"... like a needle in a haystack": Annotation and Classification of Comparative Statements

Pritha Majumdar[1], **Franziska Pannach**[1], **Arianna Graciotti**[2], **Johan Bos**[1]
[1]University of Groningen, [2]University of Bologna
Correspondence: p.majumdar@rug.nl

Abstract

We present a clear distinction between the phenomena of *comparisons* and *similes* along with a fine-grained annotation guideline that facilitates the structural annotation and assessment of the two classes, with three major contributions: 1) a publicly available annotated data set of 100 comparative statements; 2) theoretically grounded annotation guidelines for human annotators; and 3) results of machine learning experiments to establish how the–often subtle–distinction between the two phenomena can be automated. For the purpose of automatic classification, we present a baseline system (SVM), as well as experiments with large language models. We achieve 82% accuracy on the best performing model-Llama 3.3-70b-instruct following a few shot prompting strategy.

1 Introduction

The automatic processing of figurative language is a challenge that has long been a focal point of research in natural language processing (Ge et al., 2023; Rai and Chakraverty, 2020; Joshi et al., 2017; Amin and Burghardt, 2020). One of the tasks in dealing with comparative statements is to find clear boundaries between two very similar phenomena: that of comparison and simile.

The term *comparison* describes a linguistic unit that is used to convey similarities and dissimilarities between two entities. Even though comparisons in general are understood to be syntactic, they can in effect harmonize much more relevant semantic knowledge in everyday language. Comparisons are used in everyday communication, e.g., in a debate when one has to put across a point, or when pointing out a similarity or difference between entities that share some property. *Similes* on the other hand are special structures that are derived from comparisons and can also be called *figurative comparisons*. A simile can be defined as a figure of speech that is used to draw a parallel between two dissimilar entities or processes that have some shared properties. Let's consider the two examples:

(1) He is as tall as his brother.

(2) He is as tall as the tower.

In (1), the comparison is drawn between two brothers' physical size, while (2) draws a comparison between a human (he) and an object (tower) which makes the comparison figurative. This leads to an important consideration in the distinction between comparisons and similes: Comparisons can be drawn between any two entities, while a comparison only becomes figurative (simile), if the two entities belong from different semantic categories.

This paper presents an annotation methodology that allows us to distinguish if a comparative statement is a *literal comparison* or a *figurative comparison* (simile). To the best of our knowledge this is the first work that focuses on step-by-step annotation guidelines for comparison vs. simile, taking into account various features of comparative statements.

Similes are a particularly interesting phenomenon in the domain of literature, because they often carry subtle meaning that can be overlooked if a statement is treated as a simple comparison: "Poor Dorothea felt that every word of her uncle's was about as pleasant as a grain of sand in the eye to Mr. Casaubon." (Eliot, 1994, p.335). The former statement contains a simile. It carries more meaning than a simple comparison, i.e., it requires the reader to acknowledge that a grain of salt in the eye is very unpleasant and thus the statements that Dorothea's uncle makes in front of Mr. Casaubon are evoking a negative emotion in the listener.

In fiction, simile's are often used to transport subtle meaning and therefore particularly interesting to study. However, similes are a sparse phenomenon with rare occurrence in everyday usage, their annotation is time-consuming and often yields very little data per literary work. Therefore, this

Proceedings of LaTeCH-CLfL 2025, pages 261–271

work presents straightforward annotation guidelines by defining *nine* categories (subject of comparison, object of comparison, nature, categories, feature matching, symmetry, salience, broad unifying concept and domain incongruence) to distinguish similes from non-figurative comparisons, together with machine learning experiments that can help automatically annotating larger corpora of fictional texts for further studies in the domain of computational literary studies. The data set and gold standard annotation can be found here[1].

