



# Automatic annotation of similes in literary texts

Suzanne Patience Mpouli Njanga Seh

► **To cite this version:**

Suzanne Patience Mpouli Njanga Seh. Automatic annotation of similes in literary texts. Other [cs.OH].  
Université Pierre et Marie Curie - Paris VI, 2016. English. NNT : 2016PA066298 . tel-01467081

**HAL Id: tel-01467081**

**<https://tel.archives-ouvertes.fr/tel-01467081>**

Submitted on 14 Feb 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Université Pierre et Marie Curie

École Doctorale ED 130

*Laboratoire d'Informatique de Paris VI, Équipe ACASA*

## **Automatic Annotation of Similes in Literary Texts**

Par Suzanne Patience Mpouli Njanga Seh

Thèse de doctorat en Informatique

Dirigée par Jean-Gabriel Ganascia

Présentée et soutenue publiquement le 03 octobre 2016

Devant un jury composé de :

M. Walter Daelemans, Professeur, Universiteit Antwerpen – Rapporteur

M. Stéphane Ferrari, Maître de conférences [HDR], Université de Caen – Rapporteur

Mme Catherine Fuchs, Directrice de recherche, LATTICE-CNRS – Examinatrice

M. Jean-Gabriel Ganascia, Professeur, UPMC – Directeur de thèse

M. Dominique Legallois, Professeur, Université Sorbonne Nouvelle – Examinateur

Mme Vanda Luengo, Professeur, UPMC – Examinatrice



Except where otherwise noted, this work is licensed under  
<http://creativecommons.org/licenses/by-nc-nd/3.0/>







# ABSTRACT

This thesis tackles the problem of the automatic recognition of similes in literary texts written in English or in French and proposes a framework to describe them from a stylistic perspective. In this respect, in the first part of this work, we are mainly interested in circumscribing the notion of simile and giving an overview of previous works and existing annotated corpora of similes and comparisons. For the purpose of this study, a simile has been defined as a syntactic structure that draws a parallel between at least two entities, lacks compositionality and is able to create an image in the receiver's mind.

In the second and last part, we present the designed method, its evaluation, and three of its possible applications in a literary context. Three main points differentiate the proposed approach from existing ones: it is strongly influenced by cognitive and linguistic theories on similes and comparisons, it takes into consideration a wide range of markers and it can adapt to diverse syntactic scenarios. Concretely speaking, it relies on three interconnected modules:

- a syntactic module, which extracts potential simile candidates and identifies their components using grammatical roles and a set of handcrafted rules,
- a semantic module which separates creative similes from both idiomatic similes and literal comparisons based on the salience of the ground and semantic similarity computed from data automatically retrieved from machine-readable dictionaries;
- and an annotation module which makes use of the XML format and gives among others information on the type of comparisons (idiomatic, perceptual...) and on the semantic categories used.

Finally, the two annotation tasks we designed show that the automatic detection of figuration in similes must take into consideration a series of features among which salience, categorisation and the sentence syntax.



## RÉSUMÉ

Cette thèse aborde le problème de la détection automatique des comparaisons figuratives dans des textes littéraires en prose écrits aussi bien en français qu'en anglais et propose un canevas pour décrire ces comparaisons d'un point de vue stylistique. A cet effet, dans la première partie de ce travail, nous nous sommes attelés à circonscrire la notion de comparaisons figuratives et à présenter un panorama des précédents travaux réalisés dans le domaine ainsi que des pratiques hétérogènes en matière d'annotations de comparaisons dans des corpus de textes. Par comparaison figurative, il est entendu, dans le cadre de cette étude, toute structure syntaxique qui met en parallèle au moins deux entités, déroge au principe de compositionnalité et crée une image mentale dans l'esprit de ceux à qui elle est destinée.

Dans la seconde partie de cette thèse, nous présentons notre méthode, quelques résultats d'évaluation ainsi que trois de ses possibles applications à des questions littéraires. Trois éléments principaux distinguent notre approche des travaux précédents : son ancrage dans les théories linguistiques et cognitives sur les comparaisons littérales et figuratives, sa capacité à gérer des marqueurs appartenant à différentes catégories grammaticales et sa flexibilité qui lui permet d'envisager différents scénarios syntaxiques. De manière plus concrète, nous proposons une méthode s'articulant autour de trois modules complémentaires :

- un module syntaxique qui utilise la structure syntaxique et des règles manuelles pour identifier les comparaisons potentielles ainsi que leurs composantes ;
- un module sémantique qui mesure la saillance des motifs détectés et la similarité sémantique des termes comparés en se basant sur des données recueillies automatiquement dans des dictionnaires électroniques ;
- et un module d'annotation qui s'appuie sur le format XML et fournit entre autres des informations sur le type de comparaison (idiomatique, sensorielle...) et sur les catégories sémantiques employées.

Pour finir, au vu des données recueillies au cours des deux campagnes d'annotation que nous avons menées, il paraît clair que la détection automatique des comparaisons figuratives doit tenir compte de plusieurs facteurs parmi lesquels la saillance du motif, la catégorie sémantique des termes comparés et la syntaxe de la phrase.



## ACKNOWLEDGEMENTS

Doing a PhD thesis is generally described as a nerve-racking experience, but it is also fulfilling in its own way as it enables to meet new people, to learn new things and to discover one's own strengths and weaknesses as an academician.

For this mind-shaping experience, I would like to thank my supervisor, Prof. Jean-Gabriel Ganascia, for the enthusiasm he has always shown for my thesis subject and for giving me free reins to choose in which directions to steer my research.

I am, of course, very grateful to the LabEx OBVIL and its Director, Prof. Didier Alexandre for funding my PhD.

I would particularly like to convey my sincere appreciation to Prof. Walter Daelemans, Mr Stéphane Ferrari, Prof. Dominique Legallois, Prof. Vanda Luengo and Ms Catherine Fuchs who readily accepted to be members of my jury and endeavoured to fit my defense in their schedule, in some cases, despite prior engagements.

I would also like to thank my two annotators, Emmanuelle Kaas and Pauline Bruley, for their work and for sharing their insightful remarks with me. I was also positively influenced this last year by the constructive criticisms of the two members of my midterm defense, Prof. Milad Doueihy and Mr Eric de la Clergerie.

For these three years full of ups and downs, laughter, food, drinks, productive and less productive brainstorming, I would like to thank all my colleagues, past and present, of the LIP6-ACASA & the LabEx OBVIL: Alexandre, Amal, Amine, Bin, Carmen, Chiara, Elodie, Fiona, Francesca, Gauvain, Marine, Marianne, and Marissa. I am tremendously indebted to Mihnea Tufis who poured more than his soul into the Dissimilitudes project. Words cannot express how much I admire his dedication and his quest for perfection.

On a more personal note, I would like to thank my uncle and his partner for their silent support and for putting up with my odd schedule.

I am also indebted to all my friends who inquired about the progress of my thesis and patiently listened to me each time I needed it. I would specially like to express my deepest gratitude to Albert and Christelle who took time to read parts of my thesis.

Last but not least, I know that I could never have done it without my wonderful cheerleading team: my parents, Marthe and Martin, to whom I owe everything and who have always encouraged me; my brothers, Guillaume and Thierry, who are never afraid to give me advice and to point my mistakes; my in-laws, distant relatives and nephews, Raul, Emmanuella, Céline, Anatole, Melvin, Yann-Karel & Jaycee who were a breath of fresh air. My heartfelt thanks go to my sister, Prisca, who has always believed in me and helps me every day to become a better version of myself.

# CONTENTS

<b>1 INTRODUCTION</b> .....	<b>10</b>
1.1 STYLISTICS AND THE STUDY OF LITERATURE .....	10
1.2 INTRODUCING RHETORICAL FIGURES .....	12
1.3 RHETORICAL FIGURES AND COMPUTER-ASSISTED STUDIES OF LITERARY TEXTS .....	14
1.4 SCOPE OF THE THESIS .....	16
1.5 MOTIVATION OF THE STUDY.....	18
1.6 ORGANISATION OF THE THESIS.....	19
<b>2 SIMILES, COMPARISONS, METAPHORS AND FIGURATIVENESS</b> .....	<b>25</b>
2.1 COMPARISON: SEMANTICS AND SYNTAX .....	26
2.1.1 <i>Comparison in Rhetoric</i> .....	26
2.1.2 <i>Grammatical Expressions of Comparisons</i> .....	30
2.2 COMPARISONS AND SIMILES .....	34
2.2.1 <i>Comparisons of Inequality and Similes</i> .....	34
2.2.2 <i>Cognitive Accounts of Similes and Comparisons</i> .....	36
2.3 FIGURATIVE SIMILES .....	44
2.4 METAPHOR AND SIMILES .....	47
<b>3 COMPUTATIONAL APPROACHES TO SIMILE DETECTION</b> .....	<b>53</b>
3.1 CHALLENGES OF COMPUTATIONAL DETECTION OF SIMILES .....	54
3.1.1 <i>Markers' Polysemy</i> .....	54
3.1.2 <i>Comparison and Ellipsis</i> .....	57
3.2 COMPUTATIONAL APPROACHES .....	60
3.2.1 <i>Automatic Detection of Comparatives</i> .....	60
3.2.2 <i>Detection and Analysis of Non-Literal Comparisons</i> .....	69
3.2.3 <i>Automatic Detection of Similes</i> .....	73
<b>4 SIMILE ANNOTATION</b> .....	<b>81</b>
4.1 PRINCIPLES .....	82
4.1.1 <i>Types of Linguistic Annotations</i> .....	82
4.1.2 <i>The TEI as the Annotation Standard in the Humanities</i> .....	87
4.2 SIMILE DESCRIPTION IN LITERARY STUDIES .....	90
4.2.1 <i>The Structural Dimension</i> .....	90
4.2.2 <i>The Semantic Dimension</i> .....	95
4.3 EXISTING CORPORA OF ANNOTATED COMPARISONS AND SIMILES.....	99

<b>5 THE PROPOSED APPROACH.....</b>	<b>111</b>
5.1 A GRAMMAR OF THE SIMILE .....	112
5.2 THE SYNTACTIC MODULE .....	120
5.3 THE SEMANTIC MODULE.....	137
5.4. THE ANNOTATION MODULE .....	144
<b>6 TOWARDS AN ANNOTATED LITERARY CORPUS OF SIMILES .....</b>	<b>151</b>
6.1 CORPUS PRESENTATION .....	152
6.2 EXPERTS' ANNOTATION .....	154
6.3 THE CROWDSOURCING ANNOTATION PLATFORM .....	157
<b>7 CORPUS-BASED APPLICATIONS.....</b>	<b>165</b>
7.1 CORPUS DESCRIPTION.....	165
7.2 STEREOTYPICAL FROZEN LITERARY SIMILES .....	167
7.3 COLOURS AND SIMILES IN THE ENGLISH CORPUS .....	169
7.3.1 <i>Why Study Colours in relation to Similes?</i> .....	169
7.3.2 <i>Basic Colour Terms and English Literature</i> .....	171
7.3.3 <i>Fully Fledged Colour Similes vs. Noun+CT Similes: Frequency and Stylistic Usage</i> ....	174
7.3.4 <i>Creativity and Noun+CT Similes</i> .....	179
7.4 ON PROPER NOUNS IN COMPARATIVE CONSTRUCTIONS .....	183
<b>8 CONCLUSION AND FUTURE WORK.....</b>	<b>191</b>
<b>9 REFERENCES .....</b>	<b>195</b>
<b>10 APPENDICES .....</b>	<b>209</b>

## LIST OF TABLES

TABLE 2.1 <i>COMPARATIO, SIMILITUDO AND DISSIMILITUDO</i> : DEFINITIONS AND EXAMPLES.....	28
TABLE 2.2 EXAMPLES OF SIMILES AND COMPARISONS WITH THEIR RESPECTIVE VALUES .....	29
TABLE 2.3 MAIN COMPARATIVES IN ENGLISH AND IN FRENCH AND EXAMPLES OF THEIR USAGE.....	32
TABLE 2.4 TYPES OF COMPARISONS AND CORRESPONDING SIMILES .....	34
TABLE 2.5 SALIENT FEATURES OF “CHAIR”, “ARMCHAIR” AND “BOULDER” .....	40
TABLE 2.6 ANATOMY OF FOUR SIMILES.....	50
TABLE 4.1 SYNTHESIS OF THE VARIOUS SYNTACTIC STRUCTURES OF THE SIMILES FOUND IN <i>DAVID COPPERFIELD</i> AND <i>OLIVER TWIST</i> .....	95
TABLE 4.2 COMBINATIONS OF DEGREES OF ABSTRACTION AND OF ANIMACY.....	98
TABLE 5.1A SIMILE MARKERS USED AS PREDICATES: GRAMMATICAL PATTERNS AND EXAMPLES .....	113
TABLE 5.1B SIMILE MARKERS USED AS CONJUNCTIONS: GRAMMATICAL PATTERNS AND EXAMPLES .....	114
TABLE 5.2 SELECTED SIMILE MARKERS FOR BOTH LANGUAGES .....	122
TABLE 5.3 CONLL OUTPUT FOR THE SENTENCE “HIS EYES DILATED AND GLISTENED LIKE THE LAST FLAME THAT SHOOTS UP FROM AN EXPIRING FIRE.”.....	130
TABLE 5.4 CORRELATION BETWEEN EACH TYPE OF CONSTITUENT, THE CLUES TO IDENTIFY IT AND ITS GRAMMATICAL FUNCTION .....	132
TABLE 5.5 SIZE OF THE CORPUS OF FRENCH PROSE POEMS .....	133
TABLE 5.6 RESULTS OBTAINED WITH THE PROPOSED ALGORITHM (LEFT) AND WITH THE BERKELEY PARSER (RIGHT) .....	134
TABLE 5.7 MACHINE-READABLE DICTIONARIES USED .....	141
TABLE 5.8 EVALUATION OF AUTOMATICALLY EXTRACTED SALIENT TRAITS AND SYNONYMS .....	144
TABLE 5.9 CORRESPONDENCE BETWEEN OUR SEMANTIC CATEGORIES AND WORDNET’S UNIQUE BEGINNERS .....	147
TABLE 5.10 SUMMARY OF THE ANNOTATION SCHEME.....	149

TABLES 6.1 A & B. STATISTICS ON THE DISTRIBUTION OF MARKERS IN THE ENGLISH (LEFT) AND FRENCH (RIGHT) ANNOTATION CORPORA.....	154
TABLE 7.1 THE 10 MOST FREQUENT SIMILES IN BOTH CORPORA.....	167
TABLE 7.2 MOST FREQUENT TRIPLETS IN BOTH LANGUAGES WITH THE DEGREE OF FIXEDNESS OF THE VEHICLE.....	168
TABLE 7.3 MOST FREQUENT PAIRS IN BOTH CORPORA.....	168
TABLE 7.4 BASIC COLOUR TERM DEPENDING ON THE NUMBER OF COLOURS EXPRESSED IN THE LANGUAGE.....	172
TABLE 7.5 COLOUR TERMS SELECTED FOR THE EXPERIMENT .....	174
TABLE 7.6 PATTERN DISTRIBUTION OF THE 5 MOST FREQUENT COLOUR TERMS USED IN NOUN+CT SIMILES .....	175
TABLE 7.7. EXAMPLES OF REINVENTED CONVENTIONAL NOUN+CT SIMILES.....	178
TABLE 7.8 LEXICAL DIVERSITY PER COLOUR FOR BOTH LITERARY PERIODS.....	182
TABLE 7.9 TOP NAMES OF PEOPLE AND OF GEOGRAPHICAL PLACES.....	184

## LIST OF FIGURES

FIGURE 2.1 OVERVIEW OF THE CONSTRUCTION OF THE COMPARATIVE SENTENCE “THAT GIRL IS MORE GRACEFUL THAN A LILY” .....	33
FIGURE 2.2 ILLUSTRATION OF THE THREE LEVELS OF NATURAL CATEGORIES (ROSCH, 1978) .....	39
FIGURE 2.3 RESULTS OF TVERSKY’S CONTRAST MODEL FOR “THIS CHAIR IS LIKE AN ARMCHAIR” AND “THIS CHAIR IS LIKE A BOULDER” .....	41
FIGURE 3.1 SYNTACTIC VERSATILITY OF “COMME” .....	55
FIGURE 3.2 EXAMPLES OF SEMANTIC INTERPRETATIONS OF COMPARATIVE SENTENCES .....	61
FIGURE 3.3 CORRESPONDENCE BETWEEN VARIOUS TEMPERATURES AND NATURAL ELEMENTS .....	71
FIGURE 3.4 ADAPTATION OF THE STRUCTURE-MAPPING THEORY TO THE SENTENCE “THE HYDROGEN ATOM IS LIKE THE SOLAR SYSTEM” .....	72
FIGURE 3.5 EXAMPLES OF CATEGORIES FOR THE NOUN “GLADIATOR” GIVEN BY THE THESAURUS REX (VEALE & LI, 2013) .....	74
FIGURE 3.6 EXAMPLE OF A SENTENCE OUTPUT (NICULAE, 2013) .....	76
FIGURE 4.1 EXAMPLE OF AN XML DOCUMENT .....	87
FIGURE 4.2 EXAMPLE FROM THE VUAMC ONLINE .....	106
FIGURE 5.1 POSSIBLE SYNTACTIC STRUCTURES OF SIMILES .....	119
FIGURE 5.2 EXTRACTION OF POTENTIAL SIMILE CANDIDATES .....	123
FIGURE 5.3 THE IMPACT OF LEXICAL RESOURCES ON THE TENOR’S RECALL AND PRECISION .....	136
FIGURE 5.4 EXAMPLE OF AN ENTRY IN THE GCIDE .....	141
FIGURE 5.5 EXAMPLE OF NOUN SEMANTIC CATEGORISATION USING LE DICTIONNAIRE ELECTRONIQUE DES MOTS (DUBOIS & DUBOIS-CHARLIER, 2010) .....	148
FIGURE 6.1 SAMPLE OF TWO ANNOTATIONS .....	155
FIGURE 6.2 AN EXAMPLE OF THE ANNOTATOR’S HESITATION AND OF A CORRECTION .....	156
FIGURE 6.3 EXAMPLE OF AN IMAGE TO ANNOTATE .....	158
FIGURE 6.4 EXAMPLE OF AN ANNOTATED SENTENCE IN THE ORIGINAL ZOONIVERSE INTERFACE .....	159
FIGURE 6.5 STARTING QUESTION OF THE ANNOTATION PLATFORM .....	160
FIGURE 6.6 EXAMPLE OF AN ANNOTATED SENTENCE .....	161

FIGURE 6.7 EXAMPLE OF A TRANSCRIPTION TASK .....	162
FIGURE 6.8 ELEMENTS ASSOCIATED WITH EACH SUBTYPE OF PSEUDO-COMPARISON .....	163
FIGURE 7.1 DISTRIBUTION OF THE NOVELS IN THE BRITISH (TOP) AND THE FRENCH (BOTTOM) CORPORA PER DECADE FROM THE 1810S TO THE 1950S .....	166
FIGURE 7.2 COLOUR TERM DISTRIBUTION IN FULLY FLEDGED COLOUR SIMILES AND NOUN+CT SIMILES .....	175
FIGURE 7.3 RELATIVE FREQUENCY OF EACH COLOUR TERM PER DECADE .....	181



# 1 INTRODUCTION

## 1.1 Stylistics and the Study of Literature

The incredible power of language cannot be denied; after all, according to the Judaeo-Christian tradition, each and every single little thing on Earth has been created only with words. Indeed, through language, it is possible to immerse people in fictional stories and settings as well as to make them experience actual events of the past as vividly as if they were actually there. Even when no storytelling is involved, language can appeal to our emotions and our intellect when, for instance, it persuades us of the soundness of an argument or moves our hearts to tears. Therefore, it is not surprising that since the Ancient Greeks, language has been a constant object of study and scrutiny, giving birth to innumerable accounts on how it should be best practised and surveyed. Regardless of the period or of the school of thought, it is generally agreed upon that although what is said has its importance, it is mainly the language strategies used to say it that makes it powerful and enables it to touch the audience more effectively. These language strategies chosen knowingly or not among all the possibilities offered by each specific language to achieve a particular effect and that distinguish an individual's or a group's production from another's are what Bally (1909) describes as the core subject of stylistics.

As its name implies, stylistics is concerned with the study of style. If the subject matter of Bally's stylistics is obviously far from being new, its scientific nature, methods, and scope differentiate it from previous studies of style. While rhetoric is restricted to "the faculty of discovering the possible means of persuasion in reference to any subject whatever" (Aristotle, trans. 1926, Book 1, Chapter 2, p. 15), stylistics is far more ambitious as it is interested in the relationship between language elements and emotions: how emotions are

expressed through language as well as the impact of language on the emotions (Bally, 1909). In practice, a stylistic analysis implies identifying a linguistic unit that shows its user's way of thinking, finding its logical equivalent in language, comparing both of them in order to assess its affective or intellectual nature and classifying it according to its affective nature based on the different connotations (aesthetic value, exaggeration, attenuation, language register, specific domain...) it embodies. Also called linguistic stylistics, this type of stylistics examines linguistic units not only in relation to an author's style or text, but from a general perspective, so as to catalogue linguistic usages that are specific to a particular language (Bally, 1909; Jenny, 1993). In contrast, literary stylistics, which is nowadays the most predominant form of stylistics, supports the understanding and the interpretation of a particular literary text by showing how some of the linguistic elements it contains interact to produce a particular effect (Carter & Simpson, 1989). Typically, a stylistic analysis of a literary text would either focus on a chosen linguistic phenomenon and study its impact on the text, or would start from a prevalent feature or idea of the text to investigate how it is linguistically expressed (Ullmann, 1964).

At a time when literary criticism was emerging as a discipline on its own and some scholars were in search of an objective methodological approach to govern their research endeavours, stylistics and consequently linguistics appeared as the most appropriate frameworks on which they could rely on, especially if taken into account the preoccupations those disciplines have in common: their interest in the verbal structure as a whole and in the diachronic as well as the synchronic use of language (Jakobson, 1960). In addition, stylistics provides to literary studies the necessary weapons to question the aesthetic value and the uniqueness of the text(s) at hand (Fahnestock, 2011). Though linguistic creativity is not restricted to literary texts, it is often believed that its finest examples especially abound in literature, particularly in poetry. In this respect, literary stylistics focuses both on how a text adheres to general trends or reflects the speech of a specific community and on how it deviates from an implicit established norm. Therefore, literary stylistics is connected to *elocutio*, the part of rhetoric which is concerned with the artistic use of language and the study of figures of speech (Levin, 1982).

## 1.2 Introducing Rhetorical Figures

From the earliest surviving texts, rhetoric has tried to formalise, classify and enumerate the various devices that often adorn human discourse. Even though several systems of rhetorical devices have marked the history of rhetoric, the initial separation of figures into three main groups is still prevalent nowadays. This distinction has first been introduced in *Rhetorica ad Herennium* (trans. Caplan, 1954), in which figures are primarily divided into two main groups:

- figures of diction which deal with a particular arrangement of words;

### Examples

a) *Isocolon*: [The father was meeting death in battle]; [the son was planning marriage at his home] (Caplan, 1954, IV. XX. 27, p. 299).

b) *Antistrophe*: Since the time when from our state concord disappeared, liberty disappeared, good faith disappeared, friendship disappeared, the common weal disappeared (Caplan, 1954, IV. XIII. 19, p. 277).

c) *Homoeoteleuton*: You dare to act dishonourably, you strive to talk despicably; you live hatefully, you sin zealously, you speak offensively (Caplan, 1954, IV. XX. 28, p. 301).

d) *Antithesis*: To enemies you show yourself conciliatory, to friends inexorable (Caplan, 1954, IV. XV. 21, p. 283).

e) *Apostrophe*: Plotters against good citizens, villains, you have sought the life of every decent man! Have you assumed such power of your slanders thanks to the perversion of justice? (Caplan, 1954, IV. XVI. 22, p. 285).

- and figures of thought which concern the specific ideas that are conveyed, independently from how it is formulated.

### Examples

1. *Conciseness*: Just recently consul, [newt he was first man of the state]; [then he sets out for Asia], [next he is declared a public enemy and exiled]; [after that he is made general-in-chief] and [finally consul for the seventh time] (Caplan, 1954, IV. LIV. 68, p. 405).

2. *Emphasis*: Out of so great a patrimony, in so short a time, this man has not laid by even an earthen pitcher wherewith to seek a fire for himself (Caplan, 1954, IV. LIV. 67, p. 401).

3. *Personification*: But if the invincible city should now give utterance to her voice, would she speak as follows? (Caplan, 1954, IV. LII. 65, p. 399).

4. *Comparison*: Just as the swallows are with us in summer time, and when driven by the frost retire, so false friends are with us in a peaceful season of our life, and as soon as they have seen the winter of our fortune, they fly away, one and all (Caplan, 1954, IV. XLVIII. 61, p. 383).

5. *Simile*: His body was as white as snow, his face burned like fire (Caplan, 1954, IV. XXXII. 44, p. 341).

Apart from these two well-defined blocks, a subset of ten figures of diction is further set apart based on the shift in meaning that characterises them:

There remain also ten Figures of Diction, which I have intentionally not scattered at random, but have separated from those above, because they all belong in one class. They indeed all have this in common, that the language departs from the ordinary meaning of the words and is, with a certain grace, applied in another sense. (Caplan, 1954, Book IV. 42. XXXI, p. 333)

This last subset is made up of:

- the *onomatopoeia*: After this creature attacked the republic, there was a hullabaloo among the first men of the state (Caplan, 1954, IV. XXXI. 42, p. 335).

- the *metonymy*: Italy cannot be vanquished in warfare nor Greece in studies (Caplan, 1954, IV. XXXII. 43, p. 337).

==> The Italians cannot be vanquished in warfare nor the Greeks in studies.

- the *antonomasia*: Surely the grandsons of Africanus did not behave like this! (Caplan, 1954, IV. XXXI. 42, p. 335).

==> Surely the Gracchi did not behave like this!

- the *periphrasis*: The foresight of Scipio crushed the power of Carthage.

(Caplan, 1954, IV. XXXII. 43, p. 337).

==> Scipio crushed Carthage.

- the *hyperbaton*: Object there was none. Passion there was none (Poe, 1884).

- the *hyperbole*: But if we maintain concord in the state, we shall measure the empire's vastness by the rising and the setting of the sun (Caplan, 1954, IV. XXXIII. 44, p. 337).

- the *catachresis*: "a mighty speech", "to engage in a slight conversation" (Caplan, 1954, IV. XXXIII. 45, p. 343).

- the *metaphor*: The insurrection awoke Italy with sudden terror (Caplan, 1954, IV. XXXIV. 45, p. 343).

- the *synecdoche*: Were not those nuptial flutes reminding you of his marriage? (Caplan, 1954, IV. XXXII. 43, p. 337).

==> Was not this marriage party reminding you of his marriage?

- and the *allegory*: For when dogs act the part of wolves, to what guardian, pray, are we going to entrust our herds of cattle? (Caplan, 1954, IV. XXXIV. 46, p. 345).

Some decades later, Quintilian (trans. 1876) proposes the term “trope” to refer to all the rhetorical devices that require “the conversion of a word or phrase, from its proper signification to another, in order to increase its force” (Book VIII, chapter VI.1, p. 124), as opposed to the more general term, “figure”, which “is a form of speech differing from the common and ordinary mode of expression” (Book IX, chap I. 5, p. 145).

In addition to the terminology coined by traditional rhetoricians, stylistics has also inherited from rhetoric the habit of passing judgement on the soundness of an author’s figure. But, as the metaphor and by association figurative language stopped being confined to extraordinary language to become an inherent part of our way of thinking (Lakoff & Johnson, 1980), tropes also started to be studied with respect to the role they play in the understanding of the mental processes involved both in the production and the reception of literary texts. However, the introduction of cognitive sciences into stylistics does not restrict itself to the treatment of tropes, but also pave the way, especially as far as writing texts is concerned, for computer-based quantitative approaches to literature.

### 1.3 Rhetorical Figures and Computer-assisted Studies of Literary Texts

If some linguistic units of the text reflect a particular vision of the world, it seems logical to deduce that to have an impact on the reader, these units would be often repeated. This intuition, far from being new, is already suggested in the second half of the 19<sup>th</sup> century when Baudelaire (1885) quotes a critic who depicts repeated words as the ideal shortcut to a writer’s mind:

Pour deviner l’âme d’un poète, ou du moins sa principale préoccupation, cherchons dans ses œuvres quel est le mot ou quels sont les mots qui s’y représentent avec le

plus de fréquence. Le mot traduira son obsession. (p. 368)<sup>1</sup>

If some early research has explored ways to connect manually acquired frequencies to textual meaning, computers bring in a totally new dimension: not only are the results they generate more verifiable and replicable (Milic, 1991), but they also make it easier to devise or investigate new measures to account for linguistic phenomena such as vocabulary richness or text complexity so as to shed new light on overstudied texts. Moreover, with the rapidly increasing number of digitised texts and the advances in natural language processing, it became possible to compare larger sets of texts on different linguistic levels (word-level, sentence-level, phonetic level, syntactic level, ...). The use of computers to quantify style suffers nonetheless from various shortcomings: automatic analyses performed by computers often contain mistakes, the obtained results are not always easy to interpret and in most of the time, figurative language is not at all taken into consideration (Warwick, 2004).

Since figurative language is pervasive in language, tackling its automatic recognition and its understanding is perceived as a way to improve the performance of information retrieval systems and to provide sufficient grounds to create systems that can generate figurative language as naturally as human beings. If most of the research in this direction has been done on metaphors, metonymy, idioms and indirect speech acts (Martin, 1996), other rhetorical figures have been addressed mainly in relation to style (anadiplosis, epanalepsis, gradatio, kyklos, anaphora, epiphora, symploche in Dierks, 1989), creativity (irony in Veale, 2013) and to their role in argumentation (among others ellipsis, alliteration, antimetabole, apocope, epizeuxis and polysyndeton in Harris & DiMarco, 2009). In the field of digital humanities, apart from the stylistic comparison of text corpora, the automatic detection of rhetorical figures also potentially opens the door to the encoding of useful stylistic information in the texts for visualisation tasks, teaching purposes or further investigation.

As a matter of fact, with the never-ending growing number of digitised texts available, arises the need to mark up those texts with pertinent information such as metadata (author's name, publisher, date of publication...), divisions of the text (line, stanza,

---

<sup>1</sup> English translation: "To divine the soul of a poet, or at least his principal preoccupation, look for the word or words that recur in his work with most frequency. They will betray his obsession" (as cited in Mansell Jones, 1969, p. 147-148).

paragraph...) or stylistic information (sentence length, word frequency...). Of course, adding such information in a large quantity of texts requires tools to create them more or less automatically as well as to process them. Very early in the history of the digital era, descriptive markup, which points to each encoded element and identifies it using an explicit name, was adopted as the most adequate format for extra-textual information as it simplifies composition, editing, publishing and information retrieval (Renear, 2003). As far as literary computing is concerned, extracted information has mostly been used to build concordance lists and to count the occurrences of various linguistic units (Hockey, 1994). Delcourt (2002) notes, however, that the fuzziness of rhetorical figures makes them improper to be encoded in a corpus as the markup in a corpus should be “integral, uncontroversial and consistent” (p. 991). This could possibly explain why, apart from some marginal works or projects like the Augmented Criticism Lab,<sup>2</sup> literary computing has focused on easily computable statistical distributions such as the frequency of part-of-speech tags or of function words, at the expense of more established literary notions such as rhetorical figures.

## 1.4 Scope of the Thesis

The present thesis studies how a specific figure, namely the simile, can be automatically identified in prose literary texts written in English and French, and described from a stylistic perspective. Therefore, it seeks, at the macrolevel, on the one hand, to reconcile digital humanities with traditional rhetoric and on the other hand, to explore new directions in literary computing. But, why stop at one figure and more importantly, why the simile?

Even though rhetorical figures are often mentioned as a whole in relation to their combined impact on a text or an occurrence, they actually have different internal structures and are not used interchangeably in everyday life. For instance, if metaphors, similes and hyperboles share one prominent pragmatic goal, clarify a point, metaphors seem to be preferred to make a statement more interesting and unlike hyperboles, they do not achieve

---

<sup>2</sup> The Augmented Criticism Lab (<http://acriticismlab.org>) is an ongoing project helmed by Michael Ulliot (University of Calgary). Through the automatic recognition of various figures of repetition, it seeks to pinpoint authors' particular habits, to identify plagiarism as well as influences from previous authors.

emphasis and humour (Roberts & Kreuz, 1994). In addition, when looking at traditional literary scholarly works dealing with the use of rhetorical figures by one author or in a collection of texts, the focus is often either on one single figure or on a relatively small cluster of related figures as shown by titles such as Ullman's *The Image in the Modern French Novel* (1960) or Chapin's *Personification in Eighteenth-Century English Poetry* (1974).

In *Rhetorica ad Herennium* (trans. Caplan, 1954), the simile, apart from being classified as a figure of thought, is defined as "the comparison of one figure with another, implying a certain resemblance between them" (Book IV, XLIX. 62, p. 385). Casting light on the term "figure" used in this definition, Puttenham (1589) explains that a simile or what he calls "resemblance by imagery or portrait" occurs not only when a human being is likened to another in countenance, speech, quality or any other quality, but also when any natural thing is likened to another (p. 204).

Based on their syntax, it is possible to distinguish phrasal similes [s1] from clausal similes [s2].

[s1] a. Debts are now-a-days like children, begot with pleasure, but brought forth with pain [Les dettes aujourd'hui , quelque soin qu'on emploie , Sont comme les enfants, que l'on conçoit en joie, Et dont avecque peine on fait l'accouchement.] (Molière, as cited in Wilstach, 1916, p. 86).

b. Her brest fairer than the vernal bloom of valley-lily, op'ning in an show'r (Logan, as cited in Wilstach, 1916, p. 31).

c. Death... was busy as on a battle field (Skelton, as cited in Wilstach, 1916, p. 41).

[s2] a. All at once noise and light burst on me as if a window of memory has been suddenly flung open on a street in the City (Milton, as cited in Wilstach, p. 39).

b. Envy excels in exciting jealousy, as a rat draws the crocodile from its hole [L'envie excelle à exciter la jalousie comme le rat à faire sortir le crocodile] (Hugo, as cited in Wilstach, p. 113)

Grossly speaking, just by looking at their structure, the difference between these two types of similes can be summarised in terms of whether the comparison is made with a phrase or a clause. In addition, different phrases can be part of a comparison: while the comparison in both sentences s1a and s1b is built with a noun phrase, in s1c, comparing is rather done with a prepositional phrase. The scope of this thesis will be restricted to what can be called nominal phrasal similes, i.e. similes that rely on a noun phrase and that compare two entities, and not two processes as it is the case in [s1c], [s2a] and [s2b].

## 1.5 Motivation of the Study

The simile occupies a particular place in the history of rhetoric. First, the simile is one of the oldest figures of speech recognised and from the beginning of rhetoric, it has been inextricably linked to the metaphor as if trapped in its shadow. In addition, rhetoricians have never totally agreed on where to classify it and even about its true status as a figure. Bullinger (1898), for example, writes: “Indeed it can hardly be called a figure, or an unusual form of expression, seeing it is quite literal, and one of the commonest form of expression in use. It is a cold, clear, plain statement as to a resemblance between words and things” (p. 726). Strange fate for a rhetorical figure that Wilstach (1916) describes as “the handmaid of all early records of words [which] has proved itself essential to every form of human utterance” (p. vii). Similarly, Woods raves about the additional dimension similes bring about in fiction: “Every metaphor or simile is a little explosion of fiction within the larger fiction of the novel or story” (as cited in Moon, 2008, p. 153).

It is therefore not surprising that even though as far back as Aristotle’s *Rhetoric* (trans. 1926), similes were judged are being less powerful than metaphors, similes are still to be found not only in everyday language but also in literary texts. Obviously, the explicit use of analogy in similes greatly explain their endurance: they are invaluable for communication as they make new concepts easier to understand as well as succeed in building expressive innovative mental images. In this respect, commenting on the several functions and values of similes or similitudes, Pechaum (1593) writes:

The use of Similitudes is verie great, yelding both profite and pleasure, profit by their perspicuitie, and pleasure by their proportion. They serve to many and sundry endes, as to praise, dispraise, teach, to exhort, move, perswade, and to many other such like effects: of all formes of speech, they are best conceived, most praised, and longest remembered. (paragraph on *Similitudo*)<sup>3</sup>

---

<sup>3</sup> In contemporary English: The use of [similes] is very great, yielding both profit and pleasure, profit by their perspicacity, and pleasure by their proportion. They serve to many and sundry ends, as to praise, dispraise, teach, to exhort, move, persuade, and to many other such like effects: of all forms of speech, they are best conceived, most praised, and longest remembered.

As part of an author's imagery, the role of the simile is dual. On the one hand, the less creative ones, which belong to the common lore either confer more authenticity to fictional characters or depict unimportant details of the text. On the other hand, creative similes can define a particular text, an author or even a literary period. According to Abrams (1999), since Caroline Spurgeon's pioneering study of Shakespeare's image motifs in similes but also in metaphors, it became evident that clustering images by theme could not only unveil the author's personality as well as personal experiences but could also sum up the text's tonality. For instance, while in Shakespeare's *King Lear*, the animal imagery is predominant, in *Hamlet* rather prevail images related to death, disease and corruption. Similarly, by analysing 400 random similes by four different generations of Hebrew poets, Shen (1995) notices structural similarities between poets of the same generation, notwithstanding the poet and the context of production of the poem.

Outside of literary texts, similes are also alive and well. The pervasiveness of similes in everyday language is understandable when taking into account the fact that similes rely on comparing, which is a fundamental human cognitive activity. Since languages of the world can be grouped together according to the dedicated words and structures they use not only to express comparisons but also to create similes, it appears obvious that various languages share the same simile structure. Of course, if all similes are comparative structures, it is worth asking when exactly a comparative structure starts being figurative and deserves to be called a simile. Moreover, if some similes have become hackneyed, are they still considered as figurative? In fact, more than any other figure, despite its apparent simplicity, the simile flirts with the boundaries of figurative language while revealing more about our own perception of the world.

## 1.6 Organisation of the Thesis

This thesis is divided into six chapters. Since similes are derived from comparisons, it seems logical that a study devoted to similes starts with defining comparisons and their main characteristics. Chapter 2, therefore, focuses on the syntax and the semantics of comparisons. It also attempts to circumscribe the notion of simile by exploring various theories which seek to explain how it differs from comparisons, on the one hand, and from metaphors, on the other hand.

The third chapter sketches the main challenges related to simile detection and gives an overview of related work on both comparative construction and simile detection.

The fourth chapter deals with the question of annotation. First, it provides some general principles for linguistic annotations, then it outlines the different criteria used by literary scholars to discuss similes and finally, it describes existing simile annotation schemes.

The fifth chapter presents our approach to automatic detection and annotation, first by presenting a grammar of the simile and secondly by stating the different steps involved and describing the annotation process.

Chapter 6 mainly deals with simile annotation and seeks to confirm some of the hypotheses at the core of our approach to simile detection. It describes in detail first, an experiment on manual annotation, and then, the crowdsourcing platform developed for this project and the data that were collected.

Chapter 7 presents three applications of the proposed method to simile detection on a corpus of French and English novels published between the early 19<sup>th</sup> century and the first half of the 20<sup>th</sup> century.

Finally, in the conclusion, directions for further research are discussed.

*Introduction*

*PART ONE*

*THEORETICAL  
AND  
COMPUTATIONAL  
BACKGROUND*

*Introduction*



## 2 SIMILES, COMPARISONS, METAPHORS AND FIGURATIVENESS

One of the main characteristics of the existing literature on similes in English and in French is the diverse denominations that have been given to this particular figure. Two broad traditions emerge: one which uses a supra-figure to refer to similes and one which specifies the type of similes discussed. Generally speaking, the first group of authors considers similes subtypes of either comparisons or metaphors. If we look at publications in English such as “Understanding metaphorical comparisons” (Glucksberg & Keysar, 1990), “Poetic Comparisons: How Similes Are Understood” (Gargani, 2014), the title immediately clarifies what they are about: they are centred around a particular type of comparisons and intuitively, the reader knows that those comparisons are what is generally referred to as similes. In French, it is less obvious; since French does not have a specific word for similes, it must rely on the term “comparaison” which at times, can be fairly confusing. In a title such as “La structure des comparaisons dans *Madame Bovary*” (Pistorius, 1971), it is only the context of usage that can make one infer that the article is going to talk about similes because it is supposed that those are the most interesting comparative structures to study in a novel. Similarly, calling similes metaphors, could also at times be baffling and is often criticised by purists. In fact, it all goes down to the school of thought to which one adheres. Therefore, to be sure, a reader interested in similes must either peruse such a text in order to see if it includes similes or must look for a sentence which states whether or not similes, in that text, are discussed as types of metaphors. For Genette (1970), this metonymical tendency can be attributed to modern theoreticians who see similes as an elongated form of

metaphors, such as Proust who constantly labels as metaphors structures that are mere similes.

In addition, depending on the researchers, similes have been described as “non-poetic” (Fishelov, 1993), “poetic” (Cohen, 1968; Fishelov, 1993), “figurative” (Shabat Bethlehem, 1993) or “creative” (Veale, 2012; Niculae, 2013). The chosen adjectives, of course, raise some questions: does “poetic” imply that these similes are found in poetry or that they have a certain lyrical value? Are similes found in poetic texts different from those found in novels, plays and in other non-fictional texts? Furthermore, if the simile is a figure of speech, is it not redundant to call it “figurative”? Does it mean that there are also non-figurative similes and in this case, are they still figures of speech? And by the way, what does one mean by figurative? Finally, are creative similes more worthy of interest than other similes?

In order to provide suitable answers to these questions, this chapter will investigate the relationship between similes and comparisons, similes and metaphors, and similes and figurativeness.

## 2.1 Comparison: Semantics and Syntax

The term “comparison” can have several acceptations in the language: it can designate a figure of speech, and in this sense, it describes linguistic unit, but it can also refer to a cerebral act, to the psychic act of sensing dissimilarities between distinct elements (Stutterheim, 1941).

Le Guern (1973) points out how the polysemy of the term “comparison” is problematic for grammarians as it corresponds to two different Latin concepts: *comparatio* and *similitudo*. While the term *comparatio* is used in relation to the act of comparing in general, its counterpart *similitudo*, which has the same etymological root as the English term “simile”, is devoted to resemblance and in some rare cases to analogy (Berteau, 1980).

### 2.1.1 Comparison in Rhetoric

Comparisons in rhetoric oppose two concepts either based on logic or based on the syntagmatic order (Berteau, 1979). If Aristotle (trans. 1926, 1984) does not explicitly define

what a comparison is, he, however, highlights its importance by stating several of its applications in everyday life:

- in a debate, comparing one's ideas to those of the other party could help to prove a point;
- while making a value judgment, comparison helps to decide what is the better good;
- contrasting the similarities and the differences between a group of things enables to discover their distinctive or relative properties so as to classify them based on their shared attributes;
- when using inductive or analogical reasoning, a conclusion about a phenomenon can be inferred by taking into account already known similar situations.

In Latin rhetorical texts, several terms convey the idea of comparison: *comparatio*, *similitudo*, *collatio*, *simile*, *imago*, *exemplum*, with *similitudo* being by far the most used both by Cicero and Quintilian (Tucker, 1998; Norton, 2013). In *Rhetorica ad Herennium* (trans. Caplan, 1954), *similitudo* is defined as “a manner of speech that carries over an element of likeness from one thing to a different thing” (p. 377). Similarly, Cicero (trans. 1856a) sees *similitudo* as the process of stating two things as being opposed or equivalent to one another, such as in: “For as a place without a harbour cannot be safe for ships, so a mind without integrity cannot be trustworthy for a man's friends” (p. 276). Following Aristotle's steps, both Cicero (trans. 1856b) and Quintilian (trans. 1876) see the comparison not only as a proof but also as a source from which new arguments could be derived. It is also worth noting that both rhetoricians differentiate between arguments relying on a comparison, those relying on similarities and those relying on dissimilarities.

Table 2.1 *Comparatio, Similitudo and Dissimilitudo: Definitions and Examples (Peachum, 1593)*

Figures	Definition	Examples
<i>Comparatio</i>	Form of speech which by apt similitude shows you the example brought in, is either like, unlike or contrary: like things are compared among themselves, unlike from the lesser to the greater in amplifying, and from the greater to the lesser in diminishing, and contraries by opposing one against another.	Now as Jams and Jambres withstood Moses, so do these also resist the truth: men of corrupt minds, reprobate concerning the faith. (2 Timothy 3:8)
<i>Similitudo</i>	Form of speech by which the orator compares with the other by a similitude fit to his purpose.	Even as the light of a candle, is opprest with the brightnesse of the Sunne, so the estimation of corporall things must needs be darkened, drowned, and destroyed by the glorie and greatnesse of vertue.
<i>Dissimilitudo</i>	Form of speech which compares diverse things in a diverse quality.	The ox knoweth his owner, and the ass his master's crib: but Israel doth not know, my people doth not consider. (KJV Isaiah 3:8)

Despite the inconsistency of the Latin terminology, Latin rhetoricians appear to treat the *similitudo-argument* and *similitudo-ornament* (often *imago*) as the two faces of the same coin. In this respect, the restrained sense of the concept *similitudo* only becomes prevalent afterwards. Peachum (1593), for example, though heavily influenced by Cicero from whom he borrows various examples, establishes a clear distinction between the *comparatio*, the *similitudo* and the *dissimilitudo* (see Table 2.1). Moreover, he classifies under the label *comparatio*, among others the *antithesis* (“He is gone but yet by a gainful remove, from painful labour to quiet rest, from unquiet desires to happy contentment, from sorrow to joy, and from transitory time to immortality”), the *antimetabole* (“Neither was the man created for the woman; but the woman for the man”) and the *correctio* (“But now, after that ye have known God, or rather are known of God, how turn ye again to the weak and beggarly elements, whereunto ye desire again to be in bondage”). A closer look at these other figures based on comparison shows that the comparison there is rather veiled and implicit, unlike the examples given for the *comparatio*.

**Table 2.2 Examples of similes and comparisons with their respective values; the terms compared are in bold (trans. Caplan, 1954, pp. 383-387)**

	Sentences	Value
Comparison	Unlike what happens in the palaestra, where <b>he who receives the flaming torch</b> is swifter in the relay race than <b>he who hands it on</b> , the <b>new general who receives command of an army</b> is not superior to <b>the general who retires from its command</b> . For in the one case it is an exhausted runner who hands the torch to a fresh athlete, whereas in this it is an experienced commander who hands over the army to an inexperienced.	<i>Embellishment, Contrast</i>
	Neither can <b>an untrained horse</b> , however well-built by nature, be fit for the services desired of a horse, nor can <b>an uncultivated man</b> , however well-endowed by nature, attain to virtue.	<i>Proof</i>
	<b>In maintaining a friendship</b> , as in a <b>foot-race</b> , you must train yourself not only so that you succeed in running as far as is required, but so that, extending yourself by will and sinew, you easily run beyond that point.	<i>Clarity</i>
	Let us imagine a <b>player on the lyre</b> who has presented himself on the stage, magnificently garbed, clothed in a gold-embroidered robe, with purple mantle interlaced in various colours, wearing a golden crown illumined with large gleaming jewels, and holding a lyre covered with golden ornaments and set off with ivory. Further, he has a personal beauty, presence, and stature that impose dignity. If, when by these means he has roused a great expectation in the public, he should in the silence he has created suddenly give utterance to a rasping voice, and this should be accompanied by a repulsive gesture, he is the more forcibly thrust off in derision and scorn, the richer his adornment and the higher the hopes he has raised. In the same way, a <b>man of high station</b> , endowed with great and opulent resources, and abounding in all the gifts of fortune and the emoluments of nature, if he yet lacks virtue and the arts that teach virtue, will so much the more forcibly in derision and scorn be cast from all association with good men, the richer he is in the other advantages, the greater his distinction, and the higher the hopes he has raised.	<i>Vividness</i>
Simile	<b>He</b> entered the combat in body like <b>the strongest bull</b> , in impetuosity like <b>the fiercest lion</b> .	<i>Praise</i>
	<b>That wretch</b> who daily glides through the middle of the Forum like <b>a crested serpent</b> , with curved fangs, poisonous glance, and fierce panting, looking about him on this side and that for someone to blast with venom from his throat – to smear it with his lips, to drive it in with his teeth, to spatter it with his tongue.	<i>Censure</i>
	<b>That creature</b> , who like <b>a snail</b> silently hides and keeps himself in his shell, is carried off, he and his house, to be swallowed whole.	<i>Contempt</i>
	<b>That creature</b> who flaunts his riches, loaded and weighed down with gold, shouts and raves like <b>a Phrygian eunuch-priest of Cybele</b> or like <b>a soothsayer</b> .	<i>Envy</i>

In *Rhetorica ad Herennium* (trans. Caplan, 1954), whereas comparisons are presented in terms of their pragmatic purposes, similes are classified according to the emotions they wish to convey (see Table 2.2), which seems to reinforce the idea that a certain amount of

subjectivity characterises similes. In addition, from a structural point of view, these translated examples of comparisons are expressed with completely different and more diverse structures than the translated similes. This apparent dichotomy between similes and comparisons does not however mean that same construction cannot be applied to both figures as in the part dealing with the hyperbole, the following sentence, which is clearly a simile, is given as an example of a hyperbolic comparison:

[s3] From his mouth flowed speech sweeter than honey. (trans. Caplan, 1954, IV. XXXIII.44, p. 341).

### 2.1.2 Grammatical Expressions of Comparisons

Rather than being inferred by the sentence syntax, the expression of comparison in natural languages is first and foremost semantic. Phrasal comparatives can fulfil various pragmatic purposes and correspond to a whole range of syntactic structures:

- **inequality**: *Les femmes travaillent plus que les hommes.*
- **equality**: *Son livre est aussi drôle qu'un film comique.*
- **prevalence**: *Il vaut mieux un mari alcoolique qu'un mari infidèle.*
- **preference**: *Il a préféré la mort au déshonneur*
- **resolved alternative**: *Un bon croquis, plutôt qu'un long discours !*
- **similarity**: *Il ment comme un arracheur de dents*
- **analogy** : *Elle a filé, telle une flèche.*
- **identity**: *Il a le même pull que son frère.*
- **alterity**: *J'ai d'autres modèles que cette robe. (Fuchs, 2014).*

According to Cohen (1968), the canonical simile form is derived from a comparison of the type “La terre est ronde comme une orange” or “the earth is round like an orange”. Both sentences fall under what is generally called the comparative degree of the adjective. Despite the multiple structures to which the comparison may correspond, the study of the phenomenon of comparison in Indo-European grammars has been mostly focused on the morphological features of the comparative degree of the adjective and on comparative clauses (Bouverot, 1969; Rivera, 1990). The study of the comparative degree of adjectives has also nurtured linguistic research on language typology as well as on language universals.

Typically, in almost all languages of the world, apart from the marker of the comparison, a comparative structure is made up of the two elements that are compared and of the property in relation to which they are compared (Dixon, 2005). In this respect, in English and in French, comparative constructions consist of:

- (1) the “item that is compared”;
- (2) the “standard of comparison” against which (1) is compared;
- (3) the “quantity or quality” which is the property on which the comparison is based;
- (4) the “standard marker” which states the relationship between (2) and (3);
- (5) the “degree marker” which states to which extent (3) is present or absent in (1) in accordance with the amount of (3) in (2) (Ultan, 1972).

When both elements compared are noun phrases as it is in the case in the type of similes discussed in this thesis, Stassen (2013) proposes the terminology comparee NP for (1) and standard NP for (2). In addition, since English and French are both Subject-Verb-Object (SVO) languages, the syntax of their comparative constructions places the standard marker between the adjective and the standard NP (Greenberg, 1963).

The two sentences “Peaches are less sweet than pineapples” and “Mon fils est plus bavard que ma mère” can, therefore, be represented as follows:

Peaches	are	less	sweet	than	pineapples
<u>Mon fils</u>	est	<u>plus</u>	<u>bavard</u>	<u>que</u>	<u>ma mère</u>
<b>comparee</b>		<b>degree</b>	<b>quality /</b>	<b>standard</b>	<b>standard</b>
<b>NP</b>		<b>marker</b>	<b>quantity</b>	<b>marker</b>	<b>NP</b>

Degree comparisons in English and in French can denote two types of relationships;

- equality and in this respect, it makes use of an equative;
- inequality further divided into relationships of superiority and of inferiority.

Table 2.3 presents all the comparatives and equatives used in English and French as well as their usage. In both languages, for comparisons of superiority, the adjective can be inflected and the degree marker omitted. In English, except when the adjective is a compound adjective such as “faithful”, has more than three syllables and in some cases two syllables, the comparative degree of an adjective can be formed by adding the suffix -er (Mason, 1874; Bain, 1879). As far as French is concerned, some adjectives and some adverbs have particular derivational forms: bon → meilleur, bien → mieux, mauvais → pis (Grevisse, 2001). The same can also be said of the comparative of some English adjectives: good → better, bad → worse, far → farther/further.

English and French have been said to have the same comparative construction as the standard NP must always be preceded by a specific comparative particle: “than” in English and “que” in French (Stassen, 2013). It is worth noting that the equative form in English uses “as” instead of “than” and even sometimes does not require any standard marker.

