*Journal of Universal Computer Science, vol. 21, no. 4 (2015), 561-586*
*submitted: 17/3/14, accepted: 26/3/15, appeared: 1/4/15* 

# Sentiment and Behaviour Annotation in a Corpus of Dialogue Summaries

**Norton Trevisan Roman**
(University of São Paulo, São Paulo – Brazil
norton@usp.br)

**Paul Piwek**
(The Open University, Milton Keynes – United Kingdom
p.piwek@open.ac.uk)

**Ariadne Maria Brito Rizzoni Carvalho**
(University of Campinas, Campinas – Brazil
ariadne@ic.unicamp.br)

**Alexandre Rossi Alvares**
(University of São Paulo, São Paulo – Brazil
alexandre.alvares@usp.br)

**Abstract:** This paper proposes a scheme for sentiment annotation. We show how the task can be made tractable by focusing on one of the many aspects of sentiment: sentiment as it is recorded in behaviour reports of people and their interactions. Together with a number of measures for supporting the reliable application of the scheme, this allows us to obtain sufficient to good agreement scores (in terms of Krippendorf's alpha) on three key dimensions: polarity, evaluated party and type of clause. Evaluation of the scheme is carried out through the annotation of an existing corpus of dialogue summaries (in English and Portuguese) by nine annotators. Our contribution to the field is twofold: (i) a reliable multi-dimensional annotation scheme for sentiment in behaviour reports; and (ii) an annotated corpus that was used for testing the reliability of the scheme and which is made available to the research community.

**Key Words:** Corpus Annotation, Automatic Dialogue Summarisation, Natural Language Processing, Computational Linguistics, Sentiment Analysis
**Category:** I.2.7, L.1.3

## 1 Introduction

Recent years have witnessed a growing interest in sentiment analysis by the computational linguistics community. Current research topics include identification and classification of sentiment in reviews of products (*e.g.* [Hu and Liu, 2004b, Hu and Liu, 2004a, Balahur and Montoyo, 2008]), people (*e.g.* [Katagiri and Takahashi, 2003, Takahashi and Katagiri, 2003, Hijikata et al., 2007]) and open domains (*e.g.* [Beineke et al., 2004, Lloret et al., 2009, Balahur et al., 2009]), determination of the semantic orientation of words, whether within (*e.g.* [Wilson et al., 2009]) or without context (*e.g.* [Hu and Liu, 2004b, Hu and Liu, 2004a]), flame detection in e-mails (*e.g.* [Spertus, 1997]), film scene retrieval (*e.g.* [Park et al., 2011]), emotion prediction in spoken tutoring systems (*e.g.* [Forbes-Riley and Litman, 2004]), detection of biased language in texts (*e.g.* [Herzig et al., 2011]) and irony in short messages (*e.g.* [Gianti et al., 2012]).

As in other areas of computational linguistics, much of the extant work relies on corpora with gold standard annotations for training and evaluation. Unfortunately, marking up texts for sentiment has proven to be an extremely difficult task, given its natural subjectivity [Turney and Littman, 2003], with results that may even depend on the personality and humour of its executors [Alm et al., 2005]. In fact, not only may the sentiment expressed in a piece of text depend on the background knowledge [Balahur and Steinberger, 2009] and contextual information the reader has at hand [Barrett et al., 2007, Callejas and López-Cózar, 2008], but it may also be expressed indirectly, by using words that describe situations that readers might relate to specific sentiments [Balahur et al., 2012]. This subjectivity, in turn, translates to low inter-annotator agreement scores [Park et al., 2011, Bayerl and Paul, 2011], specially when continuous dimensional approaches[1] are adopted [Gunes et al., 2011].

One possible reason for such disappointing results is that sentiment, defined by the Merrian-Webster dictionary as "an attitude, thought, or judgement prompted by feeling",[2] is in fact a multi-faceted phenomenon, as illustrated by the many different sources of sentiment, ranging from properties of objects to the appearance and behaviour of people (*cf.* Appraisal Theory [Ortony et al., 1988]). Also, when it comes to dialogues, expressions of emotions are sometimes shaded, due to unstated rules of politeness [Devillers et al., 2002], making them harder to be detected. This topic is, in turn, closely related to the so-called "first-order politeness" [Watts, 2003] (also known as politeness$_1$ [Eelen, 2001]), *i.e.* people's interpretation of what constitutes a polite or impolite behaviour.

To address the issue of dealing with such phenomena in textual reports, we have identified one, in our view, important class of sentiment reports and developed and evaluated an annotation scheme for this particular class. The class of sentiment reports that we focus on concerns human behaviour reports. In other words, we deal with sentiment classification for utterances where a behaviour of or interaction between human agents is reported. Ultimately, our aim is to show that given this specific notion of sentiment and a particular design of the annotation process, reasonable scores for reliability can be achieved.