2 Related Work

The study of comparison in rhetoric can be dated as back as Aristotle (Freese et al., 1926), who highlighted the importance of using comparisons in everyday life (Seh, 2016). However, computational studies addressing the distinction between figurative and literal comparisons are scarce, since both phenomena follow a similar structure, and consist of the same constituents (Niculae and Danescu-Niculescu-Mizil, 2014). Niculae (2013) proposed a "similarity based approach" that aided in measuring the degree of figurativeness of a comparison which by extension can be used as a means of identification of similes. Since there is a lack of annotated corpora of comparisons, Niculae (2013) used the VUAMC corpora (Steen et al., 2010) to extract the comparison patterns of "like". They then annotated it for the task of simile identification. Since similes are a form of comparisons, Niculae and Yaneva (2013) contributed to computational research on simile by focusing on comparison recognition through the use of syntactic patterns. Most work on automatic detection and analysis of figurative language targets metaphors (Li et al., 2023; Boisson et al., 2024) and idioms (De Luca Fornaciari et al., 2024; Chakrabarty et al., 2022), but only few recent studies investigate similes as special comparative statements.[2] More recent exceptions are Liu et al. (2018); Wang et al. (2022), in which neural and transformer-based models are used in a multi-task setting to identify similes and their components in Chinese texts.

3 Annotation Guidelines

The guidelines that are presented here is developed based on Seh (2016), in which Seh dedicates a complete chapter in understanding comparisons and their distinction from similes. Before we introduce the annotation guidelines, we must first discuss the syntactic structure and the semantic particularities of different types of comparative statements. We will then discuss the steps involved in distinguishing a literal comparison from a simile. This paper does not focus on the merits of the individual underlying theories[3] of comparison. Instead, its main contribution is building a concise annotation guidelines that is derived from the theories for the task of identifying and distinguishing comparisons and similes.

3.1 Comparisons and Similes

Typically, a comparative structure consists of two elements that are the focus of a comparison, i.e., the (two or more) elements being compared, e.g., *he, his brother*, and the property, e.g., *tall*, with regard to which they are compared (Dixon, 2018). The other components of a comparative structure are:

(1) The item that is compared or *subject of comparison*;

(2) The standard of comparison against which the item is compared or *object of comparison*;

(3) The quantity or quality, i.e. the property used for the comparison or *parameter*;

(4) The standard marker which states the relationship between the subject and object of comparison or *mark*;

(5) The degree marker which states the extent of the comparison or *index*.

Table 1 illustrates an example of these components.[4] While a comparison is the phenomena of formally likening one thing to another that resemble each other in some properties, a simile is a figure of speech which generally relies on a linguistic marker to draw a parallel between two or more semantically distant entities or processes based on stated or implied (dis)similarities, so as to produce a particular image in a person's mind (Seh, 2016).

[1] https://github.com/prithamajumdar/Annotation-Classification-of-Comparative-Statements

[2] A comprehensive survey of computational approaches towards similes is accepted and currently undertaken by the authors in https://direct.mit.edu/coli.

[3] The theories are discussed in Seh (2016).

[4] Some sources such as Dixon (2018) and Seh (2016) refer the subject of comparison as *comparee* and object of comparison as *standard*

Table 1: Illustration of terminologies in comparisons

Sentence	Subject of Comparison	Index	Parameter	Mark	Object of Comparison
Max is more intelligent than George	Max	more	intelligent	than	George

For example[5],

(1) This book is *more* interesting than that one.

(2) The Earth is round *like* an orange.

Both the examples imply a *comparative degree of adjective*. For comparisons, the structures indicate *equality, superiority or inferiority* which means that all these comparisons are *scalable* (De Mille, 2024). In similes, we consider the similarity concept as a *spectrum*, wherein it can range from "some" similarity to "more than/less than" similarities (Cohen, 1968). Thus, comparisons are usually *quantitative*, while similes are *qualitative* by nature.

Furthermore, for a comparison to be a simile, the two elements of the simile should "differ in kind" (Bain, 1890), or to be "of different kind" (Waddy, 1889) or to be "drawn from one species of things to another" (Jamieson, 1823).

Categories This leads us to the next consideration in distinguishing a literal comparison from similes: the (semantic) *categories*. A category may be defined as "a number of objects which are considered equivalent" (Rosch, 1978). Aristotle defined ten categories into which each single concept may fit: Substance, Quantity, Quality, Relation, Where, When, Position, Possession, Action and Passion (Aristotle et al., 1889). The task of this categorisation is however not done haphazardly, but is "based on specific perceptible or known attributes and most times, it is either intuitive, used in a specialised context or rooted in a culture." (Seh, 2016). Rosch (1978) list three levels of natural categories:

(a) *basic-level category* that consists of basic objects like car,

(b) *super-ordinate category* to which the basic objects belong, like vehicle for car,

