Annotating Similes in Literary Texts

Suzanne Mpouli
Université de Lyon, Lyon 2, ERIC EA 3083
suzanne.mpouli@eric.univ-lyon2.fr

Abstract

Annotated corpora are invaluable resources for researchers in the humanities: on the one hand, for natural processing tasks, they can serve as standards against which results from new automatic methods can be measured; on the other hand, in corpus-based studies, they enable either to answer existing research questions or to explore original ones. In this respect, some annotation frameworks such as the Text Encoding Initiative (TEI) attempt to standardise annotation practices in order to facilitate data reuse and exchange. However, despite the crucial role played by figurative language in general and similes in particular in language, no consensus has been reached so far on how to comprehensively annotate them in literary texts. The present paper proposes a framework for annotating similes in literary texts which takes into consideration their semantic and syntactic characteristics as well as the challenges inherent to the automatic detection of similes.

1 Introduction

Be them written or spoken, human-produced sentences are fraught with figures of speech, which possibly explain why, in recent years, various annotated corpora and annotation schemes have been elaborated to describe several of these figures from a linguistic perspective:

- metonymy in Markert and Nissim (2002);
- metaphor in Steen et al. (2010), Shutova and Teufel (2010), and Gordon et al. (2015);
- irony in Gianti et al. (2012), Trevisan et al. (2014), Van Hee et al. (2016), and Karoui et al. (2017);
- figures of repetition in Gawryjolek (2009), and Ruan et al. (2016).

However, despite the important place that figures of speech occupy in literary analysis and the linguistic creativity that particularly characterise fictional writings, literary texts have rarely been at the centre of these research endeavours. Similarly, generally speaking, the automatic detection of figures of speech has mostly been focused on general texts, consumer reviews and social media. Although, especially as far as figurative language is concerned, the quality of the results obtained and the coverage of these detection methods still need to be further improved, it is undeniable that digital versions of literary texts could tremendously benefit from the presence of stylistic annotations of figures of speech both for research and teaching purposes .

Rather than attempting to propose a framework that takes into consideration all existing figures of speech, this paper focuses on the simile, defined as a figure of speech which relies on a linguistic marker to draw a parallel between two or more distinct entities or processes based on stated or implied (dis)similarities, so as to build up a specific image in another person's mind. If like the metaphor, the simile is based on resemblance and exemplifies figurative language, it is also structurally identical to literal comparisons, which are very effective to sustain an argumentation. In addition, the simile is flexible enough to be combined with other figures of speech, as illustrated by the following examples:

- Simile + personification/animation: Diseases like **snakes** *crawling over the earth, leaving trails of slime*. [Lowell (1916)]

- Simile + alliteration: [. . .] The spring *will come back* like **a blooming bride** [. . .] [Mason (1914)].
- Simile + humour: For statesmen are as *thick* as **fleas**, and poets, they are between [. . .] [Mason (1916)].

Therefore, with respect to their pervasiveness in language and the wide spectrum of figures of speech they can be associated with, similes provide a large framework to study in detail the issues related to the annotation of figures of speech in literary texts.

The present paper is divided into four main parts. So as to better understand how similes are described stylistically in scholarly texts, Section 2 presents the main simile types and structures. Section 3 reviews two existing annotated corpora while Section 4 describes the simile annotation scheme that has been developed to produce a standard corpus based on the data collected from the (Dis)Similitudes platform.[1] Finally, Section 5 concludes the paper and discusses future work.

2 Simile Types and Structures in Stylistics

Using the terminology introduced by Richards (1936) to designate the elements of a metaphor, excluding the marker of comparison, a simile comprises at most three elements:

- the tenor, which is the entity or process that is compared;
- the ground or "the basis on which the comparison is made" (Strachan and Terry (2000));
- and the vehicle or standard that is used to establish the comparison.

