Detecting Rhetorical Figures Based on Repetition of Words: Chiasmus, Epanaphora, Epiphora

(This Page will be Replaced before Printing)

Title page logo

Abstract Dummy Page.

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Dubremetz, Marie and Nivre, Joakim. Rhetorical Figure Detection: The Case of Chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 23–31, 2015.

II Dubremetz, Marie and Nivre, Joakim. Syntax Matters for Rhetorical Structure: The Case of Chiasmus. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 47–53, 2016.

III Dubremetz, Marie and Nivre, Joakim. Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 37–45, 2017.

IV Dubremetz, Marie and Nivre, Joakim. Rhetorical Figure Detection: Chiasmus, Epanaphora, Epiphora. Submitted

Reprints were made with permission from the publishers.

Contents

Acknowledgements

First and foremost, I am very grateful to my main advisor Joakim Nivre, who has taken an active interest in my work. I am grateful for his patience and his immense knowledge of natural language processing which, taken together, make him a brilliant mentor. I am also grateful to my assistant advisors, Marcel Cori and Mats Dahllöf, for their thoughtful guidance.

Special thanks are due to my current and former colleagues in the computational linguistics group for their support, scholarly interaction, kind messages, and company throughout the years. Special acknowledgements go to Fabienne Cap, Christian Hardmeier, Sara Stymne, Aaron Smith, Miryam de Lhoneux, Beáta Megyesi, Gongbo Tang, Yan Shao, Ali Basirat, Eva Pettersson, Nils Blomqvist, Mojgan Seraji, Oscar Täckström, and Mattias Nilsson.

My gratitude also goes to all the other members of the department for sharing meals, coffees, choir sessions or simply a good discussion. I would especially like to thank: Heinz Werner Wessler, Christer Henriksén, Karin Koltay, Eric Cullhed, Michael Dunn, Jakob Andersson, Helena Löthman, Johan Heldt, Christian Schaefer, and Harald Hammarström.

I also address my deep thanks and encouragement to all the other PhD fellows at the department, namely Rima Haddad, Josefin Lindgren, Alexander Nilsson, Marc Tang, Linnéa Öberg, Vera Wilhelmsen, Jaroslava Obrtelova, Fredrik Sixtensson, Myrto Veikou, Emil Lundin, Samuel Douglas, Mahmut Agbaht, Buket Öztekin, who contributed to making my workplace a nice place to be.

I would further like to thank the administrative and technical staff of Uppsala University for their support and professionalism, in particular Per Starbäck, Inga-Lill Holmberg, Jenny Rahbek and Ina Sörlid.

I am grateful to my friends, both in Lille and Uppsala. Thank you for staying by my side, even at times when I neglected you for the sake of my research.

Last, but not least, without my dear family I would never have completed this work. Thank you! You should know that your support and encouragement was worth more than I can express in writing.

1. Introduction

If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties.

— Francis Bacon

Language is full of repetitions of words. While reading this text, you have already encountered several of them without noticing it. And, in fact, it is normal; they are not an interesting event for our mind. However, sometimes we meet exceptions, like in the following text:

> And yet, as wondrous as it is, our lives are complex. Our emotions are complex. Our intellectual desires are complex.
>
> When it comes to complexity there is no short term fix in a pill. And your friends are long-term supports, and therefore, perhaps the most significant thing you can do to add more years to your life, and life to your years.

In this text, the repetition of the word "complex" three times, and the words "life" and "years" twice, probably caught your attention, or at least they caught your attention more than the repetition of the word "as" at the beginning of the text. We may not be conscious of making such a hierarchy in our attention to repetitions, but it is nevertheless an important part of our ability to correctly analyse, translate or read this text out loud in the most natural way. The authors of antiquity already knew about the properties of certain types of repetitions and included them in a subcategory of figures of speech which we call the figures of repetition. In this thesis we focus on three of them.

Chiasmus, which is the repetition of a pair of words in reverse order such as Example 1.

(1) If the **mountain** won't come to **Muhammad**,
then **Muhammad** must go to the **mountain**.

Epanaphora, the repetition of one or several words at the beginning of sentences as in Example 2.

(2) **My life is my** purpose.
My life is my goal.
My life is my inspiration.

Epiphora, the repetition of one or several words at the end of the text, like in Example 3.

(3) The United States, as the world knows, will never start **a war**.
We do not want **a war**.
We do not now expect **a war**.

1.1 Research Questions and Contributions

For a computer, locating all repetitions of words is trivial, but locating just those repetitions that achieve a rhetorical effect is not. Can this distinction be made automatically? The methods and models presented in this thesis focus on *how* a computer can detect those figures of repetition.

1. How should we define the task for detection of figures of repetition?
2. What challenges are posed by different types of figures with respect to features and machine learning?
3. Can we automatically learn how to weight different features, and how much manually annotated data do we need for that?
4. How do we evaluate the performance on this task?
5. Which types of linguistic features are useful?

The main contributions of this thesis are:

1. A general model for detecting repetitive rhetorical figures, including a method of evaluation.
2. An experimental evaluation of this model with respect to chiasmus, epanaphora and epiphora, exploring the usefulness of different features and comparing machine learning and hand-tuning of feature weights.
3. A pilot study comparing the frequency of different figures in different text genres.
4. The development and testing of three systems of detection designed for the three repetitive figures we are looking for.
5. A number of corpora annotated for chiasmus, epanaphora and epiphora.

1.2 Outline of the Thesis

Chapter 2 presents background information about figures of speech in general, and how they are split into categories. We describe the state of the art for their treatment in computational linguistics.

Chapter 3 introduces the general method we use for all of our repetitive figure detectors. We will explain the model used for ranking, the evaluation method, the annotation process, and the tuning process.

Chapter 4 summarises the empirical results of our experiments. We analyse the performance of each system and discuss its statistical significance.

Chapter 5 summarises the contributions of the thesis. We conclude with a discussion of promising directions for future research.

Chapter 6 provides a brief overview of the papers included in this thesis.

2. Background

This chapter provides background information on rhetoric, rhetorical figures and our three figures of speech from the perspective of literary analysis and computational linguistics. This review does not attempt to cover all relevant issues and is by no means exhaustive. Instead, we focus the discussion primarily on aspects that have some bearing on Papers I–IV.

2.1 Rhetoric

One cannot talk about figures of speech without invoking their antique theoretical framework. The concepts for the figures of speech have their historical roots in the field of rhetoric, whose aim is to teach speakers to convince and persuade their audience. The most fundamental notions of rhetoric were defined by Aristotle [Roberts, 2004] as follows:

> Of the modes of persuasion furnished by the spoken word, there are three kinds. The first kind depends on the personal character of the speaker [ethos]; the second on putting the audience into a certain frame of mind [pathos]; the third on the proof, or apparent proof, provided by the words of the speech itself [logos]. Persuasion is achieved by the speaker's personal character when the speech is so spoken as to make us think him credible. –Aristotle 1356a 2,3 (Translated by Roberts [2004])

In this extract, Aristotle gives three key notions to explain the mechanisms involved when trying to persuade someone. One can act on the ethos, which is the way one gives credit to the speaker. Then one can act on the pathos, that is, the appeal to the emotion of the audience. Finally one can act on the logos, which in Greek means both logic and discourse, and which can be defined as the way one articulates the ideas, the reasoning. While Aristotle preferred that arguments be based as much as possible on logos, he did pay considerable attention to the appeal to emotion, and discussed it extensively in his *Rhetoric*.

Ancient authors such as Hermogenes and Quintilian [Butler, 1921] were passionate about rhetoric and not only practised it but also tried to define and model it. They observed patterns of discourse that would benefit their argumentation. Among those patterns are the ornaments called figures of speech, also known as rhetorical figures. In his Oratio [Butler, 1921] Quintilian tries to describe all the figures. His classification still inspires works of classification today [Harris and DiMarco, 2009; Burton, 1996]. This classification is not uncontroversial. Indeed, the notion that figures of speech are additional ornaments that deviate from a “normal” way to express things has been repeatedly questioned in modern linguistics and philosophy [Jakobson, 1956; Derrida, 1982].

Figures of speech can, in fact, leverage any of the appeals identified by Aristotle (logos, pathos, ethos). In an extensive work of classification, Howard [2010] tries not only to define each figure of speech but also to label them with the appeal they are used for. For instance, he says that chiasmus and epanaphora appeal to logos and pathos. That can be a subject of controversy. For instance, Vandendorpe [1991] seems to criticise some examples of chiasmi for acting merely on the impression made by the author, and by extension the ethos. Indeed, for him, this figure is an over-sophisticated way to express what sometimes are very cliché ideas (for example, “the research of the meaning is the meaning of research”). Thus, for Vandendorpe [1991], this extreme ability to play with words could make the audience give too much credit to the speaker by thinking that he or she is clever. As we will see in Section 2.2, rhetorical figures are vaguely defined, and this is just one example of the controversies raised by this topic.

2.2 Rhetorical Figures

2.2.1 Schemes and Tropes

Figures of speech, also called rhetorical figures, are commonly divided into two categories, *tropes* and *schemes*. This separation between figures of speech comes from the classification of Quintilian [Fahnestock, 1999, p.196].

