# Evaluation in Discourse: a Corpus-Based Study

Farah Benamara, Nicholas Asher, Yvette Yannick Mathieu, Vladimir Popescu, Baptiste Chardon

Dialogue and Discourse 7(1) (2016) 1-49 doi: 10.5087/dad.2016.101

# Evaluation in Discourse: a Corpus-Based Study

**Farah Benamara** BENAMARA@IRIT.FR
*IRIT, Université de Toulouse. 118 route de Narbonne, 31062, Toulouse, France.*

**Nicholas Asher** ASHER@IRIT.FR
*IRIT-CNRS. 118 route de Narbonne, 31062, Toulouse, France.*

**Yvette Yannick Mathieu** YMATHIEU@LINGUIST.UNIV-PARIS-DIDEROT.FR
*Bat Olympes de Gouges, 8 place Paul Ricoeur, Case courrier 7031, 75205 Paris Cedex 13, France.*

**Vladimir Popescu** POPESCU@IRIT.FR
*IRIT, Université de Toulouse. 118 route de Narbonne, 31062, Toulouse, France.*

**Baptiste Chardon** BAPTISTE.CHARDON@SYNAPSE-FR.COM
*Synapse Développement. 33 Rue Maynard, 31100 Toulouse.*

**Editor:** Massimo Poesio

## Abstract

This paper describes the CASOAR corpus, the first manually annotated corpus exploring the impact of discourse structure on sentiment analysis with a study of movie reviews in French and in English as well as letters to the editor in French. While annotating opinions at the expression, sentence, or document level is a well-established task and relatively straightforward, discourse annotation remains difficult, especially for non experts. Therefore, combining opinion and discourse annotations pose several methodological problems that we address here. We propose a multi-layered annotation scheme that includes: the complete discourse structure according to the Segmented Discourse Representation Theory, the opinion orientation of elementary discourse units and opinion expressions, and their associated features (including polarity, strength, etc.). We detail each layer, explore the interactions between them, and discuss our results. In particular, we examine the correlation between discourse and semantic category of opinion expressions, the impact of discourse relations on both subjectivity and polarity analysis, and the impact of discourse on the determination of the overall opinion of a document. Our results demonstrate that discourse is an important cue for sentiment analysis, at least for the corpus genres we have studied.

## 1. Introduction

Sentiment analysis has been one of the most popular applications of natural language processing for over a decade both in academic research institutions and in industry. In this domain, researchers analyze how people express their sentiments, opinions and points of view from natural language data such as customer reviews, blogs, fora and newspapers. Opinions concern evaluations expressed by a holder (a speaker or a writer) towards a topic (an object or a person). An evaluation is characterized by a polarity (positive, negative or neutral) and a strength that indicates the opinion degree of positivity or negativity. Example (1), extracted from our corpus of movie reviews, illustrates these phenomena[1]. In this review, the author expresses three opinions: the first two are explicitly lexical-

1. This example has been extracted from MetaCritic website as it is, including typos and English errors.

 Submitted 04/14; Accepted 11/15; Published online 01/16

ized opinion expressions (underlined in the example) whereas the last one (in italic) is an implicit positive opinion since it contains no subjective lexical cues.

(1) What a great animated movie. I was so thrilled by seeing it that *I didn't movie a single second from my seat*.

From a computational perspective, most current research examine the expression and extraction of opinion at two main levels of granularity: the document and the sentence[2]. At the document level, the standard task is to categorize documents globally as being positive or negative towards a given topic (Turney (2002); Pang et al. (2002); Mullen and Nigel (2004); Blitzer et al. (2007)). In this classification problem, all opinions in a document are supposed to be related to only one topic[3]. Overall document opinion is generally computed on the basis of aggregation functions (such as the average or the majority) that take as input the set of explicit opinions scores of a document and output either a polarity rating or an overall multi-scale rating (Pang and Lee (2005); Lizhen et al. (2010); Leung et al. (2011)). At the sentence level, on the other hand, the task is to determine the subjective orientation and then opinion orientation of sequences of words in the sentence that are determined to be subjective or express an opinion (Yu and Vasileios (2003); Riloff et al. (2003); Wiebe and Riloff (2005); Taboada et al. (2011)). This second level also assumes that a sentence usually contains a single opinion. To better compute the contextual polarity of opinion expressions, some researchers have used subjectivity word sense disambiguation to identify whether a given word has a subjective or an objective sense (Akkaya et al. (2009)). Other approaches identify valence shifters (viz. negations, modalities and intensifiers) that strengthen, weaken or reverse the prior polarity of a word or an expression (Polanyi and Zaenen (2006); Shaikh et al. (2007); Moilanen and Pulman (2007); Choi and Cardie (2008)). The contextual polarity of individual expressions is then used for sentence as well as document classification (Kennedy and Inkpen (2006); Li et al. (2010)).

We believe that viewing opinions in a text as a simple aggregation of opinion expressions identified *locally* is not appropriate. In this paper, we argue that discourse structure provides a crucial link between local and document levels and is needed for a better understanding of the opinions expressed in texts. To illustrate this assumption, let us take the example (2), extracted from our corpus of French movie reviews. (2) contains four opinions: the first three are strongly negative while the last one (introduced by the conjunction *but* in the last sentence) is positive. A bag of words approach would classify this review as negative, which is contrary to intuitions for this example.

(2) Les personnages sont antipathiques au possible. Le scénario est complètement absurde. Le décor est visiblement en carton-pâte. Mais c'est tous ces éléments qui font le charme improbable de cette série.
*The characters are unpleasant. The scenario is totally absurd. The decoration seems to be made of cardboard. But, all these elements make the charm of this TV series.*

Discourse structure can be a good indicator of the subjectivity and/or the polarity orientation of a sentence. In particular, general types of discourse relations that link clauses together like Parallel, Contrast, Result and so on from theories like Rhetorical Structure Theory (RST) (Mann and

2. There is also a third level of granularity not detailed here which is the aspect or feature level where opinions are extracted according to the target domain features (Liu (2012)).
3. Of course, this assumption is debatable. For instance in forums, blogs and news, opinions are related to several topics.

Thompson (1988)) or Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides (2003)) furnish important clues for recognizing implicit opinions and assessing the overall stance of texts. For instance[4], sentences related by the discourse relations Parallel or Continuation often share the same subjective orientation like in *Mary liked the movie. Her husband too*. Here, Parallel (triggered by the discourse marker *too*) holds between the two sentences and allows us to detect the implicit opinion conveyed by the second sentence. Polarity is often reversed in case of Contrast and usually preserved in case of Parallel and Continuation. Result on the other hand does not have a strong effect on subjectivity and polarity is not always preserved. For instance, in *Your life is miserable. You don't have a girlfriend. So, go see this movie*, the positive polarity of the recommendation follows the negative opinions expressed in the first two sentences. In case of Elaboration, subjectivity may not be preserved, in contrast to polarity (it would be difficult to say *The movie was excellent. The actors were bad*). Finally, Attribution plays a role only when its second argument is subjective, as in *I suppose that the employment policy will be a disaster*. In this case, depending on the reported speech act used to introduce the opinion, Attribution affects the degree of commitment of the author and the holder (Asher (1993); Prasad et al. (2006)).

Discourse-based opinion analysis is an emerging research area (Asher et al. (2008); Taboada et al. (2008, 2009); Somasundaran (2010); Zhou et al. (2011); Heerschop et al. (2011); Zirn et al. (2011); Polanyi and van den Berg (2011); Trnavac and Taboada (2010); Mukherjee and Bhattacharyya (2012); Lazaridou et al. (2013); Trivedi and Eisenstein (2013); Wang and Wu (2013); Hogenboom et al. (2015); Bhatia et al. (2015)). Studying opinion within discourse gives rise to new challenges: *What is the role of discourse relations in subjectivity analysis? What is the impact of the discourse structure in determining the overall opinion conveyed by a document? Does a discourse based approach really bring additional value compared to a classical bag of words approach? Does this additional value depend on corpus genre?* The CASOAR project (a two year DGA-RAPID project (2010-2012) involving Toulouse University and an NLP company Synapse Développement) aimed to address these questions by gathering and analyzing a corpus of movie reviews in French and in English as well as letters to the editor in French. It extended our earlier work where Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides (2003)) was used to study opinion within discourse (Asher et al. (2008, 2009)).

Before moving to real scenarios that rely on automatic discourse annotations, we first wanted to measure the impact of discourse structure on opinion analysis in manually annotated data. While annotating opinions at the expression, sentence or document level achieved a relatively good inter-annotator agreements, at least for explicit opinion recognition, and opinion polarity (Wiebe et al. (2005); Toprak et al. (2010)), annotation of complete discourse structure is a more difficult task, especially for non experts (Carlson et al. (2003); Afantenos et al. (2012)). Combining opinion and discourse annotations poses several methodological problems: the choice of the corpus in terms of genre and document length, the definition of the annotation model, and the description of the annotation guide so as to minimize errors, etc. A second point was more challenging: what is the most appropriate level to annotate opinion in discourse? Should we annotate opinion texts using a small set of discourse relations? Or should we use a larger set? Should discourse annotations annotators be simply asked to follow their intuitions after having been given a gloss of the discourse relations to be used, or should we provide them with a precise description of the structural constraints regarding the underlying discourse theory?

4. In this paragraph, assertions are based on our own observations of the data. They have however been empirically validated in this corpus study, as shown later in this paper.

We developed a multi-layered annotation scheme that includes: the complete discourse structure according to SDRT, opinion orientation of elementary discourse units and opinion expressions, and their associated features. In this paper, we detail each layer, explore the interactions between them and discuss our results. In particular, we examine: the correlation between discourse and semantic category of opinion expressions focusing on the role of evaluation to identify discourse relations, the impact of discourse relations on both subjectivity and polarity analysis, and the impact of discourse on the determination of the overall opinion of a document. Our results demonstrate that discourse is an important cue for sentiment analysis, at least for the corpus genres we have studied.

The paper is organized as follows. Section 2 gives some background on annotating sentiment and discourse, and provides a brief introduction to SDRT, our theoretical framework. Section 3 presents our corpus. Section 4 details the annotation scheme, annotation campaign, and reliability of the scheme. Section 5 gives our results. We end the paper by a discussion where we highlight the main conclusions of our corpus-based study and discusses the portability and applicability of the annotation scheme.

## 2. Background

### 2.1 Existing corpora annotated with sentiment

There are several existing annotated resources for sentiment analysis. Each resource can be characterized in terms of the corpus used, the basic annotation unit and annotation levels. In this section, we overview main existing resources according to these three criteria.

#### 2.1.1 Data

Several authors have focused on annotating a single corpus genre like movie reviews (Pang and Lee (2004)), book reviews (Read and Carroll (2012)), news (Wiebe and Riloff (2005)), political debates (Somasundaran et al. (2007); Somasundaran and Wiebe (2010)) and blogs (Liu et al. (2009)). Well known resources include MPQA (Wiebe et al. (2005)), the JDPA-corpus (Kessler et al. (2010)) and the Darmstadt-corpus (Toprak et al. (2010)). Multi-domain sentiment analysis has been explored in Blitzer et al. (2007) with a corpus of product reviews taken from Amazon.com[5].

Compared to English, few resources have been developed for other languages. In French, the Blogoscopy corpus (Daille et al. (2011)) is composed of 200 annotated posts and 612 associated comments. There is also Bestgen et al. (2004)'s dataset composed of 702 sentences extracted from a newspaper[6]. In Spanish, the TASS corpus[7] is composed of 70,000 tweets annotated with global polarity as well as an indication of the level of agreement or disagreement of the expressed sentiment within the content. In German, the MLSA[8] (Clematide et al. (2012)), is a publicly available corpus composed of 270 sentences manually annotated for objectivity and subjectivity. Finally for Italian, the Senti-TUT corpus[9] includes sentiment annotations of irony in tweets (Bosco et al. (2013)). Multilingual sentiment annotation has also been explored: the EmotiBlog corpus consists of labeled blog posts in Spanish, Italian and English (Boldrini et al. (2012)), Mihalcea et al. (2007) manually

5. http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
6. https://sites.google.com/site/byresearchoa/home/
7. http://www.daedalus.es/TASS2013/about.php
8. http://iggsa.sentimental.li/index.php/downloads/
9. http://www.di.unito.it/tutreeb/sentiTUT.html

annotated 500 sentences in English, Romanian, and Spanish[10]. Finally, Banea et al. (2010) automatically annotated English, Arabic, French, German, Romanian, and Spanish news documents.

In this paper, we aim to annotate opinion in discourse in multi-genre documents (movie reviews and news reactions) in French and movie reviews in English. To our knowledge, no one has conducted a corpus-based study across genres and languages that analyzes how opinion and discourse interact at different levels of granularity (expression, discourse unit and the whole document). Thus, there is almost no extent work for us to compare ourselves to other. Even though several annotation schemes already exist for the expression/phrase level (MPQA, JDPA-corpus, Darmstadt-corpus, MLSA), the descriptive analysis investigating the interaction between sentiment and discourse is novel.

#### 2.1.2 Basic annotation unit

State-of-the art opinion annotation campaigns take the expression (a set of tokens), sentence or document as their basic annotation unit. However, annotating opinion in discourse required to move to start with elementary discourse unit (EDU) which is the intermediate level between the sentence and the document. Indeed, the sentence level is not appropriate for analyzing opinions in discourse, since, in addition to objective clauses, a single sentence may contain several opinion clauses that can be connected by rhetorical relations. Moving to the clause level is also not appropriate, since several opinion expressions can be discursively related as in *The movie is great but too long* where we have a Contrast relation introduced by the marker *but*. Therefore, we need to move to a fine-grained and semantically motivated level, the EDU.

Annotating EDUs not quite corresponding to either sentences or clauses has been standard in discourse annotation efforts for many years (see Section 2.2 for an overview). However, annotating sentiment within EDUs is still marginal. Among the few annotated sentiment corpora at the EDU level, we cite Asher et al. (2009), who analyzed explicit opinion expressions within EDUs. Somasundaran et al. (2007) used a similar level in order to detect the presence of sentiment and arguing in dialogues. Zirn et al. (2011) performed subjectivity analysis at the segment level. They used a corpus of product reviews segmented using the HILDA tool[11], an RST discourse parser. Lazaridou et al. (2013) used the SLSeg software package[12] to segment their corpus into EDUs following RST. The corpus was then used to train a joint model for unsupervised induction of sentiment, aspect and discourse information.

In this paper, documents are segmented according to SDRT principles.

#### 2.1.3 Annotation levels

Our annotation scheme is multi-layered and includes: the complete discourse structure, segment opinion orientation, and opinion expressions.

At the document level, we propose to annotate the document overall opinion as well as its full discourse structure following the SDRT framework. Global opinion annotation resembles previous document level annotation (Pang and Lee (2004)). To the best of our knowledge, this is the first sentiment dataset that incorporates discourse structure annotation.

10. http://www.cse.unt.edu/~rada/downloads.html#msa
11. http://nlp.prendingerlab.net/hilda
12. http://www.sfu.ca/~mtaboada/research/SLSeg.html

At the segment level, we propose to associate to each EDU a subjectivity type (among four main types: explicit evaluative, subjective non evaluative, implicit, and objective) as well as polarity and strength. Segment opinion type mainly follows Wiebe et al. (2005); Toprak et al. (2010) and Liu (2012). Wiebe et al. (2005) already proposed an expression-level annotation scheme that distinguishes between explicit mentions of private states, speech events expressing private states, and expressive subjective elements. Toprak et al. (2010), following Wiebe et al. (2005), distinguished in their annotation scheme (consumer reviews) between explicit opinions and facts that imply opinions. Finally, Liu (2012) has also observed that subjective sentences and opinionated sentences (which are objective or subjective sentences that express implicit positive or negative opinions) are not the same, even though opinionated sentences are often a subset of subjective sentences. In this work, we propose, in addition, to study what are the correlations between segment opinion types and the overall opinion on the one hand (cf. Section 5.3.4), and between segment types and rhetorical relations (cf. Sections 5.3.2 and 5.3.3).

The opinion expression is the lowest level and focuses on annotating all the elements associated to an opinion within a segment: (1) the opinion span, excluding operators (negation, modality, intensifier, and restrictor), (2) opinion polarity and strength, (3) opinion semantic category, (4) topic span, (5) holder span, and (6) operator span. Our annotation at this level is very similar to state of the art annotation schema at the expression level (e.g. MPQA, JDPAcorpus, Darmstadt-corpus, MLSA corpus). However, in addition, we explore the link between discourse and opinion semantic category of subjective segments (cf. Section 5.3.1).

### 2.2 Existing corpora annotated with discourse

The annotation of discourse relations in language can be broadly characterized as falling under two main approaches: the *lexically grounded approach* and an approach that aims at *complete discourse coverage*. Perhaps the best example of the first approach is the Penn Discourse Treebank (Prasad et al. (2008)). The annotation starts with specific lexical items, most of them conjunctions, and includes two arguments for each conjunction. This leads to partial discourse coverage, as there is no guarantee that the entire text is annotated, since parts of the text not related through a conjunction are excluded. On the positive side, such annotations tend to be reliable. PDTB-style annotations have been carried out in a variety of languages (Arabic, Chinese, Czech, Danish, Dutch, French, Hindi and Turkish).

Complete discourse coverage requires annotation of the entire text, with most, if not all, of the propositions in the text integrated in a structure. It includes work from two main different theoretical perspectives, either intentionally or semantically driven. The first perspective has been investigated within Rhetorical Structure theory, RST (Mann and Thompson (1988)), whereas the second includes Segmented Discourse Representation Theory, SDRT (Asher and Lascarides (2003)), and the Discourse Graph Bank model (Wolf and Gibson (2006)). RST annotated resources exist in Basque, Dutch, German, English, Portuguese and Spanish. Corpora following SDRT exist in Arabic, French and English.

To get a complete structure for a text, three decisions need to be made:

- what are the elementary discourse units (EDU)?
- how do elementary units combine to form larger units and attach to other units?
- how are the links between discourse units labelled with discourse relations?