**Table 2.3 Main comparatives in English and in French and examples of their usage**

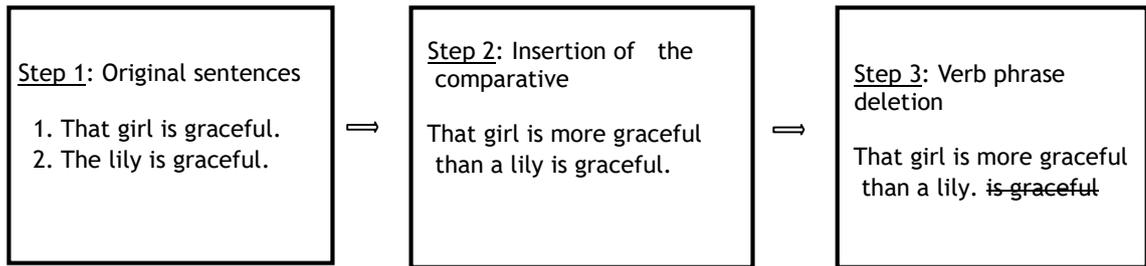
	Equality	Inequality	
		Superiority	Inferiority
<i>English</i>	- (verb, adjective, adverb) + as - (verb, adjective) + like - as + (adjective, adverb, noun phrase) + as	- (verb, adjective) + more + than - more + (noun phrase, prepositional phrase, adjective, adverb) + than - (adjective) + -er + than	- (verb) + less + than - less + (noun phrase, prepositional phrase, adjective, adverb) + than
<i>French</i>	- (verb, adjective) + comme - aussi + (adjective, adverb) + que - (verb) + autant + que - autant + (prepositional phrase) + que	- (verb) + plus + que - plus + (adjective, adverb, noun phrase, prepositional phrase) + que	- (verb) + moins + que - moins + (adjective, adverb, noun phrase, prepositional phrase) + que

At the semantic level, in a typical comparative construction such as [s4] “Jean est plus intelligent que Max”, the standard marker establishes a scale between two degrees of the quality/quantity involved in the comparison (Bouchard, 2008). From sentence [s4], the following propositions can be deduced:

- Max is intelligent to some extent
- Jean is intelligent to some extent
- The extent to which Jean is intelligent surpasses the extent to which Max is intelligent.

With regard to the syntax of the comparative constructions in French, Grevisse (2001) observes that they are elliptical by nature as what has already been said, generally, the quantity or quality at the heart of the comparison is often not repeated. In this respect, as exemplified in Figure 2.1, to transform two main clauses expressing the same quality or quantity into a comparison or a simile, two main operations must take place: first, form a single sentence by inserting a comparison marker between the two clauses, and then, delete the verb phrase after the standard of comparison.

**Figure 2.1 Overview of the construction of the comparative sentence “That girl is more graceful than a lily”**



As far as English phrasal comparatives are concerned, Bresnan (1973) also notices the same, stating that the standard NP is not simply a complement but a fully fledged clause in which one or more constituents of the head of the comparative (the part of the sentence that starts after the comparee NP and ends with the quality/quantity) have been deleted. As illustrations, here are different underlying structures of comparative constructions:

[s4]

- a) “I’ve never seen a taller man than my father” → I’ve never seen a taller man than my father is tall a man.”
- b) \*\*“I’ve never seen a taller man than my mother” → “I’ve never seen a taller man than my mother is tall a man.”
- c) John is older than Mary. → John is older than Mary is old.
- d) John read more books than Mary. → John read more books than Mary read books.
- e) More people bought books than magazines. → More people bought books than people bought magazines.
- f) Peter introduced more people to Jack than John. → Peter introduced more people to Jack than he introduced to John.<sup>4</sup>

By reconstructing the full sentence, it is possible to better understand why the second sentence is not acceptable, as it implied the impossibility of the mother being a man. The syntax and semantics of comparative clauses are therefore closely connected to their underlying structure.

---

<sup>4</sup> The first two examples are taken from Bresnan (1973, pp. 316-318) and the remaining examples from Lechner (2001, pp. 683-84, p. 720).

## 2.2 Comparisons and Similes

With respect to the relationship they infer between the compared objects, Bredin (1998) distinguishes six types of comparisons which can each be transformed into a corresponding simile (see Table 2.4). It is worth noting that, in the proposed classification, similes do not only express similarities, but also dissimilarities, be it through negated similarity statements or through comparisons of inequality.

**Table 2.4 Types of comparisons and corresponding similes (Bredin, 1998, p. 69-73)**

	Comparison	Simile
<i>A is like B</i>	Paul is <u>like</u> Mary.	Huge fragments vaulted <u>like</u> rebounding hail, / Or chaffy grain beneath the thresher's flail.
<i>A is not like B</i>	Paul is not <u>like</u> Peter.	My Mistress' eyes are nothing <u>like</u> the Sun.
<i>A is like B in respect of p</i>	Paul and Peter <u>look</u> alike.	The world is charged with the grandeur of God. /It will flame out, <u>like</u> shining from shook foil.
<i>A is unlike B in respect of p</i>	Paul is a good coo <u>but</u> Jane is a wonderful hostess.	To rust unburnished, not to shine in use! / <u>As though</u> to breathe were life.
<i>A has as much of p as B has</i>	Peter's hair is <u>as</u> black <u>as</u> Jane's.	Between my finger and my thumb / The squat pen rests; snug <u>as</u> a gun.
<i>A has a different quantity of p than B has</i>	Paul is wiser <u>than</u> Jane.	Coral is far <u>more</u> red, <u>than</u> her lips' red.

### 2.2.1 Comparisons of Inequality and Similes

At first glance, from a purely syntactic point of view, nothing differentiates comparisons from similes. However, even though Le Guern (1973) agrees that the French adverb "comme" can be used both in similes and comparisons, he affirms that other comparison markers cannot be used so freely: whereas "plus + adjective + que", "moins + adjective + que", "aussi + adjective + que" always mark a comparison, "semblable à", "pareil à" and "de même que" only introduce a simile. This difference in usage could be explained by the fact that the comparison is quantitative by nature, unlike the simile which is generally qualitative. In this respect, if in "Pierre est fort comme son père", "comme" highlights a mere comparison and means exactly the same as "aussi...que", in "Pierre est fort comme un lion", "comme" denotes a simile and cannot be understood as "Pierre est aussi fort qu'un lion". The rationale behind this distinction is that the first "comme" assesses quantitatively Pierre's and his father's strength, whereas in the second case, Pierre's

strength is described by making reference to the lion, perceived as possessing a great amount of strength.

Similarly, De Mille (1878) distinguishes between three types of comparisons: the “comparison of degree”, the “comparison of analogy” and the “comparison of similarity” and considers only the latter two as similes (p. 106). By comparisons of degree, it is meant all structures that imply equality, superiority or inferiority, which means that all these comparisons are scalable. However, some of the examples given to sustain this interpretation are far from being convincing. As a matter of fact, “He is as brave as a lion” is listed as a non-simile unlike “He is like his father”. When drawing this distinction, De Mille (1878) seems to have been wrongly influenced by grammatical considerations and the fact that in English, “as...as” is used for equality.

The whole debate on the use of degree in similes appears to have its roots in the name of the figure itself. Since, simile comes from the Latin *similis* which means “*like, similar, resembling closely, or in many respects*” (Bullinger, 1898, p. 726), many rhetoricians tend to restrict it to statements of similarity as illustrated by the following definitions of the simile:

“Simile, or Comparison consists in formally likening one thing to another that in its nature is essentially different, but which it resembles in some properties.” (Waddy, 1889, p. 221)

“A comparison, or simile, is a figure of speech in which a likeness is pointed out or asserted between things in other respects unlike.” (Kellog, 1901, p. 125)

“Simile is a comparison of objects based on resemblance [..]” (Raub, 1888, p. 187)

Moreover, the names of the general figure under which similes are generally classified also speak for themselves. As a matter of fact, calling the simile a “figure of similarity” (Bain, 1890), “a figure of similitude” (Waddy, 1889) or a “figure founded on resemblance” (Raub, 1888) suggests that dissimilarities cannot be the foundations of a simile, a point on which Bullinger (1898) insists: “*Simile* differs from Comparison, in that comparison admits of dissimilarities as well as resemblances” (p. 727).

If it can be agreed that some markers are preferred in the language to form comparisons and similes, the nature of the marker is not enough to differentiate between the two figures. In this respect, both Bouverot (1969) and Pistorius (1970) respectively find in Baudelaire’s *Les Fleurs du Mal* (1857) and in Flaubert’s *Madame Bovary* (1857) various instances of similes with comparatives of equality and of inequality:

De la mâle Sapho, l’amante et le poète,  
Plus belle que Vénus par ses mornes pâleurs !  
— L’œil d’azur est vaincu par l’œil noir que tachète

Le cercle ténébreux tracé par les douleurs  
De la mâle Sapho, l'amante et le poète !  
— Plus belle que Vénus se dressant sur le monde  
Et versant les trésors de sa sérénité  
Et le rayonnement de sa jeunesse blonde  
Sur le vieil Océan de sa fille enchanté ;  
Plus belle que Vénus se dressant sur le monde ! (Lesbos, *Les Fleurs du Mal*)<sup>5</sup>

C'est pourquoi je ne suis point délicat comme vous, et il m'est aussi parfaitement égal de découper un chrétien que la première volaille venue (Flaubert as cited in Pistorius, 1971, p. 226).<sup>6</sup>

With regard to the usage of the marker of inequality in similes, Bouverot (1969) notices that they confer to the simile a hyperbolic quality, whereas similes with the equative “aussi... que” has more or less the same meaning and value as “comme”.

### 2.2.2 Cognitive Accounts of Similes and Comparisons

According to Blair (1787), similes are agreeable to the mind because they change our view of the world by forcing us to find similitudes in things not often associated together, they illustrate the comparee NP in a clear and unforgettable way and they enable us to see the

---

<sup>5</sup> English Translation:

Of the male Sappho, lover, queen of singers,  
More beautiful than Venus by her woes.  
The blue eye cannot match the black, where lingers  
The shady circle that her grief bestows  
On the male Sappho, lover, queen of singers —  
Fairer than Venus towering on the world  
And pouring down serenity like water  
In the blond radiance of her tresses curled  
To daze the very Ocean with her daughter,  
Fairer than Venus towering on the world —  
Roy Campbell, *Poems of Baudelaire* (New York: Pantheon Books, 1952)

<sup>6</sup> Translation: [...] that is why I am not squeamish like you, and it is as indifferent to me to carve a Christian as the first fowl that turns up (trans. Marx-Aveling, 1886).

standard of comparison in a new light. For similes to accomplish that, they must follow certain rules:

In the first place, they must not be drawn from things, which have too near and obvious a resemblance to the object with which we compare them. The great pleasure of the act of comparing lies, in discovering likeness among things of different species, where we would not, at the first glance, expect a resemblance. There is little art or ingenuity in pointing out the resemblance of two objects, that are so much a-kin, or lie so near to one another in nature, that everyone sees they must be like. (Blair, 1787, p. 438)

In manuals of rhetoric influenced by this distinction, the two elements of a simile are often said to “differ in kind” (Bain, 1890, p. 138), to be “of different kind” (Waddy, 1889, p. 221) or to be “drawn from one species of things to another” (Jamieson, 1826, p. 152). In contrast, Bredin (1998) writes about the comparison: “Where comparisons are concerned, everything is fair game” (p. 69). Consequently, comparisons know neither rules nor restrictions: a parallel can be made between any two objects.

But, for a comparison to be considered a simile, how different must the compared objects be? From the examples given in manuals of rhetoric, comparisons generally occur in two main scenarios:

- the species or kind can be the object itself, that can be the case when eyes are compared with eyes, a city to another city, a mountain to another mountain, a man to another man.
- the species or kind may refer to an implied category to which belong the objects compared, for example, when one compares Jules Verne to H. G Wells, one is comparing one writer to another. Similarly, if one compares the Antiquity to the Middle Age, one historical period is compared to another.

In this respect, a simile is said to occur if the same lexeme is not used both as the comparee NP and the standard of comparison and if both lexemes do not belong to the same category. With respect to the first condition, though he admits that this form does not bring in any new knowledge, Cohen (1968) lists among possible simile forms, the redundant simile “La neige est belle comme la neige”<sup>7</sup> (p. 48). Two facts must be mentioned to better grasp this example: first, Cohen (1968) is interested in his article in studying anomalous similes and secondly, unlike other examples, this one seems to be invented by the author

---

<sup>7</sup> The snow is as beautiful as the snow.

and is not taken from an actual text. Does it, however, mean that this possibility must be ignored, especially in a literary context where authors have been known to take liberties? Besides, is it not possible that such a repetition in a literary text could be used for certain stylistic effects? If the simple repetition of a lexeme does not seem enough to completely characterise a comparison, what about categories?

### 2.2.2.1 Similes and Categorisation

A category may be defined as “a number of objects which are considered equivalent” (Rosch, Mervis, Wayne, Johnson & Boyes-Braem, 1976, p. 383). Human beings tend to try to make sense of chaos by grouping together elements that they deem similar. Aristotle (trans. Owen, 1853), for example, cites 10 categories into which each single word may fit: “Substance”, “Quantity”, “Quality”, “Relation”, “Where”, “When”, “Position”, “Possession”, “Action” and “Passion” (p. 5). Categorisation is not done haphazardly, but is generally based on specific perceptible or known attributes and most times, it is either intuitive, used in a specialised context or rooted in a culture. If we go back to Aristotle’s classification, for example, a substance would be understood as something liquid or solid that can be eaten or drunk and that is part of the composition of other elements.

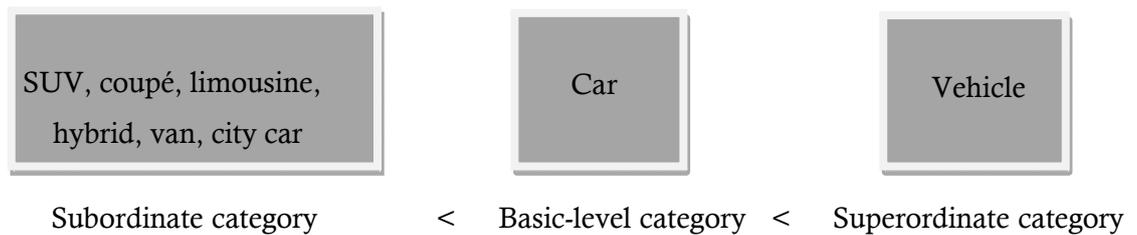
Rosch (1978) distinguishes three levels of natural categories:

- basic-level categories that consist of basic objects<sup>8</sup> such as “car” or “chair”;
- superordinate categories to which basic objects belong, for example, “furniture” for “chair” or “vehicle” for “car”;
- and subordinate categories that are types of basic objects, for example, “rocking chair” and “armchair” for “chair” and “SUV” or “coupé” for “car”.

---

<sup>8</sup> By ‘basic object’, it is meant a group of objects that have a great amount of attributes in common, share similar motor movements, have the same shapes and can be identified by averaging the shapes of the members of the category (Rosch, Mervis, Wayne, Johnson & Boyes-Braem, 1976).

**Figure 2.2 Illustration of the three levels of natural categories (Rosch, 1978)**



It is on this hierarchy that Glucksberg and Keysar (1990) base their account of the difference between similes and comparisons. Their theory postulates that unlike similes, comparisons generally concern entities at the same level of categorisation and which belong to the same superordinate category; they lose all meaning if the marker of comparison is deleted and do not posit the standard of comparison as a prototypical category. Consequently, “Spoons are like forks” would constitute a comparison because spoons and forks are basic objects that have several subordinate categories (dessert spoon, teaspoon, soup spoon, fish fork, snail fork, salad fork...) and belong to the same superordinate category, cutlery. In addition, it would not make any sense to say “spoons are forks” because the category “fork” is not included in the category “spoon”. This class-inclusion property, is however found in similes; for instance, “the girl is like a butterfly” can easily be converted into the metaphor “the girl is a butterfly” without a very significant change in meaning and “butterfly” is easily processed as the embodiment of fluidity, flittiness, transience, lightness. In addition, to find a common category that can be attached both to “girl” and “butterfly”, it is necessary to reach a very high level of abstraction.

In their proposal, Glucksberg and Keysar (1990) limit themselves to examples of the type “a is like b”. As they were mainly interested in statements of similarity and in the processes involved in metaphor comprehension, their focus on that specific structure is perfectly understandable. This choice, however, raises the question of whether their conclusions could be applied to other types of similes. If so, is the presence of an adjective in a sentence like “the girl is as flitty as a butterfly” superfluous when it comes to distinguishing between comparisons and similes and in metaphor comprehension as a whole?

To put things in their context, it is important to say that Glucksberg and Keysar’s theory (1990) results from the desire to correct two related models of similarity: Tversky’s contrast model (1977) and Ortony’s imbalance model (1979). Instead of categories, both models rely on feature matching, the rationale being that a wide range of attributes is intuitively associated with an object.

## 2.2.2.2 Tversky's Contrast Model

Tversky (1977) proposes to measure the similarity  $S$  of two elements  $a$  and  $b$  compared in the sentence “ $a$  is like  $b$ ” by taking into account their similarities and their differences:

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

where  $A \cap B$  corresponds to the set of features that are common to both  $a$  and  $b$ ,  $A - B$ , the features that belong only to  $a$  and  $B - A$ , the features that only belong to  $b$ . If all the features of  $a$  and  $b$  are known, this model enables to determine which features are the most decisive in similarity statements. Imagine we have these two sentences: [s5] “This chair is like an armchair” and [s6] “This chair is like a boulder”. According to Goatly (2011), [s6] would be a simile as  $A - B_2$  does not equal to zero. Table 2.5 lists the salient features of all the elements compared while the similarities and differences between the objects compared are rendered in Figure 2.3.

**Table 2.5 Salient features of “chair”, “armchair” and “boulder” (Goatly, 2011)**

A → chair	B1 → armchair	B2 → boulder
concrete	concrete	concrete
inanimate	inanimate	inanimate
artefact	artefact	-artefact +natural
furniture	furniture	-furniture
for sitting	for sitting	for sitting
for one person	for one person	for one person
support for back	support for back	support for back
	with arms	made of stone
	castors	covered in moss
	upholstered	
	coffee-stained	

**Figure 2.3 Results of Tversky's contrast model for "This chair is like an armchair" and "This chair is like a boulder"**

<p>"This chair is like an armchair"</p> <p><math>A \cap B1 = A</math></p> <p><math>A - B1 = 0 \implies</math> comparison</p> <p><math>B1 - A =</math> (with arms/castors/upholstered/coffee-stained)</p> <p>"This chair is like a boulder"</p> <p><math>A \cap B2 =</math> (concrete, inanimate, for sitting, for one person)</p> <p><math>A - B2 =</math> (artefact, furniture) <math>\implies A - B2 &gt; 0 \implies</math> simile</p> <p><math>B2 - A =</math> (natural, made of stone, covered in moss)</p>
---

From this example, it is also obvious that the more two elements share attributes, the more similar they are and the more they have distinct attributes, the more dissimilar they are. As far as similarity statements are concerned, this model was mainly developed to explain why some comparisons are asymmetrical.

If one says "a is like b", the comparison is said to be directional, in the sense that there a specific element a which is the comparee, another element b which is the standard of comparison and "a is like b" have a different meaning than "b is like a" (Tversky, 1977). If "a is like b" is equivalent to "b is like a", the comparison is said to be symmetrical, otherwise, it is asymmetrical. For Bredin (1998), similes, unlike comparisons are symmetrical and can be reversed because they are only made to assess likeness or unlikeness and do not seek to describe: "Spoons are like forks" means exactly the same as "Forks are like spoons" whereas "The girl is like a butterfly", gives more information about the girl and has a completely different meaning than "A butterfly is like the girl". It has, however, been shown that a change of directionality can affect the sense of a comparative sentence. For instance, "Canada is like the United States" would not be understood exactly as "The United States is like Canada" since Canada possesses most salient features of the United States such as its location and the mixedness of its population, but the reciprocal is not true as the United States cannot be said to be officially bilingual (Glucksberg & Boaz, 1990).

To explain how features are measured in our mind depending on the context, Tversky (1977) introduces the diagnosticity principle, which states that our assessment of features are based either on intensive or on diagnostic factors. Whereas intensive factors are those that are related to vision or audition (loudness, clarity, saturation...), diagnostic factors

enable to eliminate some features in order to retain only those that are the most pertinent for the task at hand. For example, when classifying animals, “real” would be an important feature of classification if and only if the class “animals” includes legendary or fictitious animals, otherwise it would be ignored as it can be applied to all existing animals. In everyday life, people rely on diagnosticity to group similar objects together and to reassess their classification when objects are deleted or added. For example, faced with a watermelon, an orange, a mango, a leech and a lettuce, people would most probably divide those items into two clusters, a fruit cluster and a vegetable cluster. If a lemon and an orange were to be added, because of the features shared by these two items with an orange, it is very likely that the fruit cluster would be further divided into two clusters: a cluster of citrus fruit and a cluster of non-citrus fruit.

From the point of view of simile understanding, however, Gentner (1983) argues that the contrast model cannot really account for non-literal similarity as among all the attributes that differ in the terms compared, only specific ones or none at all intervenes in the construction of the analogy. For instance, in the sentence “An electric battery is like a reservoir”, it is not on the colour, shape or size of each element that the analogy is constructed but rather on the fact that they both store and release energy.

#### 2.2.2.3 Ortony’s Saliency Imbalance Model

Examining the importance of similes in languages, Ortony (1975) observes that they help to achieve three main goals: compactness, vividness and formulating the inexpressible. Compactness refers, here, to the fact that similes make it possible to pack a whole range of implied meanings in a single word. These meanings are filtered in two steps: they are chosen first by saliency and then by tension elimination, so that remain only the most distinctive traits of the standard of comparison that can be transferred to the comparee NP. According to Ortony, Vondruska, Foss and Jones (1985), the term saliency has two acceptations:

- the relevance of an attribute in making a judgement in a particular domain;
- the importance given to an attribute of an object or a category.

Ortony (1979) obviously uses the latter sense when he uses saliency as a distinctive factor between comparisons and similes. According to him, for a statement “a is like b” to be a comparison, A and B must share features that are very high-salient in both elements. Spoons and forks, for instance, are both utensils, that are held with a hand, and are used to eat. In contrast, in a simile, the features that A and B have in common should be high-salient in B, the standard of comparison, and low-salient in A, the comparee NP. For

instance, in “The girl is like a butterfly”, fluidity, flittiness, lightness and transience are more readily associated with butterflies than with girls. Therefore, to take into account feature salience, Ortony (1979) transforms the contrast model into the imbalance model:

$$S(a,b) = \theta f^B(A \cap B) - \alpha f^A(A-B) - \beta f^B(B-A)$$

where  $f^A$  and  $f^B$  correspond to the measures of salience of the set of features of A and B respectively.

Ortony (1978) also specifies that even though the compared elements in a similarity statement do not come from the exact domain, they can nonetheless be grouped together under a higher specific domain. Consequently, “Billboards are like spoons” could not be called a “sensible similarity statement” as “billboards” and “warts” could not be reunited under a single domain or category (p. 36). In contrast, in “Sally is like a block of ice”, both “Sally” and “block of ice” could describe elements that can both exhibit stiffness. In addition, in this last sentence, a transfer occurs between “coldness” referring to the temperature and “coldness” associated with lack of emotional response. In the three sentences given as examples, it is also possible to notice what Ortony (1978) refers to as “domain incongruence”, i.e. the comparee NP and the standard of comparison belong to distinct semantic categories. However, instead of being described as the source of figurativeness, “domain incongruence” is perceived as enhancing figurativeness in a similarity statement.

If Ortony’s theory characterises comparisons and similes, it fails, however, to do the same for the various structures in between that both share low features (“Billboards are like pears”), no features at all (“Chairs are like syllogisms”) or where the common features are high-salient in A and low salient in B (“Sleeping pills are like sermons”). Moreover, since this last type of similarity statement is described as being metaphorical, even though its metaphoricity is very low, why can it not be considered a simile?

According to Weiner (1984), a simile cannot be recognised only in terms of its low- and high-salient attributes, but rather by the fact that the attributes shared by the comparee NP and the standard of comparison are not strictly identical: the comparee NP can never possess those attributes exactly as the standard comparison but only in an approximate way. In this respect, “Blood vessels are like aqueducts” is a literal comparison and not a simile like Ortony (1978) claims because blood vessels and aqueducts function identically as channels. Similarly, Fishelov (1993) unveils some of the limits of Ortony’s theory (1978) when he considers the sentence “Goliath is like the Empire State Building” as a simile

because although “height” is a salient attribute of both “Goliath” and “the Empire State”, the comparee NP is animate whereas the standard of comparison is inanimate.

The different cognitive theories exposed in this section are undoubtedly oriented towards simile understanding and have in common the prominence they give to the standard of comparison, which is invariably described as the element that decides whether a statement is a simile or a comparison. Not only Ortony (1978) but also Glucksberg and Keysar (1990) particularly analyse simile components in terms of the “given-new strategy” (Clark & Haviland, 1977): while the comparee NP is known, the sentence segment containing the quality/quantity and the standard of comparison contains the new information that is conveyed about it. In this respect, they agree with rhetoricians on the pragmatic use of similes.

## 2.3 Figurative similes

So far, the term “simile” has been used to refer to the figure of speech and the term “comparison” to all other syntactic structures in which a parallel is drawn between at least two objects. In some texts, to avoid confusion, what has been called comparisons until now is often called “literal” comparisons and is generally contrasted with similes that are, therefore, figurative.

A word is said to be literal if it retains its usual meaning, the one often first listed in dictionaries. In figurative expressions, however, the original meaning of the word is extended to encompass a new meaning. Some figurative meanings of words are recorded in dictionaries, but are always indicated for people to easily see their difference between them and the primary meaning of the word. How can similes be figurative if as observed by (Lord, 1855) the comparee NP and the standard of comparison are both intended literally?

Shabat Bethlehem (1996) distinguishes 2 types of figurative similes:

- deviant encoded figurative similes in which a relation of resemblance is established between two objects rather highlights how much they are dissimilar. It is generally the case with similes in which the comparee NP and the standard of comparison share low-salient attributes (“Billboards are like spoons”) or when the quantity or quality is not a high-salient feature of the standard of comparison (“La terre est bleue comme une orange”).
- multiply encoded figurative similes in which the relation of resemblance is combined with another figure of speech as in the following examples:

- a) similarity + metonymy: “Mr. McKee was asleep on a chair with his fists clenched in his lap, like a photograph of a man of action” (Fitzgerald, *The Great Gatsby*, as cited in Shabat Bethlehem, 1996, p. 221). Here the character’s occupation is used to build the simile.
- b) similarity + personification/animation: “The wave paused, and then drew out again, sighing like a sleeper whose breath comes and goes unconsciously” (Wolf, *The Waves*, as cited in Shabat Bethlehem, 1996, p. 223). Through the simile, an inanimate object, the wave, becomes animate and is personified.
- c) similarity + synaesthesia: “his words fall cold on my head like paving-stones” (Wolf, *The Waves*, as cited in Shabat Bethlehem, 1996, p. 226). The comparee NP and the standard of comparison are derived from different sense domains, audition and sight in this case. Motion can also play an essential part in this kind of simile.
- d) similarity + word polysemy “His thoughts were as gray as ashes” (Chandler, *The Big Sleep*, as cited in Shabat Bethlehem, 1996, p. 227). “Gray” does not only refer to the colour, but also imply dullness.

A whole range of similes, however, is described as being literal, depending on whether those similes rely only on similarity and on whether they could potentially be lexicalised in the future (Shabat Bethlehem, 1996). In this respect, neither “Her face was as red as a beet” nor “Tanned as an aspirin tablet” are deemed literal because the former merely asserts a resemblance and is more descriptive and the latter is seen as potentially entering the language as an idiomatic expression meaning “very tanned” (p. 218-219).

Similarly, Addison (1993) thinks that similes can have various degrees of literalness and figurativeness. 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. Examples of such dead similes used to intensify a distinctive quality abound: for example, “sleep like a top”, “as blind as a bat”, “crooked as a dog’s hind leg” in English, “sale comme un peigne”, “boire comme un trou”, “pauvre comme une souris d’église” in French. In addition, comparing “a brave man to a lion”, “a cunning man to a fox”, “time to a river”, “eternity to an ocean”, “death to night” and “woman to beauty” are so familiar associations that, when one comes across them, they fail to impress or to be seen as figurative (De Mille, 1878, p. 110). The most logical explanation to justify the change of status of these “stock similes” (Norrick, 1986) is their fossilisation in the language with passing time, so much so that they stop deviating from the norm to become the norm. In this respect, figurativeness in similes is connected to creativity and to pragmatic purposes: “The more distant, indeed, is the

subject from which any illustration is drawn, the more novelty it has, and the more surprise it causes” (Quintilian, trans. 1856, Book VIII, Chap III, 74, p. 104). Does it, however, mean that “stock similes” should be ignored, especially from a literary perspective?

Even though literary style is mainly associated with creative writing and deviations from stereotypes, some literary critics have argued that clichés can be used in literary texts for stylistic effects (Amossy & Herschberg-Perrot, 1997). Riffaterre (1964), for example, states that a cliché can either constitute a feature of the author’s style that reinforces the literary status of the text or can serve to highlight the moral as well as social behaviours of a certain group of people. Norrick (1986) notes that stock similes can often be used as support for humour through irony (“swim like a stone”, “clear as mud”), the introduction of far-fetched standards of comparison (“cold as a witch’s tit in January”) and through pun (“nutty as a fruitcake”). Moreover, literature is known to imbue new senses in cliché statements. As a matter of fact, the connotations of standards of comparison sometimes add another layer to the descriptive function of “stock similes”: when Shakespeare, in *Romeo and Juliet* depicts Tybalt’s corpse as being as “pale as ashes”, he does not refer only to the colour and the lividity of the corpse but also to the fact that all corpses ultimately become ashes (Norrick, 1986). Hence, cliché similes have their importance in literary texts, especially given the fact that creativity is often mere reinvention of what already exists.

A more general theory of figurativeness in similes takes into account its lack of compositionality and the effect of the comparative statement. In semantics, the principle of compositionality refers to the “principle that the meaning of an expression is a function of, and only of, the meanings of its parts together with the method by which those parts are combined” (Pelletier, 1994). A literal comparison respects this principle as saying “Max is more intelligent than James” means nothing else than the fact that the intelligence of Max is superior to that of James. Furthermore, the degree of intelligence of Max cannot be inferred just from this occurrence and it is, therefore, impossible to say whether Max is a genius, which places James as someone far more intelligent than the norm or if, on the contrary, Max’s intelligence is average or below average. In contrast, in “the girl is as graceful as a lily”, “graceful as a lily” gives a precise idea of the level of gracefulness: the girl is described as being particularly graceful with an idea of delicateness. Different effects would be achieved with different standards of comparisons; whereas “The girl is as graceful as a newborn calf” would mean that she is rather awkward on her two feet and clumsy, “The girl is as graceful as a butterfly” still means “very graceful” but also light, a connotation that was absent with “lily” as the standard of comparison.

As far as the effect is concerned, “Max is more intelligent than James” does not evoke anything apart from a scalar difference. Each of the other three sentences, on the contrary, conjures up a distinct mental image, the image of a lily, that of a newborn calf and that of a butterfly. Kellogg (1901) defines images as “expressions in which, departing from our ordinary style, we assert or assume [real or fancied relations between things]” (p. 125). Similes and metaphors are generally classified among elements of an author’s imagery. In fact, in classical rhetorical texts such as *Rhetorica ad Herennium* and Quintilian’s *Institutes of Oratory*, *imago*, the Latin root of image is used to refer to simile while in Greek, Aristotle (1926) designates the simile by the term *eikon*, which means icon or image. Besides, in M. H. Abrams’s *Glossary of Literary Terms*, similes and metaphors are discussed under the heading “Imagery”. As images are only successful if they can resound in the person to whom they are meant, the standard of comparison cannot be something that is too unfamiliar.

If Quintilian (1876) warns against introducing things “obscure” or “unknown” in similes, especially in an oration, he, nevertheless concedes that those types of similes should be left to poets (p. 104). Several rhetoricians, however, do not show the same latitude towards writer’s poetic license. De Mille (1878), for example, criticises Milton’s similes for their use of objects with which ordinary people are not acquainted, claiming that those similes impede their understanding of some of Milton’s best passages and prevent his works from being as popular as those of Shakespeare, Burns, Pope and the like. This vision of similes seems to be against the surprise principle that was mentioned earlier and even against the principle of literature and art in general. If everything in a text is flat-out visible, what pleasure would be derived from reading, especially since various similes echo their author’s own personal experiences? Is it really possible as manuals of rhetoric suggest to be creative while avoiding both trite and obscure similes?

With respect to the different pertinent traits of similes, we propose to redefine the simile as follows: 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.

## 2.4 Metaphor and Similes

It is impossible to discuss similes without mentioning metaphors. The strong link that unites those two figures results, of course, from the fact that both are used for imagery and are figures of speech relying on resemblance or similarity which create a parallel between

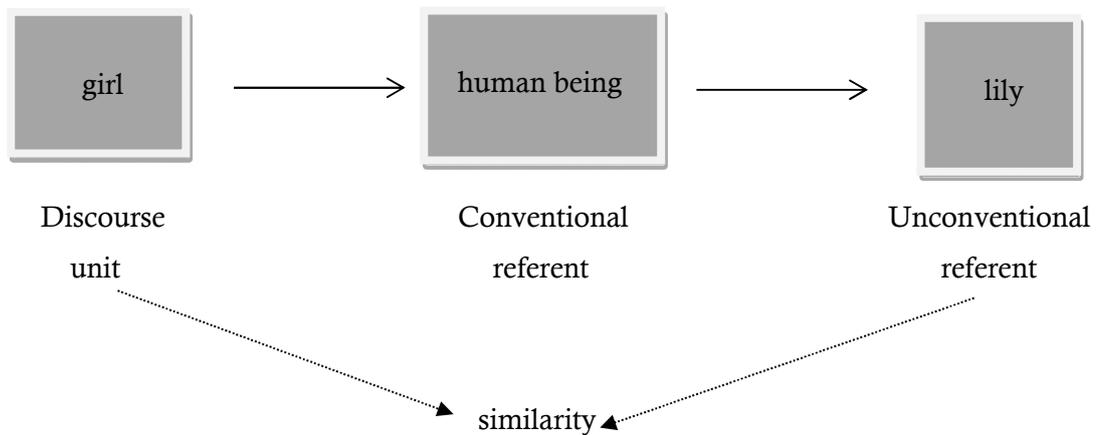
two semantically unrelated objects. Both figures are also an inherent part of everyday language and just like there are dead similes, there are dead metaphors. Besides, most similes, albeit with a certain loss in meaning, can be easily converted into metaphors: thus, Aristotle (trans. 1926) easily transforms the simile “he rushed on like a lion” into the metaphor “a lion, he rushed on” (p. 367).

There exist two predominant views on the relationship between similes and metaphors: one which sees the simile as an “explicit expression of a metaphorical mapping” and one which considers the metaphors as an elliptical simile (Israel, Harding & Tobin, 2004, p. 123). It is interesting to notice that, once again, both traditions can be traced back to early rhetorical texts.

Treating similes as a type of metaphors finds its source in Aristotle’s *The Rhetoric*: “For the simile, as we have said, is a metaphor differing only by the addition of a word, wherefore it is less pleasant because it is longer; it does not say that this is that, so that the mind does not even examine this” (trans. 1926, Book III, chapter X, p. 397). Of course, this passage has given way to various interpretations and certainly explains why similes are often taken as unattractive poor substitutes for metaphors. Genette (1970) observes that the history of rhetoric is characterised by a progressive reduction of the number of tropes, which in the case of figures of resemblance and analogy, has only been advantageous to metaphor (p. 163). The following definition of the metaphor, for example, clearly derives from this school of thought, in the sense that the described characteristics apply both to the simile and to the metaphor:

A metaphor occurs when a unit of discourse is used to refer to an object, concept, process, quality, relationship or world to which it does not conventionally refer, or colligates with a unit(s) with which it does not conventionally colligate; and when this unconventional act of reference or colligation is understood on the basis of similarity or analogy involving at least two of the following: the unit’s conventional referent; the unit’s actual unconventional referent; the actual referent(s) of the unit’s actual colligate(s); the conventional referent of the unit’s conventional colligate(s). (Goatly, 2011, p. 109).

A possible schematisation of the simile “That girl is as graceful as a lily” with respect to this definition would be:



Leech (1969), in his analysis of similes and metaphors in poetry, highlights the fact that the metaphor is far less limited than the simile because it concentrates countless possible meanings in a small space. Undoubtedly, the metaphor’s expressiveness and flexibility could explain why it is sometimes considered as “*la figure, le trope des tropes*” (Sojcher, 1969, p. 68),<sup>9</sup> while the simile must content itself with playing second fiddle to its sister figure. Even Quintilian (trans. 1876), who departs from the Aristotelian view on similes, is full of praise concerning metaphors:

Metaphor is not so only so natural to us, that the illiterate and others often use it unconsciously, but it also so pleasing and ornamental, that, in any composition, however brilliant, it will always make it apparent by its own lustre. If it be but rightly managed, it can never be either vulgar, or man, or disagreeable. It increases the copiousness of language, by allowing it to borrow what it does not naturally possess; and, what is its greatest achievement, it prevents an appellation from being wanting for anything whatever.

(...) On the whole, the metaphor is a short comparison; differing from the comparison in this respect, that, in the one, an object is compared with the thing which we wish to illustrate; in the other, the object is put instead of the thing itself. It is a comparison, when I say that a man has done something *like a lion*; it is a metaphor, when I say that he *is a lion*. (Book VIII, Chap VI 4, p. 125-126).

Whether because of the strong prevalence of the Aristotelian view or because both figures are based on a comparison, the terminology developed, in literary criticism, to discuss

<sup>9</sup> *The figure, the trope of tropes.*

metaphors has also been used in relation to similes. Following the terminology introduced by Richards (1936), three main elements are essential to analyse both similes and metaphors: the tenor, the vehicle and the ground also called tertium comparationis. While the tenor corresponds to what has so far been called the comparee NP, the vehicle refers to the standard of comparison. By ground, it is meant “the basis on which the comparison is made” (Strachan & Terry, 2000, p. 124). Thus, in sentences such as “That girl is more graceful than a lily”, the ground would be the adjective “graceful”, while in sentences such as “he rushed on like a lion”, the ground would be “rushed on”. In some cases, a whole clause or a noun phrase can serve as ground as illustrated in Table 2.6.

**Table 2.6 Anatomy of four similes**

	<i>TENOR</i>	<i>GROUND</i>	<i>MARKER</i>	<i>VEHICLE</i>
<b>That girl is more graceful than a lily.</b>	That girl	graceful	more...than	a lily
<b>He rushed on like a lion.</b>	He	rushed on	like	a lion
<b>Contempt is like the hot iron that brands criminals: its imprint is almost always indelible (Alibert, as cited in Wilstach, 1916, p. 67).</b>	Contempt	its imprint is almost always indelible	like	the hot iron that brands criminals
<b>With the grace of an antelope, the ballerina leapt.</b>	the ballerina	Grace	with the ... of	antelope

Unlike metaphors, similes are easily recognisable at their specific grammatical structures. In this respect, Israel et al. (2004) note that not only are similes less grammatically flexible than metaphors but the metaphor is first and foremost a figure of thought that can affect indistinctly nouns verbs, adjectives and prepositions. Furthermore, while discussing the differences between similes and metaphors, Leech (1969) points out that not only do similes state whether the terms compared are similar or not, but in most cases, they can make the ground of the comparison rather explicit. For example, in a metaphor such as “That girl is a lily”, it can only be supposed in which respect the girl is likened to a lily. On the contrary, in the simile “That girl is as graceful as a lily”, it is clearly stated on which aspect the girl in question resembles a lily. In addition, in a metaphor, the tenor can often be omitted to form a metaphor in absentia, as it the case when someone refers to another person as “That blood-loving hyena”. In this respect, Genette (1970) stresses the fact that the difference between these two figures could not be restricted only to the absence or presence of the tenor since a structure like “girl graceful lily” can neither be called a simile

nor a metaphor. Consequently, what distinguishes these two figures is the presence or the absence of not only the tenor, but also, of the vehicle, the ground and the comparison marker.

Similes belong to a long and ongoing rhetorical tradition. On the one side, similes are subtypes of comparison and on the other side, they are related to metaphors and images. Different theories have tried to explain how comparisons differ from similes. Most of them agree on the fact that similes are figurative while comparisons are literal and that in a simile, a certain semantic distance must exist between the tenor and the vehicle. Figurativeness in similes is, however, often biased as it only takes into account creative similes. To remedy this shortcoming, for the purpose of this study, figurativeness in similes has been redefined in terms of lack of compositionality and of the creation of a mental image.



# 3 COMPUTATIONAL APPROACHES TO SIMILE DETECTION

Even in computational linguistics, which is particularly interested in figurative language, few research works have focused exclusively on the automatic detection of similes in unrestricted texts. Though a large part of this disregard can be attributed to the peculiar structure of similes – no words exhibit a shift in its meaning –, another main reason is the complex structure of comparative statements. Friedman (1989) sums up in few words both the potential of the automatic treatment of comparative structures and the difficulties such a task raises:

An interest in the comparative is not surprising because it occurs regularly in language, and yet is a very difficult structure to process by computer. Because it can occur in a variety of forms pervasively throughout the grammar, its incorporation into a NL [natural language] system is a major undertaking which can easily render the system unwieldy. (p. 161)

In this respect, the first part of this chapter focuses on the challenges inherent to the automatic recognition of similes and comparative statements. The second part gives an overview of the various research works that have been done in the automatic detection of comparative statements. Finally, the third part presents various computational methods proposed to detect or disambiguate similes.

## 3.1 Challenges of Computational Detection of Similes

### 3.1.1 Markers' Polysemy

It is a well-known fact that words can have various meanings depending on their context of usage. This intrinsic polysemy, which makes languages interesting and worth studying, is an aspect that most natural language processing tasks must take into consideration. In the case of comparative statements, for example, the presence of the comparative marker in a sentence is not enough to determine whether that sentence is a comparison or not.

#### 3.1.1.1 The Polysemy of “comme”

Concerning “comme”, the prototypical simile marker in French (Cohen, 1968), Fuchs and Le Goffic (2005), notes that like most grammatical markers, it can fulfil several morpho-syntactical functions and consequently is semantically polysemous. Moline and Flaux (2008) note that “comme” can appear in a wide range of contexts depending on the syntactic elements taken into consideration (see Figure 3.1).

From the semantic perspective, Fuchs and Le Goffic (2005), apart from the role “comme” plays in introducing comparison, attribute the following values to it:

- *coordination*: L’homme comme la femme sont des êtres pensants = L’homme et la femme sont des êtres pensants.
- *temporal simultaneity*: Comme j’allais partir, j’entendis un grand bruit = When he was leaving, I heard a big noise.
- *causality*: Comme il avait froid, il mit un pull. = Because it was cold, I put a sweater on.
- *correlation between a statement and a subordinate clause containing a speech or attitude verb*: Comme tu l’imagines, je fus choqué.
- *identity*: Les femmes considèrent le prince comme un bon parti.  
Comme ami, tu as encore beaucoup à apprendre.
- *exclamation*: Comme le monde est joli!  
Comme tu aimes manger!

Figure 3.1 Syntactic versatility of “comme”

1/ Type of clauses	
<i>Main clause</i>	Il vit <u>comme</u> une ombre. <u>Comme</u> le monde est joli !
<i>Subordinate clause</i>	<u>Comme</u> il avait froid, il mit un pull. <u>Comme</u> tu l’imagines, je fus choqué. Il parle <u>comme</u> il mange, très vite.
2/ Position in the sentence	
<i>Inside</i>	Il vit <u>comme</u> une ombre. Il parle <u>comme</u> il mange, très vite.
<i>Detached</i>	<u>Comme</u> il avait froid, il mit un pull. <u>Comme</u> tu l’imagines, je fus choqué. <u>Comme</u> son père, il adore la mer.
3/ Governing elements	
<i>Verb</i>	Il vit <u>comme</u> une ombre.
<i>Noun</i>	Un <u>gentilhomme</u> <u>comme</u> Don Diego mérite mieux.
<i>Adjective</i>	<b>Fort</b> <u>comme</u> tu es, gagner ce combat ne sera pas difficile.
<i>Main clause</i>	<u>Comme</u> il avait froid, il mit un pull. <u>Comme</u> tu l’imagines, je fus choqué.
4/ Introduced structures	
<i>Elliptical clauses</i>	Il saute <u>comme</u> un gorille. Ton ami est sourd <u>comme</u> un pot.
<i>Clauses expressing an actual fact</i>	<u>Comme</u> il avait froid, il mit un pull. <u>Comme</u> tu l’imagines, je fus choqué.

### 3.1.1.2 The Polysemy of “like” and “as”

Prototypical simile markers in English, “like” and “as”, can also have other pragmatic meanings than comparison. Of course, “like” can be an inflected form of the verb “to like”. Besides, as a preposition or conjunction, it can introduce:

- a quotation: and then, and then Kevin came up to me and said erm ... if you if you go and see Mark this afternoon erm he would like to speak to you, I was like, he should come and speak to me.
- an approximation: My lowest ever [score] was like forty.
- an exemplification: I know but it wouldn't be any point if someone wanted to be, like, a doctor and they got into a nursery place.
- hesitation: Alright. Erm, well like, I usually take the train about... twenty past.
- a metaphor: She's like tearing the wall down.
- a hyperbole: We can like endlessly swear on it. (Andersen, as cited in Walaszewksa, 2013, p. 329-330).

According to the *Oxford Advanced Dictionary Learner's Dictionary of Current English* (Hornby, 2000, p. 54), the morpheme “as” can be used as:

- a preposition signalling what somebody or something appears to be (e.g. They were all dressed as clowns. The bomb was disguised as a package), somebody's job or role (I respect him as a doctor. Treat me as a friend) or something's function (The news came as a shock);
- an adverb to signify a similarity in a situation (As always, he said little.);
- a conjunction that marks temporal simultaneity (As she grew older, she gained in confidence), causality (As you were out, I left a message), conformity in manner (I did as he asked), a comment or an additional information (As you know, Julia is leaving soon) and contrast (“as” means “though”: Happy as they were, there was something missing.)

Deléchelle (1995) also notes that “as” can be used as a relative pronoun, for example in sentences such as “He was very rude, as was his wife” which can be rephrased into “He was very rude, which his wife was too” (p. 194).

All in all, looking at all the different uses of these markers, it is possible to notice that “comme”, on the one hand, and “like” and “as”, on the other hand, not only have similar function in both English and French but they share almost identical pragmatic values.

### 3.1.2 Comparison and Ellipsis

It is well established that comparative statements are elliptical (Desmets, 2008). Thomas (1979) defines the ellipsis as “the absence of elements from the overt form of sentences”, giving examples such as “I wouldn’t if I were you” which is implicitly understood as “I wouldn’t do that if I were you” and can be more specific depending on the context (p. 43). Shopen (1973) distinguishes two main types of ellipses: “functional ellipsis” as in “Kathy’s shop” when a predicate is omitted and “constituent ellipsis” such as in “The duke accepted”, which occurs when one or all the arguments of a predicate are missing (p. 65). For Tamba-Mecz (1983), an elliptical utterance is an abridged form, is semantically equivalent to its reconstructed form and can be rephrased into one and only one monosemous and unambiguous form. Quirk, Greenbaum, Leech and Svartvik (1985) list five criteria that must be satisfied to say that there is an ellipsis:

- the omitted words can be easily and exactly recovered;
- the presupposed elliptical phrase is grammatically incorrect;
- adding back the deleted words produces a sentence that is not only grammatical but that also has the same meaning as the original;
- the omitted words can be found word for word in the original;
- the omitted words have the same morphological form or a slightly modified form than in the original. (pp. 884-887)

How do comparative statements exemplify all these criteria? In the sentence “That girl is graceful as a butterfly” reconstructed as “That girl is graceful as a butterfly is beautiful”, the missing words “is beautiful” are already present in the abridged form, so they can easily be taken from the main clause and added without further modification. If “That girl is as graceful” constitutes a main clause, consequently the rest of the sentence is a clause, which, however, does not comply with the normal clause structure: subject + verb + object. In this respect, it can be considered as grammatically deficient. Finally, the reconstructed sentence “That girl is as graceful as a butterfly is graceful” has exactly the same meaning as the original one.

Reconstructing the full comparative clause, therefore, appears an essential step in comparative statement understanding because of the markers’ flexibility and their incidence on the meaning of the comparison. Syntactically speaking, the canonical simile can be reduced to the form: Noun phrase1 + Verb phrase + marker + Noun phrase2. But, even if all sentences in which one of these markers introduces a subordinate clause were to be deleted, the remaining sentences would not automatically qualify to be pegged as

comparisons. A good illustration would be a sentence such as “He hates beer like milk” which is closer in meaning to “He hates beer and milk” as compared to “He hates beer like he hates milk”. This type of structure, which has the structure of a simile, but semantically does not involve a comparison, is what will be referred to as a pseudo-comparison. In this respect, whereas “That girl is as graceful as a butterfly” is the equivalent of “That girl is as graceful as a butterfly is”, “The bomb was disguised as a package” cannot be said to mean “The bomb was disguised as a package was/would be disguised”. Consequently, since the second example has failed the reconstruction test, it is not a comparison, but as a pseudo-comparison.

Similarly, converting “I respect him as a doctor” into “I respect him as a doctor would respect him” would be going against the original intended meaning since the whole phrase starting with the linguistic unit “as” describes the direct object “him” and not the subject “I”. Also called “subjective-transitive construction”, this particular structure generally occurs with verbs of sensory/cognitive perception (consider, think, find...), calling (label, call, declare), volition (wish, want...) and preference (like, prefer) (Tobback & Defrancq, 2008). Of course, interpretation sometimes would be tricky as this sentence could be likened to “I respect him as I would respect a doctor”. Deléchelle (2004) argues that the semantics of the verb of the main clause strongly affects the meaning that “as” will have. In “I find myself being stared at as a wild or wilful eccentric”, “as” can be considered as introducing a property, a comparison or causality. However, if “stare” were to be replaced by ‘describe’, the implication of causality would disappear. This sentence can be assimilated to a comparison because it can be understood as “I find myself being stared at as a wild or wilful eccentric would be stared at or described”. Moreover, verbs such as “consider” mainly introduce the idea of manner rather than comparison as “I respect him as a doctor” can be conceived as an answer to the question “How do you respect him?” (Deléchelle, 2004).

As far as the canonical simile is concerned, Ultan (1972) observes that in most cases, the subject of the clause is the comparee NP or tenor, the adjective is the quantity/quality or ground and the complement introduced by the marker is the standard of comparison. This form is to be contrasted with a comparison with two full clauses. In “The girl breathes like a dog pants”, beyond the subjects of the two clauses, a parallel is first and foremost made between two processes, the girl’s breathing and the dog’s panting. Comparisons of manner between processes are, however, not restricted to fully fledged comparative clause, but can also be done with nominal ones.

Of particular interest are nominal comparative clauses that consist of a single noun phrase or of two consecutive noun phrases. Let's take a look at the following two sentences: "He threw his bride like a sack of potatoes" and "He threw the coin like a shot putter". Syntactically, these two sentences are equivalent as they have the exact same structure: subject + verb + direct object + marker + complement. But, the reconstruction phase shows that they do not share similar underlying structures:

[s7] He threw his bride like a sack of potatoes. → He threw his bride like he would have thrown a sack of potatoes.

[s8] He threw the coin like a shot putter. → He threw the coin like a shot putter would have thrown the coin.

Whereas in the first sentence [s7], the standard of comparison replaces the direct object of the main clause, in the second sentence [s8], it takes the place of its subject. Far from being trivial, this difference has an impact on the semantics of both sentences. As a matter of fact, whereas [s7] compares processes (the way he threw his bride vs the way he would have thrown a sack), [s8] compares two entities (the man and a shot putter). Thus, though both sentences are similes, they lead to different interpretations.

According to Lechner (2001), comparative statements can exhibit parataxis, i.e. clauses, words or phrases following each other without any punctuation or coordinating conjunction in sentences such as "Gary buys books more than Betty food." (p. 687). If we consider the sentence "He threw his bride like a teenager his dirty clothes", the reconstructed version would be "He threw his bride like a teenager would have thrown his dirty clothes". Once again, it is not simply the man that is compared to a schoolboy, but their manner of throwing, for the man, his bride, and for the teenager, his clothes.

Hence, because of the polysemy of the markers and the ambiguous nature of comparative statements, automatic simile detection cannot merely stop at the presence of a specific marker in a sentence or to the surface structure subject + verb phrase + marker + complement but must dig deeper, especially when the main phrase contains a direct object or when the complement is directly followed by a noun phrase.

## 3.2 Computational Approaches

Though similes are comparative statements, it is worth noting that comparative sentence detection and simile detection are completely independent research domains. In addition, the great majority of work in both research domains have been mainly conducted on English texts.

### 3.2.1 Automatic Detection of Comparatives

#### 3.2.1.1 Comparative Mining from a Semantic Perspective

Since comparative statements have been widely discussed by grammarians, it is not surprising that grammar plays a crucial role in the early computational approaches to comparative statements. Most of these proposed grammars, however, are generally oriented towards semantics and mainly geared towards comparative statement understanding. As far as such accounts are concerned, comparative detection is not meant as a separate task, but as a part of a whole system that works in combination with other language processing tools. Ballard (1988), for example, handles comparatives with “less than”, “more than”, “as long as”, “as many as” inside TELI, a question-answering system: the method he proposes uses rules and conceptual knowledge to simplify and rewrite the output of a sentence parse tree in order to obtain a logical expression that can be easily read by a computer (see Figure 3.2).