In practice, at the sentence level, apart from the marker and the vehicle, the remaining components could be omitted; for example, as opposed to a close simile that contains a ground, a simile without any ground is called an open simile as it leaves more room for imagination. Thus, a close simile such as "She is soft, crinkled like a fading rose" [Lowell (1916)] would be analysed as follows:

She	is *soft, crinkled*	like	**a fading rose**
tenor	ground	marker	vehicle

As a comparative sentence, such a simile, of course, has the canonical structure of the comparative construction in most Indo-European languages: **object of comparison + shared quantity/quality + marker of comparison + standard of comparison**. When putting side by side the simile "She is soft, crinkled like a fading rose" and the literal comparison "Her hands are soft, crinkled like her sister's", it becomes obvious that a simile and a literal comparison differ only in terms of semantics: a simile uses world knowledge to help deduce and picture specific features of an entity in relation to another entity which generally belongs to a different semantic domain while a comparison merely states whether two entities are equal or not. In addition, on the surface, a third type of grammatical constructions, referred to as pseudo-comparisons, has exactly the same syntactic structure as literal comparisons and consequently, similes. As a matter of fact, in some cases, markers of comparison convey an estimation ("approximation"), highlight a function ("identification"), introduce a hyponym ("exemplification") or coordinate terms ("coordination").

Example of an identification: And so he yielded to his fate, and came forth as a candidate. [Mason (1916)].

Unlike similes and literal comparisons, although metaphors can be introduced by some signalling words (Goatly (2011)), they do not altogether require any comparative marker as the comparison they establish is implied and mostly takes place in the mind. Furthermore, metaphors correspond to a number of rather different syntactic patterns among which:

- the stand-alone vehicle: The surgeon toiled the livelong night above the gory **wreck**; he got the ribs adjusted right, the wishbone and the neck [Mason (1916)];

[1] English version:`dissimilitudes.lip6.fr:8180`; French version: `dissimilitudes.lip6.fr:8181`

Similes comparing entities	**Similes involving processes**
. . . sad and **corpse**-like is his face, as he carves his ancient veal. [Mason (1916)]	Miniature rockets peppering the black bricks with golden stars, as though **a gala flamed a night of victorious wars.**[Lowell (1916)]
The great gift the gods bestowed on mortal was his dome of thought [. . .]; it sometimes seems a trifling thing, less *useful* than **one's lungs or slats**. [Mason (1916)]	The empty form drops from a cloud, like **a gourd from a vine** [. . .] [Williams (1920)]
No more of his triumphs he *lilted*, like **Spartacus spieling in Rome**; the steel hearted warrior wilted, and followed his conquerer. [Mason (1914)]	I gazed upon that mighty flood, that *writhed* as though **in pain or woe**. [Mason (1916)]

Table 1: Examples of similes based on the nature of the compared elements

- a nominal vehicle and a verb : Every hour that's **gone**'s a dead one, and another **comes** and **goes / Lasso**, then, the hour that's with you, **ride** it till its back is sore. [Mason (1916)];
- a genitive link: [. . .] in **the graveyard** of the ages hours will find their last repose [Mason (1916)];
- an adjective and a noun: Every hour that's gone's a **dead** one, and another comes and goes [Mason (1916)];
- a copula: [. . .] her eyes were **stars, from heaven torn**, and she was guiltless of a corn upon her sweet angelic toes. [Mason (1916)]

Similarly, within the confines imposed by the comparative sentence, similes can be expressed through various markers and syntactic constructions. In this respect, despite the lack of consensus between scholars, the stylistic analysis of similes in literary texts has mainly been developed around two non-mutually exclusive paradigms: their syntactic structure and their semantic components. While the former encompasses word order as well as the length and number of simile components, the latter describes the semantic leap at work in similes in terms of semantic categories, animacy or abstractedness.

2.1 Syntactic Description of Similes

Based solely on the simile structure and the grammatical category of the vehicle, it is possible to distinguish between similes that compare entities and those that involve processes. While the first type is mostly restricted to nominal vehicles, vehicles in the latter one can be prepositional phrases, whole or elliptical clauses (see Table 1).

