotation is to verify that identified instances of motifs are, in fact, being used in a motific way. The identified instances come from a lexical matcher that uses a bank of terms to search for potential motif instances. These verified instances are then used to test a system that can automatically identify motifs in text.

A Stricter Definition of Motif

The most basic definition of a motif is that it is "a repeated cultural element grounded in a well-known story." Popular examples in modern culture are the troll under the bridge, the knight in shining armor, and the flying magic carpet. More specifically, we define motifs as:

Motif A set of closely-related variants of a non-commonplace, specific element that is repeated across tales of the same type or from the same folkloric tradition.

This definition captures the most important aspects of motifs: they are narrative elements that appear in tales of the same type (e.g., in variants of Cinderella, we expect to see an evil stepmother [S31—*Cruel stepmother.*]) or in the same culture (across Native American tales, Coyote/Old Man Coyote as a trickster [A177.1—*God as dupe or trickster.*] is common). They persist across generations and as the tale they are in travels across cultures. They encapsulate a set of variants: a fox bringing its master a measure for money (unlisted in Thompson's motif index) and a cat bringing its master a measure for money (K1954.1—*Helpful cat burrows measure for his master's money.*)

are undoubtedly related. They are non-commonplace and specific: a table is not worthy of noting, but a table with endless food and drink is notable (D1472.1.7—*Magic table supplies food and drink.*), and rather than noting that there is a dragon (B11—*Dragon.*), it is better to note that there is a dragon with three heads (B11.2.3.2—*Three-headed dragon.*).

There are three types of motifs:

Character Animate entities with cultural significance.

Prop Inanimate entities that fulfill a role in a cultural story.

Event Commonly occurring incidents within a culture.

Importantly, motifs are typically metaphoric in usage. The same phrase used for a motif may be **unrelated** to the cultural group or it may be **referential**: a direct reference to the story or a definition based on the story. Additionally, a motif may be an **eponym** for a business or group, intended to invoke the associations generally, but not used in a metaphoric way within a conversation. Consider the following different usages of the term "troll":

Unrelated And I just... it's a really sensitive subject and I appreciate the people get **trolled** all the time online, me included.

Reference Do you know the story "Three Billy Goats Gruff"? There is **a troll that lives under the bridge** and he is preventing the three billy goats from crossing the bridge to get to greener pastures.

Eponym So I had been hearing that since Annette and Jack Slocum took over **the Grumpy Troll restaurant** and brewery in March, the menu had been expanded and the place was looking better than ever.

Metaphoric Op-ed: PATENT Act is **a bridge over patent trolls**

Detailed Description of Motif Types

Thompson defines three categories of motifs: the actors, items in the background of the action such as magic objects and unusual customs, and single incidents. In this guide we refer to these as **Actors**, **Props**, and **Events**, respectively.

Actor An actor is a character who drives some action forward. They might be the hero, or a dragon, or an evil witch, so long as they exist within the story and perform some sort of action. This may include seemingly mundane characters: for example, the stepmother who refuses to allow her stepdaughter to go to the ball (S31—*Cruel stepmother.*).

Prop Thompson treats props as a catch-all category for any motif that is neither an actor nor an event. Examples of types of props you may encounter include: magical objects (D1500.1.15—*Magic healing ring.*), customs or beliefs (C168—*Tabu: giving younger daughter in marriage before elder.*), and unusual locations (F223—*Fairy hall.*).

Event Events are "single incidents"—for example, a hero fighting a dragon to rescue a princess (B11.11.4—*Dragon fight in order to free princess.*), a person substituting a rabbit for themselves to win a race (K11.6—*Race won by deception: rabbit as "little son" substitute.*), or a witch flying around on a broomstick (G242.1—*Witch flies through air on broomstick.*).

Annotation Procedure

What to Annotate

Annotation is relatively simple. For each pre-tagged motif, verify whether the motif is used in a way that invokes its associations. If it does not, change the tag from the motif to the most relevant non-motific tag: **unrelated** for completely unrelated usage (e.g., a last name like Dove); **referential** for direct reference to the story in which the motif is grounded without the use of simile or metaphor to link it to another subject

or a definition of the motif based on the story; or **eponym** for names of things like businesses, groups, locations, etc. that may be based in the motif but aren't clearly invoking associations.

When annotating, you should make sure to take into account all the context that you can: this includes the name of the file and the rest of the text in the section containing the motif: these often contain important contextual clues.

Keep These In Mind

From the definition provided, there are four questions to keep in mind while annotating, roughly in order of importance:

1. Is the motif being used to invoke its associations?
2. Is this usage of the motif a literal reference or being used metaphorically?
3. Is this usage of the motif related to the culture?
4. Does this motif match the expected type of character, prop, or event?

(1) is the key question in determining if a usage is motific: without invoking the associations of a motif, its usage is at most referential. One way to determine this is to put yourself in the head of the author and ask what purpose the usage is within the text: if it is to draw out the meaning of the motif, typically relating it to something, then it is motific.

(2) helps determine if a culturally relevant motif instance is actually being used as a motif or if it is a direct reference to the original source. This is, in conjunction with (1), useful for differentiating between motific usage and referential usage.

(3) and (4) help to filter out spurious, non-culturally relevant matches.

When in doubt, refer to this annotation guide and the motif catalog at the end of this guide.

Special Cases and Specific Considerations

Definitions

The line between definitions (which are to be marked as **referential**) and motific usage can be somewhat unclear. As a general rule, if the marking in question directly describes the motif without the motif being used to invoke associations in some way (e.g., the definition is provided as part of a simile or metaphor), it is definitional.

Eponym vs. Motif

Motifs used as names (e.g., the Shamrock Bar) should be marked as **eponym**. In the event that a motif is invoked alongside a name (e.g., "Living up to its name, James Murphy won the lottery off a ticket given to him at the Shamrock Bar."—in this case, invoking the "lucky" aspect of the motif), then this would be considered motific.

When the motif is based on a real-world item (e.g., a shamrock), references to the real-world item with no specifically cultural context should be considered unrelated. For example, a lucky shamrock might be motific, but a botanical description of the lesser clover commonly called a shamrock is unrelated.

Motific Iconography

The usage of a motif as iconography (a logo or symbol), e.g. a shamrock on a letter or a harp on a flag, is considered motific.

Motifs in New Fiction

Generally, the repurposing of motifs for use in newer fiction should be considered motific: whether faithful to or adding a "new twist" to a motif, the intention is typically to invoke those associations and either draw from them or subvert them.

Holidays and Rituals

Where a holiday is not clearly motific (e.g., a usage like "a Christmas miracle"), they should be counted as referential if they are highly specific, or unrelated if general (e.g., the 4th of July).

Where a ritual is not clearly motific, it should always be counted as referential, as rituals are typically too culturally-specific to be unrelated.

Is it "kiddush wine" or "Kiddush Wine"?

When in doubt about whether something is being named for a motif or the motif is part of a descriptor for that thing (most common in holiday/ritual cases, such as "kiddush wine"), assume that they are being used referentially. However, clues such as capitalization and nearby context should always be taken into account.