Many theories such as RST take full sentences or at least tensed clauses as the mark of an EDU. SDRT, as developed in (Asher and Lascarides (2003)), was largely mute on the subject of EDU segmentation, but in general also followed this policy. Concerning attachment, most discourse theories define hierarchical structures by constructing complex segments (CDUs) from EDUs in recursive fashion. RST proposes a tree-based representation, with relations between adjacent segments, and emphasizes a differential status for discourse components (the nucleus vs. satellite distinction). Captured in a graph-based representation, with long-distance attachments, SDRT proposes relations between abstract objects using a relatively small set of relations. Identifying these relations is a crucial step in discourse analysis. Given two discourse units that are deemed to be related, this step labels the attachment between the two discourse units with discourse relations such as Elaboration, Explanation, Conditional, etc. For example in [*This is the best book*]$_1$ [*that I have read in along time.*]$_2$ we have *Elaboration*$(1, 2)$. Their triggering conditions rely on the propositional contents of the clauses - a proposition, a fact, an event, a situation –the so-called abstract objects (Asher (1993)) or on the speech acts expressed in one unit and the semantic content of another unit that performs it. Some instances of these relations are explicitly marked i.e., they have cues that help identifying them such as *but, although, as a consequence*. Others are implicit i.e., they do not have clear indicators, as in *I didn't go to the beach. It was raining*. In this last example to infer the intuitive Explanation relation between the clauses, we need detailed lexical knowledge and probably domain knowledge as well.

In this paper, we aim to annotate the full discourse structure of opinion documents following a semantically driven approach, as done in SDRT.

### 2.3 Overview of the Segmented Discourse Representation Theory (SDRT)

SDRT is a theory of discourse interpretation that extends Kamp's Discourse Representation Theory (DRT) (Kamp and Reyle (1993)) to represent the rhetorical relations holding between EDUs, which are mainly clauses, and also between larger units recursively built up from EDUs and the relations connecting them. SDRT aims at building a complete discourse structure for a text or a dialogue, in which every constituent is linked to some other constituent. We detail below the three steps needed to build this structure, namely: EDU determination, attachment, and relation labelling.

#### 2.3.1 EDU determination

We follow the principles defined in the Annodis project[13] (Afantenos et al. (2012)). In Annodis, an EDU is mainly a sentence or a clause in a complex sentence that typically corresponds to verbal clauses, as in [*I loved this movie*]$_a$ [*because the actors were great*]$_b$ where the clause introduced by the marker *because*, indicates a cutting point. We have here the relation Explanation$(a, b)$. An EDU can also correspond to other syntactic units describing eventualities, such as prepositional and noun phrases, as in [*After several minutes,*]$_a$ [*we found the keys on the table*]$_b$ where we have two EDUs related by Frame$(a, b)$. In addition, a detailed examination of the semantic behavior of appositives, non restrictive relative clauses and other parenthetical material in our corpora, revealed that such syntactic structures also contributed EDUs[14]. Such constructions provide semantic contents that do

13. This project aimed at building a diversified corpus of written French texts enriched with a manual annotation of discourse structures. The resource can be downloaded here `http://w3.erss.univ-tlse2.fr/Annodis`
14. In RST, embedding is handled by the "same unit" relation. To a much more limited extent, PDTB also allows for nominalizations to be arguments to relations.

not fall within the scope of discourse relations or operators between the constituents in which they occur. In Example (3), we see that the apposition in italic font does not or at least needs not fall within the scope of the conditional relation on a defensible interpretation of the text. Such "nested" EDUs are a useful feature in sentiment analysis as EDUs conveying opinions may be isolated from surrounding "objective" material, as in the movie review in (4). Finally, concerning attributions, we segment cases like "I say that I am happy" into two EDUs: "I say" and "that I am happy".

(3) If the former President of the United States, *who has been all but absent from political discussions since the 2008 election*, were to weigh in on the costs of the economic shutdown, the radical Republicans might be persuaded to vote to lift the debt ceiling.

(4) [The film [*that distressed me the most*] is CRY OF FREEDOM].

In addition to this definition, we observe in our corpora that several opinion expressions (often conjoined NP or AP clauses) could be linked by discourse relations. We thus resegment such EDUs into separate units. Annodis segmentation principles were then refined in order to take into account the particularities of opinion texts. For example, the following sentence: [*the movie is long, boring but amazing*] is segmented as follows: [*the movie is long,*]$_1$ [*boring*]$_2$ [*but amazing*]$_3$ with Continuation(1,2) and Contrast([1,2],3), [1,2] being a complex discourse unit. Even if segments 2 and 3 do not follow the EDU standard definition (they are neither sentences nor clauses), we believe that such fine-grained segmentation will facilitate polarity analysis at the sentence level.

During the annotation of EDUs, we consider that argument naming generally follows the linear order in the text. In case of embedding, the main clause is annotated first. For instance in (4), we have: [The film [*that distressed me the most*]$_2$ is CRY OF FREEDOM]$_1$.

### 2.3.2 Attachment decision

In SDRT, a discourse representation for a text $T$ is a structure in which every EDU of $T$ is linked to some (other) discourse unit, where discourse units include EDUs of $T$ and complex discourse units (CDUs) built up from EDUs of $T$ connected by discourse relations in recursive fashion. Proper SDRSs form a rooted acyclic graph with two sorts of edges—edges labeled by discourse relations that serve to indicate rhetorical functions of discourse units, and unlabeled edges that show which constituents are elements of larger CDUs. SDRT allows attachment between non adjacent discourse units and for multiple attachments to a given discourse unit[15], which means that the structures created are not always trees but rather directed acyclic graphs. These graphs are constrained by the right frontier principle that postulates that each new EDU should attach either to the last discourse unit or to one that is super-ordinate to it via a series of subordinate relations and complex segments.

One of the most important feature that makes SDRT an attractive choice for studying the effects of discourse structure on opinion analysis is the scope of relations. For instance, if an opinion is within the scope of an attribution that spans several EDUs, then knowing the scope of the attribution will enable us to determine who is in fact expressing the opinion. Similarly, if there is a contrast that has scope over several EDUs in its left argument, this can be important to determine the overall

15. In SDRT, several discourse relations can hold between two constituents if they are of the same type, i.e., either all coordinating or all subordinating and their semantics effect are compatible. So semantics puts important constraints on relations. Consider the example: [John kissed Mary.]$_1$ [She then slapped him]$_2$ [and his wife did too, at the same time.]$_3$. In this example, segment 1 and 2 are related by both a Narration and a Result. We have thus the following annotation: $\mathit{Narration}(1,[2,3]) \wedge \mathit{Result}(1,[2,3]) \wedge \mathit{Parallel}(2,3)$, [2,3] being a CDU.

contribution of the opinions expressed in the arguments of the contrast. To get this kind of information, we need to have discourse annotations in which the scopes of discourse relations are clear and determined for an entire discourse graph. Example (5) taken from the Annodis corpus (Afantenos et al. (2012)) illustrates what are called *long distance attachments*[16].

(5) [Suzanne Sequin passed away Saturday at the communal hospital of Bar-le-Duc,]$_3$ [where she had been admitted a month ago.]$_4$ [She would be 79 years old today.]$_5$ [...] [Her funeral will be held today at 10h30 at the church of Saint-Etienne of Bar-le-Duc.]$_6$.

A causal relation like Result, or at least a temporal Narration holds between 3 and 6, but it should not scope over 4 and 5 if one does not wish to make Sequin's admission to the hospital a month ago and her turning 79 a consequence of her death last Saturday.

#### 2.3.3 Relation labelling

SDRT models the semantics/pragmatics interface using discourse relations that describe the rhetorical roles played by utterances in context, on the basis of their truth conditional effects on interpretation. Relations are constrained by: semantic content, pragmatic heuristics, world knowledge and intentional knowledge. They are grouped into *coordinating* relations that link arguments of equal importance and *subordinating* relations linking an important argument to a less important one. This semantic characterization of discourse relations has two advantages for our study: first, the semantics of discourse relations makes it more straightforward to study their interactions with the semantics of subjective expressions, and secondly the semantic classification in SDRT leads to a smaller taxonomy of discourse relations than that given in RST, enabling an initial study of the interaction of discourse structure and opinion to find generalisations. Additionally, the fact that in SDRT multiple relations may relate one discourse unit to other discourse units allows us to study more complex interactions than it would be possible in the other theories.

Figure 1 gives an example of the discourse structure of the example (6), familiar from Asher and Lascarides (2003). In this figure, circles are EDUs, rectangles are complex segments, horizontal links are coordinating relations while vertical links represent subordinating relations.

(6) [John had a great evening last night.]$_1$ [He had a great meal.]$_2$ [He ate salmon.]$_3$ [He devoured lots of cheese.]$_4$ [He then won a dancing competition.]$_5$

## 3. The CASOAR corpus

We selected data according to four criteria: document genre, the number of documents per topic, document length and the type of opinion conveyed in the document. To better capture the dependencies between discourse structure and corpus genre, the annotation campaign should be conducted on different types of online corpora, each with a distinctive style and audience. For each corpus, topics (a movie, a product, an article, etc.) have to be selected according to their related number of documents or reviews. Our hypothesis was that the more attractive a topic is (i.e., it aroused a great

16. For a discussion of long-distance discourse relations in RST, see (Marcu (2000)).

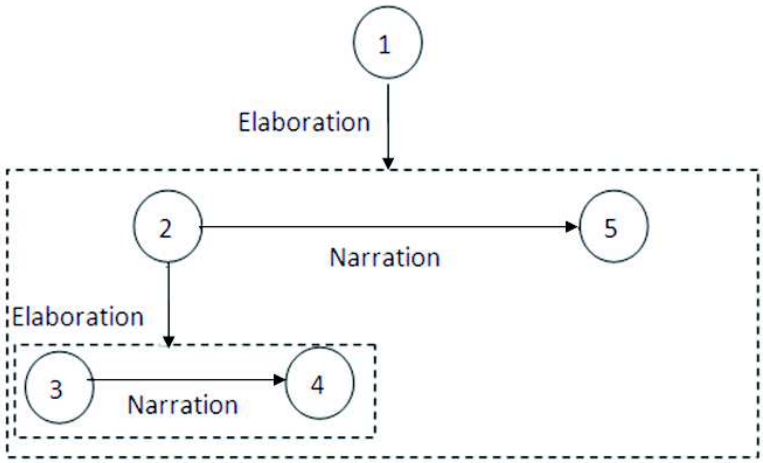


Figure 1: An example of a discourse graph.

number of reactions), the more opinionated the reviews are. In addition, the number of positive and negative documents has to be balanced. Given that discourse annotation is time consuming and error prone, especially for long texts where long distance attachments are frequent, documents should not be too long. On the other hand, documents should have an informative discourse structure and hence should not be too short either. Finally, the data should contain explicit opinion expressions as well as implicit opinions. One of our aims was to measure how these kinds of opinions are assessed in discourse.

Given these criteria, we chose to build our own corpus and not to rely on existing opinion datasets. Indeed, in French, the only existing and freely available opinion dataset (the Blogoscopy corpus Daille et al. (2011)[17]) was not available when we began our annotation campaign. In English, there are several freely available corpora already annotated with opinion information. Among them, we have studied four resources: the well known MPQA (Wiebe et al. (2005)) corpus[18], the Sentiment Polarity DataSet and the Subjectivity DataSet[19] (Pang and Lee (2004)), and the Customer Reviews Dataset[20](Hu and Liu (2004)). We chose not to build our discourse based opinion annotation on the top of MPQA for two reasons. First, text anchors which correspond to opinion in MPQA are not well defined since each annotator is free to identify expression boundaries. This is problematic if we want to integrate rhetorical structures into the opinion identification task. Secondly, MPQA often groups discourse indicators (but, because, etc.) with opinion expressions not leading to any guarantee that text anchors will correspond to a well formed discourse unit.

The Sentiment Polarity DataSet consists of 1,000 positive and 1,000 negative processed reviews annotated at the document level. However, it was not appropriate for our purposes because the documents in this corpus are very long (more than 30 sentences per document) which would have made the annotation of the discourse structure too hard. On the other hand, the Subjectivity DataSet

17. http://www.lina.univ-nantes.fr/?Blogoscopie,762.html
18. http://mpqa.cs.pitt.edu
19. www.cs.cornell.edu/people/pabo/movie-review-data
20. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

contains 5,000 subjective and 5,000 objective processed sentences. Only sentences or snippets containing at least 10 tokens were included along with their automating labelling decision (objective vs. subjective), as shown in (7). Since sentences are short (at most 3 discourse units), this corpus also did not meet our criteria. Finally, the Customer Reviews Dataset consists of annotated reviews of five products (digital camera, cellular phone, mp3 player and dvd player), extracted from Amazon.com. This corpus provides only target and polarity annotations at the sentence or the snippet level focusing on explicit opinion sentences (cf. (8) and (9) where [u] indicates that the target is not lexicalized (implicit)).

(7) nicely combines the enigmatic features of 'memento' with the hallucinatory drug culture of 'requiem for a dream . '

(8) camera[+2] ## This is my first digital camera and what a toy it is...

(9) size[+2][u] ## it is small enough to fit easily in a coat pocket or purse.

To conclude, none of the above pre-existing annotated corpora firs our objectives. We thus built the following corpora, summarized in Table 1:

- The French data are composed of two corpora: French movie/product reviews (*FMR*) and French news reactions (*FNR*). The movie reviews were taken from AlloCine.fr, book and video game reviews from Amazon.fr, and restaurant reviews from Qype.fr. The news reactions, extracted from lemonde.fr, are reactions to articles from the politics and economy sections of the "Le Monde" newspaper. We selected those topics (movies, products, articles) that are associated to more than 10 reviews/reactions. In order to guarantee that the discourse structure is informative enough, we also filtered out documents containing less than three sentences. In addition, for *FMR*, we balanced the number of positive and negative reviews according to their corresponding general evaluation (i.e., stars[21]). For *FNR*, reactions that are responses to other reactions were removed.

- The English data are movie reviews (*EMR*) from MetaCritic[22]. The choice of movie reviews is motivated firstly by the fact that this genre is widely used in the field and secondly, by our aim to compare how opinions are expressed in discourse in different languages (movie reviews were also selected for the French annotation campaign). The selection procedure (number of reviews per movie, number of sentences per review) was the same as for the one used in French data selection.

| | Number of documents | Selected topics |
|---|---|---|
| *FMR* | 180 | films (6), books and video games (6), restaurants (13), TV series (20) |
| *FNR* | 131 | politics (5), economy (6), international (2) |
| *EMR* | 110 | films (11) |

Table 1: Characteristics of our data.

21. The star scale was 1-5 and neutral reviews (3-star) were equally distributed in the positive/negative class.
22. `http://www.metacritic.com`

## 4. Methods

### 4.1 Annotation scheme

The annotation scheme is multi-layered, and includes: (1) the complete discourse structure according to SDRT, (2) opinion orientation of EDUs, and (3) opinion expressions, and their associated features. Each level has its own annotation manual and annotation guide, as described in the next sections.

In the remainder of this paper, all the examples are extracted from our corpora. Examples from *EMR* are given in English while examples from *FMR* and *FNR* are given in French along with their direct English translation (when possible). Note however that there are substantial semantic differences between the two languages.

#### 4.1.1 The document level

In this level, annotators were asked to give the document overall opinion towards the main topic using a five-level scale, where 0 indicates a very bad (negative) opinion and 4 a very good (positive) one. Then, annotators have to build the discourse structure of the document following the SDRT principles.

Our discourse annotation scheme was inspired from an already existing manual elaborated during the Annodis project, a French corpus where each document was annotated according to the principles of SDRT. This manual gives a complete description of the semantics of each discourse relation along with a listing of possible discourse markers that could trigger any particular relation. However, the manual did not provide any details concerning the structural postulates of the underlying theory. This was justified, since one of the objectives of the Annodis project was to test the intuitions of the naive annotators relevant to these issues. In CASOAR however, we aimed at testing the intuitions of naive annotators on *how discourse interacts with opinion*. We therefore modified the Annodis manual in order to make precise all the constraints annotators should respect while building the discourse graph. In particular, we made explicit the constraints concerning segment attachment and accessibility of complex segments. We stipulated in the manual that each segment in the graph should be connected and that the attachment should normally follow the reading order of the document and the right frontier principle (cf. Section 2.3). CDU constraints detailed how EDUs can be grouped to form complex units. Figure 2 shows an example of a complex discourse unit constraint. Suppose [1,2] and [2,3] are CDUs. Figures on the right and in the middle are correct configurations whereas the one on the left is not allowed for two main reasons: an EDU cannot belong to two distinct CDUs (as the EDU 2 in the CDUs [1,2] and [2,3]) and the head of a CDU[23] cannot appear as a second argument of a relation.

During the writing of this manual, we faced another decision: (1) should we annotate opinion texts using a small set of discourse relations or (2) should we use a larger set (i.e., the 19 relations already used in the Annodis project). The first solution is more convenient and has already been investigated in previous studies. For example, in Asher et al. (2008), we experimented with an annotation scheme where lexically-marked opinion expressions and the clauses involving these expressions are related to each other using five SDRT-like rhetorical relations: Contrast and Correction (introduced by signals such as: *although, but, contradict, protest, deny, etc.*), Support that

23. The head of a CDU is the first EDU that composes it. For example, 1 is the head of the CDU [1,2].

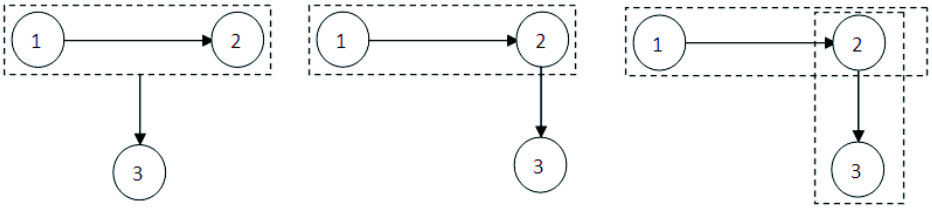


Figure 2: A CDU constraint.

groups together Explanation and Elaboration, Result (usually marked by *so, as a result*) which indicates that the second argument is a consequence or the result of the first argument, and finally, Continuation. Somasundaran (2010) proposed the notion of opinion frames as a representation of documents at the discourse level in order to improve sentence-based polarity classification and to recognize the overall stance. Two sets of homemade relations were used: relations between targets (same and alternative relations) and relations between opinion expressions (reinforcing and non-reinforcing relations). Finally, Trnavac and Taboada (2010) examined how some nonveridical markers and two types of rhetorical relations (Conditional and Concessive) contribute to the expression of appraisal in movie and book reviews. In our case, we chose not to use a predefined small set of rhetorical relations selected according to our intuitions because we *did not know in advance* what were the most frequent relations occuring in opinion texts and *how this frequency was correlated with corpus genre*. Of course, this choice made it harder to do the annotations. But we think that this was a necessary step to investigate the real effects of discourse relations on both polarity and subjectivity as well as to evaluate the impact of discourse structure when assessing document overall opinion.

