GOLEM: GOld standard for Learning and Evaluation of Motifs

W. Victor H. Yarlott*, Anurag Acharya†, Diego Castro Estrada*, Diana Gomez*, Mark A. Finlayson*

*Florida International University
Knight Foundation School of Computing and Information Sciences
11200 S.W. 8th Street, Miami, FL 33199 USA
{wyarl001, dcast230, dgome133, markaf}@fiu.edu

†Pacific Northwest National Laboratory
Advanced Computing, Mathematics, and Data Division
Physical and Computational Sciences Directorate
902 Battelle Blvd, Richland, WA 99354 USA
anurag.acharya@pnnl.gov

Abstract

Motifs are distinctive, recurring, widely used idiom-like words or phrases, often originating from folklore, whose meaning are anchored in a narrative. Motifs have a significance as communicative devices across a wide range of media—including news, literature, and propaganda—because they can concisely imply a large constellation of culturally relevant information. Indeed, their broad usage suggests their cognitive importance as touchstones of cultural knowledge, and thus their detection is a step towards culturally aware natural language processing. We present GOLEM (GOld standard for Learning and Evaluation of Motifs) the first dataset annotated for motific information. The dataset comprises 7,955 English news articles, opinion pieces, and broadcast transcripts (2,039,424 words) annotated for motific information. The corpus identifies 26,078 motif candidates across 34 motif types drawn from three cultural or national groups: Jewish, Irish, and Puerto Rican. Each motif candidate is labeled according to the type of usage (MOTIFIC, REFERENTIAL, EPONYMIC, or UNRELATED), resulting in 1,723 actual motific instances in the data. Annotation was performed by individuals identifying as members of each group and achieved a Fleiss' kappa (κ) of > 0.55. In addition to the data, we demonstrate that classification of the candidate type is a challenging task for Large Language Models (LLMs) using a few-shot approach; recent models such as T5, FLAN-T5, GPT-2, and Llama 2 (7B) achieved a performance of 41% accuracy at best, where the majority class accuracy is 41% and the average chance accuracy is 27%. These data will support development of new models and approaches for detecting (and reasoning about) motific information in text. We release the corpus, the annotation guide, and the code to support other researchers building on this work.[1]

Keywords: corpus, digital humanities, other

1. Introduction and Background

Motifs can be simply described as recurring cultural "memes" that are grounded in a story. Motifs often originate in folklore, but can be found anywhere that language is influenced by culture. Motifs are highly prominent and ubiquitous, and they are interesting and useful because they provide a compact source of cultural information: they concisely communicate a constellation of related cultural ideas, associations, assumptions, and knowledge. Thus, the ability to automatically detect motifs would grant access to a large repository of important cultural information to computational analysis, which is as yet not easily accessible to computational language processing systems (Acharya et al., 2021).

One common western motif that illustrates the importance and information density of motifs is *troll under the bridge*. One folktale containing the motif, *The Three Billy Goats Gruff*, involves at one point a troll, hiding under a bridge, who tries to devour the goats as they try to cross. The motif is found across the folklore of Northern Europe, especially Norway. To members of many western cultures, invoking this motif brings a number of related ideas to mind that are by no means directly communicated by the surface meaning of the words: the bridge is along the critical path of the heroes, and they must cross it to achieve their goal; the troll lives under the bridge, surprising those who attempt to cross it; the troll tries to kill, eat, or otherwise extract some value from the would-be crossers; the troll is a squatter, not the officially sanctioned master of the bridge; and the troll usually meets his end at the hands the hero. The utility of the motif as a communicative device is clearly

[1]The corpus and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/FYOWLQ.

LREC-COLING 2024, pages 7801–7813
20-25 May, 2024.

visible in the common term *patent troll*, a person or organization that claims illegitimate ownership over ideas and attempts to extract value from companies who have related products related to those ideas. Here we see an analogical transfer of cultural attributes from the troll of folklore to the "troll" of patents.

Because of their striking nature and 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 (Thompson, 1960), which lists more than 46,000 motifs and sub-motifs found in tales from drawn from over 600 collections. Thompson informally defined 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." (Thompson, 1960, p. 19). He notes that motifs generally fall into one of three subcategories (Thompson, 1977, pp. 415–416): events, characters, or props. Examples of each type, respectively, would be a **hero rescuing a princess**, **Old Man Coyote**, and a **magic carpet**; we discuss these examples in more detail in Section 2.