Motif Art as Motific

As per our consideration of whether or not a motif is being used to invoke associations, artistic works should be considered motific, as they are intended to evoke the associations in viewers familiar with them.

"Too Culturally Specific"

For some motifs, we have no examples of an origin or adaption of the motif outside of the culture, and this is noted in the motif catalog. For these motifs, the "unrelated" tag should not be used.

A Guide to BRAT

For this project, you will be using the BRAT Rapid Annotation Tool (BRAT). The following is a basic tutorial for how to use BRAT in the Cognac Lab. Further instructions can also be found on the BRAT homepage: `http://brat.nlplab.org/`.

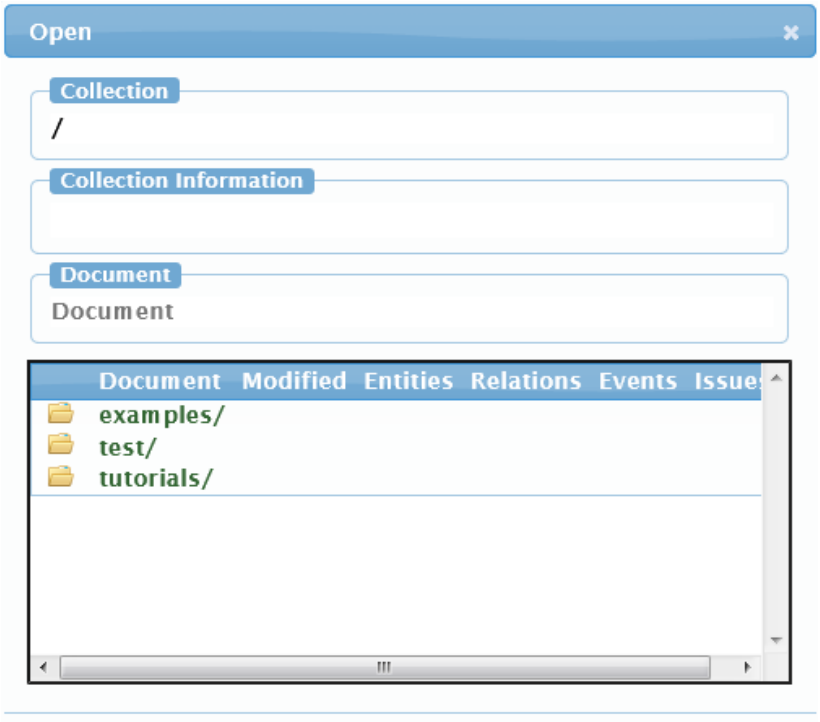

Figure 9.1: Next, open up the correct data path through the file browser. This will be given to you as part of later instructions.

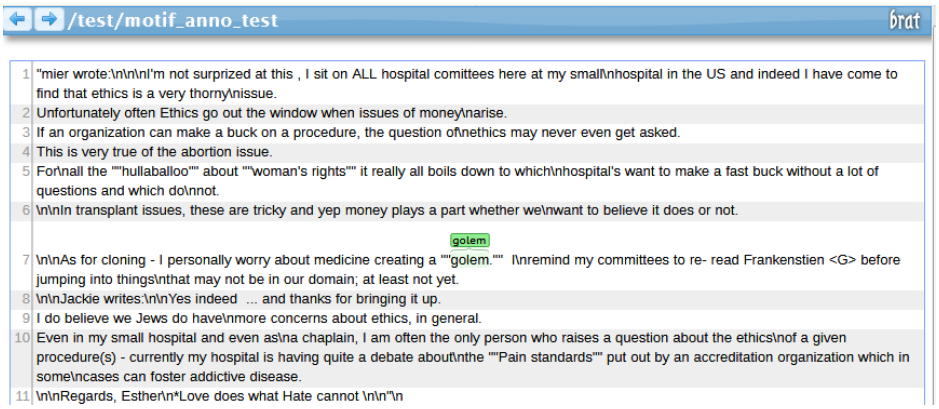

Figure 9.2: You should now see a screen showing you pre-annotated data.

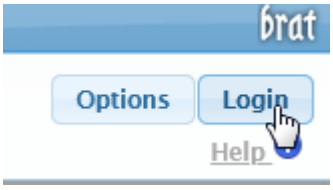

Figure 9.3: Hover over the "brat" text in the top right and choose login.

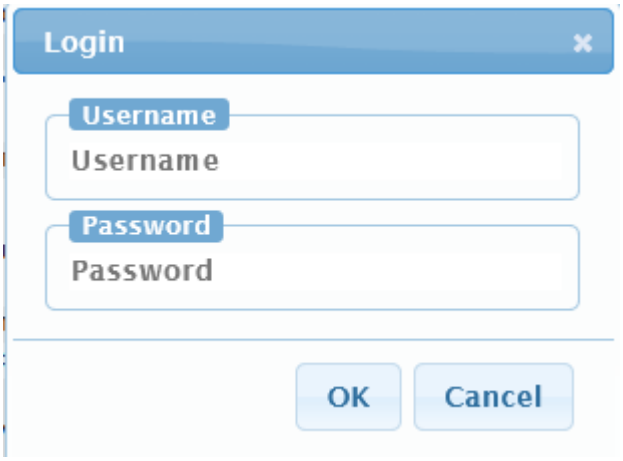

Figure 9.4: Enter your annotator credentials. This will be given to you as part of your personalized instructions.

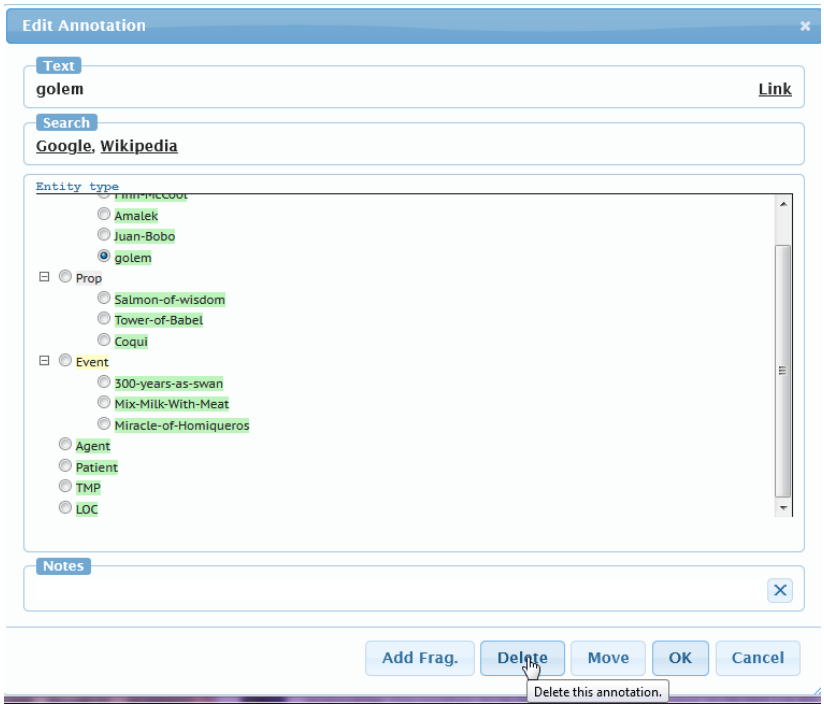

Figure 9.5: Double-clicking on an annotation allows it to be edited and deleted.