Among the set of 19 relations used in the Annodis project, we focused our study on 17 relations that involve entities from the propositional content of the clauses[24]. These relations are grouped into coordinating relations (Contrast, Continuation, Conditional, Narration, Alternative, Goal, Result, Parallel, Flashback) and subordinating relations (Elaboration, E-Elab, Correction, Frame, Explanation, Background, Commentary, Attribution). Table 2 provides a detailed list of these relations along with their definitions. In this table, $\alpha$ and $\beta$ stand respectively for the first and the second argument of a relation. $(C)$ and $(S)$ represent respectively coordinating and subordinating relations.

Annotators were asked to link constituents (EDUs or CDUs) through whichever discourse relation they felt appropriate, from our list above. In addition to this set of 17 relations, we also added the relation Unknown in case annotators were not able to decide which relation is more appropriate to link two constituents.

#### 4.1.2 The segment level

For each EDU in a document, annotators were asked to annotate its subjectivity orientation as well as its polarity and strength.

**Subjectivity orientation.** It can belong to five categories:

24. Meta-talk (or pragmatic) relations that link the speech acts expressed in one unit and the semantic content of another unit that performs it were discarded.

| **Discourse relations** | **Definitions** |
|---|---|
| Causality | |
| EXPLANATION (S) | the main eventuality of $\beta$ is understood as the cause of the eventuality in $\alpha$ |
| GOAL (S) | $\beta$ describes the aim or the goal of the event described in $\alpha$ |
| RESULT (C) | the main eventuality of $\alpha$is understood to cause the eventuality given by $\beta$ |
| Structural | |
| PARALLEL (C) | $\alpha$ and $\beta$ have similar semantic structures. The relation requires $\alpha$ and $\beta$ to share a common theme |
| CONTINUATION (C) | $\alpha$ and $\beta$ elaborate or provide background to the same segment |
| CONTRAST (C) | $\alpha$ and $\beta$ have similar semantic structures, but contrasting themes or when one constituent negates a default consequence of the other |
| Logic | |
| CONDITIONAL (C) | $\alpha$ is a hypothesis and $\beta$ is the consequence. It can be interpreted as: if $\alpha$ then $\beta$ |
| ALTERNATION (C) | $\alpha$ and $\beta$ are related by a disjunction |
| Reported Speech | |
| ATTRIBUTION (S) | relates a communicative agent stated in $\alpha$ and the content of a communicative act introduced in $\beta$ |
| Exposition/Narration | |
| BACKGROUND (S) | $\beta$ provides information about the surrounding state of affairs in which the eventuality mentioned in $\alpha$ occurs |
| NARRATION (C) | $\alpha$ and $\beta$ introduce an event and the main eventualities of $\alpha$ and $\beta$ occur in sequence and have a common topic |
| FLASHBACK (C) | is equivalent to NARRATION($\beta$,$\alpha$). The story is told in the opposite temporal order |
| FRAME (S) | $\alpha$ is a frame and $\beta$ is on the scope of that frame |
| Elaboration | |
| ELABORATION (S) | $\beta$ provides further information (a subtype or part of) about the eventuality introduced in $\alpha$ |
| ENTITY-ELABORATION (S) | $\beta$ gives more details about an entity introduced in $\alpha$ |
| Commentary | |
| COMMENTARY (S) | $\beta$ provides an evaluation of the content associated with $\alpha$ |
| Correction | |
| CORRECTION (S) | $\alpha$ and $\beta$ have a common topic. $\beta$ corrects the information given in the segment $\alpha$ |

Table 2: SDRT relations in the CASOAR corpus.

- *SE* – segments that contain explicitly lexicalized subjective and evaluative expressions, [*One of the best films I've ever seen in my life.*]

- *SI* – segments that do not contain any explicit subjective cues but where opinions are inferred from context, as in [*This is a definite choice to be in my DVD collection,*] [*and should be shared by fathers to their sons for generations.*]

- *O* – segments that contain neither a lexicalized subjective term nor an implied opinion. They are purely factual, as in [*I went to the cinema yesterday.*]

- *SN* – subjective, but non-evaluative segments used to introduce opinions. In general, these segments contain verbs used to report the speech and opinions of the author or others, as in the first segment in [*I have no doubt*][*that this movie is excellent*]. The opinion polarity (positive, negative, or neutral) is given by the verb complements. It is important to note that the *SN* category does not cover the cases of neutral opinion.

- *SEI* – that contain both explicit and implicit evaluations on the same topic or on different topics. For instance, [*Fantastic pub !*]$_a$ [*The pretty waitresses will not hesitate to drink with you*]$_b$, segment *b* contains two opinions, one explicit, towards the waitress, and the other one implicit, towards the pub.

**Polarity.** It can have five different values: *positive*, *negative*, *neutral*, *both*, and *no polarity*. Neutral indicates that the positivity/negativity of the segment depends on the context, as in [*This movie is poignant*]. Both means that the segment has a mixed polarity as in [*This stupid President made a wonderful talk*]. Finally, no polarity concerns segments do not convey any evaluation (i.e., *O* and *SN* segments).

**Strength.** Several types of scales have been used in sentiment analysis research, going from continuous scales (Benamara et al. (2007)) to discrete ones (Taboada et al. (2011)). In our case, we think that the chosen scale has to ensure a trade off between a fine-grained categorisation of subjectivity and the reliability of this categorization with respect to human judgments. For our annotation campaign, we chose a discrete 3-point scale, $[1, 3]$ where 1 indicates a weak strength. Objective segments (*O*) are associated by default to the strength 0.

#### 4.1.3 The opinion expression level

After segment annotation, the next step is to identify within each EDU at least one of these elements: the opinion expression span, opinion topic, opinion holder, and operators that interact locally with opinion expressions. Once all these elements are identified, annotators have to link every operator, topic and holder to its corresponding opinion expression using the Scope relation. This relation aims to link: an operator to an opinion expression under its scope, a holder to its associated opinion expression, and an opinion expression to its related topic. Since most opinion expressions reflect the writer's point of views (i.e., the main holder), we decided not to annotate the scope relation in this case so as not to make the annotation more laborious. Operators as well as topics are linked to the opinion in their scope only if several opinion expressions are present in an EDU. We detail below the annotation scheme.

**Opinion expression span.** Within each EDU, annotators can identify zero (in case of *SI* and *O* segments), one or several non overlapping opinion spans. An opinion span is composed of subjective tokens (adjectives, verbs, nouns, or adverbs), excluding operators[25]. Its annotation includes: a *polarity* (positive, negative, and neutral), a *strength* (on a discrete 3-point scale, cf. above), a *semantic category* and a *subcategory*. According to the opinion categorization described in Asher et al. (2008), each opinion expression can belong to four main categories: *Reporting* which provides, at

25. Operators are annotated separately. The idea is to capture both the prior and contextual polarity of opinion expressions. Contextual polarity is annotated at the segment level while prior polarity at the expression level.

least indirectly, a judgment by the author on the opinion expressed, *Judgment* which contains normative evaluations of objects and actions, *Advice* which describes an opinion on a course of action for the reader, and *Sentiment-Appreciation* containing feelings and appreciations. Subcategories include, for example, *inform, assert, evaluation, recommend, fear, astonishment, blame, etc.*

**Topics and holders.** They are textual spans within a segment that are associated with a type. The opinion topic can have three types: *main* indicating the main topic of the document, such as "the movie", *part of* in case of features related to the main topic, such as "the actors", "the music", and finally *other* when the topic has no ontological relation with the main topic, for example "theater" in *The movie was great. Shame that the theater was dirty*. Also, we distinguish between two types of holders: *main* that stands for the author's review and *other* (as in *My mother loved the movie*).

**Operators.** Finally, we deal with four types of operators: (i) *negations* that may affect the polarity and the strength of an expression, (ii) *modals* used to express the degree of belief of the holder, (iii) *intensifiers* used to strengthen (we use the operator *Int+*) or weaken (*Int-*) the prior polarity of a word or an expression, and (iv) *restrictors* that narrow the scope of the opinion in the sense that the positivity and/or negativity of the expression can be evaluated only under certain conditions, as in *the restaurant is very good for children*. Operators have to be annotated when opinion expressions are under their scope as well as in case of implicit segments when appropriate.

#### 4.1.4 A complete example

Figure 3 gives the annotation at the opinion expression and the segment level of the review (10), taken from $EMR$. In this figure, we provide for each opinion expression its polarity and strength. Similarly, we associate for each segment a triple that indicates its type (among: *SE*, *SI*, *0*, *SN*, and *SEI*), polarity (among: +, –, *neutral*, *both*, and *no polarity*), and strength (in a three level scale). Figure 4 provides the associated discourse graph.

(10) [I saw this movie on opening day.]$_1$ [Went in with mixed feelings,]$_2$ [hoping it would be good,]$_3$ [expecting a big let down]$_4$ [(such as clash of the titans (2011), watchmen etc.).]$_5$ [This movie was shockingly unique however.]$_6$ [Visuals, and characters were excellent.]$_7$

### 4.2 Annotation procedure

#### 4.2.1 Data preparation

In order to avoid errors in determining the basic units (which would thus make the inter-annotator agreement study problematic), we decided to discard the segmentation from the annotation campaign. Instead, EDUs were automatically identified. To train our segmenter, two annotators manually annotated a subset of $FMR$ (henceforth $FMR'$) by consensus. This yields a total of 130 documents and 1,420 EDUs, among which 1.33% were embedded.

Automatic segmentation was carried out by adapting an already existing SDRT-like segmenter (Afantenos et al. (2010)), built on the top of the Annodis corpus[26]. The features used in Afantenos et al. (2010) include the distance from sentence boundaries, the dependency path, and the chunk start/end. Since we used a different syntactic parser, we modified certain features accordingly, and

26. The corpus used for training the parser was composed of 47 documents extracted from *L'Est Rébublicain* newspaper. This corpus is mainly objective and contains 1,400 EDUs, among them 10% were nested.

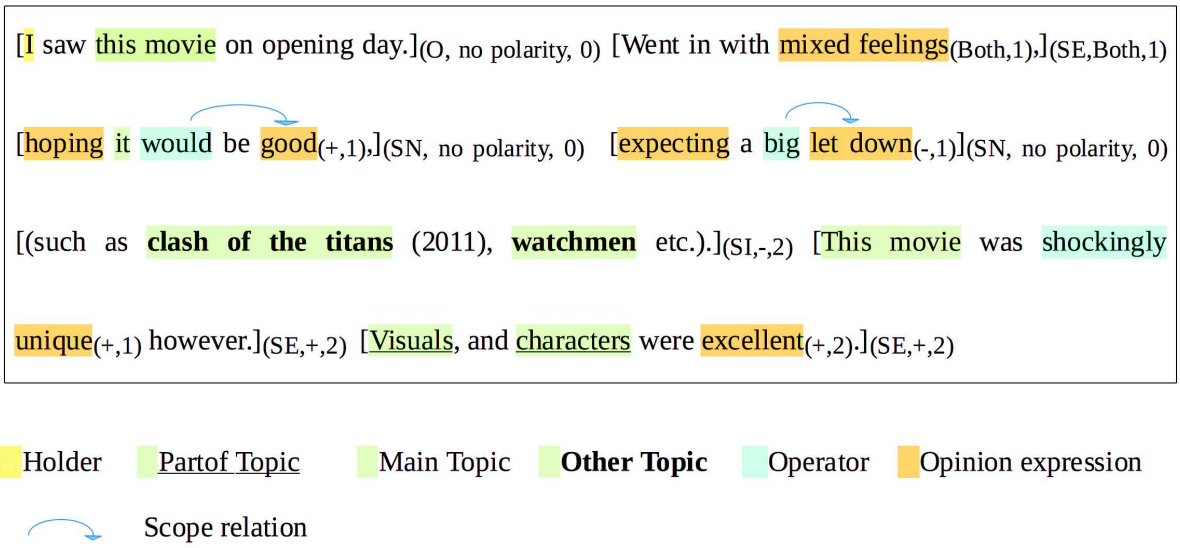


Figure 3: Annotation of (10) at the segment and the opinion expression level.

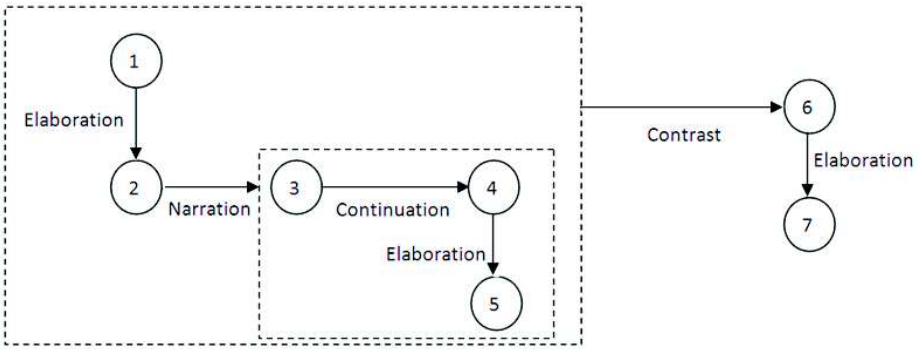


Figure 4: Discourse annotation of (10).

discarded others. We performed a two-level segmentation. First, we constructed a feature vector for each word token, which is classified into: *Right* for words starting an EDU, *Left* for tokens ending an EDU, *Nothing* for words completely inside an EDU, and *Both* for tokens which constitute the only word of an EDU. Once all EDUs were found, subjective EDUs that contain at least one token belonging to our subjective lexicon[27] are filtered out because they are good candidates for a further segmentation. The proportion of such EDUs in $FMR'$ was relatively small (around 12%). This second step was performed using symbolic rules which are mainly based on discourse connectives and punctuation marks.

27. Our lexicon is manually built and is composed of 270 verbs, 632 adjectives, 296 nouns, 594 adverbs, 51 interjections.

| | Right | | | Left | | | Nothing | | |
|---|---|---|---|---|---|---|---|---|---|
| | **R** | **P** | **F** | **R** | **P** | **F** | **R** | **L** | **N** |
| (E1) | 0.858 | 0.913 | 0.885 | 0.872 | 0.894 | 0.883 | 0.976 | 0.967 | 0.972 |
| (E2) | 0.791 | 0.927 | 0.853 | 0.752 | 0.917 | 0.827 | 0.978 | 0.926 | 0.952 |
| (E3) | 0.925 | 0.942 | 0.933 | 0.941 | 0.952 | 0.946 | 0.982 | 0.977 | 0.980 |

Table 3: Evaluation of the classifier in terms of precision (P), recall (R), and F-measure (F).

| | $FMR'$ | | | $FMR'_{Lex}$ | | |
|---|---|---|---|---|---|---|
| | **R** | **P** | **F** | **R** | **P** | **F** |
| Boundaries | 0.976 | 0.968 | 0.972 | 0.961 | 0.977 | 0.969 |
| EDU recognition | 0.821 | 0.732 | 0.774 | 0.751 | 0.772 | 0.762 |

Table 4: Evaluation of the symbolic rules in terms of precision (P), recall (R) and F-measure (F).

Our discourse segmentation followed a mixed approach using both machine learning and rule-based methods. We first evaluated the classifier and then the symbolic rules. We performed a supervised learning using Maximum Entropy model[28] in order to classify each token into *Right*, $Left$, *Nothing* or *Both* classes as described above. We conducted three evaluations: (E1) a 10-fold cross validation on the Annodis corpus in order to compare our results to the ones obtained by Afantenos et al. (2010); (E2) training on Annodis and testing on $FMR'$ to see to what extent our set of features was independent of the corpus genre; (E3) a 10-fold cross validation on $FMR'$. Table 3 shows our results for the *Right*, $Left$, and *Nothing* boundaries, in terms of precision (P), recall (R), and F-measure (F). Our results for the configuration (E1) are similar to those obtained by Afantenos et al. (2010) on Annodis. The best performance was achieved when training on our data (i.e., the configuration (E3)).

Table 4 shows the results of the symbolic rules when applied on the outputs of the configuration (E3). Results concern both segment boundaries (averaged over all the four classes) and the recognition of an EDU as a whole with a begin boundary and its corresponding end. We evaluated both on $FMR'$ when subjective EDUs are given by manual annotation and on $FMR'_{Lex}$ when they are automatically identified using our lexicon. Again, our rules performed very well.

This tool was used to automatically segment $FMR$ and $FNR$ documents. The resulting segmentation was manually corrected when necessary[29]. We did not design an automatic segmenter for English and segmentation in $EMR$ was performed manually by two annotators by consensus.

#### 4.2.2 ANNOTATION CAMPAIGN

We managed two annotation campaigns. The French one was the first and took six months. The English campaign came second and lasted three months. $FMR$ and $FNR$ was doubly annotated by three French native speakers while $EMR$ was annotated by two English native speakers. French annotators were undergraduate linguistic students while English ones were teachers. Annotators

28. `http://www.cs.utah.edu/~hal/megam/`
29. We mainly corrected unbalanced bracketing. To this end, we designed a script that recognizes if for each begin bracket, there is a corresponding end bracket. If not, we manually ensured correct bracketing. We also checked if the other segmentation cases that we defined were correctly handled. Overall, manual correction was very fast.

benefited from a complete and revised annotation manual as well as an annotation guide explaining the inner workings of the GLOZZ platform[30], our annotation tool. Since documents are already segmented, annotators first had to click on each EDU, specified its category, polarity, and strength (see Section 4.1.2), and then could isolate, within each EDU, spans of text corresponding to the annotation scheme described in Section 4.1.3. Discourse annotation was performed by inserting relations between selected constituents using the mouse. When appropriate, EDUs were grouped to form CDUs using GLOZZ schemata. GLOZZ also provides a discourse graph as part of its graphical user interface which helps the annotator to better capture the discourse structure while linking constituents. Figure 5 illustrates how a document, extracted from *EMR*, is annotated under GLOZZ. The first segment includes the spans *This* and *movie* annotated as main topics, *definitely* and *all time* annotated as intensifier operators and *the best* annotated as an opinion expression. The annotation associated to the first segment is shown in the features structure on the right. Segment 2 and 3 are related with a CONTINUATION relation, and the structure *Continuation*$(2, 3)$ is grouped into a CDU (the blue circle in the Figure).

