Towards the Automatic Mining of Similes in Literary Texts

Suzanne Mpouli **Jean-Gabriel Ganascia**

Sorbonne Universités, UPMC, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
Labex OBVIL, Université Paris-Sorbonne, 1 rue Victor Cousin 75005 Paris
mpouli@acasa.lip6.fr, jean-gabriel.ganascia@lip6.fr

Abstract

Previous studies have shown that not only similes often greatly contribute in establishing the overall tonality of a literary text, but they can also express an author's particular view of the world. This paper presents the architecture of a system geared towards simile mining in literary texts written in English and French as well as some of its early applications.

1 Introduction

Similes can be defined as comparative constructions in which a parallel is drawn between two or more semantically unrelated entities or processes, often through a shared property, so as to produce a mental image in a person's mind. As figures of speech, similes play in essential role in literary texts by making descriptions more vivid and by conveying the right mental image to the readers. While at the syntactic level, they are characterised by specific patterns shared by various languages, semantics enables to distinguish literal comparative statements such as "The pan is as heavy as the pot" from similes like "The pan is as heavy as an elephant's paw". In this respect, studying similes could help to better understand figurative language. Besides, since comparing is a fundamental cognitive activity that relies on individual judgments and associations, similes constitute an interesting basis for studying linguistic creativity.

This paper is organised as follows. Section 2 gives an overview of the different modules of our system. Section 3 summarises the results of preliminary experiments realised on a corpus of French and British novels. Finally, we conclude with perspectives for future work.

2 Overview of the Simile Annotation System

For similes to be annotated in a given text, they first need to be detected. The prototypical simile "The pan is heavy like an elephant's paw", can be represented as "A Ω *y* X B", where A is any type of noun phrase, Ω is a verb, *y* is an adjective, X is a marker of comparison and C is a noun-headed noun phrase. In scholarly works discussing similes, A is commonly referred to as the tenor, Ω or *y* as the ground and B as the vehicle (Figure 1) (Fishelov, 1993).

The	pan		is	heavy		like		an elephant's paw.
	tenor			*ground*		*marker*		*vehicle*

Figure 1. Constituents of the simile "The pot is heavy like an elephant's paw".

In practice, our system is made up of three main modules: a syntactic module, a semantic module, and an annotation module.

2.1 The Syntactic Module

This module is concerned with several preprocessing tasks (tokenisation, lemmatisation, sentence detection, part-of-speech tagging, syntactic parsing), the selection of simile candidates and the identifica-

tion of each of their components. Since similes can take various forms, a system which seeks to accurately detect similes in texts should be flexible enough to adapt to various simile structures and markers as well as to take into consideration the ambiguity inherent to some simile constructions. In addition to the grammatical markers of comparison of both languages, were also considered other types of markers implying comparisons such as verbal, prepositional and adjectival phrases (see Table 1).

	Comparatives	**Verbal phrases**	**Adjectival phrases**	**Prepositional phrases**
English	*like, unlike, as, as...as, more...than, less...than*	*resemble, remind, compare, seem, verb + less than, verb + more than, be/become… kind/sort/type of*	*similar to, akin to, identical to, analogous to, comparable to, compared to reminiscent of, noun+-like , noun+colour*	
French	*comme, ainsi que, de même que, autant que, plus...que, tel que, moins...que aussi...que*	*ressembler à, sembler, , rappeler, faire l'effet de, faire penser à, faire songer à, donner l'impression de, avoir l'air de, verb + plus que, verb + moins que, être/devenir…espèce/type/genre/sorte de*	*identique à, tel, semblable à, pareil à, similaire à, analogue à, égal à, comparable à*	*à l'image de, à l'instar de, à la manière de à l'égal de, à la manière de, à la façon de*

Table 1. Similes markers for English and French

The selection of simile candidates starts with finding one of the markers in a sentence. Then, the other elements of the potential simile are identified according to their most probable relationship to the last known component, their nature, and their grammatical function (see Figure 2). By default, if various terms can be tagged as tenors or grounds, all plausible elements are extracted. A filtering, which uses manual rules and semantic information, takes place at a later stage to keep only the most pertinent labelled elements.

Vehicle	**Marker**	**Ground**	**Tenor**
Non-subject noun head of the noun phrase following the marker	All except -like and -colour	Verb (+adjective)	Verb subject or direct object
		Non-predicative adjective	Noun phrase modified by the adjective
			Noun phrase 1
First noun of the compound adjective containing the marker	-like and -colour	Verb	Verb subject or direct object
			Noun phrase 3

Figure 2. Possible syntactic scenarios.

From these different scenarios, it can be seen that grammatical roles are crucial to the extraction of simile components. It is, therefore, not surprising that dependency parsing had been previously used for this purpose (Niculae & Danescu-Niculescu-Mizil, 2014). We choose, however, to rely on syntactic chunking because our approach is mainly phrase-based and syntactic chunking tends to be more reliable when it comes to capturing close relationships as it is often the case with the ground and the marker. As far as the filtering of wrongly extracted components is concerned, we experimented with agreement rules, lists of transitive verbs and the Sketch Engine (Kilgariff et al., 2004) to extract corpus information on the use of the vehicle as the subject or the direct object of the found verb). Our preliminary tests on a corpus of French prose poems show that our approach (Recall: 66.5%, Precision: 66.3%) performs slightly better than the one based on dependency parsing (Recall: 62.4%; Precision: 64.2%) but captures too much candidates for tenors and has problems with long dependencies. In this respect, for better coverage, we are currently working on blending the two approaches.