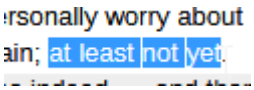

Figure 9.6: Highlighting text...

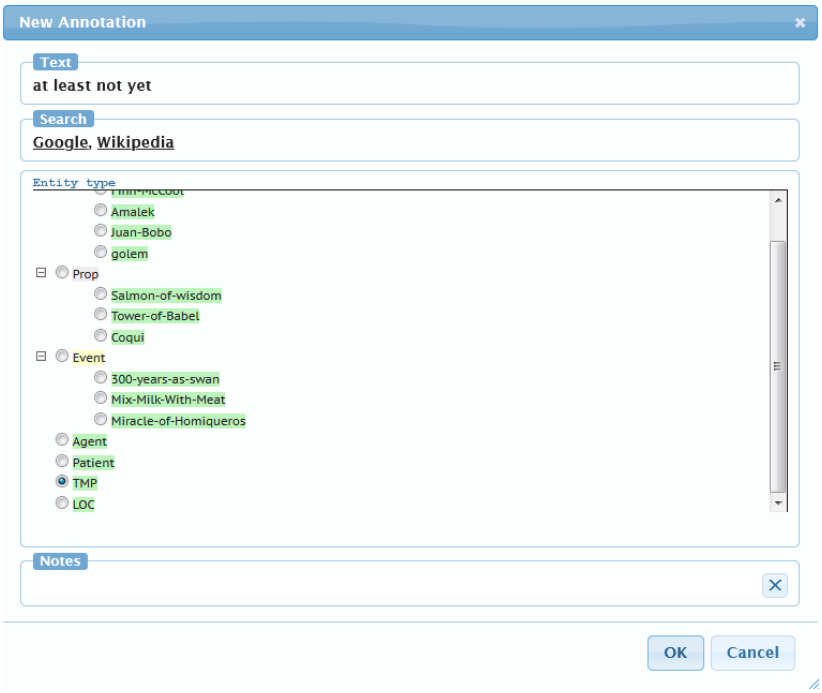

Figure 9.7: ...opens up an annotation dialogue for that text.

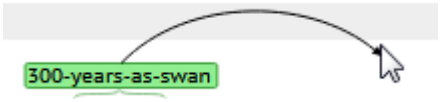

Figure 9.8: Clicking and dragging from one annotation to another...

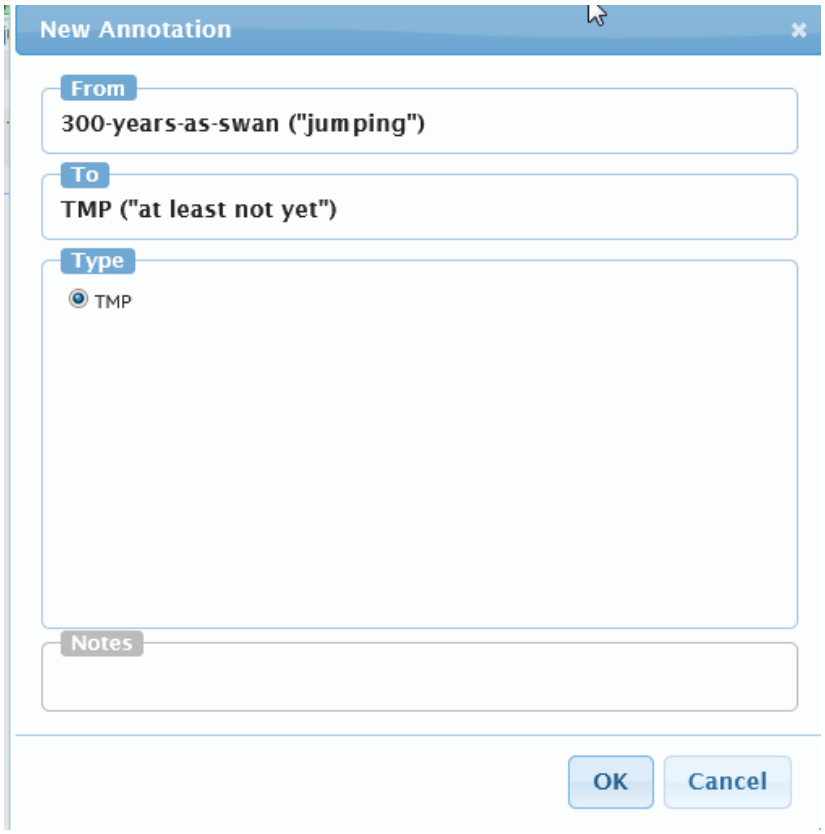

Figure 9.9: ...allows the addition of relationship annotations.

Motif Catalog

All motifs should contain as much of the following information as possible:

Motif Name	**Motif Type**
Definition	Motif Definition
Motific Example	Motif Example
Referential Example	Reference
Unrelated Example	Unrelated use
Eponym Example	Eponym

Irish Motifs

Note: where not surrounded by quotations the examples are written by the collector and are not real-world examples.

Salmon of Wisdom/Knowledge	**Prop**
Definition	Has associations with (1) wasting one's life on a fool's errand (in the story, the person who found the salmon of wisdom lost its powers to a servant boy who tasted it while cooking it), as well as (2) associations with having a lot of knowledge.
Motific Example	(1) "trying to find the salmon of knowledge. More u turns than a dodgy plumber!!" (2) "at this point i need a lick off the salmon of knowledge to pass the pre's any1 know where i can find him these days"
Referential Example	"The Salmon of Knowledge is a creature figuring in the Fenian Cycle of Irish mythology."
Unrelated Example	"And then you see the guy wearing his Salmon of Knowledge costume and it's lost the humor."
Eponym Example	"Ordered the Salmon of Knowledge and Fish 'n Chips, the food was fantastic and we licked our plates clean!"

Children of Lir	**Event**
Definition	May also be represented as "300 years as a swan." From the children of Lir, associated with (1) jealousy, (2) tragic loss, and (3) imprisonment/a terrible fate.
Motific Example	(1) "I'm getting huge Children of Lir vibes from these concept photos, especially with their forlorn looks." (2) "very sad news from Blacksod where rescue heli came down while we slept, scene of great beauty & tragedy,Children of Lir buried on Inisglora" (3) "You'll end up like the children of Lir."
Referential Example	"The Children of Lir (Irish: Oidheadh chloinne Lir) is a legend from Irish mythology."
Unrelated Example	The children of Lir, 32, from Ireland, became famous due to a viral video on YouTube.
Eponym Example	"The Children of Lir is the debut album by Irish folk rock/progressive folk group Loudest Whisper."