To our knowledge, our focus on sentiment identification and classification in behaviour reports is novel. There are, however, good reasons for investigat-

[1] *I.e.* approaches based on labels taken from a continuous-valued scale (*e.g.* from -1 to +1) [Gunes et al., 2011].
[2] http://www.merriam-webster.com/dictionary/sentiment

ing behaviour reports. Within the area of automatic summarisation there is an emerging area of dialogue summarisation (*e.g.* [Zechner and Waibel, 2000, Reithinger et al., 2000, Murray et al., 2005]). Whereas most research in this area has focused on the content of the dialogue and, for example, what the interlocutors agreed (*e.g.* [Kameyama et al., 1996, Alexandersson et al., 2000]), there is a range of applications which would benefit from also being able to report sentiment evaluations concerning the behaviour and interaction of the conversational participants.

Such applications include summaries of conversational exchanges of customer services representatives to monitor call quality, but could also, for example, be used to provide better meeting summaries (conveying not only what has been agreed, but also any animosity or positive atmosphere). To build such summarisation systems, knowledge of how behaviour is reported in a dialogue summary is required. With no theoretical account available, a corpus-based approach seems appropriate. This, however, presupposes annotated dialogue summaries and source dialogues.

In this article, we extend the results from [Roman and Carvalho, 2010], by presenting more details on the annotation process and results of a corpus of dialogue summaries, with samples in Portuguese and English, annotated with behaviour reporting information. Besides introducing this new focus on behaviour reports, this corpus also features one of the few existing corpora for languages other than English (*e.g.* [Devillers et al., 2002, Abdul-Mageed and Diab, 2011, Pápay et al., 2011, Boldrini et al., 2012, Cavicchio and Poesio, 2012, Clematide et al., 2012, Gianti et al., 2012, Roman et al., 2013]), thereby helping reduce the scarcity of resources for such languages. Finally, the entire corpus is made available, under a Creative Commons License, at `http://www.each.usp.br/norton/resdial/index_ing.html`, codified as described in [Roman, 2013].

The rest of this article is organised as follows. Firstly, in Section 2 we present some related work, pointing out the similarities and differences to ours. Then, in Section 3 we describe our method for collecting a corpus of human-authored dialogue summaries, which are based on dialogues that systematically vary in terms of the behaviour of their participants. Next, Section 4 presents our multidimensional annotation scheme for sentiment in behaviour reports. Section 5, in turn, describes how the scheme has been tested. This includes an account of the training that the annotators received and the procedure that they followed when applying the scheme. Section 6 assesses the scheme's reliability by measuring the amount of agreement between annotators. We present conclusions and directions for further research in Section 7.

## 2 Related Work

Current research on building and annotating corpora with sentiment usually fits into one out of five categories: annotation of transcribed real-life dialogues (*e.g.* [Litman et al., 2003, Craggs and Wood, 2004, Callejas and López-Cózar, 2008]), multimedia real-life dialogues, in which video and audio information is also presented to the annotators (*e.g.* [Devillers et al., 2002, Lee et al., 2002, Forbes-Riley and Litman, 2004, Morrison et al., 2007], film dialogues (*e.g.* [Mouka et al., 2012]), newswire texts (*e.g.* [Wiebe et al., 2005, Balahur and Steinberger, 2009, Abdul-Mageed and Diab, 2011]) and, more recently, social media and blogs (*e.g.* [Boldrini et al., 2012, Gianti et al., 2012]).

Despite the amount of research done, however, there seems to be no effort in annotating genres that derive from other genres, such as summaries of dialogues. So far, all efforts concentrate on identifying and annotating emotional/sentimental information by taking their sources in isolation. Although this approach does make sense for primary sources, in which the material comes directly from its producer (that is, writers, dialogue partners etc.), it makes less sense when it comes to secondary sources, which build on primary sources, as is the case with dialogue summaries that, though produced by a summariser, build on dialogues as produced by their interlocutors. This is a distinctive property of the annotation scheme and corpus presented in this article, in which we try to identify reports, by the summarisers, on behaviour demonstrated by the dialogue participants.

Hence, instead of trying to determine the specific emotion expressed by writers or dialogue participants (*e.g.* [Devillers et al., 2002, Gianti et al., 2012, Boldrini et al., 2012, Callejas and López-Cózar, 2008, Morrison et al., 2007]), or even their private state (*e.g.* [Wiebe et al., 2005]), the scheme we propose focuses on telling objective from subjective language, to further determine the polarity of the subjective part (as in [Litman et al., 2003, Lee et al., 2002, Forbes-Riley and Litman, 2004, Balahur and Steinberger, 2009, Abdul-Mageed and Diab, 2011, Mouka et al., 2012]) – a process carried out in a single step, in which a "neutral" label is assigned to objective material – along with its intensity (as in [Craggs and Wood, 2004]). Differently from these examples, however, our scheme seeks to capture not only the expression of sentiment by the summary author, but also how this sentiment can be linked to the summarised dialogue.