Although the motif examples given so far are drawn from folklore, motifs have importance beyond folktales: they occur in modern stories, news articles, opinion pieces, press releases, propaganda, novels, movies, plays—indeed, anywhere that culture impinges on language. One powerful modern example is the use of the *Pharaoh* motif in modern Middle Eastern discourse. The Pharaoh, which refers to the Pharaoh who opposes Moses in the narrative found in Qu'ran, is an arrogant and obstinate tyrant who defies the will of God and is punished for it. In modern Islamist extremist discourse, the term *Pharaoh* has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush, the last of whom Osama bin Laden referred to as the "pharaoh of the century" (Halverson et al., 2011). In applying this motif to him, bin Laden intended to condemn Bush as the worst oppressor Islamic people had seen in the past one hundred years. Without understanding the implications of the *Pharaoh* motif, we would be unable to understand both the content of this message (that these leaders are being cast as oppressors) and the cultural group for whom this message was intended.

We present GOLEM (GOld standard for Learning and Evaluation of Motifs), an English-language corpus to enable the training and evaluation of automatic techniques for detecting motifs. Creating such a corpus is a challenging endeavor, not only because of the time and labor involved in linguistic annotation generally, but because identifying motifs must be done by experts in or natives to the relevant cultural or national group. To date, there is no such corpus, and as such, there have been few efforts to use motifs as part of natural language systems to better understand culturally inflected texts.

Identifying motifs is particularly challenging because most token sequences that match the surface form of the motif don't actually correspond to an invocation of the motif itself. One clear example is the *shamrock*. As a motif in the Irish context, *shamrock* implies luck, a relation to Ireland itself, or any reference to its usage by St. Patrick to represent the holy trinity. But it may well just refer to the literal plant named *shamrock*. Therefore, identifying motifs is much harder than simple text search, and requires a nuanced understanding of the meaning of the motifs in context. This represents a substantial challenge for automatic detection.

GOLEM comprises 26,078 motif candidates (surface forms that match a motif type in our target set), with 1,723 actual motif instances. These candidates are split into 9,620 Irish (159 motific), 7,858 Jewish (1,215 motific), and 8,600 Puerto Rican (349 motific) instances. The final average agreement for the human annotation that produced this data is $\kappa > 0.55$ for the Irish team and $\kappa > 0.7$ for the Jewish and Puerto Rican teams, showing that this is a task that humans can reliably annotate.

We use GOLEM to test four modern LLMs: T5 (Raffel et al., 2020), GPT-2 (Radford et al., 2019), FLAN-T5 (Chung et al., 2022), and Llama 2 (Touvron et al., 2023). We use 0-, 3-, and 5-shot prompting to evaluate the models' ability to determine the class for a motif candidate. These models achieve an accuracy no greater than 0.41%, which is equivalent to the majority class accuracy, demonstrating that while humans are able to reliably succeed at this task, current models without any training are not.

The paper is structured as follows. We first discuss related work, covering folkloristic and other definitions of motifs and prior computational work in the field of motifs (§2). We then describe the process by which we produced GOLEM (§3) and discuss the results of the annotation (§4). We describe in detail our experiments using LLMs (§5), outline potential future work (§6) and finally list our contributions (§7).

2. Related Work

2.1. Motifs in Folkloristics

Thompson informally defined 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" (Thompson, 1960, p. 19). In folklore, motifs are preferentially retained throughout retellings and recombinations of tales due to their striking nature and the density of information they communicate. Folklorists have long hypothesized that a tale's specific composition of motifs can be used to trace the tale's lineage (Thompson, 1977, 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 (Thompson, 1960). 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).

As mentioned above, Thompson noted that motifs generally fall into one of three subcategories (Thompson, 1977, pp. 415–416): events, characters, or props. Examples of these include (with their associated Thompson's motif code):

Hero rescuing a Princess (B11.11.4) This motif is perhaps one of the most well-known event motifs in western culture. Even children know the answers to the following questions: "A princess has been kidnapped: who kidnapped her, who rescues her and how?". Common answers will be "a dragon kidnapped her, a knight must rescue her, and he must kill the dragon." This motif may be the climax of the story, with a "happily ever after" following the hero's defeat of the dragon, or it may just happen in the course of a story: in *Ivan Dogson and the White Polyanin*, a Russian folktale (Afanas'ev, 1957, 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 multiple cultures.

Old Man Coyote (A177.1) This is a character motif: known in some Native American Indian tribes as merely 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.

Magic Carpet (D1155) This motif 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" (Burton, 2009, p. 496). Solomon, the third king of Israel, was said to have had 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.