Finn McCool	**Character**
Definition	Incredibly smart, powerful savior of many people. Considered a protector of the land.
Motific Example	"When Fionn MacCumhaill was keeping his social distance from the Druid Fineglas, he ate the salmon of knowledge - be like Fionn - be wise"
Referential Example	"Fionn mac Cumhaill (usually rendered as Finn McCool or Finn MacCool in English) was an Irish mythical hunter-warrior in Irish mythology, occurring also in the mythologies of Scotland and the Isle of Man."
Unrelated Example	N/A—All unrelated examples are likely to be eponyms (e.g., unrelated real-life people).
Eponym Example	"Find out what happened when Gordon Ramsay visited Finn McCool's and read about when and why Finn McCool's closed."

Wren	**Prop**
Definition	Associated with (1) secrecy and (2) something that must be protected.
Motific Example	(1) I was able to get tickets to the sold-out concert because I'm one of the wren-boys. (2) We need to shelter children like wrens from harmful media.
Referential Example	"Well, the history of the Wren predates Christmas, its origins are in Irish mythology where birds held great prominence."
Unrelated Example	"Wrens are a family of brown passerine birds in the predominantly New World family Troglodytidae."
Eponym Example	"Wren Bistro & Bar is a hip neighborhood restaurant that boasts a relaxing, yet upscale environment."

Magic Harp	**Prop**
Definition	A symbol of Ireland, sometimes representing creativity with some associations with magical harps, such as the one owned by The Dagda, an Irish god.
Motific Example	In the Ireland vs. France match, Quinn was really playing The Dagda's harp.
Referential Example	"According to ancient Celtic folk tales, the very first harp was owned by "the Good God" Dagda."
Unrelated Example	"The harp is a stringed musical instrument."
Eponym Example	"At The Harp Irish Pub, our classic comfort food will nourish your body and your soul."

Tir na nog	**Prop**
Definition	The land of youth; often represents something you will never find.
Motific Example	My car keys are "off in Tir na nog."
Referential Example	In the story, a man went to stay in Tir na nog, leaving behind the love of his life.
Unrelated Example	N/A
Eponym Example	"Welcome to the website of Tir na nÓg Philly. Take a tour of our site to learn about the Irish culture."

Shamrock	**Prop**
Definition	Symbol of Ireland and (1) representative of luck, often used as "the luck of the Irish." Sometimes used in (2) reference to St. Patrick's usage of the shamrock to illustrate the holy trinity.
Motific Example	(1) Since Bill won the lottery, he must've been wearing a shamrock. (2) Allie's presentation was her shamrock, winning over the sponsors.
Referential Example	"A shamrock is a young sprig, used as a symbol of Ireland. Saint Patrick, Ireland's patron saint, is said to have used it as a metaphor for the Christian Holy Trinity."
Unrelated Example	Shamrock refers to either the lesser clover or the white clover, in addition to some other three-leaved plants.
Eponym Example	"Shamrock features apartments with custom-made furnishings complete with all house wares."

Leprechaun	**Character**
Definition	Tricky, grumpy fairies who steal treasure and horde it. Associated with (1) greed, (2) trickery, and (3) being short.
Motific Example	(1) That old miser is a real leprechaun. (2) That leprechaun at the used car lot really got the better of me. (3) I swear I could jump over Ethan, he's a real leprechaun.
Referential Example	A leprechaun is a type of fairy of the aos si in Irish folklore.
Unrelated Example	N/A—most uses are going to be at least somewhat culturally related.
Eponym Example	"The photo you see for Leprechaun, Inc. is a 5,000 year old Dolman, or Portal Tomb, built during the Neolithic Period."

King Conchobar	**Character**
Definition	A king of Ulster, known for bravery and great leadership.
Motific Example	Ellen, our new CEO, has affected our company as if King Conchobar had taken over.
Referential Example	Conchobar mac Nessa ruled from Emain Macha.
Unrelated Example	Most unrelated examples are likely to be eponyms (e.g., unrelated real-life people).
Eponym Example	"3 local business owners recommend Conchobar Consulting."

Fairy Fort	**Prop**
Definition	Rings in the land that are raised up, considered incredibly dangerous and sacred.
Motific Example	Talking about her ex is like messing with a fairy fort.
Referential Example	"Don't mess with fairy forts."
Unrelated Example	N/A—too culturally specific.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Aos Si	**Character**
Definition	Magical creatures that live under fairy forts, considered dangerous and not to be messed with.
Motific Example	I wouldn't get involved with Tim, he's a real-life aos si.
Referential Example	"The aos si is the Irish term for a supernatural race in Irish and Scottish mythology."
Unrelated Example	N/A—too culturally specific.
Eponym Example	"Book Aos Si Lodges Glencoe, Ballachulish on Tripadvisor."

Banshee	**Character**
Definition	A mythical creature, known for its screech. Refers to people who scream, screech, and are loud or angry. Typically used to refer to females.
Motific Example	When I got home late, my wife turned into a banshee.
Referential Example	"A banshee is a female spirit who heralds the death of a family member."
Unrelated Example	"Banshee is an American action television series,"
Eponym Example	"Banshee is a metal band that was formed in 1986 in the American Midwest."

Cu Chulainn	**Character**
Definition	An Irish demigod, born Setanta, who got his name after killing Culann's guard dog. Associated with strength as Ireland's greatest warrior.
Motific Example	Did you see how many boxes Avery was carrying by himself? He's a real Cu Chulainn.
Referential Example	"In Irish mythology, Cúchulainn ('Hound of Culann') is the pre-eminent hero of Ulster in the Ulster Cycle."
Unrelated Example	N/A—too culturally specific.
Eponym Example	"The Order of CúChulainn is the highest award for adults in Scouting Ireland."

Jewish Motifs

Note: these examples are mostly generated by the people who collected the motifs and are not real-world examples from text.

Haman	**Character**
Definition	Haman was hung on his own gallows after falsely accusing the Jewish people; used to suggest someone (esp. someone perceived as anti-Jewish or anti-Israel) will receive a fitting retribution.
Motific Example	Sorry, Bibi: Iran is bad, but it is no Haman.
Referential Example	A story in the Bible found in Esther talks about Haman who had conspired to have a Jewish man named Mordecai impaled on a "fifty cubit" pole.
Unrelated Example	The family moved to Seoul from a small town of Haman.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Golem	**Character**
Definition	A magic automaton; used either to suggest soullessness or someone or something driven by a singular purpose (often in a negative way).
Motific Example	"This is a terribly sad day. Ashley is a golem, driven by nothing more than greed and a love of money."
Referential Example	A golem is a clay creature that has been magically brought to life.
Unrelated Example	N/A—too culturally specific.
Eponym Example	"The Golem", directed by Doron and Yoav Paz, released in 2018.