A trope is defined as:

> An artful deviation from the ordinary or principal signification of a word. [Burton, 1996]

Tropes such as “the Lord is my shepherd” (metaphor), “the pen is mightier than the sword” (metonymy), and “the face that launched a thousand ships” (synecdoche) operate by changing the signification of the words “shepherd”, “pen”, “sword”, and “face”.

Greene et al. [2012] define a scheme as a change in standard word order or patterns. This includes any artful deviation from the ordinary arrangement of words. Unlike tropes, which work on the signification of words, schemes involve their arrangement [Greene et al., 2012, Art. “Scheme”]. For instance an enjambement can be considered as a scheme. It is defined as an unnatural “cut”, which can result in stylistic effects (emphasis, contrast, double interpretations) (see Example 4).

(4) I'd rather
be naked
of fake friends

Beneath the category of schemes, there are many subcategories. Among them are the categories of repetitive figures or figures of repetition, which include every figure involving repetition of linguistic units such as letters, words or concepts. This is where chiasmus, epanaphora and epiphora belong. Depending on the sources, their number can vary, as will be explained for the case of chiasmus (Section 2.2.2). As for now, we can say that there is no one unique way to name and categorise figures. We are not going to list all figures of repetition,[1] but we can present three clear examples that are representative of what repetitive figures can be.

- Alliteration: Repetition of initial or medial consonants in two or more adjacent words.
- Anadiplosis: Repetition of a word at the end of a clause and then at the beginning of the succeeding clause.
- Antanaclasis: Repetition of a word in two different senses.

Another controversy that has some bearing on our work concerns whether a figure of speech (trope or scheme) should be considered only if it is *intentional* or, in other words, is intended by the author. Such a definition of figures of speech, used in some computational linguistics work [Strommer, 2011], causes a problem in the literature field.[2] For this reason we avoid focusing too closely on the term "intentional" because we do not want to fall into any author fallacy. This choice will be reflected in how we formulate our questions during the annotation process (see Section 3.5.2).

In this first part, we have given a general overview of what rhetoric is and how rhetorical figures can be defined. We have touched on aspects of the theoretical debate that can be relevant to our approach to detection. In the next section, we give a definition of the three figures we are studying. One figure, however, will be studied more extensively: chiasmus. Throughout this study, we will see how multiple challenges already arise when searching for a definition and a list of terms adapted to the requirements of computational linguistics.

[1] For an extended list of repetitive figures, we refer the reader to `http://rhetoric.byu.edu/Figures/Groupings/of%20Repetition.htm`.

[2] The status of the author has been widely questioned in both 20th-century and contemporary literary theory [Barthes, 1984]. Indeed, it is argued that no one can really pretend to know the intention of the author. We are aware of this controversy but will not address it here.

2.2.2 Chiasmus

The word “chiasmus” comes from the Greek letter χ because of the cross this letter symbolises. Indeed, chiasmus is generally defined as the reuse of a pair of elements in reverse order, as in Figure 2.1.

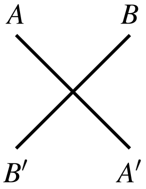

Figure 2.1. The chiasmus schema

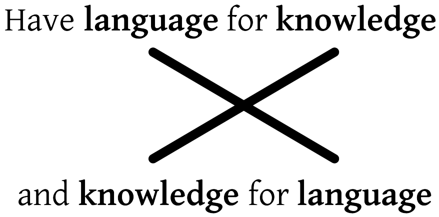

Figure 2.2. A chiasmus example

There are many forms of chiasmus. They can consist of cross-like arrangements of phonemes, letters, or syntactic elements. We will focus on the chiasmus of words, also called antimetabole. However, even within this single subcategory, we observe many variations that we need to define for the automatic treatment [Gawryjolek, 2009]. As a result of this, we present the types of chiasmus of words that we have identified.[3] This typology was obtained by reading a number of definitions given in the literature [Diderot and D’Alembert, 1782; Dupriez, 2003; García-Page, 1991; Greene et al., 2012; Pougeoise, 2001; Nordahl, 1971; Rabatel, 2008; Vandendorpe, 1991; Van Gorp et al., 2001] and observing what kinds of examples they are citing.

[3] This typology is inspired by and adapted from the one developed in French by Dubremetz [2013].

Name	Repeated Elements	Example
Antimetabole or Strict antimetabole	Identical word forms[4] (pair of strings)	**Sounds** of **poetry**, and **poetry** of **sounds**.
Flexional chiasmus or Flexional antimetabole	Inflected lemmas	An optimist **laughs** to **forget**, a pessimist **forgets** to **laugh**.
Derivational chiasmus	Derivations with same stem	**Modernise Islam** rather than **islamisation** of **modernity**.
Antimetalepse	Ideas or notions, but without necessarily reusing morphologically related words[5]	Who **dotes**, yet **doubts**; **suspects**, yet **loves**.

Table 2.1. *Typology of chiasmus of words*

Even if the term chiasmus is a recent one,[6] the figures described in this table are very old and have been known for a long time. For instance, already in Quintilian we observe the existence of antimetabole of identical terms but with different inflexions, though no name has ever been given to this phenomenon to differentiate it from the strict antimetabole. A similar lack of defining terms is noticeable for the chiasmi that allow derivations. This table is probably not exhaustive and it is dependent on the examples cited as canonical in the literature. It should not be seen as the sole possible classification of chiasmus and antimetaboles. In fact, some authors prefer to refer to antimetabole as reverted antithesis [Viklund, 2002]. Others, by contrast, want to consider those special cases a completely different figure called “reversion” [Fontanier, 1827, Art. Reversion].[7] As stated by [Watson, 1912, p. 3]:

> The importance attached to rhetorical expression among ancient writers is recognized not only by special treatises written by them, but in the compilations or writings on rhetoric by modern authors who have made special study of such treatises. These ancient writers discuss in bewildering profusion examples with subdivisions of tropes, figures of rhetoric, figures of syntax, etc. But frequently their definitions do not offer a basis for clear cut division. Hence it is often impossible to make any sharp or important distinction between tropes, figures rhetorical, and figures grammatical, and no two writers of the present day arrange them in exactly the same way.

Thus we see Table 2.1 as one possible classification that answers our specific problem of defining the task for a computer. Indeed without a clear, formal and adapted definition, it will be difficult to make any system of detection [Gawryjolek, 2009, p. 67].

[4] [Dupriez, 2003; Pougeoise, 2001, Art."Chiasme"]

[5] [Diderot and D’Alembert, 1782, Art. "Antimétabole, Antimétalepse, Antimétathèse"]

[6] This term does not appear until the second half of the 19th century [Horvei, 1985, p. 24].

[7] We can trace this back to another conflicting classification in Antiquity, as Hermogenes prefers to place chiasmus under a category of figure completely separated from other figures of repetition. For him chiasmus belongs to “gorgotes” (i.e.,“vigour”), whereas epanaphora and epiphora are a sort of parallel figures that should be placed under “kallos” (i.e., “beauty”) [Wooten, 2012].

From now on, we will restrict the term "chiasmus" to cases based on identity of lemma (i.e., flexional chiasmus in Table 2.1). This restriction is a compromise aiming to make the task feasible. Indeed, if we tried to detect all kinds of chiasmus, technical difficulties would pile up that would prevent any deep investigation. For example, the extraction process would become noisier and we would have to improve stemmers as well as ontologies [Dubremetz, 2013].

In this part we have investigated the definition(s) of chiasmus, and have proposed a typology. This typology divides chiasmi into different subtypes based on similarities between words, and can differ from some authors' definitions partly because there is a lack of consensus about what kind of figure each name should denote. This does not however mean that we completely ignore other definitions. In fact, we consider those other definitions to be observations about the figure behaviour, which provides a precious help in designing features. For instance, in Paper I, we decide to use negations like "not" in order to capture chiasmi based on antithesis. In Paper II we decide to model switching of syntactic roles between main words in order to model the notion of symmetry described by Morier [1961].

When we get to the actual experimentation in this thesis, the definition of a candidate chiasmus will be limited as follows:

1. A chiasmus is the repetition of a minimum of two pairs of words in reverse order satisfying the $ABB'A'$ schema, as in Figure 2.1.
2. The words repeated should appear within a reasonably limited context. In our actual implementation: 30 tokens.[8]
3. The word pairs (AA' and BB') should have the same lemma.

2.2.3 Epanaphora

The lack of definitions adapted to computational linguistics is not exclusive to chiasmus. In fact, any figure of speech suffers from a lack of consensus about the definitions, or from vagueness in delimiting what we are looking for. Epanaphora confirms this rule. Epanaphora[9] is defined as the repetition of a word or a group of words at the beginning of several sequences of language. Curiously, the ambiguity in the definition of epanaphora does not concern the same issue as for chiasmus. For chiasmus it is the type of element that can be criss-crossed (sounds, letters, strictly similar words or just synonyms) that is unclear. For epanaphora there seems to be some kind of consensus: only words should be the object of the repetition (not letters or sounds). And all

[8] 30 tokens is the upper bound found empirically by Dubremetz [2013]. In that corpus study, the largest chiasmus found consisted of 23 tokens.

[9] In rhetoric, *epanaphora* is better known under the competing term *anaphora*. However, in computational linguistics the term *anaphora* can be ambiguous, as it also refers to a referential pattern. For the sake of clarity, we will only use the term *epanaphora*.

examples cited by our references were concerned with repetition of words that at least share the same lemma. Where the definition gets blurry is: (1) What type of sequences should be considered? (2) How close together should they be?