While Thompson's index is the best known, there are many other motif indices focusing on specific cultures, national groups, and periods, for example, early Irish literature (Cross, 1952), traditional Polynesian narratives (Kirtley, 1971), or Japanese folk-literature (Ikeda, 1971). In addition, the idea of motif was incorporated into another useful notion, the *tale type*, which seeks to classify whole tales based on the collection of motifs present in them. Antti Aarne constructed an index of tale types in 1910 (Aarne, 1910), with translations and revisions by Thompson (1960) and Uther (2004) (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* (Thompson, 1977). 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 (Dundes, 1997). These motif indices provide a

substantial foundation 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 (Cross, 1952) and Dov Noy's Motif-index of Talmudic-Midrashic literature (Noy, 1954) to select our group of motifs.

2.2. Computational Approaches to Motifs

Declerck et al. (2012) worked on converting electronic representations of TMI and ATU to a format that enables multilingual, content-level indexing of folktale texts, building upon past work (Declerck and Lendvai, 2011). This work is focused on the descriptions of motifs and tale types, without reference to the stories.

Darányi (2010) called for attention to the automation of extraction and annotation of motifs in folklore, citing the generalizability of the idea, especially to modern forms of language such as scientific communication. Darányi et al. (2012) made headway towards using motifs as sequences of "narrative DNA", and Ofek et al. (2013) demonstrated learning tale types based on these sequences.

With regard to analyzing motif annotation schemes, Karsdorp et al. (2012) 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. Other work by Darányi and Forró (2012) suggested that motifs may not be the highest level of abstraction in narrative.

It is important to note, though, that none of these approaches provide an annotated corpus of motif usage in actual, modern text.

2.3. More Computationally Amenable Definitions of *Motif*

One necessary step towards effective computational approaches of motifs is providing a more formal definition for the term: 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 (Thompson, 1977, pp. 415–416).

This simple definition results in several problems, which have been well-known for some time (Propp, 1968, Chapter 1). More recently some researchers have repeated these complaints on the clarity of Thompson's definition of motifs, and have attempted to address these problems by providing a clearer definition (Jason, 2007). 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 several issues with this definition. First, Jason does not appear to define what simplest means beyond filling a slot of literary structure. Second, restricting motifs to characters or deeds ignores the importance of props within a story, such as *magic carpets* (D1155). Finally, context-free motifs ignore the vast wealth of cultural knowledge relevant to plots: to encapsulate cultural knowledge, motifs necessarily arise from related tales (a tale type) within a culture—undoubtedly, the previous example of the *Pharaoh* would lose much of its meaning and power were it robbed of its cultural context.

In more recent work, computationalists have provided a more precise definition of motif addressing these concerns, as follows:

> 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. (Yarlott and Finlayson, 2016)

Additional work by the same researchers has used this definition to develop and test a preliminary motif detection pipeline, achieving an F_1 of 0.35 on motifs and a macro-average F_1 of 0.21 across the four categories they chose using an off-the-shelf metaphor detector (Yarlott et al., 2022), showing the difficulty of the problem. This work leveraged metaphor detection due to the expectation for the usage of motifs "to primarily be figurative" (Yarlott et al., 2022, p. 9). This work was also made more difficult by the lack of comprehensive data available for this task. That work contained a small dataset of annotated motifs (5,006 candidates), but the data was not made available to the general research community.

Thus, until now, while there has been sporadic interest in motifs due to their versatility and ubiquity, there has been no effort to produce a corpus. GOLEM fills this gap.

3. Corpus Production Method

Our process for creating GOLEM comprised the following steps. First, given the nature of motifs

as being deeply embedded in the culture of specific groups, we first selected the cultural or national groups from which to draw motifs, then selected specific motif to annotate. As part of selecting these motifs, we consulted with informants who identified as members of those groups as to whether our selected motifs were recognizable and in current use. Next, we selected and acquired textual data containing the motifs, developed an annotation guide and scheme, selected an annotation tool, and created an annotation pipeline. We then sought and hired annotators who identified as members of the groups in question, and trained them according to the guide. After training the annotators performed the annotation, with adjudication meetings held weekly to go over each batch. We describe in detail each of these steps below, explaining at each step the considerations taken while making decisions in this process. One note is that subjectivity is an ever-present challenge in many annotation tasks, and throughout production of the corpus we aimed to address it through our clear annotation guide, inter-annotator agreement metrics, consultation with cultural informants, and use of double annotation and adjudication.

3.1. Selection of Cultural Groups

Before selecting motifs, it was important to identify groups from which to draw motifs. We had two criteria: first, there needed to be a strong authoritative source of motifs (a motif index, folklore collection, or something similar)—this made it substantially easier to identify motif candidates that may be interesting. Our starting seed was Thompson's Motif Index (Thompson, 1960), which provided the names of many other motif indices, allowing us to quickly expand our search.