Like Ballard’s methods, most early works on comparatives in computational linguistics involve two main phases: the production of an intermediary representation of the comparative sentence and the transformation of this representation into a logical expressing using interpretation or writing rules (Staab, 1998). However, apart from Staab and Hahn (1997a, 1997b), the proposed model of semantic interpretation is not evaluated. These works also underline the strong connection between the syntax of a comparative statement and its semantics. As a matter of fact, several of these early research endeavours rely on linguistic theoretical descriptions of comparative constructions.

**Figure 3.2 Examples of semantic interpretations of comparative sentences**

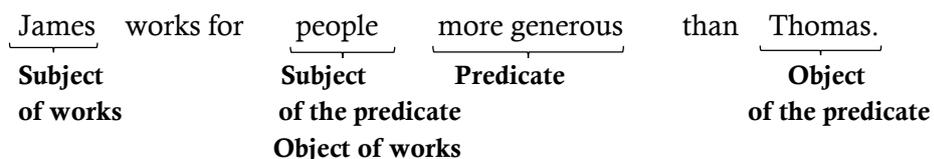
<p><i>Sentence:</i> "List the cars at least 20 inches more than twice as long as the Century is wide"</p> <p>Two representations of "at least 20 inches more than twice as long as the Century is wide" (Ballard, 1988)</p> <p><u>Normalized Parse Tree</u></p> <pre>(CAR (NOUN CAR)       (COMPAR (ADJ LONG)                (λ (P A) (≥ P (+ 20 (x 2 A) ) )                   (CAR (= CENTURY))                   (ADJ WIDE)))</pre> <p><u>Algebraic-Logical Form</u></p> <pre>(SET (CAR P1         (≥ (Length-of-Car P1)             (+20 (x2 (Width-of-Car CENTURY))))))</pre>	<p><i>Sentence:</i> John needs a bigger spanner than the No. 4.</p> <p>Representation of the whole sentence (Rayner &amp; Banks, 1988)</p> <p>needs (John,  <math>\lambda x: \text{spanner}(x)^\wedge</math>  <math>\exists y: \text{big}(x,y)^\wedge</math>      the (<math>\lambda z: \text{type\_of}(x, \text{No.4}),</math>  <math>\lambda z: y': \exists \text{big}(z,y') (y &gt; y')</math>)</p>
---	--

### 3.2.1.2 Borrowing from Grammar

Writing more from a computational perspective, Ryan (1981) relies on corepresentational grammar to define several principles for the analysis of comparative statements based on their surface syntactic structure.

Even though, these rules only take as examples sentences of the form subject + verb + object + (more) + quality/quantity + marker + standard of comparison ("Alice builds planes faster than robots") and subject + verb + more + quality/quantity + marker + standard of comparison + verb (+ object) ("John knows more doctors than lawyers debate"), they can also be extended to different markers and to other types of comparative statements. In order to be able to parse comparative constructions and to identify their different elements, Ryan (1981) elaborates a series of principles, the most important being:

- more, more + adjective and adjective+ -er function as predicates.
- the law of correspondence: the function (subject, object...) of every noun phrase or term of a sentence should be determined.



- the law of uniqueness: two terms can only share the same relation to the predicate if they are coordinate or coreferential.



The second sentence gives way to two possible explanations: either it means “John met people who are taller than Bob” and is analysed as the previous sentence or it rather means “John met taller people than Bob did” and in this case, Bob and John are both subjects of “met”, and are therefore coordinate.

John met taller people than Bob.

(1) John met taller people than Bob.

**Subject of met**      **Subject of taller**   **Object of taller**  
**Object of met**      **Object of met**

(2) John met taller people than Bob.

**Object of met and subject of taller**

⇒ John and Bob are coordinate and subjects of met

Similarly, in “John builds planes faster than robots”, “robots” can be seen as being coordinate with “John” and in this case, it is the subject of “builds” or it is coordinate with “planes” and is, consequently, the object of “builds”.

(1) John builds planes faster than robots

{“planes”: object of builds and subject of faster, “John”: subject of builds, “robots”: object of faster} → John builds planes that are faster than robots.

(2) John builds planes faster than robots

{“planes”: object of builds and subject of faster, “John”: subject of builds, “robots”: subject of builds} → John builds planes faster than he builds robots.

One of the main focal points of this grammar that can also be found in subsequent detection methods of English comparative sentences is the fact that the starting point of the analysis is the marker. As a matter of fact, regardless of the type of texts they have been applied to, computational approaches to the comparatives select potential comparative sentences by looking first for the presence of a comparison-inducing word. Then, the identified sentences are filtered to delete non-comparative sentences. Two main techniques have been used for this effect: pattern matching and supervised learning.

### 3.2.1.3 Pattern Matching

With the aim of improving machine translation, Masui, Tsunashima, Sugio, Tazoe and Shiino (1996) devise a four-step approach towards the disambiguation of comparative sentences containing the markers “more...than” and “as...as”. The initial phase relies on a corpus and consists in listing the different patterns corresponding to these comparative

structures as well as the transformation needed for them to be accurately translated. For example, “as...as” occurs in the following configurations:

- “as...as” → equality comparison: *phrase as ADJECTIVE as phrase / clause as ADJECTIVE PHRASE / ADVERBIAL PHRASE / NOUN PHRASE as clause.*
- “times as...as” → multiple comparison: *clause times as much / ADJECTIVE / NOUN PHRASE as clause / phrase times as ADJECTIVE as phrase.*
- “as...as any” → comparative emphasis: *phrase or clause as ADJECTIVE PHRASE as any.*
- “as well as” → prepositional/conjunctive phrase: *clause as well as phrase.*
- “as...as possible” → adjectival / adverbial / noun phrase: *as soon / many / quickly / early / long / much / ADJECTIVE PHRASE / ADVERB / NOUN PHRASE as possible.*
- “go as...as to + verb” → *go as far as to VERB PHRASE.*

Then, all sentences are matched with the previously identified models. Apart from improving the automatic translation of sentences containing these types of structures, the results obtained suggest that this method particularly works well with equality comparisons and comparative emphases and can be extended to other types of English sentences.

Similarly, Fiszman, Demner-Fushman, Lang, Goetz and Rindfleisch (2007) compile, for biomedical texts, a list of possible patterns that convey comparison of products. But, instead of applying them directly to raw texts, they choose to match them on a partially parsed representation of the sentence. Unlike the previous method, their pattern includes the compared terms: *compare Term1 and/versus Term2*, which makes it also possible to identify the comparee NP and the standard of comparison by looking either at the right or at the left of the trigger for known names of drugs or chemicals. Overall, on the identification of the drugs task and on the comparison recognition task, this method has a precision of 70% and a recall of 96%. It is important to notice that scalar comparisons (“as...as”, “more...than”, “superior to”, “inferior to”) have a lower recall than comparisons built with an inflected form of the verb “compare” but the precision for each type of comparisons is well above 90%. Most of the errors, however, are mostly domain-related:

- empty head noun phrases for name drugs containing dosage or release information

### Example

1/ Oxybutynin 15 mg was more effective than propiverine 20 mg in reducing symptomatic and asymptomatic IDCs in ambulatory patients. ==> 15 mg identified as the comparee NP and 20 mg as the standard of comparison

2/ Intravesical atropine was as effective as oxybutynin immediate release for increasing bladder capacity and it was probably better with less antimuscarinic side effects.

- word sense ambiguity

### Example

Retapamulin ointment 1% (bid) for 5 days was as effective as oral cephalexin (bid) for 10 days in treatment of patients with SID, and was well tolerated” ==> bid is matched here to the BID protein.

## 3.2.1.4 The Machine Learning Era

### 3.2.1.4.1 *The Jindal and Liu's Approach*

Jindal and Liu (2006a) also use patterns to identify comparative sentences of the type “*Car X is much better than car Y*” in text documents. As they are particularly interested in opinions expressed with comparative sentences, they manually compile a list of 83 triggers that includes “beat”, “exceed”, “outperform”, “number one”, “set against”, “but”, “whereas”, “on the other hand”, “favour”, “prefer”, “win”, and of course “more than”, “less than”, “as...as”. Just with this list of markers, they report that they could identify 94% of the comparative sentences in their data set, the precision, however, was far lower, 32%, which means that a lot of sentences that are captured are not really comparative sentences. To solve this issue, Jindal and Liu (2006a) investigate manual rules, sequential rules based on part-of-speech tags and machine learning techniques. To generate their sequential rules, they consider the part-of-speech tag of each three words before and after each trigger. Then, the generated sequence is labelled as either comparative or non-comparative and stored in a database. In the last step, class sequential rules with a minimum confidence threshold are derived from the dataset. However, class sequential rules alone prove to not be sufficient enough to accurately recognise comparative sentences because a single sentence can meet several conflicting rules. Machine learning classifiers such as Naive Bayes were, therefore, used to tackle this problem and combined with class sequential and manual rules, they substantially outperform all other methods with an average precision of 77.3%, a recall of 81% and an F-Score of 79% on manually labelled sentences of three types of texts: review, articles and forums. Tested on other languages such as Korean (Yang & Ko, 2009), this method has also significantly improved the precision initially obtained with triggers alone.

In a subsequent work, Jindal and Liu (2006b) mine comparative sentences in order to extract compared objects and features. More specifically, they seek to identify the relation underlying each comparative sentence so as to render it by the expression (<relationWord>, <features>, <entityS1>, <entityS2>). For instance, a sentence such as “Canon’s optics is better than those of Sony and Nikon” would be summarised as (better, {optics}, {Canon}, {Sony}). It is important to note that some constraints have been laid out: only one relation per sentence is possible, entities and features can only be nouns (proper or common) and pronouns, leaving out cases such as “Intel costs more” in which the feature is a verb.

To train a classifier to recognise each element to generate such an expression, label sequential rules have been tested and have proved to be more efficient than Conditional Random Fields (CRF). In order to generate those rules, each object and feature of all sentences were manually assigned a specific label. Then, for each label identified in the sentence, a sequence, which takes into account the four words at its right and at its left as well as tags indicating the beginning and/or the end of the sentence, is stored in the database. For the creation of class sequential rules, only the most frequent sequences containing at least one label are kept but the part of speech of that label is removed. In order to take into account the variety of grammatical functions an entity can take, all possible parts of speech are associated with the label.

**Example** (see Appendix 1A for details about the part-of-speech tags)

Canon has better optics than Nikon.

Canon\_NNP has\_VBZ better\_JJR optics\_NNS than\_IN Nikon\_NNP

==> 3 sequences corresponding to the compared entities (Canon and Nikon) and to the feature compared (optics)

{#start} {l1} {\$ES1, NNP} {r1} {has, VBZ} {r2} {better, JJR} {r3} {\$FT, NNS} {r4} {thanIN}>

{#start} {l4} {\$ES1, NNP} {l3} {has, VBZ} {l2} {better, JJR} {l1} {\$FT, NNS} {r1} {thanIN} {r2} {entityS2, NNP} {r3} {#end}

{has,VBZ} {l4} {better, JJR} {l3} {\$FT, NNS} {l2} {thanIN} {l1} {\$ES2, NNP} {r1} {#end}

The sequence {\$ES1, NNP} {r1} {has, VBZ} {r2} gives {\$ES1} {VBZ} and generates the following rules:

<{\* , NN} {VBZ}> → <{\$ES1, NN} {VBZ}>

<{\* , NNP} {VBZ}> → <{\$ES1, NNP} {VBZ}>

<{\* , NNS} {VBZ}> → <{\$ES1, NNS} {VBZ}>

<{\* , PRP} {VBZ}> → <{\$ES1, PRP} {VBZ}>

Finally, for each sentence, the rule with the highest confidence score is applied, the matched elements are replaced one after the other for each remaining rule until no rule with a high confidence is left.

**Example** (see Appendix 1A for information about the part-of-speech tags)

Suppose we have the sentence “Coke is preferred because the taste is better than Pepsi” and these tree rules:

R1:  $\{*, \text{NNP}\} \{ \text{VBZ} \} \rightarrow \langle \{ \$ES1, \text{NNP} \} \{ \text{VBZ} \}, \text{confidence: } 80\%$

R2:  $\{ \text{DT} \} \{ *, \text{NN} \} \rightarrow \langle \{ \text{DT} \} \{ \$FT, \text{NN} \}, \text{confidence: } 90\%$

R3:  $\{ \$FT \} \{ \text{VBZ} \} \{ \text{JJR} \} \{ \text{thanIN} \} \{ *, \text{NNP} \} \rightarrow \{ \$FT \} \{ \text{VBZ} \} \{ \text{JJR} \} \{ \text{thanIN} \} \{ \$ES2, \text{NNP} \}, \text{confidence } 70\%$ .

- After the preprocessing tasks, the sentence becomes:

$\{ \text{Coke}, \text{NNP} \} \{ \text{is}, \text{VBZ} \} \{ \text{preferred}, \text{VBN} \} \{ \text{because}, \text{IN} \} \{ \text{the}, \text{DT} \} \{ \text{taste}, \text{NN} \} \{ \text{is}, \text{VBZ} \} \{ \text{better}, \text{JJR} \} \{ \text{than}, \text{IN} \} \{ \text{Pepsi}, \text{NNP} \}$

- First, the rule R2 is applied since it has the highest confidence and the compared feature is identified

$\{ \text{Coke}, \text{NNP} \} \{ \text{is}, \text{VBZ} \} \{ \text{preferred}, \text{VBN} \} \{ \text{because}, \text{IN} \} \{ \text{the}, \text{DT} \} \{ \$FT, \text{NN} \} \{ \text{is}, \text{VBZ} \} \{ \text{better}, \text{JJR} \} \{ \text{than}, \text{IN} \} \{ \text{Pepsi}, \text{NNP} \}$

- Then, the rule R1 enables to found the first entity compared:

$\{ \$ES1, \text{NNP} \} \{ \text{is}, \text{VBZ} \} \{ \text{preferred}, \text{VBN} \} \{ \text{because}, \text{IN} \} \{ \text{the}, \text{DT} \} \{ \$FT, \text{NN} \} \{ \text{is}, \text{VBZ} \} \{ \text{better}, \text{JJR} \} \{ \text{than}, \text{IN} \} \{ \text{Pepsi}, \text{NNP} \}$

- Finally, the standard of comparison is labelled:

$\{ \$ES1, \text{NNP} \} \{ \text{is}, \text{VBZ} \} \{ \text{preferred}, \text{VBN} \} \{ \text{because}, \text{IN} \} \{ \text{the}, \text{DT} \} \{ \$FT, \text{NN} \} \{ \text{is}, \text{VBZ} \} \{ \text{better}, \text{JJR} \} \{ \text{than}, \text{IN} \} \{ \$ES2, \text{NNP} \}$

Overall, this method makes it possible to retrieve complete relations in all 32% of the sentences tested. However, as compared to the precision (entityS1: 100%, entityS2: 85%, features: 98%), the recall is less good (entityS1: 68%, entityS2: 59%, features: 43%). The good results obtained for entity S1 can be explained by the fact that they are easily recognisable as they often occur at the beginning of a sentence or before a verb. In contrast, entity S2 can appear anywhere in the sentence.

#### 3.2.1.4.2 Other Approaches

In order to avoid writing manual rules which could be domain-dependent and to capture more than one comparative in a sentence as well as comparison involving various features, Xu, Liao, Li and Song (2011), who also work on consumer reviews, design a CRF model

that link relations and entities, relations and words as well as entities and words. The focus is mainly on the identification of relations in comparisons and not a lot is said about the identification of comparative sentences itself. The features they use include capital letters, part-of-speech tags, affixes, linguistic triggers (“unlike”, “then”, “same”, “similar”, “improvement over”, “in contrast to”), syntactic paths derived from syntactic trees and grammatical roles (predicate, subject, attribute...). In this respect, first some preprocessing tasks such as part-of-speech tagging, dependency parsing and named entity recognition are performed. Then, the model relies on probability to recognise each element of the comparison based on its part of speech and on the neighbouring words. In comparison to Jindal and Liu’s method (2006b), this graphic model generates a visualisation of the comparison and can capture more than two compared terms as well as the direction of the comparison (better (>), worse (<), same (=)). In addition, although it yields better recall and precision than the previous method, its accuracy is lower as far as the extracted relations are concerned.

### **Example**

In the sentence “N95 has better reception than RAZR2V8 and Blackberry Bold 9000”, the model would identify two comparative relations  $r_1$  and  $r_2$  that could be summarised as follows:

$r_1$ : > (N95, Motorola RAZR2 V8, reception, better) ==> the N95 and the RAZR28 are compared in terms of their respective reception

$r_2$ : > (N95, Blackberry Bold 9000, reception, better) ==> the N95 is compared to the Blackberry Bold 9000.

Thus if N95, Blackberry Bold 9000 and RAZR2V8 denote the three products, better refers to the sentiment and reception to the attribute that is compared.

Still with the aim of mining opinions in reviews, Kessler and Kuhn (2013) use an existing semantic role labelling system to correctly tag, in sentences that supposedly contain a comparison, the compared items as well as the feature in terms of which they are compared. Semantic role labelling consists in identifying all the arguments of a specific verb or predicate and assigning them a role denoting the relationship that links them to that verb (Giver, Agent, Patient, Reason, Speaker, Message, Judge, Evaluatee...) (Gildea & Jurafsky, 2002; Màrquez, Carrerars, Likowski & Stevenson, 2008). The labelling task is done in two steps: first, the predicate is identified and then, its argument. For the first task, it is important to note that even though the precision is good, the recall is always lower than the one achieved with the list compiled by Jindal and Liu (2006 a & b). The results concerning the identification and the classification of the arguments, though better than the

baseline, are extremely low, with an F1-measure of 48-54% and 37-45% approximately. One of the main reasons given to explain those scores is data sparseness, as most predicates and arguments occur only once in the corpora used.

Although Park and Blake (2012) consider comparative sentence detection as a classification problem like the preceding methods, they are not interested in what comparative sentences can tell them about consumer reviews but rather about the ideas they contrast or sustain in scientific articles. For this purpose, they train three classifiers (Naive Bayes, Bayesian Network and Support Vector Machine) with a set of features made up of lexical cues (“versus”, “times that of”, “fewer”, “similar”, “different”) and syntactical rules (parse tree structure of a particular comparative sentence) such as [prep  $W_{1\_than}$ ], which in the sentence “DBP is several orders of magnitude more mutagenic/carcinogenic than BP” enables to tag “BP” as the standard of comparison. The three classifiers perform rather well overall, the Bayesian Network having the highest accuracy score (93.2%) and the highest area under the ROC curve (95.8%).

From the various works that have been exposed in this section, it is possible to see that as far the detection and the analysis of comparative sentences are concerned, the potential figurative nature of comparative constructions has been totally ignored.

### 3.2.2 Detection and Analysis of Non-Literal Comparisons

Unlike its counterpart, simile detection has given a prominent place to comparisons. Some of the early computational works on similes tackle the issue of disambiguating similes from literal comparisons and as such, provide interesting insights as they are based on existing cognitive theories.

#### 3.2.2.1 Weiner’s Proposal

According to Weiner (1984), simile computation must take into account the following elements:

- salience: in accordance with Ortony (1979), the high-salient attributes of the vehicle should be low-salient in the tenor;
- the context of usage as several properties can be epitomised by the vehicle depending on the context. In this respect, the meaning of a simile is shaped by the fact that the vehicle could be presented as the prototypical example of a particular quality and by its probable values or acceptations. For example, while in “John’s hair is like a carrot”, “carrot” would

refer to the colour orange, in “My cat’s tail is like a carrot”, “carrot” refers to an elongated pointed shape.

- some sentences seem more literal because they are less hyperbolic, especially when the vehicle and the tenor share high-salient features (Ortony, 1979).
- the position of the vehicle and the tenor in a taxonomic hierarchy and incongruity (the semantic distance between the tenor and the vehicle).

### **Example**

In these two sentences “Penguins are like wolves” and “Dogs are like wolves”, even though vehicles and tenors in both sentences belong to the same semantic category, namely “animals”, the first sentence can be considered figurative unlike the second one as the terms compared are more distant. In this respect, Weiner (1984) suggests that the best metaphors are “those presented by the best poets, those in which a vague experience is clarified through the predicates, salient or otherwise, of B terms” (p. 7).

- a high number of predicates shared between the tenor and the vehicle: Jane’s eyes are like stars (beauty, brightness, ...).

As simile understanding requires people to intuitively know the salient features of the compared items as well as their position in a hierarchical taxonomy, Weiner (1984) states that the semantic relations expressed in a knowledge representation system such as KL-ONE (Brachman & Schmolze, 1985) can give insight into the incongruity and the semantic distance characterising similes. Knowledge representation can be defined as the “field of study concerned with using formal symbols to represent a collection of propositions believed by some putative agent” (Brachman & Levesque, 2004, p. 4). Moreover, as such a representation includes known facts often called Roles about concepts, it can also be used to grade the prototypical features of the concept and ultimately to weigh salience. For a proper name such as “John”, for example, Roles would specify the fact that it is the name of a human being, and human beings are rational creatures that have eyes, hair, hands, feet... The nature of the Roles could be further circumscribed by using a Value Restriction (V/R), for example, if temperature is measured with a scale from 1 to 8, 1 being “more than extremely hot”, and 8 being “more than extremely cold”, the temperature of ice is restricted to 7 (extremely cold) while the hand’s temperature oscillates between 3 (hot) and 6 (cold), as shown in the figure below.

**Figure 3.3 Correspondence between various temperatures and natural elements**

(1) Beyond linguistic description (hot)	(2) Extremely hot	(3) Hot	(4) Lukewarm	(5) Cool	(6) Cold	(7) Extremely cold	(8) Beyond linguistic description (cold)
Mars	Boiling water	Plausible temperatures of JOHN'S HANDS			Temperature of ICE		

In this respect, knowledge representation can account for hyperbolic similes. In a sentence like “John’s hands are like ice”, a conceptual representation would show that although “TEMPERATURE” is a feature common to both ICE and HANDS, the temperature expressed by EXTREMELY COLD is out of the normal range of temperature of HANDS, which makes the statement hyperbolic.

Weiner (1984) also acknowledges that due to the existence of technical languages and of different language registers, one semantic network could not be enough to process similes: the semantic network should be adapted to the situation at hand, taking into account the context of utterance. For example, “John is an animal”, would have a literal meaning for a veterinarian but not elsewhere.

Weiner (1987) argues that approximation and inequality as far as the features of the tenor and the vehicle are concerned play an important role in differentiating similes from literal statements. In this respect, not only distant domains but also predicate inequality/hyperbole characterise figurative language. A sentence such as “The lane has the shape of a disk with its edges warped in opposite directions, like the brim of a fedora”, therefore, is a literal analogy as opposed to “a novel by, as it happened, a young writer who had, in the words of one critic, ‘made all previous American Jewish writing look like so much tasteless matzo dough.’” Based on these observations, Weiner (1987) proposes a three-step algorithm to analyse similarity statements which has, however, actually not been implemented.

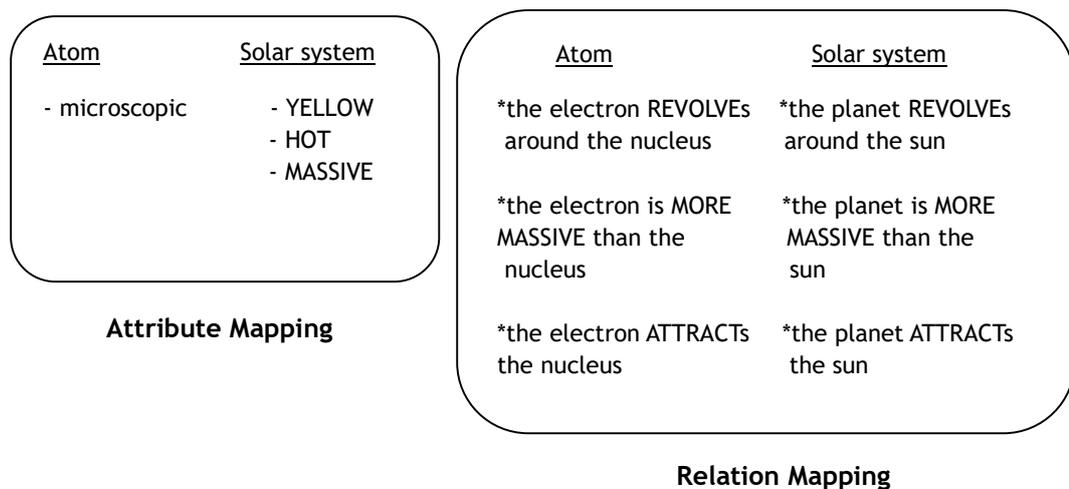
1. If the topic is an individual concept (IC), such as John, establish restrictions (using the Restricts link), if any, on the Role in question (for example, JOHN’s HANDS). If there are restrictions, note these; otherwise, note inherited V/RS.
2. Establish those salient predicates for the vehicle for which the topic also has a Role (for example, TEMPERATURE for ICE and JOHN’s HANDS).

3. If the V/Rs for these Roles are extreme in the vehicle but not in the topic, the utterance is hyperbolic. If on the other hand, the predicates are approximately the same, the vehicle is assumed to epitomise those predicates in a non-hyperbolic way. If in addition, the vehicle and topic are in the proper relationship to one another with respect to the taxonomy (i.e., there exists domain distance as described above), the utterance is metaphorical. Given that these conditions hold, raise the salience of the relevant Roles of the topic.

### 3.2.2.2 Structure-Mapping Theory

Structure-mapping (Gentner, 1982, 1983) is mainly interested in studying analogical reasoning. In this framework, objects are described by their attributes (predicate taking only an argument, green, thin, tall) and by their relations (predicate taking 2 arguments, REVOLVE (x,y), COLLIDE (x,y), GREATER-THAN(x,y). The number of mapped attributes and relations between the items compared enables to distinguish the different types of comparative constructions. In the case of a literal statement, the terms compared overlap significantly not only in terms of their attributes but also in terms of their relations. Consequently, whereas “The helium atom is like the neon atom” will be a literal statement because both atoms share exactly the same attributes and relations, “The hydrogen atom is like the solar system” is a simile because although the compared items have a small number of attributes in common, their relations overlap (see Figure 3.4). When it comes to anomalies like “Coffee is like the solar system”, the terms compared have simultaneously very few attributes and very few relations that overlap.

**Figure 3.4 Adaptation of the structure-mapping theory to the sentence “The hydrogen atom is like the solar system”**



Gentner’s theory (1982, 1983) presupposes that all the knowledge about concepts as well as their attributes and relations have already been given. If some examples have been

simulated in the structure-mapping engine (Falkenhainer, Forbus & Gentner, 1989), an automatic system devoted to the interpretation of analogical statements, there has been no real evaluation of its ability to disambiguate comparative statements. Furthermore, Ferrari (1997) criticises its bias for often putting aside in the mapping process some attributes such as colours, arguing that they can also play a central role in similes.

### 3.2.3 Automatic Detection of Similes

Simile detection can be divided into two main types: partial simile detection which only focuses on the ground and the vehicle and full simile detection which seeks to retrieve all the components of the simile. In addition, some methods have been proposed to measure the figurativeness of comparisons. As it was the case with the detection of comparative sentences, the detection of similes also relies primarily on the presence of a specific marker in a sentence.

#### 3.2.3.1 Partial Simile Detection

The recognisable pattern that some similes follows make them easy to be found by a search engine, as shown by Roncero, Kennedy and Smyth (2006). Taking advantage of this regularity, Veale and Hao (2007) use Google search engine to create a large database of similes: first, for each adjective of a pre-compiled list, they keep the 200 first results of the query “as ADJ as a|an NOUN” and then, for adjectives not in the list, they look for the form “as \* as a NOUN”. While the vehicle has been disambiguated automatically to arrive at the most adequate sense, similes have been manually filtered by a human judge. Interestingly, the range of adjectives associated with some vehicles can be used to derive the most salient traits of a particular word.

#### **Example**

Gladiator ==> manly, violent, competitive...

Through these salient traits, it is possible to group together words belonging to the same semantic category as the vehicle. To capture automatically those words, the first step is to obtain the superordinate term or hypernym through the query “P \* such as C” (“manly \* such as gladiators). Then, the words at the same level can be retrieved by repeating the following query: “P S such as C and \*”, C being the last item identified, (“manly men such as gladiators and \*” => “manly men such as soldiers and \*”). Finally, by associating a

salient trait to each C, it has been possible to generate a database that, for each word gives a net of fine-grained categories of a salient property and of the word belonging to the same category<sup>10</sup>.

**Figure 3.5 Examples of categories for the noun “gladiator” given by the Thesaurus Rex (Veale & Li, 2013)**

strong:fighter, fighting:fighter,  
**fighting:unit,** strong:unit,  
athletic:strength,  
bestial:strength, thuggish:violence,  
bestial:violence,  
heroic:strength, brute:muscle,  
brutish:muscle,  
stormy:violence, brute:strength,  
brutish:strength, equine:muscle,  
godlike:strength, brutish:beet,  
tyrannical:violence,  
godly:strength,

In addition, Veale and Hao (2009) show that by modifying the query to retrieve hedged similes of the form “about as \* a|an \*”, a completely different set of similes can be extracted. Furthermore, the presence of “about” before the marker seems often to signal the beginning of an ironic simile.

This simile extraction method has inspired other research works such as Li, Huang, Zhang, Chen and Tang (2012) who use it for sentiment analysis to retrieve extract similes in English and in Chinese from the web, relying on adjectives found in WordNet (Fellbaum, 1998) and in HowNet (Dong, Dong & Hao, 2010).

### 3.2.3.2 Full Simile Detection

#### 3.2.3.2.1 Full Simile Detection in French

Few research works have been done on the automatic detection of similes or comparative sentences in French. Ferrari (1997) proposes a system relying on linguistic cues and syntactic patterns to detect metaphors in French texts. The term “metaphor”, in this

---

<sup>10</sup> <http://ngrams.ucd.ie/therex2/>

research, is taken in its broad sense to also include similes. A first list of linguistic markers was drawn from a collection of texts written by students. Then, to that list were added synonyms of the identified linguistic markers and lexemes which share the same morphological root. Finally, all linguistic markers were classified into groups, depending on their syntax and the kind of relation they introduce. For instance, class A contains linguistic clues that express an explicit comparison (“plus ... que”, “aussi ... que”, “comme”, “à la manière de”, “pareillement à”, “similaire à”, etc.) whereas class B comprises other linguistic cues that establish a comparison through their meaning (“ressembler”, “sembler”, “paraître”, etc.). Ferrari (1997) notices that the grammatical category strongly influences the syntax of the sentences, making it easy to tag the vehicle and the tenor. In practice, if a marker is found in a sentence, the proposed system attempts to gather the words in the sentence into phrases and then, seeks to determine which group is metaphorical depending on the type of marker used. In addition, because of the specific attributes of the vehicle that they transfer temporarily to the tenor, unlike literal comparisons, similes exhibit what is referred to as “tension” or “duality”. This idea of tension proper to metaphorical structures is already evoked by Richards (1936) in *The Philosophy of Rhetoric*: “As the two things put together are more remote, the tension created is, of course, greater. That tension is the spring of the bow, the source of the energy of the shot, but we ought not to mistake the strength of the bow for the excellence of the shooting; or the strain of the aim” (p. 125). Nothing, however, in Ferrari’s work (1997) is proposed to effectively handle this “tension” in the detection process.

### *3.2.3.2.2 Full Simile Detection in English*

The following methods attempt to take advantage of the correlation between the function of terms in the sentence and their role in the simile. In a prototypical simile such as “A is B like C” or “A is like C”, while C is a complement, B is a predicate adjective, “is” is the verb which has for subject A, the tenor of the simile.

In this respect, Niculae and Yaneva (2013) propose an approach for extracting and mining similes using GLARF (Grammatical and Logical Argument Representation Framework) (Meyers, Grishman, Kosaka & Zhao, 2001), a framework for predicate-argument structure which regularises the output of parse trees (see Appendix 3 for an example). In addition to the tenor and the vehicle, this approach looks for the eventuality or the verb of the main clause of the simile, and only for adjectival grounds. Furthermore, it only takes into account sentences that have nominal tenors and nominal vehicles. For each sentence of the text, the procedure is as follows:

- explore the nodes of the sentence to find one of the listed markers;
- if the marker is found and has a common noun as complement, label that common noun as the vehicle;
- look for a verb that is syntactically connected to the marker;
- if such a verb exists and has a common noun as subject, that common noun is the tenor;
- label the verb as the event;
- if an adjective is connected to the verb, label that adjective as the ground.

This method had been tested with the markers “as” and “like” on two datasets. Whereas a precision of 70.5% and a recall of 41.7% is reported for partial matching of comparisons with “like”, a precision of 29.6% and recall of 64.8% is mentioned with comparisons with “as”. It is worth noting that a correct partial matching concerns 12% of the analysed sentences with “like” and 16% of the sentences in “as”. In most cases, the method does provide a lot of wrong and null matches: only 22% sentences with “like” have all their simile/comparison elements correctly identified as opposed to 66% without any correct match. Similarly, 37% of sentences with “as” are fully matched as opposed to 48% without any correct match.

Figure 3.6 gives an example of a sentence tagged with this method (see Appendices 1A and 2.III for more information about the part-of-speech and dependency tags used). By replacing GLARF with TurboParser (Martins, Smith, Xing, Aguiar & Figueiredo, 2010), a dependency parser, Niculae (2013) notices a slight decrease of the precision (30% vs 24%) but a significant increase in the recall (43% vs 71.1%). Generally speaking, several reasons can be advanced to explain those results: tagging errors due to the polysemy of “like” or parsing errors, as sometimes it happens that some complements of the markers are not markers, the wrong term is identified as complement of the markers, or the wrong verb is connected to the marker.

**Figure 3.6 Example of a sentence output (Niculae, 2013)**

The	the	DT	1	2	NMOD	_		
characters	character	NNS	2	3	SUB	_		
are	be	VBP	3	9	VMOD	_		
well	well	RB	4	3	VMOD	_		
drawn	draw	VEN	5	3	VC	_		
and	and	CC	6	9	VMOD	_		
the	the	DT	7	8	NMOD	_		
pace	pace	NN	8	9	SUB	TOPIC		
is	be	VBZ	9	0	ROOT	EVENT		
like	like	IN	10	9	PRD	COMPARATOR		
a	a	DT	11	12	NMOD	_		
locomotive	locomotive	NN	12	10	PMOD		VEHICLE	
.	.	P	13	9				



### **Example**

\*\*Semantic relatedness (Niculae, 2003)

The piano ripples like patent leather. [ DM (piano, leather) = 0.076]

Ink, like paint, uses subtractive colour mixing while the video monitor uses the additive colours; red, green and blue, to produce the same effect. [ DM (ink, paint) = 0.502]

In addition, Niculae and Danescu-Niculescu-Mizil (2014) apply this measure to identify similes extracted from a collection of Amazon consumer reviews and show that by combining it to other attributes, it is possible to predict similes almost as accurately as humans. Their experiment focuses on three markers: “as... as”, “more / less ... than”, “like”.

The additional attributes or “linguistic insights” they choose to consider can be divided into two groups:

- domain-dependent features: the specificity of the word to the domain (Electronics, Books, Jewelry and Music), the domain itself;
- domain-independent features: the semantic similarity between the tenor and the vehicle, the presence of an article before the vehicle, and three features previously used for metaphor identification: the degree of abstraction of the term, its degree of imageability and its supersenses (the superordinate categories to which it belongs).

As far as the domain of the reviews is concerned, it is possible to notice that the domain of use of the comparison/simile plays a great role in their detection task as a whole: not only are domain-related vehicles more frequent in literal comparisons but as similes tend to be more common in some domains, knowing the domain could help for the identification task. Furthermore, Niculae and Danescu-Niculescu-Mizil (2014) contrast topic-vehicle similarity in literal and in figurative comparisons and confirm that indeed, tenors and vehicles are most often semantically connected in literal comparisons. Of the three metaphor-inspired features, imageability, especially, proves itself particularly interesting to distinguish similes from literal language.

According to the results in the classification task, only the previously described “linguistic insights” perform almost as good as the best system which takes in addition to these insights, a slotted bag of words (distribution of each word as part of a simile). Further investigations show that whereas the semantic similarity of the tenor and the vehicle, the vehicle specificity and the use of a vehicle belonging to the category of communication mark a literal comparison, the absence of an indefinite article before the vehicle and the

imageability of the vehicle generally characterise a simile. Similarly, the use of “picture” and of “other(s)” as vehicles generally predict literal comparisons whereas the use of “crap” or “life” is a good indicator for a simile. Those observations appear, however, too connected to both the type of texts and the domain to be successfully generalised to similes from a more generic text.

It is also important to note that errors most often made by the system are due to metonymy (“the typeface was larger than most books” actually means larger than the typefaces found in most books), ellipsis (“a lot [of songs] are even better than sugar”) or polysemy (“the rejuvenac formula is about 10 times better than yogurt”).

Simile detection is mainly challenging because of the polysemy of the markers and because of its elliptical nature. Until now, if most algorithms geared towards the detection of comparative sentences in general and similes in particular have looked for ways to disambiguate the marker, they have mostly ignored the ambiguity inherent to some similes structures. In addition, the fact that most computational methods have been designed and tested on specific types of texts (consumer reviews, biomedical articles...) raises the question of their successful application to other domains.



## 4 SIMILE ANNOTATION

According to Ide and Romary (2003), an annotation corresponds to “[t]he process of adding linguistic information to language data (‘annotation of a corpus’) or the linguistic information itself (‘an annotation’), independent of its representation” (p. 2). In the creation of a reference corpus, annotation plays a crucial role as it enriches the texts with readily available trustworthy information that can be used for different research purposes. For instance, if one is interested in the frequency or the context of usage of a particular word, phrase or part of speech, such information could be effortlessly retrieved if one can automatically query a representative corpus in which the genre or domain of text, sentence boundaries, the lemmas and the part of speech of each word are indicated. In addition to facilitating research and providing a reliable source for scientific investigation, manual or automatic textual annotations also serve as standards against which results from future experiments can be measured. In this respect, data annotation enables to reduce costs as annotations from experts are generally expensive as well as time-consuming, and encourages objective comparisons of methods aimed at tackling a specific natural language processing task.

The first section of this chapter is concerned with general principles of linguistic annotations. The second section presents the main features used for simile analysis in literary studies. Finally, the last section gives an overview of some existing corpora containing annotated comparisons or similes.

## 4.1 Principles

Annotating texts has been practised for quite some time in the computational linguistics community. Unfortunately, as underlined by Ide and Sperberg-McQueen (1995), most annotation schemes from the 1960s to the 1980s were either not reusable or needed consequent alterations because of their conceptual bias. In this respect, arose the need to centralise those individual endeavours to come up with standards so as to spend less time, money as well as manpower in developing annotation schemes or tools and to more easily exchange, merge, exploit and compare language resources (Kahrel, Barnett & Leech, 1997; Ide & Romary, 2003). Essential questions that must be considered when building an annotation scheme framework relate to the representation format and the purpose of these annotations.

Among the various maxims listed by Leech (1993) concerning corpus annotation schemes, four are particularly important as far as this work is concerned:

- the corpus must be easily stripped of all annotations so as to obtain only the original corpus in its entirety;
- the annotations should be retrievable from the corpus in order to be written in a different document;
- the annotation scheme should not be presented as “God’s truth”;
- the annotation scheme should be consensual and should not exclusively be influenced by a particular theoretical framework (p. 275).

In practice, annotations may appear in the same document as the data or in a separate document. In the latter case, they are often referred to as “stand-off annotations” (Ide & Romary, 2003). Despite the numerous advantages of this type of annotations, for instance, non-overlapping tags describing distinct features, lighter files and effortless merging of different annotation schemes, Ide (2004) notes a great number of corpus still incorporate annotations with the original data because of the development of inter-document linking within the XML framework.

### 4.1.1 Types of Linguistic Annotations

Depending on the level of linguistic analysis, Leech (1993, 2005) distinguishes the following types of annotations:

- orthographic annotations that can be used to disambiguate some linguistic tokens, to give more information about their spelling or to correct it.

### Example

\*\* Information about capitalisation (TEI P5 Guidelines, 2016).

```
<entry>
  <form>
    <orth>academy</orth>
  </form>
  <cit type="example">
    <quote>The Royal <oRef type="cap"/> of Arts</quote>
  </cit>
</entry>
```

- phonetic/phonemic annotations that are related to word pronunciation.

### Example

\*\* Transcription of the words “mackle” and “macule” (TEI P5 Guidelines, 2016).

```
<form>
  <orth>mackle</orth>
  <pron>"makəl</pron>
</form>
<form>
  <orth>macule</orth>
  <pron>"makju:l</pron>
</form>
```

- prosodic annotations that generally describe spoken data and can underline stress, pitch, intonation, pauses, etc.

### Examples

1/Transcription of pauses (TEI P5 Guidelines, 2016).

```
<u>
  <seg>we went to the pub yesterday</seg>
  <pause/>
  <seg>there was no one there</seg>
</u>
<u>
  <seg>although its an old ide´a</seg>
  <seg>it hasnt been on the mar´ket very long</seg>
</u>
```

## 2/Transcription of the voice volume

```
<u>  
<shift feature="loud" new="f"/>Elizabeth  
</u>  
<u>Yes</u>  
<u>  
<shift feature="loud" new="normal"/>Come and try this <pause/>  
<shift feature="loud" new="ff"/>come on  
</u>
```

- grammatical tagging which associates each word or token with its grammatical category or word class.

**Examples** (See Appendix 1B for details about the tags used)

1/ Brown Corpus output for the sentence “The spray rails are first glued on the outside and fastened from the inside with screws.”

The/at spray/nn rails/nns are/ber first/rb glued/vbn on/in the/at outside/nn and/cc fastened/vbn from/in the/at inside/nn with/in screws/nns ./.

2/ TEI P5 markup for the same sentence:

```
<s n="85"><w type="AT">The</w> <w type="NN">spray</w> <w type="NNS">rails</w> <w  
type="BER">are</w> <w type="RB">first</w> <w type="VBN">glued</w> <w type="IN">on</w> <w  
type="AT">the</w> <w type="NN">outside</w> <w type="CC">and</w> <w type="VBN">fastened</w> <w  
type="IN">from</w> <w type="AT">the</w> <w type="NN">inside</w> <w type="IN">with</w> <w  
type="NNS">screws</w> <c type="pct">./.</c> </s>
```

- syntactic annotations which give information about phrases and the overall sentence structures.

### Example

Penn Treebank output for the sentence “The biggest firms still retain the highest ratings on their commercial paper.” (See Appendix 2.I for details about the labels used)

```
((S (NP-SBJ The biggest firms)  
  (ADVP still)  
  (VP retain  
    (NP (NP the highest ratings)  
      (PP on  
        (NP their commercial paper))))  
  .))
```

- semantic annotations which can be done at the world level and may concern named entities or word senses. They can also be done at the sentence level to express the semantic relations between words and phrases (Kübler & Zinsmeister, 2015).

### Example

Word sense tagging from SemCor: Each noun, verb or adjective is associated with its corresponding meaning in WordNet (Fellbaum, 1998). The part-of-speech tags are taken from the Penn Treebank tagset (see Appendix 1A).

```
<s snum="12">
<wf cmd="ignore" pos="DT">Every</wf>
<wf cmd="done" pos="NN" lemma="policy" wnsn="2" lexs="1:10:00:">policy</wf>
<wf cmd="done" pos="NN" lemma="officer" wnsn="2" lexs="1:18:02:">officer</wf>
<wf cmd="ignore" pos="MD">cannot</wf>
<wf cmd="done" pos="VB" lemma="help" wnsn="4" lexs="2:42:08:">help</wf>
<wf cmd="ignore" pos="CC">but</wf>
<wf cmd="done" pos="VB" lemma="be" wnsn="1" lexs="2:42:03:">be</wf>
<wf cmd="ignore" pos="DT">a</wf>
<wf cmd="done" pos="NN" lemma="planning" wnsn="2" lexs="1:04:02:">planning</wf>
<wf cmd="done" pos="NN" lemma="officer" wnsn="2" lexs="1:18:02:">officer</wf>
<punc>.</punc>
</s>
```

- discourse annotations which regroup a wide range of annotations that express relations that exist beyond sentence boundaries. It, therefore, concerns among others pragmatic, coreference, and anaphora annotations.

### Examples

1/ A pragmatic annotation from the ELC XML (Alsop & Nesi, 2014)

```
<u who= "m2001"><summary type= "preview content of future lecture">you're going to need to be able to do all of those moment questions that are in the book<gap reason= "pause"/>because we're going to start using them next week to work out beam reactions</summary><gap reason= "pause"/><humour type= "sarcasm">thank you for the yawn</humour></u>
```

2/ Coreference annotation (Mitkov, as cited in McEnery, Xiao & Tono, 2006)

```
<COREF ID= "100">The Kenya Wildfire</COREF>estimates <COREF ID= "101" TYPE=IDENT REF= "100">it</COREF> loses $1.2 million a year in park entry fee because of fraud.
```

- stylistic annotations which have mainly focused on labelling the forms and functions of speech and thought.

### **Example**

Stylistic annotation of a short passage (Leech, McEnery & Wynne, 1997).

<sptag cat=NRSAP next=NRS s=1 w=10>  
He also called for **an immediate end to the fighting**.  
<P>  
<sptag cat=NRS next=IS s=0.48 w=15>  
Foreign Secretary Diyglas Hurd - who flew to Belgrade in a new push for peace - said  
<sptag cat=IS next=NRS s=0.52 w=16>  
**the West was just weeks away from pulling out if the Bosnian Serb warlords rejected peace.** <P>  
<sptag cat=NRSAP next=IS s=0.07 w=2>  
He warned  
<sptag cat=IS next=NI s=0.93 w=28>  
**that if the warring factions refused to talk, the allies would have no choice but to pull their troops out and lift the arms embargo on Bosnia's Moslems.**

---

NRSAP= narrative report of speech act with topic  
IS=indirect speech; NRS=narrative report of speech  
NI=narrative report of internal state

Annotations can be done manually, semi-automatically with the help of a computer program, or fully automatically. Of all the aforementioned annotation types, grammatical tagging annotations are most commonly used because they are easily done automatically as there exists a wide range of reliable part-of-speech taggers. Besides, since most of these tools achieve an accuracy of 95% to 98%, the percentage of error is so insignificant that, for most research tasks, it could be ignored, and post-editing could be overlooked (Leech, 2005). In contrast, pragmatic and stylistic annotations still strongly rely on human insights.

Levels of annotation are not mutually exclusive. For example, for a corpus of spoken speech, it could be interesting to mark prosody, but also parts of speech, phonetics and even spelling mistakes or corrections. Moreover, apart from linguistic information, other useful information could be stocked as annotations such as information about the author but also information about the structure of the text.

## 4.1.2 The TEI as the Annotation Standard in the Humanities

The Text Encoding Initiative (TEI) was born in the late 1980s in reaction to the growing need for clear, exhaustive and simple guidelines that combine the best existing annotation practices, do not require a dedicated software, could be customised and could serve as reference for scholars unfamiliar with annotations or in search of information on how to annotate a particular element (Barnard & Ide, 1997). In addition to addressing a variety of annotation categories and types of supports (dictionaries, language corpora, drama, literary prose, spoken data...), one of the main characteristics of the TEI guidelines are that they allow flexibility and are not prescriptive (Ide & Sperberg-McQueen, 1995). More than being simply guidelines, the TEI is a collaborative effort by a whole community of scholars. Consequently, its guidelines are evolutive and often rely on external feedback or propositions for improvement (Mylonas & Renear, 1999).

Initially written in SGML (Standard Generalised Markup Language), the TEI has adopted since its fourth version the XML (Extensible Markup Language), as there exist many tools that can create or support an XML text (Cummings, 2008). Both languages, of course, share a number of similarities since, in fact, XML is derived from SGML, but is a lighter version of SGML which makes it easier to use. In addition, over the years, various standards such as XQuery, XML Namespace, XML Schemas, XSL have been developed to facilitate the rendering, the processing and the transformation of XML documents. Structurally speaking, an XML document always starts with a document type declaration which states the specific XML version used and makes use of start- and end-tags or sometimes empty tags that enclose a particular portion of the called element and that provide information about the name and the attributes of that element (Bray, Paoli, Sperberg-McQueen, Maler & Yergeau, 2008).

**Figure 4.1 Example of an XML document**

```
<?xml version="1.0" encoding="utf-8"?>
<fitness record date="June-26-2012" number="1">
  <activity>dancing</activity>
  <style>energetic<style>
  <duration="minutes">45</duration>
  <burned calories="105"/>
  <breaks>
    <water quantity=50cl>1<water>
    <loo>0</loo>
  </breaks>
</fitness record>
```

As far as literature is concerned, the debate is open about the impact and the importance of markup in digital texts. Cummings (2008), for example, emphasises the link between markup and tenets of modern literary criticism currents, especially structuralism, because it associates the text with its interpretation, whereas McGann (as cited in Cummings, 2008) sustains that by reducing imaginative works to structural data, markup does not conform to reading habits in vogue in the humanities community. But, despite markup not being universally accepted by literary scholars, a quick glimpse at the table of contents of the TEI guidelines show on the one hand, the great range of literary materials that the TEI could potentially describe (verse, critical apparatus, manuscripts..) and on the other hand, the variety of the different elements that it can identify for research purposes (verse structure, meter, stage direction, names, dates, places, paragraphs, quotations and even the level of certainty of the annotation). More in detail, annotating the name of a character or an authority does not solely consist in marking each and any occurrence of its name but could differentiate its surname from its given name(s), contain information about its profession, link its name to the corresponding element in the DBpedia database, or say whether the name is in its abbreviated form, is the birth name, the usual name, or a pseudonym.

Despite their exhaustive coverage of some textual elements, the TEI guidelines fail to address figurative language as a whole but mention the metaphor in passing. Faithful to their original spirit, they give some general markup tags for such types of annotations, but ultimately entirely leave 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. (TEI P5 Guidelines, 6.7, 2016)

As can be seen from the two examples below, every markup scheme is, therefore, fair-game as far as figurative language is concerned, provided it respects the TEI principles, relies on or expands some of its tags. But, is this lack of clear rules not also symptomatic of the heterogeneity that characterises literary analyses of figurative language at large as be seen from the following examples?

## Examples

1/TEI by Example (Metaphorical Language, <http://teibyexample.org/modules/TBED04v00.htm#metaphor>)

```

<lg xml:id="p001" type="poem">
  <lg xml:id="s001" type="stanza">
    <l xml:id="l001">Poppadom</l>
    <l xml:id="l002">Oatmeal</l>
    <l xml:id="l003">Bubble gum</l>
    <l xml:id="l004">Cut of veal</l>
  </lg>
  <lg xml:id="s002" type="stanza">
    <l xml:id="l005">Mince for pie</l>
    <l xml:id="l006">Frozen peas</l>
    <l xml:id="l007">Video for Guy</l>
    <l xml:id="l008">Selection of teas</l>
  </lg>
  <lg xml:id="s003" type="stanza">
    <l xml:id="l009">Paper towels/garbage bags</l>
    <l xml:id="l010">Pasta sauce and Parmesan</l>
    <l xml:id="l011">Pumpkin seed and olive oil</l>
  </lg>
  <lg xml:id="s004" type="stanza">
    <l xml:id="l012">Cheesy crisps and favourite mags</l>
    <l xml:id="l013">Kidney beans (1 large can)</l>
    <l xml:id="l014">Cling film and kitchen foil</l>
  </lg>
</lg>
<spanGrp resp="RvdB" type="imagery">
  <span from="#l001" to="#l006">food</span>
  <span from="#l007">non-food</span>
  <span from="#l008">food</span>
  <span from="#l009">non-food</span>
  <span from="#l010" to="#l013">food</span>
  <span from="#l014">non-food</span>
</spanGrp>
resp=person responsible for the annotation

```

2/ Encoding for the first two lines of John Keats' "Ode on a Grecian Urn" (Singer, 2013)

```

<theme type="immortality">
  <| n="1"><litFigure type="apostrophe">THOU</litFigure> <ambiguity
  item1="immobile" item2="continuation">still</ambiguity>
  <negation>un</negation><connotation item="aggression"
  value="neg">ravish'd</connotation><litFigure type="metaphor"
  tenor="urn">bride</litFigure> of <w reg="silence">quietness</w>, </|>

  <| n="2"><litFigure type="apostrophe">Thou</litFigure>
  <connotation item="adopted" value="neg">foster-</connotation>
  <litFigure type="metaphor" tenor="urn">child</litFigure> of <litFigure
  type="personification"><w reg="silence">Silence</w>
  </litFigure> and slow <litFigure type="personification">Time</litFigure>, </|>
</theme>

```

## 4.2 Simile Description in Literary Studies

Literary scholars generally intermingle two levels of description to analyse and classify similes in literary texts: first, the number of simile components that are involved and secondly, the linguistic dimension that can be either structural or syntactic. It is worth noting that in some cases, the separation between the linguistic aspects of the simile is not so clear-cut.

### 4.2.1 The Structural Dimension

One of the most universally accepted categorisations of similes is the one based on the absence or presence of the ground. Beardsley (1950) introduces this distinction by distinguishing between "closed" and "open" similes. In a closed simile such as "Your smile is precious as a jewel", the scope is narrowed, as it is explicitly stated the respect in which the compared elements are similar or not. On the contrary, in an open simile like "Your smile is as a jewel", one must resort to the context, world knowledge and cultural background to be able to guess the source of the comparison.