(c) *sub-ordinate category* are the types of basic objects, like SUV for car,

Therefore, comparisons generally concern entities that are at the same level of categorization and belong to the same super-ordinate category, while similes involve entities that are on different levels of categorization. For example,

(3) *Spoons are like forks.*

(4) *The girl is like a lily.*

In (3), spoons and forks are basic objects that have several subordinate categories (dessert spoon, teaspoon, soup spoon, fish fork, salad fork etc.) and belong to the same super-ordinate category, cutlery. Thus (3) is a *comparison*. On the other hand, girl and lily in (4) do not belong to a common category and a very high level of abstraction is required to find a shared super-ordinate category. Thus, (4) is a *simile* (Seh, 2016).

Feature Matching The next important step is to perform *feature matching*. On this level of annotation, similes can be identified by measuring similarity of two elements, taking into account their similarities and differences. For example,

(5) *The chair is like an armchair.*

(6) *This chair is like a boulder.*

Example (5) is a *comparison*, since they share many similar features (both are used for sitting and belong to the same super-ordinate category-furniture). The similarities are more prominent than the differences. However in (6), the similarity between a chair and boulder is much lower, the differences are more prominent than similarities. Therefore, it shows features of a *simile* (Seh, 2016).

Symmetry The next concept that we use to differentiate comparison and simile is *symmetry*. Comparisons are symmetrical in nature, which means that you can alter the order of subject of comparison and object of comparison. However, similes are asymmetrical in nature which means that changing the position of the subject and object can affect the meaning. For example,

(7) *Spoons are like forks* has the same meaning as *forks are like spoons* making the statement a *comparison*.

(8) *A girl is like a lily* is not the same as *A lily is like a girl*, making the statement a *simile*.

In (8), a descriptive quality of *a girl* is conveyed, but less so a quality of *a lily*.

Salience The next distinction between a comparison and simile is *salience*. In similes, the shared features of subject and object should show low

[5]All the examples discussed in this section are taken from the thesis of Seh (2016).

salience in the subject of comparison, and high salience in the object of comparison.

(9) *Spoons are like forks*, both concepts show high salience, i.e., both are utensils and both are held by hand and are used for eating). Thus, this is a *comparison*.

(10) *The girl is like a butterfly*, the concepts have different levels of salience, e.g. the *butterfly* signifies fluidity, flittiness, lightness and transience, features that are more readily associated with butterflies than with girls. Thus, this makes it a *simile*.

Meaningfulness For a statement to be considered a simile, it should also be meaningful. That is, the items compared–while potentially from different domains–should still be relatable under a broader, unifying concept or category, e.g.,

(11) *Billboards are like spoons.*

(12) *Sally is like a block of ice.*

From the above example, (11) lacks a meaningful semantic connection because billboards and spoons cannot be easily grouped under a shared domain or concept–at least not without a further explanation. This makes the statement a *comparison*. While in (12), even though *Sally* and *a block of ice* come from different domains, they can still be compared through an abstract quality, e.g. *stiffness/metaphorical or actual coldness*. This broader concept allows for a reasonable connection between the two and makes the comparison a *simile*.

Domain Incongruence The last phenomenon to consider is *domain incongruence*. In our case, this means that the elements of comparison must belong to distinct categories or semantic domains, e.g. person and object). A statement can only qualify as a simile when the attributes shared by the subject of comparison and the object of comparison are not strictly identical.

(13) *Max is like the Empire State Building* is a *simile*.

(14) *Max is as tall as George* is a *comparison* because both are human.

3.2 Annotation Methodology

In this section we present the annotation methodology that allows us to decide if a statement is a *comparison, simile* or if the distinction is *Not Applicable* (see Table 2).

3.2.1 Identification

The first step in the annotation process is to identify the *subject of comparison* and *the object of comparison*. For example,

Max is as tall as George.

Tom is as fast as a leopard.

The subject of comparisons are *Max* and *Tom*, while the objects they are being compared to are *George* and *a leopard*.

Contextual Span We need to consider how much context should be included as the *subject* and *object of comparison*. In our annotation, we include the *noun*, the whole *noun phrase* or even the complete *clause* in situations where it is applicable. For example,

(1) In *Tom is as fast as a leopard*, we annotate *Tom* as the subject of the comparison, and *a leopard* as the object of the comparison.