Authors can also create very simple stylistic effects by changing the canonical sentence word order, for instance by inverting the vehicle or even the tenor.

Examples: And behind her *came, slowly* as **a hunter**, a young man who wore a cloak of two colours. [Wilde (2000)]

His brawn stands out in hummocks, he like **a lion** *treads*; he sits on foemen's stomachs and stands them on their heads. [Mason (1914)]

Furthermore, so as to emphasise a particular point or to make an image more vivid, a simile can have more than one tenor, ground or vehicle. In this respect, Pistorius (1971) calls a simile with two grounds such as "The big black dog *went stalking on*, as *calm and tranquil* as **the dawn** [. . .] [Mason (1916)]" a "doubled simile", whereas Kirvalidze (2014) refers to it as a "polymotivated simile". If the simile rather has a ground related to more than one vehicle, or two grounds with different vehicles, a different effect is created as a second image is created to reinforce or to enhance the first one.

Examples: Like **a blow**, **a kiss**, **a caress**, my songs *shall came*. [Anderson (1918)]

It's freighted with a gentle woe as *old* as **all the seas that flow**, as *young* as **yesterday**; as *changeless* as **the stars above**, as *yearning* as **a woman's love for true knight far away**. [Mason (1916)]

Degree of abstraction	**Degree of animacy**
abstract tenor-concrete vehicle Lord has no use for the twenty-cent skate, whose courage is *weak* as **the foam** [. . .] [Mason (1916)]	inanimate tenor-animate vehicle So the hours like **spotted ponies** *trot along in single file* [. . .] [Mason (1916)]
abstract tenor-abstract vehicle . . . when failure is as *certain* as **the coming of the dusk**, then it's wise to take your fiddle. [. . .] [Mason (1916)]	inanimate tenor-inanimate vehicle So does his fame, like **that lone mountain**, *rise, cleaving the mists and reaching the skies*[. . .] [Mason (1916)
concrete tenor-abstract vehicle Time to make a showing that your trade is growing, time to show your grit and *rustle round* like **the sin**. [Mason (1914)]	animate tenor-inanimate vehicle There's the man with hands so horny that they feel like **chunks of slate** [. . .] [Mason (1916)]
concrete tenor-concrete vehicle There's the man whose hand is *clammy* as **a fish that lately died** [. . .] [Mason (1916)]	animate tenor-animate vehicle And the neighbors come and chaff me, *laugh* like **horses at the door** [. . .] [Mason (1916)]

Table 2: Possible semantic combinations in similes

Besides increasing the number of similes components, creating a lack of balance between the length of the tenor and that of the vehicle often enables to shape a full-fledged image, typically by extending the vehicle with a relative clause.

Example: [. . .] I looked as *slick* as **a cabbage rose that's kissed by the nice wet dew**. [Mason (1916)]

Still on the structure of similes, Quintilian (Watson (1856)) observes that "sometimes the simile stands by itself and is unconnected; sometimes, as is preferable, it is joined with the object of which it is the representation, resemblances in the one answering to resemblances in the other". In the latter case, one would easily recognise the prototypical simile of the type "The spring will come back like a blooming bride" while the former case corresponds to elliptical similes devoid of tenor such as in: "*Firm* as **that mountain in the day of dread, when Freedom wept**, and pointed to her dead; *grim* as **that mountain to the ruthless foe, wasting the land that wearied of its woe** ; *strong* as **that mountain, 'neath his load of care, when brave men faltered in a sick despair**."[Mason (1916)]

In some cases, anticipation can be induced by running the simile on more than one sentence.