Finally, another noteworthy point in the field of corpora annotation, and which is still open for debate (*e.g.* [Taboada and Das, 2013]), relates to the number of annotators that actually applied the proposed scheme, since it is a common belief that the more annotators one has, the more one can be confident about the annotation results (*cf.* [Artstein and Poesio, 2005, Bayerl and Paul, 2011]). Current numbers for this variable vary from a single annotator (*e.g.* [Litman et al., 2003]), to as many as 11 (*e.g.* [Craggs and Wood, 2004]), with two being the most frequent choice (*e.g.* [Devillers et al., 2002, Abdul-Mageed and Diab, 2011, Mouka et al., 2012, Boldrini et al., 2012, Lee et al., 2002, Forbes-Riley and Litman, 2004, Maks and Vossen, 2012]). As described in Section 5, and similar to [Callejas and López-Cózar, 2008, Morrison et al., 2007], we gathered data from a set of nine independent annotators, thereby reducing the odds that the annotation results are distorted (as a result of the work of a single annotator) [Bayerl and Paul, 2011].

## 3 A Corpus of Human-authored Dialogue Summaries

Our aim was to collect a corpus of dialogue summaries where summaries include not only information about the conversational exchange content, but also evaluations of the way this content was presented by the dialogue participants. In doing so, our intention was primarily to build up the basis for determining the existence of bias in the way these evaluations were presented (in this scenario, we identified bias by determining whether reports on the dialogue participants' behaviour varied according to the summariser's assumed viewpoint). This information could then be used not only to detect bias in existing summaries, but also to identify the way bias is introduced, thereby developing strategies to automatically generate more balanced, ideally unbiased, summaries. More detail on these results can be found in [Roman et al., 2006, Roman and Carvalho, 2010].

For this purpose we required, as a basis for the human-authored summaries, dialogues which involve behaviours that are amenable to evaluative reporting. In particular, we focused on (im)politeness of interlocutors. Though we considered using naturally-occurring dialogues (*e.g.* [Devillers et al., 2002, Craggs and Wood, 2004, Callejas and López-Cózar, 2008]), we eventually decided to work with machine-generated material. The main reason for this decision was the ability to systematically control for the dialogue participants' behaviours, an almost impossible goal to achieve with naturally-occurring dialogues, and which is paramount if we are to draw any correlation between both genres (that is, summary and dialogue).

We used the eShowroom NECA system [van Deemter et al., 2008], a system for automatically generating scripted dialogues between a virtual sales person and a buyer, to create a basis of four sale dialogues with (i) two dialogues in which both participants act politely, and (ii) two dialogues in which one of the participants is impolite (see Figure 1 for a sample dialogue). We had 30 independent summarisers produce one summary for each of the 4 dialogues. Summarisers were asked to produce both a summary where there was no limit on the number of words that could be used and one with a size restriction (10% of the words of the original dialogue).[3] This limit was arbitrarily chosen so we could

[3] The requests to summarise unrestricted and restricted size summaries were separated

verify the summarisers' behaviour both when facing a very loose size restriction and a much harder one which, although more restrictive, still fall within the 5-30% compression rate claimed by some authors to produce quality summaries (*cf.* [Yeh et al., 2005]). Finally, our group of summarisers was split up into three subgroups: we asked one subgroup to summarise the dialogue as if they were a neutral observer, whereas each of the other two was asked to summarise it from the perspective of one of the interlocutors.

R : "Hello! I am Ritchie."
T : "Can you tell me something about this car?"
R : "It is very safe."
R : "It has anti-lock brakes."
R : "It has airbags."
T : "How much does it consume?"
R : "It consumes 8 litres per 60 miles."
T : "Interesting."
T : "What kind of luggage compartment does this car have?"
R : "It has a spacious luggage compartment."
T : "Excellent!"
T : "What kind of interior does this car have?"
R : "It has a spacious interior."
T : "Excellent!"
T : "How much does this car cost?"
R : "It costs 25 thousand Euros."
T : "Well, well."
T : "All in all this is a perfect car. It is a deal!"
R : "Really? I am sure you wont regret it."

Figure 1: Sample dialogue, with both participants (Ritchie and Tina) acting politely.

As a result, we obtained a total of 240 different dialogue summaries, with 10 summaries for each experimental condition, *i.e.* for each possible combination of the variables maximum allowed size, viewpoint and summarised dialogue. Since source dialogues were in English and summarisers were native speakers of Portuguese,[4] they were allowed to produce their summaries in the language they felt more comfortable with in this task, resulting in 36 (15%) summaries in English and 204 (85%) in Portuguese (see Figure 2 for some sample summaries[5], covering each of the three possible viewpoints, for the dialogue in Figure 1). As for summary lengths, Table 1 presents the total and mean number of words used in the summaries within each experimental condition.[6] As expected, the

in time by a couple of months, in order to reduce any experimental bias that might occur as a result of repeated runs on the same task.

[4] Although all of them were also fluent speakers of English.

[5] Since two of the summaries were in Portuguese, we had them translated in the Figure.

[6] The number of words in a document was measured with the "wc" Unix command.

group with no restriction in the summary size was the one producing the longest summaries (7,049 words, versus 1,316 in the 10% restriction group). We refer the interested reader to [Roman et al., 2006] for more detail on the construction of the dialogue summary corpus, including transcripts of the four source dialogues.