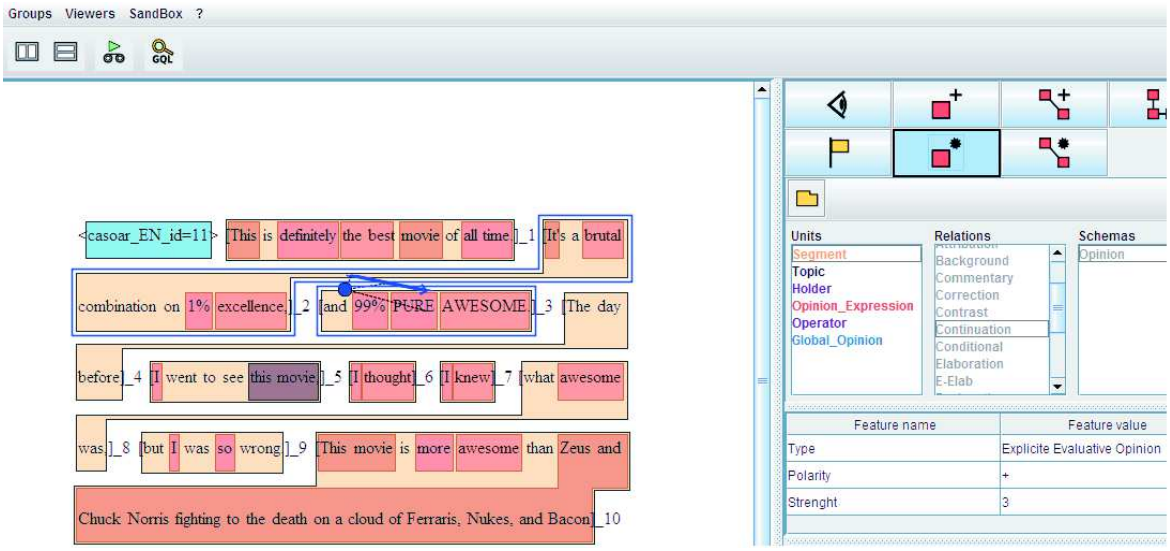


Figure 5: The annotation of an English movie review under the Glozz platform.

The French annotation proceeded in two stages. First, the annotation of the movie reviews; then, the annotation of news reactions. For each stage, we performed a two-step annotation where an intermediate analysis of agreement and disagreement between the three annotators was carried out. Annotators were first trained on 12 movie reviews and then they were asked to annotate separately 168 documents from *FMR*. Then, they were trained on 10 news reactions. Afterwards, they continued to annotate separately 121 documents from *FNR*. The training phase for *FMR* was longer than for *FNR* since annotators had to learn about the annotation guide and the annotation tool. Similarly, the English annotation campaign was done in two steps. Annotators were trained on 10 *EMR* and then the rest of the corpus (100 documents) was annotated separately. The time needed to annotate entirely one text was about 1 hour.

30. `www.glozz.org`

During training, we noticed that annotators often made the same errors. At the segment and the opinion expression level, these errors included: segments labelled as opinionated (*SE* and *SEI*) with no opinion expression inside; *O* or *SI* segments with an opinion expression inside; *O* and *SN* segments with a prior polarity; opinion expressions with no associated semantic category, etc. For example, if one annotator considered the following segment *I am a huge fan of Tintin* to be subjective, he should annotate the span *fan* as being an opinion expression. Some of the discourse-level errors include: violation of the right frontier constraint, cycles, overlapping CDUs, segments not attached to the discourse graph, etc. To ensure that the annotations were consistent with the instructions given in the manual, we designed a tool to automatically detect these errors. Among all the provided annotations, 15% of the French documents contained errors at the segment and the opinion level vs. 12% for the English documents. The annotators were asked to correct their errors before continuing to annotate new documents. With respect to discourse structure, just a few French documents were ill-formed. However, the English annotators felt uncomfortable with discourse annotation, and their annotations were full of errors. We retrained them but finally decided to annotate discourse in *EMR* by consensus.

### 4.3 Reliability of the annotation scheme

In this section, we report on inter-annotator agreements at the document, segment, and opinion expression levels. All statistics have been computed using the IRR library under R[31].

#### 4.3.1 At the document level

Recall that the document annotation level consists of two tasks: assigning to each document an overall opinion (on a discrete five-level scale) and then a discourse structure.
Agreements have been computed on 152 *FMR* documents, 100 *EMR*, and 120 *FNR*.

**Agreements on overall opinion.** We used two different measures. First, Cohen's Kappa which assesses the amount of agreement between annotators. Second, Pearson's correlation that measures the linear correlation between two vectors variables: the annotators' overall opinions (variable 1) and the original overall opinions as given by Allociné or MetaCritic users (variable 2). The aim is identify whether the first variable tends to be higher (or lower) for higher values of the other variable. Pearson's correlation gives a value between $[-1, +1]$ where $+1$ indicates a total positive correlation, 0 no correlation, and 1 total negative correlations.

Table 5 gives our results in terms of Cohen's Kappa when overall opinion has to be stated on the five level scale 0 to 4 (Kappa multi-scale), the weighted Kappa (weighted Kappa multi-scale), and the Kappa after collapsing the ratings 0 to 2 and 3 to 4 into respectively positive and negative ratings (Kappa polarity). Compared to a non weighted version, weighted Kappa allows to compute agreements on ordinal labels. Hence, a disagreement of 0 vs. 4 is much more significant that a disagreement of 1 vs. 2. We also give the average Pearson's correlation between the overall opinion given by our annotators and the overall ratings already associated to each movie review documents[32].

Our results are good in movie reviews in polarity rating and weighted Kappa but moderate in multi-scale rating, with a lower value obtained for news reactions. This shows that news reactions

31. https://cran.r-project.org/web/packages/irr/irr.pdf
32. Correlations are given only for *FMR* and *EMR* documents since in news reactions (*FNR*), authors are not asked to give the overall opinion of their comments.

| | Kappa multi-scale | Weighted Kappa multi-scale | Kappa polarity | Pearson Correlation |
|---|---|---|---|---|
| $FMR$ | 0.48 | 0.66 | 0.68 | 0.83 |
| $EMR$ | 0.53 | 0.72 | 0.70 | 0.79 |
| $FNR$ | 0.40 | 0.51 | 0.55 | – |

Table 5: Inter-annotator agreements on document overall opinion rating.

are more difficult to annotate. Finally, when evaluating the correlation between the annotators' overall opinions and the authors overall scores, we observe that correlations are good.

**Agreements on discourse structure.** As described in Section 4.1, discourse annotation depends on two decisions: a decision about where to attach a given EDU, and a decision on how to label the attachment link via discourse relations. Two inter-annotator agreements have thus to be computed and the second one depends on the first because agreements on relations can be performed only on common links. For attachment, we obtained an F-measure of 69% for $FMR$ and 68% for $FNR$ assuming attaching is a yes/no decision on every EDUs pair, and that all decisions are independent, which of course underestimates the results. When commonly attached pairs are considered, we get a Cohen's Kappa of 0.57 for the full set of 17 relations for $FMR$ and 0.56 for $FNR$, which is moderate. Here again, this Kappa is computed without an accurate analysis of the equivalence between rhetorical structures[33]. Figure 6 shows two discourse annotations for the French movie review in Example (11). We observe that the annotator (on the left) formed more CDUs than the other annotator (on the right) which causes both attachment and relation labeling errors. Our goal being to study the effects of discourse on opinion analysis, a detailed analysis of inter-annotator attachment agreements is out of the scope of this study and is left for future work.

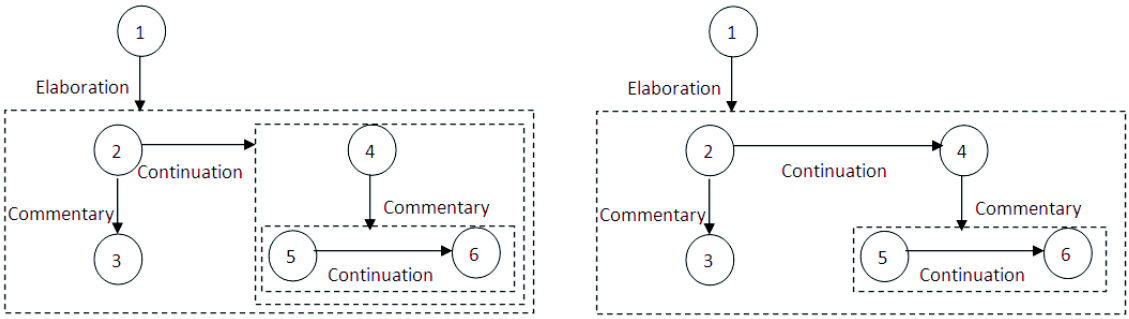


Figure 6: Two discourse annotations of Example (11).

Overall, our results are higher than those obtained by Annodis (66% F-measure for attachment and a Cohen's Kappa of 0.4 for relation labeling) mainly for two reasons. First, our annotation manual was more constrained since we provided annotators a detailed description of how to build

33. See (Afantenos et al. (2012)) for an interesting discussion on the difficulty on how to compare rhetorical structures, especially when CDU are have to be taken into account.

the discourse structure. Second, our documents are smaller (an average of 20 EDUs compared to 55 EDUs in Annodis) which implies less long distance attachments.

(11) [Bonne série.]$_1$ [Petits épisodes plus ou moins bien ficelés]$_2$ [(mais n'est-ce pas le cas dans les autres aussi ?).]$_3$ [Le tout tenant en une 20e de minutes...]$_4$ [Rapide]$_5$ [et sans temps morts.]$_6$
[Good TV series.]$_1$ [Small serials more or less well done]$_2$ [(but it isn't the case in the others too ?)]$_3$ [All within 20 minutes time...]$_4$ [Fast]$_5$ [and without time out.]$_6$

#### 4.3.2 At the segment/opinion level

Table 6 shows the inter-annotator agreements on segment opinion type, segment polarity and segment strength averaged over all the annotators. Agreements have been computed on 1706 *FMR* segments, 1260 *EMR*, and 1060 *FNR*. When computing these statistics for segment polarity, we have discarded the *neutral* category since we do have few instances of it in our data. In addition, since the *both* category means that the segment conveys at the same time negative and positive opinions, we decided to count it only once by conflating it with the *positive* category. Similarly, we have also counted the *SEI* class (which indicates that segments contain both implicit and explicit opinions) with *SE*.

| | *FMR* | *EMR* | *FNR* |
|---|---|---|---|
| Kappa on segment opinion type | 0.66 | 0.60 | 0.50 |
| Kappa on segment polarity | 0.76 | 0.71 | 0.48 |
| Kappa on segment strength | 0.35 | 0.27 | 0.27 |
| Weighted Kappa on segment strength | 0.49 | 0.43 | 0.34 |

Table 6: Inter-annotator agreements on segment opinion type, polarity, and strength per corpus genre.

We observe that the inter-annotators agreements are better for movie reviews than for news reactions and that *FMR* achieves the best scores. We get very good Kappa measures for both explicit opinion segments *SE* (0.74) and the polarity (positive and negative) of a segment in French movie reviews (respectively 0.78 and 0.77). We get similar results in English with as an example a Kappa of 0.67 for the *SE* class and a Kappa of 0.75 and 0.74 for respectively positive and negative segment opinion type. These results are in agreement with state-of-the-art results obtained in contemporary annotation campaigns (see e.g. Wiebe et al. (2005)). The Kappa for the *SN* class is also very good: 0.74 in *FMR* and 0.64 in *EMR*. Finally, the agreements for the *SI* and *O* classes were respectively 0.56 and 0.63 in *FMR*, and 0.52 and 0.58 in *EMR*. They are moderate because annotators often fail to decide whether a segment is purely objective and thus if it conveys only facts or if a segment expresses an implicit opinion. Here are two examples illustrating annotators disagreement on segment opinion type:

(12) [As mentioned elsewhere,]$_1$ [the romance in the movie was painful]$_2$ [but helped tie things up at the end.]$_3$ [Good way to burn 2h of your life and 15$]$_4$.

(13) [the production company had any idea how to market this film.]$_1$ [The trailer looks like a non-stop action thriller set in a train station,]$_2$ [when in fact it is far slower]$_3$ [but well-paced,]$_4$ [and its best moments come away from the station.]$_5$

In (12), one annotator (A) considered that segments 3 and 4 conveyed positive implicit opinions towards the movie while the second annotator (B) has labeled these segments as explicit by selecting the spans *Good way* and *tie things up* as being positive opinion expressions. In (13), (A) and (B) agreed to put the segments 4 and 5 into the *SE* category but disagreed on the category of the first three segments: for (A), segments 1 and 2 are implicit negative segments whereas for (B) they are purely objective. Similarly, for (B) segment 3 is objective and for (A) it is an explicit opinion because it contains the word *slower* which has been annotated as a negative opinion expression.

The difficulty to discriminate between explicit, implicit, and objective segments can also be explained by the lower Kappa measure obtained for *no polarity* with 0.60 in $EMR$ and 0.68 in $FMR$ compared to the Kappa obtained on positive and negative segment polarity. This difficulty is, we believe, an artifact of the length of the texts. Indeed, the longer a text is, the greater the difficulty for human subjects to detect discourse context. However, the study of this hypothesis falls out of the scope of this paper and is therefore left for future work. Nonetheless, these results are good in the range of state-of-the-art research reports in distinguishing between explicit and implicit opinions. For instance, Toprak et al. (2010) obtained a Kappa of 0.56 for *polar fact* sentences which are close to our *SI* category.

In $FNR$, our results were moderate for the *SE* and *SN* classes (respectively 0.56 and 0.58) and weak for the *SI* and *O* classes (respectively 0.48 and 0.40). We have the same observations for the agreements on segment polarities where we obtain moderate Kappas on all the three classes (*positive, negative,* and *no polarity*). This shows that the newspaper reactions were more difficult to annotate because the main topic is more difficult to determine (even by the annotators) – it can be one of the subjects of the article, the article itself, its author(s), a previous comment or even a different topic, related to various degrees to the subject of the article. Implicit opinions, very frequent, can be of a different nature: ironic statements, jokes, anecdotes, cultural references, suggestions, hopes and personal stances, especially for political articles. Here is an example of implicit segments extracted from $FNR$. Annotators disagreed on how to annotate the first segment: for (A), 1 is negative implicit while for (B) it is explicit (with the spans *vraiment/really* and *plaindre/pity* annotated respectively as an operator and an opinion expression):

(14) [Les enseignants sont-ils vraiment à plaindre ?]$_1$ [Avec 6 mois de vacances par an]$_2$ [et la possibilité de prendre une retraite à 45 ans dans certains cas...]$_3$
[Are teachers really to be pitied ?]$_1$ [With 6 months vacation per year]$_2$ [and the opportunity to retire at 45...]$_3$

Finally, the Kappa for segment strength averaged over the scale $[0, 3]$ is bad. However, the Kappas are good on the extreme values of this scale, and moderate when using a weighted measure. For example, we get a Kappa of 0.67 and 0.58 in respectively $FMR$ and $EMR$ on the strength 0 vs. 0.4 in $FNR$. These results confirm that multi-scale polarity annotation is a difficult task, as already observed in similar annotation schema (cf. Toprak et al. (2010)). We think that low agreements were mainly due to the annotation manual that failed to clearly explain strength annotation. Indeed, for the same "basic" opinion expression, we got different annotations. For example, in similar contexts, the adjective *good* got different scores (+1 or +2). We think that the manual can be

improved by explicitly stating the prior score of "basic" expressions (e.g., *good* (+1), *brilliant* (+2) and *exceptional* (+3)) and then asking annotators to score new expressions by comparing their strength to these expressions.

## 5. Results

We give now the results of the annotation campaign focusing on quantitative results on each annotation level, and more importantly on the impact of discourse on sentiment analysis.

### 5.1 Quantitative analysis at the document level

Our discourse annotations contain a total of 3,453 discourse relations for $FMR$, 1,740 for $FNR$ and 1,677 relations for $EMR$. We analyzed our results according to two main axis: the distribution of relations per corpus genre and the importance of CDUs for sentiment analysis.

#### 5.1.1 Distribution of discourse relations per corpus genre

Figure 7 shows these distributions, sorted according to their frequency in $FMR$, from the most frequent (on the left) to the less frequent one (on the right). The frequencies of each discourse relation across corpus genres are statistically different from what would be expected by chance using the $\chi^2$ test. Note however that the difference between the observed and the expected frequencies of Conditional were not statistically significant. In this figure, we discarded the frequencies of the relations Flashback and Unknown for two reasons. First, Flashback was highly infrequent in all the corpora (0.12%, 0.06% and 0% for respectively $FMR$, $FNR$, and $EMR$) and second, the relation Unknown was not used in $EMR$ since the discourse annotation in this corpus has been performed by consensus. It is however interesting to note that this relation was more frequent in $FMR$ (around 2.06%) than in $FNR$ (0.69%) mainly because the annotators were more experienced with respect to the "Reviews" corpus (annotated first).