Example: He lay still, for the ash stick held him in place. Six months! Then her face came out of a mist of green. *Pink and white and frail* like **Dresden china**, lilies-of-the-valley at her breast, puce-coloured silk sheening about her. [Lowell (1916)]

2.2 The Semantic Dimension of Similes

Traditionally, the semantics of the similes is concerned with measuring the semantic distance between the tenor and the vehicle using semantic categories which can me more or less well-defined. Brooke-Roose (2002) summarises the various predominant theories that classify similes based on its content by distinguishing: first, Aristotle with the species/genus classification, then Aristotle's successors among whom Quintilian, who introduced the animate/inanimate classification. Afterwards, came the classification by domain of thought or activity used in the 19th and the 20th century for linguistic and literary analysis, and finally, the analysis by dominant trait which focused on the resemblances between the vehicle and the tenor. Therefore, it can be said that describing similes based on the semantic traits of the tenor and the vehicle or specific only to the vehicle has been a fixed feature of literary studies. Table 2 illustrates the various types of simile that are often used in scholarly literary texts about literature, depending on the degree of abstraction or animacy of the tenor and the vehicle.

3 Review of Existing Simile Annotation Schemes

Obviously, when talking about digital publishing and the annotation of literary texts, the first resource that comes to mind is the Text Encoding Initiative (TEI), which, in recent years, has imposed itself as the standard in the humanities for encoding additional information in texts. However, despite their exhaustive coverage of the encoding of various textual elements such as places or characters, the TEI P5 Guidelines[2] only address figurative language briefly and leave entirely the choice to the encoder:

> For other features it must for the time being be left to encoders to devise their own terminology. Elements such as `<metaphor tenor="..." vehicle="...">...</metaphor>` might well suggest themselves; but given the problems of definition involved, and the great richness of modern metaphor theory, it is clear that any such format, if predefined by these Guidelines, would have seemed objectionable to some and excessively restrictive to many. (6.7)

In the case of the simile, apart from the fact that the ground needs to be added, determining a posteriori how the different components are related in a sentence with more than one simile could be an issue.

Unlike what is suggested by the TEI, existing annotated corpora of similes have operated mainly at the word level. Niculae and Danescu-Niculescu-Mizil (2014) propose a corpus of annotated similes identified in Amazon product reviews which consists of 2,400 sentences in which a comparison between two common nouns has been automatically detected.[3] Each sentence is presented in the CoNLL format, the output format of the dependency parser used (TurboParser),[4] to which the mentions “TOPIC” for the tenor, “EVENT” for the verbal ground, “PROPERTY” for the adjectival ground, “COMPARATOR” for the marker and “VEHICLE” have been added when suitable. Before each sentence, metadata are given about the domain of the review, the annotators score about its figurativeness, the title of the review, the price of the article, the author of the comment…

Example: {”category”: ”Music”, ”figurativeness”: [4, 4, 4], ”title”: ”Siempre”, ”price”: ”unknown”, ”userId”: ”A20AEO9CWVD7JY”, ”score”: ”5.0”, ”helpfulness”: ”0/0”, ”time”: ”1182902400”, ”profileName”: ”Hilda Gonzalez Gonzalez”, ”productId”: ”B000NI3G8W”}

Their	their	PRP$	1	2	NMOD	_
voices	voice	NNS	2	3	SUB	TOPIC
blend	blend	VBP	3	0	ROOT	EVENT
like	like	IN	4	3	VMOD	COMPARATOR
magic	magic	NN	5	4	PMOD	VEHICLE
....	....	:	6	3	P	_

Clearly, because of the chosen output and of the information it contains, this corpus is mostly aimed at NLP researchers interested in the automatic detection of similes. Besides, in addition to its restrictive scope and the fact that all sentences are presented as stand-alone entities devoid of any context, this annotation scheme suffers from various other shortcomings: a principle of unicity (one simile per sentence and only one simile component annotated) and incorrect annotations resulting from parsing errors which were not rectified because the crowdsourcing task only dealt with figuativeness. Moreover, still in compliance with the parser’s output, auxiliaries and modal verbs preceding another verb are wrongly tagged as main verbs.