The second criteria was that groups needed large populations near the authors or our collaborators. This restriction was deemed necessary for an ongoing survey study that was part of the same project, to ensure that we were easily able to find participants (although the COVID-19 pandemic quickly made these concerns obsolete).

The three cultural groups we selected that satisfied these criteria were *Irish*, *Puerto Rican*, and *Jewish*. For Irish, we used T.P. Cross's *Motif-Index of Early Irish Literature* (Cross, 1952) as a main source; for Puerto Rican, we drew motifs from S.R. Lamarche's *The Mythology and Religion of the Tainos* (Hurley et al., 2021), R.E. Alegría's *The Three Wishes: A Collection of Puerto Rican Folktales* (Alegria et al., 1969), and J. Ramírez-Rivera's *Puerto Rican Tales: Legends of Spanish Colonial Times* (Ramírez-Rivera et al., 1977); and for Jewish motifs, we referenced D.N. Noy's *Motif-index of Talmudic-Midrashic literature* (Noy, 1954).

3.2. Selection of Motifs

Once we selected the groups, we needed to identify individual motifs relevant to those groups. This was necessary as there are an unmanageably large number of motifs identified in the indices: Thompson's motif index alone lists over 46,000, drawn from many different cultures. To assemble a tractable list of motifs for annotation, we developed three selection criteria:

1. **Clearly identifiable source narrative**: By this we mean a well-known story of which we can find a telling drawn from the same body of folklore as other motifs in the group. This criteria is intended to provide evidence of relevance for the motif to the group, as well as a source for identifying potential associations that the motif calls to mind in group members. If a motif had no definitive source within the folklore of the group, it was excluded.

2. **In common use**: This criteria was intended to simplify the process of findings motifs widely known within the group. 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. If we couldn't find it in our target media sources, there was no point in including it in the study. We also consulted in-group informants to assess how well-known the motifs were: this was necessary to ensure that they actually meant something to in-group members. These individuals were contacted through our own individual networks and participated in brief interviews to discuss the motifs.

3. **Commonly used *qua* motif**: Even though the specific surface form of a motif might be in common use, it may not commonly be used to call to mind the cultural associations captured in the source narrative. This criterion was not hard and fast: it was a subjective judgement based on our observations of *how* the motifs were used when found in social media or news, as well as the discussions with in-group informants. If a motif seemed to be used in a way to allude to a motific associations (i.e., in a metaphorical or analogical fashion) rather than a simple reference or usage as a name, this suggested it's relevance for actual motific usage.

We used these criteria, in combination with the motif indices, to create a selection of motifs. During the selection, we aimed for a total of 30 motifs, roughly 10 from each group. From the initial selection phase, the following 34 motifs were chosen:

Irish (13) 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, the children of lir

Jewish (9) Haman, golem, Amalek, babel, leviathan/behemoth, 70 languages, name in vain, the ark of the covenant, kiddush hashem

Puerto Rican (12) Reyes Magos/Three Kings, Agueybana, Atabey, Roberto Cofresi, Divina Providencia, Guanina, Juan Bobo, Yocahu, the coqui, Hormigueros, jibaro/jibarito, chupacabra

Many motifs we considered were eliminated for a variety of reasons. For example: the Jewish motif *sukkot*, though listed in our reference Jewish motif index, 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 narrative; and the Puerto Rican motif *three camels come for grass on January 5th* was removed for being too difficult to find in modern texts. In practice, some motifs that we selected were also simply not present in the data: the Jewish motif *milk with meat* is one such case, where it was clearly recognizable to our informants, but simply did not appear when we searched for it.

3.3. Selection and Acquisition of Texts

We obtained texts through NexisUni[2], a university version of LexisNexis, a tool for searching through news articles, which provides world-wide scope for news and related text. We searched for motif terms and batch downloaded these articles, as allowed by the University's license. These articles were then further processed by a 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 for use by the annotators. In total, we collected 7,955 articles. As has been stated previously, all of the data was collected in English.

3.4. Annotation Guide & Scheme

The annotation guide, provided to our annotators, describes what text annotation is, the purpose of the corpus, the idea of a motif, the annotation procedure, the annotation tool, and provides a catalog of the selected motifs. Additionally, the guide contains a subsection 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 we felt needed to be noted for future work. Discussions that were not a simple resolution of a disagreement or correction of a mistake, but resulted in a decision about how specific cases should be handled, were noted in this subsection. The annotation scheme was developed over the course of many pilot annotations done by the first four authors in small batches of around 100 samples each and also helped to refine the annotation guide and scheme (for example, the addition of the EPONYMIC tag).