Overall, the frequencies can be grouped into three classes: (1) Continuation, Elaboration and Commentary (more than 10%), (2) Contrast, Entity-Elaboration, Result, Explanation, Attribution and Frame (from 3% to 10%) and (3) Correction, Goal, Narration, Parallel, Background, Conditional and Alternation (less than 3%). We noticed that some relations are more present in certain corpora. For instance, Commentary, Entity-Elaboration, Explanation, Attribution, Frame, Goal, Parallel and Alternative are more frequent in news reactions than in reviews. The frequencies of Parallel, Alternative and Frame are consistent with a logically more structured discourse for news reactions than for movie reviews. Also Goal and Explanation are more frequent which confirms that $FNR$ contains more argumentative structures than in reviews. The same goes for the Attribution relation, which denotes that in $FNR$ people tend to make reference to what other people said., e.g. *The president thinks that...*, or even that people tend to be more reserved when stating opinions, e.g. *I guess that this is a good measure*, unlike in the reviews, where people might tend to be more categorical, e.g. *This movie is great*, without modalizing the statement. Also, Entity-Elaboration is more frequent in $FNR$ (more than 10%), which confirms that news reactions are multi-topic opinion documents. Another interesting comparison between corpus genres is the frequency of Commentary, more frequent in news reactions where commentaries are often ironic. Finally, the proportions of Elaboration, Contrast, Background, Narration and Result in the En-

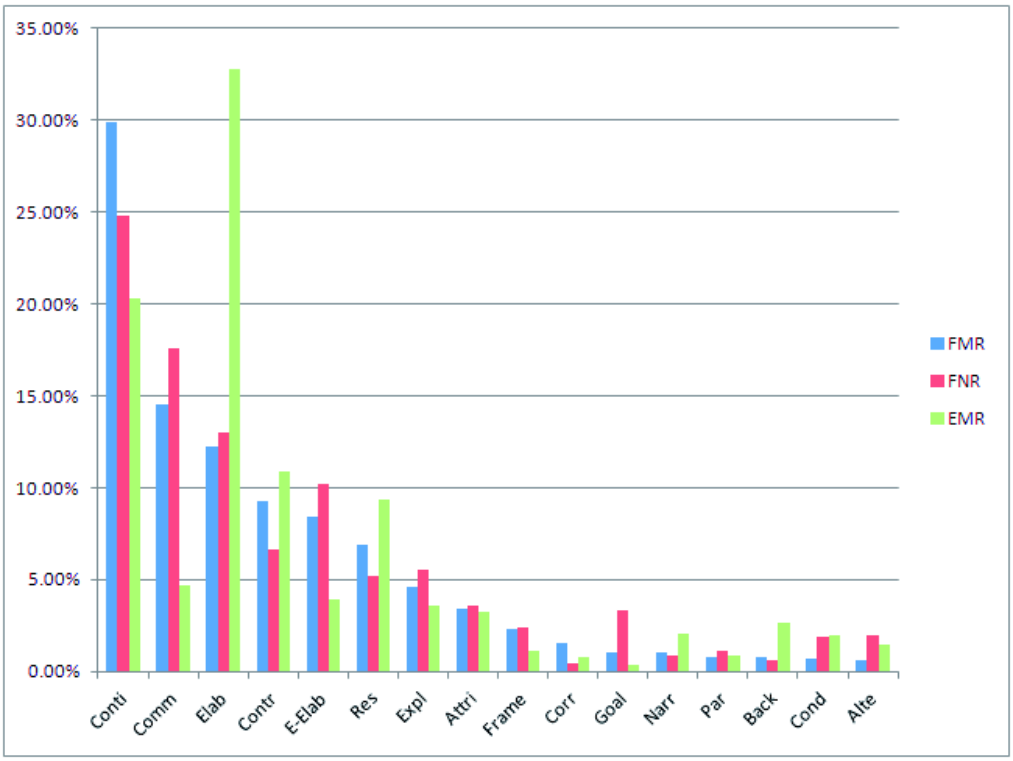


Figure 7: The distribution of discourse relations per corpus genre.

glish corpus were higher compared to the two other corpora, may be because English reviews tend to be more verbose.

### 5.1.2 IMPORTANCE OF CDUS

We have also analyzed the ratio of complex segments to the total number of rhetorical relation arguments in our annotations. Figures 8, 9, 10 show the proportions of relations between EDUs, between an EDU and a CDU, and between CDUs, sorted according to the increasing frequencies of relations between EDUs (all the relations are shown except UNKNOWN and FLASHBACK). First, we see that some relations are *local* and tend to appear more often between EDUs (more than 70%), as in Example (15) taken from *EMR*. In news reactions, these *local* relations have the same distributions except for ATTRIBUTION and CONDITIONAL which link simple segments in 60% of cases. This is more salient for BACKGROUND with only 45% of instances. We will see in Section 5.3 that some of these *local* relations are very important for sentiment analysis while others can simply be ignored.

BACKGROUND and COMMENTARY have different behaviors in English reviews compared to French documents: BACKGROUND seems to be more local in French documents whereas COMMENTARY tends to be more local in English reviews. On the other hand, the following relations often have CDUs in at least one of their arguments: ELABORATION, EXPLANATION, FRAME, RESULT, CONTRAST, CORRECTION, NARRATION and COMMENTARY. For example, CORRECTION concerns CDUs in most of 55% of cases. This relation links segments sharing a common topic and such that the second argument corrects

the information given in the first argument (which is often at a long distance attachment) (see the Correction in Example (16)). Another interesting behavior comes from the Contrast relation. Contrary to our expectations, only 40% of instances of this relation link EDUs in all the corpora. Example (17) illustrates a Contrast with scope over two CDUs.

(15) [One of the worst movies ever !]$_1$ [It's just terrible !]$_2$
Explanation(1,2)

(16) [The day before,]$_1$ [I went to see this movie,]$_2$ [I thought]$_3$ [I knew]$_4$ [what awesome was,]$_5$ [but I was so wrong.]$_6$
Frame(1,2)
Background([1,2],[3,4,5,6])
Continuation(3,4)
Attribution([3,4],5)
Correction([3,4,5],6)

(17) [The dialogue is stodgy]$_1$ [and the drama slows the pace,]$_2$ [but the violent action]$_3$ [and the imaginative look make it fun to watch.]$_4$
Continuation(1,2)
Continuation(3,4)
Contrast([1,2],[3,4])

### 5.2 Quantitative analysis at the segment and opinion expression level

The total number of annotated segments was 3,825 for *FMR*, 2,071 for *FNR* and 2,578 for *EMR*. The histogram in Figure 11 gives a comparative analysis of how segments are distributed over the five classes (i.e., *SE* (explicit opinion), *SI* (implicit opinion), *O* (objective), *SN* (subjective non evaluative) and *SEI* (explicit and implicit segment)). A similar analysis is given in Figure 12, this time for segment polarity (i.e., *positive, negative, neutral, no polarity* and *both*). The frequencies of each segment opinion type and each segment polarity type across corpus genres are statistically different from what is expected by chance using the $\chi^2$ test.

We observed that the frequencies of the segments containing implicit opinions (*SI*) depend on the corpus genre: for *FMR* and *EMR*, frequencies are less important (respectively 26.5% and 24.5%) compared to *FNR* (47.1%). Moreover, in the three corpora, the purely objective segments are not very widespread (less than 20% of all segments). The same goes for segments that contain at the same time an explicit and an implicit opinion (*SEI*), with a yet lower frequency for *ENR*. As for the subjective non-evaluative segments (*SN*), they are rather infrequent as well, especially in French and English movie reviews. However, they are slightly more numerous for *FNR*, which shows that the reported speech constructions are more frequent in reactions to newspaper articles than in movie reviews. Another interesting genre bias concerns the polarity of the segments: whereas in French movie reviews positive segments are a majority in spite of balancing the corpus between overall positive and overall negative documents (in terms of their star counting), this is not the case for the reactions to newspaper articles, where negative segments are a majority. In *EMR* however, segment polarity distribution is more balanced than for *FMR*. We also observe that non evaluative segments (mainly from the objective and the subjective non evaluative segment type) are more numerous in

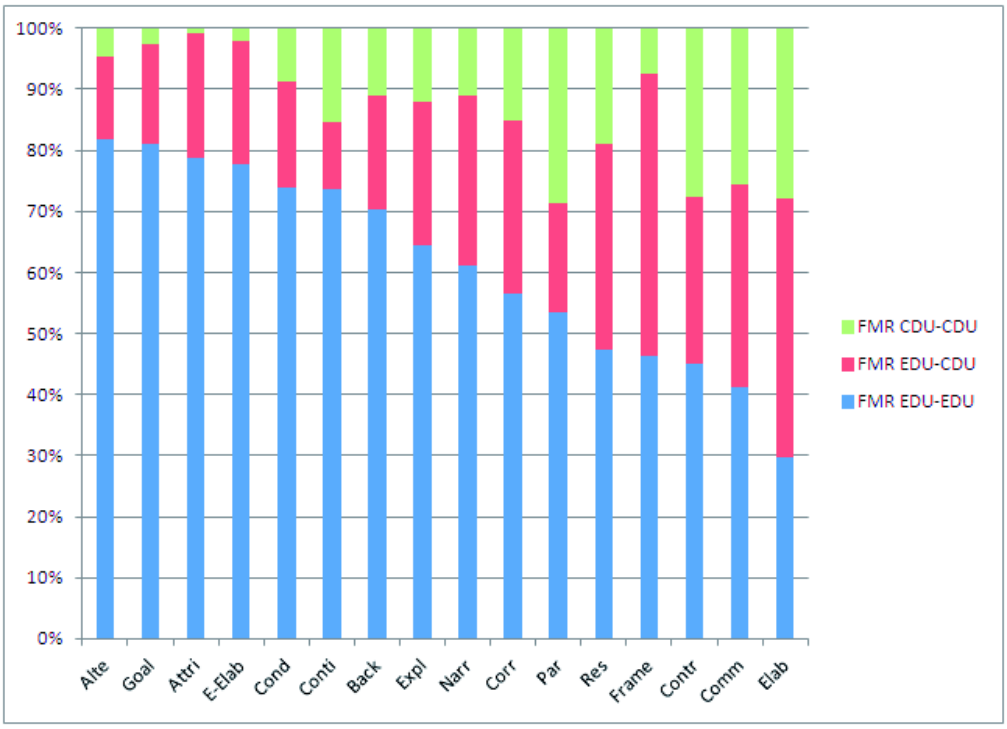


Figure 8: The distributions of discourse relations in *FMR* according to the type of their arguments.

English reviews than in French reviews. Finally, the proportion of *both* and *neutral* are a minority in all the corpora (respectively less than 3% and 2%). The last segments in Examples (18) and (19) respectively illustrate segments from the *both* and *neutral* category.

(18) [I am very torn about this film,] [as I think] [it contains some really bad directing by a great director.]

(19) [As some one commented already] [it is a combo of ”Black Beauty” and ”All Quiet on the Western Front.”]

Within evaluative segments (i.e., *SE, SEI* and *SN*), 2,329 opinion expressions were annotated for *FMR*, 743 for *FNR* and 1,610 for *EMR*. Among explicit segments (i.e., *SE* and *SEI*), 97% contain a single opinion expression for *FMR* and *EMR* vs. 94% for *FNR*. This confirms the usefulness of the per-segment analysis since this simplifies opinion fusion with respect to a per-sentence analysis for instance. We further discuss this important result in Section 6.

The semantic categories of opinion expressions are similarly distributed for *FMR* and *EMR* with around 3% for *Advice*, and between 5 and 8% for *Reporting*. However, we observe that in English movie reviews, most opinion expressions are from the *Sentiment-Appreciation* category (48.2% vs. 24.2% for French) while, in *FMR*, opinion expressions are mostly judgments and evaluations (66.4% vs. 36.4% for English). As expected, we get different distributions of semantic categories for *FNR*, with a greater number of *Reporting* (27.5%) and *Advice* expressions (6.9%) and no instances of the *Sentiment-Appreciation* category.

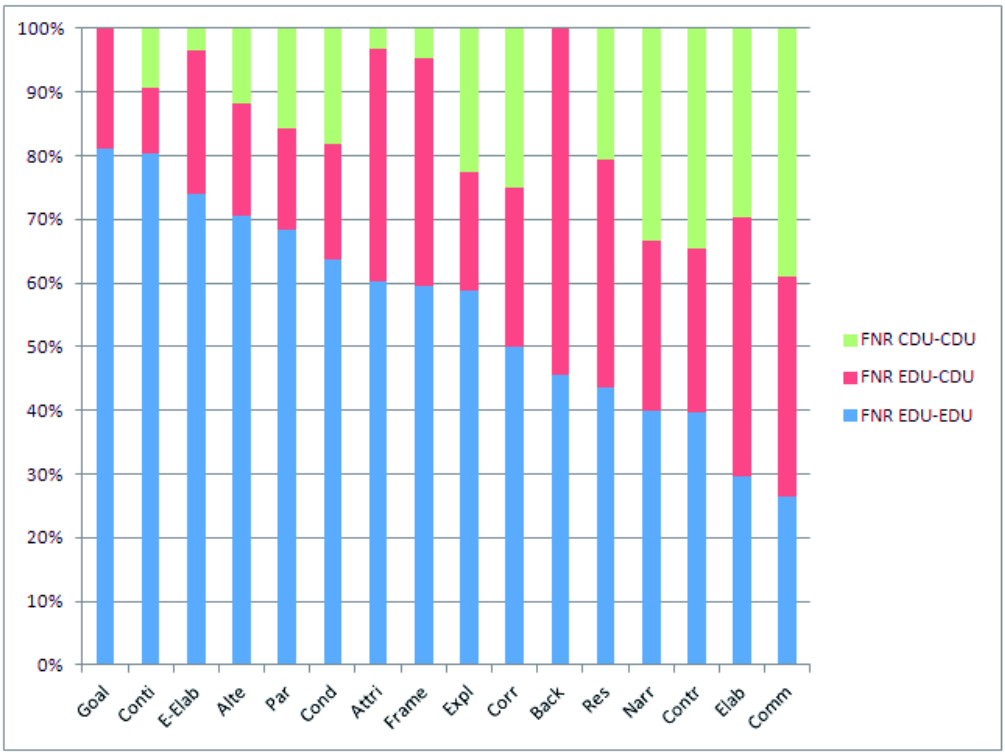


Figure 9: The distributions of discourse relations in *FNR* according to the type of their arguments.

Concerning the annotations of topics and holders, the total number was respectively: 2,939 and 754 for *FMR*, 1,915 and 262 for *FNR*, and 1,981 and 499 for *EMR*. For movie reviews, topics are mainly from the *part of* category (around 60%) whereas few of them are out of topic (*other*) (around 10%). However in *FNR*, we observe a different distribution: the number of topics from the *main* category are lower (around 9%) whereas the number of *other* topic are greater (around 19.4%). For the holders, we get similar distributions over all the corpora: 2/3 of annotated holders are from the *main* category.

Lastly, we also noticed the importance of opinion operators: 1,371 for *FMR*, 924 for *EMR* and 488 for *FNR*. At least one such operator is present in 32% of subjective segments in news reactions vs. 40% for movie reviews. These operators are also present in implicit segments (18% for the French corpus vs. 25% for the English documents and 17% for news reactions) which indicates that valence shifter terms are good cues for detecting implicit opinions. The distribution of operators per category is shown in Figure 13. Most of them are intensifiers. Restrictors are from different types: they can be temporal (as *some* in *[Some scenes are beautifully shot]* and *at times* in [*It can be entertaining at times*]) or topic restrictions as in [*This movie is made for 10 year old kids.*].

In our previous work on using discourse in sentiment analysis, we have annotated opinion semantic categories at the segment level in movie reviews and letters to the editor in English and French. Our past results, reported in (Asher et al. (2008)), showed that the distribution of semantic categories in these corpora are comparable to those observed in the corpora annotated in this current

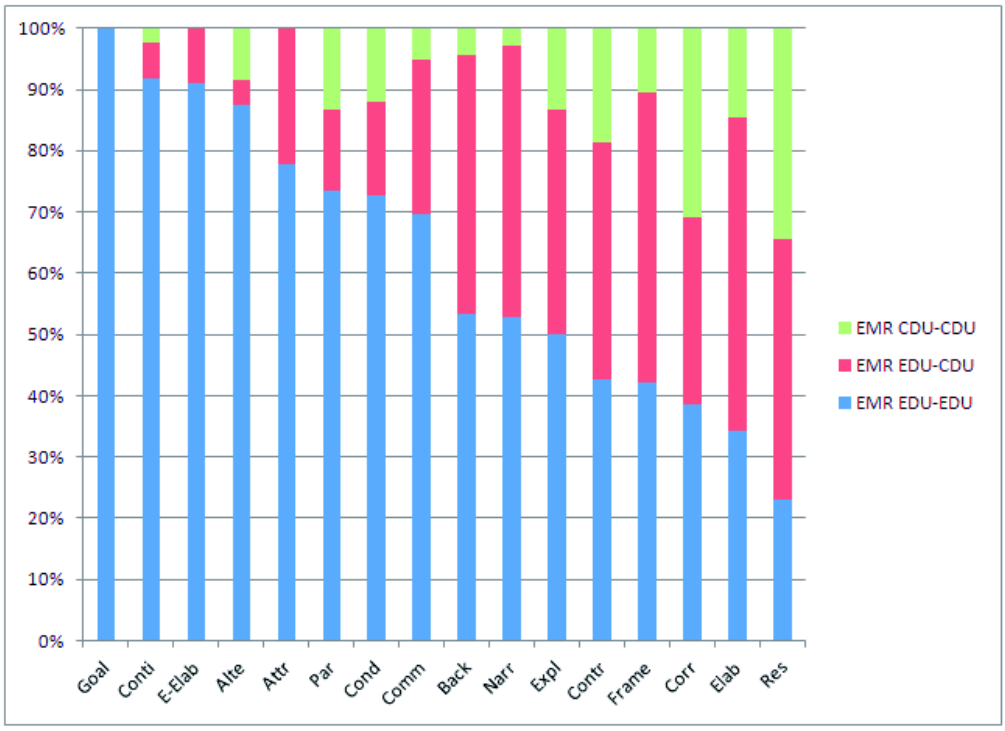


Figure 10: The distributions of discourse relations in *EMR* according to the type of their arguments.

study. As far as the semantic categories are concerned, we can conclude that our observations are valid to French and English movie reviews and news reactions in general. We believe that our results on segment polarity and segment type can also be generalized. More annotations are however needed to validate this assertion.

### 5.3 Impact of discourse on sentiment analysis

In this section, we attempt to answer the challenges mentioned in the introduction of this paper: *What is the role of discourse relations in subjectivity analysis? What is the impact of the discourse structure in determining the overall opinion conveyed by a document? Does a discourse based approach really bring additional value compared to a classical bag of words approach? Does this additional value depend on corpus genre?* To this end, we explored the interactions between the discourse, the segment, and the opinion expression annotation layer. In particular,

- Section 5.3.1 investigates the correlation between discourse and opinion semantic category of subjective segments (mainly from the *SE, SEI* and the *SN* category). Recall that an opinion expression can belong to four semantic categories, namely: *Sentiment-appreciation, Judgment, Advice* and *Reporting*. Our aim is to analyze to what extent semantic categories of opinion expressions can be an indicator for predicting discourse relations.