Though these similes require from the audience more imagination and thinking, they may utterly fail to serve their purpose if they leave the door too wide open for any kind of interpretation.

Based on semantics, Fishelov (2007) further subdivides each of these main types of similes into conventional, non-conventional, opaque or confusing, and ironic similes. A simile is said to be conventional if it uses an obvious ground which is culturally perceived as an attribute of the vehicle, such as in “Peter is as wily as a fox”. Even if this sentence were to be transformed into the open simile, “Peter is as a fox”, the cunningness of foxes is so embedded into the English language that it would immediately make sense. In contrast, a sentence such as “Peter is like an old Chevrolet” is non-conventional in the sense that it defies usual simile associations and gives way to various interpretations. In confusing or opaque similes, the ground lends to the vehicle unfamiliar attributes, which hinders the understanding of similes like “Peter is joyful like a fox”. Finally, in an ironic simile, to achieve a particular effect, the ground is the antonym of the normally expected distinct trait of the vehicle; this occurs if one writes “Peter is as genuine as a fox” instead of “Peter is as wily as a fox”.

Two main observations should be made about the above simile framework. First, even if it obviously has its root in literary criticism, it was not used to describe specific literary texts but rather to make an experiment on the individual understanding challenges raised by each subtype of simile. In addition, the status of the irony as a figure of speech and the fact that similes are often combined with other figures of speech (Shabat Bethlehem, 1996) make it possible to imagine another classification of similes that rather takes into account the figure of speech that is used together with the simile.

As a matter of fact, other figures of speech are often considered in similes analysis when they enhance a particular stylistic feature or aspect of that simile. In this respect, Pistorius (1971), for example, finds in Flaubert’s *Madame Bovary* cases in which a metaphor is part of a simile: “ses yeux commençaient à disparaître dans une pâleur visqueuse qui ressemblait à une toile mince, comme si des araignées avaient filé dessus”<sup>11</sup> (p. 236) and those in which a metaphor builds on or contributes to the image created by the simile: “leur grand amour, où elle vivait plongée, parut se diminuer sous elle, comme l’eau d’un fleuve qui s’absorberait dans son lit ; et elle aperçut la vase”<sup>12</sup> (p.239). It is important to observe that in the first example, the first simile “une pâleur visqueuse qui ressemblait à une toile mince” is

---

<sup>11</sup> English translation: “her eyes commenced to disappear in a viscous pallor which resembled a thin sheet, as if the spiders had been spinning above them.” (Flaubert, 1896, Vol II, p. 154).

<sup>12</sup> English Translation: “...their great love in which she lived immersed seemed to diminish under her like the water of a river which sinks into its bed; and she perceived the slime at the bottom.” (Flaubert, 1896, Vol I, p. 266).

followed by a second simile which prolongs the existing image. Generally speaking, it is, therefore, possible to distinguish between similes combined with other similes or other figures of speech at the sentence level or at the level of the simile proper, for instance, when the ground itself is used metaphorically: “Sa pensée, sans but d’abord, vagabondait au hasard, comme sa levrette, qui faisait des cercles dans la campagne (...)”<sup>13</sup> (p. 234).

Similes are also known to often rely on repetitions, either repetitions of the whole simile (“it’s as good as a play — as good as a play!”, Dickens, *Oliver Twist [OT]*, as cited in Tomita, 2008b, p. 9) or phonetic repetitions. In everyday life, sound repetition in similes can be found in various idiomatic similes such as “cool as a cucumber” or “busy as a bee”. Literary examples of this type of similes include “I am glad to remember, as mute as a mouse about it” (Dickens, *David Copperfield [DC]*, as cited in Tomita, 2008a, p. 5), “for the old Scholar —what an excellent man !—is as blind as a brickbat” (Dickens, *DC*, as cited in Tomita, 2008a, p.5), “we would have put you a clean collar on, and made you as smart as sixpence!”(Dickens, *OT*, as cited in Tomita, 2008b, p.7) and “A many, many, beautiful corpses she laid out, as nice and neat as waxwork” (Dickens, *OT*, as cited in Tomita, 2008b, p. 8). This last example is particularly interesting because apart from the alliteration, this simile contains two separate grounds, which corresponds to what Pistorius (1971) calls a “doubled simile” or what Kirvalidze (2014) refers to as a “polymotivated simile”.

Similarly, so as to emphasise a particular point or to create a particular impact, a simile can be made up of two vehicles that could share the same ground or not.

### Examples

- Simile with at least two vehicles that share the same ground: “... la vieille cité normande s’étalait à ses yeux, comme une capitale démesurée, comme une Babylonie où elle entraît.”<sup>14</sup> (Flaubert, *Madame Bovary*, as cited in Pistorius, 1971, p. 230).

- Similes with two vehicles having each their respective ground: “...un regret immense, plus doux que la lune et plus insondable que la nuit.”<sup>15</sup> (Flaubert, *Madame Bovary*, as cited in Pistorius, 1971, p. 230).

---

<sup>13</sup> English translation: “Her thoughts, without an aim at first, wandered at hazard like her greyhound who ran around in the fields in circles, barking after the yellow butterflies, chasing the shrew-mice, or snapping at the wild poppies on the edge of a grain field. (Flaubert, 1896, Vol. I, p. 73)

<sup>14</sup> English translation: “... the old Norman city stretched itself out before her eyes like an immeasurable capital, like a Babylon which she was entering.” (Flaubert, 1896, Vol II, p. 53).

Structural descriptions of similes also concern the syntax of one particular element or that of the simile as a whole. For instance, in the case of open similes, the word class (verb, adjective...) of the ground could constitute a good basis for differentiating between various similes. In addition to the nature of one of the elements of the simile, its length could also be a decisive factor of classification. One of the earliest and most enduring distinctions that have been made in literature concerns the epic or Homeric simile. Here is an example of an open simile translated from Homer's *Iliad*:

Before the lofty gates the champions twain  
Stood, as two oaks upon the mountain stand  
Rearing their heads on high, that through all time  
Bide brunt of wind and rain, by mighty roots  
Far spreading through the soil full firmly set.  
So these, on hand and strength reliant, bode  
Great Asius as he came, and fled him not. (Green, 1877, p. 91).

Why is that simile particular? Undoubtedly because of its lyricism and the fact that it runs on a considerable amount of lines of the poem, greatly extending the initial image. Chateaubriand (1739) contrasts comparisons in *The Bible* with comparisons in Homer's *The Iliad* and *The Odyssey*, stating that comparisons though generally simple can also be written in detailed form to personify an object, whereas Homeric similes are akin to paintings hung around an edifice to stop people from seeing elevation work occurring on its dome by presenting pastoral or landscape scenes. Though devoid of the epic nature of the Homeric simile and of its formulaic structure, some similes resemble them in their construction as the vehicle is often lengthened more than usual. In this respect, Pistorius (1971) makes a distinction between "simple" and "developed" similes on the one hand and between "symmetric" and "asymmetric" similes on the other hand. In a simple simile, the ground and the vehicle are used in their simplest form, i.e. the ground consists only of a single word and the vehicle of a minimal noun phrase or one expanded by one adjective or one prepositional phrase. In contrast, in a developed simile, phrases that make part of the simile components are extended to furnish more details. Typically, the vehicle is extended by the means of a relative clause.

---

<sup>15</sup> English translation: "...an immense regret, softer than the moon and more unfathomable than the night" (Flaubert, 1896, Vol II, p. 169)

### Examples

- Simple simile: “La conversation de Charles était plate comme un trottoir de rue.”<sup>16</sup> (Flaubert, *Madame Bovary*, as cited in Pistorius, 1971, p. 230).

- Simile with a developed vehicle: “...il se trouvait dans une de ces crises où l’âme entière montre indistinctement ce qu’elle enferme, comme l’Océan, qui, dans les tempêtes, s’entr’ouvre depuis les fucus de son rivage jusqu’au sable de ses abîmes.”<sup>17</sup> (Flaubert, *Madame Bovary*, as cited in Pistorius, 1971, p. 240).

This last simile is also considered as being asymmetric as the vehicle is much longer than the tenor. Asymmetry, therefore, is measured in terms of balance or lack thereof between the length of the tenor and that of the vehicle. To better illustrate the difference, here is an example of a perfectly symmetric simile: “le plancher de la sellerie luisait à l’œil comme le parquet d’un salon”<sup>18</sup> (Flaubert, *Madame Bovary*, as cited in Pistorius, 1971, p. 229).

Still regarding syntax, similes can also be classified according to the syntactic order of its elements. As Quintilian (trans.1876) notes, though the order of some elements is fixed, it is possible to conform to the standard order in English and French by writing the tenor before the vehicle or to opt for some variations by choosing to put the vehicle first and then the tenor: “In every comparison, either the simile precedes and the subject of it follows, or the subject precedes and the simile follows.” (Book VIII, Chap. III, 77, p. 105). In this quotation, the term “simile” is, of course, used metonymically to refer to the phrase formed by the marker and the vehicle. Of course, despite the inversion, the sentence must remain grammatically correct, which excludes non-sensical constructions of the type “was like a butterfly graceful the girl”.

Finally, a more thorough syntactic approach to simile analysis can take into account not only the position of distinct elements but the overall syntactic composition of the similes. An example of such a classification is presented in Table 4.1 which shows the finer distinctions that such a system enables to make, namely the type of verb on which the

---

<sup>16</sup> English translation: “The conversation of Charles was as flat as the pavement of the street...” (Flaubert, 1896, Vol. II, p. 67).

<sup>17</sup> English translation: “...he was in one of those crises in which the entire soul displays indistinctly all that it incloses, like the ocean which, in its tempests, opens up, from the fucus on its shore to the sand of its abysses” (Flaubert, 1896, Vol. II, p. 32).

<sup>18</sup> English translation: “The flooring of the saddle room was polished like the parquet of a salon” (Flaubert, 1896, Vol. II, p. 88).

simile is built and the semantic role played by the different similes in the text. In addition, a closer look at the structure associated with each type of simile suggests that the marker plays a non-negligible part in shaping the meaning of the simile.

**Table 4.1 Synthesis of the various syntactic structures of the similes found in *David Copperfield* and *Oliver Twist* (Tomita, 2008 a [p. 2-4] & b [p. 4-6])**

Intensifying similes	Descriptive Similes
<i>be (+as) + adjective + as + N</i> “...my head is <u>as</u> heavy <u>as</u> so much lead...” [DC]	<i>V + like + N</i> “...the mist rolled along the ground <u>like</u> a dense cloud of smoke.” [OT]
<i>be (+as) + adjective + as + Clause</i> “...she sat there, playing her knitting-needles as monotonously <u>as</u> an hour-glass might have poured out its sands. [DC]	<i>look/seem/appear + like + N</i> He certainly did look uncommonly <u>like</u> the carved face on the beam outside my window [DC]
<i>verb + as + adjective/adverb + as + clause</i> “Oliver was not altogether <u>as</u> comfortable <u>as</u> the hungry pig was...” [OT]	<i>-like + N</i> “...the death- <u>like</u> stillness came again” [OT]
<i>verb + as + adjective/adverb + as + N</i> “...they’ll come back for another, the day after to-morrow, <u>as</u> brazen <u>as</u> alabaster.” [OT]	<i>look + -like</i> “...the sombre shadows thrown by the tree upon the ground, looked sepulchral and death- <u>like</u> , from being so still.” [OT]
	<i>not + unlike + N</i> “They were not <u>unlike</u> birds, altogether...” [DC]
	<i>look + as + adjective + as + clause</i> “She had a little basket-trifle hanging at her side, with keys in it; and looked <u>as</u> staid and <u>as</u> discreet a housekeeper <u>as</u> the old house could have.” [DC]
	<i>verb + as + adjective/adverb + as if + clause</i> “The sun shone brightly: <u>as</u> brightly <u>as if</u> it looked upon no misery or care...” [OT]
	<i>verb + as if + clause</i> “...she controlled it soon, and spoke in whispers, and walked softly, <u>as if</u> the dead could be disturbed.” [DC]
	<i>look/seem/appear + as if /as though + clause</i> “he seemed to breathe <u>as if</u> he had been running...” [DC]

#### 4.2.2 The Semantic Dimension

The idea of a correlation between specific markers and a particular meaning attached to the simile seem to be shared by various authors. Bouverot (1969), for instance, distinguishes between images of type I built with a finite number of comparatives, prepositions or conjunctions (“ainsi que”, “de même que”, “comme”, “plus que”, ...) and which express a comparison as opposed to images of type II that are observed after a verb or an adjective

phrase (“pareil à”, “semblable à,” “on dirait”, “faire penser à” ...) which only semantically induces the idea of similitude or difference and convey a weakened identification. Similarly, Leech and Short (2007) separate conventional similes of the form “X is like Y” from quasi-similes which revolve around all other linguistic constructions expressing the idea of similitude or comparison. Examples of such quasi-similes in Conrad’s *The Secret Sharer* include: “here were lines of fishing stakes resembling a mysterious system of half-submerged bamboo fences” and “To the left a group of barren islets, suggesting ruins of stone walls (...)” (p. 66-67).

Traditionally, the semantics of the similes is concerned with measuring the semantic distance between the tenor and the vehicle. As a matter of fact, as far back as Aristotle’s *The Poetics*, has been elaborated a theory of metaphor relying on semantics: “Metaphor is the application of an alien name by transference either from genus to species, or from species to genus, or from species or species, or by analogy, that is proportion” (treans 1898, XXI. 4, 1457b, p. 78-79). Perpetuating this tradition, Quintilian (1876) too distinguishes four main types of metaphors: a living thing combined with another living another, an inanimate thing with another inanimate thing, an inanimate with living things, a living thing with an inanimate thing. Brooke-Rose (2002), aptly summarises the various predominant theories that classify similes based on its content by distinguishing: first, as we have seen Aristotle with the species/genus classification, then Aristotle’s successors among whom Quintilian, who introduce the animate/inanimate classification, afterwards the classification by domain of thought or activity used in the 19<sup>th</sup> and the 20<sup>th</sup> century for linguistic and literary analysis, and finally, the “analysis by dominant trait” which focused on the resemblances between the vehicle and the tenor (p. 9). Therefore, one way or another, describing similes through the different semantic categories or groups that they put together has been a fixed feature of literary studies. The interest that scholars have always shown towards this aspect of simile is far from being gratuitous as, as mentioned in Chapter 2, it is often believed that the more semantically far apart the compared elements are, the more resonating and surprising is the simile:

L’image est une création pure de l’esprit.

Elle ne peut naître d’une comparaison mais du rapprochement de deux réalités plus ou moins éloignées.

Plus les rapports des deux réalités rapprochées seront lointains et justes, plus

l'image sera forte — plus elle aura de puissance émotive et de réalité poétique... etc.  
(Reverdy, as cited in Breton, 1924).<sup>19</sup>

In scholarly texts about literature, the degree of abstraction or of animacy as well as semantic categories are generally used to describe the distant realities that are joined in a simile.

As far as the degree of abstraction or animacy is concerned, depending on the perspective that is adopted, it is possible to distinguish four types of similes (see Table 4.2). Grossly speaking, something is said to be concrete if it has a physical existence, is measurable and can be seen or touched whereas something is animate when it can move on its own volition. Inanimate objects, consequently, can be either concrete or abstract: although both “car” and “impression” are inanimate words, only “car” is concrete. Morinet (1995) criticises the use of such semantic labels to describe similes, claiming that they are unreliable as they are not as fixed as one would have thought. For instance, if a car can talk or drive by itself like KITT in the American TV show Knight Rider, is it still inanimate? And what about personified abstract entities so common in literature such as Death or Love?

---

<sup>19</sup> English translation:

The image is a pure creation of the mind.

It cannot be born from a comparison but from a juxtaposition of two more or less distant realities.

The more the relationship between the two juxtaposed realities is distant and true, the stronger the image will be -- the greater its emotional power and poetic reality...etc. (Stewart as cited in Tigges, 1988, p. 118).

Table 4.2 Combinations of degrees of abstraction and of animacy

Degree of abstraction	Degree of animacy
<i>abstract tenor - concrete vehicle</i> “In the town like Mason City... time gets tingled in its own feet and lies down <u>like</u> an old hound and gives up the struggle” (Warren, as cited in Kirvalidze, 2014).	<i>inanimate tenor - animate vehicle</i> “After making one or two sallies to her relief, which were rendered futile by the umbrella’s hopping on again, <u>like</u> an immense bird, before I could reach it, I came in, went to bed, and slept till morning” (Dickens, <i>DC</i> , as cited in Tomita, 2008a, p. 6).
<i>abstract tenor - abstract vehicle</i> “The answer, a little while in coming was fragile <u>as</u> the flight of a moth” (Capote, as cited in Kirvalidze, 2014, p. 28).	<i>inanimate tenor - inanimate vehicle</i> “... the two stone steps descending to the door were as white <u>as</u> if they had been covered with fair linen...” (Dickens, <i>DC</i> , as cited in Tomita, 2008a, p. 6).
<i>concrete tenor - abstract vehicle</i> “cette grande nef, qui s’étendait devant elle <u>moins</u> profonde <u>que</u> son amour...” (Flaubert, as cited in Pistorius, p. 237).	<i>animate tenor - inanimate vehicle</i> “Here he shook hands with me; not in the common way, but standing at a good distance from me, and lifting my hand up and down <u>like</u> a pump handle that he was a little afraid of” (Dickens, <i>DC</i> , Tomita, 2008a, p. 6).
<i>concrete tenor - concrete vehicle</i> “And here, in the very first stage, I was supplanted by a shabby man with a squint, who had no other merit than swelling like a living-stables, and being able to walk across me, more <u>like</u> a fly than a human being, while the horses were at a canter!” (Dickens, <i>DC</i> , as cited in Tomita, 2008a, p. 6).	<i>animate tenor - animate vehicle</i> “And here, in the very first stage, I was supplanted by a shabby man with a squint, who had no other merit than swelling like a living-stables, and being able to walk across me, more <u>like</u> a fly than a human being, while the horses were at a canter!” (Dickens, <i>DC</i> , Tomita, 2008a, p. 6).

Indeed, regarding semantic categories, it is rather difficult to find a standard, even though some categories such as “humans” and “animals” seem to be quite agreed upon. The semantic categories defined for a particular analysis, therefore, mostly appear to be dictated by the literary text itself. In this respect, while some categories remain rather general (“natural categories”, “abstract objects”), others are more fine-grained (“supernatural beings”, “vegetal elements”). The description of similes using semantic categories could either indicate the shift from one semantic category to another or could be centred on a specific thematic shared by a large group of similes.

### Examples

a) Human → Supernatural beings:

‘An angel,’ continued the young man, passionately, ‘a creature as fair and innocent of guile as one of God’s own angels, fluttered between life and death. (Dickens, *OT*, Tomita, 2008b, p. 13).

b) Vegetal similes in Proust's *À la recherche du temps perdu* (1913-1927) :

“Et ainsi l'espoir du plaisir que je trouverais avec une jeune fille nouvelle venant d'une autre jeune fille par qui je l'avais connue, la plus récente était alors comme une de ces variétés de roses qu'on obtient grâce à une rose d'une autre espèce.”

“...dans un de ces longs tuyautages de mousseline de soie, qui ne semblent qu'une jonchée de pétales roses ou blancs (qui) donnaient à la femme (...) le même air frileux qu'aux roses”. (Proust, as cited in Trousson, 1981, p. 5-6).

In addition to the two dimensions that have been discussed so far, another characteristic of similes that is often taken into consideration is their idiomaticity, what Tomita (2008 a & b) refers to as proverbial similes. An apt example of this type of similes would be “The walls were whitewashed as white as milk, and the patchwork counterpane made my eyes quite ache with its brightness” (Dickens, *DC*, as cited in Tomita, 2008 a, p. 5).

### 4.3 Existing Corpora of Annotated Comparisons and Similes

The annotation corpora presented in this section concern not only similes but also comparative constructions and can be divided into two groups: corpora resulting from computational experiments and that can be used as baselines in subsequent research works and manually annotated corpora. Furthermore, annotations can be made at the sentence- or at the world-level.

In Jindal and Liu's dataset (2006 a & b),<sup>20</sup> each identified comparison is marked and is immediately followed by its basic structure specifying its main components. The corpus is separated according to the different reviews analysed, and is in lowercase, with one sentence per line. Each component is numbered: 1\_ for the comparee, 2\_ for the standard of comparison and 3\_ for the quality/quantity.

In addition, a distinction is made between the different types of comparative structures defined during the experiment: non-equal gradable, equative, superlative and non-gradable.

---

<sup>20</sup> Freely downloadable at <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets>

### Example

\*\* Non-equal gradable and equative comparative constructions

```
<cs-1><cs-2>
```

the new 30GB iPod is 30 percent thinner than the previous 20GB color model, but the height and width are the same

```
</cs-1></cs-2>
```

1\_new 30GB iPod 2\_previous 20GB color model (thinner)

1\_new 30GB iPod 2\_previous 20GB color model 3\_height 3\_width (same)

It is important to note that this corpus gives the automatically generated results without any post-editing and therefore, some errors or omissions can be found.

### Example

```
<cs-2>
```

this thing, while looking pretty cool, is not as sexy as the ipod.

```
</cs-2>
```

2\_ipod 3\_sexy (as sexy as)

The second relevant dataset is the J. D. Power and Associates (JDPA) Corpus which is made up of blog comments about cars and digital cameras, in which entities, semantic relations between entities (part-of, feature-of, instance-of...), modifiers such as intensifiers (“quite”, “top”, “!”) and negators (“didn’t”, “no”, “without”, etc.), and words denoting a sentiment (“nice”, “poor”, “fun”) have been manually annotated (Kessler, Eckert, Clark & Nicolov, 2010). Therefore, unlike the previous corpus, comparative constructions are not the focus of the study but are perceived as part of a broader problem, the expression of sentiment.

### Example

*Text*

For right at \$13,000, I get a car that’s smaller and lighter than the Fit, and has a few fewer options, but still provides stability peppy performance and decent room for people and cargo (though admittedly not a lot of both at the same time) in an incredibly gas-efficient package.

*Corresponding annotations*

```
<annotation>
```

```
  <mention id="StructuralSentiment_Instance_270655" />
```

```
  <annotator id="1">1</annotator>
```

```
  <span end="978" start="968" />
```

```
</annotation>
```

```
<classMention id="StructuralSentiment_Instance_270655">
```

```
  <mentionClass id="Comparison">Comparison</mentionClass>
```

```
  <hasSlotMention id="StructuralSentiment_Instance_270660" />
```

```
<hasSlotMention id="StructuralSentiment_Instance_270659" />
<hasSlotMention id="StructuralSentiment_Instance_270657" />
<hasSlotMention id="StructuralSentiment_Instance_270658" />
</classMention>
<annotation>
  <mention id="StructuralSentiment_Instance_270624" />
  <annotator id="1">1</annotator>
  <span end="988" start="983" />
</annotation>
<classMention id="StructuralSentiment_Instance_270624">
  <mentionClass id="Mention.Units.Money"> Mention.Units.Money </mentionClass>
  <hasSlotMention id="StructuralSentiment_Instance_270689" />
</classMention>
<annotation>
  <mention id="StructuralSentiment_Instance_270634" />
  <annotator id="1">1</annotator>
  <span end="994" start="989" />
</annotation>
<classMention id="StructuralSentiment_Instance_270634">
  <mentionClass id="Mention.Organization"> Mention.Vehicles.Cars </mentionClass>
  <hasSlotMention id="StructuralSentiment_Instance_270731" />
  <hasSlotMention id="StructuralSentiment_Instance_270643" />
</classMention>
<annotation>
  <mention id="StructuralSentiment_Instance_270645" />
  <annotator id="1">1</annotator>
  <span end="994" start="989" />
</annotation>
<classMention id="StructuralSentiment_Instance_270645">
  <mentionClass id="Mention.Organization"> Mention.Organization </mentionClass>
</classMention>
[...]
```

```
<complexSlotMention id="StructuralSentiment_Instance_270660">
  <mentionSlot id="Less" />
  <complexSlotMentionValue value="StructuralSentiment_Instance_270634" />
</complexSlotMention>
<complexSlotMention id="StructuralSentiment_Instance_270659">
  <mentionSlot id="Dimension" />
  <complexSlotMentionValue value="StructuralSentiment_Instance_270583" />
</complexSlotMention>
<stringSlotMention id="StructuralSentiment_Instance_270657">
  <mentionSlot id="Same" />
  <stringSlotMentionValue value="true" />
</stringSlotMention>
<complexSlotMention id="StructuralSentiment_Instance_270658">
  <mentionSlot id="More" />
```

```
<complexSlotMentionValue value="StructuralSentiment_Instance_270543" />
</complexSlotMention>
<classMention id="StructuralSentiment_Instance_270634">
  <mentionClass id="Mention.Vehicles.Cars">Mention.Vehicles.Cars</mentionClass>
  <hasSlotMention id="StructuralSentiment_Instance_270731" />
  <hasSlotMention id="StructuralSentiment_Instance_270643" />
</classMention
```

The third corpus available online also deals with product reviews but is more interested in similes. For their experiment reported in the previous chapter, Niculae and Danescu-Niculescu-Mizil (2014) relied on Amazon Mechanical Turk to annotate a sample of around 2, 400 similes<sup>21</sup>. The corpus only consists of sentences in which a comparison between two common nouns has been found. The sentences are presented in the CoNLL format, the output format of the dependency parser used, to which the mentions “TOPIC”, “EVENT”, “PROPERTY”, “COMPARATOR” 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...

---

<sup>21</sup> <http://vene.ro/figurative-comparisons/>

**Example**

An annotated sentence from the Amazon Corpus (Niculae and Danescu-Niculescu-Mizil, 2014). The part-of-speech tags and dependency relations are detailed in Appendices 1A and 2.III.

```
# {"category": "Electronics", "figurativeness": [4, 2, 4], "title": "Sony VCT870RM Tripod w/Remote for Sony MiniDV, DVD, HDR-HC5 & HC7 Camcorders", "price": "unknown", "userId": "A2FG90RW53W8WS", "score": "5.0", "helpfulness": "1/1", "time": "1174694400", "profileName": "Hello", "productId": "B000063W8Q"}
```

The	the	DT	1	2	NMOD	_
tripod	tripod	NN	2	3	SUB	TOPIC
is	be	VBZ	3	0	ROOT	EVENT
like	like	IN	4	3	PRD	COMPARATOR
a	a	DT	5	6	NMOD	_
magnet	magnet	NN	6	4	PMOD	VEHICLE
because	because	IN	7	3	VMOD	_
it	it	PRP	8	10	SUB	_
always	always	RB	9	10	VMOD	_
brings	bring	VBZ	10	7	SBAR	_
questions	question	NNS	11	13	NMOD	_
and	and	CC	12	13	NMOD	_
people	people	NNS	13	10	OBJ	_
to	to	TO	14	10	VMOD	_
the	the	DT	15	16	NMOD	_
tripod	tripod	NN	16	14	PMOD	_
.	.	.	17	3	P	_

The chosen structure clearly makes this corpus more useful for NLP researchers. In addition, a certain bias has been introduced since the elements of the comparison were already indicated and were not asked to be corrected. As a matter of fact, since the algorithm design does not take coordination into account, in some cases, only partial results are found. Similarly, no distinction is made between the various possible simile structures which are treated exactly the same.

**Examples** (See Appendices 1A and 2.III for details about the part-of-speech tags and the dependency relations used)

## 1/ Coordination

# {"category": "Books", "figurativeness": [4, 4, 4], "title": "Ulysses", "price": "unknown", "userId": "A2IV0VON1EO9LE", "score": "1.0", "helpfulness": "24/57", "time": "1151884800", "profileName": "N. E. Cobleigh \"Fast Eddie\"", "productId": "0613175719"}

if	if	IN	1	13	VMOD	_
,	,	,	2	1	P	_
to	to	TO	3	8	VMOD	_
you	you	PRP	4	3	PMOD	_
,	,	,	5	8	P	_
a	a	DT	6	7	NMOD	_
book	book	NN	7	8	SUB	TOPIC
walks	walk	VBZ	8	1	SBAR	EVENT
like	like	IN	9	8	VMOD	COMPARATOR
a	a	DT	10	11	NMOD	_
duck	duck	NN	11	9	PMOD	VEHICLE
,	,	,	12	13	P	_
<b>looks</b>	<b>look</b>	<b>VBZ</b>	<b>13</b>	<b>0</b>	<b>ROOT</b>	<b>_</b>
<b>like</b>	<b>like</b>	<b>IN</b>	<b>14</b>	<b>13</b>	<b>VMOD</b>	<b>_</b>
<b>a</b>	<b>a</b>	<b>DT</b>	<b>15</b>	<b>16</b>	<b>NMOD</b>	<b>_</b>
<b>duck</b>	<b>duck</b>	<b>NN</b>	<b>16</b>	<b>14</b>	<b>PMOD</b>	<b>_</b>
,	,	,	17	13	P	_
<b>and</b>	<b>and</b>	<b>CC</b>	<b>18</b>	<b>13</b>	<b>VMOD</b>	<b>_</b>
<b>sounds</b>	<b>sound</b>	<b>VBZ</b>	<b>19</b>	<b>13</b>	<b>VMOD</b>	<b>_</b>
<b>like</b>	<b>like</b>	<b>IN</b>	<b>20</b>	<b>19</b>	<b>VMOD</b>	<b>_</b>
<b>a</b>	<b>a</b>	<b>DT</b>	<b>21</b>	<b>22</b>	<b>NMOD</b>	<b>_</b>
<b>duck</b>	<b>duck</b>	<b>NN</b>	<b>22</b>	<b>20</b>	<b>PMOD</b>	<b>_</b>
,	,	,	23	19	P	_
then	then	RB	24	19	VMOD	_
it	it	PRP	25	26	SUB	_
was	be	VBD	26	19	VMOD	_
most	most	RB	27	29	VMOD	_
likely	likely	RB	28	29	VMOD	_
written	write	VC	29	26	VC	_
by	by	IN	30	29	VMOD	_
a	a	DT	31	32	NMOD	_
quack	quack	NN	32	30	PMOD	_
.	.	.	33	13	P	_

2/ Wrong tags

```
# {"category": "Books", "figurativeness": [1, 1, 2], "title": "Rulers of the Darkness (The World at War, Book 4)", "price": "unknown", "userId": "A2EJP1CB7YGPNK", "score": "4.0", "helpfulness": "2/2", "time": "1098316800", "profileName": "Philip B. Yochim", "productId": "B0009WLSW8"}
```

The	the	DT	1	2	NMOD	_		
characters	character		NNS	2	3	SUB	TOPIC	
make	make	VB	3	0	ROOT	EVENT		
the	the	DT	4	5	NMOD	_		
<b>story</b>	<b>story</b>	<b>NN</b>	<b>5</b>	<b>7</b>	<b>SUB</b>	<b>_</b>		
more	more	RBR	6	7	AMOD	_		
interesting		interesting	JJ	7	3	VMOD	PROPERTY	
than	than	IN	8	7	AMOD	COMPARATOR		
the	the	DT	9	10	NMOD	_		
action	action	NN	10	8	PMOD	VEHICLE		
.	.	.	11	3	P	_		

The last corpus to be discussed in this section, the VUAMC (Vrije Universiteit Amsterdam Metaphor Corpus) Online,<sup>22</sup> is a manually annotated corpus which mainly deals with the metaphor in its broadest sense, and, therefore, devotes a rather small space to similes. This corpus is made up of fragments of academic texts, conversations, fiction and news taken from the BNC Baby, which is itself a subset of the British National Corpus (BNC). Consequently, the final output reuses the part-of-speech tags already present in the BNC.

At the basis of the identification of all these metaphorical linguistic units, lies MIPVU, which itself derives from the MIP (Metaphor Identification Procedure) (Pragglejaz Group, 2007). The MIP recommends four main steps to decide whether a lexical unit is used metaphorically or not: first, read the text to gather a general understanding of its content, identify each lexical unit, determine the meaning of each one of text in the text and compare it to its historical meaning and then label the unit as metaphorical if its current meaning differs from its basic meaning. If the MIPVU (Steen et al., 2010) still scan each word of the text to find out if it is used metaphorically or not, it also distinguishes between direct metaphor, implicit metaphors and words signalling metaphors also called “metaphor flags”. Comparison markers enter in this last category. All the signals considered for this annotation task are: “appearance”, “as”, “as...as”, “as if”, “as though”, “call”, “constitute”, “-ish”, “just as...so”, “like”, “-like”, “metaphorical”, “no more than”, “reminding”, “reminiscent”, “resembling”, “seemed”, “shaped”, “-shaped”, “so-called”, “some sort of”,

---

<sup>22</sup> <http://www.vismet.org/metcor/search/>

“sort of”, “symbolically”, “types”, “with the ... of a(n)...”. Even if the greater part of these signals introduces similes, some of them only precede an analogy or a metaphor. A filtering by signal words should, therefore, be done to extract only similes.

The corpus is searchable online<sup>23</sup> or can be freely downloaded as an XML file.<sup>24</sup> The online version proposes to search the corpus using metaphor-related words, signals or conceptual mappings. Similarly, in the XML file, in addition to the part-of-speech tags, each metaphorical word and metaphorical signals are tagged respectively with “mrw” and “mFlag”. If the corpus constitutes a good basis to study metaphoricity in general, it does not say much about the reason why a particular word is metaphorical or give information on the structure of the identified similes. An example of an annotated simile is presented in Figure 4.2, details about the part-of-speech tags used can be found in Appendix 1C.

**Figure 4.2 Example from the VUAMC Online**

```
<s n="65">
  <w lemma="the" type="AT0">The </w>
  <w lemma="effect" type="NN1">effect </w>
  <w lemma="be" type="VBZ">is </w>
  <w lemma="rather" type="AV0">rather </w>
  <w lemma="like" type="PRP">
    <seg function="mFlag" type="lex">like</seg>
  </w>
  <w lemma="an" type="AT0">an </w>
  <w lemma="extended" type="AJ0">
    <seg function="mrw" type="lit" vici:morph="n">extended</seg>
  </w>
  <w lemma="advertisement" type="NN1">
    <seg function="mrw" type="lit" vici:morph="n">advertisement</seg>
  </w>
  <w lemma="for" type="PRP">for </w>
  <w lemma="marlboro" type="NP0">
    <seg function="mrw" type="lit" vici:morph="n">Marlboro</seg>
  </w>
  <w lemma="light" type="NN2">
    <seg function="mrw" type="lit" vici:morph="n">Lights</seg>
  </w>
  <c type="PUN">.</c>
</s>
```

<sup>23</sup> <http://www.vismet.org/metcor/search/showPage.php?page=start>

<sup>24</sup> <http://ota.ahds.ac.uk/headers/2541.xml>

Annotated corpora are valuable resources for researchers as they enable to store knowledge that can be later retrieved and compared with other data. In this respect, some frameworks such as the TEI attempt to standardise these annotations in order to facilitate data reutilisation and exchange. As far as figurative language in general and similes in particular are concerned, no consensus has, however, been reached at, because of theoretical divergences. Consequently, literary practices of simile description which rely on syntactic (overall structure, nature of the ground, doubled simile) and semantic properties (type of marker, degree of abstraction, semantic categories...) have, for the moment, not been applied to the manual or automatic annotation of similes.



# *PART TWO*

## *FROM RAW TEXT TO STRUCTURED DATA*



## 5 THE PROPOSED APPROACH

Obviously, similes cannot be annotated automatically without at least prior information about the presence of a simile in a specific sentence and about the anatomy of that simile (tenor – ground – vehicle). However, such information can only be obtained after mining the text and separating potential similes from pseudo-comparisons and literal comparisons. In this respect, the proposed approach to the automatic annotation of similes in (literary) texts is made up of four main stages: extract all comparative constructions, identify their components, decide whether these constructions are literal or figurative, and annotate them accordingly. As stated in most of the previous research works presented in Chapter 3, despite comparison being construed by meaning, comparative constructions are subordinated to the specific syntax of the language in use, which makes them easily recognisable and makes it possible to generalise their composition to unseen structures.

The first part of this chapter deals with the syntactic structure of phrasal similes in English and French. In the second section, the three modules of the annotation system are reviewed in detail: the syntactic module, which is mainly concerned with the preprocessing tasks, the selection of potential simile candidates and the identification of each of their components, the semantic module which takes part in choosing the most plausible components in ambiguous cases and in distinguishing literal statements from figurative ones and, finally the annotation module which adds descriptive tags to similes based on the annotation framework designed for this purpose.

## 5.1 A Grammar of the Simile

Defining the grammatical category of the marker is essential to determine which syntactic patterns comparative constructions would follow. A quick look at definitions of the simile in English and French shows the existence of an Anglo-Saxon bias towards “like” and “as” as simile markers, as exemplified by the fact that most definitions in English both in non-specialised and in rhetorical dictionaries go along this line: “In a simile, a comparison between two distinctly different things is explicitly indicated by the word ‘like’ or ‘as’” (Abrams, 1999, p. 97). On the contrary, in French, although “comme” is generally presented as the prototypical marker, no definition of the simile is centred around the use of that specific marker. As proof of this Anglo-Saxon “reductionism” of the simile, Shabat Bethlehem (1996) observes that in about fifteen scholarly articles on similes, the general trend is indeed to restrict the scope of markers to “like” and “as” and when it is not the case, to not cite directly other markers but to put them under the rather vague “etc.” (p. 210-211).

Because of the semantic nature of comparisons and the conflicting views on what unequivocally constitutes a simile, it seems difficult to provide a full inventory of all the existing simile markers both in English and in French. To remedy this fact, authors either give examples of alternative forms of simile markers (Israel et al. 2004), classify similes according to their effects (Goatly, 2011) or regroup similes in clusters while specifying that the list of markers given is far from being finite (Bouverot, 1969; Pistorius, 1971; Moon, 2011). Grossly speaking, simile markers can be made up of a single word or of a whole phrase, acting as a verb or as a conjunction/preposition (see Table 5.1 A & B).

On the structure of similes, Quintilian (trans. 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” (Book, VIII. Chap III. 77, p. 105). In the latter case, one would easily recognise the prototypical simile of the type “The girl is as graceful as a lily” while the former case corresponds to elliptical similes such as in:

-What’s them plants, ma’am?

-Oh, those are chrysanthemums, giant whites and yellows, I raise them every year, bigger than anybody around here.

-Kind of a long stemmed flower? Looks like a quick puff of colored smoke?—he asked. (Steinbeck, as cited in Kirvalidze, 2004, p. 27).

If we also take into account the fact that similes can be open or closed, at the sentence level and including the marker, a simile may contain from two to four components:

two-component-similes: marker + vehicle

three-component-similes: tenor + marker + vehicle or ground + marker + vehicle

four-component-similes: tenor + ground + marker + vehicle

Although an ellipsis of the vehicle is linguistically possible and admitted by some scholars like Cohen (1968), it is not considered here as it seems to defeat the whole purpose of comparison. In addition, the verses that Cohen (1968) gives as example are far from being totally convincing as it is more a matter of implying the whole comparison than of simply omitting the vehicle: “Nous aurons des lits pleins d’odeurs légères, / Des divans profonds comme des tombeaux,/ **Et d’étranges fleurs sur des étagères, / Éclores pour nous sous des cieus plus beaux**.”<sup>25</sup>Baudelaire (1857, p. 243, “La Mort des amants”).

Table 5.1A Simile markers used as predicates: Grammatical patterns and examples

Corresponding Markers	Grammatical Patterns	Examples
Verbs Verbal phrases	marker + Vehicle verb complement noun-headed noun phrase (NH_NP)	<u>On dirait</u> ton regard d’une vapeur couvert; / Ton œil mystérieux (est-il bleu, gris ou vert?) / Alternativement tendre, rêveur, cruel, / Réfléchit l’indolence et la pâleur du ciel. (Baudelaire, 1857 p.107) La nuit <u>parut</u> une blessée. (Rodenbach, n.d., p.89)
	Tenor Subject of the marker + marker + Vehicle marker complement noun-headed noun phrase (NH_NP)	[...] je le lisais dans les gestes de toutes ces marionnettes bourgeoises [...], dans les moindres détails de cet affreux salon jonquille [...] que l’uniformité de ses soirées faisaient <u>ressembler à</u> un tableau à musique. (Daudet, 1909, p.246)

Tables 5.1.A & B give an overview of the possible structures that a simile can take, depending on the nature of the marker. In all the listed patterns, the vehicle is described as being both a complement and a noun-headed noun phrase. In this respect, it is implied that

---

<sup>25</sup> English translation: “We shall have beds round which light scents are wafted,/ Divans which are as deep and wide as tombs; / Strange flowers that under brighter skies were grafted / Will scent our shelves with rare exotic blooms. Roy Campbell, *Poems of Baudelaire* (New York: Pantheon Books, 1952)

the vehicle could not fulfil another function such as being the subject of another verb in the sentence as when in this case, the whole clause beginning with the vehicle would be the complement, creating a comparison between processes.

**Table 5.1B Simile markers used as conjunctions: Grammatical patterns and examples**

Corresponding Markers	Grammatical Patterns	Examples
Adjective phrases Conjunctions Prepositional phrases Noun phrases Affixes	Open similes	
	Marker + Vehicle complement NH_NP	Ça va très bien. J'ai dormi comme un prince. <u>Comme</u> un prince ! (Pagnol as cited in Cazelles, 1996, p.86)
	Tenor NP + marker + Vehicle complement NH_NP	Vous êtes hébété de fatigue. Sale. Les cheveux <u>comme</u> les poils d'un vieux balai. Les ongles cassés. (de Buron, as cited in Cazelles, 1996, p. 65)
	Tenor Subject of the verb+ verb + marker + Vehicle complement NH_NP	Ideas are <u>like</u> shadows - substantial enough until we try to grasp them. (Butler, as cited in Wilstach, 1916, p. 208)
	Closed similes	
	Verbal phrase ground + marker + Vehicle complement NH_NP	- Très important... de ne pas s'accrocher <u>comme</u> une nouille ! (de Buron, as cited in Cazelles, 1996, p. 10)
	Tenor Subject of the verbal phrase ground + verbal phrase ground + marker + Vehicle complement NH_NP	He leaped <u>like</u> a man shot. (Stevenson, as cited in Wilstach, 1916, p. 229)
	Adjectival ground + marker+ Vehicle complement NH_NP	Le dénommé Marc entre. Immense, très maigre, blanc <u>comme</u> un ver de pomme, une curieuse coiffure, - des cheveux rasés sur les côtés mais une longue mèche désordonnée lui recouvrant le front et même le nez. (de Buron, as cited in Cazelles, 1996, p. 50)
	Tenor NP modified by the adjectival ground + adjectival ground + marker + Vehicle complement NH_NP	Le cuivre, sous l'effet de la chaleur, fondait et coulait en ruisseaux rouges frangés de scories spongieuses et dures <u>comme</u> de la pierre (Vian, as cited in Cazelles, 1996, p. 91)
	Tenor Subject of the verb + verb + adjectival ground +marker + Vehicle complement NH_NP	Her eyes are grey <u>like</u> morning dew (Yeats as cited in Wilstach, 1916, p. 186)
Tenor Object of the verb+ verb + adjectival ground +marker + Vehicle complement NH_NP	La bienveillance de madame Vieuxnoir avait affranchi ce garçon et le rendait hardi <u>comme</u> un coq. (Duranty, as cited in Cazelles, 1996, p. 95)	

### **Examples**

I suspect that there's in an Englishman's brain **a valve that can be closed as pleasure, as an engineer shuts off steam** (Emerson, as cited in Wilstach, 1916, p. 31).

==> The comparison here is between the manner in which a valve in an Englishman's man brain can be closed and the manner an engineer shuts off steam.

I suspect that there's in an Englishman's brain **a valve that can be closed as pleasure, as steam on an engine**.

==> The comparison here it is between the valve and the steam with respect to the fact that they can be switched off.

For more readability of the Tables, all forms follow the canonical syntactic order of each language: Subject – Verb (– Object). By subject, it is understood here the entity that typically does the action expressed by the verb which, consequently, can be conjugated or be used in the infinitive form.

### **Examples**

1/ Infinitive verb

On sentait le froid emmagasiné **refluer** entre les jambes ainsi que de l'eau glacée. (Carrière, as cited in Delabre, 1984, p. 15)

2/ Present participle

I was running on, very fast indeed, when my eyes rested on little Em'ly's face, which was bent forward over the table, listening with the deepest attention, her breath held, her blue eyes **sparkling like** jewels, and the color mantling in her cheeks. (Dickens, *DC*, as cited in Tomita, 2008a, p. 9)

It is worth noting, however, that the subject may be inverted, the marker and the vehicle could be inserted between the subject and the verb, or the marker and the vehicle could be detached before the tenor and the verb with or without the adjective ground.

### **Examples**

\* Inverted subject

Jusqu'au format, oiseux; et vainement, concourt **cette extraordinaire, comme** un vol recueilli mais prêt à s'élargir, intervention du pliage ou le rythme, initiale cause qu'une

feuille fermée, contienne un secret, le silence y demeure, précieux et des signes évocatoires succèdent, pour l'esprit, à tout littérairement aboli.<sup>26</sup> (Mallarmé, 1897, p. 275)

\*Marker and vehicle inserted between the tenor and the ground:

Fame, like a new mistress of the town, is gained with ease, but then she's lost as soon. (Dryden, as cited in Wilstach, 1916, p. 132)

\* Marker and vehicle detached at the beginning of the clause:

Children are never too tender to be whipped; like tough beef-steaks, the more you beat them the more tender they become. (Poe, as cited in Wilstach, 1916, p. 52)

\*Ground, marker and vehicle placed before the tenor:

Également blanche comme neige, une barbe de fleuve, divisée en deux branches, descendait sur le gilet de velours noir à fleurs grenat. (de Vogüé, as cited in Cazelles, 1996, p. 51)

In addition, because of its impact on the meaning of the simile structures, the presence of the direct object of the verbal ground could not be simply ignored as it is the case in the methods described in Chapter 3. Therefore, a simile which makes use of a verbal ground to compare two entities may contain a direct object whose only role is to restrict the meaning of the verb. Consequently, direct objects of the identified verbal grounds must also be identified alongside the other simile components, and a filtering stage must occur in which the most plausible meaning would prevail: do the vehicle semantically replace the subject of the verb or its object? This question is certainly far from being trivial as it could help in distinguishing similes involving entities of the form Tenor NP + marker + Vehicle complement NH\_NP from similes involving processes with the structure Verbal ground + Verb Object + marker + Vehicle complement NH\_NP. Furthermore, as in the example below, instead of the verb subject, the verb direct object can be the true tenor of a simile comparing entities.

### Example

Ce drôle a les **jambes** comme des pincettes. (Bertrand, 1842, p. 75)

---

<sup>26</sup> English translation: And since even the book's format is useless, of what avail is that extraordinary addition of foldings (like wings in repose, ready to fly forth again) which constitute its rhythm and the chief reason for the secret contained in its pages? Of what avail the priceless silence living there, and evocative symbols following in its wake, to delight the mind which literature has totally delivered? (Reynard Seifert, *Nothing ever happens*, <http://htmlgiant.com/excerpts/nothing-ever-happens/>)

In this respect, the various positions that can occupy the direct object of the verbal ground of a simile must be listed:

- it could be placed after the verb and generally before the marker if the latter does not precede the verb.

### **Examples**

Books, like men their authors, have **no more than one way of coming into the world**, but there are ten thousand to go out of it and return no more. (Swift, as cited in Wilstach, 1916, p. 29)

Le marchand d'antiquités fredonnait un contre-chant d'une simplicité pastorale et balançait **sa tête** de côté comme un serpent à sonnettes. (Vian, as cited in Cazelles, 1996, p.33)

Wear **your leaning like** your watch, in a private pocket, and do not pull it out and strike it merely to show that you have one. (Chesterfield, as cited in Wilstach, 1916, p. 229)

- it could be placed before the verb if the direct object is a personal or a relative pronoun

### **Example**

Les heures vides [...] étaient devenues vraiment vides parce qu'elle ne l'attendait plus comme un miracle mais comme une habitude. (Sagan, as cited in Cazelles, 1996)

Elle resta perdue de stupeur, et n'ayant plus conscience d'elle-même que par **le battement de ses artères**, qu'elle croyait entendre s'échapper comme une assourdissante musique qui emplissait la campagne<sup>27</sup>. (Flaubert, 1885, p. 360-361).

- it could be placed after the vehicle

### **Example**

J'aime Dijon comme l'enfant **sa nourrice** dont il a sucé le lait, comme le poète **la jouvencelle qui a initié son cœur**. (Bertrand, 1842, p.3)

---

<sup>27</sup> She remained lost in stupor, having no longer consciousness of herself, excepting through the pulsation of her arteries, which she thought she heard escaping from her like a deafening music which filled the country side (Flaubert, 1896, Vol II, p. 129)

There also exists another form of elliptical subordinate clause denoting similes involving processes, in which it is the preposition that follows the verbal ground which is repeated, generally with a new complement.

### **Example**

I live **in** the town like a lion **in** his desert, or an eagle **in** his rock, too great for friendship or society, and condemned to solitude by unhappy elevation and dreaded ascendancy. (Dr Johnson, as cited in Wilstach, 1916, p. 239)

Apart from the direct object, often to achieve particular stylistic effects, the tenor subject of the verb or the adjectival ground could also be placed after the vehicle.

### **Examples**

Des sifflements de mort et des cercles de musique sourde font monter, s'élargir et trembler comme un spectre **ce corps adoré**, des blessures écarlates et noires éclatent dans les chairs superbes.<sup>28</sup> (Rimbaud, 1922, p. 97)

His glance was like a gimlet, **cold and piercing** [Son regard était une vrille, cela était froid et cela perçait] (Hugo, as cited in Wilstach, 1916, p. 173).

Finally, the absence of the direct object does not necessarily imply that the subject of the verb, if present, is compared to the vehicle as in some elliptical sentences, a transitive verb could be used without any direct object to convey an idea of approximation.

### **Examples**

J'entendis Ø comme des soupirs et des sanglots, tandis que la flamme, livide maintenant, décroissait le foyer attristé. (Bertrand, 1842, p. 150)

So far, the different patterns of similes have been described using their grammatical function; this approach, however, suffers from the fact that various syntactic orders and scenarios are possible. Alternatively, to represent the mechanisms at work in these patterns without the boundaries imposed by the linear, dependency grammar can be used.

---

<sup>28</sup> English translation: Whistling of death and the circling of faint music make this adored body rise, expand and quiver like a spectre; wounds of scarlet and black burst from superb flesh (trans. A. S. Kline, 2002, [http://www.poetryintranslation.com/PITBR/French/Rimbaud2.htm#anchor\\_Toc202067618](http://www.poetryintranslation.com/PITBR/French/Rimbaud2.htm#anchor_Toc202067618)).

According to Tesnière (1959), each word in a sentence is linked to its neighbouring words by a series of links which constitute the skeleton of the sentence. Each of this link is hierarchical in the sense that there is always a term A that depends on the other term B, in the sense that A, also called the dependent or the modifier may be optional while B, the head, regent or governor, is compulsory and also determines the form and the linear position of A (Nivre, 2005). In a simple sentence of the form Subject – Verb – Object, the subject and the object will typically be the dependents of the same governor, the verb.

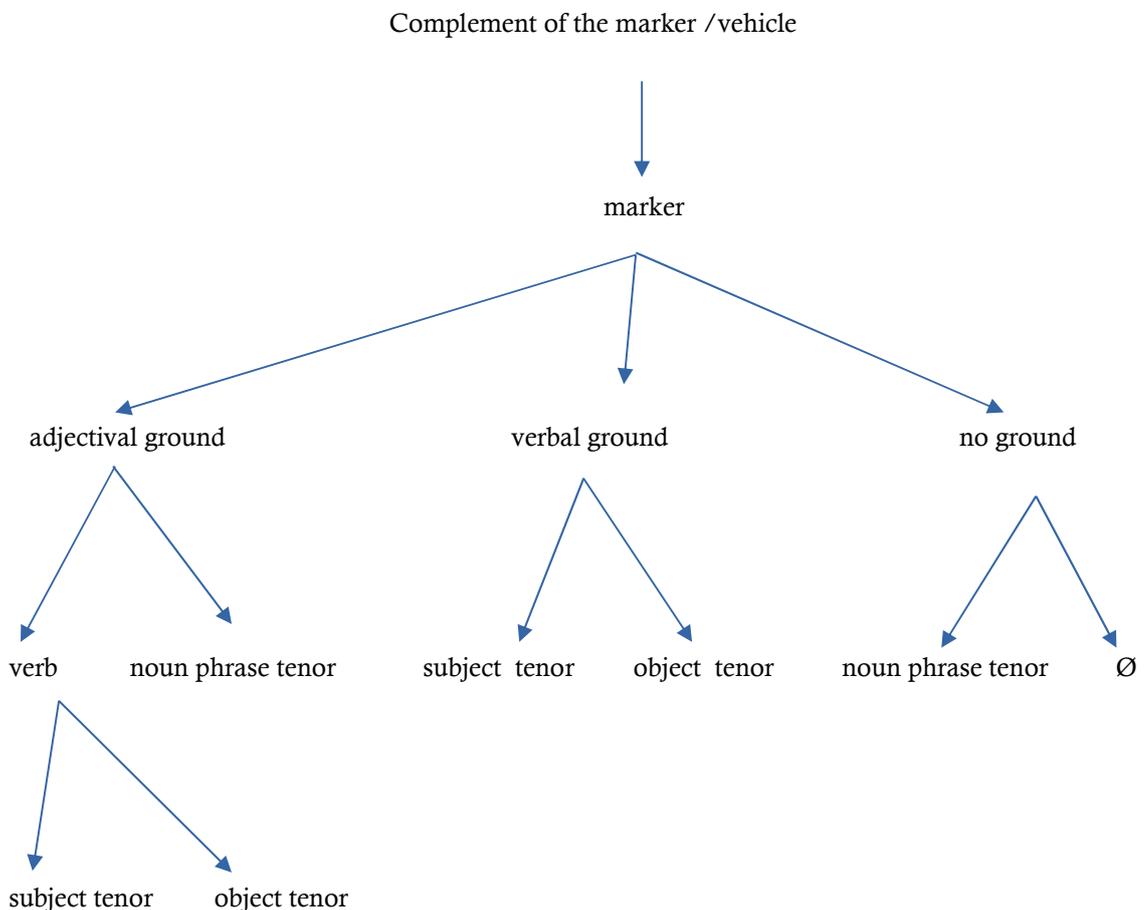
Figure 5.1. summarises the main simile syntactic structures. These structures need, of course, to be slightly adapted when the marker is a suffix combined with the vehicle of the form “noun + -like” or “noun + colour term”.