Amalek	**Character**
Definition	The devil; used to refer to someone who intends harm to Jewish people.
Motific Example	Sorry, Bibi: Iran is bad, but it is no Amalek.
Referential Example	After the victory over Amalek, Moses built an altar which showed his gratitude to God.
Unrelated Example	N/A—too culturally specific.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Kiddush	**Event**
Definition	A prayer; more commonly invoked as "kiddush hashem," which is representing God, the Torah, and the Jewish people in the best light possible.
Motific Example	This article is a great kiddush hashem.
Referential Example	Next time Israel launches a spy satellite, recite kiddush and pray to God it succeeds.
Unrelated Example	N/A—too culturally specific.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Milk with Meat	**Event**
Definition	Two things that don't, or shouldn't, go together. Usually only used with direct references.
Motific Example	N/A — no example found yet.
Referential Example	Chukim are laws which have no explanation attached to them such as the prohibition against eating milk with meat or wearing shaatnez.
Unrelated Example	"On a hot stove, cook the meat with milk, potatoes, and salt."
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

God's name in Vain	**Event**
Definition	Using the lord's name in vain
Motific Example	Thou shalt not take Paul Volcker's name in vain.
Referential Example	Saying God approves of wrongdoing is taking His name in vain.
Unrelated Example	N/A—don't seem to be any other usage of the phrase.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Opening the Ark of the Covenant	**Event**
Definition	An act which unleashes disaster; most references appear to be to the movie Raiders of the Lost Ark.
Motific Example	Diet tip: pretend your fridge is the Ark of the Covenant so you're never tempted to open it!
Referential Example	
Unrelated Example	That scene from Indiana Jones where they open the ark of the covenant and everbodys face melts? That's my life!
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Seventy Languages	**Prop**
Definition	God can speak "70 languages"; used as a prop to indicate something universally understandable, e.g. "yelling in seventy languages."
Motific Example	"*cries in seventy languages* love you so much jord , tysm for following me."
Referential Example	What were the names of the 70 languages that Moshe spoke?
Unrelated Example	Instead of tweeting you should be practicing how to say "I told you so" in seventy languages.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Leviathan/Behemoth	**Prop**
Definition	A gargantuan monster; references all use the term to represent size.
Motific Example	Facebook has turned into a Behemoth in less than two decades.
Referential Example	"God will construct canopies to shelter the righteous, who will eat the meat of the Behemoth and the Leviathan amid great joy and merriment."
Unrelated Example	N/A—too culturally specific.
Eponym Example	Behemoth is a Polish extreme metal band from Gdańsk, considered to have played an important role in establishing the Polish extreme metal underground.

Dove: Symbol of Israel	**Prop/Character**
Definition	Sometimes used to represent Israel; used both positively or satirically to refer to Israeli politicians.
Motific Example	Israel's former defense minister Shaul Mofaz is no dove; he explicitly advocated for the use of Palestinian civilians as human shields.
Referential Example	A Hebrew-language site, Sod1820, pointed out to the fact that Israel is sometimes referred to as a dove in literature.
Unrelated Example	People used doves as a messenger for a long time.
Eponym Example	N/A—don't seem to be any examples of businesses, etc. using the term.

Tower of Babel	**Prop**
Definition	A tower representing hubris of man; motif may be too widely-spread to be culturally relevant.
Motific Example	"Tower Of Babel: Part II, The Solar System! If we keep sending garbage into our Atmosphere, we will destroy Our World, and Our Universe!"
Referential Example	Tower of Babel was built in the land of Shinar some time after the Deluge
Unrelated Example	N/A—too culturally specific.
Eponym Example	The Tower of Babel by Bruegel is a masterpiece.

Puerto Rican Motifs

Note: these examples are mostly generated by the people who collected the motifs and are not real-world examples from text.

Reyes Magos/Three Kings	**Event/Character**
Definition	The three kings that arrive on a camel and leave presents in the traditional Catholic holiday Three Kings Day. Associated with Catholicism and presents.
Motific Example	It's January 6th! I hope the camels ate our grass and the three kings left presents.
Referential Example	Three Kings Day is a traditional Catholic Holiday celebrated on January 6th.
Unrelated Example	In English history, "The Year of the Three Kings" may refer to the years 1066, 1483, or 1936.
Eponym Example	Three Kings Consulting Firm is a business development and consulting firm for small and medium-sized businesses.

Agueybana	**Character**
Definition	The most powerful and well-known Taino chief/cacique known for doubting the Spaniards and leading a rebellion when they were beginning to conquer the island. Associations with rebellion and being the principal cacique.
Motific Example	Jose stood up to his teacher and got the whole classroom to start rebelling—he's a real Agueybana.
Referential Example	Agueybana was a principal Taino chief in Puerto Rican history.
Unrelated Example	N/A—will still be culturally specific referring to the Taino chief.
Eponym Example	The Agueybana de Oro was the most prestigious award in Puerto Rico's annual Superior Council of Arts awards.

Atabey	**Character**
Definition	Supreme goddess in Taino folklore. Associated with giving birth and fertility (said to have given birth to herself; often depicted in a birthing position).
Motific Example	Monica has given birth to triplets after already giving birth to twins – the spirit of Atabey is within her.
Referential Example	Atabey is the supreme goddess in Taino folklore.
Unrelated Example	N/A – will be eponyms
Eponym Example	Atabey Hospital is looking into expanding its facilities by building two new cancer research centers.

Coqui	**Prop**
Definition	Small frog species native to Puerto Rico. Known for being the national symbol of Puerto Rico.
Motific Example	You can't think about Puerto Rico without hearing the sound of coquis.
Referential Example	Coquis are a frog species native to Puerto Rico.
Unrelated Example	N/A – will be eponyms
Eponym Example	The Coqui soccer team won their first match today.

Roberto Cofresi	**Character**
Definition	Famous Puerto Rican pirate who was against the Spanish colonization of Puerto Rico. Purposely targeted American and Spanish ships due to personal despise. Known for inventing pina colada and for his Robin Hood-like attitude. Popular belief that his treasure is hidden in a cave in Cabo Rojo.
Motific Example	She made her pina colada so well that even Cofresi would be proud.
Referential Example	Roberto Cofresi was a famous pirate from Puerto Rico.
Unrelated Example	N/A – will be eponyms
Eponym Example	Come and relax with your family at the Cofresi Palm Beach Resort & Spa.

Our Lady of Providence/Divina Providencia	**Character**
Definition	Puerto Rico's main Virgin Mary figure (same concept as Mexico's Guadalupe). She was declared the patroness of Puerto Rico by Pope Paul VI. Her feast day is celebrated by many Puerto Ricans.
Motific Example	Despite the recurring earthquakes, Puerto Rico will always be safe as long as Our Lady of Providence is there.
Referential Example	Our Lady of Providence is a Virgin Mary title whose devotion originated in Italy.
Unrelated Example	N/A – will be eponyms
Eponym Example	Our Lady of Providence Church is open every day from 8am-7pm with daily mass starting at 10am.