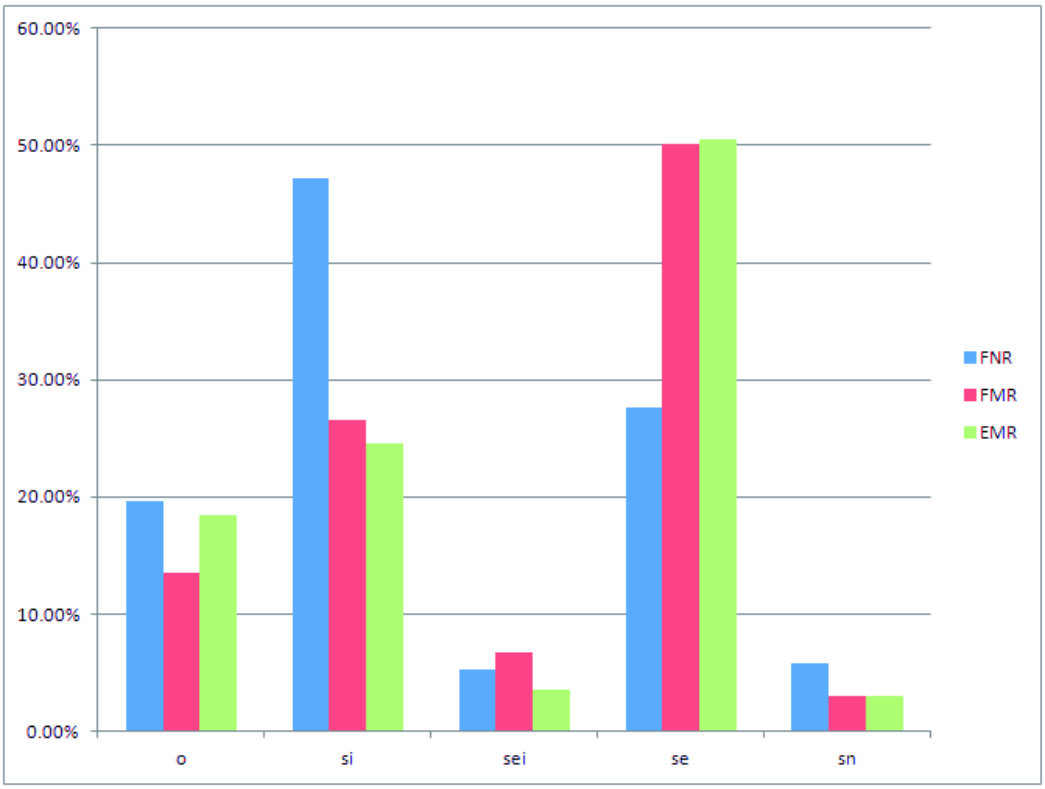


Figure 11: Frequencies of segments per opinion type.

- Section 5.3.2 focuses on the impact of discourse on subjectivity analysis. Can discourse relations be used to predict subjectivity orientation of elementary discourse units?
- Section 5.3.3 analyzes the impact of discourse on polarity analysis. Can discourse relations be used to predict polarity of elementary discourse units?
- Section 5.3.4 studies the impact of segment opinion type and segment polarity on the determination of the document overall opinion. Do segments with implicit opinions contribute to the author's global opinion on the main topic of the document?

This section details experiment aspect addressing each of these challenges while Section 6 summarizes the conclusions answering these questions.

### 5.3.1 DISCOURSE AND OPINION SEMANTIC CATEGORIES

We tested two hypotheses: (H1) there is an association between the relative position of segments within the document and the semantic category of the opinion expressions they contain. If a correlation is found, then the position can be used for example to identify the semantic category of segments conveying implicit opinions. (H2) there is an association between discourse relations and the semantic categories of the opinion expressions that appear within the relation arguments.

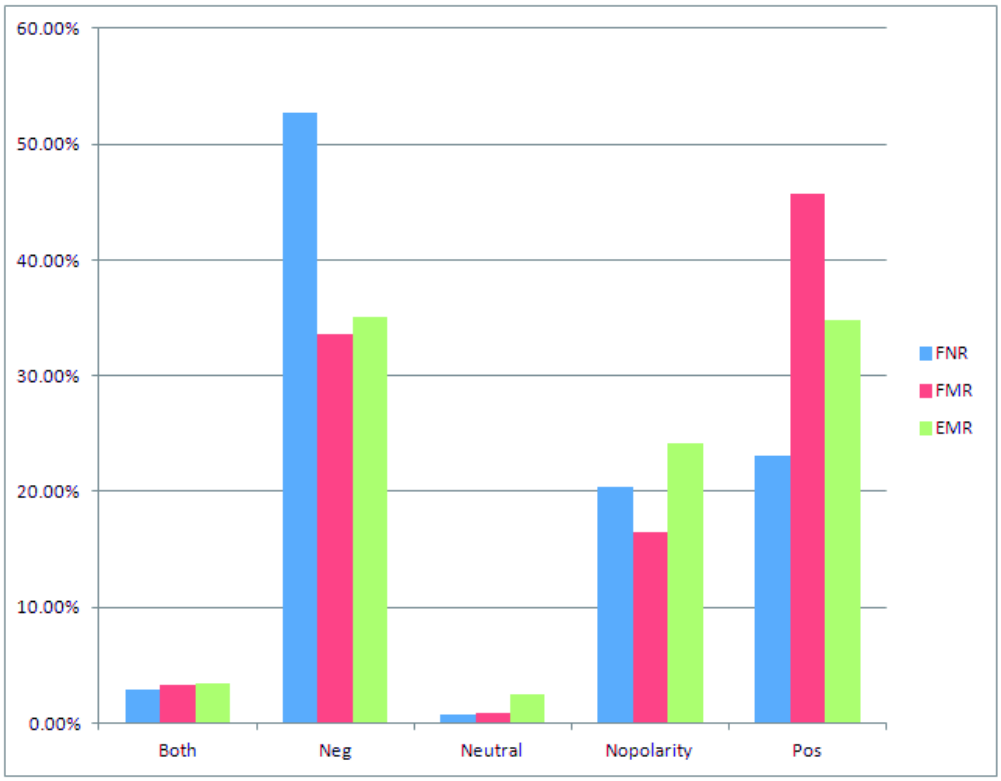


Figure 12: Frequencies of segments per polarity type.

**Position of segments vs. semantic categories.** Table 7 gives the proportions (in percent) of opinion semantic categories according to the relative position of the segment they belong to. We considered two positions: beginning and end of the document. To compute them, we simply divided a document into 3 parts (beginning, middle, end). The first two segments being the beginning while the last two the end. In the Table 7, the configurations Begin-x (resp. End-x) stand for segments containing an opinion expression from an x category.

When using the $\chi^2$ test, the hypothesis (H1) is confirmed at $p < 0.05$. We see that the proportion of the *Advice* category is higher when expressions of this type appear in segments at the end of the document. The proportion of the other categories is relatively stable. This increase is more impressive in reviews (more than 10%) than in news reaction (around 5%) which confirms that users in reviews tend to end their reviews by expressions of recommendations, hopes, or suggestions.

**Discourse relations vs. semantic categories of their arguments.** For each corpus, we constructed three contingency tables:

- (T1) gives the number of discourse relations that have a right argument containing an opinion expression from a given semantic category. For each discourse relation $R$ and for each semantic category $c \in \{Sentiment - Appreciation, Judgment, Advice, Reporting\}$, we counted all the pairs $R(se\_c, all)$ where $se\_c$ is an *SE* segment containing an opinion expression from a category $c$ and *all* stands for an EDU whatever its type (i.e., *SE*, *SEI*, *O*, *SN* or *SI*).

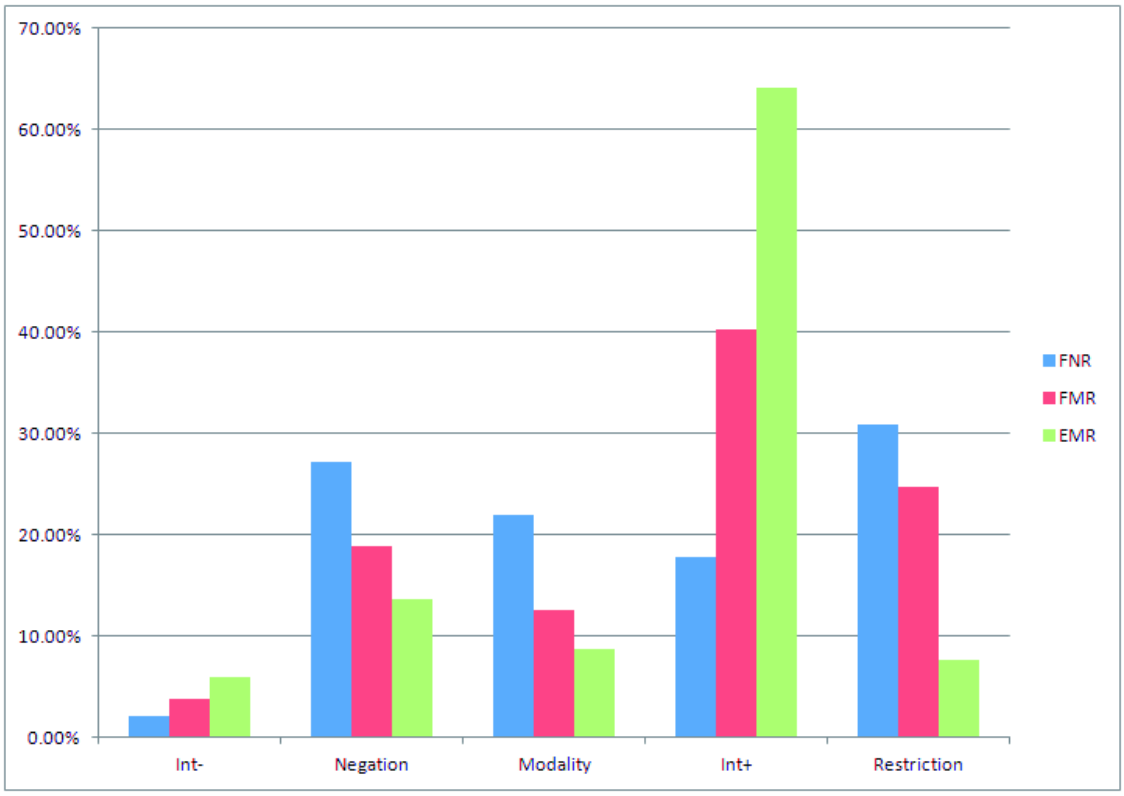


Figure 13: The distribution of operator per category.

| | *EMR* | *FMR* | *FNR* |
|---|---|---|---|
| Begin-Reporting | 0.10 | 0.06 | 0.32 |
| Begin-Judgment | 0.46 | 0.64 | 0.60 |
| Begin-Sentiment-appreciation | 0.43 | 0.29 | 0.00 |
| Begin-Advice | 0.01 | 0.01 | 0.08 |
| End-Reporting | 0.09 | 0.02 | 0.22 |
| End-Judgment | 0.38 | 0.58 | 0.65 |
| End-Sentiment-appreciation | 0.43 | 0.26 | 0.00 |
| End-Advice | 0.10 | 0.14 | 0.13 |

Table 7: Proportions (in percent) of opinion semantic categories according to the relative position of the segment they belong to.

- In Table (T2), we do the same by counting all the pairs $R(all, se\_c)$.
- Table (T3) provides the frequencies for each relation $R$ and the frequencies of $R(se\_c, se\_c)$.

Tables 8, 9, and 10 give respectively the results of (T1), (T2) and (T3) for the French movie reviews corpus. The tables associated to the other two corpora looked similar.

| | Advice_Right | SentiApp_Right | Reporting_Right | Judgment_Right |
|---|---|---|---|---|
| Elaboration | 3 | 49 | 13 | 170 |
| Attribution | 5 | 16 | 15 | 27 |
| Goal | 0 | 3 | 0 | 7 |
| Continuation | 10 | 164 | 16 | 511 |
| Frame | 1 | 7 | 9 | 13 |
| Conditional | 2 | 4 | 0 | 2 |
| E-Elab | 2 | 35 | 9 | 89 |
| Parallel | 1 | 5 | 0 | 13 |
| Explanation | 2 | 44 | 14 | 134 |
| Result | 16 | 82 | 18 | 102 |
| Background | 1 | 4 | 1 | 7 |
| Narration | 0 | 5 | 1 | 9 |
| Commentary | 24 | 95 | 21 | 175 |
| Alternative | 2 | 2 | 2 | 4 |
| Correction | 1 | 4 | 3 | 17 |
| Contrast | 5 | 47 | 14 | 136 |

Table 8: Frequency of discourse relations that have a right argument containing an opinion expression from a given semantic category.

| | Advice_Left | SentiApp_Left | Reporting_Left | Judgment_Left |
|---|---|---|---|---|
| Elaboration | 5 | 78 | 18 | 184 |
| Attribution | 9 | 13 | 47 | 15 |
| Goal | 0 | 5 | 0 | 9 |
| Continuation | 12 | 155 | 26 | 509 |
| Frame | 0 | 2 | 1 | 7 |
| Conditional | 0 | 3 | 0 | 1 |
| E-Elab | 4 | 34 | 8 | 99 |
| Parallel | 0 | 6 | 2 | 15 |
| Explanation | 0 | 46 | 26 | 110 |
| Result | 20 | 64 | 8 | 138 |
| Background | 0 | 6 | 2 | 0 |
| Narration | 0 | 5 | 2 | 12 |
| Commentary | 4 | 76 | 14 | 193 |
| Alternative | 2 | 3 | 0 | 7 |
| Correction | 2 | 7 | 1 | 23 |
| Contrast | 6 | 41 | 15 | 153 |

Table 9: Frequency of discourse relations that have a left argument containing an opinion expression from a given semantic category.

Given the frequencies in these tables, the hypothesis (H2) was rejected using the $\chi^2$ test. For each corpus genre, there is no statistically significant relationship between discourse relations and the opinion category of their arguments. However, in the French corpora, after removing the relations GOAL, CONDITIONAL, FRAME, BACKGROUND and ATTRIBUTION from the contingency table (T1), the

| | Advice_Same | SentiApp_Same | Reporting_Same | Judgment_Same |
|---|---|---|---|---|
| Elaboration | 0 | 3 | 0 | 10 |
| Attribution | 1 | 0 | 0 | 0 |
| Goal | 0 | 0 | 0 | 2 |
| Continuation | 4 | 25 | 3 | 134 |
| Frame | 0 | 0 | 0 | 1 |
| Conditional | 0 | 1 | 0 | 0 |
| E-Elab | 0 | 6 | 0 | 25 |
| Parallel | 0 | 3 | 0 | 7 |
| Explanation | 0 | 6 | 6 | 34 |
| Result | 0 | 6 | 0 | 16 |
| Background | 0 | 0 | 0 | 0 |
| Narration | 0 | 1 | 0 | 4 |
| Commentary | 0 | 7 | 0 | 17 |
| Alternative | 0 | 1 | 0 | 3 |
| Correction | 0 | 0 | 0 | 4 |
| Contrast | 0 | 6 | 0 | 47 |

Table 10: Frequency of discourse relations that have arguments containing opinion expressions from the same semantic category.

association between discourse relations and opinion category of right arguments was significant at $p < 0.05$ using the $\chi^2$ test[34]. For *EMR*, the association is significant when removing the same set of relations as above and when discarding, in addition, the categories *Advice* and *Reporting*. In (T2), the association between discourse relations and left arguments was significant when removing the *Advice* category and the same set of relations as above except ATTRIBUTION. Finally, for (T3), we get a statistically significant association when removing both the same set of relations as above and the categories *Advice* and *Reporting*.

Overall, the absence of a strong correlation between discourse relations and opinion categories can be due to the categories themselves that were not adequate to capture that relations well. To confirm or reject hypothesis (H2), it would be interesting to conduct a similar study using different categories.

Concerning the distribution of relations with regard to the opinion semantic category, the proportion of ATTRIBUTION relations is relatively high when the first argument of this relation is from the *Reporting* category. We also have instances from CONTINUATION and ELABORATION. Similarly, the proportion of RESULT is high when its second argument contains an *Advice* expression. Examples like (20) are very frequent in our reviews corpora (here segments 4 and 5 contain explicit recommendations to see the movie and they are related to the first part of the document by a RESULT relation):

34. Note that the $\chi^2$ test cannot be computed if some frequencies are less than 5. To overcome this problem, some relations that have similar semantic effects on opinion were grouped, like CONTRAST with CORRECTION, CONTINUATION, PARALLEL with ALTERNATIVE, etc.

(20) [It it is the best adventure movie of our time]$_1$ [and whats in bonus its an awesome joy-full adventure for all ages.]$_2$ [Its a full family entertainer.]$_3$ [So go]$_4$ [and watch the movie]$_5$ [and uncover the secret of Hugo Cabret.]$_6$

On the other hand, several *Advice* expressions in the *EMR* corpus are related with Conditional, like in (21) where the author recommends the movie under certain conditions:

(21) [If you're after a film]$_1$ [that doesn't employ too much thinking]$_2$ [and is enjoyable to watch]$_3$ [I would recommend going to see this.]$_4$

Finally, *Advice* can also come under a Commentary, as in the news reaction in (22):

(22) [Quand on en est à emprunter de l'argent frais]$_1$ [pour payer les intérêts des emprunts précédents,]$_2$ [c'est que le mur se rapproche.]$_3$ [Un conseil :]$_4$ [achetez de l'or...]$_5$
[When we are borrowing money]$_1$ [to pay past loan interest,]$_2$ [it means that the wall is approaching.]$_3$ [An advice:]$_4$ [buy gold...]$_5$

### 5.3.2 Discourse relations and subjectivity analysis

We also assessed, for each relation instance linking EDUs only (except for Unknown and Flashback since they have the lowest frequencies), whether they preserve subjectivity or not. We computed statistics on the stability of the subjectivity class (for the (*SE*, *SE*), (*SI*, *SI*) and (*SI*, *SE*) pairs) or the variation of the stability class (for the (*O*, other) pairs, where "other" spans the set of subjectivity classes, other than *O*). Figures 14, 15 and 16 summarize our results. The relations in these figures are sorted according to the decreasing frequencies of subjectivity preservation. The subjectivity preservation frequencies of each discourse relation across corpus genres are statistically different from what is expected by chance using the $\chi^2$ test. Note however that the difference between the observed and the expected frequencies of Attribution and Conditional were not significant.