The second corpus, the VUAMC (Vrije Universiteit Amsterdam Metaphor Corpus) Online[5] (Steen et al. (2010)) is a manually annotated corpus of fragments of academic texts, conversations, fiction and news taken from the BNC Baby, a subset of the British National Corpus (BNC). It contains 16,202 sentences in which each word has been scanned to find out if it is used metaphorically or not. In addition, it

[2] `http://www.tei-c.org/Guidelines/P5/`

[3] `http://vene.ro/figurative-comparisons/`

[4] `http://www.cs.cmu.edu/~ark/TurboParser/`

[5] This corpus is searchable online (`http://www.vismet.org/metcor/search/showPage.php?page=start`) or can be freely downloaded as an XML file (`http://ota.ahds.ac.uk/headers/2541.xml`).

distinguishes between direct metaphors, implicit metaphors and words signalling metaphors also called “metaphor flags” into which fall simile markers. However, some of these metaphor flags such as “appearance”, “call”, “symbolically” and “types”, only precede an analogy or a metaphor. As a matter of fact, although as its name implies this corpus deals with the metaphor in its broadest sense, it devotes a rather small space to similes as exemplified by the discrepancy between the number of identified metaphorical words (more than 25,000) and the number of true similes (113 in total, 40 in the fiction fragments).

The downloadable version of this corpus is TEI-compliant and makes use of XML tags to delimit each sentence. Those tags also indicate for each word or punctuation mark, its part-of-speech tag, its lemma and whether it is a metaphorical word (`function="mrw"`) or a metaphorical signal (`function="mFlag"`). Furthermore, cases in which doubts subsist are also clearly indicated with the attribute `WIDLII` (when in doubt, leave it in). If this corpus constitutes a good basis to study metaphoricity in general, it does not say much about the reason why a particular word is metaphorical in a specific context or give information on the semantic structure of the identified similes.

```
 <s n="88">
    <w lemma="the" type="AT0">The </w>
    <w lemma="result" type="NN2">results </w>
    <w lemma="be" type="VBB">are </w>
    <w lemma="terse" type="AJ0">
        <seg function="mrw" type="met" vici:morph="n">terse</seg>
    </w>
    <w lemma="and" type="CJC">and </w>
    <w lemma="sharply" type="AV0">sharply </w>
    <w lemma="etch" type="VVN">
        <seg function="mrw" type="met" vici:morph="n">etched</seg>
    </w>
    <c type="PUN">, </c>
    <w lemma="like" type="PRP">
        <seg function="mFlag" type="lex">like</seg>
    </w>
    <w lemma="the" type="AT0">the </w>
    <w lemma="good" type="AJS">best </w>
    <w lemma="line" type="NN1">
        <seg function="mrw" xml:id="a1h-fragment05-cn2" type="lit" vici:morph="n">line</seg>
     </w>
     <w lemma="drawing" type="NN2">
        <seg function="mrw" corresp="#a1h-fragment05-cn2" type="lit" vici:morph="n">drawings
        </seg>
    </w>
    <c type="PUN">.</c>
</s>
```

4 The Proposed Annotation Scheme

This annotation framework has been designed with two main purposes in mind: to constitute the final output of a simile detection algorithm for prose literary texts written in English or in French (Mpouli (2016)) and to describe the similes and (pseudo-)comparisons that volunteers have annotated online. Built with scribeAPI [6], the (Dis)Similitudes crowdsourcing platform proposes 1,456 fragments of French, British and American prose poems published between the 18th and the 21st century. A fragment, here, refers to a sentence which contains one or more comparison markers and its surrounding sentences. Each fragment is presented as an image accompanied by a series of questions concerning the structure to analyse. To facilitate the annotation process, the corresponding marker has been coloured in blue beforehand.

Each volunteer can choose between two main tasks:

- answer questions on the structure to analyse (decide whether it is a comparison or a pseudo-comparison, give its function or pragmatic value, identify and describe its components);

- transcribe already annotated elements and indicate their semantic categories.