**Table 1:** Corpus length (number of words) for each experimental condition

| *Restriction* | *Dialogue* | *Dialogue Length* | *Viewpoint* | | | *Total* | *Mean* |
|---|---|---|---|---|---|---|---|
| | | | *Customer* | *Vendor* | *Observer* | | |
| No | $D_1$ | 99 | 334 | 647 | 564 | 1,545 | 51.5 |
| | $D_2$ | 182 | 471 | 1,010 | 950 | 2,431 | 81.03 |
| | $D_3$ | 124 | 321 | 868 | 815 | 2,004 | 66.8 |
| | $D_4$ | 53 | 228 | 446 | 395 | 1,069 | 35.63 |
| 10% | $D_1$ | 99 | 92 | 113 | 97 | 302 | 10.07 |
| | $D_2$ | 182 | 148 | 184 | 158 | 490 | 16.33 |
| | $D_3$ | 124 | 118 | 129 | 116 | 363 | 12.1 |
| | $D_4$ | 53 | 51 | 61 | 49 | 161 | 5.37 |

*Viewpoint*: customer

> Tina liked the car because it's safe and economic, it has spacious luggage compartment and interior, and it isn't expensive.

*Viewpoint*: observer

> Tina was interested in a car Ritchie was selling. She asked for information on many aspects of the car, such as safety items, fuel consumption, internal and luggage compartment space and, finally, the price. After this short negotiation Tina was convinced that car was ideal for her and closed the deal.

*Viewpoint*: attendant

> A woman bought a car from me today. She came in asking some of the car's features, such as fuel consumption, luggage compartment space, internal space and, of course, the price. She seemed happy with what she saw and we closed the deal.

Figure 2: Sample summaries, from different points of view, with no size constraints.

As a final note, even though 240 summaries may seem quite limited a resource, in the realm of human produced summaries, corpora range from collections as short as 15 summaries (*e.g.* [Jing and McKeown, 1999]) up to 1,000 summaries (*e.g.* [Amini, 2000]). Also, another important feature to take into account is the number of human summarisers involved. With current initiatives ranging from a single summariser (*e.g.* [Hasler, 2007]) to as many as 202 (*e.g.* [Teufel and Moens, 1997]), this corpus, with its set of 30 summarisers, does not seem off the scale. Finally, questions may arise on how genuine an example of language machine generated dialogues might be. We think this is not an issue, since the resulting summaries are all produced by humans. In this case, the fact that people's reactions are based on what they expect [Balahur and Steinberger, 2009], and not on any particular innate feature of the source material, mitigates any "lack of naturality" that artificial dialogues might present, thereby shifting the focus to how summarisers interpreted these dialogues.

## 4 Multi-dimensional Annotation Scheme

In our scheme, the clause functions as the basic unit of annotation, defined as a text span consisting, as a minimum, of a verb and its complements [Miller, 2002]. Since we do not require verbs to explicitly appear in sentences, text spans with elliptical verbs, such as the elliptical "to be" in "The perfect buyer", for example, are taken as a separate clause. Also, we do not tell apart different types of clauses, such as subordinated, coordinated and main clauses, for example. As a unit, clauses have a number of desirable properties for annotation, notably their independence of a specific linguistic theory [Mann and Thompson, 1988] and their purely syntactic nature [Krippendorff, 2004], which makes them less prone to ambiguities.

Even though clauses are not so common a unit of annotation in the related literature, they also have the advantage of being longer than words (*e.g.* [Wiebe et al., 2005]) and shorter than full sentences (*e.g.* [Devillers et al., 2002, Abdul-Mageed and Diab, 2011, Clematide et al., 2012]), utterances (*e.g.* [Craggs and Wood, 2004, Callejas and López-Cózar, 2008, Mouka et al., 2012]) or even annotator-defined text spans (*e.g.* [Gianti et al., 2012]). In choosing clauses, our intention was to reduce the effect that too long or short units might have on the annotation effort, also avoiding disagreements on the units' boundaries, should we decide for free-length text spans.

The annotation scheme we propose classifies every clause in a summary according to five distinct dimensions, shown in Figure 3. At the root of the scheme is a distinction between clauses that do and those that do not report behaviour or interaction with sentiment. The other dimensions classify clauses according to whether the reported sentiment was positive or negative, its intensity, who or what was reported and any inferential relation of the current report with other statements in the summary. All dimensions are summarised in Table 2.

Finally, since dealing with a high number of categories might put higher demands both on memory load and annotator's ability to differentiate between possible choices [Bayerl and Paul, 2011], we decided to keep the number of categories in each dimension low, as an attempt to increase the reliability of the scheme, specially in light of existing evidence that points to a decrease in inter-annotator agreement levels as the number of categories in an annotation scheme grows (*e.g.* [Gut and Bayerl, 2004, Callejas and López-Cózar, 2008, Bayerl and Paul, 2011, Maks and Vossen, 2012]). On this account, it has been observed that adding just a single category to a subjective dimension dropped agreement in almost 29% [Craggs and Wood, 2004]. In what follows, we define each of the dimensions in detail.