### Examples

With a stealthy, leopard-like pride Ciccio went through the streets of London in those wild early days of war. (Lawrence, 1921, p. 323).

The lovely translucent pale irises, tiny and morning-blue, they lasted only a few hours. (Lawrence, 1921, p. 373).

**Figure 5.1 Possible syntactic structures of similes**



With regard to the different points aforementioned, it seems important for any system that pretends to recognise similes in raw texts to be flexible enough to take into consideration the various possibilities offered by each sentence and to select the most adequate one. For instance, in French, in front of a sentence with the structure subject + verb + noun phrase object + adjective + marker + vehicle (“Il récitait des phrases inintelligibles telles des puits de sagesse”), three interpretations are possible:

- 1/ the verbal (“récitait des phrases inintelligibles”) is the ground and the subject (“Il”) is the tenor,
- 2/ the adjective (“inintelligibles”) is the ground and the object (“phrases”) is the tenor,
- 3/ the simile compares two processes “récitait des phrases inintelligibles” and “récitait + “des puits de sagesse”.

In this respect, two distinct dimensions must come together to analyse such types of similes: a syntactic one which identifies the comparative statement and its components based on the pattern(s) it matches and a semantic one which determines the most plausible interpretation semantically as well as assesses how figurative it is.

## 5.2 The Syntactic Module

As far as texts are concerned, different levels of analysis are possible, the word-level, the sentence-level, the paragraph-level... In this study, with regard to simile automatic detection, the search has been restricted to the sentence-level, even though it can be argued that in most elliptical similes of the form marker+vehicle, the scope of the simile largely goes beyond the boundaries of a single sentence (see the Steinbeck example, p. 102). Of course, the texts first need to be preprocessed in order to facilitate the retrieval of new information. Some of these basic preprocessing tasks include: tokenisation, lemmatisation, sentence segmentation and part-of-speech tagging. To determine sentence boundaries, both the type of punctuation and its context of usage must be taken into consideration. Whereas the comma and the semi-colon never signal the end of the sentence, the full stop can also be used in abbreviations (Mr., i.e.), software names (Python 2.7.10) or numbers (40.6%). If natural language processing can disambiguate the full stop with more or less success, it is far from being the case for other equally challenging punctuation marks such as the ellipsis, the question or the interrogation mark. In this respect, these punctuation marks have been taken as marking the end of a sentence when the next segment starts with a capital letter. This rule would not, however, work in some cases such as the example below.

### **Example**

The foreign policy of France, like its cuisine, should be unmistakably, ineffably . . . French.<sup>29</sup>

Of course, a crucial point before starting to look for similes in texts is to define the simile markers to be considered. In order to be able to grasp a wide range of similes and in accordance with practices in literary stylistics, different categories of markers were taken into account. Because of the semantic nature of comparisons in general and similes in particular, it appears difficult to compile an exhaustive and finite list of simile markers. In this respect, a first list of simile markers in English and French was drawn out by compiling markers cited in existing research works on metaphors and similes. Then, the synonyms of these markers and in some cases, words of the same family were added; for example, from the adjective phrase “comparable à” and “semblable à” mentioned by Bouverot (1969), it is possible to add to the list “comparer à”, “identique à” and “similaire à”.

After the first experiments on test data, the list was narrowed down according to the figurative potential of each marker, i.e. its ability to reunite both a tenor and a vehicle in a simile. Consequently, forms with implicit tenors such as “on dirait ...” did not make the final list. Similarly, to comply with this rule, some markers were amended. For example, “a kind of” was transformed into “be / become + (determiner) + kind of”. Other markers found in the literature such as “noun + -shaped” or “en forme de” were judged too narrowed in meaning to be used figuratively and were therefore removed from the initial list. Moreover, verbs such as “rappeler”, “simuler” and “paraître” were judged too polysemous to be part of the final list. All the remaining markers can be found in Table 5.2.

---

<sup>29</sup> Examples taken from the Similepedia Blog (<http://similepedia.blogspot.fr/>)

Table 5.2 Selected simile markers for both languages

	Comparatives	Verbal phrases	Adjectival phrases	Prepositional phrases
<i>English</i>	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	with the ... of a/an
<i>French</i>	comme, ainsi que, de même que, autant que, plus...que, tel que, moins...que aussi...que	ressembler à, sembler, 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

For a sentence to be extracted as a simile candidate, it needs to fulfil the following criteria:

- it must contain at least one marker or a variant of a marker which is either directly followed by a noun-headed noun phrase or separated from the noun-headed noun phrase it introduces by a parenthetical expression. A variant of a marker in this case refers both to inflected verb forms and to slight alterations that are often made to noun phrase and verb phrase markers through the addition of an adjective phrase and an adverbial respectively, for example using “different kind of” instead of simply “kind of”.

- the noun head of the noun phrase which completes the marker must not also be the subject of a verb. This last condition particularly holds for comparatives and for noun phrase markers and is useful to eliminate both clausal similes and non-relevant uses of polysemous markers such as temporal clauses introduced by “as” and “comme”. Are typically considered as subjects, head nouns that are not separated from a conjugated verb by a personal pronoun subject, a relative pronoun subject, a coordinating or a subordinating conjunction. Some false positives, however often occur with coordinating conjunctions and with the past participle in English which is sometimes wrongly tagged as the past tense, especially after “like”.

### Examples

When he suddenly remembers to smile -- as he did, quite awkwardly, outside No. 10 -- his face bursts into an unnatural glare, like [a fluorescent light] flicked on in a dark room, as opposed to the warm, glowing grin of Blair. (Similepedia Blog)

Falsehood, like the dry-rot, flourishes the more in proportion as [air] and light are excluded. (Whately as cited in Wilstach, 1916, p. 132)

Once a simile candidate has been found, the marker, the vehicle and the vehicle word phrase are extracted. All these steps are summarised in Figure 5.2.

**Example**

He looked pretty good, with a pair of cheeks like [big fat juicy **apples**].

\*\* Presence of a marker (“like”)

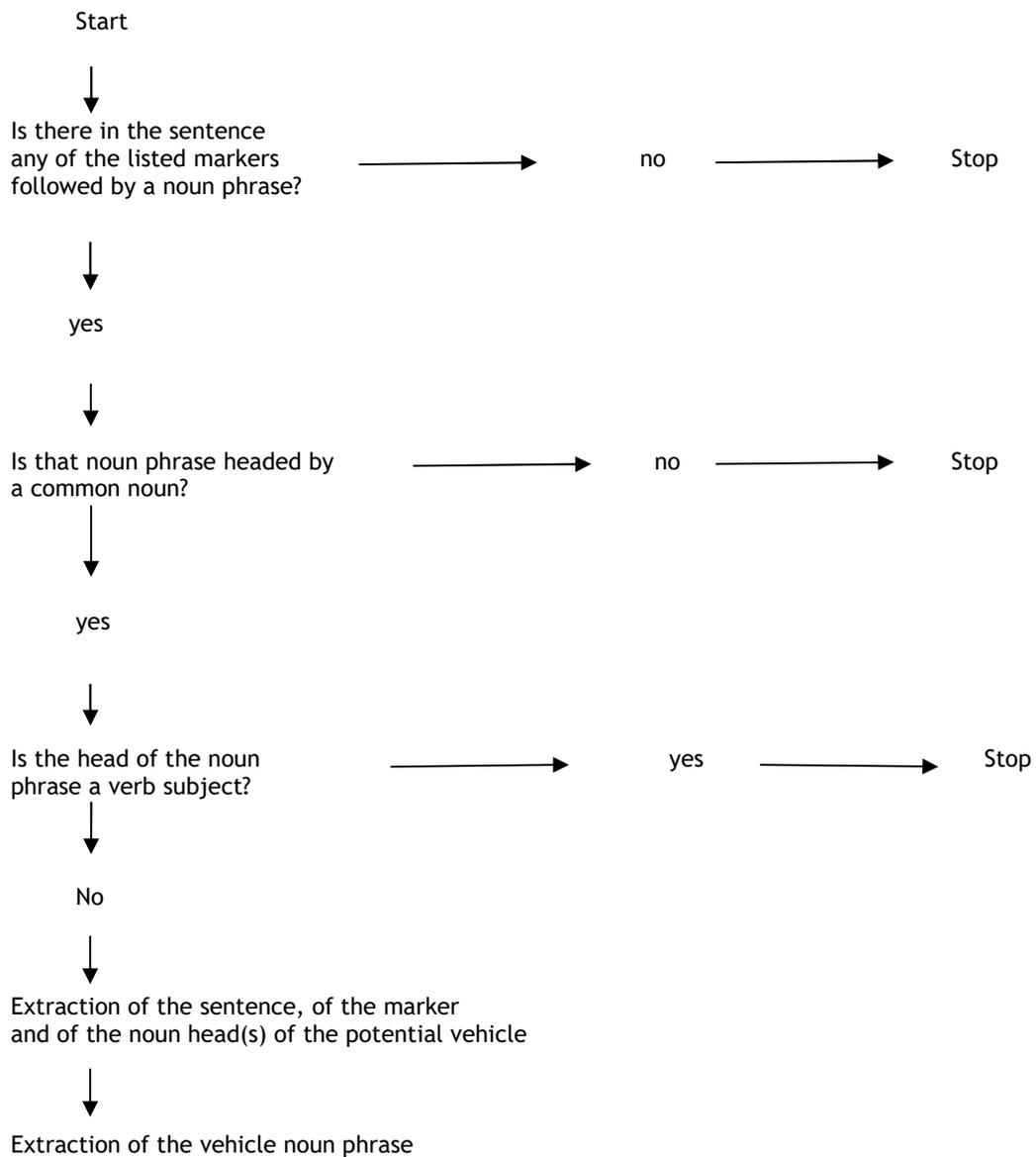
\*\* The marker is followed by a noun phrase (“big fat juicy apples”)

\*\* That noun phrase is headed by a common noun (“apples”)

\*\* That noun is not a subject

→ The sentence is considered a simile candidate.

**Figure 5.2 Extraction of potential simile candidates**

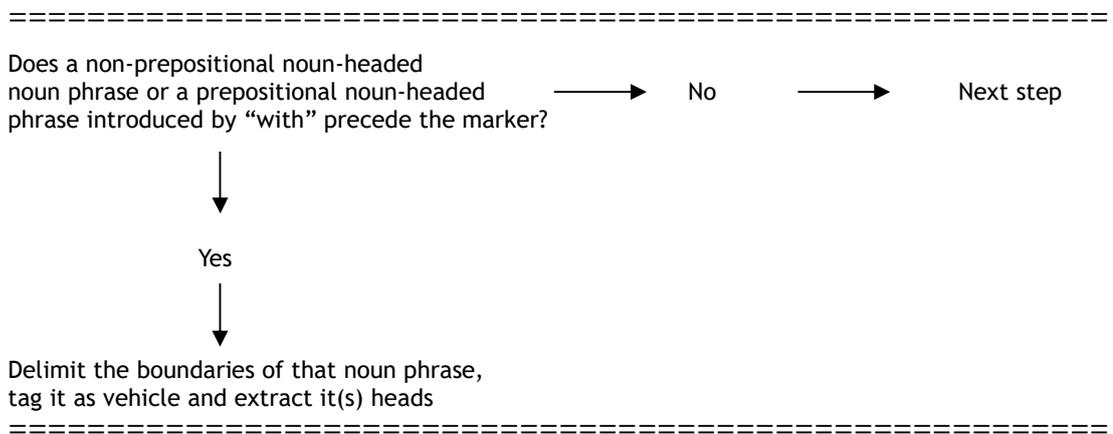


Then, the other simile components are identified respecting a specific order:

1/ the noun tenor phrase and its head(s): the noun tenor phrase is defined as any noun-headed noun phrase which immediately precedes the marker, is not preceded by any proposition except “with” and is neither the direct object of the first verb form, if any, left to the marker nor the subject of any conjugated form, on the right side of the marker.

### Example

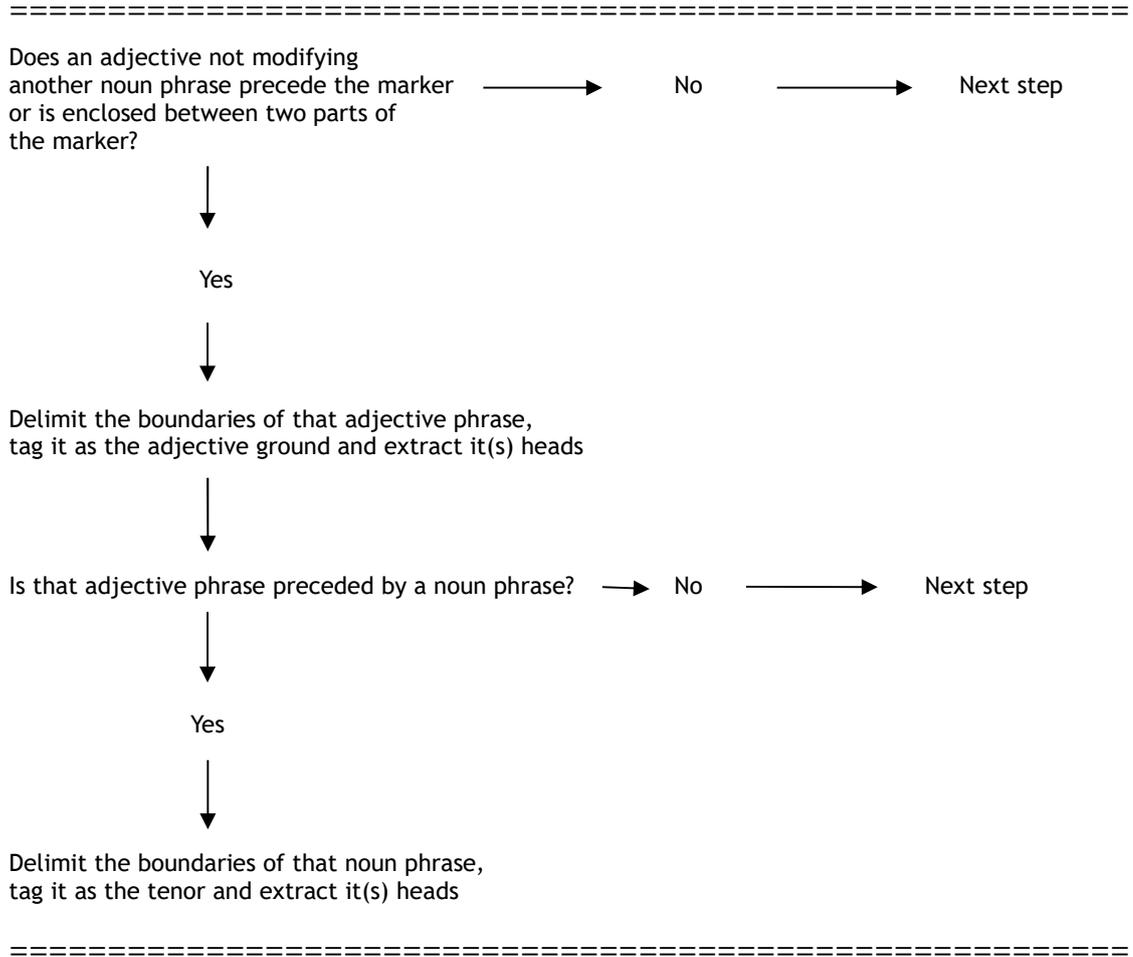
Heaped onto the street and the sidewalk are tons of the flimsy stuff of American housing – fiberglass insulation like poisonous cotton candy; sheets of warped plywood; mock-pine pressed sheathing; pulverized plasterboard [...] (Similepedia Blog).



2/ the adjectival phrase ground and its head which can have three functions according to both the English and French grammar:

- an attributive adjectival ground that modifies a nominal tenor which is not a verb subject
- a predicate adjective that is linked to the tenor noun phrase subject or object of the verb
- an appositive adjective often linked to the tenor noun phrase subject of the verb.

The two last tenors are only sought when looking for the first type of tenor did not yield any result. Of course, it implies that the verb that links the adjective and the phrase it modifies has previously been identified. By default, predicate adjectives that occur immediately after the noun the modifies are considered as being attributive as in “He thought [the painting] **ugly** like hell”.

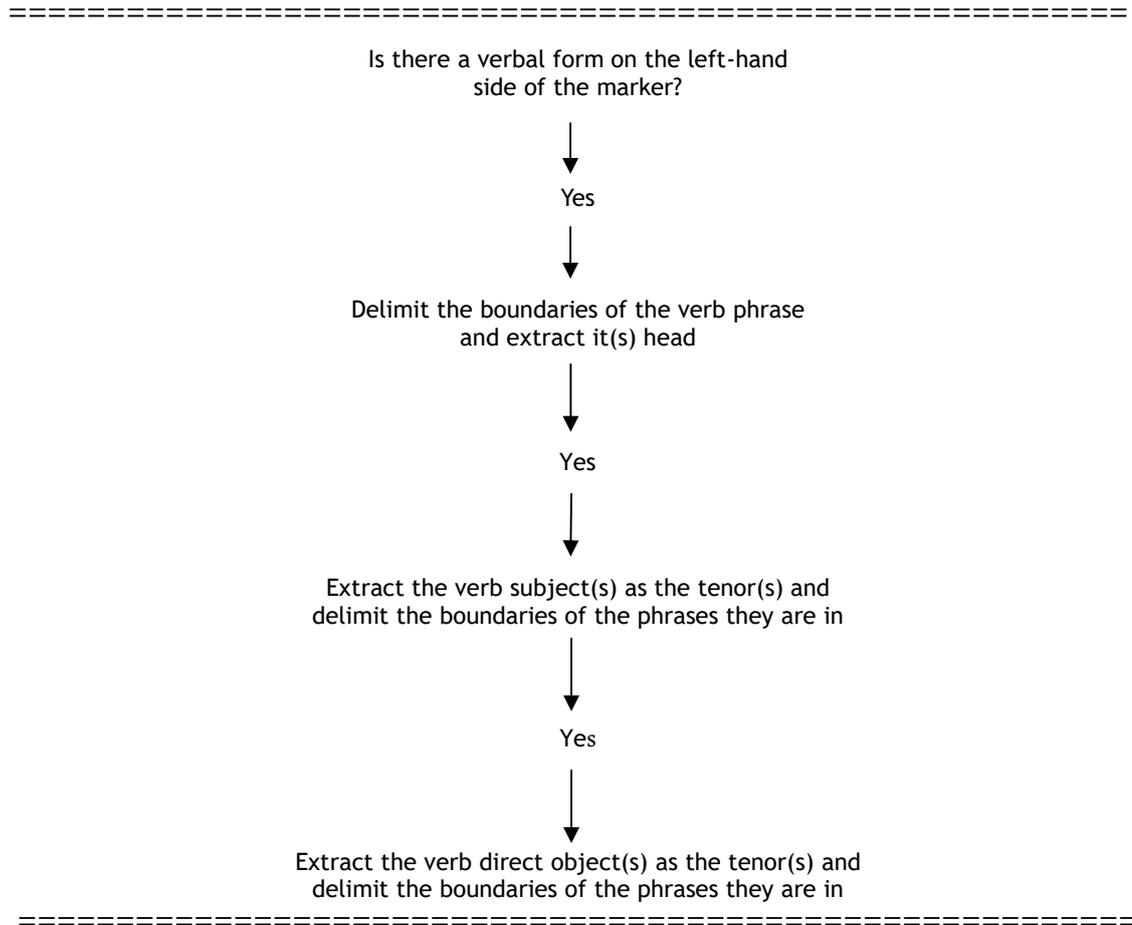


3/ the verbal ground enables to detect other sentence constituents:

- the direct object of the verb;
- and the tenor(s) noun or verb phrase subject.

**Example**

[Trying] to describe it is a bit like four blind men trying to describe an elephant... (Similepedia Blog).



When the subject or the direct object is a relative pronoun, instead of marking the pronoun as the subject, it seems more beneficial for the sentence analysis to rather search for its antecedent.

Generally speaking, coordination plays a crucial role in the detection task, principally because all components can be coordinated and therefore must be retrieved accordingly but also because, in the case of coordinated clauses, it is necessary to follow up the chain of coordinated verbs to be able to find their common subject. In addition, sometimes, in such types of constructions, the direct object can be attached to one of the verbs, which is not necessarily the closest one to the marker.

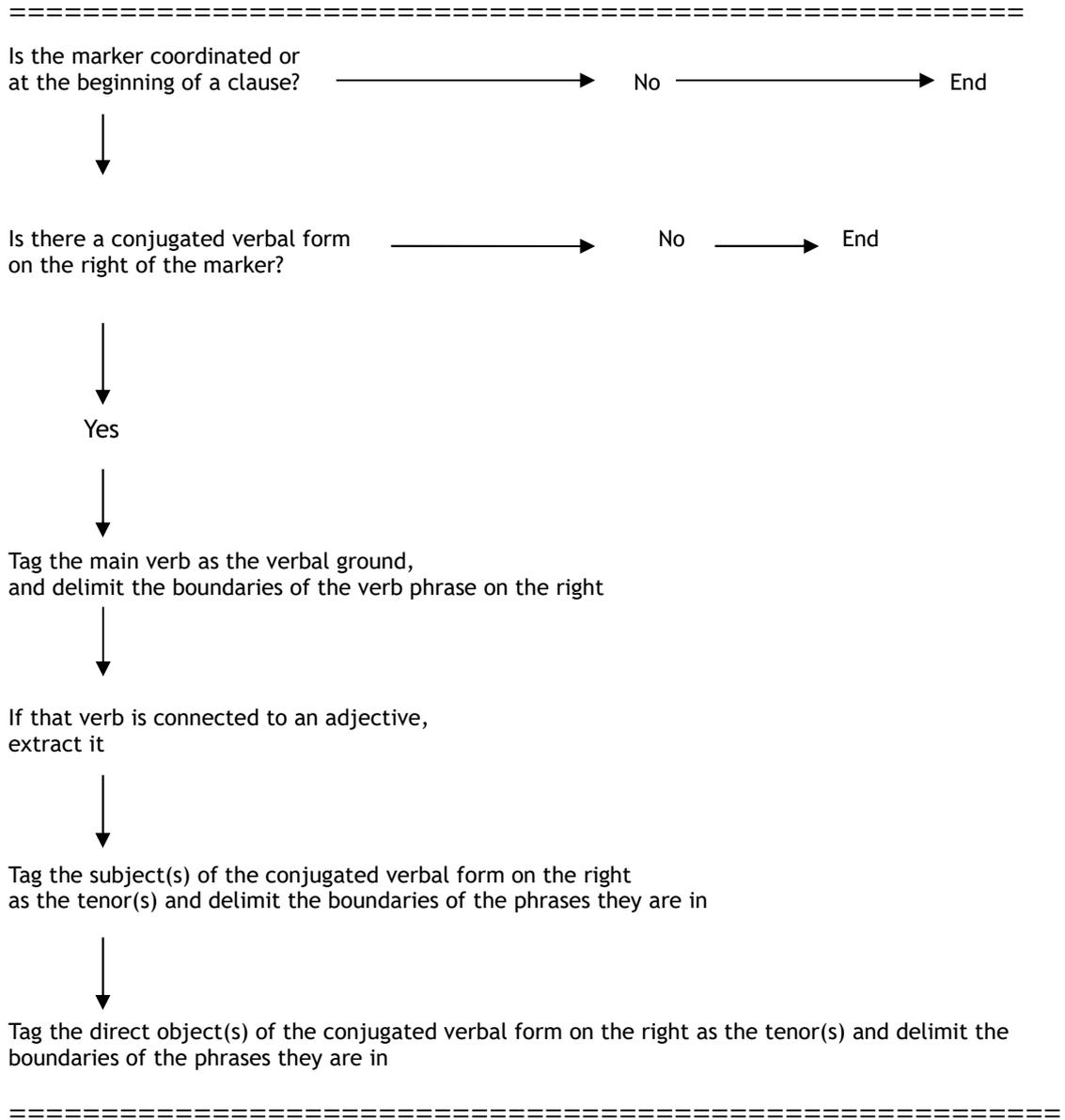
### Example

Je *l'*ai laissé se **multipliant**, **ramassant** les balles que Madame de Ligny rate à chaque coup, **courant** comme un perdu, ruisselant et ravi (Gyp as cited in Cazelles, 1996, p.72).

In case of coordinated marker, it is often necessary to search for a ground and a tenor on the right of the marker as a coordinated marker can indicate an inversion.

**Example**

But with puberty divergence begins; and, like the radii of a circle, (we) [**go** further and further apart]. (Schopenhauer as cited in Wilstach, 1916, p. 96)

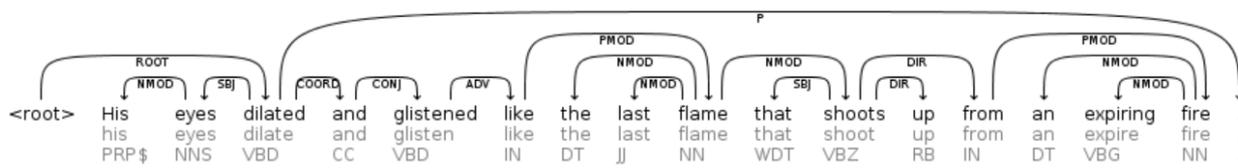


Even though, the verb subject is identified in the last stage, finding a verbal form is a necessary step to create a search interval in which to look for the noun tenor and the adjectival ground. As such, the presented algorithm does not take into consideration grounds which are clauses or noun phrases and those who occur outside the interval that starts with the verb.

Obviously, the fact that most simile components can be identified by their grammatical function explains why previous research works (Niculae & Yaneva, 2013; Niculae, 2013) have predominantly used dependency parsing. Based mainly on dependency grammar, dependency parsing represents a sentence structure as a dependency tree consisting of a unique root, generally the main verb, and of links connecting each head to its dependents (Covington, 2001).

### Example

Dependency tree of the sentence “His eyes dilated and glistened like the last flame that shoots up from an expiring fire.”<sup>30</sup>



From the tree above, it is easy to see that “like” is the head of “flame” and is itself a dependent of “glistened”. Furthermore, “glistened” and “dilated” are connected through the coordinating conjunction “and”, which implies that they share the same subject “eyes”. It can therefore easily be deduced that while “flame” is the head of the noun phrase vehicle, “glistened” is the verbal ground and “eyes” is the head of the noun phrase tenor. Once the head has been determined, dependency parsing can also be used to reconstruct the whole phrase under scrutiny by taking into consideration grammatical restrictions and the elements labelled as dependents of each identified head. For example, although the head of noun phrase subject depends on the verb, it cannot be considered as being part of the verb phrase unlike the noun phrase direct object which is itself also a head.

If this dependency tree enables to accurately visualise the different relations existing between the words of a sentence, it makes it difficult to directly retrieve the head and the dependents. In this respect, another more query-friendly output, the CoNLL data format is often used for computational purposes (see Table below; cf. Appendices 1A and 2.III for details about the tags).

<sup>30</sup> Obtained with Mate Tools (<http://en.sempar.ims.uni-stuttgart.de/>).



**Table 5.3 CoNLL output for the sentence “His eyes dilated and glistened like the last flame that shoots up from an expiring fire.”**

ID	FORM	LEMMA	CPOSTAG	PHEAD	PDEPREL
1	His	His	PRP\$	2	NMOD
2	eyes	eye	NNS	3	SBJ
3	dilated	dilate	VBD	0	ROOT
4	and	and	CC	3	COORD
5	glistened	glisten	VBD	4	CONJ
6	like	like	IN	5	ADV
7	the	the	DT	9	NMOD
8	last	last	JJ	9	NMOD
9	flame	flame	NN	6	PMOD
10	that	that	WDT	11	SBJ
11	shoots	shoot	VBZ	9	NMOD
12	up	up	RB	11	DIR
13	from	from	IN	11	DIR
14	an	an	DT	16	NMOD
15	expiring	expire	VBG	16	NMOD
16	fire	fire	NN	13	PMOD
17	.	.	.	3	P

Of course, the correctness of the dependency links is of utmost importance for the identification of the simile components. Different reasons could explain why some dependencies could be wrongly established: part-of-speech tagging errors, the fact that the dependent is considered to play another grammatical function in the sentence which takes precedence as well as non-linear constructions, parenthetical expressions or other types of long distance dependency. In addition, the final output of a dependency parser often lacks flexibility: for instance, in the example below, based on our definition of the noun tenor, despite the link connecting the marker “like” to “war”, “war” cannot be identified as the noun tenor. Subsequently, by following the links from “war”, it can be assumed that “like” is connected to the verb “snaked” which has for subject “that”, but the trail ends there, as despite the link connecting the verb “snaked” to its real subject “river”, no conclusion can be made solely by using this parser output.

**Example** (The tags used are detailed in appendices 1A and 2.III)

1	Weeks	week	NNS	2	AMOD
2	away	away	RB	21	SUBJ
3	and	and	CC	2	COORD
4	hundreds	hundred	NNS	3	CONJ
5	of	of	IN	4	NMOD
6	miles	mile	NNS	5	PMOD
7	up	up	IN	3	DEP-GAP
8	a	a	DT	9	NMOD
9	river	river	NN	7	AMOD
10	that	that	WDT	11	SBJ
11	snaked	snake	VBD	9	NMOD
12	through	through	IN	11	ADV
13	the	the	DT	14	NMOD
14	war	war	NN	12	PMOD
15	like	like	IN	14	NMOD
16	a	a	DT	19	NMOD
17	main	main	JJ	18	NMOD
18	circuit	circuit	NN	19	NMOD
19	cable	cable	NN	15	PMOD
20	-	-	:	9	P
21	plugged	plug	VBD	0	ROOT
22	straight	straight	RB	23	PMOD
23	into	into	IN	21	DIR
24	Kurtz	kurtz	NNP	23	PMOD
25	.	.	.	21	P

As a matter of fact, for the purpose of simile component identification, dependency parsing seems to lack flexibility in the case no link exists between the marker and the ground. In the first stage, to remedy the shortcomings of dependency parsing, syntactic chunking was combined with hand-crafted rules. Also called shallow parsing, syntactic chunking is often presented as an alternative to full parsing and delimits the boundaries of each phrase, making it possible to infer the grammatical relations between them with a set of rules.

**Example**

TreeTagger chunker (Schmid,1994) output for the sentence “His jealousy rises and falls like the wind.” (For information on the tags used, see appendices 1A and 2.II)

```

<NC>
His    PP$    his
jealousy NN    jealousy
</NC>
<VC>
rises  VVZ    rise
and    CC    and
falls  VVZ    fall
</VC>
,
<PC>
like  IN    like
<NC>
the   DT    the
wind  NN    wind
</NC>
</PC>
.     SENT .

```

Obviously, one of the main advantages of syntactic chunking is the fact that it marks phrase boundaries, which makes it easy both to identify the head of a phrase and to directly take into account these boundaries during the automatic analysis. However, unlike dependency parsing, syntactic chunking does not provide the grammatical function of the sentence words. Consequently, a set of rules and definitions has been made to be able to identify each simile component using textual clues.

**Table 5.4 Correlation between each type of constituent, the clues to identify it and its grammatical function**

Constituent	Grammatical category	Informative Clues	Governor
Adjectival ground	Adjective, past or present participle	Not separated from the marker by a coordinating conjunction, a relative pronoun, a preposition or a noun phrase	/
Tenor - head of the noun phrase that the adjectival ground modifies	Noun	Part of the noun phrase before or after the adjective	Non-predicative adjectival ground
Tenor - head of the noun phrase		Not after a preposition Head of the noun phrase directly before the marker	/
Tenor - Postposed direct object		Not after a preposition Follows a verb or a prepositional phrase that follows a verb	Verb
Tenor - Preposed direct object antecedent of a relative pronoun		Part of the noun phrase directly before “que”, “that”, “which” and the subject	
Tenor - objective personal pronoun (direct object)	Personal and demonstrative pronouns	Directly before a verb	
Tenor - subjective personal pronoun		Directly before or after a conjugated verb	
Tenor - subject head of the noun phrase	Noun	Before a verb and not after a preposition	
Verbal ground	Verb	Not separated from the marker by a colon or a semi-colon	/
Copular verb	Verb		Predicative adjectival ground
Vehicle - common noun	Common noun	Separated from the verb that follows it by a punctuation mark, a relative pronoun subject, a subjective personal pronoun, a coordinating or subordinating conjunction	Marker

The chunking-based implementation was tested on a corpus of French prose poems written by four authors: Aloysius Bertrand (1807-1841), Charles Baudelaire (1821-1867), Stéphane Mallarmé (1842-1898), and Arthur Rimbaud (1854-1891). Table 5.5 give some details about the size of that corpus and the number of simile candidates it contains. This number does not correspond to the number of comparative sentences, but to the number of markers followed by a non-subject noun-headed noun phrase. In this respect, a sentence such as

“Les étoffes parlent une langue muette, comme les fleurs, comme les ciels, comme les soleils couchants” is taken as three distinct simile candidates.

**Table 5.5 Size of the corpus of French prose poems**

Authors	Number of sentences	Number of tokens	Number of simile candidates
Aloysius Bertrand	1,167	25,298	67
Stéphane Mallarmé	1,746	92,661	44
Charles Baudelaire	1,153	41,299	126
Arthur Rimbaud	1,379	24,608	26
	5,445	183,866	262

TreeTagger (Schmid, 1994), a part-of-speech tagger that relies on decision trees, was used for tokenisation, part-of-speech tagging and syntactic chunking. The results obtained were compared on the manually annotated corpus with an improved version of the method based on dependency parsing described in Niculae (2013). Improvements mainly consisted in capturing a wider range of markers, subjects as well as direct objects, antecedents of relative pronouns and subjects of coordinated verbs. To parse the corpus in French, the Berkeley Parser (Candito, Nivre & Anguiano, 2010) was used.

The performance of both methods is detailed in Table 5.6. It can be said on the overall that the algorithm that relies on chunking and rules yields better results than the one based on dependency parsing.

In each of the simile candidates, the algorithms looked for the marker, the vehicle, the adjective ground, the verbal ground, the linking verb that is connected to a predicate adjective, and the vehicle. In this respect, in Table 5.6, the column “event” refers to those two types of verbs, in accordance to Niculae & Yaneva’s experiments (2013).

For each method, the recall and the precision are given. They were calculated based on these two formulas:

$$\text{Recall} = \text{True Positives (TP)} / \text{True positives} + \text{False negatives (FN)}$$

$$\text{Precision} = \text{True Positives} / \text{True positives} + \text{False positives (FP)}$$

**Table 5.6 Results obtained with the proposed algorithm (left) and with the Berkeley Parser (right)**

	Rc (%)	Pr (%)	TP	FP	FN
<i>Tenor</i>	61.9	46.9	163	184	100
<i>Eventuality</i>	55.5	52.8	75	67	60
<i>Ground</i>	58	69.1	83	37	60
<i>Vehicle</i>	90.8	96.7	238	8	24

	Rc (%)	Pr (%)	TP	FP	FN
<i>Tenor</i>	54.3	50.1	143	142	120
<i>Eventuality</i>	64.4	47.8	87	95	48
<i>Ground</i>	44	69.2	63	28	80
<i>Vehicle</i>	87	90	228	23	34

Overall, the implementation that relies on chunking and rules yields better results than the one based on dependency parsing. The dependency-based algorithm, in particular, is less good at detecting vehicles because of part-of-speech tagging errors, faulty sentence segmentation, vehicles wrongly identified as being subjects or a wrong dependency. However, if syntactic chunking works well for close-distance dependency, it does not perform well with long-distance dependencies, for example when a parenthetical expression separates the verb from its subject or when it comes to retrieving coordinated syntactic elements.

Furthermore, generally speaking, some structures are highly problematic for both methods:

- past participles used as nouns: "... *coupable à l'égal d'un faux scandalisé*";
- a succession of comparisons in the same sentence: "*ses cheveux longs comme des saules et peignés comme des broussailles*";
- inverted subjects: "*cette solide cage de fer derrière laquelle s'agite, hurlant comme un damné, secouant les barreaux comme un orang-outang exaspéré par l'exil, imitant, dans la perfection, tantôt les bonds circulaires du tigre, tantôt les dandinements stupides de l'ours blanc, **ce monstre poilu** dont la forme imite assez vaguement la vôtre*";
- comparisons without tenors: "*Ce soir à Circeto des hautes glaces, grasse comme le poisson, et enluminée comme les dix mois de la nuit rouge, - (son cœur ambre et spunk), - pour ma seule prière muette comme ces régions ...*";
- the use of an adjective which is not a ground before the marker: "*Il est aussi difficile de supposer une mère sans amour maternel qu'une lumière sans chaleur*";
- a succession of more than two adjectives: "*Les meubles sont vastes, curieux, bizarres, armés de serrures et de secrets comme des âmes raffinées*";
- long dependencies between the verb and its subject: "*Tel qui, craignant de trouver chez son concierge une nouvelle chagrinante, rôde lâchement une heure devant sa porte sans oser rentrer, tel qui garde quinze jours une lettre sans la décacheter, ou ne se résigne qu'au bout de six mois à opérer une*

*démarche nécessaire depuis un an, se sentent quelquefois brusquement précipités vers l'action par une force irrésistible, comme la flèche d'un arc".*

With regard to the strengths and weaknesses of both outputs, the next logical step, in order to improve the performance, is to merge the two approaches, so as to be able to take advantage of grammatical functions and phrase boundaries at the same time. For English, the Stanford Core NLP seems perfect for this task as it performs tokenisation, part-of-speech tagging, lemmatisation, dependency and constituency parsing. It is, however, extremely verbose and its implementation for French does not lemmatise words and the output of the dependency parser contains too many mistakes to be objectively exploitable.

Globally speaking, since the proposed algorithm gives multiple solutions and consequently tends to generate more noise than the dependency-based one, it is crucial to find how to reduce the noise.

### **Examples**

1/ Un immense bruissement de vie remplissait l'air -- la vie des infiniment petits, -- coupé à intervalles réguliers par la crépitation des coups de feu d'un tir voisin, qui éclataient comme l'explosion des bouchons de champagne dans le bourdonnement d'une symphonie en sourdine. {marker: "comme", vehicle: "explosion", verb\_ground: "éclataient", tenor\_subject: ["crépitation", "coups", "feu", "tir"]}

2/ A travers ces barreaux symboliques séparant deux mondes, la grande route et le château, l'enfant pauvre montrait à l'enfant riche son propre joujou, que celui-ci examinait avidement comme un objet rare et inconnu. {marker: "comme", vehicle: "objet", verb\_ground: "examinait", tenor\_subject: "celui-ci", tenor\_object: "joujou"}

As far as the French language is concerned, checking agreement could enable to delete wrong nouns modified by an adjective or wrong subjects. In this respect, Morphalou (Romary, Salmon-Alt & Francopoulo, 2004) was used to check noun-adjective and subject-verb agreement. It is important to note that since a verb can have more than one subject, the agreement is not checked when the verb is in the third person plural and the potential subject is a noun.

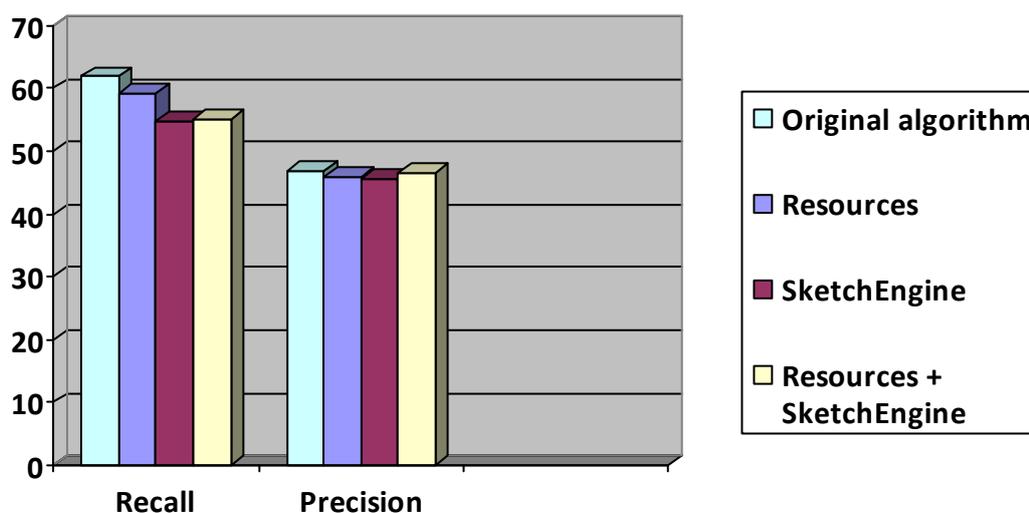
In addition, by adding information about the transitivity of the verb, it is possible to further delete all extracted direct objects when the identified verbal ground or linking verb is a non-transitive verb. When, on the contrary, the verbal ground or linking verb is indeed a

transitive verb, to decide whether the true tenor is the subject of the verb or its object, the fact that the vehicle has the same grammatical function as the tenor could help to decide. As a matter of fact, since the verbal ground does not apply only to the tenor, but also to the vehicle, this method is based on the idea that if the vehicle is generally used in English or in French as the subject of the identified verbal ground, the extracted subject of that verb would be the tenor. If otherwise, the vehicle is generally used as the object of the identified verbal ground, the extracted direct object would be the tenor.

In order to specify the relationship between the vehicle and the eventuality, VerbNet (Kipper, Dang & Palmer, 2000) and Les Verbes français (Dubois & Dubois-Charlier, 1997), two lexical databases that organise verbs into different semantic classes, are used respectively for English and French in combination with the SketchEngine (Kilgariff, Rychly, Smrz & Tugwell, 2004). The first step is to extract from VerbNet and Les Verbes français, a list of verbs that have almost identical meaning as the verbal ground. Then, the SketchEngine enables to determine if in fact in language the vehicle tends to appear more as the subject or as the direct object of one of those verbs.

Figure 5.3 shows that on the French corpus of prose poems, the different resources (Morphalou and the list of transitive verbs) combined with the SketchEngine (Kilgariff et al., 2004) achieves slightly better precision than the original algorithm, although it is possible to notice a significant drop as far as the recall is concerned.

**Figure 5.3 The impact of lexical resources on the tenor's recall and precision**



As grammatical resources could not be used in all cases and do not seem to make a great difference, we also explored simpler rules. For example, despite the absence of a

conjugated verb after the vehicle, it is possible to conclude that the marker introduces a clausal simile when the first preposition in the verbal phrase except “of” is also the first preposition after the vehicle.

### **Example**

Uncertainty and even despair hover **over** Shea Stadium and the neighbouring shell of Citi Field like smog **over** Beijing (Similepedia Blog).

## 5.3 The Semantic Module

Once potential similes have been identified and their components disambiguated, the next step is to determine whether they express a simile or a literal statement. It is possible to reduce the number of comparisons to disambiguate by eliminating structures corresponding to pseudo-comparisons such as: verb of perception/judgement + “comme”, “il y + avoir + comme”, “ce + être + comme”, “there + be + like”. While the first type of structures corresponds to “identification” as it confers a role to a thing or a person (“I see her as a friend”), the remaining three structures all fall under the label “approximation”. As the other values of pseudo-comparisons, “exemplification” and “coordination”, generally concerns words that are semantically related, they are treated as literal comparisons.

### **Examples**

There was suddenly like a commotion in the street. → approximation

Il y eut soudain comme un grand bruit. → approximation

Les hommes comme les femmes se sont mobilisés en cette période difficile. → coordination

For the remaining sentences, in order to apply the categorisation theory, it is necessary to determine how the category of the tenor and the vehicle would be determined. Dictionary definitions have been known to contain relevant information from which the taxonomy of a specific word can be drawn (Amsler, 1980). But definitions of traditional dictionaries are not dependable enough to enable to always retrieve directly the word hypernym as well as to extract semantic information such as abstractedness and animacy, unlike dictionaries that rely on an ontology or an ontological system. In this respect, WordNet (Fellbaum, 1998) and Le Dictionnaire électronique des mots (Dubois & Dubois-Charlier, 2010) seem the most promising resources for English and for French respectively.

WordNet's ambitions, since its beginning, has been to avoid circularity in word definitions and to relate each noun to its superordinate. In this respect, a set of 25 unique beginners or semantic categories was determined (Miller, 1990) so that the definition of a term is complemented by its semantic category.

**Examples:**

Two definitions in WordNet (The unique beginner is the second element of the definition and is located between the number of the synset and the part-of-speech of the defined term)

00345817 **04** n 01 toss 2 002 @ 00331950 n 0000 + 01890792 v 0107 | an abrupt movement; "a toss of his head" ▫ noun denoting an act or an action

08557976 **15** n 01 viscounty 0 001 @ 08556491 n 0000 | the domain controlled by a viscount or viscountess ▫ noun denoting a spatial space

We tested the effectiveness of simply contrasting these semantic categories on 30 tenor-vehicle couples used in a scientific experiment by Ortony et al. (1985): 96.6% of the literal comparisons and 58% of the similes were correctly identified. Nearly all the errors as far as similes are concerned are caused by polysemy. This sample also raises the issue of compound nouns such as "shopping centre": is it only the core noun that should be considered or the compound noun as a whole? It makes a difference in the sentence "Shopping centres are like jungles" since both "jungles" and "centres" are classified as nouns denoting groupings of people or objects [14] and nouns denoting spatial position [15] whereas as "shopping centre" appears only under man-made objects [6]. The simile is therefore detected with the compound noun and not with the head of the noun phrase.

Just like WordNet (Fellbaum,1998), Le Dictionnaire électronique des mots (Dubois & Dubois-Charlier, 2010) aims to give more semantic information than mere definitions. It, therefore, indicates the animacy of noun terms by specifying whether the noun refers to an animal, a human being or a non-animate. If the first two semantic features can be considered as reliable semantic categories, the category "non-animate", however, is too vague and too broad to make a significant difference. As a matter of fact, "non-animate" can apply to objects, abstract concepts, plants, food and locations. In order to differentiate between all those types of nouns, it is possible to use another element provided by the dictionary: the tag <OP> that further delimits the category of a word inside a given domain terminology.

### Example

A definition taken from Le Dictionnaire électronique des mots (Dubois & Dubois-Charlier, 2010)

```
<mot mot="glaïeul" nb="1" id="glaïeul">
  <entree ligne="63650">
    <M mot="glaïeul" mot-initial="glaïeul"/>
    <CONT>culture N </CONT>
    <DOM nom="plantes">PLA</DOM>
    <OP>herb</OP>
    <SENS>iridacée,grdes fleurs</SENS>
    <OP1>R3a1</OP1>
    <CA categorie="N" type="non-anime" genre="M">-1</CA>
  </entree>
</mot>
```

An experimental test conducted on similes extracted from the French corpus presented in the previous section called attention to two main points. First of all, the head of the noun phrase that complements the marker is not always the semantic tenor, especially if it is a collective noun followed by a prepositional phrase introduced by “of”. In a sentence such as “Her hair was blazing like a myriad of colours”, “colours” is clearly the semantic tenor and not the head of the noun phrase “myriad”. Secondly, the categorisation should take into account context and more specifically grounds and verify whether it is a salient attribute of the vehicle. For instance, in the following sentence “A côté de lui, gisait sur l’herbe un joujou splendide, aussi frais que son maître, verni, doré, vêtu d’une robe pourpre”, despite the change in categorisation from “joujou” to “maître”, this sentence is not a simile because “frais” does not denote an intrinsic characteristic of “maître”.

With regard to all that have been said above, for a comparative construction to be considered a simile, at least one of the following conditions must be fulfilled:

1/ the ground + vehicle combination is recorded in a precompiled list of idiomatic similes, or if only one of them can be found in the precompiled list, the other is a synonym of the word the idiomatic ground or vehicle is generally paired with;

### Examples

This kid is more obstinate than a mule.

Cet enfant est plus obstiné qu’une mule.

2/ the ground expresses common conceptions about the vehicle, for example, “calm” and “lake”;

### Examples

I touched her cheeks, [soft] like flower's petals.

Je touchai ses joues, [douces] telles des [pétales] de fleurs.

3/ the vehicle is part of an extended noun phrase;

### Examples

Her smile is rare as [snow in June].

Son sourire est rare comme [la neige en juin].

4/ 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” [Weiner, 1984]).

### Examples

[Your bedroom] reminds me of [a battefied].

[Ta chambre] me fait penser à [un champ de bataille].

The two last conditions do not apply when the marker is a comparative of degree, because in this case, the salience of the ground prevails. In addition to these conditions, other textual clues can be added:

- If the marker is like and is preceded by “just”, it introduces a comparison;
- In compliance with Niculae and Danescu-Niculescu-Mizil (2014), “other” used as a standard of comparison indicates a comparison and the same goes for “nothing”, “something”, “rest”, “somebody”, “anything” which are also indicators of comparison when they form the comparee NP.
- By default, “as” introduces a pseudo-comparison when it is used only with a verbal ground;
- The use of a possessive pronoun before the vehicle indicates a comparison;
- When all rules have failed, the presence of an indefinite article (or the absence of an article for English sentences) before the vehicle indicate a simile.

To build the database necessary for the recognition task, first, idiomatic similes of the form “verb + marker + vehicle” and “adjective + marker + vehicle” were retrieved from two simile dictionaries: *Les Comparaisons du français* (1996) by Nicolas Cazelles and the *English/French Dictionary of Similes* (2002) by Michel Parmentier. Then, hypothesising that salient features commonly associated with a certain word are connected to its usage, and are therefore embedded in language, we compiled a corpus of machine-readable

dictionaries (see Table 5.7) to automatically retrieve specific linguistic pairs: nominal subject-verb, verb-nominal direct object, nominal subject-predicative adjective, adjective-noun. In addition, when indicated in the dictionary (see Figure 5.4), all synonyms as well as antonyms of verbs, adjectives and nouns were also extracted, so as to capture not only rewriting of idiomatic similes but also variants of the frequent salient traits.

**Table 5.7 Machine-readable dictionaries used**

	Dictionaries	Tokens
English	GCIDE (Collaborative International Dictionary of English) <sup>31</sup>	8,187,172
	Wiktionary (Navarro et al., 2009)	11,564,739
	WordNet (Fellbaum, 1998)	1,717,911
French	Littré (1873-1874) <sup>32</sup>	2,657,996
	Wiktionary (Navarro et al., 2009) <sup>33</sup>	9,649,312
	Dictionnaire de l'Académie Française, 6 <sup>e</sup> édition (1835)	3,994,518
	Dictionnaire de l'Académie Française 8 <sup>e</sup> édition (1932-1935) <sup>34</sup>	3,239,560

**Figure 5.4 Example of an entry in the GCIDE**

```

<entry key="Murderous">
  <hw source="1913 Webster">Mur"der*ous</hw>
  <wordforms>
    <wf>Mur"der*ous*ly</wf>
    <pos>adv.</pos>
  </wordforms>
  <pos>a.</pos>
  <def>Of or pertaining to murder; characterized by, or causing, murder or bloodshed; having the
purpose or quality of murder; bloody; sanguinary; <as>as, the <ex>murderous</ex> king;
<ex>murderous</ex> rapine; <ex>murderous</ex> intent; a <ex>murderous</ex> assault.</as>
  </def>
  <q>
    <ex>Murderous</ex>coward.</q>
  <au>Shak.</au>
  <syn source="1913 Webster">Bloody; sanguinary; bloodguilty; bloodthirsty; fell; savage;
cruel.</syn>
</entry>

```

<sup>31</sup> The GCIDE is made up from definitions from the 1913 Webster Dictionary supplemented with some definitions from WordNet (Fellbaum, 1998) and is freely available at the following address: <https://www.ibiblio.org/webster/>

<sup>32</sup> <https://bitbucket.org/Mytskine/xmlittre-data>

<sup>33</sup> Both versions of Wiktionary can be downloaded at: <http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html>

<sup>34</sup> The two versions of the Dictionnaire de l'Académie Française can be found at the XDXF dictionaries repository (<https://sourceforge.net/projects/xdxf/files/>)

That idea of the ground expressing a salient trait of the vehicle or the vehicle being the archetypal exemplification of the quality denoted by the ground, can be of course found in various articles such as Ortony (1977) and Fishelov (1993). In addition, from a structural point of view, Fishelov (1993) mentions a difference between the vehicle length and the tenor length as one of the criteria that may differentiate a poetic from a non-poetic simile. This abnormally lengthy vehicle is generally constructed either by extending the noun phrase with a relative clause, with an adjective or with a prepositional phrase. This strategy is particularly useful when the vehicle only exemplifies the salient trait expressed by the ground in particular situations.