Regarding the first question, "sequences" can be defined in different ways. One can speak of epanaphora of paragraphs [Bacry, 1992, Art. L'anaphore et l'épiphore], verses [Suhamy, 2004, Chap. III], clauses or phrases [Dupriez, 2003, Art. Anaphore]. All authors have at least invoked or cited in their examples the sentence as a possible sequence for epanaphora production. In addition, sentence splitting is a nearly solved problem in natural language processing. Sentences should not represent a major challenge for definition and extraction. That is why, in this thesis, we limit the scope to epanaphora of sentences, as exemplified in Example 5.

(5) **I am** an actor.
I am a writer.
I am a producer.
I am a director.
I am a magician.

On the second question, some authors, like Suhamy [2004, Chap. III] and Bacry [1992, Art. L'anaphore et l'épiphore], talk about "successive" sequences, while others seem to consider repetitions that skip one or several sequences before producing a repetition to be an epanaphora [Fontanier, 1827, p.329]. To keep the problem simple, we will consider only epanaphora relying on immediately successive sentences.

As for chiasmus we limit the definition of candidate epanaphora as follows:

1. An instance of epanaphora consists of two or more adjacent sentences.
2. One or more words are repeated at the beginning of each sentence.[10]
3. To count as repetitions, the words should have the same lemma.[11]

2.2.4 Epiphora

Epiphora[12] is a figure of speech with repetition at the end of a sequence (see Example 6).

[10] We started our exploration by requiring at least one word, but for feasibility we later had to restrict it to at least two words (see Paper IV).

[11] As for the previous constraint, we have later restricted it to a strict identity of word or string chain.

[12] Epiphora is also known under the term *epistrophe*, but for consistency with *epanaphora* we will only use the term *epiphora*.

(6) I'm so **gullible**.
I'm so damn **gullible**.
And I am so sick of me being **gullible**.

As for epanaphora, one speak of epiphora of paragraphs, lines, clauses or phrases, but as with epanaphora we limit the scope to sentences, as exemplified in Example 6. Epiphora is usually defined as merely epanaphora in reverse. In most of the sources that we have consulted [Bacry, 1992; Greene et al., 2012; Suhamy, 2004; Dupriez, 2003; Fontanier, 1827], this term does not even have a separate paragraph. There could be several reasons for this; perhaps its theoretical proximity to epanaphora makes it not useful to study this figure on its own. Perhaps those authors also suffer from a lack of examples to refer to. Indeed, one author [Bacry, 1992] claims that this figure is less frequent than its counterpart. He argues that the repetition of epanaphora is more visible than any other repetition because it is placed first. He assumes that this visibility makes this figure more popular. Bacry [1992] also seems to look for some specific property that would distinguish epiphora from epanaphora. For instance, he thinks that epiphora is less violent than epanaphora, and that it is particularly suitable in the context of melancholia.[13] As with chiasmus and epanaphora, we limit the definition of candidate epiphora:

1. An instance of epiphora consists of two or more adjacent sentences.
2. One or more words are repeated at the end of each sentence.
3. To count as a repetition, the words should have the same lemma.

2.3 Related Work in Computational Linguistics

In this section we will give an overview of the work related to our topics. A first part will discuss research targeting a variety of stylistic phenomena related to figures of speech. The second part focuses on one particular figure of speech frequently treated in computational linguistics: metaphor. The last part discusses our three rhetorical devices in the context of computational linguistics.

2.3.1 Stylistic Phenomena in Computational Linguistics

In a study about rhetorical figures, Harris and DiMarco [2009] claim:

> Too much attention has been placed on semantics at the expense of rhetoric (including stylistics, pragmatics, and sentiment). While computational approaches

[13] For an example of epiphora in the context of melancholia the reader can refer to the poem "The Raven" by Edgar Allan Poe [Poe, 1898].

> to language have occasionally deployed the word “rhetoric”, even in quite central ways (such as Mann and Thompson’s Rhetorical Structure Theory [Mann and Thompson, 1988]), the deep resources of the millenia-long research tradition of rhetoric have only been tapped to a vanishingly small degree.

Reading Harris and DiMarco [2009] one might think that natural language processing has only modelled semantic aspects of texts without much consideration for the form. However, this is probably not what they mean. It is certainly true that the use of very specialised terms describing rhetorical figures such as “antithesis”, “antimetabole” or “elision” is rare in our community.[14] However, one should not believe that NLP has totally ignored style and effect. In fact, we believe that a lot of research treats rhetoric in some way that involves figures of speech, but with slightly different focus and/or using different terms to talk about it. For instance, Booten and Hearst [2016] and Bendersky and Smith [2012] work on finding what makes sentences “quotable”. More precisely, what are the fundamental properties of pieces of text that push Internet users to cite them on specialised quotation websites [Bendersky and Smith, 2012] and on Tumblr under the label “#quote” [Booten and Hearst, 2016]? Other researchers talk about “memorability” of sentences [Danescu-Niculescu-Mizil et al., 2012]. There is even work on poetry that asks the question of what makes up the “quality” of a poem [Kao and Jurafsky, 2012], and others that look for how to model “humour” in a double entendre [Kiddon and Brun, 2011].

Considering these research studies, we believe that NLP is treating rhetoricity (i.e., how the text produces an effect on the reader through its form, the way it is written), but is exploring it with its own concepts (here, “memorability”, “humour”, “quotability”). There may be several reasons for this, but after doing the research on how to define chiasmus, we definitely believe that rhetorical figures are not easy to define. They suffer from a lack of consensus among rhetoricians and linguists. This is a real obstacle for NLP researchers; even before facing any implementation problem, they need to handle ambiguous definitions in a domain, rhetoric, that they do not feel qualified to discuss. This might change in the future with the growing interest in digital humanities. Indeed, in recent years, some work within machine learning has targeted very specific literary concepts such as free indirect speech [Hammond et al., 2013], similes [Niculae and Yaneva, 2013], enjambment [Ruiz et al., 2017], or rhymes [McCurdy et al., 2015]. In particular, the studies by Hammond et al. [2013] and Brooke et al. [2015], inspired by other research on subjectivity, like that of Morris and Hirst [2004], tend to allow subjectivity, multiplicity

[14] Observation based on key word search results on the ACL anthology `https://aclanthology.coli.uni-saarland.de`.

of text interpretation and variance in the intensity of the phenomena that they automatically detect.[15] This is an essential feature when analysing literature.

2.3.2 A Particular Case of Trope: Metaphor

When talking about figures of speech in computational linguistics, the first figure that comes to mind is metaphor.[16] In this domain, extensive work has been done within from computational linguistics, relying on cognitive science and corpus studies. Metaphor attracts the attention of computational linguists because it is seen as a massive and recurrent language phenomenon. This cognitive phenomenon [Lakoff and Johnson, 1980] commonly appears in language in the form of metaphoric expressions [Deignan, 2005]. The most comprehensive manual study of metaphoric expressions in large corpora [Steen et al., 2010] found that up to 18.5% of words in the British National Corpus were used metaphorically. This explains why metaphors are treated to a much greater extent in computational linguistics than any other figure: they are too frequent to ignore. Beginning with Wilks [1978], the issue of metaphor has been approached as an identification task: first look for metaphoric expressions and then (1) prevent them from interfering with the computational treatments of literal expressions and (2) use them to increase the understanding of the content (e.g., Carbonell [1980], Neuman and Nave [2009]). The task is generally defined like this: for a given unit of language (e.g., word, phrase, sentence) decide whether it is metaphoric or non-metaphoric. Neuman and Nave [2009] used selectional restrictions for this purpose; Mason [2004] used hand-crafted knowledge resources to detect similar selectional mismatches. Another approach is to detect selectional mismatches using statistically created resources (e.g., Shutova et al. [2013], Shutova and Sun [2013]). A second general approach to this classification problem has been to use mismatches in properties like abstractness [Gandy et al., 2013; Assaf et al., 2013; Tsvetkov et al., 2013; Turney et al., 2011], semantic similarity [Li and Sporleder, 2010b,a], and domain membership [Dunn, 2013b,a] to identify metaphoric units of language. A third approach has been to use forms of topic modelling to identify linguistic units which represent both a metaphoric topic and a literal topic [Bracewell et al., 2013].

All of these approaches view the task as a binary classification problem of distinguishing metaphoric language from non-metaphoric language. This binary distinction assumes a clear boundary between the two; in other words, it assumes that metaphoricity is a discrete property. However, three strands of

[15] See their website on automated analysis of Voices in T.S. Eliot's The Waste Land http://www.hedothepolice.org/. Their division of the poem into voices integrates the degree of confidence by the computer and has directly inspired our method [Hammond et al., 2013].

[16] According to our research on the key term "metaphor" in the ACL Anthology `https://aclanthology.coli.uni-saarland.de`, the term definitely stands out with more than 100 articles.

theoretical research show that metaphoricity is not a discrete property. First, psycholinguistic studies of metaphor processing show that there is no difference between the processing of metaphoric and non-metaphoric language [Coulson and Matlock, 2001; Gibbs, 2002; Evans, 2010]. The most plausible interpretation of this psycholinguistic evidence is that most linguistic units fall somewhere between metaphoric and literal, so that metaphoricity is a scalar value which influences processing gradually. Thus, the high frequency of metaphorically used language implies that it is hard to set a boundary beyond which a word is used metaphorically [Dunn, 2014]. Thus Dunn [2014] claims that 18.5% of the BNC is not highly metaphorical, but rather is the sort of slightly metaphoric language of which speakers are not consciously aware because it is used so frequently.