(2) In *Few treasures are worth as much as a friend who is wise and helpful*, *Few Treasures* is the subject of the comparison, and the whole clause *a friend who is wise and helpful* is the object of comparison.

(3) *Better is the poor who walks in his integrity, than he who is perverse in his ways, and he is rich.* Here, we annotate *The poor who walks in his integrity* as the subject and *The rich who is perverse in his ways* as the object of the comparison.

Contraction In cases of contraction, we reduce the form to the root word. For example,

(4) In *I'm as hungry as a bear*. The subject of comparison is *I* instead of I'm.

Co-reference In cases of co-reference, we identify the subject/object of comparison as the noun/noun phrases. For example,

(5) *Tom is a solid and determined man, but sometimes he's as impetuous as a river of molten lava*. We resolve *Tom* as the subject of comparison instead of *he's*.

Multiple Components In cases of statements with multiple components in the subject or object of comparison, we mark all of them. For example,

(6) *Her mouth is smoother than oil, but in the end she is as bitter as wormwood, and as sharp as a two-edged sword.* Here, *her mouth* is the subject of comparison and *oil, wormwood, two-edged sword* are the objects of comparison.

Dialogues In case of dialogues, we reconstruct the subject/object of comparison to the most meaningful form. For example,

(7) *So February's policy note is a stunning reversal – as close as an institution can come to*

Table 2: Snippet of the method of annotation

Sentence	Subject of comparison	Object of comparison	Nature	Categories	Feature matching	Symmetry	Salience	Broad, unifying concept	Domain Incongruence	Result
Tom is as fast as a leopard	Tom	a leopard	Qualitative	Different basic level category (human, animal)	More prominent differences	Asymmetrical	High salience in object of comparison	Meaningful	Distinct	Simile
An elephant isn't as big as a whale	An elephant	a whale	Quantitative	Same superordinate category (animal)	More prominent similarities	Asymmetrical	Same salience	Meaningful	Similar	Comparison
I'll send it out as soon as the machine is available	It	-	-	-	-	-	-	-	-	Not Applicable

recanting without saying, "Sorry, we messed up". Here, we annotate the subject of comparison as *February's policy note*, and assign *an institution can come to recanting without apologizing* instead of *an institution can come to recanting without saying, "Sorry, we messed up."* as the object of comparison.

Exceptional cases In statements such as:

(8) *He paid as much as a million dollars for the painting*. There is no object of comparison. This statement merely is a form of emphasis and the marker *as much as* in this context does not compare two entities. In such cases, we mark the subject of comparison (if it is clear, i.e. *he*) and the object of comparison as *Not Applicable*.

However, this does not mean that all statements that contain the phrase *as much as* should be discarded. For example,

(9) In *She enjoys reading as much as watching movies*, we have a subject of comparison *reading* and an object of comparison *watching movies* which highlights and quantifies what *she* likes doing better by the phrase *as much as*.

3.2.2 Annotating the Characteristics

The second step of the annotation process is to annotate the characteristics derived from the subject and object of comparison. In this step we consider the factors introduced in Section 3.1 and establish them as the seven categories to make the final judgement of whether a statement is a simile or not. These seven categories are: *nature, categories, feature matching, symmetry, salience, broad unifying concept and domain incongruence*. See Table 3.

From the aforementioned considerations, the most important characteristics that enable us to decide if a statement is a comparison or a simile are:

(1) *Categories*: If the subject and object of comparison belong to the same super-ordinate category, more often than not, the statement is a *comparison*.

The category of domain incongruence is directly dependent on the characteristic of the category.

(2) *Feature matching*: A statement can be a simile if there are more prominent differences than similarities.

(3) *Broad unifying concept*: Since comparisons can practically be drawn from any two concepts, we need to establish if the comparison makes sense for it to be a simile.

However, as mentioned above, we still need to annotate the other characteristics since in some cases, we need to go beyond main three characteristics to assess the comparative statement.

Based on these relevant characteristics, we decide if a statement is a comparison or a simile.

For example,

(1) *Better is the poor who walks in his integrity, than he who is perverse in his ways, and he is rich* (see Table 4). The subject of comparison is *The poor who walks in his integrity* and the object of comparison is *The rich who is perverse in his ways*. This statement is a *comparison*. In this example, we can see there is an equal number of characteristics for the statement to be a comparison or simile. In such cases, we will concentrate more on the characteristics *category, feature matching* and *broad unifying concept* to determine if we annotate it as a comparison or a simile. Through this example we can see that not all comparisons have to be symmetrical or quantitative in nature.