### **Example**

He saw just one yellow gleam of the mast-head light high up and blurred like a last star ready to dissolve. (Conrad, as cited in Wilstach, 1913, p. 26).

==> Stars are generally associated with brightness and not to blurriness and the quotation suggests that stars are blurry when they are on the verge of disappearing.

If it is generally agreed upon in the literature that the vehicle and the tenor should not belong to the same semantic categories, the exact semantic category that has to be considered remain fuzzy. As a matter of fact, since semantic categories are hierarchical, the question is whether to decide based on the hypernym, the top semantic category, or on sub-categories. For example, as previously stated, Weiner (1984) describes “Penguins are like wolves” as a simile unlike “Dogs are like wolves”, arguing that “penguins” and “wolves” are different types of animals and are farther on the animal taxonomy. This distinction is only possible if one goes beyond the fact that both the tenor and the vehicle are animals. Consequently, to capture this close relationship between words, coordinated nouns were also automatically extracted.

Coordination, in a way, can also be considered as a kind of ellipsis, as it enables to avoid repetition. In this respect, instead of saying “I bought a car and I bought a dress”, one can say “I bought a car and a dress”. Thus, coordinated words are often semantically close, are frequently used in the same context and belong to the same semantic category, in the previous example, “objects”.

Following the same rationale, coordinated verbs and adjectives were also clustered as synonyms and added to the list of already extracted verb and adjective synonyms.

A particular set of rules apply to comparative constructions with “have” and a direct object since the verbal phrase can constitute the ground or the direct object could the tenor of the simile.

### **Examples**

He had a face like a benediction (Cervantes as cited in Wilstach, 1913, p. 122).

Honor that is gained and broken upon another hath the quickest reflection, like diamonds with facets; and therefore let a man contend to excel any competitors of his in honor, in outshooting them, if he can, in their own bow (Cervantes as cited in Wilstach, 1913, p. 204).

Since “to have” implies possession, if the direct object refers to an attribute or a part of the identified vehicle, the subject of the verbal ground would be considered as the tenor of the simile. This is the case of the second sentence, in which “reflection” is an attribute of diamonds. If on the contrary, the direct object does not refer to an attribute or a part of the identified vehicle, it would be taken as the tenor of the simile. For English, meronyms and their corresponding holonyms were extracted from WordNet (Fellbaum, 1998). For French, are considered as meronyms body parts, and abstract qualities and attributes.

In the past years, various methods have been proposed to extract automatically collocates from a corpus based on syntactic relations (Lin, 1998; Curran & Moens, 2002; Almuhareb & Poesio, 2004; Kilgarriff, Rychly, Smrz & Tugwell, 2004; Padó & Lapata, 2007; Rothenhäusler & Schütze, 2009). Unlike most of these methods, due to the small size of our corpus (between 18 and 20 million tokens per language), we decided not to rank the obtained pairs. We evaluated the reliability of the automatically extracted salient traits by contrasting them with data from two shared tasks of the Lexical Semantics Workshop (ESSLI 2008):<sup>35</sup> correlation with free association norms and comparison with speaker-generated featured. The results of this evaluation are presented in Table 5.8; only the 531 words found in both datasets were considered.

### **Example**

Words associated with “lettuce” in the ESSLI dataset: large, **enormous**, great, heavy, big, size, mane, **animal**, **beast**, **African**, carnivore, meat, meat-eater, **roar**, furred, furry, pelt, feline, **ferocious**, **fierce**, furious, savage, wild, wilderness

---

<sup>35</sup> <http://wordspace.collocations.de/doku.php/workshop:essli:task>

Words associated with “lion” in the machine-readable dictionaries:

- Adjectives: female, social, famous, sculptured, winged, southern, northern, black, tawny, Numidian, **African**, small, stylized, Asian, male, marauding, **enormous**, dead, rampant, **ferocious**, **fierce**, full-grown, hungry, young, cougar, American, maneless
- Verbs: **roar**, disperse, perish, design, paw, begin, eat, hamper, groan, believe, say, give, raise, leap, find, think, devour, catch, lie, show, see, jump, raven, lash, incline, appropriate, seize
- Nouns: tigress, wolf, seal, head, cat, **beast**, eagle, leopard, panther, tiger, adder, dragon, cheetah, jaguar, horse, man, object, bear, catamount, **animal**, wing, crocotta, aspect

**Table 5.8 Evaluation of automatically extracted salient traits and synonyms**

	Recall (%)
Exact Matches ESSLi dataset $\cap$ Automatically generated database	<b>48.8</b>
ESSLi Dataset $\cap$ Exact Matches + synonyms of the terms associated with each word in the dictionaries	<b>80.3</b>
Exact Matches ESSLi dataset + synonyms of the terms associated with each word in the ESSLi data $\cap$ Automatically generated database	<b>89.3</b>

We tested this set of rules on the similes encoded in the VUAMC Online (Steen et al., 2010) that have been identified in literary texts (see Appendix 4). Of the 43 similes in that dataset, our system correctly retrieved 40 similes, the remaining two were not found because of incorrect part-of-speech tags. 8 similes were wrongly considered as comparisons, either because the ground was not captured as being salient and the tenor and the vehicle belong to the same grammatical category (“tent-like coat”, “...outriggers splayed from her upperworks like antennae of some outlandish insect”) or are synonyms (“the conditions were like the feeling of a tomb”) or because of the rule that imposes an adjectival ground in similes with “as” (“In the catalogue John House quoted Monet’s description of the painted light around the snowy haystacks as an enveloping veil”). We also noticed that, even though some obvious salient traits are absent from the database (“fall” for “parachutist”, “fret” for “hen””, “unmoving” for “lizard”), the system can, in most cases, successfully rely on other rules.

## 5.4. The Annotation Module

One of the main challenges of any annotation scheme is, of course, to decide what exactly should be annotated. Generally speaking, as far as comparative constructions are

concerned, it is possible to annotate not only the whole structure but also to tag each of its components. Since our main interest is in similes, the first question to answer is if we should ignore all retrieved instances of literal comparisons and of pseudo-comparison. Although there is no clear-cut answer, if those types of constructions do not seem to be very indicative stylistically or ideologically in a fictional text, they could be useful in non-fictional texts such as critical literary texts in which they are often used to contrast ideas or for argumentation. Besides, still in this type of texts, it could be interesting to study pseudo-comparisons denoting identification or exemplification. Furthermore, as the proposed method also often captures some clausal similes, those should undoubtedly be taken into account. In this respect, depending on the type of text, at the sentence level, four types of structures could be annotated: literal comparisons (<comparison>... </comparison>), pseudo-comparisons (<pseudo-comparison value = “...”>... </pseudo-comparison>), clausal similes (<simile nature= “clausal”>...</simile>), and phrasal similes (<simile nature= “phrasal”>...</simile>).

Another important characteristic of this annotation frame as far as the component of each identified concerned is concerned, is the fact that the mark up occurs both at the phrase level and at the word level. For example, in a sentence such as “His jealousy rises and falls like the wind”, the tenor would be rendered as:

```
<tenor marker_id= “6”>His <head lemma= “jealousy” postag= “NN” category=
“abstract, attributes and qualities”> jealousy</head></tenor>
```

In this example, apart from descriptive annotations derived from the extraction, it is possible to notice additional information such as the position of the marker in the sentence and the semantic category of the head noun.

As it often occurs that a single sentence contains several similes introduced by the same or by different markers, it seems essential to clearly identify around which marker is centred the (pseudo-)comparison or the simile. For semantic categories, our aim was to have a set of categories not too broad and too refined.

After consulting, ontologies such as WordNet (Fellbaum, 1998) and the SIMPLE-CLIPS,<sup>36</sup> we decided on the following 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	

Once these categories were defined, we had to match them to the common nouns in each language. As far as English is concerned, we took advantage of the unique beginners attached to each WordNet entry and made it correspond to one of our semantic categories (see Table 5.9). To cope with polysemy, a word – one semantic category principle was adopted. Consequently, for each word with more than one possible semantic category, its monosemous synonyms and hypernyms were used to determine its most frequent category. Once a semantic category was allocated to a word, the process was reiterated so as to disambiguate all polysemous words.

---

<sup>36</sup> <http://webilc.ilc.cnr.it/clips/Ontology.htm>

**Table 5. 9 Correspondence between our semantic categories and WordNet's unique beginners  
(Fellbaum, 1998)**

Man-made objects	06 nouns denoting man-made objects 13 nouns denoting foods and drinks 15 nouns denoting spatial position 21 nouns denoting possession and transfer of possession
Natural objects	17 nouns denoting natural objects (not man-made) 27 nouns denoting substances
Body parts	08 nouns denoting body parts
Human beings	18 nouns denoting people
Animals	05 nouns denoting animals
Plants, fruits and vegetables	20 nouns denoting plants
Temporal elements	28 nouns denoting time and temporal relations
Concepts	09 nouns denoting cognitive processes and contents 10 nouns denoting communicative processes and contents 19 nouns denoting natural phenomena 23 nouns denoting quantities and units of measure 24 nouns denoting relations between people or things or ideas 25 nouns denoting two- and three-dimensional shapes
Feelings and emotions	12 nouns denoting feelings and emotions
Acts and processes	04 nouns denoting acts or actions 11 nouns denoting natural events 16 nouns denoting goals 22 nouns denoting natural processes
Attributes and qualities	07 nouns denoting attributes of people and objects 26 nouns denoting stable states of affairs
Collective nouns	14 nouns denoting groupings of people or objects

For French, we used *Le Dictionnaire électronique des mots* (Dubois & Dubois-Charlier, 2010) which clearly indicates whether a noun designates an animal or a human being and for other types of words, as mentioned in the previous section, gives its syntactic behaviour and its semantics by providing its context (i.e. an abbreviated type of use, for example, the context of everything that can be counted is “compt P N” whereas the context of something that could be drunk is “boire N”), its semantic category, (everything that can be drunk is tagged as “liq”, acts are identified as “acte”) and the type of verbs with which it typically comes. Using these three elements, it is, therefore, possible to cluster together words that occur in the same syntagmatic and syntactic context and, consequently, belong to the same semantic domain.

## Examples

```

<mot mot="milk-shake" nb="1" id="milk-shake">
  <entree ligne="87803">
    <M mot="milk-shake" mot-initial="milk-shake"/>
    <CONT>boire N</CONT>
    <DOM nom="boisson">BOI</DOM>
    <OP>liq</OP>
    <SENS>boisson à base d lait</SENS>
    <OP1>S3j1</OP1>
    <CA categorie="N" type="non-anime" genre="M">-1</CA>
  </entree>
</mot>

<mot mot="calvados" nb="1" id="calvados">
  <entree ligne="21781">
    <M mot="calvados" mot-initial="calvados"/>
    <CONT>boire N</CONT>
    <DOM nom="boisson">BOI</DOM>
    <OP>liq</OP>
    <SENS>eau-de-vie d cidre</SENS>
    <OP1>S3j1</OP1>
    <CA categorie="N" type="non-anime" genre="M">-1</CA>
  </entree>
</mot>

```

Figure 5.5 Example of noun semantic categorisation using Le Dictionnaire électronique des mots (Dubois & Dubois-Charlier, 2010)

insécurité	f épro p N	SOC	sent	P2a1	Entités abstraites – Emotions et sentiments
intimidation	f épro p N	PSY	sent	P2a1	Entités abstraites – Emotions et sentiments
leurre	f épro p N	PSY	sent	P2a1	Entités abstraites – Emotions et sentiments
mésestimation	f épro p N	PSY	sent	P2a1	Entités abstraites – Emotions et sentiments
mésinterprétation	f épro p N	PSYt	sent	P2a1	Entités abstraites – Emotions et sentiments
mieux-être	f épro p N	SOC	sent	P2a1	Entités abstraites – Emotions et sentiments
minotaurisation	f épro p N	SOCv	sent	P2a1	Entités abstraites – Emotions et sentiments
modération	f épro p N	PSY	sent	P2a1	Entités abstraites – Emotions et sentiments
mystification	f épro p N	PSY	sent	P2a1	Entités abstraites – Emotions et sentiments
casernement	f habi p N	MIL	bât	L1a1	Entités abstraites - Actions et procédés
castramétation	f habi p N	MIL	loc	L1a1	Entités abstraites - Actions et procédés
cohabitation	f habi p N	SOC	loc	L1a1	Entités abstraites - Actions et procédés
conurbation	f habi p N	GEG	loc	L1a1	Entités abstraites - Actions et procédés
cooccupation	f habi p N	DRO	loc	L1a1	Entités abstraites - Actions et procédés
domiciliation	f habi p N	ADM	loc	L1a1	Entités abstraites - Actions et procédés
habitabilité	f habi p N	BAT	loc	L1a1	Entités abstraites - Actions et procédés
habitation	f habi p N	SOC	loc	L1a1	Entités abstraites - Actions et procédés
abaissement	f mvt p N	QUA	type	M1a1	Entités abstraites - Actions et procédés
abduction	f mvt p N	SOM	phys	M3a1	Entités abstraites - Actions et procédés
abord	f mvt p N	SOC	type	M1a1	Entités abstraites - Actions et procédés
aboutissants	f mvt p N	TPS	mes	M3a1	Entités abstraites - Actions et procédés
aboutissement	f mvt p N	TPS	mes	M3a1	Entités abstraites - Actions et procédés
accélération	f mvt p N	TPS	mes	M3a1	Entités abstraites - Actions et procédés
accès	f mvt p N	QUA	état	M1a1	Entités abstraites - Actions et procédés
accès	f mvt p N	LOC	voie	M1a1	Entités abstraites - Actions et procédés
accompagnement	f mvt p N	SOC	état	M1a1	Entités abstraites - Actions et procédés

To some extent, using rather broad categories decreases word polysemy as most lexemes are either objects, abstract entities or living beings. In case of doubt, however, the first sense registered in the dictionary prevails.

**Table 5.10 Summary of the annotation scheme**

Structure	Substructure	Components
<i>Pseudo-comparison</i>	Identification Exemplification	Complement of the marker Element identified/exemplified
<i>Literal comparisons</i>		Comparee NP Quantity/quality Standard NP
<i>Clausal similes</i>		Tenor Vehicle
<i>Phrasal similes</i>	- Perceptual - Proverbial - Idiomatic - Reinvented - Original	Tenor Vehicle Ground

In the proposed annotation scheme, as shown in the Table above, we distinguish five types of similes:

- 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 tag <source> under goes which the typical form of the idiomatic simile.
- original similes (<type= “original”>...</type>)

From a stylistic point of view, we found it interesting to also add syntactic information about the marker and about the vehicle noun phrase. In this respect, it is stated whether the marker occurs at the beginning of a sentence or a clause, or after a comma and whether the vehicle noun phrase is extended by a relative clause.

### **Example**

```
<simile type= "original">
<tenor marker_id="4">The<head lemma="pan" postag="NN" category="concrete, man-
made object">pan</head></tenor>is<ground marker_id="4"><head lemma="heavy"
postag="JJ "> heavy </head> <ground><marker lemma="like" marker_id="4"
syntax="null">like </marker> <vehicle marker_id="4">an elephant's<head
lemma="paw" postag="NN" category="concrete, body part">paw</head></vehicle>.
</simile>
```

In this chapter, we detailed a grammar of similes on which we based our method for simile detection and annotation. Unlike previous algorithms, the proposed one is flexible enough to take into consideration a wide range of markers and several types of comparative structures. More concretely, the method we proposed relies on syntax, semantics and a set of rules to extract simile candidates, to identify their components, to judge the degree of literalness of each captured structure and to enrich raw texts with valuable information.

# 6 TOWARDS AN ANNOTATED LITERARY CORPUS OF SIMILES

Evaluation in natural language processing goes hand in hand with annotated datasets or corpora. When there exist freely available datasets suitable for a particular task, it is simpler and generally recommended to reuse them so as to compare the new results with previous ones. Otherwise, the next logical step is to build an annotated dataset either by relying on the know-how of a small number of experts or through crowdsourcing which consists in collecting annotations from the largest number of non-experts, with the belief that correct answers would emerge by aggregating all the propositions. As compared to expert annotations, crowdsourced ones are not only less costly and less time-consuming, but are qualitatively as good on certain linguistic tasks (Snow, O'Connor, Jurafsky & Ng, 2008). Another interesting aspect of crowdsourcing is the new light it could shed on tasks that are often deemed as being easy for human beings as well as on commonly made mistakes. As a matter of fact, annotated corpus of this nature can also be used for psycholinguistic purposes and can serve to verify some working hypotheses. In this respect, in this study, first, experts and then, crowdsourcing were used to collect annotations on a prebuilt corpus. Before discussing what we learned from these annotations, we will first of all give more information about the corpus used and describe the design conceived in both cases to collect annotations.

## 6.1 Corpus Presentation

As our focus is on prose texts, it would appear obvious to choose novels to be part of the corpus. But, similes being a sporadic linguistic phenomenon whose use varies from one other to another, recording all the comparisons and pseudo-comparisons in a set of novels would have been tedious, whereas selecting picking a random number of similes per text could have been subjective. In addition, both solutions make it difficult to check the recall afterwards. Consequently, instead of novels, we choose to restrict the corpus to prose poems, which have the double advantage of sometimes being short so as to put in entirety and of being susceptible to contain a lot of similes as they pertain to the poetic genre. Lehman (2003) defines a prose poem as:

... a poem written in prose rather than verse. On the page it can look like a paragraph or fragmented story, but it acts like a poem. It works in sentences rather than lines. With the one exception of the line break, it can make use of all the strategies and tactics of poetry. Just as free verse did away with meter and rhyme, the prose poem does away with the line as the unit of composition. It uses the means of prose towards the ends of poetry. (p. 13)

Often considered as prose borrowing poetic features or as poetry written as prose, by defying traditional writing norms, the prose poem is a historically subversive genre which does not let itself being confined by specific rules and conventions (Murphy, 1992). Although the genre was mainly popularised by Charles Baudelaire's *Petits Poèmes en Prose* (1869), the paternity of the prose poem is often attributed to Aloysius Bertrand whose *Gaspard de la Nuit* which was published amidst total indifference 27 years before. From then onwards, this new form strongly influenced subsequent generations of writers inside as well as outside France. It is, however, worth noting, if we take into account only the form and not the author's deliberate intent, that various examples of prose blended with poetry can be found in British literature in texts such as *Hamlet* (1602), *The King James Bible* (1611), or William Blake's *Marriage of Hell* (1793) (Lehman, 2003). This could explain why the prose poem has struggled to impose itself as a genre in British literature and has been more wholeheartedly embraced by avant-garde American authors.

For the purpose of this study, we selected a total of eleven French-speaking authors who published collections of prose poems between 1842 and 1920: Aloysius Bertrand (*Gaspard de la Nuit*, 1842), Charles Baudelaire (*Petits poèmes en prose*, 1869), Arthur Rimbaud (*Une Saison en Enfer*, 1873; *Les Illuminations*, 1895), Jules Barbey d'Aureville (*Amaléc*, 1890), Ephraïm Mikhaël (*Œuvres, poésie, poèmes en prose*, 1890), Stéphane Mallarmé (*Divagations*, 1897), Gabriel de Lautrec (*Poèmes en prose*, 1898), Albert t'Serstevens (*Poèmes en prose*,

1911), Jean de Bère (*Au fond des yeux: Petits poèmes en prose*, 1911), Louis-Joseph Doucet (*Au bord de la clairière: Petits poèmes en prose et autres*, 1916), and Jean Aubert Loranger (*Les Atmosphères*, 1920). With regard to the short span of time that seems to cover the corpus, it is important to add that, in the 19<sup>th</sup> century, French poetry witnessed the birth of various literary currents such as the Parnasse or the symbolism, and that this experimental trend went through the beginning of the 20<sup>th</sup> century. In this respect, most of these poets not only have radically different styles and themes but are essentially interested in redefining and stretching the boundaries of the prose poem.

With the exception of three British texts (William Blake, *Marriage of Heaven and Hell*, 1790-1793; Oscar Wilde, *Poems in Prose*, 1894; Ernest Dowson, *The Poems of Ernest Dowson*, 1911), the corpus in English consists of American prose poems: Ralph Waldo Emerson's "Woods, A Prose Sonnet" (1839), Edgar Allan Poe (*Eureka*, 1848), Gertrude Stein (*Tender Buttons*, 1914), Amy Lowell (*Men, Women and Ghosts*, 1916), Sherwood Anderson (*Mid-American Chants*, 1918), Williams Carlos Williams (*Kora in Hell: Improvisations*, 1920), Edna Kingsley Wallace (*The Stars in the Pool: A Prose poem for lovers*, 1920), Charles Freeland (*Albumen*, 2014). Several poems were also taken from collections of prose poems by Walt Mason, who between 1907 and 1939 daily furnished thousands of newspapers in USA, Canada, Great Britain and even India in prose poems on subjects ranging from sports to economics and society and was purportedly read by around 10 million people (White, 1910; French, 1929).

For the annotation task, all sentences containing comparisons or pseudo-comparisons were manually identified. Tables 6.1 A & B below give an overview of the distribution of the different markers in each corpus. As far as the English corpus is concerned, whereas "markers of inequality" refers to "more... than", "less...than" and "-er...than", "others" regroups suffixes and verbs. In the French corpus, mainly "ainsi que" and "tel que" are found under "others". Unsurprisingly, if all British and American poets use at least once "like", all French-speaking poets use "comme" profusely. Some authors, however, tend to vary more often the markers they use: for example, 15 different markers have been registered in Mallarmé's selected poems.

**Tables 6.1 A & B. Statistics on the distribution of markers in the English (left) and French (right) annotation corpora**

Markers	Instances	Frequency per author
like	255 (42%)	21.2
as	215 (36%)	17.9
as...as	48 (8%)	4
Inequality comparatives	67 (11%)	5.5
Others	16 (3%)	1.3
	601	/

Markers	Instances	Frequency per author
Comme	654 (76.5 %)	59.4
Verbal phrases	24 (3 %)	2.1
Adjective phrases	62 (7.2 %)	5.6
Degree comparatives	70 (8.1%)	6.3
Prepositional phrases	13 (1.5%)	1.1
Others	31 (3.6%)	2.8
	855	/

## 6.2 Experts' Annotation

Two specialists of French literary stylistics were asked to annotate an author of their choice in the corpus: one of the them, Annotator A chose Baudelaire, and the second one picked Aloysius Bertrand. They each received the text with highlighted sentences containing the structure marker + non-subject noun-headed noun phrase (the marker was always underlined), a list of the markers that are part of the experiment and a series of instructions. More explicitly, they were asked to:

- Identify the compare NP/tenor

### Examples

Le **ciel** est triste et beau comme un grand reposoir.

Le **ciel** est triste et beau comme ton regard.

J'ai cité le **ciel** comme un élément atmosphérique.

- Identify the standard of comparison/vehicle

### Examples

Le ciel est triste et beau comme un grand **reposoir**.

Le ciel est triste et beau comme ton **regard**.

J'ai cité le ciel comme un **élément** atmosphérique.

- Identify the ground, i.e. the adjective or the verb that expresses the relationship uniting the compare NP to the standard NP.

### Examples

Le ciel est **triste** et **beau** comme un grand reposoir.

J'ai **cité** le ciel comme un élément atmosphérique.

- specify whether the underlined marker introduces a simile, a literal comparison or a pseudo-comparison;
- explain why the marker introduces that type of structure

### Examples

The standard of comparison designates a part of the comparee NP or vice versa.

The standard of comparison and the comparee NP are connected by a possession relationship.

The marker does not convey comparisons and is used with a verb of the type “consider, judge, elect...”

The standard of comparison and the comparee NP belong to distinct semantic categories and in this case, those categories must be added.

The standard of comparison is part of an extended phrase either by means of an adjectival phrases or a relative clause.

As illustrations, some semantic categories were proposed: Actions, Animals, Natural phenomena, People, Objects, Plants, Body parts...

The annotators, therefore, had at their disposal a colour code which facilitated their task and made it easier to see correlations between annotations.

### Figure 6.1 Sample of two annotations

#### Annotator 1

**Les danseuses**, **belles** comme **des fées** ou **des princesses**, sautaient et cabriolaient sous le feu des lanternes qui remplissaient leurs jupes d'étincelles.

*Comparaison figurative*

*Comparaison entre manières d'être*

*Catégorie sémantique : humain / personnages fabuleux*

#### Annotator 2

**Mon florin** que tu examines avec défiance à travers la loupe est moins **équivoque et louche** que **ton petit** **oeil gris**, qui **fume** comme **un lampion mal éteint**.

*Littérale puis figurative*

*3 catégories sémantiques différentes objet / partie du corps humain, objet*

In total 133 syntactic structures were annotated among which 81 were classified as being figurative, 25 as literal and 27 problematic cases, pseudo-comparisons or comparisons between processes. Annotator 2 had to review all her annotations as she first based her analysis on a more restrictive definition of similes, which shows that simile annotation is not so trivial as one would have thought. In addition, both annotators expressed their doubts at various occasions.

### Figure 6.2 An example of the annotator's hesitation and of a correction

Annotator 2

et sous les murs de Dijon, au-delà des meix de l'abbaye de St-Bénigne, le cloître de la Chartreuse, blanc comme le froc des disciples de saint Bruno.

Littérale ?

Relation métonymique : comparant et comparé sont en relation d'inclusion

Annotator 1

le souvenir des choses terrestres n'arrivait à mon cœur qu'affaibli et diminué, comme le son de la clochette des bestiaux imperceptibles qui paissaient loin, bien loin, sur le versant d'une autre montagne.

Comparaison littérale (le narrateur est dans la montagne) OUI mais on n'est pas dans le même système référentiel, mais la comparaison porte sur un procès, en fait, c'est « arriver à mon cœur » qui est déjà métaphorique (double sens entre parvenir dans l'intériorité et se déplacer dans un espace) je dirais FIGURATIVE

Cé (Ct déterminatif) et Ca (relative déterminative) GN étendus

Catégories sémantiques différentes : réalité mentale / son

Our final analysis of this dataset is based on 99 sentences, 77 similes and 22 literal comparisons. Interestingly, a change of semantic categories does not occur in 18 out of these 77 similes and in 6 of the 22 literal comparisons, which suggests that in these texts, a change of semantic categories is as frequent in similes as in literal comparisons and therefore, does not always signal figuration. Furthermore, although minimal noun phrases are more frequent in similes, they tend to be used more in closed similes. In this respect, it can be supposed that adding the ground salience as a simile feature could further help in the discrimination process.

### Example

Soudain le jaune papier de la lanterne s'enflamma, crevé d'un coup de vent dont gémirent dans la rue des enseignes pendantes comme des bannières.

Figurative

Catégorie sémantique commune : objet / enseigne et bannière présentent des propriétés communes (Annotator 2)

All in all, this dataset seems to imply that figuration does not have one single source but results from a combination of several factors, depending on the sentence. For instance, the

fact that the vehicle is part of an extended noun phrase can be important if the vehicle is the same word as the tenor:

### **Example**

Livre fermé comme le livre de sa destinée !

Figurative

Deux catégories sémantiques différentes : objet matériel / symbole religieux (Annotator 2)

## **6.3 The Crowdsourcing Annotation Platform**

According to Sabou, Bontcheva, Derczynski and Scharl (2014), crowdsourcing methods for corpus creation can be divided into three main types: tasks that are remunerated, games with a purpose and tasks which count on the goodwill of non-paid volunteers. Our annotation platform, which owes a lot to the previous experiments, falls in the last section.

As crowdsourcing became popular and more trustworthy, more and more online platforms were created to manage, store and distribute data to be analysed. As such platforms often already have readily made templates and a dedicated community, we decided to use one of them rather than investing in creating our own platform. We, therefore, chose to exploit the Zooniverse infrastructure which has proved itself very successful in the past in projects related to space, environment and the humanities (Simpson, Page & De Roure, 2014). As the only output accepted in Zooniverse and Zooniverse-related projects are images, we had to convert all the poem fragments into images. As much as possible, we aimed to give not only the sentence that contains a potential simile but also its surrounding sentences so that the annotator would have as much context as possible. Furthermore, a colour code was elaborated to differentiate the meaningful elements in the image from the background: so as to highlight the sentence to be analysed, it is featured in black in contrast to the other sentences that are in grey and the marker is easily recognisable by its blue colour. To avoid the repetitiveness of reading and annotating the same sentence various times, in case of multiple comparisons sharing the same tenor, all the markers are marked and that image is only presented once.

### Figure 6.3 Example of an image to annotate

A quick spin and shudder of brakes on an electric car, and the jar of a church-bell knocking against the metal blue of the sky. I am a piece of the town, a bit of blown dust, thrust along with the crowd. Proud to feel the pavement under me, reeling with feet. Feet tripping, skipping, lagging, dragging, plodding doggedly, or springing up and advancing on firm elastic insteps. A boy is selling papers, I smell them clean and new from the press. They are fresh like the air, and pungent as tulips and narcissus.

Amy Lowell (1874-1925) – "Spring Day".

The advantage of the Zooniverse platform is that coding is unnecessary: one just has the choice between questions with suggested answers and those that require an action such as delimiting an important space in the image (see Figure 6.4). However, as that platform does not give the possibility to transcribe texts and to avoid doing OCR at a later stage, we switched to the Scribe project,<sup>37</sup> which specifically tackles projects requiring transcriptions and offers more tools and liberties to set up a customised crowdsourcing project. In the final version of our project dubbed (Dis)Similitudes, we were, therefore, able to propose to users two types of tasks: the marking task which deals with marking an element of the text and/or with selecting applicable answers and the transcription task which ask to reproduce and/or to give additional information on the marked elements. The platform also allows users to concentrate on the task with which they feel the most comfortable: for example, as long as there is something marked, one can choose to dedicate oneself to transcribing and to completely ignore the marking task. In addition, a user can decide to transcribe a term immediately after it has been marked or to transcribe all the elements marked in the image once that image is completely done.

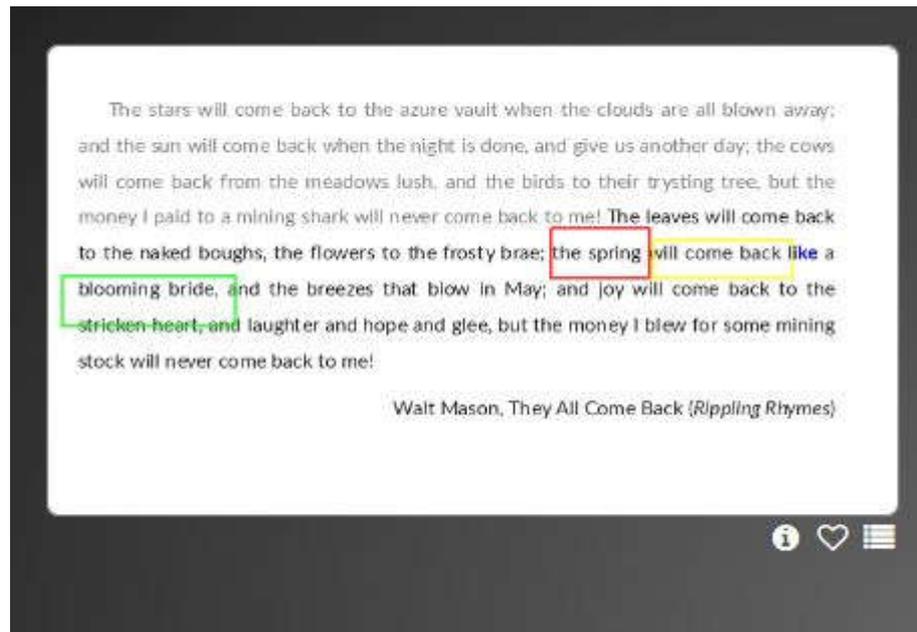
As we are working on two languages, we developed two versions of the platform:

- one in English for the English corpus: `dissimilitudes.lip6.fr:8181`
- and the second one in French for the French corpus: `dissimilitudes.lip6.fr:8180`

---

<sup>37</sup> <http://scribeproject.github.io/>

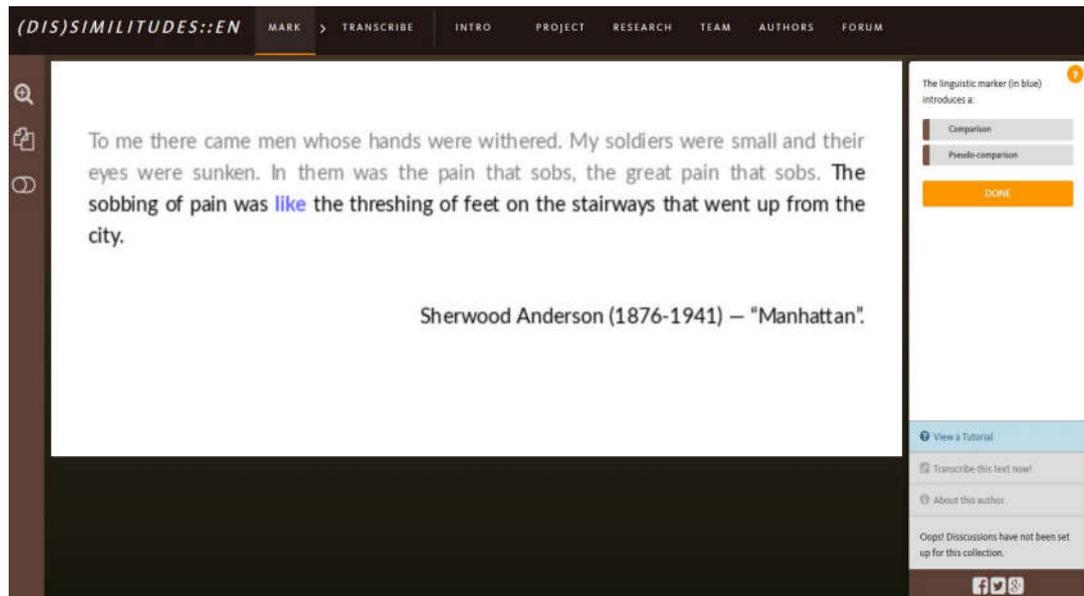
Figure 6.4 Example of an annotated sentence in the original Zooniverse interface



The main challenges of the design of the annotation tasks were to decide which information was required to be annotated and how to formulate questions as simply as possible for non-specialists. Depending on the focus of the question, we distinguished identification questions from descriptive ones. While the identification questions require to recognise a specific structure, the descriptive questions ask to further describe the nature of the sentence to be analysed, or whether it is a comparison or a pseudo-comparison.

This general question which deals with identifying the syntactic structure presented is the first question of the task (Figure 6.5) and as such, determines the subsequent information one will have to provide.

Figure 6.5 Starting question of the annotation platform



Possibility 1: The structure to analyse is a comparison.

Once a routine question has been asked on the presence or the reliability of existing annotations, each annotator has to answer to the four questions chronologically listed below:

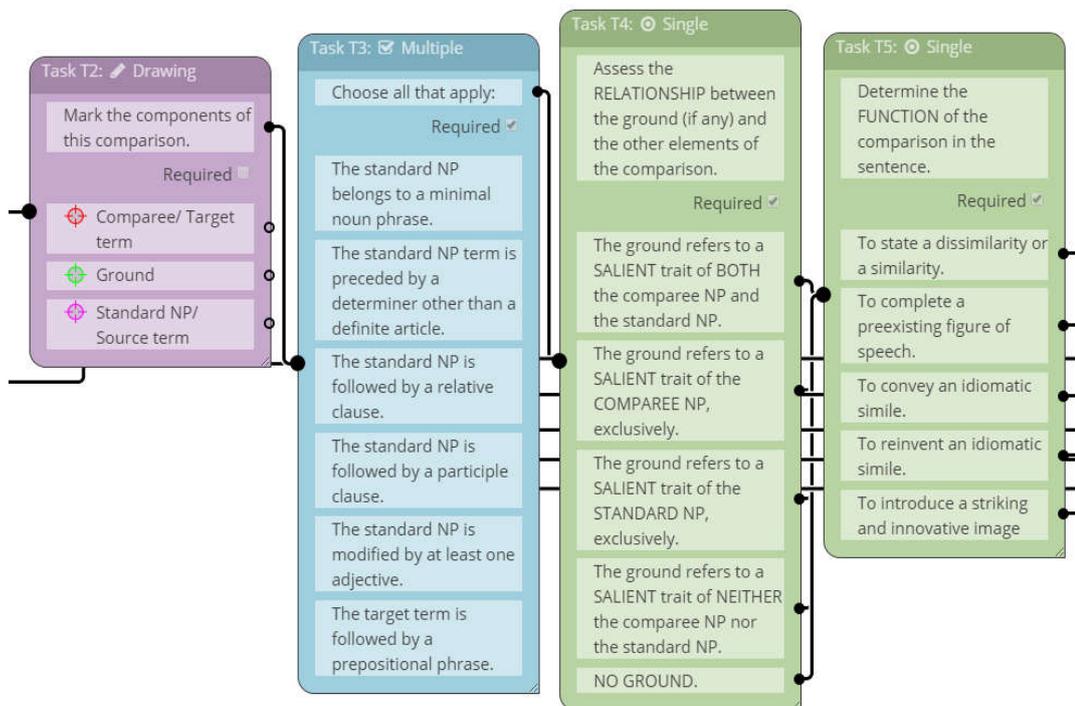


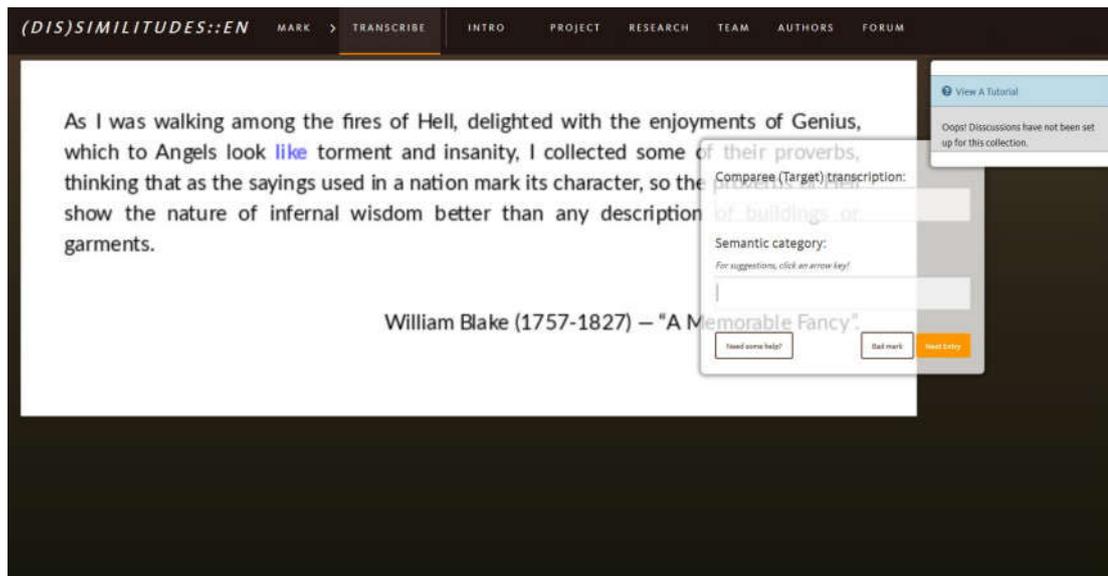
Figure 6.6 Example of an annotated sentence

The screenshot displays a web interface for annotating literary text. The main content area shows three paragraphs of text. The second paragraph, "The corn stood up like armies in the shocks.", is annotated with colored circles: a red circle over "corn", a green circle over "stood up", a blue circle over "like", and a purple circle over "armies". To the right of the text is a sidebar with instructions and a list of semantic categories. The instructions state: "It seems that there is no previous marking or that the existing marking is wrong. Mark the components of this comparison. Confirm by clicking on DONE, next to each mark you placed." The sidebar lists three categories: "Compare / Target term (1)" with a red icon, "Ground (1)" with a green icon, and "Standard NP / Source term (1)" with a purple icon. Below the list is an orange "NEXT" button. At the bottom of the sidebar are links for "View a Tutorial", "Transcribe this text now!", "About this author.", and a message: "Oops! Discussions have not been set up for this collection." Social media icons for Facebook, Twitter, and Google+ are at the very bottom.

In the case of a comparison, the transcription task concerns each of its elements and the annotator has to indicate the semantic category of each of them. To narrow the possibilities, the semantic categories mentioned in the previous chapter are proposed. In addition, if the simile is motivated, the type of ground used must be filled, whether it is an adjective phrase, a clause, an adverbial phrase or verbal phrase. As the final aim of this work is to produce a gold annotated corpus, we found it reasonable to go beyond the actual capacities of the developed method, by letting, for example, people identify adjectival placed after the vehicle.

We deliberately put the part concerning semantic categories in the transcription task because we noticed, in our different tests, that it tended to puzzle and to discourage various annotators when it was asked, for comparative structures, at the beginning of the marking task.

Figure 6.7 Example of a transcription task

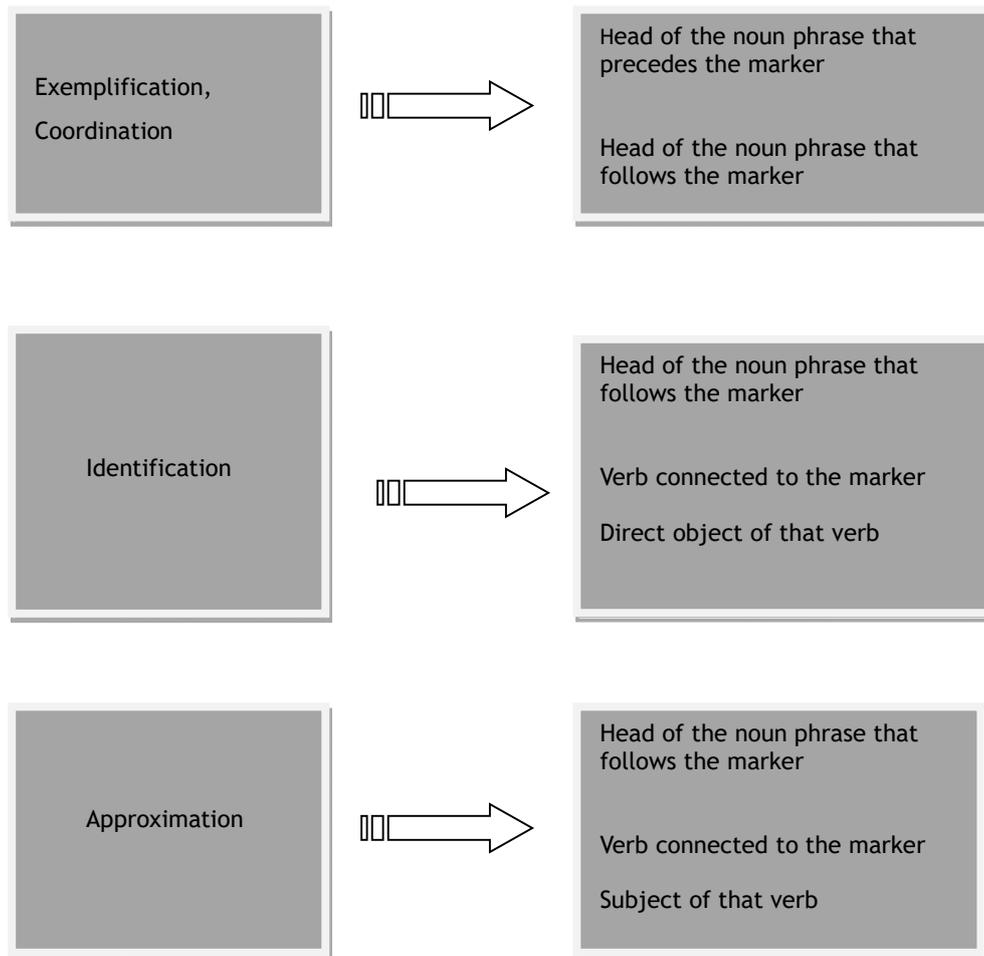


Possibility 2: The structure to analyse is a pseudo-comparison.

Here, still after the routine question, the annotator is first asked to choose the value of the pseudo-comparison and to decide whether it is an exemplification, a coordination, an approximation or an identification. Then, the components need to be marked. The semantic categories are asked in the transcription task only for pseudo-comparisons expressing exemplification and coordination; for the rest, only a transcription of the related word or phrase is enough. Figure 6.8 gives for each semantic value of a pseudo-comparison, the sentence elements that need to be provided.

It is worth noting that for the marking task, the user has to put a target somewhere on the text, as it is that mark that would launch the transcription task. If on the one hand, it can be argued that a target is not the ideal form particularly to capture phrases made up of multiple words, on the other hand, it is interesting to see whether the transcribers would consider the whole phrase or only the head. Most annotators faced with this choice tend to mark phrases and not simple words, which confirms that the whole phrase has its importance.

**Figure 6.8 Elements associated with each subtype of pseudo-comparison**



As the platform has only been recently launched, few conclusions can already 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, still at this embryonary level, it could already provide some valuable information on the perception of semantic categories and on the validity of our dictionary-based matching between lexemes and the preselected semantic categories. For the French corpus, for 15 similes containing 28 terms among which 6 were annotated by more than one person, if the broader semantic categories fit almost perfectly with human annotations (98%), the score is slightly lower (67%) when it comes to further semantic distinctions. In addition, annotations seem particularly to oscillate between different subcategories 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.

**Example**

- Cri      Abstract entities – Concepts (1)
- Abstract entities – Acts and processes (3)
- Abstract entities – Feelings and emotions (2)
- Abstract entities – Others (2)

The correlation observed between the English annotations and the semantic categories confirm the same tenency but is less promising. On a total of eleven similes, 14 terms were described semantically by only one annotator: while 64% of the broad semantic categories matches with human annotations, it is almost divided in half when it comes to refined semantic categories (35%). These results could be explained by wrong annotations ({{'sourceTranscript': 'smoke', 'sourceCategory': 'Collective nouns'}}) and faulty semantic categories due to polysemy: for example, the semantic category assigned to “cat” is “Living beings - Human beings” instead of “Living beings - Animals” whereas “hair” is labelled as “Objects - Man-made”. These mistakes should therefore be taken into consideration or corrected before the final evaluation.

This chapter describes our efforts to create a corpus which could be used to validate simile recognition methods and to study among others the perception and the origin of figuration. In this respect, we started with a small-scale experiment before shifting to an online platform. We hope in the near future, not only to be able to make the resulting annotated corpus freely available for other researchers but also to use it to have a more global view of our method’s strengths and weaknesses.

# 7 CORPUS-BASED APPLICATIONS

As the goal of this thesis was not to focus on a particular author but to examine similes as a whole, in order to explore the relevance of the automatic extraction of similes to literary scholars, the following topics were investigated with the help of corpus-based methods: stereotypical frozen literary similes, colours in similes and the use of proper nouns in comparative constructions. In this respect, in a cross-linguistic perspective, a corpus of novels in English and French first had to be compiled. After explaining the corpus design, each application and its results will be presented and discussed.

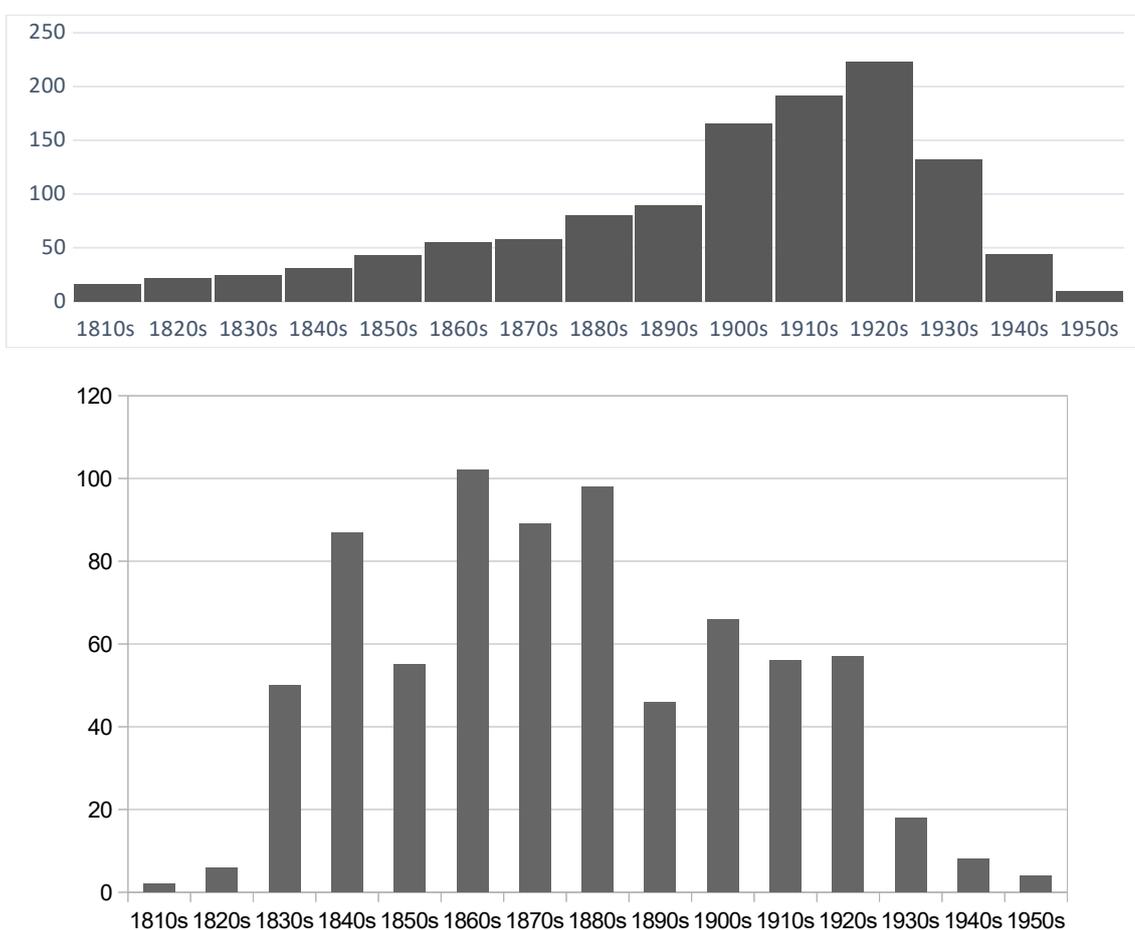
## 7.1 Corpus Description

In order to ensure linguistic homogeneity, only novels published between the 19th and the mid-20th century were included in the corpus, spanning about 150 years. Apart from covering different literary periods, the resulting corpus combines different literary genres (historical novels, detective novels, adventure novels, novels of manners, science-fiction novels...). In addition, a ratio of at least three novels per novelist was observed. This method enabled to create a corpus of 1,191 British texts (1,188 novels and three short

stories) authored by 62 writers and a corpus of 745 French fictional texts penned by 55 novelists (see Appendices 6 & 7). All novels were downloaded from the Project Gutenberg's main website,<sup>38</sup> from the Project Gutenberg Australia's website,<sup>39</sup> and from the Bibliothèque électronique du Québec website.<sup>40</sup> In terms of size, the British corpus contains 152,941,750 tokens and its French counterpart, 119,914,914 tokens.

Since only already digitised novels were considered, there exists, of course, a certain discrepancy in their distribution, as can be seen in Figure 7.1: the bulk of the British corpora has been published between 1900 and 1920 while more than half of the French corpora is made up of books published between 1850 and 1880.

**Figure 7.1 Distribution of the novels in the British (top) and the French (bottom) corpora per decade from the 1810s to the 1950s**



<sup>38</sup> <https://www.gutenberg.org/>

<sup>39</sup> <http://gutenberg.net.au/>

<sup>40</sup> <http://beq.ebooksgratuits.com>

## 7.2 Stereotypical Frozen Literary Similes

In most of the literature (Wilstach, 1916; Cazelles, 1996; Parmentier, 2002), idiomatic similes generally follow two specific patterns: adjectival ground + simile marker + nominal vehicle (e.g. *cunning as a fox*) and verbal ground + simile marker + nominal vehicle (e.g. *cry like a baby*). In addition to these patterns, we propose to investigate triplets of the forms nominal tenor + adjectival ground + simile marker + nominal vehicle and nominal tenor + verbal ground + simile marker + nominal vehicle as well as strong associations between a nominal tenor and a nominal vehicle irrelevant of the ground used.

Generally speaking, independently from the type of similes involved, the rather low overall frequency of usage of the frozen similes extracted in each corpus tends to confirm the fact that literature is indeed a place where linguistic innovation is typically prioritised and preferred. Another interesting fact is what a closer look at the most recurring frozen similes in both languages reveals. Not only are the same similes featured prominently in both corpora, but “death” and “whiteness” are the most frequently used themes. As a matter of fact, both concepts are connected in “pale + marker + death” or “white + marker + death” in which pallidness is transferred from a corpse to death itself, which is personified in the process. In addition, as far as the British corpus is concerned, it is important to note the presence of three very fixed expressions with little figurative potential: “as good as one’s word”, “worse than death” and “as good as gold”.