The approach of Dunn [2014] is actually something that we relate to in our consideration of the figures of repetition. Starting from Paper I and throughout Paper IV we repeatedly refer to the non-discrete property of rhetorical figures. This property will directly influence how we implement, evaluate, and annotate chiasmus, epanaphora, and epiphora.

2.3.3 Chiasmus, Epanaphora, Epiphora

Gawryjolek [2009] was the first to tackle the automated process of repetitive figures, and he built an annotation tool called JANTOR. As it was the first work ever done on this topic, he mostly focused on how to extract candidates. The result of this research effort is a graphical interface for human annotation [Gawryjolek, 2009, p.94]. Thus it is never really the machine that makes the distinction between true instances and accidental repetitions. For instance, for chiasmus, following the general definition of the figure, he proposed to extract every repetition of words that appear in a criss-cross pattern. At the end of his research, he concludes that the recall is perfect (100%) but that the precision is low. Based on our own investigations, we can give an idea of what he means by "low", using the example of *The River War* by Winston Churchill, a book consisting of 150,000 words, with 66,000 examples of criss-cross patterns but where we have only found one real chiasmus.[17] Hromada [2011] then proposed to restrict the procedure for candidate extraction; he drastically reduced the number of false positives by requiring three pairs of words to be repeated in reverse order without any variation in the intervening material. However, in the example of Churchill's book, this also eliminates the one real example, and the user ends up with a completely empty output. This does not mean that the filter chosen by Hromada [2011] is bad. In reality this algorithm is extremely precise, according to Dubremetz [2012]. What it means is that chiasmus, unlike metaphor, can be so rare (e.g., one in a book)

[17] **Ambition** stirs **imagination** nearly as much as **imagination** excites **ambition**.

that any sharp filtering attempt might leave a literature analyst with no material to analyse at all.

Like chiasmus, epiphora was treated from the perspective of extracting every candidate but without focusing on rhetorical versus accidental repetitions. Gawryjolek [2009] and Hromada [2011] both extract candidates at the end of sentences, clauses, and phrases. The main difference between the systems of Gawryjolek [2009] and Hromada [2011] is that they do not use the same technologies (Hromada [2011] works with regular expressions; Gawryjolek [2009] develops an infrastructure based on the Stanford Parser) and that the system of Hromada [2011] is supposed to be multilingual.[18] Their research is pioneering, and many questions and problems therefore remain or are not yet raised. For instance, they do not give precise data on the evaluation of their systems (e.g., recall and precision). In fact, they could not give such an evaluation. In contrast to the case of metaphor, no existing corpus was available in which those figures were annotated.

In fact, only epanaphora has been the object of deeper and more focused study in computational linguistics. Strommer [2011] is the first to explicitly work not only on the extraction of candidates, but also on the distinction between true and false instances. He is also the first to give some ideas of the core problems. For instance, thanks to him we know that the figure is rare. When he had two annotators annotate 152 candidates of epanaphora, only two of the candidates were judged as true by both annotators. He thus raises the problem of the imbalanced corpus, with a lot of false examples but few true candidates to train on. He is the first to have applied machine learning to repetitive figures of speech. His underlying aim is to use epanaphora as a metric of genre, and for this task, his detection needs to make binary distinctions. Finally, his choice of definition is slightly different from ours. Strommer [2011] starts from a broader definition of epanaphora than we do: he accepts that some epanaphora could have sentence gaps, as in Example 7.

(7) **I felt** moody and irritable.
I felt squished inside, I felt like standing in a field and twirling in circles [...].
Is it the driver's license?
I felt overwhelmed by it tonight.

This definition is definitely acceptable, but we think that it makes the task even more complicated. Strommer [2011] reports technical difficulties, mainly in getting enough annotations. Despite these difficulties, he describes some features useful for epanaphora, some of which we think are easy to transpose or to use with epiphora detection as well. These are the number of sentences, the presence of "strong" punctuation marks (! and ?) and the length of sentences (shorter than ten words).

[18] Though its effectiveness across languages may vary, as discussed by Dubremetz [2012].

In the next chapter, we not only consider the extraction of candidates for each of our three repetitive figures, but also go beyond the simple extraction of candidates towards feature engineering and machine learning. The general approach common to our four papers will be explained.

3. Methodology

In this chapter, we describe the approach common to all our papers. We motivate the most important decisions about the model we defend. We briefly describe the tools and corpora we use, and finally we discuss the annotation instructions.

3.1 Defining the Task

So far, the state of the art describes two approaches for rhetorical figures. The first consists of trying to make systems for all repetitive figures [Gawryjolek, 2009; Hromada, 2011]. This approach is the most general, but it prevents us from going deeply into the problem, the development of a corpus, the exploration of data. The second approach consists of keeping one figure into focus: epanaphora [Strommer, 2011]. Our way of proceeding is closer to the second approach; we focus initially on chiasmus, but with the aim to later apply our best method to the two other figures: epanaphora and epiphora.

One important element in our approach is that we redefine it as a ranking task. The reason for this is that we believe that ranking figures is more close to the reality of the situation than just binary detection. Indeed, some candidates for chiasmus, epanaphora and epiphora are easy to classify as real examples (8, 9, 10); and others are easy to classify as irrelevant instances (11, 12, 13); while some are unclear cases that combine properties of real examples and irrelevant instances (14, 15, 16).

Clear Cases of Chiasmus, Epanaphora, Epiphora

(8) **Comedy** without **darkness** rapidly becomes trivial.
And **darkness** without **comedy** rapidly becomes unbearable.

(9) **Did I** offer peace today?
Did I bring a smile to someone's face?
Did I say words of healing?
Did I let go of my anger and resentment?
Did I forgive?
Did I love?

(10) The first **is to be kind.**
The second **is to be kind.**
And the third **is to be kind.**

Non-Figure Repetitions

(11) The **copies** can only repeat themselves **word** for **word**. A virus is a **copy**.

(12) **It's** like fashion, like flares go out then skinny jeans come in, people want something fresh.
It's the strongest ever urban scene at the moment and I hope it can progress and keep getting stronger and be the base for something larger.

(13) Those powers that control the tent are not threatened at all by any activity that you engage in, in the shadows, that's not moving toward the **tent**.
And I am rather convinced that we have a generation that is so preoccupied with life in the shadows, they never even focus on getting to the sunlight where you open up the big **tent**.

Unclear Cases

(14) No **nation** can have a monopoly on **God**, but **God** will bless any **nation** whose people seek and honour His will as revealed by Christ and declared through the Holy Spirit.

(15) **I'm** not good at hiding my feelings.
I'm also not good at lying.
I'm very open about everything.

(16) It feels intimate, doesn't **it**?
I love **it**.

The fact that borderline cases like Examples 14, 15 and 16 exist is not surprising, and is not necessarily a problem in literature. Hammond et al. [2013] underlines that the study of literature is nourished by the plurality of interpretations of texts. Examples 14, 15 and 16 can be interpreted as either rhetorical figures or random repetitions by a literary analyst. The choice depends on the interpretation that someone wishes to make of the text, and is outside our control as researchers. Thus, eliminating those examples would be an arbitrary choice made by us, and the machine would not facilitate the plurality of interpretation desired by humans. If overused, a detector with only a binary output could even create a bias toward the machine that would normalise the interpretation assigned to repetitions of words.

To solve this issue, and make an effective detector that gives extended control to the literary analyst, we decide to see the problem not as a binary task but as a ranking task. The machine should give all the instances of repetitions, but in a sorted manner: from very prototypical true instances (like Examples 10, 11, 12) to less and less likely instances. Thus users may benefit from the help of the machine without relinquishing their autonomy to choose which borderline cases are useful for their literary interpretation.

3.2 Evaluation

Redesigning the task as one of ranking is the easiest way to take into account the gradedness of the phenomena we search for. However, it makes the evaluation less straightforward. In an ideal world, we would like to have a set of thousands of repetitions of each category (chiasmus, epanaphora, epiphora) all ranked by degree of rhetorical effect. Then we would try to achieve this exact ranking with a machine. The problem is that creating such a corpus would be very difficult and time consuming. Annotation time, given the noise generated by repetition extraction, is the real bottleneck of the detection problem. Apart from the fact that it is very time-consuming to annotate all candidates, it is very challenging for an annotator to sort them into a complete ranking.

As a practical compromise, we therefore limit annotation to three categories: True, False and Borderline. And instead of evaluating only by precision and recall, we use average precision, which measures not binary decisions, but whether true instances have been ranked higher than irrelevant cases. Moreover, when using data annotated by multiple annotators we count as True only those instances that have been annotated as True by all annotators. In this way, we make sure that systems are evaluated with respect to their capacity to rank good, prototypical instances of a figure above less good instances. We consider this a reasonable compromise between the theoretical ideal of having a complete ranking of candidates and the practical necessity of making annotation and evaluation feasible. Finally, our use of a three-way categorisation into True, False and Borderline makes it possible to apply more fine-grained evaluation methods at a later time.