(2) *The root of a flower is as weak as a baby's finger.* In this case, the subject of comparison is *The root of a flower* and the object of comparison is *a baby's finger*. This statement is a *simile*. In this example, we can see that the subject of comparison and the object of comparison are

Table 3: Characteristics of Comparisons and Similes

Characteristic	**Comparison**	**Simile**
Nature	Quantitative	Qualitative
Categories	Can belong to the same superordinate category	Should belong from different basic objects
Feature matching	More prominent similarities than differences between entities	High prominent differences than similarities between entities
Symmetry	Symmetrical	Asymmetrical
Salience	High salient in subject of comparison than object of comparison	High salient in object of comparison than subject of comparison
Broad concept	Can be any comparison (even nonsensical)	Should be a meaningful comparison
Domain incongruence	Similar semantic domains	Distinct semantic domains

Table 4: Example 1: Better is the poor who walks in his integrity, than he who is perverse in his ways, and he is rich

Attribute	**Value**
Nature	Qualitative
Category	Same superordinate category (human nature)
Feature matching	More prominent differences than similarities
Symmetry	Asymmetrical
Salience	Both have the same salience
Broad concept	Meaningful
Domain incongruence	Similar

symmetrical, i.e. they can be used interchangeably. This is a typical characteristic of comparisons. We also have the same salience for this statement. For example, the root of a flower is small and fragile, which are also both typical characteristics of a baby's finger. In such cases (as discussed above), we prioritize the characteristics *category, feature matching* and *broad, unifying concept* to aid us in deciding. According to those three characteristics, the statement qualifies as being a simile.

(3) *So February's policy note is a stunning reversal – as close as an institution can come to recanting without saying, "Sorry, we messed up." But it parallels a general shift in economists' opinion* (see Table 5). The subject of comparison is *February's policy note* and the *object of comparison* is *an institution can come to recanting without apologizing*. As we mentioned above, the most important characteristics when deciding between comparison or simile are *categories, feature matching, broad unifying concept*. If we have a different basic level category, more prominent differences and a meaningful concept, we annotate the statement as a *simile*. However, here we have another step that we need to consider before deciding if the statement is a simile or not, i.e. *the nature*. If the nature is quantitative, chances are high that there is no shared property, and the comparison between the subject and object of comparison are drawn just to quantify the relevance of the comparators. In such cases, we would identify the statement as a *comparison*. This is especially easy if we have an "as ... as" construction.

Table 5: Example 3: So February's policy note is a stunning reversal – as close as an institution can come to recanting without saying, "Sorry, we messed up." But it parallels a general shift in economists' opinion

Attribute	**Description**
Nature	Quantitative
Category	Different basic level category (politics, human nature)
Feature matching	More prominent differences
Symmetry	Symmetrical
Salience	Same salience
Broad concept	Meaningful
Domain incongruence	Distinct

(4) *It's as lovely as a rose.*

In such cases, we cannot annotate the subject of comparison in a meaningful way, since "it" and could signify anything. In such cases we will leave the annotation of the characteristics blank and classify the statement as *Not Applicable*.

However, in statements where we have a context following the undefined subject of comparison, we might be able to resolve it. For example,

(5) *"What are the twelve signs of the Zodiac, in the order in which the sun passes them by in the course of a year?" - "Um, let me think for a minute!" - "No thinking! It's got to come as quick as a shot!"* In this case, we can reconstruct the unspecified subject of comparison *it* to *the answer*.

(6) *He is a figment as much as a figure.*

This example is an idiomatic expression. Even though they have the structure of a comparison, subject and object of comparison cannot be derived in a meaningful way. We annotate such cases as *Not Applicable*.

4 Data

The data for the annotation study was extracted from the English data present in the Parallel Meaning Bank (PMB) (Abzianidze et al., 2017) and filtered by the simple regular expression: `as [a-z]* as an?`. We then manually clean the data to remove duplicate instances, shorten the sentences to simplify annotation and split complex sentences with multiple comparative structures into shorter sentences. For example, *"I am as light as a feather, I am as happy as an angel, I am as merry as a school-boy"* was split into three simple comparative sentences. Furthermore, all instances of "as well as" were removed as those are usually synonymous to statements containing *too* or *also*. We eventually gather a data set of 100 sentences. The statistics of our gold standard annotation can be found in Table 6.