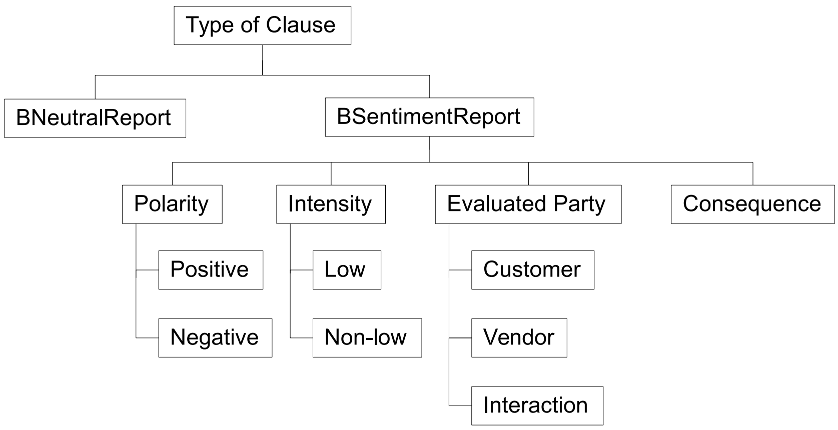


**Figure 3:** Dimensions hierarchy in the annotation scheme.

### 4.1 Type of Clause

The basic distinction in this scheme rests on the dimension *Type of Clause*, between clauses classified as **BNeutralReport** and **BSentimentReport**. In contrast with the related literature (*e.g.* [Batliner et al., 2000, Pang and Lee, 2004, Abdul-Mageed and Diab, 2011, Clematide et al., 2012]), the distinction underlying our notion of *Type of Clause* is more specific than the distinction between subjective/objective clauses, or clauses that convey sentiment and those that do not. Our distinction is grounded in what [Mills, 2003] calls social politeness (or political behaviour), that is behaviour, humour, feelings and sentiments concerning an interactional feature (akin to what [Keenan et al., 1977] call the interactional content of the clause).

In particular, we propose to label a clause *BSentimentReport* only when it evaluates, either positively or negatively, the individual dialogue participants'

**Table 2:** Dimensions that build up the scheme

| *Dimension* | *Values* | *Example* |
|---|---|---|
| Type of Clause | BNeutralReport<br>BSentimentReport | I asked a vendor about a car<br>I was in an awful mood |
| Polarity | Positive<br>Negative | Ritchie is a great vendor<br>[she] seemed rather impatient |
| Intensity | Low<br>Non-low | I contradicted myself<br>The perfect buyer |
| Evaluated Party | Customer<br>Vendor<br>Interaction | A nutty client came in<br>I [vendor] took a deep breath<br>Today was not a good selling day |
| Consequence | [Clause id] | [Since I was in such a bad mood] I ended it up with yet another scolding |

behaviour or the interaction as a whole. This covers all the evaluation forms proposed by [Martin, 1999], to wit, affect (sentimental responses), judgements (moral evaluations of behaviours) and appreciation (an aesthetic quality of semiotic processes and natural phenomena characterized as, for example, harmonious or elegant); applying them to the interaction and its participants. Within this context, *BNeutralReport* characterizes the complement of *BSentimentReport*. It applies to clauses with no evaluations of the interaction or its participants. Note that our notion of *BSentimentReport* is also more specific than emotions as classified by the appraisal theory of emotions [Ortony et al., 1988]. While the latter covers reactions to events, agents and objects, our scheme focuses on agents and events (but only as far as these are about the interaction between agents). In particular, sentiment reports regarding objects (e.g. "The car was beautiful") are classified as *BNeutralReport*.

### 4.2 Polarity

Applied only to *BSentimentReport* clauses, polarity ranges over two values only: **Positive** and **Negative**. According to this dimension, and following [Lehnert, 1982] and [Martin, 1999], a clause is classified as *Positive*, when it describes pleasant actions or feelings, or when it assesses them positively, like in "She was patient" and "The service was fine". A clause is otherwise classified as *Negative* when it describes actions or feelings that are unsatisfactory/inappropriate (and so must be avoided) as in "He is rude" or "Tossed all her dissatisfaction on me". It is worth noticing that, even though our scheme differs from some of the related literature (*e.g.* [Hovy, 1990,Litman et al., 2003,Forbes-Riley and Litman, 2004, Clematide et al., 2012, Maks and Vossen, 2012]), which allow for three polarity values (positive, negative and neutral), all these values are contemplated here, since neutral reports are captured at the *Type of clause* level, through the *BNeutralReport* value.

### 4.3 Intensity

With only two values in its scale: low intensity and non-low intensity (i.e., normal to high), this dimention is intended to capture the degree of intensity or strength of the reported sentiment, so as to determine whether sentiment reports are biased. Hence, instead of applying a dichotomy of extreme values, such as "low" (or "<NORM") and "high" (or ">NORM") [Dyer, 1983, Turney and Littman, 2003], or a more fine-grained scale, such as "low", "medium", "high", or "extreme" [Wiebe et al., 2005], or numbers from 0 to 4 (or 5) [Craggs and Wood, 2004], or even relying on a continuum of values, from "calm" to "excited" [Lang, 1995, Picard, 1995, Martin, 1999], for example, this dimension seeks to highlight situations where reporters tone down the intensity of, for example, a negative report, if this suits their purposes (*cf.* [Knee and Zuckerman, 1996, Higgins and Bhatt, 2001, Blackwood et al., 2003, Kaplan and Ruffle, 2004]). Just like Polarity, this dimension applies only to *BSentimentReport* clauses.