As mentioned above, the most important metric for our task measures the ranking capacity of the machine and is called *average precision*. Average precision is calculated on the basis of the top n results in the extracted list, where n includes all positions in the list until all relevant instances have been retrieved [Zhang and Zhang, 2009]. The average precision is expressed by the following formula:

$$\sum_r \frac{P@r}{R} \tag{3.1}$$

r = rank for each relevant instance
$P@r$ = precision at rank r
R = number of relevant instances in gold standard

3.3 Ranking Model

We propose a standard linear model to rank candidate instances:

$$f(r) = \sum_{i=1}^{n} x_i \cdot w_i$$

where r is a candidate pattern, x_i is a set of feature values extracted from r, and w_i is the weight associated with feature x_i. Given candidates r_1 and r_2, $f(r_1) > f(r_2)$ means that r_1 is more likely to be a true figure of speech than r_2, according to the model.

We choose the linear model for its simplicity. As it just adds (weighted) features, a human can easily interpret the results. This allowed us in Paper I to design detectors using manual tuning when no data was yet available for automatic tuning. Once we have accumulated enough training data, we train the system through a special case of the linear model: logistic regression [Pedregosa et al., 2011]. This algorithm assigns a probability to each instance. This not only allows us to do ranking (like we did with the manually tuned system) but also to give a precision and relative recall score, because every instance with a score above 0.5 is considered a true instance by the model. Moreover, we can adjust the probability threshold if we want to favour precision over recall or vice versa.

3.4 Corpora and Tools

All corpora are in English. In all papers we use extracts from Europarl [Koehn, 2005] as our main training and test sets. This is a corpus of political discussions commonly used in natural language processing because it is large (several million words), and is written in a consistent English generic enough to make the model applicable to other genres like novels. In Paper II and Paper IV we additionally use some other corpora, but only for test purposes; one is an anthology of Sherlock Holmes stories (Paper II) and the others are lists of titles and quotations downloaded from the Internet (Paper IV). If the experiments work on those other corpora, this will mean that the choice of Europarl was generic enough to make a robust system over several genres.

In Paper I our tool of implementation is the lemmatiser treetagger, for lemmatisation and tokenisation [Schmid, 1994]. In Papers II, III, IV we use Stanford CoreNLP [Manning et al., 2014] and the Stanford parser in order to get not only lemmatisation and tokenisation but also information about syntactic structure that is important for detection of chiasmus. The implementation of our own system is done in Python, and we use scikit-learn [Pedregosa et al., 2011] for all machine learning tasks.

3.5 Annotation

3.5.1 *What* Are We Annotating?

An important common trait of all our studies (Papers I–IV) is the selective way in which we choose to annotate our training and test data. Indeed we do not annotate all the candidates extracted by the machine. Part of the contribution of our research consists in exploring methods for saving time during annotation. For that, we explore how to preselect the annotation set during the training phase, and how to limit the evaluation set during the test phase.

Annotation of Training Data
For training data, two slightly different approaches have been explored for annotating chiasmus (Papers I–III) and annotating epanaphora/epiphora (Paper IV). This is because we had to adapt to two different types of problems. For chiasmus, the ratio of true instances to candidates (which is already very small for all the figures), is much smaller than for epanaphora and epiphora. Additionally, unlike epanaphora, chiasmus never benefited from research exploring relevant filtering features. Thus, for chiasmus we began by designing different ad hoc detectors with weights entered manually. For instance, we first assumed that stopwords were important discriminative features and started with an arbitrary large negative number (-100), then we empirically reduced this number to balance it against other important features that we gradually added after observing the type of false instances up-ranked by the machine. Each time a new set of features or weights was tried we annotated all the instances ranked above 200 by the machine. This approach is actually close to active learning, except that all the tuning is done manually. We started from a corpus with no annotation at all and thus no positive examples to train on. At each iteration our number of examples increased but not enough to train automatically. (The number of new positive examples is very small at each iteration, one or two at best.) That is why, in Papers I and II, we re-evaluate the weighting and the features manually instead of automatically, during what we call the tuning phase. Once our ranking seemed satisfactory, and once we could not rank up new positive examples, we stopped this tuning process on the training set and moved on to the test phase. In Paper III, we had finally accumulated

a sufficient number of training instances to use machine learning to weight features.

For epanaphora/epiphora, the situation was slightly better than for chiasmus, and our preselection protocol for training was thus simpler. We already had, thanks to Strommer [2011], an idea of three positive features working for epanaphora and transposable to epiphora (number of sentences equal to or greater than three, presence of "strong" punctuations (!,?), sentence length less than ten words). Thus, we directly annotated every instance that matched any of those criteria, without trying manual weighting on each of them. This gave us enough examples to use machine learning from the start.

Annotation of Test Data

In the test phase, the annotation protocol was kept the same for each figure and all papers (Papers I–IV). We apply the system on an independent two million word corpus. On this test corpus, the machine ranks the candidates for chiasmus/epanaphora/epiphora. Depending on the figure we are looking for, the output ranges from a couple of thousand candidates to a million, of which we choose to annotate only the top 200 instances given by the systems we are testing. The system that not only has the largest number of positive examples above this limit but also manages to highly rank those positive examples is then considered the best. All instances used for evaluation were annotated by two annotators, and only instances considered as True by both annotators were counted as positive instances.[1]

The decision to annotate only a partial quantity of instances in the corpus has drastically reduced the annotation time. Instead of annotating a couple of million examples of chiasmus, we only had to annotate a couple of thousand. And instead of annotating a couple of thousand candidates for epanaphora/epiphora, we annotated a couple of hundred. The drawback of this approach is that we will never really know what the recall of our machine on our corpus is, because the recall at test time is only relative to the union of the top 200 candidates of each system.

3.5.2 *How* Are We Annotating?

To the best of our knowledge, there are no existing guidelines for annotation of repetitive figures. In fact, it is an extremely controversial question because of the high valuation of ambiguity in literature described by Hammond et al. [2013]. Another reason why such guidelines have not been developed is that under each definition of repetitive figures, there is such a diversity of linguistic phenomena, effects and intensity that we are never certain if we have a full description of all the existing subtypes. Thus a case-by-case discussion, at least

[1] Double annotation was occasionally also used for training data, to check inter-annotator agreement, but most of the training data was only annotated by one person.

when the example is not prototypical, as in Examples 8, 9, 10, is more appropriate. Strommer [2011] asks the annotators if the repetition is "intentional" by checking if it produces "emphasis, clarity, amplification or emotional effect" [Strommer, 2011, p. 40]. This instruction leaves some liberty to the annotators as it is up to them to decide if they feel an emotional effect or not. We do not know what the precise criterion is that separates a rhetorical repetition from a non-rhetorical one, so we leave it to the intuition of the annotators, as Strommer [2011] did. However, we can definitely report that when discussing the annotations, some questions were repeatedly asked to help with decisions. These are:

- Does this instance have a similar structure to another example previously annotated? In particular, is this example reminiscent of an example often cited in dictionaries and stylistic handbooks?
- If we replace one of the repeated words with a synonym, do we lose an effect in the rhythm, the meaning, or the emotional impact?
- If we had to translate the sentences that contain this repetitive figure into a very different language, like Japanese, would we have to reproduce this repetition at almost any cost?

The annotation was done by the authors of Papers I–IV.[2] It is an expert annotation (as opposed to a crowdsourcing one). Both annotators have studied literature analysis but at different schools, in different languages, and at different times. It is interesting to see through this annotation whether two experts, not belonging to the same school, can agree on the interpretation of the candidate repetitions. Because multiple valid interpretations of a given repetition candidate are often possible, above all in borderline cases, each candidate considered as True or Borderline by the first and main annotator was assigned to the second annotator during the training phase.[3] This allowed us to track subjective differences in interpretation, as well as to identify repetitive figure candidates about which there was an interpretative consensus.

In this chapter we have addressed the questions of the task definition and summarised the approach common to our papers. Some details of the implementation, extraction of candidates and annotation process vary between papers, but these are explained in more detail in each article. The next chapter summarises the main outcomes of our investigations.

[2] To avoid bias toward the machine or between annotators we incorporated randomisation into the annotation process.

[3] During the test phase this does not apply, because all candidates above rank 200 were annotated by both annotators, including the ones annotated as False.

4. Summary of Empirical Results

This chapter summarises the results obtained in Papers I, II, III and IV. Our objective is not to review all the results that the reader can find in our articles, but to highlight the main trends and findings.

4.1 Exploration

One of the first simple but important results of this research is the exploration of the problems generated by the figures. From Paper I to Paper IV we illustrate by examples or annotation of random samples (Paper IV) how rare the figures are, and thus what problems we can anticipate: imbalanced corpus, difficulties of annotation. From the first paper we state how chiasmus is not just slightly infrequent, but is an extreme needle-in-the-haystack problem and thus a challenge for machine learning. We point to the real issue, which is not to extract candidates, but how in practice we can design a decent system for detection.

Another interesting finding concerns epanaphora. In Paper IV, we observe that out of 100 randomly selected epanaphora candidates, only one is a genuine example, the rest being obviously false examples and a couple of borderline cases. A few years ago, Strommer [2011] conducted a similar exploration. In order to measure statistical significance, he asked two annotators to annotate 156 randomly selected epanaphora. The interesting finding is that of those 156 examples, only 2 (i.e., $\sim$1%) were considered as True by both annotators, and the rest were a majority of false instances annotated as such by both annotators.[1] Of course, the fact that we find the exact same ratio could be a coincidence. Nevertheless, it is an interesting finding because it means that both we and this researcher agree on the rarity of the phenomenon, at least for the most prototypical cases, with a surprisingly close ratio. These exploratory annotations have been done by two independent researchers with different annotators, and different theoretical frameworks. (Strommer, unlike us, prefers to speak of "intentional" repetition.) Yet the empirical observation remains the same.