We observe that our predictions (as stated in the introduction) are by and large confirmed. Some relations preserve subjectivity in all corpora (with more than 70% of instances): Continuation, Parallel, Alternative, Contrast, Elaboration, Explanation, Commentary, Result, Narration, and Contrast. For some relations, the preservation is more salient for reviews than for news reaction. For instance, Commentary preserves subjectivity in 80% of cases in *FNR* vs. between 60 and 72% for reviews where examples like (23) are less frequent. Result however gets a different distribution, with more than 80% preservation in reviews vs. 70% in reactions.

(23) [J'ai découvert la vie de Piaf,]$_1$ [on a l'impression d'être avec elle tout le long du film.]$_2$
[I discovered Piaf's life,]$_1$ [I felt that I was with her all along the movie]$_2$

Other relations do not preserve subjectivity across our corpora: Background, Attribution and Frame. In news reaction, Attribution preserves subjectivity in 50% of cases whereas in reviews the proportion is about 20%. This might be because examples like [*The chairman thought*] [*that it rained in his town yesterday*] are more frequent in the first corpus genre (movie reviews) than in the second (news reactions) where attributions are more often used to introduce opinions and point of views. Subjectivity preservation in the case of Frame is about 40% in French document vs. 87% in English reviews because in French corpora, this relation often relates non evaluative segments to evaluative ones. Correction seems to preserve subjectivity in reviews (60% in English

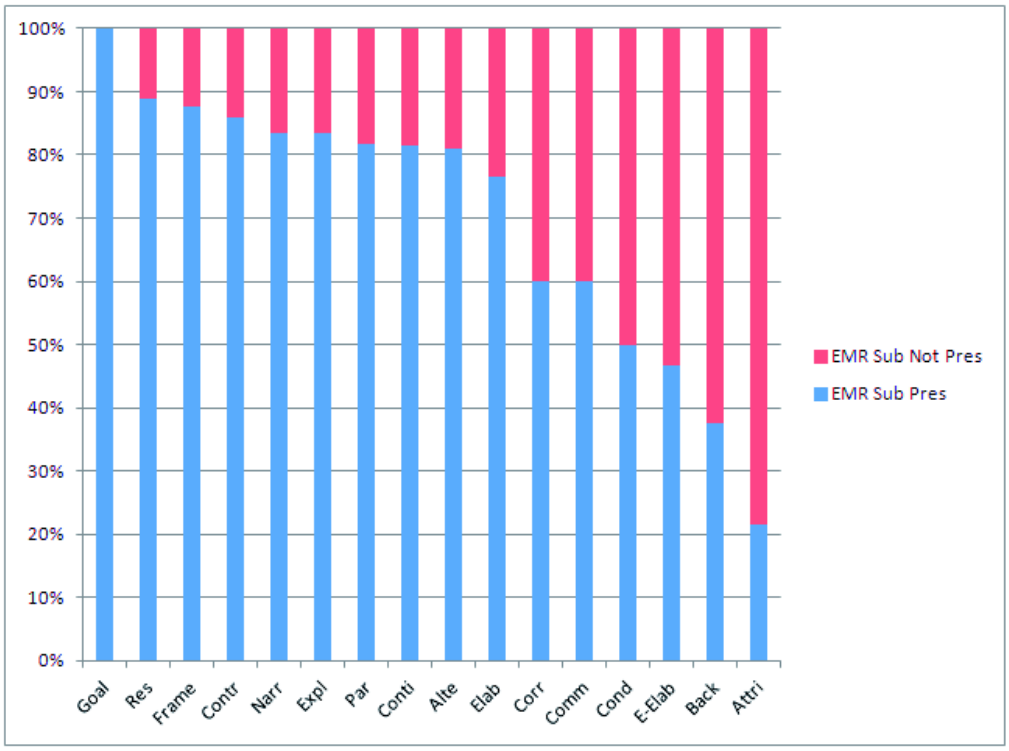


Figure 14: Discourse relations and subjectivity in *EMR*.

reviews and 83% in French reviews) but not in news reactions where the proportion is about 50%. We observe the contrary for Conditional and Entity-Elaboration where subjectivity preservation is more frequent in news reactions. Indeed, in *FMR*, consequences are often objective even when their corresponding conditions are evaluative as shown in (24).

(24) [c'est long,]$_1$ [froid,]$_2$ [pas bon.]$_3$ [si vous y allez une fois]$_4$ [ce sera bien la seule]$_5$.
[It's long,]$_1$ [cold,]$_2$ [not good.]$_3$ [if you go once]$_4$ [it will be the only time]$_5$.

Continuation(1,2)
Continuation(2,3)
Result([1,2,3],[4,5])
Conditional(4,5)

### 5.3.3 Discourse relations and polarity analysis

We finally computed similar statistics for the polarities, but between subjective (*SN*, *SE*, *SEI*, *SI*) EDUs only: the (+, +) and (–, –) for stability and (+, –) for polarity change. We assess in Figures 17, 18 and 19 the behavior of our relations with respect to polarity preservation and non-preservation. Only relations preserving subjectivity are taken into account (Background, Attribution, Frame and Entity-Elaboration have been discarded[35]). They are presented by decreasing order of polarity

35. Relations that do not preserve subjectivity are necessarily relations that do not preserve polarity.

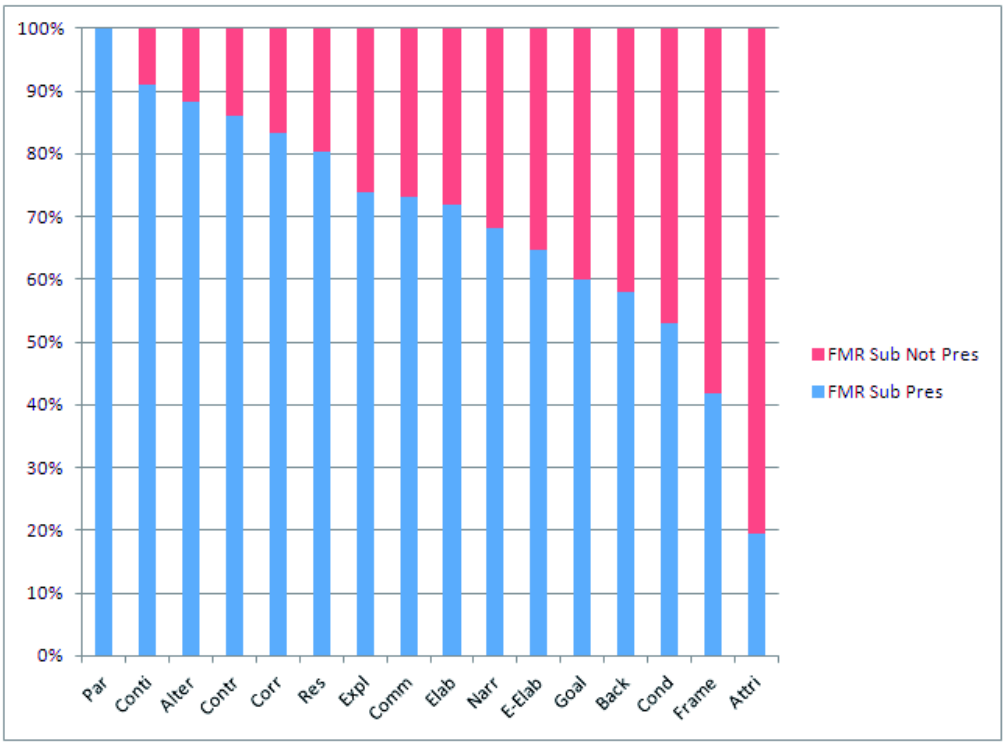


Figure 15: Discourse relations and subjectivity in *FMR*.

preservation frequencies. The polarity preservation frequencies of each discourse relation across corpus genres are statistically different from what is expected by chance using the $\chi^2$ test. Note however that the difference between the observed and the expected frequencies of Conditional, Correction and Contrast were not significant.

As far as polarity is concerned, our hypotheses seem by and large verified as well, for all corpora. However, contrary to expectations, Contrast seems to change polarity in reviews but not in news reactions. In reactions, this can be explained by examples of the type: *[The economical situation is grim,] [but the cultural life is grim as well]* where there is the *but* connective linking the two segments, which makes the annotators place a Contrast between the two segments. However, in this particular case it would be more appropriate to link the two segments by the Parallel relation or with both Parallel and Contrast[36], which is possible in SDRT and provides the right semantics for such relations (Asher (1993)). Note however that the frequencies of Contrast and Correction in all the corpora were not significant. We need more annotations to establish the relationships between these relations and polarity analysis.

### 5.3.4 Segment type, segment polarity, and overall opinion

We investigated whether implicit opinion segments contribute to the author's global opinion on the main topic of the document. We have computed the Pearson's correlation between the global

36. When preparing the gold standard, we reconsidered the relation labels only in 5% of the cases.

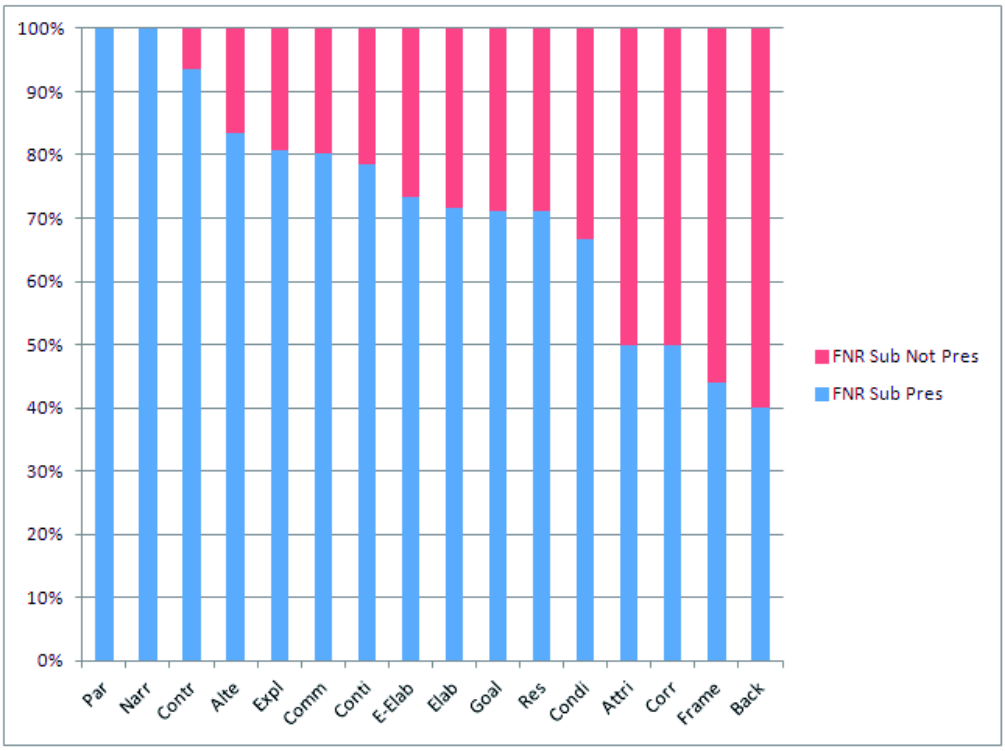


Figure 16: Discourse relations and subjectivity in *FNR*.

opinion score (on a scale going from 0 for a strongly negative opinion, to 4 for a strongly positive opinion) and the subjectivity class and polarity of the segments. More specifically, for each of the three corpora, we have constructed a vector with the global opinion scores for all the annotated document instances[37]. Then, another set of four vectors has been built for each corpus, with the counts of segments of a given subjectivity class and polarity: SE_Pos for explicit positive opinion segments (*SE* and *SEI*) class with a positive polarity; SE_Neg for explicit negative opinion segments; SI_Pos for implicit positive opinion segments (*SI* class with positive polarity); and SI_Neg for implicit negative opinion segments. Similarly, we have computed the correlation between the overall opinion and segment polarity regardless of their types: All_Pos for positive segments and All_Neg for negative segments. In addition, we have measured the correlation between the overall opinion vector and the average segments scores (given between −3 and +3) of each document (All_Avg). The results are shown in Table 11, averaged over all the annotators.

In movie reviews (*FMR* and *EMR*) there is a better correlation between global opinion score and explicit subjective segment counts (of both positive and negative polarities – for negative polarities, a good correlation means a negative Pearson's correlation of high absolute value) than between global opinion score and implicit subjective segment counts. In *FNR*, a different behavior is observed: the correlation is better for segments which contain implicit opinions. This brings us to the conclusion that the importance of implicit opinions varies, depending on the corpus genre: in

37. If one input document has been doubly annotated, we thus obtained two document annotation instances.

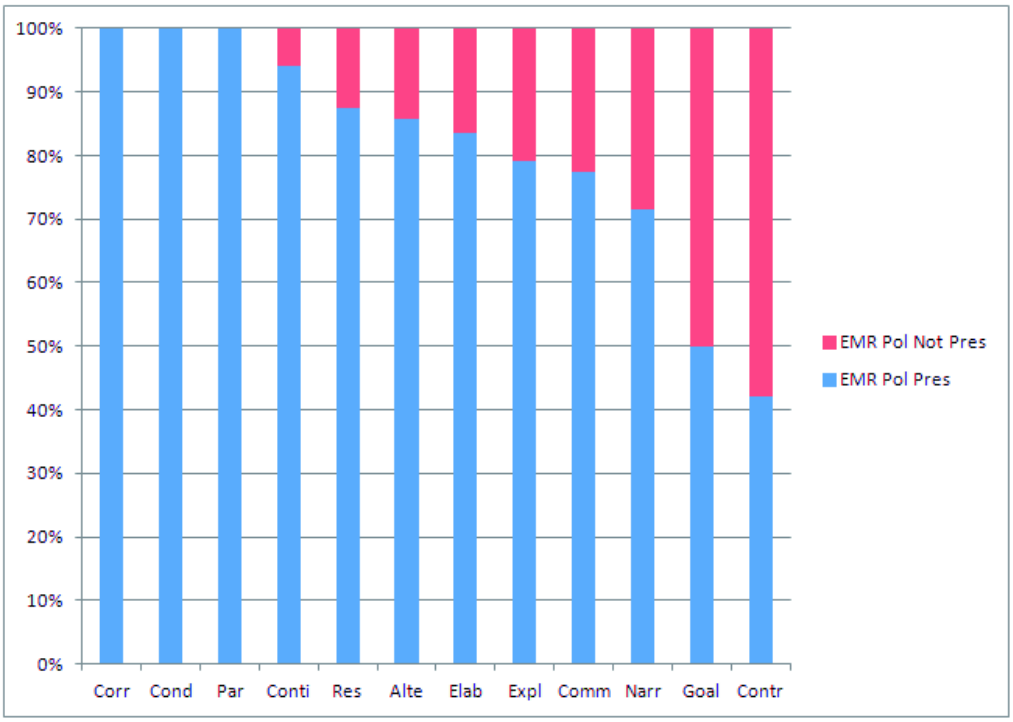


Figure 17: Discourse relations and polarity in *EMR*.

| | *FMR* | *EMR* | *FNR* |
|---|---|---|---|
| SE_Pos | 0.54 | 0.72 | 0.3 |
| SE_Neg | -0.64 | -0.74 | -0.19 |
| SI_Pos | 0.42 | 0.59 | 0.42 |
| SI_Neg | -0.45 | -0.67 | -0.40 |
| All_Pos | 0.64 | 0.61 | 0.52 |
| All_Neg | -0.6 | -0.57 | -0.38 |
| Avg_All | 0.19 | 0.17 | 0.10 |

Table 11: Correlations between overall opinion and segment opinion type/polarity.

movie review, more direct and sometimes terse, explicit opinions are better correlated to the global opinion score, whereas in news reactions, implicit opinions are more important. This could indicate a tendency to "conceal" negative opinions as apparently objective statements, which can be related to social conventions (politeness, in particular) (Pang and Lee (2008)). Now, when we have grouped segments by polarity (cf. All_Pos and All_Neg), we observe that the correlation with positive segments are better compared to those with negative polarity. The *politeness bias* is more salient in news reactions than in movie reviews where users tend to express their opinions in a more positive way. Finally, we see that correlations in All_Avg are the lowest, which confirms that overall opinions is not only a simple aggregation of opinions taken in isolation. A more elaborated way of aggregation is needed.

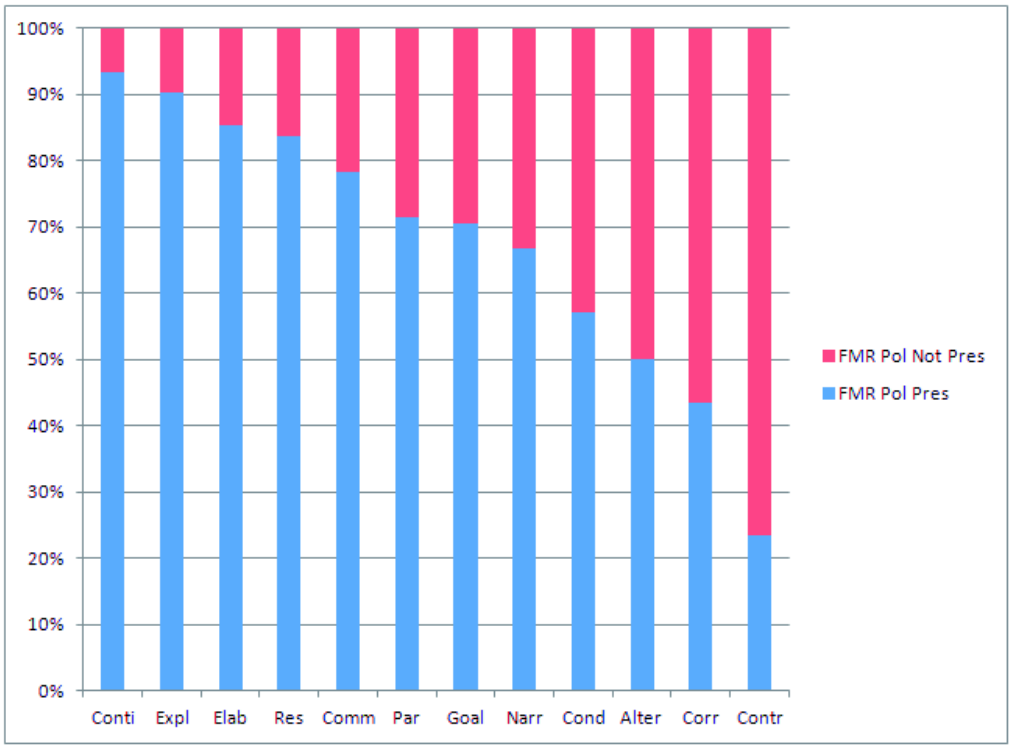


Figure 18: Discourse relations and polarity in *FMR*.