### 4.4 Evaluated Party

Since one of our goals is to identify behaviour reports with sentiment, with this dimension we intend to record the dialogue participant whose behaviour or sentiment is reported (*cf.* [Hunston, 1999]), both explicitly, as in "the vendor treated me very well", or implicitly, as in "the service was very good" (in which case a service necessarily implies a server). In the context of our sales dialogues this dimension takes one of three values: (a) **Vendor**, meaning that the evaluation concerns the vendor's sentiments and behaviour; (b) **Customer**, when it concerns the client's; and (c) **Interaction**, which must be used whenever the clause does not evaluate either of the dialogue participants individually but, instead, their interaction with each other.

As for this last value, *Interaction* means that the clause in question presents a summary of what happened, as in "what a horrid day". Clauses should only be classified as *Interaction* if they refer to the interaction between the participants. For example, if the fact that the day was horrid had something to do with the weather (as may be inferred from context), then "what a horrid day" should not be labelled *BSentimentReport* (and consequently Evaluated Party would not apply).

### 4.5 Consequence

*Consequence* should be applied to *BSentimentReport* clauses describing situations or feelings that were caused by something reported in another clause, as in "I was so badly served *that I lost my nerve*", where "I lost my nerve" describes a consequence of "I was so badly served". The motivation for this dimension, beyond determining a cause, is to establish whether blame may have been transferred, thereby resulting in a probably biased text. As such, this dimension labels those clauses in which one party tries to justify his/her actions in terms of something else that happened.

### 4.6 Multiple classifications

It has already been reported in the related literature (*e.g.* [Devillers et al., 2002, Craggs and Wood, 2004, Wiebe et al., 2005, Balahur and Steinberger, 2009, Gianti et al., 2012, Maks and Vossen, 2012]) that some units (be they sentences or utterances) can express several different sentiments, with different polarities, towards different targets. The corpus we describe here is no exception to that rule. Hence, from the five dimensions described above, three can take multiple classifications (there being no limit to the number of classifications that can be assigned to a clause). These are *Polarity*, *Intensity*, and *Evaluated Party*. As a result, clauses like "I gently served the rude client", for example, are classified as *BSentimentReport*, holding different evaluations about two participants (*Vendor* and *Customer*, respectively), with opposite polarities (*Positive* and *Negative*) and with the same *Intensity*.

## 5 Applying the Scheme: Corpus Annotation

The test bed for our annotation scheme comprises 1,773 clauses from the 240 human generated summaries described in Section 3. The entire set of summaries was independently annotated by nine volunteers (seven male and two female, all graduate students from a Brazilian university). As an additional measure to increase reliability, and to make annotators more familiar with the annotation scheme, annotators had to independently go through a training stage of about one and a half hour, before doing the annotation. This is a commonly adopted procedure (*e.g.* [Devillers et al., 2002, Craggs and Wood, 2004, Wiebe et al., 2005, Cohn et al., 2008, Balahur and Steinberger, 2009, Abdul-Mageed and Diab, 2011, Clematide et al., 2012, Maks and Vossen, 2012, Petukhova and Bunt, 2012]) that was already found to raise agreement (*cf.* [Balahur and Steinberger, 2009, Bayerl and Paul, 2011]).

During the training stage, annotators were given a description of the annotation scheme, along with a set of guidelines to help them understand what each category meant. The participants were then asked to annotate a set of 18 summaries, with 128 clauses in total, constructed specifically for the training phase (see Figure 4 for some test summaries[7]). These training summaries were carefully written to provide a number of specially designed overly prototypical examples, making it easier for annotators to get the grips with the task.

I served a horrid person today. Jeez, she was really rude. Treated me so savagely. I don't know how I managed to keep my nerve... it's good that she went away soon.

Today I was well treated when I went to buy a car. The vendor was very nice and gave me all attention I needed. The price for the car I wanted wasn't bad at all. Deal.

I'd like to buy a car, that's why I went to a store. I found a really beautiful and cheap car. I had no doubt and bought it.