2.2 The Semantic Module

Once all the simile components have been found, the next step is to determine whether they express a simile or a literal statement. At least one of the following conditions must be fulfilled for a comparative construction to be considered a simile:
- the ground + vehicle combination is recorded in a precompiled list of idiomatic similes;
- the ground expresses common conceptions about the vehicle, for example, 'calm' and 'lake';
- the vehicle is part of an extended noun phrase in a comparison of equality;
- the vehicle and the tenor are nouns belonging either to distinct semantic categories or to different subcategories of a broad semantic category (e.g 'penguins' and 'wolves').

Since accepted ideas about a particular word are connected to its usage, they are embedded in language. Consequently, we put together various French and English machine-readable dictionaries[1] to automatically retrieve specific linguistic pairs: nominal subject-verb, verb-nominal direct object, nominal subject-predicative adjective, adjective-noun. In addition, coordinated nouns, verbs, and adjectives are clustered together as synonyms.

Example: Adjectives frequently associated with 'biscuit': flavoured, crisp, rectangular, hard, crescent-shaped, flat, thin, crushed, dry, individual, burnt, unleavened, soft, oblong, German, small.

2.3 The Annotation Module

Although there exist some annotated corpora of similes, none has been devoted to the description of figures of speech for literary studies. Furthermore, despite briefly touching the question of metaphor annotation, the TEI guidelines do not provide a definitive framework, leaving the choice to the encoder.

We distinguish two levels of annotations:
- descriptive annotations which indicate the boundaries of the different elements of the simile, and state for the tenor, the ground, and the vehicle, both the whole phrase they are part of and its head;
- and analytical annotations which provide information about the semantic category of the tenor and/or the vehicle, the idiomaticity of the simile, its frequency in literary texts, and the fixedness of the couple tenor-vehicle or of the triplet tenor-ground-vehicle.

The semantic categories will be derived from coupling each common noun with the clusters of coordinated nouns and other lexicographical information obtained in the previous step. In contrast, for the remaining annotations, a series of experiments were designed to produce a basis to sustain future simile annotations.

Example: The pan is heavy like an elephant's paw.

```
<creative>
<tenor marker_id="4"> The <head lemma="pan" postag="NN" category="object"> pan
</head></tenor> is <ground marker_id="4"><head lemma="heavy" postag="JJ "> heavy
</head><marker lemma="like" marker_id="4" syntax="null"> like </marker> <vehicle
marker_id="4" >an elephant's<head lemma="paw" postag="NN" category="body part"> paw
</head></vehicle>.
</creative>
```

3 Examples of Experiments

For the purpose of this project, a reference corpus was put together. In total, 746 French novels and 1190 British novels written between the 1810s and the 1940s were downloaded using online digital libraries such as the Project Gutenberg[2] and the Bibliothèque électronique du Québec.[3]

The first experiment (Mpouli & Ganascia, 2016a) is focused on finding frozen similes in the reference corpus and on determining which ones are used as literary clichés. Two main patterns of frozen similes

[1] French: 8th and 9th editions of Le Dictionnaire de l'Académie française, Littré, Wiktionary; English: GCIDE, Wordnet, Wiktionary

[2] www.gutenberg.org

[3] beq.ebooksgratuits.com

were predefined: adjectival ground + simile marker + nominal vehicle (e.g. *happy as a lark)* and verbal ground + simile marker + nominal vehicle (e.g. *sleep like a top*). The generated results suggest that frozen similes are not so frequent in literary texts, which tends to sustain the idea that creativity plays a central role in literature. Interestingly, English and French share the same most frequent simile "pale as death" or "pâle comme la mort" with 152 and 182 occurrences respectively. In addition, to give new to frozen similes, novelists are very fond of replacing the verbal or the adjectival ground by a synonym or the canonical vehicle by a related noun or an extended noun phrase.
The second and last experiment (Mpouli & Ganascia, 2016b) studies noun+colour term (CT) similes of the type "storm-green sky" in order to investigate if their use of colours correlates the Berlin and Kay's hypothesis (1969) and how they differ from other similes. From the obtained results, there is no doubt that both noun+CT similes and fully-fledged similes function differently: while traditional similes are strongly governed by collocations and can be used figuratively more easily, noun+CT similes typically describe background elements. This dichotomy could perhaps explain the difference in their use of colours, confirming the idea that colours should not be taken in abstraction, but must be studied in a specific context. Moreover, it is possible to notice an increase in the number of noun+CT similes between the 19th and the 20th century, which suggests that noun+CT similes actively participate in shaping depictions in Modernist novels.

4 Conclusion

Simile annotation is an interesting natural language processing problem that requires not only syntactic but also lexical and semantic information. Apart from improving the identification of simile components, our system also proposes solutions to analyse different types of simile structures. The next phase, besides completing the implementation of the last modules, concerns the evaluation of the system. In this respect, in order to create a gold standard, a platform has been developed to collect annotations of similes in preselected prose poems.[4]

Acknowledgements

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

References

Berlin, B, & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley and Los Angeles: University of California Press.

Fishelov, D. (1993). Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric. *Poetics Today*, 14(1), 1-23.

Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004) The Sketch Engine. *Proceedings of EURALEX*, 105-115.

Mpouli, S., & Ganascia, J.-G. (2016a). "Pale as death" or "pâle comme la mort": Frozen similes used as literary clichés. *EUROPHRAS 2015: Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, 179-187.

Mpouli, S., & Ganascia, J.-G. (2016b; forthcoming). Another Face of Literary Similes: A Study of Noun+Colour Term Adjectives. Selected Papers of the Corpus Linguistics in France Conference.

Niculae, V., & Danescu-Niculescu-Mizil, C. (2014). Brighter than Gold: Figurative Language in User Generated Comparisons. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2008-2018.

[4] French version: dissimilitudes.lip6.fr:8180.