Table 6: Results of Gold Standard Annotation

Class	Count
Simile	63
Comparison	19
Not Applicable	18

4.1 Annotation procedure

Subsequently, we conducted annotations based on the above presented annotation guidelines with two expert annotators[6]. It is to be mentioned here that the first-language of the annotators are Bengali and Italian, and none of them use English as their first language. This led to variation in understanding and interpreting many statements caused by a language barrier. The annotators were presented with 100 sentences and were asked to annotate the nine categories. Table 2 presents a snippet of such an annotation. After independent annotation, the gold standard was derived through resolving cases where Annotator 1 and 2 disagreed in their judgement by discussion between the experts.

4.2 Inter-annotator Agreement

We have analyzed the inter-annotator agreement using Cohen's κ across the following pairs (see Table 7). The annotation by the LLM is the result of prompting (that is discussed in section 5). The highest agreement is achieved between the LLM using different prompts, i.e. 64%. We have noted interesting differences of opinion between our human annotators, see subsection 6.1.

Table 7: Inter-annotator agreement

Comparison	Cohen's κ
Annotator 1 vs Annotator 2	0.62
Annotator 1 vs Zero-shot	0.47
Annotator 2 vs Zero-shot	0.39
Annotator 1 vs Few-shot	0.52
Annotator 2 vs Few-shot	0.55
Zero-shot vs Few-shot	0.64

5 Experiments

In correspondence to the human annotation, we also conduct machine learning experiments to help determine if and how the process of classifying a comparative statement into comparison or simile can be automated. For that purpose, we use a simple support vector machine (SVM) as our baseline (with support vector classification (SVC), a linear kernel, and the default regularization parameter (C=0.1)). The data was split into a training set and test set of 80%-20% and Tf-idf vectorizer was used as the feature extractor.

[6]Author 1 and Author 3

We then conduct two experiments with the Large Language Model (LLM) LLama-3.3-70b-instruct[7].

We perform the first experiment using zero-shot prompting, in which the LLM is asked to judge if a comparative statement is a *simile, comparison* or *Not Applicable*, see Table 9. In the second experiment, we apply a few-shot prompting method to the same model, see Table 10.

We test the performance against the gold standard annotated data.

6 Results

In this section, we report the results on the annotation task (inter-annotator agreement, Cohen's κ), and the machine learning experiments, i.e. the SVM baseline and LLM annotations that were conducted on the curated data set.

6.1 Error Analysis for Human Annotations

In this section, we will examine interesting differences noticed between the judgements of the two annotators. We categorize the differences into the following:

Stock similes: Certain comparisons are perceived as a proverb to one annotator while the other perceives it simply a simile (according to the annotation guideline). In figurative language such proverbial comparisons are called stock similes (Norrick et al., 2010). As Seh (2016) says, "The simile is so ancient a figure of speech that several comparee NP/quantity or quality-standard of comparison combinations have become an integral part of the language, losing in the process their initial figurative flavour". Stock similes thus have such familiar associations through the passing of time that they fail to impress or not even seen as figurative to the common folk (De Mille, 2024). Some of the examples of such disagreements from our data are:

(1) I am as healthy as a horse.

(2) Tom is as fast as a fiddle.

Cultural implications: Different cultural background has affected the decision of annotators in some cases. In such instances we see one annotator labels a comparison as a simile and the other (by perceiving the comparison quite literally) labels the same as Not Applicable. In such cases, difference in interpreting the construction literally vs. figuratively plays a role in the decision of the annotator.

(3) The child is as neat as a pin.

(4) He is as nutty as a fruitcake.

For (3), the annotator cannot associate the shared property *neat* with the object of comparison *pin*. The annotator perceives them as very different concepts and fails to have a meaningful relationship, i.e. a child can can be neat, but neat cannot be associated with a pin. Here, we can see how one annotator has annotated the sentences strictly according to the guidelines, while the other favored a more holistic perspective.