The guide defines the following terms. When we use the term **motif**, we are referring to the general idea of motifs, without referring to a specific motif. We use the term **motif type** to refer to a specific motif, e.g., *magic carpet* or a *Old Man Coyote*. When we use the term **motif candidate**, we are referring to a span of text that matches one of the possible lexical forms of a motif. Since motif candidates were found by keyword search, the main task of the annotators was to determine if the motif candidate was actually being used to express the cultural ideas associated with the motif. The guide further defines the following mutually exclusive classes that are applied to each motif candidate to capture this:

MOTIFIC Invokes the cultural associations of a motif (e.g. referring to something large and monstrous as a "behemoth").

EPONYMIC 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).

REFERENTIAL Directly refers to the folklore origin of the motif or its definition (e.g., discussing the origin of the "behemoth" motif itself).

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

When we use the term **motific instance**, we are referring to a motif candidate that has been marked as MOTIFIC and thus is used to express the associations found in the source narrative of the motif.

3.5. Annotation Tooling

Selecting an annotation tool was a relatively simple matter: while we explored several tools, including an annotation tool developed for a similar task, Story Workbench (Finlayson, 2011), we eventually settled on brat (Stenetorp et al., 2012) as the simplest and easiest to deploy tool for the annotation.

[2] https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page

Since displaying full articles to the annotators was inefficient (some articles may contain only a single motif candidate, or have motif candidates spread far apart), we decided to instead show snippets of articles for context and display multiple subsections with motif candidates per annotation file. To enable this, it was necessary to develop scripts to extract portions of the texts contains motif candidates and combine them into files for presentation to the annotators in the brat UI. Figure 1 shows an example of annotated text as seen by the annotators.

3.6. Selection & Training of Annotators

We hired annotators who identified as members of the groups in question, with a strong background in the culture, which was determined through an interview. We also required annotators to possess a college degree and be fluent in English. We hired six annotators total (two annotators per group) to perform the double-blind annotations. They were paid $20.80 per hour. The Irish and Puerto Rican annotators were born in those places, and of the two Jewish annotators, one was born in Israel and the other takes yearly trips there. All of our annotators either currently live in the US or had lived here for some time, as they required work permits to be paid.

We gave annotators an initial 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, we 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. Further, the period of time before annotators reached "reasonable agreement" are considered part of the training regiment.

3.7. 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 batch was completed. We limited annotators to a total of 10 hours of work a week, as annotation can be a tedious task and we wanted to avoid annotation fatigue, which would reduce data quality. Annotation batches started at 300 motif candidates for the first week and was increased as annotators became more accustomed to the task, rising to over 1,000 candidates in the final weeks. The exact numbers varied depending on the articles selected for the week, as articles were not split between batches. Batch sizes and per-batch agreement measures are listed in Table 4 in the Appendix.

We monitored inter-annotator agreement continuously through the process. 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. The annotation and adjudication itself took a total of 11 weeks and cost roughly $15,000.

3.8. Adjudication

Adjudication was a relatively simple process: the first author (the adjudicator) met with each pair of annotators to discuss the previous batch of annotations. The meetings focused solely on disagreements and annotators were allowed to come to a decision on the correct annotation except for times when the adjudicator was asked for 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. These special cases were used to capture phenomenon where annotators experienced significant disagreement.

Examples of special cases are the use of motifs in new fiction (e.g., the re-purposing of a motif is considered motific because it aims to invoke and subvert associations) and the inclusion of a motific term as part of a descriptor (e.g., "kiddush wine" is not motific, but referential as the usage of "kiddush" is strictly to specify what the wine is for).

4. Corpus Description

This dataset comprises 26,078 motifs candidates across 7,955 texts. The data we release[3] takes the form of a CSV file featuring the title, source, author, and publication date of the source article, as well as the motif, its appropriate label as one of MOTIFIC, REFERENTIAL, EPONYMIC, or UNRELATED, window of up to 50 tokens on either side of the candidate. The candidate itself is enclosed in `<motif>` tags. Table 1 lists the number of texts, tokens in the full texts, and tokens in the released windows for each group.

[3]The corpus and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/FYOWLQ.