**Figure 4:** Sample constructed summaries used at the training stage.

To help annotators in their endeavour we have developed an annotation tool specifically designed for this effort. This tool was important not only to standardise the annotation results, making it easier to compare outputs from different annotators, but also to further reduce the cognitive load on them, by grouping together *Type of Clause*, *Polarity* and *Evaluated Party* into a single overall category. Thus, instead of classifying some clause as, for example, <*Type of Clause*: BSentimentReport; *Polarity*: Positive; *Evaluated Party*: Vendor>, annotators are offered the "Positive Report about the Vendor" alternative. This procedure leaves annotators with a single choice to make, as opposed to three separate choices, without loss of generality.[8]

The idea in following this procedure was to achieve better inter-annotator agreement, since evidence has emerged that such an approach may indeed improve agreement (*cf.* [Bayerl and Paul, 2011]). In particular, [Wilson et al., 2009] noticed that executing annotation tasks in a single step performed about as well or better than doing it in separate steps.[9] Finally, as a way to verify the acceptance of our labels by annotators, the program interface also featured a "None of the above" value, which was meant to be used whenever the annotator does

[7] Adapted from Portuguese.
[8] Note, however, that dimensions were unified at the interface level only.
[9] In this case, they compared a two-step approach, where annotators first decide on whether some word is neutral, and then classify it further in case it is not, to a one-step approach, where annotators do the entire classification in a single go.

not agree with any of the existing categories.

Upon finishing the training stage, annotators were given the annotation program and the set of instructions, and asked to complete the annotation whenever they found it most convenient (*cf.* [Birnbaum, 2004]). At this point, they were explicitly instructed not to rush on it, since we would rather them to do it in the best-suited surroundings, so we did not run the chance of having them overlook the data, and consequently produce incorrect classifications (*i.e.* classifications that would not correspond to their real beliefs). Also, before giving the program away, we shuffled the summaries in the database, ensuring that each annotator had a different summary order so as to avoid any bias that might result from the order of data presentation.

### 5.1 Results

Figure 5 shows the distribution of labels for *Type of Clause*, *Polarity* and *Evaluated Party*. In this figure, we have followed [Moens, 2002] and used data from all nine annotators. The total number of data points is then 1,773 clauses $\times$ 9 annotators = 15,957. However, with some clauses involved in multiple classifications, we ended up with the slightly higher number of 16,011 items. With around 0.34% of the corpus presenting multiple annotations,[10] our results astray from the current literature, which reports as disparate values as 3.2% [Craggs and Wood, 2004] and 44% [Wiebe et al., 2005] of all units bearing multiple ratings. We conjecture that such a discrepancy may come out as a result of the idiosyncrasies of each annotation scheme. Finally, the dimensions *Intensity* and *Consequence* were left out of this and the next figure because of their low reliability (we will discuss this further in Section 6).

As can be seen in the figure, *BNeutralReport* clauses clearly outnumber *BSentimentReport* clauses. This is in line with research by [Callejas and López-Cózar, 2008], who found that over 85% of the utterances in their corpus were annotated as "neutral". Similarly, [Devillers et al., 2002] report an over 86% amount of neutral sentences, there also being reports on rates around 60% [Morrison et al., 2007] and 73% [Forbes-Riley and Litman, 2004]. In our research, neutral units make up around 80% of the corpus, if we take data from all annotators altogether.

Interestingly, our corpus also shows a strong predominance of negative labels (the set of Negative Reports, represented in the second, fourth and sixth column), even though two of the source dialogues were more balanced. In this case, out of the 3,096 non-neutral (except for "None of the Above"[11]) labels given, 2,197

[10] There were 54 units in total with double classifications. We observed no units with three or more classifications.

[11] "None of the above" was assigned to nine units (around 0.06% of the corpus) by two annotators only, probably as a result of deviant readings of the annotation instructions.

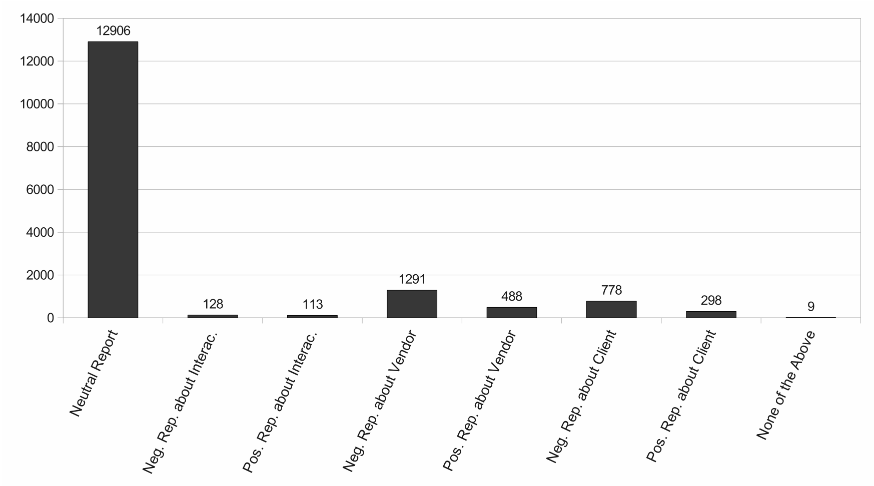


**Figure 5:** Overall distribution of category labels, amongst 9 annotators.

(71%) were negative labels. A very similar result was also reported by [Wiebe et al., 2005], who found that negative labels were assigned to 73% of their non-neutral units. This is also somewhat in line with psychological findings about a natural tendency, both in humans and animals, to give more importance to negative data [Rozin and Royzman, 2001].