**Table 7.1 The 10 most frequent similes in both corpora**

French	English
pâle + marker + mort [283]	speak + marker + man [283]
pleurer + marker + enfant [185]	good + marker + word [227]
blanc + marker + neige [162]	pale + marker + death [164]
immobile + marker + statue [154]	treat + marker + child [144]
tomber + marker + masse [139]	cold + marker + ice [131]
aimer + marker + frère [138]	bad + marker + death [121]
pâle + marker + morte [121]	stand + marker + statue [119]
tuer + marker + chien [120]	white + marker + death [112]
trembler + marker + feuille [114]	good + marker + gold [112]
passer + marker + éclair [112]	white + marker + sheet [103]

From the most frequent triplets (tenor + ground + marker + vehicle), it can be seen that formulaic similes like “blood is thicker than water” or “one’s bark is worse than one’s bite” are indeed most commonly found in English as well as the fact that this type of constructions often concern similes describing a body part at a particular moment in time (see Table 7.2). For example, in French, novelists seem to revert to the same image when it

comes to describing a shrill voice with the means of a simile. Besides, for certain elements, only a restricted number of the tenor's attributes appear to be considered and developed with a simile. For instance, in both languages, the eyes are mainly described in terms of their brightness.

**Table 7.2 Most frequent triplets in both languages with the degree of fixedness of the vehicle**

English	French
bark + bad + marker + bite [1]	voix + bas + marker + souffle [1]
blood + thick + marker + water [0.93]	voix + faible + marker + souffle [0.95]
face + set + marker + flint [0.91]	(intonation, voix) + doux + marker + chant [0.88]
heart + beat + marker + hammer [0.85]	dent + blanc + marker + perle [0.84]
vein + stand + marker + whipcord [0.8]	(dent, main, front, squelette..) + blanc + marker + ivoire [0.75]
money + spend + marker + water [0.77]	voix + léger + marker + souffle [0.73]
eye + wide + marker + saucer [0.76]	(minute, seconde) long + marker + siècle [0.66]
vein + stand + marker + cord [0.76]	(cheveu, chevelure, boucle, sourcil) + noir + marker + jais [0.65]
face + be + marker + mask [0.71]	œil + briller + marker + charbon [0.65]
eye + burn + marker + coal [0.71]	œil + briller + marker + escarboucle [0.63]
eye + glow + marker + coal [0.64]	(geste, mouvement) + prompt + marker + pensée [0.63]
skin + be + marker + parchment [0.64]	(pensée, idée, souvenir) + traverser + marker + éclair [0.61]
eye + bright + marker + star [0.57]	

The relationship between descriptions of some body parts and stereotyped images is further confirmed by the most frequent pairs identical nominal tenor – nominal vehicle and nominal tenors of the same semantic domain – nominal vehicle, which, in some cases, bring about new images (see Table 7.3). If in French, “voix” is still present, associated with “clairon” it does not imply an idea of shrillness like before, but rather clarity and wide range. Similarly, in English, the eyes are no more depicted in terms of brightness but rather in terms of their size. It is also important to note in both languages the image of the vice which is always linked with parts of the body that are tightly pressed together or that firmly press something else (arms, hands, fingers...).

**Table 7.3 Most frequent pairs in both corpora**

English	French
eye + gimlet [0.7]	(bras, main, tempe, crâne, mâchoire) + étau [0.77]
(skin, face) + parchment [0.6]	(voix, hennissement, parole, cri) + clairon [0.73]
(finger, hand) + clay [0.59]	(cheveu, chevelure) + crinière [0.72]
face + mask [0.54]	nez + bec [0.7]
(arm, grip, hand) + vice [0.53]	oeil + escabourcle [0.6]
eye + saucer [0.52]	
(vein, muscle) + whipcord [0.5]	

From a cognitive point of view, it is, therefore, possible to suggest that according to our corpus of novels, novelists tend to fall back on common imagery when they are talking about a permanent or a temporary state of a body part, typically the eye. From this study, it

can also be inferred that frozen similes do not concern a single form but rather a family of similar similes that renders the same idea. For instance, “voix + bas + marker + souffle”, “voix + faible + marker + souffle” and “voix + léger + marker + souffle” are three renditions of the same simile.

## 7.3 Colours and Similes in the English Corpus

### 7.3.1 Why Study Colours in relation to Similes?

One of the main reasons that fictional texts succeed in resonating with their readers is their use of strong visual images which enable one to re-create a scene or to picture a character as if he or she were physically there. Colours, in particular, play a crucial role in shaping those visual images, not only because they make descriptions more vivid but also because a wide range of connotative meanings is culturally associated with specific colour terms. For example, whereas in the Western world, the colour “black” is generally associated with death and mourning, in the Eastern world, this role is devoted to the colour “white”. In addition, if the colour “white” is generally linked to purity and goodness, its opposite “black” evokes evil as well darkness and the colour “red” can, depending on the circumstances, refer to fury, flame or even embarrassment (Philip, 2006). In this respect, the scarlet letter in Nathaniel Hawthorne’s eponymous novel does not only indicate to others that the woman who wears it has committed adultery, but also keeps her in a state of perpetual shame, the colour scarlet being presented in the Bible as the colour of sin, the colour of the garment of the prostitute depicted in Revelations 17:4. Similarly, in *La Comédie Humaine*, Balzac adheres to a popular medieval belief by systematically assigning green, yellow or orange garments as well as physical attributes to his malevolent characters (Vanoncini, 2004).

As with word arrangement, writers have notably been known for how they manipulate colours either by giving them new connotative meanings or by exemplifying idiosyncratic colour usage worthy of a painter’s palette. If we go back to what has been said earlier about the colour “white” in the Western culture, the title of Webster’s play, *The White Devil* sounds at first like an oxymoron. However, it takes all its meaning in the whole colour imagery of the play in which “whiteness” is depicted as being the colour of hypocrisy and as such, far more deceptive than “blackness” (Connolly & Hopkins, 2015). Moreover, the semantic field of both concepts in that play shows that the idea of ‘blackness’ is conveyed not only through the colour adjective “black” but also through its synonym ‘dusky’ and

several compound nouns containing the word “black” (**blackbird**, **blacklust**, **blackthorn**), whereas the adjective “pale” is used as a synonym of “white”. As a matter of fact, expressions of colours can take various forms in literary texts, from single nouns (*the green of her eyes*), verbs (*embrown*) and adjectives (*sulphurous light*) to compound adjectives (*fiery-red complexion*), noun phrases (*the colour of tallow*) and fully fledged similes (*brown as a gipsy*), depending on the impact and hue the author seeks to achieve. In terms of pictorial precision, it can be hypothesised that complex expressions of colours offer more creative liberties to writers as they make it possible to blend different colours (*large eyes violet-bluey-blackish*), to circumscribe the coloured area (*red-faced*) and even to pinpoint the intended shade of a particular colour by mentioning a prototypical object or phenomenon which possesses it (*gem-green*).<sup>41</sup>

As far as colour similes are concerned, it can, however be argued that the degree of figurativeness of an occurrence such as “Her cheeks are as red as roses” is lower than that of an open simile like “Her cheeks are roses”, which could rely on various other possible salient traits of roses such as their beauty, their delicateness, their warmth or their softness. Even though Ortony (1979) agrees that the similarity between the vehicle and the tenor in colour similes can also be built on attributes inherent to the colour itself such as hue, saturation and intensity, he argues that colour attributes are so high-salient that they tend to eclipse other common attributes shared by the vehicle and the topic and reduce the figurativeness so much so that the resulting simile is very close to a literal comparison. Addison (1993), on the contrary, still considers colour similes as similes albeit *literal* ones, since the compared entities do not belong to the same semantic field.

With regard to colour and figurative language, Philip (2006) notes that the high saliency of colours makes them rather adequate to be used in a figurative sense as they can only be successfully applied to apt and valid comparisons. Consequently, it is possible to distinguish between on the one hand, idiomatic colour similes which are fixed collocations in a given language whose meaning cannot be inferred from the combined meaning of its constituents, and, on the other hand, their variations, creative colour similes, which explore more widely the spectrum of shades a particular colour can take. Moreover, for a creative colour simile to be easily understandable, it must ideally rely on shared cultural beliefs and use as vehicle an object that typically exemplifies that particular colour. In this respect, “whiter than dried rice” would be considered a fairly accessible variation of “whiter than white”, the canonical English to express extreme righteousness, as it requires far less

---

<sup>41</sup> All examples are taken from Peprnik (1996) and Peprnik (2000).

thought and processing than “whiter than next week’s improved detergent”, a more opaque variation of the same simile (Philip, 2006).

The present subsection is focused on particular forms of complex expressions of colours that can either take the form of fully fledged similes (*brown as a gipsy*) or synthetic similes (*gem-green*). Even though, unlike fully fledged similes, synthetic ones are not built around a comparison marker, writers equally use both types of similes to communicate subtly with their readers by soliciting their imagination as well as their own perception of the colours of elements of the world. Thus, this subsection intends to answer the following questions: Are synthetic similes use differently and do they fulfil a different stylistic purpose than fully fledged ones? What specific features distinguish creative synthetic similes? Finally, since synthetic similes combine a noun and a colour term, what can be said about the choice and the distribution of colour terms used by British writers?

### 7.3.2 Basic Colour Terms and English Literature

In the field of optics, colours can be described as the way an eye’s retina interprets light wavelengths: “light can be made up of a mixture of these colours, and can occur at varying intensities. Hence the perceived phenomenological colour of light depends on which wavelength are present, and on the intensity of each wavelength” (Dowman, 2001). Of course, in each natural language, some words also called colour terms (CTs) have been coined to discriminate between these perceived light wavelengths. Consequently, as language universals, colour names have often been used to compare conceptual systems of different languages and therefore, to fuel the debate on the arbitrary nature of meaning. Up until the influential work of Berlin and Kay (1969), the inconsistency and differences characterising colour separation in various languages were considered sufficient proof of the interdependence of linguistic semantic systems (Leech, 1981; Steinvall, 2002). Berlin and Kay (1969), however, notice that some colours are too easily translated between unrelated languages and set out to investigate whether it is simply coincidental. From their experiment on the mapping of basic colour terms in about 20 languages, they conclude that although the number of basic colour terms may vary from one language to another, these colour terms are always taken from a fixed set of eleven basic colour categories: “white”, “black”, “red”, “green”, “yellow”, “blue”, “brown”, “purple”, “pink”, “orange” and “grey”. A language such as English, for example, possesses all of these eleven basic colour terms, i.e. colour terms that are monolexemic, whose meaning is not included in another colour term, which can be applied to an unrestricted range of objects and are psychologically salient (Berlin & Kay, 1969). In addition, these basic colour terms,

irrelevant of the language, invariably follow the same order of appearance illustrated in Table 7.4.

**Table 7.4 Basic colour term depending on the number of colours expressed in the language**

LANGUAGE CHARACTERISTICS	COLOUR TERMS
LANGUAGES WITH 2 COLOURS	white and black
LANGUAGES WITH 3 COLOURS	white, black and red
LANGUAGES WITH 4 OR 5 COLOURS	white, black, red, green <i>and/or</i> yellow
LANGUAGES WITH 6 COLOURS	white, black, red, green, yellow and blue
LANGUAGES WITH 7 COLOURS	white, black, red, green, yellow, blue and brown
LANGUAGES WITH 8 OR MORE COLOURS	white, black, red, green, yellow, blue, brown, purple, pink, orange <i>and/or</i> grey

Several research works, however, have highlighted important biases that put into question the veracity and the objectivity of Berlin and Kay's results. The conception of the experiment, in particular, is often criticised, mainly because of the initial subjectivity of the researchers, the lack of geographical diversity as well as adequate scanning of the participants (McIntyre, 2009), and the use of the Munsell chart which, besides being an American standard, limits the participants to an already predefined conception of colours (Dubois & Grinevald, 1999). Moreover, Berlin and Kay's definition of colours has been qualified as being rather ethnocentric for two main reasons:

- it does not consider that, unlike English, some languages could have more than one term for a colour as in the case in Russian for blue or that some colour terms could be polysemous as it is the case in Scottish Gaelic (McIntyre, 2009);
- it restricts colours and consequently basic colour terms to chromatic properties, eliminating words denoting material entities used as colour terms such as it is the case in Jale, a language spoken in New Guinea, which does not possess a particular word for "green" but uses "pianó", the name of a plant used to dye yarn, to refer the particular hues of green (Dedrick, 1998).

Despite these flaws in the methodology adopted, the influence of the Berlin and Kay's hypothesis (1969) on the linguistic community as a whole cannot be undermined, especially since their results are particularly precise with regard to the order in which colour terms become part of languages of the world. In this respect, its veracity has been tested against different languages and even in written material such as literature and newspaper articles. By compiling the frequencies of colour terms from Pratt's analysis of the usage of colours by 17 British Romantic poets (1898) and those of two corpora, one of Chinese poetry and the other one of modern novels, McManus (1983) observes that all these frequencies strongly correlate the hierarchical order proposed by Berlin and Kay (1969). Besides, as far

as poetry is concerned, the earlier a colour term has entered the English language, the more it is used by poets, which tends to suggest that more dated colour terms are more psychologically salient for poets and thus, are favoured either because they are more connoted and richer in meaning or either because of their synaesthetic properties. Moreover, even though some authors use some colours more extensively than others, the relative frequency of each colour per author remains constant overall.

In contrast, the same experience performed diachronically on a corpus of French novels, poems and plays published between 1500 and 2000 reveals two main facts:

- apart from “*blanc*” (“white”), “*noir*” (“black”) and “*rouge*” (“red”), which are always respectively the first, the second and the third most frequent colour term used by French authors, the frequency order of the remaining colour terms does not really respect the Berlin and Kay’s hypothesis (1969) and changes from one century to another;
- as time goes by, colour terms are more and more used in literary texts, which could be seen as proof either that literature reflects a world that produces a mass of objects that have to be differentiated by their colour or that conveying sensory experiences has gained more importance in literature (Cheminée, Dubois & Resche-Rigon, 2006).

With the method described in Appendix 5, around 2,280 pertinent noun+CT similes were found in all the novels in the corpus except the ones written by Jane Austen and Lewis Carroll. In addition, we also plotted colour similes ending with the suffix “-coloured” and noticed that this suffix is mostly used after one of the related non-basic CTs.

**Table 7.5 Colour terms selected for the experiment (the asterisk signals a colour term that also refers to the hue of another colour term)**

Basic CT	Related non-basic CTs
white 	silver*, eggshell, ivory, magnolia
black 	ebony, raven, sable
red 	cardinal, carmine, carnation, cerise, cherry, cinnabar, claret, crimson, fuchsia, garnet, magenta, maroon, murrey, roan, rubby, sandy, scarlet, stammel, vermeil, vermilion, wine, heather*, coral*
green 	celadon, chartreuse, jade, myrtle, pistachio, verdigris, turquoise*
yellow 	amber*, apricot*, buff, champagne, citrine, crocus, daffodil, flaxen, gilt, gold, golden, jasmine, lime, maize, ocher, peach, primrose, saffron, sand, straw, sulfur, sulphur, camel*, rust*, coral*
blue 	aqua, azure, cerulean, indigo, lavender, mauve, periwinkle, sapphire, teal, violet, indigo, azure, turquoise*
brown 	amber*, auburn, bay, biscuit, bistre, bole, bronze, chestnut, chocolate, cinnamon, coffee, copper, dun, ecru, fallow, fawn, ginger, hazel, khaki, liver, mahogany, russet, tan, tawny, umber, beige*, burgundy*, rust*, camel*
purple 	burgundy*, eggplant, heliotrope, lilac, mulberry, orchid, petunia, plum, puce, heather*, mauve
pink 	bisque, blush, damask, rosy, salmon, apricot*, rose
orange 	Tangerine
grey 	beige*, ash, dove, pewter, slate, cineritious, drab, platinum, taupe, silver*

### 7.3.3 Fully Fledged Colour Similes vs. Noun+CT Similes: Frequency and Stylistic Usage

Three main facts concerning the results of the extraction task show that fully fledged colour similes and noun+CT similes are not used interchangeably:

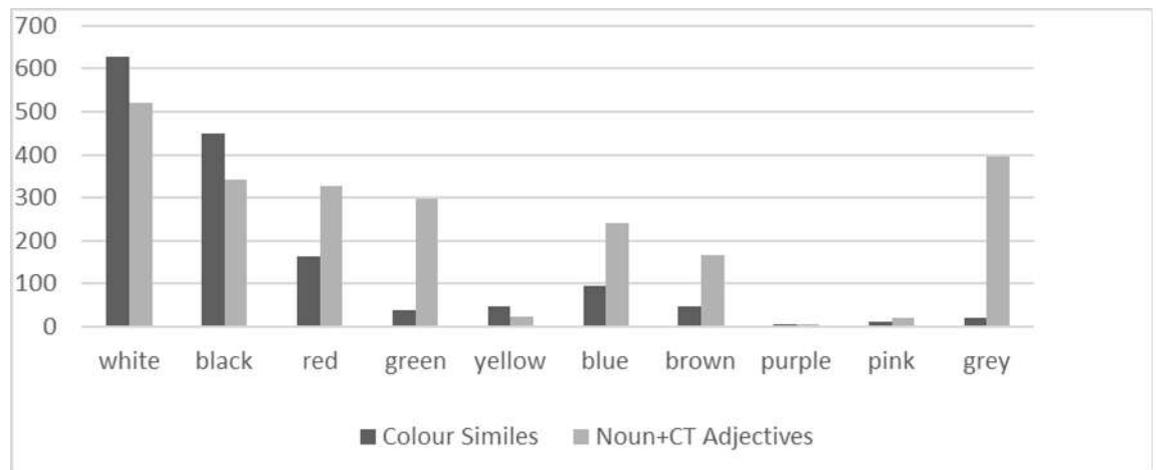
- the corpus contains far more noun+CT similes than fully fledged colour similes (about 1,550 occurrences);
- the frequency of some colours terms differs drastically from one type of similes to another (see Figure 7.2);
- the most frequent prototypical vehicle for the same colour term also often varies from one type of simile to another as illustrated in Table 7.6. Similarly, some topics seem to be preferred depending on the type of similes used.

**Table 7.6 Pattern distribution of the 5 most frequent colour terms used in noun+CT similes**

CT	Prototypical example(s) given in the <i>GCIDE</i>	Noun+CT Similes		Fully Fledged Colour Similes	
		Most Frequent Noun+CT Adjective	Most Frequent Tenor(s)	Most Frequent Tenor(s)	Most Frequent Vehicle(s)
white	snow	snow-white (283/511)	hair	People and body parts	death (109/629)
black	soot, coal	coal-black (135/337)	hair and horses		night (71/448)
red	blood	blood-red (244/326)	complexion and light		blood (18/164)
gray/grey	pepper, salt, ashes, hair whitened by age	iron-grey (163/315)	hair	-	wing (4/22)
green	growing plants or grass	sea-green (69/288)	linen and clothes	places	emerald, glass (7/37)

As far as non-basic colour terms are concerned, apart from the fact that they are less used in similes, there is no correlation between the noun+CT similes and fully fledged colour similes: for the first type of similes, only instances of “rose” (10), “bay” (1), “gold”(5), “silver” (2), “scarlet” (1), “gilt” (1) and “mauve” (2) were found, whereas for the second type, the ground of the simile is generally inflected forms of existing colour terms (“rosy”(17), “scarlet”(4), “bluish” (1), “blond” (1), “silver” (1) and “greeny” (1)). It is worth noting that in both types of similes, the most frequent colour terms refer to the colour pink, surpassing even the frequency of the colour term “pink” itself.

**Figure 7.2 Colour term distribution in fully fledged colour similes and noun+CT similes**



Grossly speaking, the frequency of occurrences of basic CTs in both types of similes (Figure 7.2) does not confirm the trend observed in McManus (1983) and thus, questions the Berlin and Kay's hypothesis (1969). The frequency of the CTs used in noun+CT similes, especially, differs widely from their hypothesis: although "white" is the most frequent CT, it is followed directly by "grey/gray" which should logically be the least frequent CT, while "blue" and "brown" are far more recurrent than "yellow" and "green". On the contrary, as far as fully fledged colour similes are concerned, the hierarchical order proposed by Berlin and Kay (1969) is more or less strictly respected, apart from the frequency of "blue" which surpasses that of "yellow".

It seems obvious that the bulk of the selected non-basic CTs are not used in similes because they are not really connoted in the English language. Alongside connotation, the role of collocation and foregrounding cannot be undermined. Similes are successful when they are creative, surprise the readers and outline a behaviour or characteristic that the author wants to stress. In this respect, a simile that makes use of a prototypical vehicle is not very creative. Therefore, instead of wasting similes on minor background elements such as body parts, clothes or animals, authors tend to transform them into noun+CT similes, and in the process, make them less prominent in the sentence. That could explain why the frequency of "blood-red" is more than ten times that of "red as blood" or why there is a perceptible difference in meaning between (7) "But there was no lack of animation in her little steel-grey eyes, nor of decision in her manner" (J. Galsworthy, *The Patrician*, 1911) and (8) "Uncomfortable under those stern searching eyes that were as grey as steel and as cold, Pablo shifted on his feet, shrugged and put on a sneering brag" (R. Sabatini, *Columbus*, 1941): whereas in the first sentence, the focus is on the colour of the eyes that resembles the colour of steel, in the second one, the simile highlights the coldness of the eyes that is reminiscent of the coldness of steel.

Apart from the semantic and conceptual differences between noun+CT and fully fledged colour similes, a close study of the extracted sentences also shows some restrictive uses of each type of similes. Unlike their counterparts, fully fledged similes make it possible:

- to combine the colour term with a second ground that enhances its meaning: (9) The room was vacant; the room was black and silent as a dungeon. (G. Meredith, *The Tragic Comedians*, 1880);
- to emphasise the purity or the brightness of the colour through a hyperbole: (10) They went home in a motor-bus and a cloud of dust, with the heaven bluer than blue above, the hills dark and fascinating, and the land so remote seeming. (D. H. Lawrence, *Kangaroo*, 1923);
- to take advantage of the connotative meaning of the colour term: (12) He looked as black as night when he caught sight of us (E. Glyn, *Red Hair*, 1905).

In contrast, since noun+CT adjectives are compound adjectives, writers exploit this structure to create innovative and striking associations that go far beyond merely describing a colour hue. The corpus therefore contains:

a) Metonymical similes

In these similes, the topic refers to a part or a quality of the vehicle. In examples (12) and (13), it would not make any sense to simply transform the noun+CT as CT+noun simile to obtain “teeth as white as an animal” or “eye as grey as a fish”. In both sentences, the flexibility of English syntax is used to coin new adjectives and create a dual meaning. As a matter of fact, apart from meaning that the teeth are as white as an animal’s, “*animal-white teeth*” also implies sharpness and ferocity. Similarly, while the expression “*rat-brown eyes*” suggests slyness because of the connoted meaning of the word “rat”, in “*velvet-green moss*”, velvet evokes the inherent softness of the fabric which is lent to the moss.

**Examples**

(12) They longed—or dreaded—to stand within that huge cavern of blue lonely ice and hear the waves of the Polar Sea lick up the snow; to taste that sugary cane with animal-white teeth, and feel the fluffy cotton between thick, lumpy fingers; to swim under water and look up instead of down; to crawl fearfully a little nearer to the molten centre of the planet through smoke and fire and awful thundering explosions. (A. Blackwood, *The Promise of Air*, 1918)

(13) The Reverend Mr. Arbroath started indignantly, and stared so hard that his rat-brown eyes visibly projected from his head. (M. Corelli, *The Treasure of Heaven*, 1906)

(14) They had rested here; he sitting on the weatherworn parapet of the bridge; she leaping over it, and idly dropping bits of velvet-green moss into the whirl of clear brown water below. (W. Black, *MacLeod of Dare*, 1878)

b) Cause and effect similes

This type of similes generally associates a natural element or phenomenon such as the sun or the winter to the colour it casts on a particular entity.

**Examples**

(15) "I think I shall go and bathe," said Miss Inger, out of the cloud-black darkness. (D. H. Lawrence, *The Rainbow*, 1915)

(16) With strong weather-brown fingers she tried to close the tiger’s staring eyes. (T. Mundy, *Full Moon*, 1935)

(17) She basked with him on the edge of a rock and gazed over the ten—or was it twenty? — miles of snowy wilderness; then they turned their tinted glasses on the knife-edge of the

Jungfrau summit, its outline crystal-yellow against a storm-green sky. (J. Hilton, *Contago*, 1932)

c) Reinvented conventional noun+CT similes

These similes play on the various colours some objects can have in the universe and call into question the supremacy of the colours predominantly used with a particular vehicle (cf. Table 7.7).

**Table 7.7. Examples of reinvented conventional noun+CT similes**

	Proposed Alternative(s)
iron-grey/iron-gray	(18) The blue sky settled against them nakedly; they were leafless and lifeless save for the <u>iron-green</u> shafts of the organ cactus, that glistened blackly, yet atmospherically, in the ochreous aridity. (D. H. Lawrence, <i>The Plumed Serpent</i> , 1926)  (19) A single tent stood in a gully running from one of the gravel-pits of the heath, near an <u>iron-red</u> rillet, and a girl of Kiomi's tribe leaned over the lazy water at half length, striking it with her handkerchief. (G. Meredith, <i>The Adventures of Harry Richmond</i> , 1871)
peacock-blue	(20) The only light other than stars glowed through one <u>peacock-green</u> curtain in the upper part of the building, marking where Dr. Emerson Eames always worked till morning and received his friends and favourite pupils at any hour of the night. (G. K. Chesterton, <i>Manalive</i> , 1912)
ink-black	(21) The lush, dark green of hyacinths was a sea, with buds rising like pale corn, while in the riding the forget-me-nots were fluffing up, and columbines were unfolding their <u>ink-purple</u> ruches, and there were bits of blue bird's eggshell under a bush. (D. H. Lawrence, <i>Lady Chatterley's Lover</i> , 1928)
bottle-green	(22) He looked up towards the ingenuous, protruding, shining, liquid, <u>bottle-blue</u> eyes of Thomas Johnson... (Ford, <i>No More Parades</i> , 1925)

d) Unfamiliar noun+CT associations

The last type concerns the more opaque noun+CT similes which make use of unusual colour associations, so much so that the readers need to use a significant amount of imagination to understand them.

**Examples**

(23) The dark forest of karri that ran to the left of Wandoo away on the distant horizon, cut a dark pattern on the egg-green sky. (D. H. Lawrence, *The Boy in the Bush*, 1924)

(24) The fly-driver touched his age-green hat with his whip. (F. M. Ford, *Some Do Not*, 1924)

(25) "Then he stooped down, and put his lips to the cold clay-blue forehead." (A. Trollope, *Ralph the Heir*, 1871)

Overall, while on the one hand, noun+CT adjectives are used to perpetuate prevailing noun-colour combinations, on the other hand, they seem to provide to British novelists

more latitude to play around and to show that colours are not as fixed as we think in the surrounding world. Like fully fledged similes, are noun+CT similes also made up of traditional and creative similes? What can they tell us about the period in which they were written?

### 7.3.4 Creativity and Noun+CT Similes

Creativity is a key question in literature, especially as far as stylistic devices are concerned. At first glance, frequency and fixedness seem to be pretty good criteria to judge how creative a noun+CT adjective is. The rationale, in this case, is fairly simple: if a term is only used by one writer, it is highly plausible that it is a creative noun+CT adjective. However, it is important to also take into consideration the collocations existing in the language. For example, “ebony-black” appears only once in the corpus whereas “black as ebony” occurs 18 times, which suggests that it is a fairly common expression. Consequently, so as to objectively measure creativity in noun+CT adjectives, fully fledged colour similes must definitely be considered in order to get the broadest picture.

From the extracted results, we distinguish three main groups of noun+CT adjectives. The first group is made up of lexicalised noun+CT compounds that have entered the dictionary; they are generally the most frequent ones and their vehicle is almost never combined with another CT. Apt examples would be compounds such as “*jet-black*” (119 occurrences), “*bottle-green*” (43 occurrences) and “*nut-brown*” (48 occurrences).

The second group comprises semi-lexicalised noun+CT compounds that convey images that are shared by different authors without exhibiting the same fixedness as adjectives of the first group. Examples of adjectives of this group include “*amber-brown*” (3 occurrences/3 authors), “*coffee-brown*” (3 occurrences /3 authors) and “*apple-red*” (4 occurrences/3 authors).

The last group contains creative or original noun+CT compounds. These are compound adjectives that appear generally once in the corpus or are used several times by the same author and do not correspond to a fully fledged colour simile. Some examples are “*death-blue eyes*”, [D.H. Lawrence, *Aaron's Rod* (1922)], “*phantom-grey yacht*” [W. Black, *Donald Ross of Heimra*, 1891], “*lamp-black lashes*” [J. Galsworthy, *Fraternity* (1909)]. This group can be further divided into three subgroups: reinvented noun+CT similes, literal noun+CT similes and metaphorical noun+CT similes. As reinvented noun+CT similes have already been explained in the previous section, we will focus here only on the two remaining noun+CT similes.

Literal noun+CT similes are essentially descriptive as the colour term expresses a salient trait of the vehicle. The vehicle must normally be fairly well known to the readers and is often a vegetal entity or an animal. Examples include “pansy-blue eyes” [G. Griffith, *The World Masters* (1902)], “reseda-green chiffon” [A. E. W. Mason, *At the Villa Rose* (1910)], “lizard-green emerald” [E. P. Oppenheim, *The Ostrekoff Jewels* (1932)], “plumbago-grey suit” [D. H. Lawrence, *The Lost Girl* (1920)].

In contrast, metaphorical noun+CT similes occur mainly when the vehicle or the tenor is an abstract entity and is, therefore, by definition colourless such as in:

(26) It was a weird scene, full of definite detail, fascinating detail, yet all in the funeral-grey monotony of the bush. (D. H. Lawrence, *Kangaroo*, 1920)

(27) That brilliant bird the Baron, whose velvet coat and knickerbockers were the astonishment of Boscastle, instinctively drew near to Christabel, whose velvet and sable, plumed hat, and point-lace necktie pointed her out as his proper mate—Little Monty, Bohemian and *décousu*, attached himself as naturally to one of the Vandeleur birds, shunning the iron-grey respectability of the St. Aubyn breed. (M. E. Braddon, *Mount Royal*, 1882)

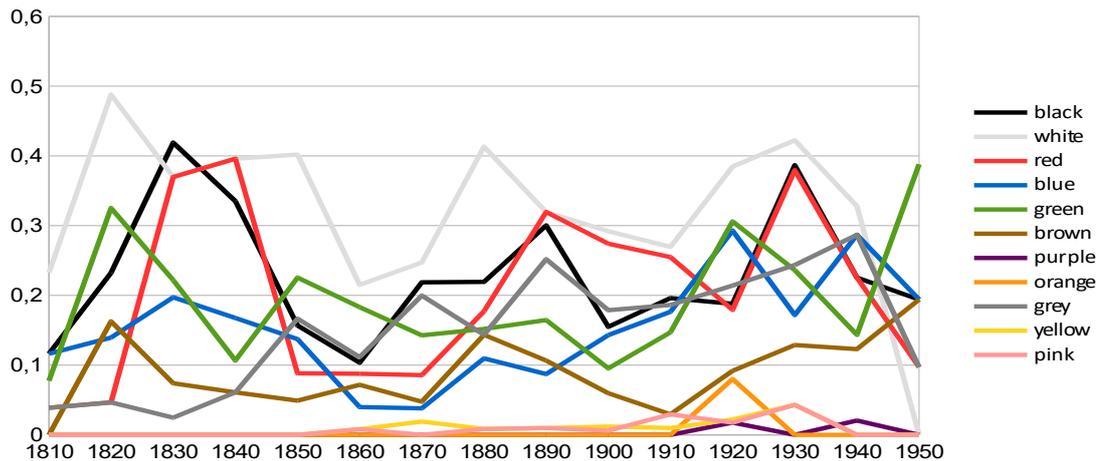
Another interesting point about creative similes is the discrepancy between the 17 contained in the 647 novels by 19th-century writers and the 130 found in the 1682 texts written during the 20<sup>th</sup> century, a fact that could be correlated to the different periods to which these novels belong. In fact, while the first group of novels corresponds roughly to the Victorian period (1837-1900) which saw the emergence of a true novelistic tradition, the second group falls under the Modernist period (1901-1953), a period in which writers feel committed to depicting the world as it is, as accurately as possible:

And art itself may be defined as a single-minded attempt to render the highest kind of justice to the visible universe, by bringing to light the truth, manifold and one, underlying its every aspect. It is an attempt to find in its forms, in its colours, in its light, in its shadows, in the aspects of matter and in the facts of life what of each is fundamental, what is enduring and essential—their one illuminating and convincing quality—the very truth of their existence. (Conrad, 1914, p. vii)

In order to verify whether the frequency of noun+CT similes as a whole differs from the Victorian to the Modernist period, we compared the relative frequency of colour terms per decades, the relative frequency of noun+CT similes, the relative frequency of creative

noun+CT similes and the lexical diversity of vehicles used. All these relative frequencies were computed each time by dividing the number of occurrences by the number of tokens in the novels pertaining to each literary period. The lexical diversity of vehicles is based on the type/token ratio, a formula often used to measure lexical diversity, and has been measured by dividing, for each period and for each basic colour term, the number of unique vehicles by the total number of noun + CT similes containing that colour term.

**Figure 7.3 Relative frequency of each colour term per decade**



In Figure 7.3, it is possible to see that the 1920s especially appear as the noun+CT similes golden era of as it is in the only decade in which all basic colour terms can be found in the corpus. Furthermore, whereas there is only a minor difference as far as the relative frequency of noun+CT similes is concerned, it is possible to witness an increase in the number of creative similes between the two periods:

- Relative frequency Noun+CT similes – Victorian period:  $12 \times 10^{-6}$
- Relative frequency Noun+CT similes – Modernist period:  $16 \times 10^{-6}$
- Relative frequency creative Noun+CT similes – Victorian period:  $0,2 \times 10^{-6}$
- Relative frequency creative Noun+CT similes – Modernist period:  $1,4 \times 10^{-6}$

An author such as D. H. Lawrence perfectly exemplifies the importance of colours in Modernist novels as he is not only the writer with the largest palette of colours but also the one who uses the most noun+CT colours in his writings. As a matter of fact, of the 10 authors (Lawrence, Galsworthy, Orwell, Buchan, Brontë, Walpole, Mundy, Hilton, Griffith, Corelli) with the highest relative frequency of noun+CT similes, only one novelist, Charlotte Brontë, was not published in the 20<sup>th</sup> century.

According to the obtained results, the Modernist novelists are also more innovative than their predecessors (cf. Table 7.8); this is particularly noticeable with colours that are cognitively associated with rather salient vehicles such as “white”, “black”, “red” and “grey”. For all these colours, even though the predominant vehicle is still the most frequently used in Modernist novels, there is also a wide range of new vehicles introduced (e. g. “hearse” with “black”, “egg” with “white”, “fire” with “red” and “skeleton” with “grey”). In contrast, it is possible to infer from its very high lexical diversity that the colour “purple” is the least connoted.

**Table 7.8 Lexical diversity per colour for both literary periods**

	Victorian	Modernist
<i>white</i>	0.06	0.14
<i>black</i>	0.06	0.12
<i>red</i>	0.08	0.12
<i>green</i>	0.11	0.21
<i>yellow</i>	0.53	0.6
<i>blue</i>	0.12	0.25
<i>brown</i>	0.2	0.5
<i>purple</i>	0	1
<i>pink</i>	0.33	0.23
<i>orange</i>	0	0.5
<i>grey</i>	0.09	0.15
Average	0.17	0.34

If all these results tend to confirm the initial hypothesis about the important place of noun+CT similes in Modernist novels and confirm Lawrence’s obsession with colours, they also suggest that the use of colours, in general, is far from being static, especially in diachronic experiments.

From the results obtained, there is no doubt that although both structures are similes, they function differently: while traditional similes are strongly governed by collocations and can be used figuratively more easily, noun+CT similes typically provide background information. This dichotomy could perhaps explain why they also differ in their use of colours, confirming the idea that colours should not be taken in abstraction, but must be studied in a specific context. Furthermore, from the extracted noun+CT similes, a classification has been drawn that takes into account their originality. It has also been shown that despite being relegated to background elements, noun+CT similes actively participate in shaping descriptions in Modernist novels. Besides, writers often take

advantage of the fact that noun+CTs are compound adjectives in order to propose new daring semantic associations and to project the connotative meaning of the colour term on the word modified by the compound adjective. Whether the images thus created are really accessible to the readers should be interesting to test as they require in some cases either a solid culture or a vivid imagination.

## 7.4 On Proper Nouns in Comparative Constructions

This study differs from the two previous ones in the sense that it focuses on how the developed method could be modified to tackle other kinds of simile constructions, in this case, similes which use a proper noun as a vehicle. By proper nouns, of course, it is meant here, nouns of people or places external to the narrative. Proper nouns in similes constitute an interesting research question as this type of similes has not been closely examined in the literature of similes and seems to be underrepresented in idiomatic similes. As far as French is concerned, out of the 1022 idiomatic similes cited by Cazelles (1996) and Parmentier (2002), only 27 make use of proper nouns. Worse, Parmentier (2003) lists only 7 similes that makes use of proper nouns in his dictionary that contains 454 similes in English. With regard to the insignificant proportion of proper nouns in idiomatic similes, the question that therefore arises is whether the use of proper nouns in comparative constructions is trivial, especially if taken into consideration the role proper nouns play in alluding and in interconnecting texts. In terms of literary tradition, as the Bible has often been one of the main source of inspiration for writers, it would be worthwhile to determine if the names of locations and places used in these comparisons are drawn from particularly identifiable sources. Furthermore, does the presence of proper nouns affect in some way the understanding of this type of comparisons? Finally, are there particular stylistic strategies involved?

In order to eliminate as much noise as possible and to avoid extracting comparisons that involve characters or places from the novel, only were extracted comparisons which had as standard NP a proper noun that appear twice in the novel. No system specialised in named entity recognition was used and we relied entirely on the output produced by the part-of-speech tagger, in this case TreeTagger (Schmid, 1994). From the obtained results, it is possible to distinguish four main types of peoples and entities used as standard NPs:

- Personified entities: It was Polly Sims, who was incontinently made as blind as **Fortune** or **Justice**, or any other of the deities who dispense benefits to man. (M. E. Braddon, *Vixen*, 1879)

- nationalities/ethnicities: And he sat for the most part impassive and abstract as a **Red Indian**. (D. H. Lawrence, *The Lost Girl*, 1920)
- divinities and religious figures: Out of the provinces came Waldemar, like **Mahomet** from the desert, to preach a new gospel. (J. Buchan, *The Gap in the Curtain*, 1932)
- historical figures: Abbot, I think, gave me credit for being a sort of infantine **Guy Fawkes**. (C. Brontë, *Jane Eyre*, 1847).
- artistic figures and productions which include all sorts of artists (painters, sculptors, comedians, opera writers, writers...) but also their work (paintings, characters, novels...): You can die like **Keats** or survive to be a pompous old ass like **Tennyson**. (H. Walpole, *Hans Frost*, 1929)

As with frozen similes, the frequency of the most frequent names of people and of geographical places is rather low. Unsurprisingly, most of the top proper names are used in idiomatic similes. However, a quick glance at the top names of people confirms our first hunch on novelists alluding prominently to biblical stories. The role played by Greek mythology in providing external references, however, cannot be undermined.

**Table 7.9 Top names of people and of geographical places**

French		English	
People	Places	People	Places
Job [45]	Pont-Neuf [16]	Lucifer [34]	Paradise [11]
Turc [26]	Orient [15]	Job [33]	Sahara [9]
Samson [24]	Rhône [6]	Croesus [29]	Jerusalem [7]
Jean [22]	Louvre [5]	Apollo [28]	Thames [6]
Crésus [21]	Rhin [5]	Solomon [27]	Styx [6]
Madeleine [21]		Samson [25]	
Jésus [17]		God [25]	
Achille [15]		Madonna [23]	
Hercule [15]		Indian [23]	
Ajax [14]		Jew [22]	

More important than who is alluded to, it is what is said about that person in order to see whether the same scene is repeated over and over again, which could suggest either that it belongs to the common knowledge, was popular at that time or that it is strongly associated with that particular place or character. By studying the structure of the standard NP, it was possible to unveil networks of intertextuality but also different stylistic strategies used by the authors to allude to a historical figure or a fictional character.

Consider the following sentences from the corpus:

- a) La vieille hôtesse était là comme **Marius sur les ruines de Carthage**. (H. de Balzac, *Le Père Goriot*, 1835)
- b) Il est là comme **Marius sur les ruines de Carthage**, les bras croisés, la tête rasée, Napoléon à Sainte-Hélène, quoi ! (H. de Balzac, *La Cousine Bette*, 1846)
- c) Gaston et moi, nous nous sommes assis sur ces débris comme **Marius sur les ruines de Carthage**. (Z. Fleuriot, *En Congé*, 1874)
- d) Le général était arrivé à Paris le front penché, l'âme en deuil, le désespoir au cœur, résolu à vivre seul, comme **Marius debout sur les ruines de Carthage**. [P. P. du Terrail, *Les Exploits de Rocambole*, 1859]
- e) Ukridge sat like **Marius among the ruins of Carthage**, and refused to speak. (P.G. Wodehouse, *Love among the Chickens*, 1906)
- f) Having reached the bottom, he sat amid the occasional china, like **Marius among the ruins of Carthage**, and endeavored to ascertain the extent of his injuries. (P. G. Wodehouse, *Something New*, 1915)
- g) Sammy had by this time disposed of the clock-work rat, and was now standing, like **Marius, among the ruins barking triumphantly**. (P. G. Wodehouse, *Mike*, 1909)
- h) Robinson, as he descended into the darkened shop, and walked about amidst the lumber that was being dragged forth from the shelves and drawers, felt that he was like **Marius on the ruins of Carthage**. (A. Trollope, *The Struggles of Brown Jones and Robinson*, 1862)
- i) "You look like **Marius sitting amidst the ruins of Carthage**, my dear!" (E. Gaskell, *Wives and Daughters*, 1864-1866)
- j) He found Jannath glowering like **Marius in a dungeon**. (T. Mundy, *Jungle Jest*, 1932)
- k) He felt like **the boy Marius on his way to his bed in the mountain monastery**, with the life of the cities far behind and the purity and sweetness of the country already like a sweet tonic in his blood... (E. P. Oppenheim, *Murder at Monte-Carlo*, 1933)

The first nine examples, both in English and in French, all use almost word for word the same standard NP. In addition, it is very surprising to notice that Balzac and Wodehouse, unknowingly or not, plagiarise their own sentences. The formulation being too precise for it to be a simple coincidence, we looked for a possible original source and found this passage concerning Marius, consul of Rome:

Then, when asked by him what he had to say, and what answer he would make to the governor, he answered with a deep groan: "Tell him, then, that thou hast seen Caius Marius a fugitive, seated amid the ruins of Carthage." Plutarch, *The Parallel Lives* (trans. 1923)

One of the most plausible explanation for the reuse of the same image would be to hypothesise that the different novelists came across the original texts in their studies or readings. For modern readers, however, unless they are versed in the history of Ancient Rome, this reference would most probably remain obscure as they need to know specifically to which episode it refers and what was the state of mind of Marius, information that are necessary to comprehend the simile here used. As put by Perri (1978), allusions cannot be separated from a number of pragmatic considerations:

[A]llusion is a way of referring that takes into account and circumvents the problem of what we mean when we refer: allusion-markers act like proper names in that they denote unique individuals (source texts), but they also tacitly specify the property(ies) belonging to the source text's connotation relevant to the allusion's meaning. (p. 290)

Unlike the first nine examples, the remaining two examples are less clear, but still by knowing the life of Marius and the fact that he was imprisoned at some point, it is possible to extrapolate by assimilating his prison to a dungeon and therefore, to be able to conclude in that specific sentence that it is the same Marius.

If the text or any comparison/simile in general is seen as a dialogue between the reader and the author, the question of the audience and of the reception of these comparisons/similes is essential. In this respect, we investigated the stylistic choices in comparisons involving literary characters to see whether the novelists often ease the task of their readers.

We distinguished two main ways of introducing an allusion in a comparison/simile:

- the plain reference which is a priori the most difficult for the readers as nothing could help them to situate the person or the place mentioned, especially if the allusion is completely unknown.

### Example

He is such a perfect stick; but then certainly there is no other single man in the parish under forty. He is like **Robinson Crusoe**. It is an awfully deceptive position for a young man to occupy. (E. Braddon, *The Golden Calf*, 1883).

The surrounding sentences, in this example, help very little to deduce on which aspects the person mentioned is similar to Robin Crusoe, which is paradoxically a very well-known literary figure.

- the contextual reference in which complementary information is given about the alluded term so as to better clarify the comparison. This type of reference can be further divided into three subtypes:

a) the reference accompanied by the author's or the book's name;

**Example**

Cob-Lafleur avait su se montrer doux, souriant, timide. Quand il veut, il peut ressembler au Gringoire de Banville. Bref, Passavant se montrait séduit et était sur le point de l'engager. (A. Gide, *Les Faux-monnayeurs*, 1925).

Unlike the previous example, even though the reference remains obscure for those unfamiliar with Banville's plays, it is possible from the context to imagine the kind of person Gringoire: kind, smiling and shy. This strategy is particularly useful to differentiate between homonyms as in:

J'aime beaucoup sa fille, la pastoresse. Madame Vedel ressemble à l'Elvire de Lamartine; une Elvire vieillie. Sa conversation n'est pas sans charme. (A. Gide, *Les Faux-monnayeurs*, 1925).<sup>42</sup>

b) general characteristic + reference: this is the case for closed similes built with a ground which is typically associated with a specific standard NP. In addition, in this type of simile, the standard NP tends to come from the Greek mythology. For example, beauty, bravery and strength are generally linked to Greek gods and characters from *The Odyssey*. In this respect, they are usually easy to interpret.

**Example**

She had risen from the ground more lovely than Helen of Troy and now he was blinder than Homer. (H. Walpole, *Katherine Christian*, 1944)

c) specific characteristic or behaviour + reference + specific episode: the ground is not permanent but is associated with the standard NP at a particular moment and knowing that episode is generally required to be able to understand all the nuances of the simile.

---

<sup>42</sup> By mentioning Lamartine, no confusion could be made among others with Molière's Elvire in *Le Festin de pierre* (1682) and Corneille's Elvire in *Le Cid* (1648).

### Examples

Like Adam when God breathed into his nostrils the breath of life, she had become a living soul, and that of which she was the living soul was his work. (E. Von Arnim, *The Pastor's Wife*, 1914). → She did not exist before just like Adam.

Votre douce voix qui m'appelle me rend malade de fureur si elle est un piège ... mais plus mou qu'Hercule aux pieds d'Omphale si elle vibre d'une véritable tendresse, comme, parfois, j'ai osé l'espérer et comme je veux le croire ce soir! (G. Leroux, *La Poupée sanglante*, 1923).

In the last example, the usual image of Hercules the invincible strong warrior is turned upside down and Hercules is rather associated with powerlessness, a condition he briefly endures when he has to serve Omphale, doing all sorts of menial works.

Direct or indirect quotations may also be used to establish similarities between the situation depicted in the novel and words of a literary character.

### Examples

\*\* Intertextuality with Shakespeare's *Richard III*, Act 5, scene 4 (1592): "A horse! a horse! my kingdom for a horse!" (1916, p. 183).

She is at this moment shouting for her governess, as **King Richard** (I am a great reader of Shakespeare) once shouted for his horse (W. Collins, *The Evil Genius*, 1886).

– Que pouvez -vous attendre d'un homme qui à tout moment s'écrie comme Richard III: Mon royaume pour un cheval! dit Emmanuel. (H. de Balzac, *La Recherche de l'absolu*, 1834).

Si son cheval eût manqué, il eût crié comme Richard III: Ma couronne pour un cheval ! (A. Dumas, *La Reine Margot*, 1845).

... Et, de même que Richard III, dans un moment suprême, avait crié: « **Ma couronne pour un cheval!** (A. Dumas, *La Comtesse de Charny*, 1853).

– Volontiers, comme Richard III, il aurait crié: « **Ma fortune pour un fiacre !** » (A. Gaboriau, *L'affaire Lerouge*, 1863)

When the quotation is exactly given as in the original as in the Balzac's example, it is not imperative to know the source but it could help when that quotation has been transformed for humoristic effects as in the fourth sentence. Moreover, as illustrated by the first example,

all these strategies for alluding are not mutually exclusive. In the following sentence, the reference is restricted with both the author's name and the description of a specific episode: Et, semblable au Silène de Virgile qui, barbouillé du suc des mûres, chantait à des bergers de Sicile et à la naïade Églé l'origine du monde, il se répandit en paroles abondantes : — Appeler un malheureux à répondre de ses actes ! (A. France, *Histoire comique*, 1903).

Another interesting impact that the use of proper nouns in comparative constructions has on the writing is that it enables to create network of comparisons and either to create subsequent images or to compare more broadly the universe of the novel and that universe of the text from which the standard NP has been taken from.

### Examples

1/ Abraham had two or three wives and several concubines, and he was the very soul of virtue according to sacred lore, —whereas my Lord Tom-Noddy in London to-day has one wife and several concubines, and is really very much like Abraham in other particulars, yet he is considered a very dreadful person. (M. Corelli, *The Sorrows of Satan*, 1895)

2/ Jackal, aussi pensif, aussi morne qu'Hippolyte, la tête aussi basse que les coursiers du héros classique, absorbé dans une pensée non moins triste que celle qui occupait l'esprit de ces nobles animaux, se dirigea vers la rue du Puits-qui-Parle. (A. Dumas, *Les Mohicans de Paris*, 1854-1859)

In the first sentence, a contrast is created from the beginning between Lord Tom-Noddy and Abraham, contrast that is further accentuated with the comparison. On the contrary, in the second example, the comparison is introduced first and two other comparisons are built by exploiting known attributes of the first standard NP.

In some respect, proper nouns in comparative constructions participate to the creation of intertextuality between texts and as such, perpetuate of an existing literary tradition. Therefore, through the use of proper nouns, authors are able to add new layers of meaning to comparative constructions. In addition, they have developed some strategies to make those comparative constructions at times less obscure for the common reader or to make them blend with the rest of the narrative.

The three corpus-based applications presented show that automatic simile annotation could be used to explore diverse general literary questions as well as to focus on a particular aspect of similes. In this respect, it could constitute a valuable tool for literary scholar

interested in these types of questions. Moreover, the proposed method can easily be adapted to other kinds of simile structures.

## 8 CONCLUSION AND FUTURE WORK

Similes are so common and evocative that we use them every day without even thinking about it. If some rhetoricians fail to see any appeal in similes because there is no change in word meaning, similes are particularly interesting as they grammatically fall under comparative constructions and can be transformed through deletion into metaphors. In this respect, similes seem to be particularly adequate for studying the source of figuration in languages.

In this work, we attempted to take advantage of the syntactic similarities between French and English on the one hand, and between their comparative constructions, on the other hand, to propose a method to detect and mine similes in literary texts written in any of these two languages. As different traditions of similes and metaphors exist, we first refined our definition of the simile and explored various theories explaining how similes differ from literal comparisons. Furthermore, we also attempted to briefly enumerate some of the main challenges that are inherent to simile constructions: the polysemy of the markers and ellipsis. If the former is often mentioned as far as the detection methods described in this thesis are concerned, the latter is generally ignored.

With regard to the annotation of similes found in literary texts, we reviewed annotation guidelines and practices in the humanities and noticed that, apart from some individual efforts, they generally do not agree with the two axes with which literary scholars describe and evaluate similes: the syntactic axis and the semantic axis (degree of animacy, degree of abstraction).

As basis for our detection method, we elaborated a grammar of the simile which lists the different syntactic forms it can take and establishes for each case, a correlation between the grammatical function or positions of constituents of the sentence and their role in the simile / comparison. This step, in our opinion, is crucial as both simile recognition and simile annotation rely on these components.

With simile annotation from a stylistic perspective as our ultimate goal, we designed a method which first focuses on syntax to identify simile candidates and their components, then, on semantics to decide whether retrieved structure are similes or not, and finally tags those components and highlights particular features of the simile structure such as the position of the marker or the semantic categories involved in the simile. Besides, as far as the simile recognition task is concerned, we tried to incorporate salience and levels of categorisation by extracting noun-adjective and noun-verb pairs in machine-readable dictionaries. We also propose a two-level set of semantic categories which tackles word polysemy and enable in case of doubt to stop at the broadest level, which is particularly useful for abstract entities.

In addition to proposing an annotation scheme, we also let other people annotate a corpus of prose poems in order to:

- help evaluate our own automatic simile detection method;
- produce an annotated corpus for future research;
- gather data on simile perception, specifically on figuration and creativity in similes.

The results obtained so far suggest that indeed, people tend to less disagree on broad semantic categories and that simile annotation is not so simple for human beings, especially in front of poetic texts. In addition, most annotators knowingly or not, identify for each simile component, the whole phrase, which legitimate our decision of tagging the whole phrases in our annotation as they are important both semantically and stylistically.

Finally, we applied to another corpus, a corpus of British and French novels published between 1810 and 1950 with the following questions: which idiomatic similes can be considered as literary clichés and what do they refer to? What is the difference in usage between fully fledged colour similes and noun + CT similes? How are proper nouns used in comparative constructions and what can they tell us about literary tradition?

In a nutshell, we were able to find out that stereotypical literary similes are made up of formulaic similes such as “bark + bad + marker + bite” and mainly of similes describing the state of a body part or of a corporeal attribute. As far as colours are concerned, fully fledged colour similes are mainly used for figurative contexts, whereas noun + CT similes are generally used for background descriptions. However, so as to create striking contrasts, novelists do not often hesitate to innovate by creating improbable colour associations. With proper nouns, the effect is a little bit different as it is mostly the readers’ cultural knowledge which is challenged. In addition, our study has shown interesting networks of intertextuality conveyed through the use of proper nouns in similes and the various techniques used by the authors to allude to their predecessors or to classical texts.