Guanina	**Character**
Definition	Guanina was a Taino princess who fell in love with a Spanish officer. Her brother murders the Spaniard, Guanina is branded as a traitor, and she then dies by her lover. They are both buried together under a ceiba tree. Associations with love, betrayal, and being the Puerto Rican Romeo and Juliet.
Motific Example	Despite the growing opposition from both of their families, Roberto and Maria got married just like Guanina would have done.
Referential Example	In Puerto Rican folklore, Guanina was a Taino princess who fell in love with a Spanish officer.
Unrelated Example	Guanina is Spanish for 'Guanine' which is a nucleobase found in both DNA and RNA.
Eponym Example	N/A – not many examples of her name being used for establishments.

Hormigueros	**Event/Prop**
Definition	Municipality in Puerto Rico known as "El Pueblo de Milagros" (the town of miracles) due to folklore story titled the Miracle of Hormigueros. In the folklore story the mayor of the town finds his daughter, safe and unharmed, after searching for her for 15 days.
Motific Example	Nicolas have been searching for his dog for 15 days but still no luck. Perhaps if he was in Hormigueros he would've found him.
Referential Example	Hormigueros is one of the 78 municipalities in Puerto Rico.
Unrelated Example	Hormigueros is Spanish for 'ant-hill'.
Eponym Example	Hormigueros Airport is open 24/7 every day.

Juan Bobo	**Character**
Definition	Famous character in Puerto Rican folklore known for being dumb and lazy.
Motific Example	Gabriel fails all his classes and never does the chores – a real Juan Bobo.
Referential Example	Juan Bobo is a popular comedic character in Puerto Rican folklore.
Unrelated Example	N/A – will still be culturally specific referring to the character himself
Eponym Example	N/A – not many businesses named after him

Yocahu	**Character**
Definition	Yocahu is the supreme god and is the son of Atabey and the brother of Guacar. He is associated with cassava and is known for guarding the national park of Puerto Rico, el Yunque.
Motific Example	Cassava production will always be bountiful as long as Yocahu remains watching over them.
Referential Example	In Taino folklore, Yocahu is the supreme god and the son of Atabey.
Unrelated Example	N/A – All refer directly to the Taino character.
Eponym Example	N/A – Not many businesses named after Yocahu.

Data Preparation Considerations

Specific to Lexis Nexis

LexisNexis data can only be downloaded 100 articles at a time, does not filter duplicates particularly well, and can only provide articles as PDF, DOCX, or RTF format.

For the purposes of data collection, all data from LexisNexis was downloaded 100 at a time as RTF formats, non-combined, as a ZIP file. The ZIP files were processed with Python's zipfile library [1] and the documents are processed using a Python package [2] to strip the RTF information from the document.

Issues with BRAT

Before input into BRAT, several changes need to be made:

- Filenames cannot contain unicode and must not contain spaces.
- All files need to be encoded as UTF-8.
- Although UTF-8 seems to be supported, text generally works better if non-unicode. There is a Python library to decode unicode [3].
- Newlines must be standardized as only newlines (`\n`), not carriage return + newline (`\r \n`).

When generating annotations for BRAT, annotations are in the format: `T<#>\t<label> <start> <text>` (where \t is a tab character.) The label numbers must start at 1 and the raw text must exactly match the text that is present in the raw text file.

[1] `https://docs.python.org/3/library/zipfile.html`
[2] `https://pypi.org/project/striprtf/`
[3] `https://pypi.org/project/Unidecode/`

APPENDIX: INCEL ANNOTATION GUIDE V1.1

Purpose of the Project

Note: for the purposes of this appendix, repeated elements from the previous guide have been removed.

What are you going to do? (Pilot Edition)

For the pilot, it will also be helpful for you to provide input into the ideas contained within this document on the categories of motif-like elements (e.g., are the categories correct? How can the definitions be improved? Are there more categories?) and the definition of motif-like elements itself (especially with regard to the relation to a *canonical narrative*).

Content Warning

The "incel" community is a self-identifying online fringe group of self-proclaimed "incels" or involuntary celibates. Some of the content posted by members of this community can be exceptionally vitriolic, with racist, misogynistic, or generally misanthropic sentiments. If at any point during the annotation process you feel uncomfortable proceeding, it is recommended that you stop annotating.

Motifs and Motif-like Elements

In previous annotations, we have been concerned primarily with motifs, which we have defined as follows:

Motif A set of closely-related variants of a non-commonplace, specific element that is repeated across tales of the same type or from the same folkloric tradition. These elements can be prop, character, or event.

In contrast to this definition, the data collected from the incel community suggests that there are a wider range of narrative elements at work. Further, the incel community is rather new, with no real shared folkloric background. This seems to mean that there are less elements rooted in a specific *story*, although there are sources for some of the terminology (for example, alpha/beta/omega come from theories on pack hierarchy).

In lieu of a better term, we are currently calling these narrative elements motif-like. These motif-like elements fall under roughly four categories:

Props Inanimate objects that fulfill a role in a story.

Character Roles Animate entities that fulfill a general role within a story or have stereotypical traits.

Character Descriptors Descriptions of traits that a person within a story might have.

Event Scripts General plotlines describing stereotypical events that are considered common or relevant to the community.

However, it should be noted that some motif-like elements can act as these different categories: e.g., redpill might be a prop ("take the redpill"), a descriptor ("he's redpilled"), or an event script ("you need to redpill them").

From these categories, and from sampling of the data, we may begin to formulate a rough definition for the motif-like elements we encounter within the incel community. Interestingly, it appears that a majority of the motif-like elements from the incel community are centered around supporting what might be called a *canonical narrative*—in this case, that the reason for the members' status as involuntary celibates is rooted in a (usually) external problem. Thus, there exist what might be considered narrative-preserving narratives that use these motif-like elements.

Motif-like element A set of narrative elements that are related by their involvement in a central, *canonical* narrative important to the community and typically take

the form of fragmentary parts of the narrative (e.g. a role, a trait, an event, or a meaningful prop).

Like motifs, we anticipate that motif-like elements embody a large amount of culturally-relevant information. The purpose of this annotation is twofold. First, to verify that identified instances of motif-like elements are, in fact, being used in a motific way. The identified instances come from a lexical matcher that uses a bank of terms to search for potential motif-like instances. These verified instances are then used to test a system that can automatically identify motifs in text. Second, to determine the what additional forms motif-like elements might take and verify their relation to *canonical narratives.*

Specifics of Motif-like Elements

Beyond the four categories of motif-like elements (props, character roles, character descriptors, and event scripts), we anticipate that motif-like elements will fall under five different categories: **unrelated**, **referential**: which are direct references, **eponym**: a name based on the motif-like element, **figurative**: drawing upon the associations, or **grounded instance**: when a typically figurative term is used to refer to something specific and grounded in reality rather than used to allude to the qualities of the figurative motif; often, this may be referring to a community or a specific person. Consider the following different usages of the term "troll," a motif, as illustrations of these four categories:

Unrelated And I just... it's a really sensitive subject and I appreciate the people get **trolled** all the time online, me included.