[6] http://scribeproject.github.io/

As the annotation process is still in progress, few conclusions can presently be drawn with certitude on the difficulty of the task as the whole or on the relevant information that it will reveal about the origin of figuration in similes. However, the data collected has enabled to confirm that indeed the whole phrase plays a role in creating and understanding the image conveyed by the simile since most annotators tend to mark phrases and not simple words.

To accurately render the different levels of analysis traditionally found in stylistics, similes in the developed annotation framework are annotated at a more general level and at the level of each of their components. The sentence constitutes the upper level of analysis and as such, each sentence is numbered and enclosed within a tag `<sentence>`. Then, the nature of the simile is specified: clausal similes (`<simile nature="clausal" >...</simile>`) or nominal similes (`<simile nature="nominal">... </simile>`). Five types of similes are distinguished:

- idiomatic similes (`<type= "idiomatic">...</type>`);
- perceptual similes which occur with a verb of perception like “look”, “sound”, “taste”, “smell” (`<type="perceptual">...</type>`);
- proverbial similes which occur with the verb “to be”, a nominal tenor and a nominal vehicle (`<type="proverbial">...</type>`);
- reinvented idiomatic similes (`<type="reinvented">...</type>`) in which the adapted form is of course mentioned with the attribute `source` followed by the typical form of the idiomatic simile.
- original similes (`<type="original">...</type>`) which are the creative ones.

Unlike what has been done so far, the tenor, the ground and the vehicle are annotated both at the word and at the phrase level. Concretely for each of these components, the mark up delimits the boundaries of the phrase to which it belongs, links it to the corresponding marker in the sentence and gives the grammatical category and the lemma of its head(s) as well as its position. Each marker in a sentence is identified by its position in the sentence which is encoded by the attribute `marker_id`. For multiword markers, only the position of the head lexeme is considered. Furthermore, additional stylistic information is given about the marker, whether it occurs at the beginning of the sentence, after a comma or a coordinating conjunction. Such information is introduced by an attribute `syntax`. Similarly, a tag `<rel>` signals vehicles that are followed by a relative clause while the tag `<neg>` indicates a negated ground.

The semantic category of the tenor and the vehicle is also specified via an attribute `category`. Even though some categories such as “humans” and “animals” seem to be quite agreed upon, it remains rather difficult to define semantic categories. After consulting lexical resources such as Fellbaum (1998) and the SIMPLE-CLIPS,[7] we opted for a set of categories neither too broad nor too refined which consists in the following semantic categories:

Concrete	Man-made objects Natural objects Body parts Human beings Animals Plants, fruits and vegetables
Abstract	Temporal elements Concepts Feelings and emotions Acts and processes Attributes and qualities
Collective nouns	