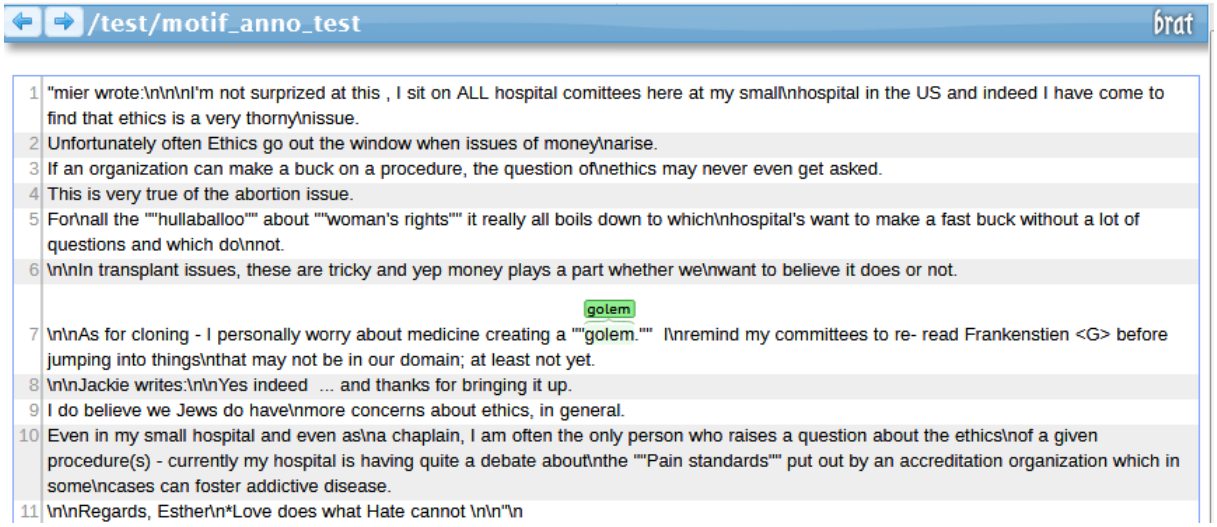

Figure 1: A sample annotation from within brat. Annotators saw similar samples, although they saw multiple snippets per annotation file.

Group	# Texts	# Text Tokens	# Window Tokens
Jewish	2,422	2,388,310	799,397
Irish	2,178	1,817,278	975,423
PR	3,355	2,752,380	860,341
Total	7,955	6,957,968	2,635,161

Table 1: Number of texts, total tokens across the texts, and tokens in the window in the released data, by group. The sum of the rows exceeds the totals because some texts contain motifs from more than one group.

	Motif	Ref.	Eponym	Unrel.	Total
Irish	159	3,197	4,341	1,923	9,620
Jewish	1,215	3,422	2,977	244	7,858
PR	349	4,328	2,214	1,709	8,600
Total	1,723	10,947	9,532	3,876	26,078

Table 2: Motif candidate types per group

Table 2 shows a breakdown of the candidate classes for each of the three groups. The Irish and Puerto Rican groups annotated for slightly longer than the Jewish group, and so they produced more data. These breakdowns show that motific usages are, as expected, relatively rare. Interestingly, motific instances of Jewish motifs are far more common. Class distributions for each individual motif type are shown in Table 6 in the Appendix.

4.1. Inter-Annotator Agreement

We used Fleiss' kappa (Fleiss, 1971) to calculate inter-annotator agreement. The Jewish and Puerto Rican teams participated produced annotations with an average agreement of $\kappa > 0.7$ while the Irish team produced annotations with an average agreement of $\kappa > 0.55$. Detailed annotator agreements per batch as annotation progressed is shown in Table 4 in the Appendix.

4.2. Discussion

The study demonstrated, first and foremost, that humans can reliably identify motific usage of motif candidates in text and distinguish these from other types of usage. The Jewish annotators had consistently high agreement throughout. Many of the Jewish motifs had very specific and distinct meanings that are not in use outside of the group (e.g., *Amalek* or *Haman*) which we believe is responsible for this high 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 resulted in annotator fatigue. We believe that both of the reasons likely contributed: many Irish motifs, as they become more common, have a diluted meaning—for example, the *leprechaun* is perceived differently outside vs. inside Ireland, where it is viewed as mischievous or naughty; 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. The Puerto Rican team also expe-

rienced a dip in agreement in the last two batches, which we infer is for similar reasons: the Puerto Rican annotators suggested some new motifs, which have less stable or well-agreed-upon meanings.

5. Experimental Results

We performed a basic evaluation of four large language models (LLMs)—T5 (Raffel et al., 2020), GPT-2 (Radford et al., 2019), FLAN-T5 (Chung et al., 2022), and Llama 2 (Touvron et al., 2023)—at the task of classifying a motif candidate into our scheme (i.e., MOTIFIC, EPONYMIC, REFERENTIAL, UNRELATED). We used the standard version of GPT2 (124M parameters), the base versions of T5 and FLAN-T5 (220M), and Llama 2 (7B). We performed the evaluation using 0-, 3-, and 5-shot prompting, providing hand-crafted examples to the models for the 3- and 5-shot experiments.