Syntactic Structure: In this category, we see that sometimes the syntactic structure of having "like" or "as" leads to misinterpretation. For example,

(5) Having eluded killers like malaria and AIDS, one should not then be killed prematurely by cancer – especially a form of cancer that could have been prevented with something as simple and as affordable as a vaccine.

In the (5), one annotator annotates it as *Not Applicable*, while the other annotates it as a simile. The annotator choosing simile as a category was also influenced by the widely spread metaphorical use of "illness as a killer", "illness as a war", which is also attested in the cognitive metaphor literature (Sontag, 1978; Lakoff and Johnson, 2008).

Metaphorical Influence: We have some interesting cases of metaphorical influence. For example,

(6) He is as innocent as a child.

(7) Her skin is as firm as a teenager's.

While on the surface level it seems like a comparison (since they belong to the same category), it is not always simple even though the subject and object of comparison are both humans. Here, we are comparing an adult to a child. The annotators disagree in this case, wherein one perceives it as a mere comparison, while the other thinks it's a simile. During discussion, the annotator said that metaphorical expression had an influence on the decision. We plan to look into more of these cases the future. For that purpose, we need to find more fine-grained way of annotating such cases, e.g. by looking at similar forms of expression from different domains like fiction.

[7] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

6.2 Results of the Machine Learning Experiments

In this section, we report the results of our SVM baseline and the LLM- Llama 3.3-70b-instruct on our curated data set. In Table 8 we compare the results of all three experiments that we evaluate on our gold standard data. We see that in the third experiment, i.e. prompting the model with proper examples as illustration (see Section 5), the LLM is able to massively improve the accuracy from 72% to 82%.

Table 8: Baseline vs. LLM performance

Model	Accuracy
SVM	75%
Llama-3.3-70b (zero-shot)	72%
Llama-3.3-70b (few-shot)	82%

While the performance is encouraging, we also see some cases where the LLM takes some unexpected decisions (see Table 11). Even with clear prompts such as Instruction Prompt 2 (*If the subject and object of comparison belong to the same category, you should mark as a Comparison*), the LLM annotated the Example 1 as a *simile*. Interestingly, even though Instruction Prompt 1 says (*if there is unspecified subject or object of comparison you should mark it as Not Applicable*), the LLM judges Example 2 as *simile* and Example 4 as *comparison*.

6.3 Discussion

As pointed out by the Perspectivist Data Manifesto[8], linguistic annotation follows four basic components. A *set of instances* to annotate, followed by a *target phenomena* which is described in detail with guidelines and examples, an *annotation schema* that defines the phenomenon to annotate and finally a *group of annotators* who are deemed fit to carry out the annotation based on their expertise. In this paper, we follow the same procedure to make a distinction between when a comparative structure is called a comparison, and when it becomes it's figurative counterpart, simile. We begin with first defining the phenomena, comparison and simile, followed by the illustrations on what to annotate and a step-by-step process on how to annotate the comparative structures. Our fine-grained annotation guideline allows annotators to take a well-formed decision on whether a comparative statement is a literal comparison or a simile.

As discussed in section 6.1, the annotation of figurative language can be influenced by many factors, with cultural differences playing a significant role in shaping perspectives. This phenomenon of difference in perspective is reflected in the score of our inter-annotator agreement between our human annotators. We use the Cohen κ metric to track how similar the answers of our annotators are to the same set of questions. The final data set contains 63 instances of *Simile*, 19 instances of *Comparison* and 18 instances of *Not Applicable* on our gold standard annotation. Subsequently, our machine learning experiments also yield interesting results. From the performance of our baseline (SVM) and LLM (Llama-3.3-70b-instruct), we can clearly see that our baseline performs better than the zero-shot prompt with the LLM. This raises the interesting question of how well we can trust the judgments of LLMs, especially in subjects that require taking world knowledge into account. Our best performing model is the few-shot prompting with an accuracy of 82% which clearly indicates that by prompting a few examples the performance of the LLM can be boosted for such a classification task, showing the benefit of prompt engineering.

7 Conclusion and Future Work

This work is the first step towards building a pipeline to automatically detect and annotate similes in fiction. It is essential to first draw a clear distinction between a comparative structure as a literal comparison and as a simile, which is what we aimed through this work. The next focus of our project is to develop a fine-grained annotation guideline to annotate similes in literature. We also aim to make the guidelines largely language-agnostic, with a focus on English that will be refined for other languages, such as Bengali, that come from a completely different language family with a different word order. Furthermore, the final objective is to perform a quantitative and contrastive analysis to uncover cultural narratives and values depicted in simile usage in literature and the way of expression of humans in general.