[7] `http://webilc.ilc.cnr.it/clips/Ontology.htm`

At the moment, for automatically generated annotations, when the tenor or the vehicle is not a common noun, the value of its semantic category is marked as ”undetermined”. Based on the annotations that have been collected so far, we checked the relevance of the selected semantic categories. In the French counterpart of the (Dis)Similitudes platform, the broader semantic categories fit almost perfectly with human annotations (98%) whereas the score drops significantly (67%) when it comes to further semantic distinctions especially as far as abstract entities are concerned. Such differences, of course, could be attributed to the polysemy of some words but also to personal sensibility. For instance, although all annotators agree that the term “cri” (“shout”) is an abstract entity, for some it denotes an act or process and for others an emotion. It is worth noting that annotators can also choose “Others” if they disagree with all the listed subcategories.
Example of an annotated simile:
** It’s freighted with a gentle woe as *old* as **all the seas that flow**, as *young* as **yesterday**; as *changeless* as **the stars above**, as *yearning* as **a woman’s love for true knight far away** [Mason (1916)].

```
<sentence id="2">
It’s freighted with
<simile nature="nominal" type="original">
<tenor marker_id="8,17,22,29">a gentle <head id="7" lemma="woe" postag="NN" category="abstract
    , emotions and feelings">woe</head>
</tenor>
<marker marker_id="8" lemma="as" syntax="null">as</marker>
<ground marker_id="8"><head id="9" lemma="old" postag="JJ">old</head>
</ground>
<marker marker_id="8" lemma="as">as</marker>
<vehicle marker_id="8">all the<head id="13" lemma="sea" tag="NNS" category="concrete,natural
    objects">seas</head><rel>that</rel></vehicle> flow,
</simile>
<simile nature="nominal" type="original">
<marker marker_id="17" lemma="as" syntax="juxt">as</marker>
<ground marker_id="17"><head id="18" lemma="young" postag="JJ">young</head></ground>
<marker lemma="as" marker_id="17">as</marker>
<vehicle marker_id="17"><head id="19" lemma="yesterday" tag="NN" category="abstract, temporal
    elements">yesterday</head>;</vehicle>
</simile>
<simile nature="nominal" type="original">
<marker marker_id="22" lemma="as" syntax="juxt">as</marker>
<ground marker_id="22"><head id="23" lemma="changeless" postag="JJ">changeless</head></ground>
<marker marker_id="22" lemma="as"  syntax="null">as</marker>
<vehicle marker_id="22">the<head id="26" lemma="star" tag="NN" category="concrete, natural
    objects">stars</head>above,</vehicle>
</simile>
<simile nature="nominal" type="original">
<marker marker_id="29" lemma="as"  syntax="null">as</marker>
<ground marker_id="29"><head id="30" lemma="yearning" postag="JJ">yearning</head></ground>
<marker marker_id="29" lemma="as" syntax="null">as</marker>
<vehicle marker_id="29">a woman’s<head id="36" lemma="love" postag="NN" category="abstract,
    emotions and feelings">love</head> for true knight far away.</vehicle>
</simile>
</sentence>
```

In addition, specifically for simile detection gold standards, two main types of structures can be distinguished:

- literal comparisons (<comparison>) and their three components <comparee_NP>, <quantity_quality> and <standard_NP>;
- and pseudo-comparisons with their respective values (exemplification, identification, coordination and approximation) and components.

Like simile components, all these components are described with the attributes id and marker_id as well as a child element head.

Examples:

1/ It was ten cents cheaper than suits I'd bought, from local dealers... [Mason (1916)]

```
<sentence id="2">
<comparison>
<comparee_NP marker_id="6">
<head id="1" lemma="it" postag="PRP" category= concrete, man-made objects">It</head>
</comparee_NP>
was
<quantity_quality  marker_id="6">ten cents <head id="5" lemma="cheap" postag="JJR">cheaper</
    head>
</quantity_quality>
<marker marker_id="6" lemma="than" syntax="null">than</marker>
<standard_NP marker_id="6"><head id="7" lemma="suit" tag="NNS" category="concrete, man-made
    objects">suits</head></standard_NP>
</comparison>
I'd bought, from local dealers...
</sentence>
```

2/ You yet may have a chance to serve as juryman, in court [Mason (1916)].

```
<sentence id="7">
<pseudo_comparison type= identification >
<identified_element marker_id="9">
<head id="1" lemma="you" postag="PRP" category= concrete, human beings">You</head>
</identified_element>
yet may have a chance to
<verb marker_id="9"><head id="8" lemma="serve" postag="VB">serve</head>
</verb>
<marker marker_id="9" lemma="as" syntax="null">as</marker>
<complement_marker marker_id="9"><head id="10" lemma="juryman" tag="NN" category="concrete,
    human beings">juryman</head></complement_marker>
</pseudo-comparison>
, in court.
</sentence>
```

5 Conclusion

This paper proposed a simile annotation framework that has been developed to produce a gold standard from the (Dis)Similitudes crowdsourcing platform and which takes into account stylistic practices as well the challenges specific to the automatic detection of similes. In this respect, it presents a multi-layered annotation scheme that describes each simile in the sentence and its respective components. As other figures of speech are not currently taken into consideration, the next step is to consider how to combine the current framework with the annotation scheme for irony, metaphor, metonymy and figures of repetition mentioned in the Introduction so as to accurately represent how these rhetorical figures are interconnected. Furthermore, it could be interesting to go beyond the sentence level as some similes can cover more than one sentence. Finally, to adequately single out creative similes, by mining literary corpora, it will be possible to separate original similes from cliché or frozen ones, i.e. those similes that are widely used among authors without being idiomatic such as the combination heart + beat + marker of comparison + hammer.