Table 3 shows the performance of each model, with the highest performance occurring for Llama 2 in the 0-shot trial. This seems to be due to Llama 2 almost always outputting REFERENTIAL, which happens to be the majority class. T5 almost exclusively produced output that did not fit one of the labels. All the models (except T5) had a tendency to choose one of the four categories and output solely that. We believe this is influenced, in part, by the order in which they are listed as part of the prompt, the order of examples, and the tags placed within the examples.

	0-shot	3-shot	5-shot
	Accuracy		
T5	0.03	0.12	0.07
GPT-2	0.06	0.38	0.38
FLAN-T5	0.15	0.15	0.15
Llama 2	0.41	0.35	0.37
	Macro F_1		
T5	0.03	0.08	0.07
GPT-2	0.11	0.24	0.25
FLAN-T5	0.07	0.07	0.08
Llama 2	0.15	0.20	0.17

Table 3: Results of evaluating four LLMs at classifying a motif candidate using few-shot prompting.

Overall, we believe this experiment demonstrates that motif classification remains a challenging task that cannot be solved by naïve approaches using off-the-shelf LLMs. As such, we believe GOLEM provides a valuable resource to those who wish to produce stronger LLMs that demonstrate a more nuanced understanding of the cultural information contained within text.

6. Future Work

An obvious next step is to examine fine-tuning LLMs to better suit the task of classifying the identified motif instances. Additionally, the task of classifying the motifs is just one of many, including identifying the location of motifs within a body of text and generalizing the concept of a motif beyond the specific motif types given in the training data: both of these tasks are enabled by GOLEM.

There are further applications of GOLEM beyond the domain of motifs. The weak performance of LLMs already suggests one potential application of GOLEM: in strengthening the performance of general language tools, both in understanding the underlying meaning of motific language and in effectively delivering messages for specific audiences by using motifs. Further, GOLEM could be used in information extraction tasks to better access the underlying cultural information inherent in them.

Another potential avenue for expanding this work is collaboration with universities and teams that speak a given language of interest for a set of motifs. We acknowledge that many of the motifs are likely to be more effective and more recognizable in the language they originate; as an English-speaking team, this work was done in English. Similarly, we are expanding motific annotation to Arabic texts (albeit only in English, although we hope to transfer this annotation to Arabic itself).

7. Contributions

The ubiquity and information density of motifs makes them important to consider for anyone working with culturally influenced texts. Here we have provided GOLEM, the first dataset of annotated motifs. The annotation process demonstrated that human annotators can reliably annotate the type of usage (MOTIFIC, EPONYMIC, REFERENTIAL, or UNRELATED) of motifs within a text. We have also demonstrated the difficulty of this task by showing that four off-the-shelf, modern LLMs struggle with classifying motif candidates, suggesting that the task is a challenging one.

GOLEM will enable the development of more robust, culturally aware information extraction, knowledge, and language understanding models. Additionally, we hope that this paper will demonstrate the wealth of knowledge that can be missed through mass collection efforts that are not aware of the necessity for a more nuanced approach to culturally rich knowledge.

8. Acknowledgements

We would like to acknowledge our annotators: James Conlon, Orpaz Levy, Natasha Maldonado, Sivan Manoah, Robert McKendry, and Jean Mendez, for their work on annotating this data. This work was supported in part by DARPA via SBIR Phase II Prime contract FA8650-19-C-6017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA. One of the authors of this research, Dr. Acharya, is now at Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830, but this work was completed while he was at FIU.

Ethical Considerations

Our work avoids the pitfall of crowd-sourcing data: namely, that typical crowd-sourcing can be exploitative of poorer populations. However, we intentionally operated outside of large crowd-sourcing platforms, choosing instead to work through a hiring agency to pay a competitive rate for the task of data annotation, in order to ensure that our annotators were appropriately compensated, and to achieve a high quality of data.

Additionally, our work, in a modest way, supports approaches to computing sensitive to diverse backgrounds: we provide a dataset of diverse cultural knowledge annotated by experts from the groups from which the data is sourced. We hope that this effort will not only inspire others to create similar resources, but assist in making current systems more robust and culturally aware.

9. References

Antti Amatus Aarne. 1910. *Verzeichnis der Märchentypen*. Suomalainen tiedeakatemia.

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2021. An atlas of cultural commonsense for machine reasoning. In *AAAI Conference on Artificial Intelligence*.

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

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

Richard Francis Burton. 2009. *The Arabian nights*. Barnes & Noble.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Tom Peete Cross. 1952. *Motif-index of early Irish literature*. Indiana University.