Reference Do you know the story "Three Billy Goats Gruff"? There is **a troll that lives under the bridge** and he is preventing the three billy goats from crossing the bridge to get to greener pastures.

Eponym So I had been hearing that since Annette and Jack Slocum took over **the Grumpy Troll restaurant** and brewery in March, the menu had been expanded and the place was looking better than ever.

Figurative Op-ed: PATENT Act is **a bridge over patent trolls**

Grounded Instance So that's your trolling tactic, misinterpreting incels' posts to have positivity where there is none.

While *motifs* are typically metaphoric in usage, it's less clear that *motif-like* elements are always metaphoric. In some cases, they may be used as hypotheticals that aren't clearly metaphoric: e.g., "You should totally talk to that Stacy and redpill her." may suggest literal action (speaking to a girl and attempting to convince her of a specific viewpoint) but is deeply rooted in the *canonical narrative* that the speaker wishes to preserve.

Grounded Instance

Of particular interest is the grounded instance category for motif-like elements present in the Incel dataset. The grounded instance tag should be used when a motif-like element is used without any intention of referencing the associations typically associated with it. Most commonly, this is used to refer to something specific: a person, a group of people, or a community. What follows are some illustrative examples:

> All low LMS (looks, money, status) males are treated like dirt by female human organisms. An example is every **normie** and female on this sub, we repeatedly tell them to fuck off and they stay

Label: Grounded Instance — in this case, the usage of normie is clearly aimed at specific individuals (users on the forum). While referring to a vague group of normies might be motif-like, here it is just a reference to the group of people to whom that label might apply, with no further implication.

> EDIT: lmao he deleted the comment, what a **beta cuck**.

Label: Motif-like — here, while the term "beta cuck" is aimed at a specific person, its usage implies that the action of deleting his (contentious) comment is a result of him being a beta cuck: therefore, it is invoking motif-like associations of what it means to be a beta cuck.

> Even if you think the **incels** who post here are exaggerating about their own personal obstacles

Label: Grounded Instance — here, again, it refers to a specific group of posters, even if it is referring to

> I have a similar issue. Why are you asking **normies**, they'll just tell you to lift and grow a beard. And beards are fucking stupid and a sign of overcompensating so it's useless.

Label: Motif-like — while referring to a group, the advice given by these "normies" is considered useless. The unspoken association: that the lives of normies are so vastly different from the incel experience that their advice is completely misguided.

What is clear from these examples is that the distinction between grounded instances and motif-like elements is a thin, blurry line. While annotating, try to keep in mind the following question: **does this usage invoke, or does understanding it rely on, any associations that aren't present in the text?**

Annotation Procedure

What to Annotate

Annotation is relatively simple. For each pre-tagged motif-like element, verify whether it is being used in a way that invokes its associations. If it does not, change the tag to either **unrelated** for completely unrelated usage, **referential** for direct reference to the source of the motif (e.g. directly talking about the studies of pack hierarchy for alpha/beta), or **eponym** for names based on the motif-like element (e.g. /r/redpill).

After determining whether or not it is invoking the associations, next you need to determine what motif-like category it falls under and tag it as such: **character role**, **prop**, **character descriptor**, or **event script**.

When annotating, you should make sure to take into account all the context that you can: this includes the name of the file and the rest of the text in the section containing the motif: these often contain important contextual clues.

Keep These In Mind

From the definition provided, there are four questions to keep in mind while annotating, roughly in order of importance:

1. Is the motif-like element being used to invoke its associations?
2. Is this usage of the motif-like element related to the community's canonical narrative?

(1) is the key question in determining if a usage is appropriate: without invoking the associations of a motif-like element, its usage is at most referential. One way to determine this is to put yourself in the head of the author and ask what purpose the usage is within the text: is it to draw out the meaning of the motif-like element?

(2) helps to filter out spurious, non-relevant matches.

When in doubt, refer to this annotation guide and the motif-like element catalog at the end of this guide.

Special Cases and Specific Considerations

Grounded Instances: The Mythical CHAD and You

One particularly illustrative example of determining whether a usage of a motif-like element is figurative or not is the Chad character. One example of Chad as a grounded instance is:

> You could have played that last bit better. When the guy was joking about you not having a girlfriend you could have just played along and been like "yeah man you'd probably guess right" and just laughed along. Suddenly you are the guy who can take a joke (+1 socialization point with the Chad) and you also give the impression to the girl that you can take an insult and be upbeat and unscathed (+1 Socialisation+intrigue with the girl). Then you ask what his name is and put your hand out to shake his hand. He shakes it and says your pretty chill. You walk you the elevator together and say you have to go to wherever it is you are going. Wave them both goodbye and you have successful won the confrontation. The bf looks like the weaker person and you the stronger one.

In this case, the Chad being referred to is specific: it's someone making some sort of joke, who is accompanied by a girl. There's no hint of what makes him a Chad and the figurative aspect doesn't play in. In contrast, there is the Mythical CHAD:

> It's important to remember: Women don't actually want Chad to change. They like imagining what Chad "could be," not that actually becoming reality.

Here, Chad is not a specific person, but closer to a platonic ideal of the figurative element: the reality of Chad may be bad but women still desire him and have no intent to change him. Looking specifically to see how grounded a motif is is one way of determining when to use figurative vs. grounded instance markings.

Motif-like Catalog

All motif-likes should contain as much of the following information as possible:

Motif-like Name	
Motif-like Type	
Definition	Motif-like element Definition

Motif-like elements

Note: where not surrounded by quotations the examples are written by the collector and are not real-world examples.

Alpha	
Character Role, Character Descriptor	
Definition	The opposite of a beta male. Takes on risk and confrontation. Confident and a leader. Generally viewed with negativity, but may also embody traits that incels want to change about themselves.

Beta	
Character Role, Character Descriptor	
Definition	A weak male, opposite to an alpha. Can also mean subservience in general.

Chad	
Character Role	
Definition	A physically attractive male, typically white. Viewed with negativity from within the community. The object of desire for women.

Incel	
	Character Role
Definition	An involuntary celibate man. Refers to state in which a person who is willing and physically able to engage in sexual relations is unable to find a partner. The term applies to people who have not yet engaged in sex, those who have had sex at least once but are unable to find another partner, and those in a relationship with a partner who is unwilling to have sex. Used to describe member of the incel community, used as an insult outside of it.

Normie/Normalfag	
	Characer Role
Definition	An average, unextrordinary person. Typically used as an insult to deride people who have a normal life, friends, etc.

(Beta) Provider	
	Character Role, Event Script
Definition	A man who financially supports the woman in a relationship. It could also be a man who emotionally supports her by being overly romantic, sometimes to the detriment of his personal boundaries and to the health of the relationship. Considered something that women want after having relationships with "Chad", to settle down with someone who will give them money.

Cuck/Cuckold	
Character Role, Character Descriptor, Event Script	
Definition	A man with an unfaithful girlfriend/wife. Used as an insult. Sometimes refers to anyone who defends women, sometimes used as a generic insult. Can also talk about the actual meaning of cuckoldry; sometimes used as a portmanteau.