The last observation that we can make about our exploration concerns both epanaphora and epiphora. We observed in Paper IV that they are not just

[1] Although their number of what they called "debated" cases (14%) is definitely larger than our number of borderlines (3%)

mirror images of each other. Yes, theoretically they might be similar, but computationally and statistically they are not the same problem. This is due to grammar. The minimum definition of epanaphora (i.e., one word identical at the beginning of a sentence) allows the extraction of at least three times as many candidates as that for epiphora. English grammar favours the use of function words at the beginning of sentences (like pronouns or articles). This is nothing exceptional and it should not be seen as the excessive frequency of a figure of speech, but rather as standard use of the grammar.

To conclude this section, we can say that these observations, as simple as they may seem, are among the most important contributions of this thesis: we do not just propose solutions to detect figures of speech; we also help identify the most important research problems.

4.2 Inter-Annotator Agreement

As part of our investigations, we have estimated inter-annotator agreement for chiasmus, epanaphora and epiphora. We used the unweighted Cohen's Kappa coefficient to quantify agreement. Strommer [2011] has already claimed that the agreement for epanaphora is good. We confirm it ($\kappa = 0.85$ for epanaphora). The inter-annotator agreement is also good for epiphora ($\kappa = 0.88$) and chiasmus ($\kappa = 0.69$). This measurement is obtained by measuring the agreement between what our annotators think are prototypical cases (annotated as True) and are not prototypical cases (annotated as False or Borderline).

Note that the inter-annotator agreement, while still good, is definitely lower for chiasmus than for the other figures. We see two possible reasons for that. The first is practical: chiasmus was the first figure studied. We were therefore less experienced in the annotation process than for the two other figures. A possible consequence of this may be a lack of consistency in annotations. Another factor is the difficulty of the task for chiasmus. For epanaphora and epiphora the annotation question is simple: is the repetition at the beginning/end of the sentence an instance of epanaphora/epiphora? For chiasmus the question was dual, because we had to distinguish between true instances and duplicates of true instances.[2] The human and the computer have to decide not only if the extract is a good example but also if the words chosen for the repetitions are the main words of the chiasmus. As a result, annotators sometimes agree that the extract is correct but do not agree about which word the chiasmus is playing on. A concrete example of this kind of conflict is presented by Example 17 extracted from our training corpus.

[2] '**Foul** is **fair** and **fair** is **foul**' is a true instance whereas 'Foul **is fair** and **fair is** foul' is a duplicate

(17) In recent years in Africa, we have not seen a shift from **totalitarian regimes** to **democracy**, but quite the contrary – a shift from **democracy** to **totalitarian regimes**.

Most canonical examples of chiasmus display a strong hierarchy in their terms. Often the grammatical heads of the phrases are by default the main words and are all of the same nature. However, in this particular example, there is a conflict between semantic and grammatical hierarchy. On the one hand, "regimes" is the noun and thus is grammatically equivalent to "democracy". On the other hand, this chiasmus plays on the contradictions between the notions of democracy and totalitarianism, and one is tempted to prefer "totalitarian" as the main word of the chiasmus. Examples like this can influence the inter-annotator agreement.

4.3 Contributions of Different Types of Features

In this section we discuss the contribution of our thesis in terms of discovering relevant features for chiasmus, epanaphora and epiphora detection. Since our study on chiasmus is more extended, we discuss two subtypes of features: shallow and deep ones. Then we report the performance obtained with shallow features on epanaphora and epiphora.

4.3.1 Chiasmus

Shallow Features

The first empirical finding is how dramatic the improvement of the chiasmus detection can be when applying some basic features. The simple extraction of every candidate in our corpus leads to a list of one million candidates. Intuitively, we start by applying the most simple filtering features we could think of: identifying stopwords and locating punctuation marks. Thanks to this, we understand that filtering is good but not discriminative enough. Too many candidates neither involve stopwords nor are divided by major punctuation marks, and thus all of these are given the same maximum possible score by the model. As a consequence, the general precision of such a simple system is far below 2% (16 out of 1180). And, it would still take at least two hours for a trained user to read all this material before finding the real chiasmi hidden in it. This confirms how non-trivial the problem of chiasmus detection is, especially when the corpus is large (in our case, over two million words).

Three other categories of features, size-related ones (e.g., the longer the chiasmus, the lower it scores), measurement of similarity of context, and detection of recurrent lexical patterns (for instance, the presence of negation underlying a contrast) are enough to make a system useful for the user. With these features, the precision at ten candidates is 70%, which means that to find seven chiasmi the user only had to read the top ten answers given by the machine, and most of the candidates (16 out of 19) could be found among the top 200 candidates.

Deep Features
A category of features that are particularly expected in chiasmus detection are syntax features. Indeed, if we think about the top-three most well-known chiasmi, presented in Examples 18, 19, 20, the first salient common point we notice is their perfectly symmetrical switch of syntactic roles.

(18) It is not the **beginning** of the **end**; it is the **end** of the **beginning**.

(19) Ask not what your **country** can do for **you**; ask what **you** can do for your **country**.

(20) **All** for **one**, **one** for **all**.

That is why we test syntactic features (Papers II and III) and obtain an overall improvement of 14% absolute average precision over shallow features. However, the difference does not prove to be statistically significant (p>0.05).[3]

4.3.2 Epanaphora and Epiphora

As we said in Section 3.1, we start from our experience of chiasmus and apply it to other figures. Thus, the features we test for epanaphora and epiphora are similar to the basic shallow features used for chiasmus. These features are related to number of words, n-gram similarities, and n-gram differences. In the end, we observe an overall improvement of 21% for epiphora and 38% for epanaphora in average precision compared to basic features inspired by Strommer [2011]. This time, the corpus and the differences prove to be large enough, and we obtain definitely significant results (p<0.05).

One overall trend we would like to underline at this stage is that none of the experiments in Papers I–IV need very deep semantic features to build a system with reasonable performance. Despite the fact that some of the examples in our corpus seem to appeal to some sophisticated cognitive process (such as understanding a pun, a parallelism, a paradox, etc.) there was no need to model these deep semantic features to detect the figures. For instance, to detect “it is not the beginning of the end, but the end of the beginning” there is no need for a wordnet that would model the notion of opposition between “beginning” and “end”. While this finding is hardly surprising for a well-trained computational linguist, it probably sounds counter-intuitive from a literary point of view. Finally, it is an essential finding from a multilingual perspective: we might be able to develop detectors for under-resourced languages.

[3] To measure statistical significance, we apply the bootstrapping method of Berg-Kirkpatrick et al. [2012].

4.4 Machine Learning with Partially Annotated Corpora

The bottleneck issue with rhetorical figure detection is the lack of annotated data. To this problem, we must add the fact that the extraction of candidates is noisy and leads to an imbalanced corpus where the true instances are rare, or even extremely rare in the case of chiasmus. Thus we decided to annotate the corpora only partially and in a selective way. This results in only 21% of the epanaphora candidates being annotated in the training corpus, 15% in the epiphora training corpus, and not even 1% in the chiasmus training corpus. The remaining parts of the corpus were labelled as False by default without being seen by any annotator. Despite this, we obtained good results. In particular, with chiasmus, we show that the machine has comparable results to the human in the attribution of weights and in the average precision score (71% average precision for the computer against 68% for the human).

4.5 Experiments with Additional Genres

An important question thoughout our study has been to find out if our system can perform on a different type of corpus than the one it has been trained on. A first experiment was performed to detect chiasmus in an anthology of Sherlock Holmes stories. In this case, we discovered that the performance of our system is better than the baseline (+17% average precision) and the overall result is quite good (70% average precision with all features of the hand-tuned model). This score has been obtained without any particular tuning targeted at this genre. That could mean either of two things. First, it could mean that our initial choice of corpus, Europarl, was generic enough to allow our system to adapt to a different genre, in this case, novels. Secondly, it could mean that the characteristics of chiasmus, epanaphora and epiphora are not sensitive to the genre in which they appear.

The last test we performed was a real case study between three genres. We crawled web resources to extract titles of works of fiction, titles of scientific articles, and quotations. We ran our detectors for chiasmus, epanaphora and epiphora on the three corpora obtained. Despite the fact that the genres are different from Europarl, the average precision scores remained high (no result below 65% on any figure or any corpus). Finally, the main finding is that the use of figures differs between genres. First, the quotation genre contained the largest number of figures in general. This was expected, because quotations are by definition appealing and well-written parts of texts, and therefore are likely to contain special stylistic features like figures of speech. What was less expected was the apparent predominance of chiasmus in scientific titles compared to fiction titles. Additionally, we observe the reverse phenomenon for epanaphora and epiphora in fiction titles. Such observations are definitely an interesting finding for literature analysis and our knowledge of genres in general.