Sándor Darányi. 2010. 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.

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

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

Thierry Declerck and Piroska Lendvai. 2011. 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.

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

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

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

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

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

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

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

Heda Jason. 2007. 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.

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

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

Dov Neuman Noy. 1954. *Motif-index of Talmudic-Midrashic literature*. Indiana University.

Nir Ofek, Sándor Darányi, and Lior Rokach. 2013. Linking Motif Sequences with Tale Types by Machine Learning. In *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 166–182, Hamburg, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

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

Stith Thompson. 1960. *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.

Stith Thompson. 1977. *The folktale*. Univ of California Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

W Victor H Yarlott and Mark A Finlayson. 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.

W Victor H Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark A Finlayson. 2022. Finding trolls under bridges: Preliminary work on a motif detector. *arXiv preprint arXiv:2204.06085*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. 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.

A. Further Corpus Description

Table 4 provides a full listing of the per-batch results. The Irish group had the most difficulties in achieving the baseline of $F_k > 0.55$, the cutoff selected for demonstrating the viability of human annotation of motifs. 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.

	Irish		Jewish		Puerto Rican	
Week	κ	#	κ	#	κ	#
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	-	-	-	-
Final	0.562	-	0.742	-	0.729	-

Table 4: The week-by-week agreement in Fleiss' kappa of the annotation process. The final average is a macro average calculated from batches starting after an annotation team reached an initial $\kappa > 0.55$ (week 5 for Irish, week 2 for Jewish, and Week 3 for Puerto Rican). The # column indicates the number of candidates included in that week's batch.

Table 6 shows a further breakdown per each motif found within the data. There are, of course, some motifs that are exceedingly rare in the data: while there was an attempt to control for this in the data selection process, the nature of motifs mean that certain motifs are more likely to be expressed. Further, many motifs have found their way into popular culture (e.g., *leprechaun*) or are also the exact name of a real world entity that is commonly mentions (e.g., *coqui* or *shamrock*).

B. 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, we expect that in editorial articles, as compared to non-editorial articles, motifs would occur more frequently.

While we release 7,899 articles as part of this study, the genre code was run over a slightly larger set of 7,946 articles, some of which were removed from this release due to difficulties in retrieving the source artice. Of the 7,946 articles that were used in this genre test, 5,109 had either editorial tags or other genre tags; the remaining 2,678 articles did not. Using a sentence-level opinion classifier (Yu and Hatzivassiloglou, 2003) modified 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 115 editorial and 7,831 non-editorial pieces.

Calculating the rate of motifs per article, sentence, and token, we found that motifs were roughly three times as frequent (3.75x, 3.17x, and 3.04x, respectively) in editorial articles than in non-editorial articles. The detailed results of this experiment are present in Table 5.

We 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. We believe that these results strongly suggest the important of motifs for understanding human communication.

	Op-Ed	Non-Op-Ed	Ratio
Motif/Article	0.75652	0.20181	3.75
Motif/Sent.	0.01840	0.00580	3.17
Motif/Token	0.00076	0.00025	3.04

Table 5: Comparison of motif frequency per article, sentence, and token between editorial and non-editorial articles.

Motif Name	Motific	Referential	Eponym	Unrelated	Total
chupacabra	66	882	299	0	1247
arc of the covenant	5	21	0	1	27
agueybana	1	6	27	0	34
kiddush	207	1543	574	15	2339
leprechaun	53	1098	898	13	2062
wren	0	6	0	1	7
golem	112	477	1321	0	1910
gods name in vain	0	3	0	0	3
aos si	1	0	0	34	35
tower of babel	495	713	267	1	1476
shamrock	65	590	2047	1805	4507
cu chulainn	10	185	9	1	205
atabey	5	4	109	341	459
guanina	1	1	6	38	46
tir na nog	3	69	118	0	190
children of lir	1	65	2	3	71
haman	11	258	103	53	425
reyes magos	68	52	144	12	276
banshee	2	99	11	31	143
roberto cofresi	9	42	38	5	94
divina providencia	1	0	2	0	3
behemoth	269	7	727	0	1003
coqui	169	2443	1405	1243	5260
salmon of wisdom	1	23	0	0	24
seventy languages	2	13	0	131	146
hormigueros	1	439	7	19	466
amalek	127	408	2	43	580
fairy fort	0	2	2	0	4
jibaro	11	1	65	9	86
finn mccool	30	1061	1269	20	2380
yocahu	7	6	24	24	61
king conchobar	0	27	0	0	27
juan bobo	10	35	78	0	123
magic harp	0	2	0	16	18
Totals					

Table 6: Candidate classes per motif found in the data.