Acknowledgements

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

References

Anderson, S. (1918). *Mid-American Chants*. New York: B. W. Huebsch, Inc.

Brooke-Roose, C. (2002). *Invisible Authors: Last Essays*. Ohio State University Press.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Gawryjolek, J. J. (2009). Automated annotation and visualization of rhetorical figures. Master's thesis, University of Waterloo.

Gianti, A., C. Bosco, V. Patti, A. Bolioli, and L. Di Caro (2012). Annotating irony in a novel Italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pp. 1–7.

Goatly, A. (2011). *The Language of Metaphors*. London & New York: Routledge.

Gordon, J., J. R. Hobbs, J. May, M. Mohler, F. Morbini, B. Rink, M. Tomlinson, and S. Wertheim (2015). A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pp. 56–66.

Karoui, J., B. Farah, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles (2017). Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Volume 1, pp. 262–272.

Kirvalidze, N. (2014). Three-dimensional world of similes in English fictional writing. *Sino-US English Teaching 11*(1), 25–39.

Lowell, A. (1916). *Men, Women and Ghosts*. New York: Macmillan.

Markert, K. and M. Nissim (2002). Towards a corpus annotated for metonymies: The case of location names. In *LREC*, pp. 1385–1392.

Mason, W. (1914). *Rippling Rhymes*. Chicago: A. C. McClurg & Co.

Mason, W. (1916). *His Book*. New York: Barse & Hopkins Publishers.

Mpouli, S. (2016). *Automatic annotation of similes in literary texts*. Ph. D. thesis, Université Pierre et Marie Curie-Paris VI.

Niculae, V. and C. Danescu-Niculescu-Mizil (2014). Brighter than gold: Figurative language in user generated comparisons. In *EMNLP*, pp. 2008–2018.

Pistorius, G. (1971). La structure des comparaisons dans Madame Bovary. *Cahiers de l'Association internationale des études francaises 23*(1), 223–242.

Richards, I. A. (1936). *The Philosophy of Rhetoric*. London, Oxford & New York: The Oxford University Press.

Ruan, S., C. Di Marco, and R. A. Harris (2016). Rhetorical figure annotation with XML. In *Proceedings of the 16th Workshop on Computational Models of Natural Argument*, pp. 23–33.

Shutova, E. and S. Teufel (2010). Metaphor corpus annotated for source-target domain mappings. In *LREC*, Volume 2, pp. 3255–3261.

Steen, G. J., A. G. Dorst, J. B. Herrmann, A. Kaal, T. Krennmayr, and T. Pasma (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*, Volume 14. Amsterdam & Philadelphia: John Benjamins Publishing.

Strachan, J. R. and R. Terry (2000). *Poetry*. Edinburgh University Press.

Trevisan, B., M. Neunerdt, T. Hemig, E.-M. Jakobs, and R. Mathar (2014). Detecting ironic speech acts in multilevel annotated German web comments. In *Proc. Workshop NLP 4 CMC*, pp. 34–41.

Van Hee, C., E. Lefever, and V. Hoste (2016). Exploring the realization of irony in Twitter data. In *LREC*, pp. 1795–1799.

Watson, J. S. (1856). *Quintilian's Institutes of Oratory*. London: George Bell and Sons.

Wilde, O. (2000). *The Complete Works: Poems and Poems in Prose*. Oxford University Press.

Williams, W. C. (1920). *Kora in Hell: Improvisations*. Boston: The Four Seas Company.