## 6. Discussions

### 6.1 Interim conclusions

In this paper, we aimed at measuring the impact of discourse on sentiment analysis with a study of three corpora: French and English movie reviews as well as French news reactions. Here are the main conclusions of our corpus-based study:

(a) *Segment-based opinion analysis is more appropriate to study opinions in discourse.* Our results showed that more than 90% of segments contain only one opinion expression. This demonstrates that the segment level will make polarity analysis easier compared to the sentence or the clause level. In addition, our automatic discourse segmentation is feasible and yielded very good results.

(b) *Complex discourse units (CDUs) are an important part of the discourse structure of a document.* In the whole corpora, our results showed that the proportion of relations involving CDUs is higher compared to the proportion of relations linking EDUs. In particular, we observed that the arguments of the relations Contrast, Elaboration, and Result are CDUs in more than 55% of cases. CDUs related with a Frame are more frequent in movie reviews (more than 57%) whereas those related with a Commentary are more frequent in the French corpora (more than 64%). These results demonstrate

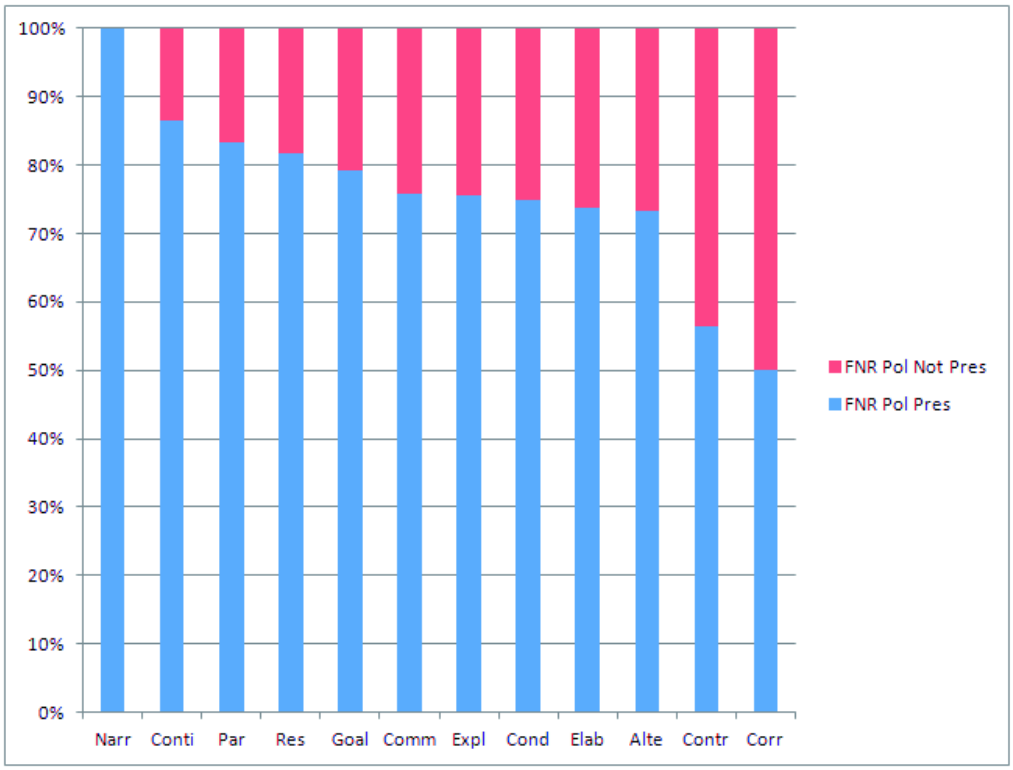


Figure 19: Discourse relations and polarity in *FNR*.

that CDUs are important for assessing the overall opinion of a document.

(c) *Implicit opinions are important*. Our results showed that the importance of implicit opinions varies, depending on the corpus genre: for movie reviews, explicit opinions are better correlated to the global opinion score, whereas for news reactions, implicit opinions are more important when negative opinions are concerned.

*(d) Semantic categories of opinion expressions can be good indicators for identifying some discourse relations.* Indeed, we observed that the discourse relations CONTRAST, CONTINUATION, NARRATION, ALTERNATIVE, RESULT, PARALLEL, ELABORATION, ENTITY-ELAB, CORRECTION, EXPLANATION, and COMMENTARY are correlated with the semantic categories (*Reporting, Judgment, Advice*, and *Sentiment-appreciation*) of the opinion expression within their arguments.

*(e) Discourse relations can be grouped according to their effects on the opinion orientation of elementary discourse units.* We studied 17 discourse relations that involve entities from the propositional content of the clauses: 9 coordinating relations (CONTRAST, CONTINUATION, CONDITIONAL, NARRATION, ALTERNATIVE, GOAL, RESULT, PARALLEL, FLASHBACK) and 8 subordinating relations (ELABORATION, E-ELAB, CORRECTION, FRAME, EXPLANATION, BACKGROUND, COMMENTARY, ATTRIBUTION. Among these relations, some can be grouped according to their similar effects on both subjectivity and po-

larity analysis: Correction and Contrast, Elaboration and Explanation, Continuation, Parallel, Narration, and Alternative. Table 12 summarizes the effects of these relations. For a given relation (or group of relations), "√" (resp. "*X*") indicates that the relation preserves (resp. does not preserve) subjectivity (resp. polarity) in more than 75% of cases in at least two corpora. This table shows that some relations have no effect at all on sentiment analysis: Frame, Goal, Background, Conditional and Flashback while others impact on subjectivity analysis, on polarity analysis or influence both these two tasks. These results confirm that discourse relations can help in identifying segments conveying implicit opinions or retrieving segment contextual polarity which, for instance can be very useful in identifying ironic statements.

| Discourse Relation | Frequency in at least two corpora | Impact on sentiment analysis | |
|---|---|---|---|
| | | Subjectivity analysis | Polarity analysis |
| Parallel | ≤ 5% and ≥ 1% | √ | √ |
| Alternative | ≤ 5% and ≥ 1% | | |
| Continuation | ≥ 5% | | |
| Narration | ≤ 5% and ≥ 1% | | |
| Explanation | ≤ 5% and ≥ 1% | √ | √ |
| Elaboration | ≥ 5% | | |
| Commentary | ≥ 5% | *X* | √ |
| Contrast | ≥ 5% | √ | √ |
| Correction | ≤ 5% and ≥ 1% | | |
| E-Elab | ≥ 5% | *X* | √ |
| Result | ≥ 5% | √ | √ |
| Attribution | ≤ 5% and ≥ 1% | √ | *X* |
| Frame | ≤ 5% and ≥ 1% | *X* | *X* |
| Goal | ≤ 5% and ≥ 1% | *X* | *X* |
| Background | ≤ 5% and ≥ 1% | *X* | *X* |
| Conditional | ≤ 5% and ≥ 1% | *X* | *X* |
| Flashback | ≤ 1% | *X* | *X* |

Table 12: Discourse relations and sentiment analysis: interim conclusions.

### 6.2 Portability of the annotation scheme

The results reported in this study were obtained on manually annotated discourse structures when the annotation scheme was instantiated on two corpus genres: movie/product reviews and news reactions. These corpora have similar characteristics: they are texts and not discussions/dialogues (remember that letters to the editor that responded to other letters were removed from *FNR*), they are relatively small (less than 30 EDUs per document), opinions are about one main topic and its related subtopics and are the viewpoints of one holder (mainly the author of the review). More important, the overall opinion is the result of a bottom-up aggregation process, from local opinions at the segment level to the global opinion at the document level. However, several other corpus genres do not meet these characteristics. Some are *author-oriented* like blogs where all the documents (posts and comments) are associated to the blogs' owners, others are both *multi-topic* and *multi-holder* documents like news articles, while others are composed of *follow-up* opinions as in discussion forums. *To what extent is the CASOAR annotation scheme portable to these other sources of opinion?*

Concerning blogs, we believe that our scheme can be easily applied. Blog comments are generally short, they are the point of view of one author towards the main topic of the blog article which is quite similar to news reactions. For news documents, things are more complicated since several viewpoints by several opinion holders are mentioned. Consider the following scenario. The author introduces and elaborates on a topic, 'switches' to other topics or reverts back to an older topic. This is known as *discourse popping* where a change of topic is signaled by the fact that the new information does not attach to the prior clause, but rather to an earlier one that dominates it (Asher and Lascarides (2003)). In this case, our three-level annotation scheme needs to be adapted. Though the discourse annotation model incorporates discourse pops, their effects on topics for opinions is presently not taken into account. Discourse pops often indicate shifts in topic, and so, instead of one topic, we will have to deal with many. At the expression level, we have to take this multi-topicality into account, by modifying the annotation of topic spans. At the segment level, we would have to link each opinion expression to its topic. At the document level, the notion of overall opinion has to evolve towards *(topic, holder)* overall opinion scores. Each score can be computed using a bottom-up aggregation procedure over a discourse sub-graph focusing only on those segments that convey the opinions on a specific holder. This procedure needs however to be tested on news documents to show its feasibility.

Finally, adapting our scheme to discussion forums will require to us adapt our scheme to handle dialogues. A thorough linguistic analysis of the link between opinion and discourse in dialogue will be very interesting.

### 6.3 Towards discourse-based sentiment analysis

The CASOAR corpus is a first step towards automated discourse-based opinion analysis. We have already used a subset of this corpus in order to investigate how discourse can help in different sentiment analysis stages. In Benamara et al. (2011), we investigated how discursive features could improve subjectivity analysis. We automatically distinguished between subjective non-evaluative (*SN*) and objective segments (*O*) and between implicit (*SI*) and explicit opinions (*SE*), by using both local and global context features. Chardon et al. (2013) exploited the French gold standard corpus to determine what are the best strategies that need to be implemented to automatically compute a document overall opinion. Here we have made a complementary, in depth multi-lingual and multi-genre analysis of a new corpus study for English and provided new results concerning the French corpus.

A final issue is *how to validate our results on automatically parsed data.* Since review style documents are relatively short, we believe that building such a discourse parser becomes easier. As far as we know, the only existing powerful discourse parser based on SDRT is the one that has been developed on the top of the Annodis corpus (Muller et al. (2012)). This parser achieves between 47 and 66% accuracy on the structure for the full set of 17 relations. We plan to adapt this parser to opinion texts. In particular, given our observations (cf. Table 12), we propose to discard certain relations from the learning process and to group others according to their similar effect on both subjectivity and polarity analysis. This will reduce the number of relations to be predicted to 10 instead of 17 actually which, we believe, will make our discourse parser more reliable.

## 7. Conclusion

In this paper, we presented the CASOAR corpus, a multi-layered annotation scheme for analyzing opinion in discourse that includes: the complete discourse structure according to the Segmented Representation Discourse Theory, the opinion orientation of elementary discourse units and opinion expression annotation. For each layer, we presented the annotation model, annotation guide, and results of its annotation campaign. We explored the interactions between these different layers—in particular, the impact of discourse structure on the overall opinion of a document and implicit opinions, the link between discourse and opinion semantic category, and the role of discourse relations on both subjectivity and polarity analysis. Our results demonstrate that opinion and discourse structure are strongly related and that discourse is an important cue for sentiment analysis, at least for the corpus genres we have studied.

## Acknowledgments

This work was supported by a DGA-RAPID project under grant number 0102906143. We also thank our annotators: Simon Leva, Nicolas Nouno, Anny Soubeille, Lisa Petersen, and Julie Hall for their efforts during the annotation campaign. The authors also thank ERC grant 269427 for research support. We finally thank the editor and three anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

## References

Stergos Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. Learning recursive segments for discourse parsing. In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*, pages 3578–3584, 2010.

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paul Pery-Woodley, Laurent Prevot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 2012.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. Subjectivity word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–199, 2009.

Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht, 1993.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. Distilling opinion in discourse: A preliminary study. In *Proceedings of International Conference on Computational Linguistics (COLING)*, pages 7–10, 2008.

Nicholas Asher, Farah Benamara, and Yannick Mathieu. Appraisal of opinion expressions in discourse. *Linguisticae Investigationes 32:2*, 32(2):279–292, 2009.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual subjectivity: are more languages better. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 28–36, 2010.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, and Vladimir Popescu. Towards context-based subjectivity analysis. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1180–1188, 2011.

Yves. Bestgen, Cédrick. Fairon, and Laurent. Kevers. Un barométre affectif effectif. In *G. Purnelle, C. Fairon, and A. Dister (Eds.), Actes des septiéme Journées internationales d'Analyse Statistique des Données Textuelles*, pages 182–191, 2004.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, 2015.

John Blitzer, Dredze Mark, and Pereira Fernando. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447, 2007.

Ester Boldrini, Alexandra Balahur, Patricio Martnez-Barco, and Andrés Montoyo. Using emotiblog to annotate and analyse subjectivity in the new textual genres. In *Data Mining and Knowledge Discovery, Volume 25, Issue 3*, pages 603–634, 2012.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. In *IEEE Intelligent Systems, special issue on Knowledge-based approaches to content-level sentiment analysis. 28:2*, 2013.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.

Baptiste Chardon, Farah Benamara, Yvette Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. Measuring the effect of discourse structure on sentiment analysis. In *Proceedings of the Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 25–37, 2013.

Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 793–801, 2008.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. Mlsa. a multi-layered reference corpus for german sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 2012.

Béatrice Daille, Estelle Dubreil, Laura Monceaux, and Mathieu Vernier. Annotating opinion-evaluation of blogs : the blogoscopy corpus. In *Language Resources and Evaluation Springer, 45(4)*, pages 409–437, 2011.

Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1061–1070, 2011.

Alexander Hogenboom, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. Using rhetorical structure in sentiment analysis. *Commun. ACM*, 58(7):69–77, 2015.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, 2004.

Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Dordrecht, 1993.

Alistair Kennedy and Diana Inkpen. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

Jason S. Kessler, Miriam Eckert, Lyndsay Clark, and Nicolas Nicolov. The 2010 icwsm jdpa sentment corpus for the automotive domain. In *4th Int'l AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*, 2010.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1630–1639, 2013.

Cane Wing-Ki Leung, Chi-Fai Chan Stephen, Fu-Lai Chung, and Grace Ngai. A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web*, 14(2):187–215, 2011.

Shoushan Li, Sophia Y. M. Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 635–643, Beijing, China, 2010.

Bing Liu. *Sentiment Analysis and Opinion Mining (Introduction and Survey)*. Morgan Claypool Publishers, 2012.

Feifan Liu, Li Bin, and Liu Yang. Finding opinionated blogs using statistical classifiers and lexical features. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM-2009)*, 2009.

Qu Lizhen, Georgiana Ifrim, and Gerhard Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 913–921, 2010.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text, Vol. 8, No. 3.*, 1988.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA, 2000.

Rada Mihalcea, Carmen Banea, and Jan Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Association of Computational Linguistics (ACL)*, 2007.

Karo Moilanen and Stephen Pulman. Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 378–382, 2007.

Subhabrata Mukherjee and Pushpak Bhattacharyya. Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1847–1864, 2012.

Tony Mullen and Collier Nigel. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418, 2004.

P. Muller, Afantenos, S., Denis P., and N. Asher. Constrained decoding for text-level discourse parsing. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1883–1900, 2012.

Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 271–278, 2004.

Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Association of Computational Linguistics (ACL)*, pages 115–124, 2005.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

Livia Polanyi and Martin van den Berg. Discourse structure and sentiment. In *Data Mining Workshops (ICDMW)*, pages 97–102, 2011.

Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10, 2006.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the penn discourse treebank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 31–38. Association for Computational Linguistics, 2006.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, 2008.

Jonathon Read and John Carroll. Annotating expressions of appraisal in english. In *Language Resources and Evaluation.*, 46(3), pages 421–447, 2012.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL*, pages 25–32, 2003.

Al Mostafa Shaikh, Helmut Prendinger, and Ishizuka Mitsuru. Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 191–202, 2007.

Swapna Somasundaran. *Discourse-level relations for Opinion Analysis*. PhD Thesis, University of Pittsburgh, 2010.

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line debates. In *In Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. North American Association for Computational Linguistics (NAACL)*, pages 116–124, 2010.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 26–34. Association for Computational Linguistics, 2007.

Maite Taboada, Voll Kimberly, and Brooke Julian. Extracting sentiment as a function of discourse structure and topicality. In *School of Computing Science Technical Report 2008-20*, 2008.

Maite Taboada, Julian Brooke, and Manfred Stede. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 62–70, 2009.

Maite Taboada, Brooke Julian, Tofiloski Milan, Kimberly Voll, and Manfred Stede. Lexicon based methods for sentiment analysis. In *Computational Linguistics, 2011. 37(2)*, pages 267–307, 2011.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 575–584, Morristown, NJ, USA, 2010.

Rakshit S. Trivedi and Jacob Eisenstein. Discourse connectors for latent subjectivity in sentiment analysis. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 808–813, 2013.

Radoslava Trnavac and Maite Taboada. The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, 34 (3):301–318, 2010.

Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.

Fei Wang and Yunfang Wu. Exploiting hierarchical discourse structure for review sentiment analysis. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 121–124, 2013.

Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Lecture Notes in Computer Science, pages 486–497, 2005.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210, 2005.

Florian Wolf and Edward A. Gibson. *Coherence in Natural Language: Data Stuctures and Applications*. MIT Press, 2006.

Hong Yu and Hatzivassiloglou Vasileios. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 129–136s, 2003.

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171, 2011.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. Fine-grained sentiment analysis with structural features. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 336–344, 2011.