From a computational point of view, this last experiment aptly proves that the proposed method is flexible enough to be extended to other types of similes. First, come to mind, of course, clausal similes that undoubtedly need their own annotation scheme. Though they seem to be less used, it would also be interesting to explore phrasal similes built around prepositional phrases. Similarly, it could be challenging but certainly worthwhile to try to adapt the method to languages sharing the same comparative constructions as English and French and to see to which extent, the described method could tackle the detection and the analysis of similes in verse poems.

Furthermore, in order for the results to better reflect choices, different levels other than the sentence-level should be considered as well as other figures of speech, as it has been repeatedly said that similes can easily blend with other figures of speech. In this respect, figures of repetition constitute a good start as they are easily found through pattern matching. In the same vein, instead of extracting all possible similes in the text, it could be interesting to restrict the search to a particular cluster of pertinent similes connected to the main themes of the text, to a specific character or situation.



## 9 REFERENCES

- Abrams, M.H. (1999). *A Glossary of Literary Terms*. 7<sup>th</sup> ed. Boston: Heinle & Heinle.
- Addison, C. From literal to figurative: An introduction to the study of simile. *College English*, 55, 4, 402-419.
- Almuhareb, A., & Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. *Proceedings on Empirical Methods in Natural Language Processing*, 158-165.
- Alsop, S., & Nesi, H. (2014). The pragmatic annotation of a corpus of academic lectures. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1560-1563.
- Amossy, R., & Herschberg-Perrot, A. (1997). Stéréotypes et clichés. *Langue – Discours – Société*. Paris: Nathan.
- Amsler, R. A. (1980). The structure of the *Merriam-Webster Pocket Dictionary*. (Unpublished doctoral dissertation). The University of Texas, Austin.
- Aristotle. (1898). *The Poetics*. Trans. S. H. Butcher. London: MacMillan and Co.
- Aristotle. (1926). *The Art of Rhetoric*. (Trans. J. H. Freese). London & New York: William Heinemann & G. P. Putnam's Sons.
- Aristotle. (1984). *Topics*. Trans. W. A. Pickard-Cambridge. In J. Barnes (Ed.), *The Complete Works of Aristotle* (). Princeton: Princeton University Press.
- Bain, A. (1879). *A Higher English Grammar*. London: Longmans and Co.
- Bain, A. (1890). *English Composition and Rhetoric*. New York: D. Appleton and Company.
- Ballard, B. W. (1998). A general computational treatment of comparatives for natural language question answering. *Proceedings of the 26<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 41-48.
- Bally, C. (1909). *Traité de Stylistique française*. 2<sup>nd</sup> edition. Paris: Librairie C. Klincksieck.

- Barnard, D. T., & Ide, N. M. (1997). The Text Encoding Initiative: Flexible and extensible document encoding. *Journal of the American Society for Information Science*, 48 (7), 622-628.
- Baudelaire, C. (1857). *Les Fleurs du Mal*. Paris: Poulet-Malassie et De Broise.
- Baudelaire, C. (1885). *L'Art romantique*. Paris: Calmann Lévy.
- Beardsley, M. C. (1950). *Thinking Straight: A Guide for Readers and Writers*. New York: Prentice-Hall.
- Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley and Los Angeles: University of California Press.
- Berteau, R. (1979). Similitudo. *L'antiquité classique*, 48 (1), 154-160.
- Berteau, R. (1980). L'opposition "comparatio" vs "similitudo" dans la rhétorique latine. *Latomus. T.*, 2(2), 393-398.
- Bertrand, L. (1842). *Gaspard de la nuit : Fantaisies à la manière de Rembrandt et de Callot*. Angers : V. Pavie.
- Blair, H. (1787). *Lectures on Rhetoric and Belles Lettres*. Vol. I. London: A. Strahan, T. Cadell & W. Creech.
- Bouchard, D.-E. (2008). Comparaison discontinue et quantification de degrés. Proceedings of the 2008 annual conference of the Canadian Linguistic Association. Retrieved from [http://homes.chass.utoronto.ca/~cla-acl/actes2008/CLA2008\\_Bouchard.pdf](http://homes.chass.utoronto.ca/~cla-acl/actes2008/CLA2008_Bouchard.pdf)
- Bouverot, D. (1969). Comparaison et métaphore. *Le Français Moderne*, 37 (2, 3 & 4), 132-147, 222-238, 301-307.
- Brachman, R. J., & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. San Francisco: Elsevier.
- Bray, T., Paoli, J, Sperberg-McQueen, J. M., Maler, E., Yergeau, F. (2008). Extensible Markup Language (XML) 1.0. (Fifth edition). <https://www.w3.org/TR/xml/>
- Bredin, H. (1998). Comparisons and similes. *Lingua*, 105, 67-78.
- Breton, A. (1924). *Manifeste du Surréalisme*. Retrieved from [http://wikilivres.ca/wiki/Manifeste\\_du\\_surr%C3%A9alisme#cite\\_ref-6](http://wikilivres.ca/wiki/Manifeste_du_surr%C3%A9alisme#cite_ref-6)
- Brooke-Roose, C. (2002). *Invisible Authors: Last Essays*. Ohio State University.
- Bullinger, E. W. (1898). *Figures of Speech used in the Bible: Explained and Illustrated*. London & New York: Messrs. Eyre & Spottiswoode, Messrs. E & J. B. Young.
- Candito, M., Nivre, J. & Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for French. *Proceedings of COLING 2010*, 108-116.
- Caplan, H. (1954). Trans. *Ad C. Herennium De Ratione Dicendi (Rhetorica ad Herennium)*. London: William Heinemann Ltd.
- Carter, R., & Simpson, P. (1989). Introduction. In *Language, Discourse and Literature: An Introductory Reader in Discourse Stylistics*. London and New York: Routledge.

- Cazelles, N. (1996). *Les comparaisons du français*. Paris: Belin.
- Chateaubriand, F. R. (1739). Le génie du christianisme. *Œuvres Complètes de Monsieur le Vicomte de Chateaubriand*. Vol III. (pp. 1-360). Paris: Chez Firmin Didot Frères.
- Cheminée, P., Dubois, D. and Resche-Rigon, P. (2006). Couleur de pensée, couleur du temps. Penser la couleur et variations diachroniques du lexique de la couleur. *Les Couleurs en question*, 23-45.
- Cicero. (1856a). *De inventione (Treatise on Rhetorical Invention)*. Trans C. D. Yonge. London: Henry G. Bohn.
- Cicero. (1856b). *On Topics. (Treatise on Topics)*. Trans C. D. Yonge. London: Henry G Bohn.
- Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed), *Discourse Production and Comprehension* (pp. 1-40). Norwood: Ablex Publishing Corporation. 1-40.
- Cohen, J. (1968). La comparaison poétique : Essai de systématique. *Langages*, 3 (12), 43-51.
- Connolly, A. and Hopkins, L. (2015). A Darker Shade of Pale: Webster's Winter Whiteness. *E-rea*, 12(2). Retrieved from <http://erea.revues.org/4483>. doi: 10.4000/erea.4483.
- Conrad, J (1914). Preface. *The Nigger of the Narcissus: A Tale of Forecastle*. New York: Doubleday, Page & Company.
- Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In J. A. Miller & J. W Smith (Eds.), *Proceedings of the 39<sup>th</sup> Annual ACM Southeast Conference*, 95-102.
- Crystal, D. (1970). New perspectives for language study. 1: Stylistics. *ELT Journal*. 24 (2), 99-106. doi: 10.1093/elt/XXIV.2.99.
- Cummings, J. (2008). The Encoding Text Initiative and the study of literature. In S. Schreibman & R. Siemens (Eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell. Retrieved from [http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-6&toc.depth=1&toc.id=ss1-6-6&brand=9781405148641\\_brand](http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-6&toc.depth=1&toc.id=ss1-6-6&brand=9781405148641_brand)
- Curran, J. R., & Moens, M. (2002). Improvements in Automatic Thesaurus Extraction. *Proceedings of the Workshop of Unsupervised Lexical Acquisition*, 59-66.
- Daudet, A. (1909). *Le Petit Chose*. Paris: Bibliothèque-Charpentier.
- Dedrick, D. (1998). The Foundations of the Universalist Tradition in Colour-Naming Research (and their Supposed Refutation). *Philosophy of the Social Sciences*, 28 (2), 179-204.
- Delcourt, C. (2002). Stylometry. *Revue belge de philologie et d'histoire*, 80, 979-1002.
- Delabre, M. (1984). Syntaxe de ainsi que et de même que en français contemporain. *L'Information grammaticale*, 23(1), 11-17.

- Deléchelle, G. (1995). Emploi de *as* en anglais, comparaison et identification. *Faits de langue*, 3(5), 193-200.
- Deléchelle, G. (2004). Causalité et phrase complexe: prédications et circonstances concomitantes. *Cercles*, 9, 121-142. Retrieved from <http://www.cercles.com/n9/delechelle.pdf>
- De Mille, J. (1878). *The Elements of Rhetoric*. New York: Harper & Brothers.
- Desmets, M. (2008). Ellipses dans constructions comparatives en comme. *Linx*, 58, Retrieved from <http://linx.revues.org/328>. doi: 10.4000/linx.328
- Dierks, K. (1986). Automatic stylistic analysis of lyrical texts. *Literary and Linguistic Computing*, 1(3), 129-135.
- Dixon, R. M. W. (2005). Comparative constructions in English. *Studia Anglica Posnaniensia*, 41, 5-27.
- Dong, Z., Dong, Q., & Hao, C. (2010). HowNet and its computation of meaning. *COLING 2010: Demonstration Volume*, 53-56.
- Dowman, M. (2001). *A Bayesian Approach to Colour Term Semantics*. Technical Report: The University of Sydney.
- Dubois, J. & Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, 3(179-180), 31-56.
- Dubois, D., & Grinevald C. (1999). Pratiques de la couleur et dénominations. *Faits de langues*, (14), 11-25.
- Fahnestock, J. (2011). *Rhetorical Style: The Uses of Language in Persuasion*. New York: Oxford University Press.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The Structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge : The MIT Press.
- Ferrari, S. (1997). Méthodes et outils informatiques pour le traitement des métaphores dans les textes écrits. Ph. D thesis, Paris XI.
- Fishelov, D. (1993). Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric. *Poetics Today*, 14(1), 1-23.
- Fishelov, D. (2007). Shall I compare thee? Simile understanding and semantic categories. *Journal of Literary Semantics*, 36 (1), 71-87.
- Fizman, M., Demner-Fushman, D., Lang, F. M., Goetz, P., and Rindfleisch, T. (2007). Interpreting comparative constructions in biomedical text. *Proceedings of BIONLP 2007: Biological, translational, and Clinical Language Processing*, 137-144.
- Flaubert, G. (1885). *Madame Bovary: Mœurs de Province*. Paris: A. Quantin.

- Flaubert, G. (1896). *Madame Bovary*. Trans. William Walton. Vol. I & II. Philadelphia: George Barrie & Sons.
- Flaubert, G. (1886). *Madame Bovary*. Trans. Eleanor Marx-Aveling. Mienola: Dover Publications.
- French, L. M. (2008). *History of Emporia and Lyon County*. Emporia: Emporia Gazette Press.
- Friedman, C. (1989). A general computational treatment of the comparative. *Proceedings of the 27<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 161-168.
- Fuchs, C. (2014). *La comparaison et son expression en français*. Paris: Orphrys.
- Fuchs, C., & Le Goffic, P. (2005). La polysémie de “comme”. In O. Soutet (Ed.), *La Polysémie* (pp. 267-292). Paris: Presses de l'Université Paris-Sorbonne.
- Gargani, A. (2014). Poetic comparisons: How similes are understood. PhD Dissertation, University of Salford.
- Genette, G. (1970). La rhétorique restreinte. *Communications*, 16(1), 158-171. doi: 10.3406/comm.1970.1234
- Gentner, D. (1982). Are scientific analogies metaphors? In D.S. Miall (Ed.), *Metaphor: Problems and Perspectives* (pp. 106-132). Brighton: Harvester Press.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gildea, D & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3); 245-288. doi: 10.1162/089120102760275983.
- Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97(1), 3-18.
- Goatly, A. (2011). *The Language of Metaphors*. 2<sup>nd</sup> ed. London and New York: Routledge Taylor Francis Group.
- Green, W. C. (1877). *The Similes of Homer's Iliad*. London: Longmans & Co.
- Greenberg, (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of Language* (pp. 73-113). London: MIT Press.
- Grevisse, M. (2001). *Le Bon Usage: Grammaire française*. 13<sup>th</sup> ed. Bruxelles: Duculot.
- Harris, R., & DiMarco, C. (2009). Constructing a rhetorical figuration ontology. Symposium on Persuasive Technology and Digital Behaviour Intervention, Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB). Retrieved from <https://cs.uwaterloo.ca/~cdimarco/pdf/publications/AISB2009.pdf>
- Hockey, S. (1994). Evaluating electronic texts in the humanities. *Library Trends*, 42 (4), 676-693.

- Hornby, A. S. (2000). *Oxford Advanced Learner's Dictionary of Current English*. 6<sup>th</sup> ed. Oxford: Oxford University Press.
- Ide, N. (2004). Preparation and analysis of linguistic corpora. In S. Schreibman & R. Siemens (Ed.), *A Companion to Digital Humanities*. John Unsworth. Oxford: Blackwell.
- Ide, N. M., & Sperberg-McQueen, S. (1995). The TEI: History, goals and future. *Computers and the Humanities*, 29, 5-15.
- Ide, N., & Romary, L. (2003). Outline of the International Standard Annotation Framework. *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, 1-5.
- Israel, M, Riddle Harding, J., & Tobin, V. (2004). On simile. In M. Archer & Suzanne Kemmer (Eds.), *Language, Culture and Mind* (pp. 124-135). CSL Publications.
- Jakobson, R. (1960). Linguistics and poetics. In T. Sebeok (Ed.), *Style in Language* (pp. 350-377). Cambridge, MA: M.I.T. Press.
- Jamieson, A. (1826). *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 Illustrated by Appropriate Examples, Selected Chiefly from the British Classics, for the Use of Schools, or Private Instruction*. New-Haven: A. H. Maltby and Co.
- Jenny, L. (1993). L'objet singulier de la stylistique. *Littérature*. 89, 113-124.
- Jindal, N., & Liu, B. (2006a). Identifying comparative sentences in text documents. *Proceedings of the 29<sup>th</sup> Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, 244-251.
- Jindal, N., & Liu, B. (2006b). Mining comparative sentences and relations. *Proceedings of the 21<sup>st</sup> National Conference on Artificial Intelligence*, Vol. 2, 1331-1336.
- Kahrel, P. Barnett, R., & Leech, G. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. In R. Garside, G. Leech, T. McEnery (Eds), *Corpus Annotation* (pp. 231-242). London and New York: Routledge.
- Kellogg, B. (1901). *A Text-book on Rhetoric supplementing the development of the science with exhaustive practice in composition*. New York: Maynard, Merrill & Co.
- Kessler, J. S., Eckert, M., Clark, L. & Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. *Proceedings of the 4<sup>th</sup> International AAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*. Retrieved from [http://www.icwsm.org/2010/papers/icwsm10dcw\\_8.pdf](http://www.icwsm.org/2010/papers/icwsm10dcw_8.pdf)
- Kessler, W. & Kuhn, J. (2013). Detection of product comparison – How far does and out-of-the-box semantic role labeling system take you? *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1892-1897.

- Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, 105-115.
- Kipper, K., Dang, H. T. & Palmer, M. (2000). Class-based construction of a verb lexicon. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Application of Artificial Intelligence*, 691-696.
- Kirvalidze, N. (2014). Three-dimensional world of similes in English fictional writing. *Sino-US English Teaching*, 11(1), 25-39.
- Kübler, S., & Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London and New York: Bloomsbury.
- Lakoff, G., & Johnson, M. (1980). *Metaphor We Live by*. Chicago and London: The University of Chicago Press.
- Lawrence, D. H. (1921). *The Lost Girl*. New York: Thomas Seltzer.
- Lechner, W. (2001). Reduced and phrasal comparatives. *Natural Language and Linguistic Theory*, 19, 683-735.
- Leech, G. N. (1969). *A linguistic Guide to English Poetry*. London and New York: Longman.
- Leech, G. (1981). *Semantics: The Study of Meaning*. Bungay: Penguin Books.
- Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistics Computing*, 8(4), 275-281.
- Leech, G. (2005). Adding linguistic annotation". In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 17-29). Oxford: Oxbow Books. Retrieved from <http://ahds.ac.uk/linguistic-corpora/>
- Leech, G. McEnery, T., & Wynne, M. (1997). Further levels of annotation. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus Annotation* (85-101). London and New York: Routledge.
- Leech, G., & Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional prose*. Harlow : Pearson Longman.
- Le Guern, M. (1973). *Sémantique de la métaphore et de la métonymie*. Paris: Larousse.
- Lehman, D. (2003). *Great American Prose Poems: From Poe to the Present*. New York: Scribner Poetry.
- Levin, S. R (1982). Are figures of thought figures of speech? In H. Byrnes (Ed.), *Georgetown University Round Table on Languages and Linguistics 1982* (pp. 112-123). Washington: Georgetown University Press.
- Li, B., Kuang, H., Zhang, Y., Chen, J., and Tang, X. (2012). Using similes to extract basic sentiment across languages. *Web Information Systems and Mining, Volume 7529 of the series Lecture Notes in Computer Science*, 536-542.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational Linguistics*, Vol. 2, p. 768-774.

- Lord, D. N. (1855). *The Characteristics and Laws of Figurative Language*. New York: Franklin Knight.
- Mallarmé, S. (1897). *Divagations*. Paris: Bibliothèque-Charpentier.
- Mansell Jones, P. (1969). Baudelaire as a Critic of Contemporary Poetry. In T. E. Lawrenson, F.E. Sutcliffe and G. F. A. Gadoffre (Eds.), *Modern Miscellany Presented to Eugène Vinaver by Pupils* (pp. 137-153). Manchester: Manchester University Press.
- Màrquez, L., Carreras, X., Likowski, K. & Stevenson, S. (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2), 145-159.
- Martin, J.H. (1996). Computational approaches to figurative language. *Journal of Metaphor and Symbolic Activity*, 11(1), 85-100.
- Martins, A., Smith, N., Xing, P., Aguiar P., & Figueiredo, M. (2010). Turbo parsers: Dependency parsing by approximate variational inference. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 34-44.
- Mason, C.P. (1874). *English Grammar including the Principles of Grammatical Analysis*. London: Bell & Sons.
- Masui, F., Tsunashima T., Sugio T., Tazoe, T., & Shiino, T. (1996). Analysis of lengthy sentences using an English comparative structure model. *Systems and Computers in Japan*, 27(8), 39-52.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.
- McIntyre, W. J. M. (2009). A Retrospective Survey of the Problems with Berlin and Kay (1969). *California Linguistic Notes*, 24(1). Retrieved from <http://english.fullerton.edu/publications/clnArchives/pdf/berlin%20%20kay-R.pdf>
- McManus, I. C. (1983). Basic Colour Terms in Literature. *Language and Speech*, 26, 243-252.
- Miller, G. A. (1990). Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4), 245-264.
- Meyers, A., Kosaka, M., Sekine, S. Grishman, R., & Zhao, S. (2001). Covering treebanks with Glarf. *Proceedings of the ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*, 51-58.
- Moline, E., & Flaux, N. (2008). Constructions en comme : homonymie ou polysémie ? Un état de la question. *Langue française*, 159, 3-9. doi : 10.3917/lf.159.0003
- Moon, R. (2011). Simile and dissimilarity. *Journal of Literary Semantics* (40), 133-157. doi: 10.1515/jlse.2011.008
- Morinet, C. (1995). La comparaison en aval ou en amont de la métaphore. *Faits de langue*, 5, 201-208. doi: 10.3406/flang.1995.995
- Murphy, M. S. (1992). *A Tradition of Subversion: The Prose Poem in English from Wilde to Ashbery*. Amherst: The University of Massachusetts Press.

- Mylonas, E., & Renear, A. (1999). The Text Encoding Initiative at 10: Not Just an interchange format anymore - but a new research community. *Computers and the Humanities*, 33, 1-9.
- Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P. & Huang, C.-H (2009). Wiktionary and NLP: Improving synonymy networks. *Proceedings of the ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, p. 19-27.
- Niculae, V. (2013). Comparison pattern matching and creative simile recognition. *Proceedings of the Joint Symposium on Semantic Processing. Textual inference and Structure in Corpora*, 110-114.
- 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.
- Niculae, V., & Yaneva, V. (2013). Computational considerations of comparisons and similes. *Proceedings of the ACL Student Research Workshop*, 89-95.
- Nivre, J. (2005). *Dependency grammar and dependency parsing*. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.
- Norton, L. (2013). *Aspects of Ecphrastic Technique in Ovid's Metamorphoses*. Newcastle: Cambridge Scholars Publishing.
- Norrick, N. R. (1986). Stock similes. *Journal of Literary Semantics*, 15(1), 39-52.
- Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational Theory*, 25(1): 45-53.
- Ortony, A. (1978). *Beyond literal similarity*. Technical Report No. 105. Cambridge: Bolt Berank & Newman.
- Ortony, A. (1979). Beyond literal similarity. *Psychologic Review*, 86(3), 161-180.
- Owen, O. F. (1853). *The Organon, or Logical Treatises of Aristotle with the Introduction of Prophyry*. Vol. I. London: Henry G. Bohn.
- Parmentier, M. (2002). *Dictionnaire français/anglais des comparaisons = English/French dictionary of similes*. Québec: Stanké.
- Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2), 161-199.
- Peachum, H. (1593). *The Garden of Eloquence*. New York: Scholars' Facsimiles & Reprints.
- Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13, 11-24.
- Peprnik, J. (1996). Cold Colours in 19th-century English Literature. *Contemporary Linguistics*, 41-42 (1/2), 503-517.
- Peprnik, J. (2000). Warm Colours in 19th-century English Literature. *Philosophica*, 73, 13-31.

- Perri, C. (1978). On Alluding. *Poetics* 7(3), 289-307.
- Philip, G. (2006). Connotative meaning in English and Italian Colour-Word Metaphors. *Metaphorik*, 10, 59-93.
- Plutarch. (1923). *The Parallel Lives*. Trans. Bernadette Perrin. V IX. Loeb Classical Library edition (Cambridge, MA and London)
- Pistorius, G. (1971). La structure des comparaisons dans "Madame Bovary". In: *Cahiers de l'Association internationale des études françaises*, 1971, 23. 223-242. doi: 10.3406/caief.1971.985.
- Poe, E. A. (1884). The Tell-Tale Heart. *The Works of Edgar Allan Poe*. 568-575. New York: A. C. Armstrong & Son. Vol II.
- Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1-39.
- Pratt, A. E. (1898). *The Use of Color in the Verse of the English Romantic Poets*. Chicago: The University of Chicago Press.
- Puttenham, G. (1589). *The Arte of English Poesie*. London: Richard Field.
- Qadir, A., Riloff, E., & Walker, M. A. (2015). Learning to recognize affective polarity in similes. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 190-200.
- Quintilian. (1876). *Institutes of Oratory*. (Trans J.S. Watson). Vol. I & II. London: George Bell and Sons.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London & New York: Longman.
- Raub, A. (1888). *Practical Rhetoric and Composition: A Complete and Practical Discussion of Capital Letters, Punctuation, Letter-Writing, Style, and Composition*. Philadelphia: Raub & Co.
- Rayner, M., & Banks (1988). *Parsing and Interpreting Comparatives*. Proceedings of the 26<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, 49-60.
- Renear, A. J. (2003). Text Encoding. In S. Schreibman, R. Siemens, J. Unsworth (Eds), *A Companion to Digital Humanities*. Oxford: Blackwell. Retrieved from <http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-5&toc.depth=1&toc.id=ss1-3-5&brand=default>
- Richards, I. A. (1936). *The Philosophy of Rhetoric*. London, Oxford & New York: Oxford University Press.
- Riffaterre, M. (1964). Fonctions du cliché dans la prose littéraire. *Cahiers de l'Association internationale des études françaises*, 16, 81-95.

- Rimbaud, A. (1922). *Œuvres Complètes d'Arthur Rimbaud : Les Illuminations*. Paris: Éditions de la Banderole.
- Rivara, R. (1990). *Le Système de la comparaison: Sur la construction du sens dans les langues naturelles*. Paris: Les Editions de Minuit.
- Roberts, R. M., & Kreutz, R.J. Why do people use figurative language? *Psychological Science*, 5 (3), 159-163.
- Rodenbach, G. (n. d.) *Le Rouet des brumes; contes*. Paris: Ernest Flammarion.
- Romary, L., Salmon-Alt, S., & Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. *Proceedings of the Workshop Enhancing and Using Electronic Dictionaries*, 22-28.
- Roncero, C, Kennedy, J. M., & Smyth, R. (2006). Similes on the Internet have explanations. *Psychonomic Bulletin & Review*, 13(1), 74-77.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Rosch, E., Mervis, B., Gray, W. D., Johnson, M.D., Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.
- Rothenhäusler, K, & Schütze, H. (2009). Unsupervised classification with dependency based word spaces. *Proceedings of the EACL Workshop on GEMS: Geometrical Models of Natural Language Semantics*, 17-24.
- Ryan, K. (1981). Corepresentation grammar and parsing English comparatives. *Proceedings of the 19<sup>th</sup> Annual Meeting of the Association of Computational Linguistics*, 13-18.
- Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, 859-866.
- Schmid, H. (1994). Probabilistic part-of-tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44-49.
- Shabat Bethlehem, L. (1996). Simile and figurative language. *Poetics Today*, 17(2), 203-240.
- Shakespeare, W. (1916). *King Richard III*. New York: The Macmillan Company.
- Shen, Y. (1995). Cognitive constraints on directionality in the semantic structure of poetic vs non-poetic metaphors. *Poetics*, 23, 255-274.
- Shopen, T. (1973). Ellipsis as grammatical indeterminacy. *Foundations of Language*, 10, 65-77.
- Simpson, P. (2004). *Stylistics: A Resource Book for Students*. London and New York: Routledge.
- Simpson, R., Page, K. R. & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. *Proceedings of the 23<sup>rd</sup> International Conference on World Wide Web*, 1049-1054.

- Singer, K. (2013). Close reading: TEI for teaching poetic vocabularies. *The Journal of Interactive Technology and Pedagogy*, 3. Retrieved from <http://jitp.commons.gc.cuny.edu/digital-close-reading-tei-for-teaching-poetic-vocabularies/>
- Snos, R., O'Connor, B., Jurafsky, D., & Ng, A. Y., (2008). Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254-263.
- Sojcher, J. (1969). La métaphore généralisée. *Revue Internationale de Philosophie*, 23, 87(1), 58-68.
- Stassen, L. (2013). Comparative Constructions. M. S. Dryer & M. Haspelmath (Eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from <http://wals.info/chapter/121>
- Staab, S. (1998). *Grading Knowledge: Extracting Information from Texts*. Berlin, Heidelberg and New York: Springer-Verlag.
- Staab, S., & Hahn, U. (1997b). Comparatives in context. *Proceedings of the 14<sup>th</sup> National Conference on Artificial Intelligence*, 616-621.
- Staab, S., & Hahn, U. (1997b). “Tall”, “good”, “high” - Compare to what? *Proceedings of the International Joint Conference on Artificial Intelligence*, 996-1001.
- Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasa, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam & Philadelphia: John Benjamins.
- Steinvall, A. (2002). English Colour Terms in Context. Ph.D. Dissertation. Umeå University: Skrifter från moderna språk 3.
- Strachan, J., & Terry, R. (2000). *Poetry*. Edinburgh: Edinburgh University Press.
- Stutterheim, C. F. P. (1941). Het Begrip Metaphoor: Een taalkundig en wijsgerig onderzoek. Amsterdam: H. J. Paris. Doctoral Dissertation Universiteit van Amsterdam.
- Tamba-Mecz, I. (1983). L'ellipse, phénomène discursif et métalinguistique. *Histoire Epistémologie Langage*, 5(1), 151-157.
- TEI: P5 Guidelines. Retrieved from <http://www.tei-c.org/Guidelines/P5/>
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.
- The Bible*. King James Version.
- Thomas, A. L. (1979). Ellipsis: The interplay of sentence structure and context. *Lingua*, 47, 43-68.
- Tigges, W. (1988). *An Anatomy of Literary Nonsense*. Amsterdam: Editions Rodopi.
- Tobback, E., & Defrancq, B. (2008). “Comme” devant l'attribut de l'objet, une approche constructionnelle. *Linx*, 58, 97-117.

- Tomita, S. (2008a). Rhetorical expressions by simile in David Copperfield. *On-line Proceedings of the Annual Conference of the Poetics and Linguistics Association (PALA)*. Retrieved from <http://www.pala.ac.uk/uploads/2/5/1/0/25105678/tomita2008.pdf>
- Tomita, S. (2008b). Similes in Oliver Twist: Humanisation and dehumanisation. *ERA*, 25 (1 & 2), 25-42.
- Trousseau, R. (1981). La fonction des images végétales dans *À la recherche du temps perdu*. Académie royale de langue et de littérature françaises de Belgique. Retrieved from <http://www.arllfb.be/ebibliotheque/communications/trousseau10011981.pdf>
- Tucker, J. (1998). *Example stories: Perspectives on four parables in the gospel of Luke*. Sheffield: Sheffield Academic Press.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.
- Ullman, S. (1964). Style et expressivité. *Cahier de l'Association internationale des études françaises*. 16 (1), 97-108.
- Ullman, R. (1972). Some features of basic comparative constructions. *Working Papers on Language Universals*, 9, 117-162.
- Vanoncini, A. (2004). Balzac et les couleurs. *L'Année balzacienne* 1/2004 (5), 355-366.
- Veale, T. (2012). A computational exploration of creative similes. In F. MacArthur, J. L. Oncins-Martinez, M. Sanchez-Garcia & A. Maria Piquer-Piriz (Eds.), *Metaphor in Use: Context, culture, and communication* (pp. 329-344). Amsterdam & Philadelphia: John Benjamins.
- Veale, T. (2013). Strategies and tactics for ironic subversion. In M. Dynel (Ed.), *Developments in Linguistics Humour Therapy* (pp. 321-340).
- Veale, T., & Hao, Y. (2007). Learning to understand figurative language: from similes to metaphors to irony. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, 683-688.
- Veale, T., & Hao, Y. (2009). Support structure for linguistic creativity: A computational analysis of creative irony in similes. *Proceedings of CogSci 2009, the 31st annual meeting of the cognitive science society*, 1376-1381.
- Veale, T., & Li, G. (2013). Creating similarity: Lateral thinking for vertical similarity judgments. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 660-670.
- Waddy, V. (1889). *Elements of Composition and Rhetoric*. New York, Cincinnati, Chicago: American Book Company.
- Walaszewska, E. (2013). Like in similes – A relevance-theoretic view. *Research in Language*, 11(3), 323-334.

- Warwick, C. (2004) Print scholarship and digital resources. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Humanities*. Oxford: Blackwell.
- Weiner, E. J. (1984). A Knowledge representation approach to understanding metaphors. *Computational Linguistics*, 10(1), 1-14.
- Weiner, E. J. (1987). Computational considerations for the processing of explanatory literal analogies. *Computers and the Humanities*, 21, 91-101.
- White, A. H. (1910). A Poet of the People. In *Uncle Walt [Walt Mason]: The Poet Philosopher*, (pp. 13-14). Toronto: The Musson Book Company.
- Wilstach, F. J. (1916). *A Dictionary of Similes*. Boston: Little, Brown, and Company.
- Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Consumer Intelligence. *Decision Support Systems*, 50, 743-754.
- Yang, S., & Ko, Y. (2009). Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques. *Proceedings of the ACL-IJCNLP Conference Short Papers*, 153-156.

# 10 APPENDICES

APPENDIX 1 PART-OF-SPEECH TAG SUBSETS	210
APPENDIX 2 SYNTACTIC LABELS	211
APPENDIX 3 EXAMPLE OF AN ANNOTATED GLARF OUTPUT	212
APPENDIX 4 SENTENCES TESTED FOR THE SEMANTIC MODULE	214
APPENDIX 5 EXTRACTION METHOD FOR NOUN + CT SIMILES	218
APPENDIX 6 BRITISH AUTHORS IN THE CORPUS	220
APPENDIX 7 FRENCH NOVELISTS IN THE CORPUS	222

## APPENDIX 1 PART-OF-SPEECH TAG SUBSETS

### A – Corresponding Penn Treebank part-of-speech tags

<b>CC</b> Coordinating conjunction	<b>TO</b> Infinitival to
<b>DT</b> Determiner	<b>RB</b> Adverb
<b>IN</b> Preposition	<b>RBR</b> Adverb, comparative
<b>JJ</b> Adjective	<b>VB</b> Verb, base form
<b>JJR</b> Adjective, comparative	<b>VBD</b> Verb, past tense
<b>MD</b> Modal	<b>VBG</b> Verb, gerund/past participle
<b>NN</b> Noun, singular or mass	<b>VBN</b> Verb, past participle
<b>NNS</b> Noun, plural	<b>VBP</b> Verb, non-3rd person singular present
<b>NNP</b> Proper noun, singular	<b>VBZ</b> Verb, 3rd person singular present
<b>NNPS</b> Proper noun, plural	<b>WDT</b> Wh-determiner
<b>PP\$</b> Possessive pronoun	, Comma
<b>PRP</b> Personal pronoun	. Sentence-final punctuation
	: colon, semi-colon

### TreeTagger's variants

<b>VVZ</b> Verb, 3rd person singular present	<b>SENT</b> Sentence-final punctuation
--	--

### B – Corresponding Brown Corpus part-of-speech tags

<b>AT</b> article	<b>VBN</b> Verb, past participle
<b>BER</b> Verb “to be”, present tense, 2 <sup>nd</sup> person singular/all persons plural	
<b>CC</b> Coordinating conjunction	<b>IN</b> Preposition
<b>NN</b> Noun, singular or mass	<b>NNS</b> Noun, plural
<b>RB</b> Adverb	. Sentence-final punctuation

### C – Corresponding British National Corpus (BNC) part-of-speech tags

<b>AJ0</b> Adjective (general or positive)	<b>NP0</b> Proper noun
<b>AT0</b> Article (the, a, an, no)	<b>PRP</b> Preposition
<b>AV0</b> General adverb	<b>PUN</b> Punctuation
<b>NN1</b> Singular common noun	<b>NN2</b> Plural common noun
<b>VBZ</b> Verb “to be”, present tense, 3 <sup>rd</sup> person singular	

## APPENDIX 2 SYNTACTIC LABELS

I – Bracket labels at the phrase and the clause levels

**ADV** Adverb phrase

**NP** Noun phrase

**SBJ** Subject

**VP** Verb phrase

**S** Simple declarative clause

**PP** Prepositional phrase

II – Chunk labels

**NP** Noun phrase

**VC** Verb phrase

**PP** Prepositional phrase

III – Dependency relations

**NMOD** Modifier of nominal

**ADV** Adverb

**AMOD** Modifier of adjective or adverb

**DEP-GAP** Gapping

**DIR** Direction

**CONJ** Between conjunction and second conjunct in a coordination

**OBJ** Object

**P** Punctuation

**PMOD** Modifier of preposition

**PRD** Predicative complement

**ROOT** Root

**SBJ** Subject

**SBAR** Subordinate clause

**SUB** Subordinated clause (dependent on subordinating conjunction)

**VC** Verb chain

**VMOD** Modifier of verb

**VOC** Vocative



*Appendices*

```
(FACTIVITY DEFINITE) (SEM-TENSE PAST) (INDEX 18))
(PTB2-POINTER |10+1|) (SEM-TENSE PAST) (FACTIVITY
DEFINITE)))
(PTB2-POINTER |10+2|) (INDEX 12) (SEM-TENSE PAST)
(FACTIVITY DEFINITE)))
(PTB2-POINTER |9+2|) (INDEX 13) (SEM-TENSE PAST)
(FACTIVITY DEFINITE)))
(PTB2-POINTER |6+2|) (INDEX 10) (FACTIVITY DEFINITE)))
(PTB2-POINTER |5+1|) (INDEX 16)))
(PTB2-POINTER |4+1|) (SEM-TENSE PAST) (FACTIVITY DEFINITE)))
(PUNCTUATION2 (|. | |. | 13)) (PTB2-POINTER |0+2|) (TREE-NUM 39)
(FILE-NAME "lw1") (INDEX 0) (SEM-TENSE PAST) (FACTIVITY DEFINITE)
(SENTENCE-OFFSET 3285)))
```

## APPENDIX 4 SENTENCES TESTED FOR THE SEMANTIC MODULE

<i>Sentences</i>	<i>Remarks</i>
Only once had he returned after they all left and that had been bad enough, <u>like a dream</u> -- no, like stepping into the set and scenario of some frightening film, a Hitchcock movie perhaps.	Simile; the adjective expresses a salient feature of the vehicle.
This boat was called Dream Baby, and she was clearly an expensive infant for rods and whip-aerials and outriggers splayed from her upperworks <u>like the antennae of some outlandish insect</u> .	Comparison, both the vehicle and the tenor belong to the same semantic category: man-made objects. Recorded as a simile in the corpus.
And now he was passing a second and more dilapidated pillbox and it struck him that the whole headland had the desolate look of an old battlefield, the corpses long since carted away but the air vibrating still with the gunfire of long-lost battles, while the power station loomed over it <u>like a grandiose modern monument to the unknown dead</u> .	Simile; the verb expresses a salient feature of the vehicle.
And now he was passing a second and more dilapidated pillbox and it struck him that the whole headland had the desolate look of an old battlefield, the corpses long since carted away but the air vibrating still <u>with the</u> gunfire <u>of long-lost battles</u> , while the power station loomed over it like a grandiose modern monument to the unknown dead.	Comparison.  This comparison is not tagged in the corpus and is therefore, deemed literal.
Usually the slightest whisper travelled <u>like jungle drums through the world of fashion</u> .	Simile; the verb expresses a salient feature of the vehicle.
In the fitting rooms at Taylors she fussed and fretted over her creations <u>like a mother hen</u> and though Paula was overawed by the great designer she also liked her on sight.	Simile; extended vehicle.
In spite of the rain, the earth was still <u>as hard as iron</u> .	Simile; idiom.
I like the secretiveness of a boat in the blackness, when the only thing to dislike is the prospect of dawn, which seems <u>like a betrayal</u> because, at night, in a boat under sail, it is easy to feel very close to God -- for eternity is all around.	Simile; extended vehicle.
Confronted with the need to proceed, Delaney took risks,	

plummeting feet first through the hatchways, and partly breaking his descent with the handrails, falling <u>like</u> a <b>parachutist</b> , rolling instantly deploying his Uzi against... Against what?	Simile; vehicle preceded by an indefinite article.
Relief surged through her <u>like</u> a <b>physical infusion of new blood</b> .	Simile; the verb expresses a salient feature of the vehicle.
`And tell me, sweet creature, do you count <u>as</u> a <b>toy</b> ?	Pseudo-comparison; not in the corpus and is therefore, judged literal.
Although he was fourteen years younger than Alexander, Daniel too was in the habit of thinking of himself <u>as</u> a <b>survivor</b> , a battered and grizzled survivor.	Pseudo-comparison; considered as a simile in the corpus.
You look, how you say?, <u>as</u> a <b>raccoon</b> .'	Simile; vehicle preceded by an indefinite article.
He turned on me <u>like</u> a <b>snake</b> .	Simile; the verb expresses a salient feature of the vehicle.
But it struck <u>with the</u> speed <u>of</u> a <b>snaking snake</b> .	Simile; abstract attribute possessed by a concrete entity.
She had known him since he was a very small five-year-old, perched <u>like</u> a <b>mosquito on one of the placid beginners' ponies</b> , so she told the class to carry on walking their ponies while she came to him.	Simile; extended vehicle
Madame Mattli might be a stickler for detail, with a generous helping of the artistic temperament which kept her tight-coiled <u>as</u> a <b>spring</b> and which would explode into frenzy if the smallest detail was not as it should be, but she also had a kind face and deep perceptive eyes.	Pseudo-comparison; considered as a simile in the corpus.
The encounter he now saw <u>as</u> a <b>omen</b> , a shadow cast by a coming event.	Pseudo-comparison considered as a simile in the corpus.
Hanged <u>like</u> a <b>chicken by his neck</b> , in town.'	Simile; the verb expresses a salient feature of the vehicle.
Was it even now shadowing them, moving soundlessly from cover to cover, <u>like</u> a <b>tiger in the steel jungle</b> ?	Simile; the verb expresses a salient feature of the vehicle.
When she checked through the spyhole it was standing in exactly the same spot, unmoving, <u>like</u> a <b>lizard</b> .	Simile; vehicle preceded by an indefinite article.

Nemesis had still come down <u>like</u> <b>the wolf on the fold</b> .	Simile; the verb expresses a salient feature of the vehicle.
Perfectly groomed from head to toe and with all that assurance, she was ready to take on the world, Arlene thought with satisfaction, for she looked on Paula <u>as</u> <b>her very own creation</b> .	Pseudo-comparison considered as a simile in the corpus.
If the approach was that way he would get no warning at all, and it would be on top of George -- his name for the dummy sitting <u>like</u> <b>a drunken son-of-a-bitch</b> -- before he knew it.	Simile; the verb expresses a salient feature of the vehicle.
At the top they came out into uncompromising, bright grey light, the bleak, hedgeless lane, the flat meadows where here and there stunted trees squatted <u>like</u> <b>old men in cloaks</b> .	Simile; the verb expresses a salient feature of the vehicle.
He looked, in this setting, a little <u>like</u> <b>some painter</b> .	Perceptual simile.
And beyond, green grass and geraniums <u>like</u> <b>splashes of blood</b> .	Simile; the vehicle and the tenor belong to distinct semantic category.
He'd never been one to exercise an over-imagination, yet the conditions were <u>like</u> <b>the feeling of a tomb -- of an interment</b> .	Comparison, both the vehicle and the tenor are synonyms in the database. Recorded as a simile in the corpus.
In the catalogue John House quoted Monet's description of the painted light around the snowy haystacks <u>as</u> <b>an enveloping veil</b> .	Pseudo-comparison considered as a simile in the corpus.
`One in ambush, with the rest of us acting <u>like</u> <b>beaters</b> .	Simile; vehicle not preceded by any article.
This sombre giant -- like a defeated proud man -- contrasts, when considered in the nature of a living creature, <u>with the</u> pale smile <u>of</u> <b>a last rose on the fading bush in front of him...</b>	Simile; the vehicle and the tenor belong to distinct semantic category. However "with" here is part of a phrasal verb; it is therefore not a simile.
This sombre giant -- <u>like</u> <b>a defeated proud man</b> -- contrasts, when considered in the nature of a living creature, with the pale smile of a last rose on the fading bush in front of him...	Simile; vehicle preceded by an indefinite article.
Anthony, recognizing incompetence, grasped Dalglish's hair firmly with a sticky hand and he felt the momentary touch of a cheek, so soft that it was <u>like</u> <b>the fall of a petal</b> .	None
He was a skilful lover: tender and gentle in the beginning, then powerful, persistent, rough almost -- until, his passion rising in harmony with hers, the climax came like the bursting of a thousand stars, <u>like</u> <b>the beginning and ending of the world</b> .	Simile; the verb expresses a salient feature of the vehicle.

<p>He was a skilful lover: tender and gentle in the beginning, then powerful, persistent, rough almost -- until, his passion rising in harmony with hers, the climax came <u>like</u> <b>the bursting of a thousand stars</b>, like the beginning and ending of the world.</p>	<p>Simile; the verb expresses a salient feature of the vehicle.</p>
<p>John House, who had organised the exhibition, came almost leaping down the stairs accompanied by a smallish woman in a pine-green <b>tent-like</b> coat.</p>	<p>Comparison, both the vehicle and the tenor belong to the same semantic category: man-made objects. Recorded as a simile in the corpus.</p>
<p>John House, who had organised the exhibition, came almost leaping down the stairs accompanied by a smallish woman in a <b>pine-green</b> tent-like coat.</p>	<p>Comparison, the vehicle is a concrete entity. Not recorded in the corpus.</p>
<p><u>Like a chameleon</u>, it moved out of the aisle between machines, then stopped, and became utterly motionless.</p>	<p>Simile; the verb expresses a salient feature of the vehicle.</p>
<p>At supper, as at lunch, Robin-Anne ate <u>with the</u> appetite <u>of a horse</u>, though her brother hardly touched his chicken and pasta salad.</p>	<p>Simile; abstract attribute possessed by a concrete entity.</p>
<p>It was too large for her and the wide sleeves of limp cotton hung from her freckled arms <u>like</u> <b>rags thrown over a stick</b>.</p>	<p>Simile; the verb expresses a salient feature of the vehicle.</p>
<p>They briefly appeared on deck for lunch; a meal which Rickie hardly touched, while Robin-Anne, despite her apparent frailty, attacked the sandwiches and salad <u>with the</u> savagery <u>of a starving bear</u>.</p>	<p>Simile; the verb expresses a salient feature of the vehicle.</p>
<p>In seconds, poor old George would be spread around the room <u>like</u> <b>an explosion in Harrod's window</b>, and the thing would be away.</p>	<p>Simile; the verb expresses a salient feature of the vehicle.</p>
<p>He strokes its side, which is white and marked with round patches of black, <u>like</u> <b>islands on a naïvely drawn map</b>.</p>	<p>Simile; the vehicle and the tenor belong to distinct semantic category.</p>
<p>Frederica kissed him too, reflecting that he was dressed <u>like</u> <b>a man who smelled dirty</b>, but in fact didn't.</p>	<p>Simile; vehicle preceded by an indefinite article.</p>
<p>They prepared chicken pies, the pastry <u>as</u> light <u>as</u> <b>Ruth's heart</b>, turtle soup, a haunch of venison, jellies, blancmanges, syllabubs, trifles, and a host of other dishes, with still more to be done on the day of the party itself.</p>	<p>Simile; the adjective expresses a salient feature of the vehicle.</p>

## APPENDIX 5 EXTRACTION METHOD FOR NOUN + CT SIMILES

The method described here can also be applied with some modifications to similes built with the suffix “-like”, the main difference being that for noun+colour term (CT) similes, the colours to consider must first be selected. In addition to the basic English colour terms defined by Berlin and Kay (1969), 6 other colour terms were haphazardly chosen: “turquoise”, “violet”, “azure”, “mauve”, “indigo” and “rose”. Then, sentences that contain a noun+CT adjectives were extracted in each novel. For this phase, each text was pre-processed with TreeTagger, a freely available multilingual tokeniser, lemmatiser, part-of-speech tagger and chunker (Schmid, 1994). The obtained output first served to determine sentence boundaries. Afterwards, all words of the form “X-CT”, i.e. words that end with one of the selected colour terms preceded by a hyphen are identified. Since all words of this form correspond not only to noun+CT adjectives but also to noun+CT nouns, CT+CT adjectives or nouns, a filtering took place. In the first stage, all sentences in which X refers to another colour term, an adjective or a word denoting colours that possesses more than one lexical form such as “light” or “deep” were deleted using the GCIDE and a manually compiled list of unwanted words. The second and last stage concerns the removal of all cases in which the X-CT word is used as a noun (“Will you wear the smoke-grey, tonight?”) or designate either a prefix or a specific colour shade (“field-grey uniform”).

Next, the different components of these similes were automatically identified using hand-crafted rules. For obvious reasons, once the noun+CT adjective is known, the vehicle and the ground of the simile are very easy to determine: the former constitutes the first part of the noun+CT adjective while the latter is the colour term. Since the topic is imperatively the noun that the adjective modifies, based on the English syntax, it is possible to derive the function of the topic in the sentence from the position of the adjective. For example, if the adjective is used attributively, it is generally immediately followed by the vehicle. All plausible scenarios are summarised below (the noun+CT is underlined and the topic is in bold):

Type of adjective	Position or Function	Example
Appositive adjective	Head of the noun phrase that precedes or follows the adjective	<i>The <b>heavy</b> face, now <u>brick-red</u> with summer suns, did not change.</i>
Attributive adjective		<i>The circle round the <u>silver-grey</u> mare narrowed slowly.</i>
Predicative adjective	Verb subject or complement	<i>Andrew's <b>back</b> was <u>blood-red</u> in the brazier light.</i>

The retrieved topics were then reviewed manually and corrected when necessary.

## APPENDIX 6 BRITISH AUTHORS IN THE CORPUS

<b>Name</b>	<b>Interval of publication of chosen texts</b>	<b>Number of texts in the corpus</b>
Jane Austen	1811-1818	6
Walter Scot	1814-1829	23
Mary Shelley	1818-1837	6
Benjamin Disraeli	1826-1880	14
Edward Bulwer-Lytton	1827-1873	16
William Ainsworth	1834-1876	13
Charles Dickens	1837-1870	14
William Makepeace Thackeray	1840-1859	10
Charlotte M. Yonge	1844-1900	38
Charlotte Brontë	1847-1857	4
Anthony Trollope	1847-1884	47
Elizabeth Gaskell	1848-1863	6
Charles Kingsley	1848-1866	7
Wilkie Collins	1850-1890	23
George Meredith	1856-1910	19
Frederic Farrar	1859-1895	5
George Eliot	1860-1876	7
Elizabeth Braddon	1862-1896	15
R. D. Blackmore	1864-1897	10
Lewis Carroll	1865-1889	3
William Black	1869-1891	10
Thomas Hardy	1871-1897	14
R. L. Stevenson	1883-1893	7
George Gissing	1884-1905	17
H. Rider Haggard	1884-1929	62
E. Nesbit	1885-1924	13
Maria Corelli	1886-1921	13
Fred White	1886-1943	69
Arthur Conan Doyle	1887-1906	16
Philips Oppenheim	1887-1943	96
George Griffith	1893-1906	11
John Buchan	1894-1940	29
Joseph Conrad	1895-1920	14
H. G. Wells	1895-1941	73
A. E. W. Mason	1895-1946	21
Bram Stoker	1897-1911	6
Elizabeth von Arnim	1898-1940	13
E. W. Hornung	1899-1909	7

*Appendices*

Elinor Glyn	1900-1927	16
Arnold Bennett	1902-1922	18
P. G. Wodehouse	1902-1934	23
Rafael Sabatini	1902-1944	31
G. K. Chesterton	1904-1927	6
John Galsworthy	1904-1933	17
Ford Maddox Ford	1906-1926	7
Harold Edward Bindloss	1906-1927	10
J. S. Fletcher	1907-1924	8
Jeffery Farnol	1907-1940	16
Edgar Wallace	1908-1936	90
Algernon Blackwood	1909-1918	9
D. H. Lawrence	1911-1929	12
Talbot Mundy	1913-1940	38
Virginia Woolf	1915-1941	9
Sapper	1919-1937	14
Arthur Gask	1921-1950	30
Hugh Walpole	1923-1943	32
Warwick Deeping	1923-1946	10
James Hilton	1924-1953	15
Josephine Tey	1929-1952	11
Charles Williams	1930-1945	7
Olaf Stapledon	1930-1950	8
George Orwell	1934-1949	6

## APPENDIX 7 FRENCH NOVELISTS IN THE CORPUS

<b>Name</b>	<b>Interval of publication of chosen texts</b>	<b>Number of texts</b>
Paul de Kock	1812-1832	4
Victor Hugo	1818-1874	8
Stendhal	1825-1894	4
Honoré de Balzac	1827-1848	45
George Sand	1832-1875	30
Théophile Gautier	1835-1863	4
Gustave Flaubert	1838-1869	4
Alexandre Dumas	1838-1872	58
Eugène Sue	1841-1849	7
Paul Féval	1843-1896	34
Pierre Ponson Du Terrail	1852-1879	26
Pierre Zaccone	1853-1882	4
Edmond About	1857-1862	4
Comtesse de Ségur	1858-1871	18
Octave Feuillet	1858-1872	5
Gustave Aimard	1858-1887	17
Erckmann-Chatrian	1862-1874	7
Henri-Émile Chevalier	1862-1879	13
Émile Gaboriau	1862-1881	13
Jules Verne	1863-1919	61
Émile Zola	1865-1903	31
André Gide	1865-1936	10
Alphonse Daudet	1868-1890	12
Hector Malot	1869-1896	17
Henry Greville	1876-1901	35
Jules Lermina	1876-1913	6
Zénaïde Fleuriot	1877-1882	5
René de Pont-Jest	1877-1889	4
Pierre Loti	1879-1906	8
Louis-Henri Bousсенard	1880-1912	7
Fortuné du Boisgobey	1881-1889	9
Anatole France	1881-1912	10
Guy de Maupassant	1883-1890	6
Octave Mirbeau	1883-1900	6
René Bazin	1884-1926	15
Paul Bourget	1885-1934	9
Jules Mary	1886-1898	4
Roger Dombre	1889-1910	4
Paul d'Ivoi	1895-1912	9
René Boylesve	1896-1920	6
Georges Le Faure	1896-1934	6
Gaston Leroux	1903-1927	29
Romain Rolland	1904-1912	10

*Appendices*

Paul-Jean Toulet	1904-1923	4
Gustave Le Rouge	1904-1927	9
Delly	1905-1913	4
Michel Zevaco	1906-1926	27
Arnould Galopin	1906-1930	6
Maurice Leblanc	1909-1935	18
Marguerite Audoux	1910-1920	4
Marcel Proust	1913-1927	7
Colette	1919-1941	8
Roger Martin du Gard	1922-1940	9
Georges Bernanos	1926-1950	8
Boris Vian	1946-1953	8