Last but not least, our tools prove able to facilitate the collection of enough examples to allow substantial studies of the figures themselves. Indeed, in a tentative study on chiasmus, Rabatel [2008, p. 22] explains:

> Nombre de nos exemples ont été rassemblés lors de la campagne des présidentielles. Le faible rendement de la collecte nous a contraint à intégrer d'autres exemples et à limiter l'analyse pragmatique contextuelle des antimétaboles inscrites dans un genre [...] spécifiques.[4]

If possible Rabatel [2008] initially would have limited his study of the figure to one genre. However, he could not, probably because he had no tool that could speed up the collection process. In total, he quotes 22 different chiasmi extracted from literature, political discourse and newspapers. This is close to the number we have from scientific titles (21 within the top 100 candidates output by the machine). Thus, a linguistic study on chiasmus targeting one genre was not possible before, but has now been made possible by our system. Thanks to our test on other corpora in Papers II and IV, our systems prove able to be a real help for the literary analyst. It assists with doing the most repetitive part of the work: collecting examples. That leaves more time for analysis. For more concrete insights on what this collection looks like, the reader can find all our input and output files of the genre study files in Paper IV[5] at this address: `http://stp.lingfil.uu.se/~marie/corpus/`.

In this chapter, we have reported our main results and their impact. The next chapter concludes and summarises our contributions and points to future directions of study.

[4] Many of our examples have been collected during the campaign for the presidency. The low yield of the collection has forced us to integrate other examples and to limit the pragmatic contextual analysis of antimetabole falling within a specific [...] genre.

[5] Waterstones (list of titles of fiction books: `https://www.waterstones.com`), DBLP (list of titles of scientific articles [Ley, 2002]) and quotations: `https://github.com/alvations/Quotables`

5. Conclusion

"God is dead." –Nietzsche
"Nietzsche is dead." –God

— Anonymous Graffiti

In this chapter, we summarise what we take to be the main contributions of the thesis and point to promising directions for future research.

5.1 Main Contributions

We sought to detect three figures of repetition called chiasmus, epanaphora, and epiphora. Our main question was to what extent this detection is possible for a computer, and our research questions revolved around the following notions: the definition of the task, the challenges raised by each figure, the amount of data needed to build a reliable model, the weighting techniques, the evaluation methods, and finally the categories of features to be used.

Our first contribution is to redefine the task itself. Based on empirical observations, we propose to rank the figures, instead of just classifying them as either rhetorical or not. Therefore we do not propose a system that would output just the figures *we* think are relevant. More generally, we propose a system that extracts all the repetitions and ranks them with the most unanimously relevant instances presented at the top of the list. By not sharply and arbitrarily cutting off the borderline cases, we manage to design a detector likely to be relevant for any user, regardless of individual preferences.

Our second contribution is to reveal the particular needle-in-the-haystack character of the figures and thus to identify the real problem: many candidates but very few real instances. This made the task of annotating a corpus particularly hard, and there was no previously annotated data for us to use. Our third contribution is to propose a method for annotating the training data but in a selective way; most obviously false instances are removed from annotation by manually tuned systems for chiasmus and by state of the art filters for epanaphora and epiphora. This method reduces by three orders of magnitude the annotation work for chiasmus, and one order of magnitude that for epanaphora and epiphora.

Our fourth contribution is to define an evaluation scheme and apply it to our system. Since ranked evaluation data is difficult to obtain, and since we do not have a completely annotated test corpus, we have to evaluate by measuring precision on clear cases. To nevertheless make the evaluation sensitive to the ranking, we propose to use average precision, which favours systems that rank clear cases at the top. This evaluation reveals that, even with a very incompletely annotated corpus, a system for repetitive figure detection can be trained to achieve reasonable accuracy.

Our fifth contribution is a systematic study of different linguistic features and their impact on detection. The study shows, first of all, that we can detect all three figures using relatively shallow features that do not require semantic analysis, although syntactic analysis is beneficial for chiasmus. The study also shows that different features are useful for different figures. In particular, despite the close similarity between epanaphora and epiphora, their detection requires partially different features.

Finally, we answer our main research question, as we demonstrate that our system works on four different types of text: political discourse, fiction, titles of articles and novels, and quotations. Besides showing that the system is robust to genre shifts, this exploration also reveals differences in the frequency of different figures for different genres.

5.2 Future Directions

While our system has been proven to be useful for a user, the statistical significance of each feature is not yet proven (especially for chiasmus). For that, we would need larger datasets for evaluation, and thus more annotation by more annotators.

It would be interesting to see if this system is applicable to other languages than English, and compare the usefulness of the features across languages. Another improvement would be to broaden the types of chiasmus, epanaphora and epiphora we detect. For instance, can we detect chiasmus with only semantic links, or epanaphora and epiphora that do not consist of adjacent sentences? Finally, an important goal could be to detect more than three types of repetitive patterns.

As for now, we have already released the chiasmus detector[1] and we have been contacted by other universities wishing to apply it to digital humanities projects[Ullyot, 2017]. Our epanaphora and epiphora detectors will be released as well, and it will be interesting to see the user feedback on these. An online demo[2] is available for chiasmus detection. It is an early version dating from 2015, but nevertheless it is the first time any repetitive figure detector has been made available to the general public. If we continue developing tools accessible for an audience from the humanities, comparative corpus studies

[1] `https://github.com/mardub1635/chiasmusDetector`

[2] `http://stp.lingfil.uu.se/~marie/cgi/demo.html`

like the one in Paper IV could be performed on a larger scale by different researchers. Ultimately, this would increase our insight into not only the use of these figures, but also genres, author styles and literature in general.

6. Overview of the Papers

In this chapter, we give an overview of the four papers included in the thesis. Marie Dubremetz has been mainly responsible for all four papers, and is solely responsible for all work on implementation and evaluation. Joakim Nivre has contributed to design, analysis, annotation and preparation of the final text.

Paper I

In this paper we develop our first hand-tuned system for chiasmus detection. We introduce the notion of chiasmus as a graded phenomenon as well as our method of annotation with True, False and Borderline labels. We test shallow features based on stopwords, punctuation, n-grams, and lexical clues. We show that our model does fairly well on this task and improve on pre-existing methods, both in terms of recall and precision.

Paper II

This paper builds on Paper I in order to improve detection. The system we develop reuses the same schema of annotation and the same type of model: linear regression. We test two categories of deep features: the part-of-speech tag similarity of words and syntactic role labels. These features appear to increase the performance of the detection on our political discourse corpus by nearly 25% in absolute terms compared to the previous system. By extending the annotation to two annotators we show that the inter-annotator agreement on chiasmus is good. That proves that chiasmus is not an idiosyncratically defined notion, at least on the prototypical level.

Paper III

This paper uses data collected in our first two studies with hand-tuned models to explore a machine learning approach to tuning feature weights. We then compare the weights given by the machine and the weights given by the human. We discover two things. A partially annotated corpus (less than 1% manually annotated) can be enough for training a system as effective as the hand-tuned one, and computers and humans generally agree on which features are positive or negative.

Paper IV

This journal article starts from the machine learning approach developed in Paper III and extends it in two ways: detecting new figures and exploring new text genres. We apply our machine learning approach not only to chiasmus but also to epanaphora and epiphora. We comparatively explore the three figures together on the same corpus and discover that, at extraction time, they display huge differences in number of candidates. Nevertheless, all of them suffer from a common problem: the overwhelming number of false instances generated by extraction. Therefore, we selectively annotate candidate epanaphora and epiphora using rule-based filters. After training the system on those annotated instances, we can outperform a model using only the basic features by 20 percentage points on almost all evaluation metrics. We discover that, despite their theoretical proximity, epanaphora and epiphora do not react in the same way to the same features. Thus, they are not just mirror images of each other. Finally, we illustrate how our detectors can be used to explore texts from three different genres. Thanks to the automatic detection of rhetorical figures, we discover that chiasmus is more likely to appear in scientific contexts, whereas epanaphora and epiphora are more common in fiction.

References

Assaf, Dan, Neuman, Yair, Cohen, Yohai, Argamon, Shlomo, Howard, Newton, Last, Mark, Frieder, Ophir, and Koppel, Moshe. Why 'dark thoughts' aren't really dark: A novel algorithm for metaphor identification. In *Proceedings of the 2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain, SSCI 2013*, pages 60–65, 2013.

Bacry, Patrick. *Les Figures de style : Et autres procédés stylistiques*. Belin, 1992.

Barthes, Roland. *Le bruissement de la langue (Essais critiques IV)*. Seuil, 1984.

Bendersky, Michael and Smith, David. A Dictionary of Wisdom and Wit: Learning to Extract Quotable Phrases. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 69–77, 2012.

Berg-Kirkpatrick, Taylor, Burkett, David, and Klein, Dan. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL*, pages 995–1005, 2012.

Booten, Kyle and Hearst, Marti A. Patterns of Wisdom: Discourse-Level Style in Multi-Sentence Quotations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1139–1144, 2016.

Bracewell, David B, Tomlinson, Marc T, and Mohler, Michael. Determining the Conceptual Space of Metaphoric Expressions. In Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, pages 487–500. Springer, Berlin, Heidelberg, 2013.

Brooke, Julian, Hammond, Adam, and Hirst, Graeme. Distinguishing Voices in The Waste Land using Computational Stylistics. *Linguistic Issues in Language Technology (LiLT), Special Issue on Computational Linguistics for Literature*, 12 (2):1–41, 2015.