Manlet	
Character Role	
Definition	A short male. Used as an insult.

Betabucks/betabux/betabuxing	
Character Role, Event Script	
Definition	When a male financially provides for a partner. Sometimes used to refer to the male providing the money.

Cope	
Prop, Event Script	
Definition	Adopting an untrue belief, to avoid the pain that comes from facing a harsh, unpleasant reality. Sometimes used as a stand-alone "rebuttal" for posts to suggest that they are cope.

Omega (Male)	
Character Role, Character Descriptor	
Definition	A man considered to be on the lowest tier of social hierarchies. An insult. Sometimes used to describe actions seen as typical of an omega male.

Roastie	
Character Role	
Definition	A sexually promiscuous woman. An insult. Considered dumb, slutty, etc.

Sperg	
Character Role, Event Script	
Definition	A person with aspergers. Typically used as an insult, sometimes used to describe something that could be considered an episode for people with aspergers.

Take the -pill

-pill	
Prop, Character Descriptor, Event Script	
Definition	Refers to embracing a type of philosophy. Taken from the movie The Matrix. Used as a suffix, denotes a specific philosophy. Has several specific subtypes that follow ()

Redpill	
Prop, Character Descriptor, Event Script	
Definition	Considered to be accepting a harsh reality.

Bluepill	
Prop, Character Descriptor, Event Script	
Definition	Considered to be rejecting reality for a comforting or convenient lie.

Blackpill	
Prop, Character Descriptor, Event Script	
Definition	A fatalistic, depressed version of the redpill.

No Examples

Foid/Femoid	
Character Role	
Definition	A derogatory term for female. Stemmed from female and -oid (as in android or humanoid). Used to suggest that females are not fully human. Not always used as an insult, sometimes just used to describe women.

IncelTears/IT	
Character Role	
Definition	The subreddit inceltears, a virgin-shaming terrorist group on Reddit. Also just used to describe incels being upset.

Kissless, hugless, handholdless virgin (KHHV)	
Character Role, Character Descriptor	
Definition	Someone who is a virgin and has no experience with women. Considered a descriptor more than an insult.

	Landwhale
	Character Role
Definition	An overweight woman. Sometimes used as an insult, sometimes used as a generic term for overweight women. Generally considered unattractive.

	Neurotypical/NT
	Character Role
Definition	People without autism, anxiety, or other mental disability. May be used as an insult.

	Numale
	Character Role
Definition	A man who has feminist qualities. Used as an insult. Often thought of as balding. Associated with low testosterone/high estrogen.

	Stacy
	Character Role
Definition	Female counterpart of Chad. Generally hated within the incel community. Dates/paired with Chad, generally seen as unattainable by the incel community.

	Thirstie
	Character Role
Definition	An incel with an exceptionally high libido.

Volcel	
Character Role	
Definition	A voluntary celibate. Sometimes viewed with disdain from the community, as they aren't "denied" sex.

Wizard	
Character Role, Character Descriptor	
Definition	A man who is a virgin until the age of 30 (or beyond). Associated with a meme where remaining a virgin until 30 grants supernatural powers. Generally used as shorthand for being relatively old and still a virgin. Sometimes used to describe behavior that might result in being a virgin at this age.

Hypergamy	
Event Script	
Definition	Refers to a facet of evolutionary psychology suggesting women prefer to date above their league. Implies that women are never satisfied with their male partners and are looking for a "better" male.

Oneitis	
Character Role, Event Script	
Definition	To fall in love with someone. Sometimes used to describe the object of someone's attraction or unrequited love.

(High) Inhibition	
	Character Descriptor
Definition	A person who sacrifices their happiness due to fear that pursuing personal goals will cause people to view them negatively.

VITA

W. VICTOR H. YARLOTT

June, 2014	B.S., EECS Massachusetts Institute of Technology Cambridge, MA
June, 2014	M.Eng., EECS Massachusetts Institute of Technology Cambridge, MA
Jan.–June 2014	Teaching Assistant Massachusetts Institute of Technology Cambridge, MA
June–Aug. 2017	Summer Intern IBM Research Yorktown Heights, NY
June–Aug. 2016	Summer Intern IBM Research Yorktown Heights, NY
June–Aug. 2012	Summer Intern Vecna Technologies Cambridge, MA
June–Aug. 2011	Summer Intern MIT Lincoln Laboratory Lincoln, MA
Feb.–May 2011	Undergraduate Researcher MIT CSAIL: EVO-DesignOpt Group Cambridge, MA

PUBLICATIONS AND PRESENTATIONS

Yarlott, W. V. H., & Finlayson, M. A. (2016). *Learning a better motif index: Toward automated motif extraction.* In 7th Workshop on Computational Models of Narrative (CMN 2016). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Yarlott, W. V. H., & Finlayson, M. A. (2016). *ProppML: A complete annotation scheme for proppian morphologies.* In 7th Workshop on Computational Models of Narrative (CMN 2016). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Eisenberg, J. D., Yarlott, W. V. H., & Finlayson, M. A. (2016). *Comparing extant story classifiers: Results & new directions.* In 7th Workshop on Computational Models of Narrative (CMN 2016). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Yarlott, W. V. H., Cornelio, C., Gao, T., & Finlayson, M. (2018, August). *Identifying the discourse function of news article paragraphs.* In Proceedings of the Workshop

Events and Stories in the News 2018 (pp. 25-33).

Banisakher, D., Yarlott, W. V. H., Aldawsari, M., Rishe, N., & Finlayson, M. (2020, January). *Improving the identification of the discourse function of news article paragraphs.* In 1st Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020).

Jahan, L., Mittal, R., Yarlott, W. V. H., & Finlayson, M. (2020, December). *A straightforward approach to narratologically grounded character identification.* In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6089-6100).

Jahan, L., Yarlott, W. V. H., Rahul, M., & Finlayson, M. A. (2021, January). *Confirming the Generalizability of a Chain-Based Animacy Detector.* In 1st Workshop on Artificial Intelligence for Narratives (AI4N 2020).

Yarlott, W. V. H., Ochoa, A., Acharya, A., Bobrow, L., Estrada, D. C., Gomez, D., Zheng, J., McDonald, D., Miller, C., & Finlayson, M. A. (2021). *Finding Trolls Under Bridges: Preliminary Work on a Motif Detector.* In Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems.

Yarlott, W. V. H., Ochoa, A., Acharya, A., Bobrow, L., Estrada, D. C., Gomez, D., Zheng, J., McDonald, D., Miller, C., & Finlayson, M. A. (2021). *Arguing With Trolls: Motif Frequency and its Relation to Genre.* Abstract presented to the Literature & Culture and/as Intelligent Systems University of Stuttgart Digital Workshop.

Yarlott, W. V. H. & Finlayson, M.A. (April 23, 2018). Invited talk at the AI Culture Symposium at Morgan State University. Baltimore, MD.

Yarlott, W. V. H. (May 31, 2019). Invited talk at the Global AI for Good Summit.

Geneva, Switzerland.