[8] https://pdai.info/

8 Appendices

Table 9: Zero-shot prompt

Zero-shot Prompt:
Does the sentence contain a comparison, a simile, or not applicable? Answer with "Comparison," "Simile," or "Not Applicable" only. Do not write anything else.

Table 10: Prompt for the few-shot experiment

Few-shot prompting:
Here are some examples to guide your response: 1. Tom is as fast as a rabbit – *Simile* 2. He donated as much as 50,000 dollars to the charity – *Not Applicable* 3. An elephant isn't as big as a whale – *Comparison*
Instruction:
1. If there is an unspecified subject or object you should mark it as *Not Applicable* Some examples: a. Nothing is as good as a breath of fresh air b. It's as beautiful as ever
2. If the subject or object of comparison belongs to the same category (human-human, animal-animal, celestial body, social gathering) **you should mark it as *Comparison*** Some examples: a. I am as beautiful as my mother b. She is as strong as her father c. He was as drunk as the guitarist d. The Earth looks as round as the Sun e. Her eyes are as beautiful as a child's f. The surface was as white as the wall
3. If there is an idiomatic expressions you should mark it as *Not Applicable*[9]. Some examples: a. I am feeling under the weather today
4. If there is "like" as an example in the sentence you should mark it as *Not Applicable* Some examples: a. I feel like an ice cream

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik Van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *arXiv preprint arXiv:1702.03964*.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41.

Octavius Freire Aristotle et al. 1889. The organon, or logical treatises, of aristotle: With introduction of porphyry.

Alexander Bain. 1890. *English composition and rhetoric*. Longmans, Green & Company.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2024. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. *arXiv preprint arXiv:2412.15375*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Jean Cohen. 1968. La comparaison poétique: essai de systématique. *Langages*, (12):43–51.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

James De Mille. 2024. *The elements of rhetoric*. BoD–Books on Demand.

Robert MW Dixon. 2018. Comparative constructions in english. In *Language at Large*, pages 472–493. Brill.

George Eliot. 1994. *Middlemarch*. Blackwood. Public domain.

John Henry Freese et al. 1926. *Aristotle, with an English Translation: The" Art" of Rhetoric*, volume 22. W. Heinemann.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Table 11: Some mismatched examples of few shot prompting

No	Sentences	LLM	Gold Standard
1	I am as giddy as a drunken man	Simile	Comparison
2	It's as lovely as a rose	Simile	Not Applicable
3	He is a figment as much as a figure	Comparison	Not Applicable
4	Nothing is as hard as a diamond	Comparison	Not Applicable
5	Tom isn't as naive as a lot of people think he is	Comparison	Not Applicable

Alexander Jamieson. 1823. *A Grammar of Rhetoric and Polite Literature: Comprehending the Principles of Language and Style, the Elements of Taste and Criticism, with Rules for the Study of Composition and Eloquence. Illus by Appropriate Examples, Selected Chiefly from the British Classics, for the Use of Schools Or Private Instruction*. G. & WB Whittaker.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium. Association for Computational Linguistics.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Vlad Niculae and Victoria Yaneva. 2013. Computational considerations of comparisons and similes. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 89–95.

Neal R Norrick, Armin Burkhardts, and Brigitte Nerlich. 2010. *Pear-shaped and pint-sized: Comparative compounds, similes and truth*. na.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Eleanor Rosch. 1978. Principles of categorization. In *Cognition and categorization*, pages 27–48. Routledge.

Suzanne Patience Mpouli Njanga Seh. 2016. *Automatic annotation of similes in literary texts*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.

Susan Sontag. 1978. Illness as metaphor. *Farrar, Straus and Giroux*, 3.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins Publishing Company.

Virginia Waddy. 1889. *Elements of Composition and Rhetoric: With Copious Exercises in Both Criticism and Construction*. American Book Company.

Xiaoyue Wang, Linfeng Song, Xin Liu, Chulun Zhou, Hualin Zeng, and Jinsong Su. 2022. Getting the most out of simile recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3243–3252, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.