Burton, Gideon O. The Forest of Rhetoric: silva rhetoricae. `http://rhetoric.byu.edu/`, 1996. Accessed: 2017-11-07.

Butler, Harold Edgeworth. *The Institutio oratoria of Quintilian*. Harvard University Press, 1921.

Carbonell, Jaime Guillermo. Metaphor: a key to extensible semantic analysis. In *Proceedings of the 18th annual meeting of the Association for Computational Linguistics*, pages 17–21, 1980.

Coulson, Seana and Matlock, Teenie. Metaphor and the Space Structuring Model. *Metaphor & Symbol*, 16(3/4):295–316, 2001.

Danescu-Niculescu-Mizil, Cristian, Cheng, Justin, Kleinberg, Jon, and Lee, Lillian. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, 2012.

Deignan, Alice. *Metaphor and Corpus Linguistics*. Converging evidence in language and communication research. John Benjamins Publishing Company, 2005.

Derrida, Jacques. *Margins of Philosophy*. 1982.

Diderot, Denis and D'Alembert, Jean le Rond. *Encyclopédie méthodique: ou par ordre de matières, volume 66*. Panckoucke, 1782.

Dubremetz, Marie. Détecter les chiasmes dont les termes principaux entretiennent une forte proximité morphologique ou sémantique. Master's thesis, 2012.

Dubremetz, Marie. Vers une identification automatique du chiasme de mots. In *Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)*, pages 150–163, 2013.

Dubremetz, Marie and Nivre, Joakim. Rhetorical Figure Detection: Chiasmus, Epanaphora, Epiphora. Submitted.

Dubremetz, Marie and Nivre, Joakim. Rhetorical Figure Detection: The Case of Chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 23–31, 2015.

Dubremetz, Marie and Nivre, Joakim. Syntax Matters for Rhetorical Structure: The Case of Chiasmus. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 47–53, 2016.

Dubremetz, Marie and Nivre, Joakim. Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 37–45, 2017.

Dunn, Jonathan. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, 2013a.

Dunn, Jonathan. How linguistic structure influences and helps to predict metaphoric meaning. *Cognitive Linguistics*, 24(1):33–66, 2013b.

Dunn, Jonathan. Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 745–751, 2014.

Dupriez, Bernard. *Gradus, les procédés littéraires*. Union Générale d'Éditions 10/18, 2003.

Evans, Vyvyan. Figurative language understanding in LCCM theory. *Cognitive Linguistics*, 21(4):601–662, 2010.

Fahnestock, Jeanne. *Rhetorical Figures in Science*. Oxford University Press, 1999.

Fontanier, Pierre. *Les Figures du discours*. Flammarion, 1977 edition, 1827.

Gandy, Lisa, Allan, Nadji, Atallah, Mark, Frieder, Ophir, Howard, Newton, Kanareykin, Sergey, Koppel, Moshe, Last, Mark, Neuman, Yair, and Argamon, Shlomo. Automatic Identification of Conceptual Metaphors with Limited

Knowledge. In *AAAI Conference on Artificial Intelligence*, pages 328–334, 2013.

García-Page, Mario. El “retruécano léxico" y sus límites. *Archivum: Revista de la Facultad de Filología de Oviedo*, 41-42:173–203, 1991.

Gawryjolek, Jakub J. Automated Annotation and Visualization of Rhetorical Figures. Master thesis, Universty of Waterloo, 2009.

Gibbs, Raymond W. A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, 34(4):457–486, 2002.

Greene, Roland, Cushman, Stephen, Cavanagh, Clare, Ramazani, Jahan, and Rouzer, Paul, editors. *The Princeton Encyclopedia of Poetry and Poetics: Fourth Edition*. Princeton University Press, 2012.

Hammond, Adam, Brooke, Julian, and Hirst, Graeme. A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8, 2013.

Harris, Randy and DiMarco, Chrysanne. Constructing a Rhetorical Figuration Ontology. In *Persuasive Technology and Digital Behaviour Intervention Symposium*, pages 47–52, 2009.

Horvei, Harald. *The Changing Fortunes of a Rhetorical Term: The History of the Chiasmus*. The Author, 1985.

Howard, Gregory T. *Dictionary of Rhetorical Terms*. Xlibris Corporation, 2010.

Hromada, Daniel Devatman. Initial Experiments with Multilingual Extraction of Rhetoric Figures by means of PERL-compatible Regular Expressions. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 85–90, 2011.

Jakobson, Roman. Two aspects of language and two types of aphasic disturbances. In *Fundamentals of language*, pages 115–133. The Hague & Paris: Mouton, 1956.

Kao, Justine and Jurafsky, Dan. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, 2012.

Kiddon, Chloé and Brun, Yuriy. That’s What She Said: Double Entendre Identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 89–94, 2011.

Koehn, Philipp. Europarl: A Parallel Corpus for Statistical Machine Translation. In *The Tenth Machine Translation Summit*, pages 79–86, 2005.

Lakoff, George and Johnson, Mark. *Metaphors we Live by*. University of Chicago Press, Chicago, 1980.

Ley, Michael. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *String processing and information retrieval*, pages 481–486. 2002.

Li, Linlin and Sporleder, Caroline. Linguistic Cues for Distinguishing Literal and Non-Literal Usage. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING’10)*, pages 683–691, 2010a.

Li, Linlin and Sporleder, Caroline. Using Gaussian Mixture Models to Detect Figurative Language in Context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, 2010b.

Mann, William C. and Thompson, Sandra A. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Manning, Christopher D, Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J, and McClosky, David. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

Mason, Zachary J. CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30:23–44, 2004.

McCurdy, Nina, Srikumar, Vivek, and Meyer, Miriah. RhymeDesign: A Tool for Analyzing Sonic Devices in Poetry. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 12–22, 2015.

Morier, Henri. *Dictionnaire de poétique et de rhétorique*. Presses Universitaires de France, 1961.

Morris, Jane and Hirst, Graeme. The Subjectivity of Lexical Cohesion in Text. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications*, pages 3–6, 2004.

Neuman, Yair and Nave, Ophir. Metaphor-based meaning excavation. *Information Sciences*, 179(16):2719–2728, 2009.

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

Nordahl, Helge. Variantes chiasmiques. Essai de description formelle. *Revue Romane*, 6:219–232, 1971.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, and Duchesnay, Édouard. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Poe, Edgard Allan. *The Raven*. Botolph classics. R. G. Badger & Company, 1898.

Pougeoise, Michel. *Dictionnaire de Rhétorique*. Armand Colin, 2001.

Rabatel, Alain. Points de vue en confrontation dans les antimétaboles PLUS et MOINS. *Langue française*, 160(4):21–36, 2008.

Roberts, Rhys W. *Rhetoric*. Dover thrift editions. Dover Publications, 2004.

Ruiz, Pablo, Martínez Cantón, Clara, Poibeau, Thierry, and González-Blanco, Elena. Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32, 2017.

Schmid, Helmut. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, 1994.

Shutova, Ekaterina and Sun, Lin. Unsupervised Metaphor Identification Using Hierarchical Graph Factorization Clustering. In *Proceedings of Annual the Conference of the North American Chapter of the Association for Computational Linguistics*, number June, pages 978–988, 2013.

Shutova, Ekaterina, Teufel, Simone, and Korhonen, Anna. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353, 2013.

Steen, Gerard J., Dorst, Aletta G., Herrmann, J. Berenike, Kaal, Anna A., and Krennmayr, Tina. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796, 2010.

Strommer, Claus Walter. *Using Rhetorical Figures and Shallow Attributes as a Metric of Intent in Text*. PhD thesis, University of Waterloo, 2011.

Suhamy, Henri. *Les figures de style*. Presses universitaires de France. Paris, 2004.

Tsvetkov, Yulia, Mukomel, Elena, Gershman, Anatole, and Gershman, Ytema. Cross-Lingual Metaphor Detection Using Common Semantic Features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, 2013.

Turney, Peter D, Neuman, Yair, Assaf, Dan, and Cohen, Yohai. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690, 2011.

Ullyot, Michael. Get with the programming. `http://ullyot.ucalgaryblogs.ca/2017/10/16/programming/`, 2017. Accessed: 2017-11-07.

Van Gorp, Hendrik, Delabastita, Dirk, and D'Hulst, Lieven. *Dictionnaire des termes littéraires*. Honoré Champion, 2001.

Vandendorpe, Christian. Lecture et quête de sens. *Protée*, 19:95–101, 1991.

Viklund, Jon. Chiasmus as an Argumentative Figure in C.J.L. Almqvist's "The Idea of History" (1819). In *Ten Nordic Studies in the History of Rhetoric*, number 1 in Nordic Studies in the History of Rhetoric. Nordisk netværk for retorikkens historie, 2002.

Watson, Helen Gertrude. *A study of the rhetorical figures of anaphora, chiasmus, and alliteration in Vergil's Aeneid, books I–VI*. PhD thesis, University of Illinois, 1912.

Wilks, Yorick. Making preferences more active. *Artificial Intelligence*, 11(3): 197–223, 1978.

Wooten, Cecil W. *Hermogenes' On Types of Style*. University of North Carolina Press, 2012.

Zhang, Ethan and Zhang, Yi. Average Precision. In Liu, Ling and Özsu, M Tamer, editors, *Encyclopedia of Database Systems*, pages 192–193. Springer US, Boston, MA, 2009.