When analysing annotator mode-based numbers[12] (*cf.* [Vieira and Poesio, 2000]), we come to a similar distribution, as shown in Figure 6. The only difference is that "None of the above" has dropped from nine to naught, meaning that, even though some annotators might have found the alternatives for classification inappropriate for some units, that view was not the most popular amongst them. Just like in Figure 5, there is a strong predominance of neutral labels (83.4% of all 1,773 clauses) and, amongst the 294 non neutral labels, 223 (75,9%) comprised negative reports.

## 6 Evaluating the Scheme: Inter-Annotator Agreement

In this research, we used Krippendorff's $\alpha$ [Krippendorff, 2004] as our coefficient of agreement. This choice was guided by three important features of $\alpha$, to wit, (i) it accounts for the amount of agreement that is expected by chance; (ii) it calculates expected agreement by looking at the overall distribution of annotations,

[12] *I.e.*, when taking, for each clause, the most commonly assigned label, as opposed to summing up all the labels separately. However different, these figures approximate absolute majority, since such a majority could not be established in only 24 of the 1,773 clauses (around 1.35%), for this category.

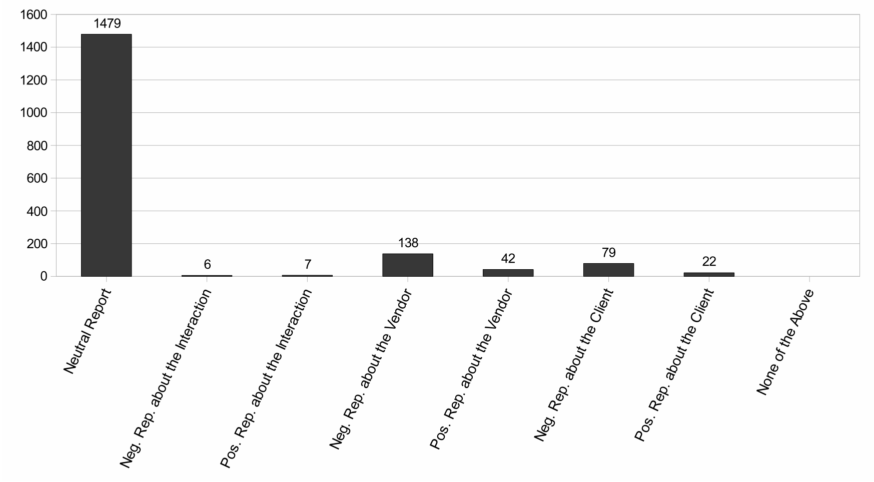


**Figure 6:** Distribution of most assigned category labels.

as opposed to inspecting each annotator's individual distribution [Artstein and Poesio, 2008]; and (iii) it is not influenced by systematically biased distributions, *i.e.*, distributions slanted towards some of the categories, as is the case with Cohen's $\kappa$ [Eugenio, 2000, Eugenio and Glass, 2004, Krippendorff, 2004], for example.

With $\alpha$, agreement is measured by taking the proportion of the difference between the observed disagreement and the disagreement expected by chance, related to the disagreement expected by chance, as follows [Krippendorff, 2004]

$$\alpha = \frac{D_e - D_o}{D_e} = 1 - \frac{D_o}{D_e} \tag{1}$$

where $D_o$ stands for the amount of observed disagreements and $D_e$ represents the amount of expected disagreements by chance. To estimate $D_e$, $\alpha$ looks at the overall distribution of annotations, with no regards to which coders produced them [Artstein and Poesio, 2008], that is, it builds a frequency distribution from all labels assigned by all coders, and takes this distribution as an estimate for the underlying probability distribution of labels.

For reliability purposes, values for $\alpha$ range from 0 to 1 [Krippendorff, 2004], where 1 means total agreement (*i.e.* when $D_o = 0$) and 0 implies that analysed data cannot be told apart from random events (*i.e.* when $D_o = D_e$). Negative $\alpha$ values may also occur, arising from two sources: sampling errors, as a consequence of using small data sets; and systematic disagreement, where observers "agree on disagreeing", that is when they systematically choose opposite labels, due to opposite interpretation of the annotation instructions, for example.

Table 3 shows the results for Krippendorff's $\alpha$, in its version for nominal categories, many observers and allowing for missing values (see [Krippendorff, 2004, pp. 230]), for all nine annotators. Assuming $\alpha \geq 0.8$ as good reliability, with $0.67 \leq \alpha < 0.8$ allowing tentative conclusions [Krippendorff, 2004, Artstein and Poesio, 2008], only *Polarity* turns out to be a reliable dimension, with *Type of Clause* and *Evaluated Party* permitting but tentative conclusions. Although these values may look rather modest, they are very encouraging when compared to existing results on sentiment classification, such as those presented by [Craggs and Wood, 2004], for example, who report $\alpha = 0.25$ for *Intensity* and $\alpha = 0.26$ for *Polarity*[13] as their highest values, and [Callejas and López-Cózar, 2008], who report an overall agreement of $\alpha = 0.3382$.

Regarding *Type of Clause*, it is worth noticing that the existence of a "Neutral" category which, however necessary, has been described as a common source of confusion [Alm et al., 2005], just makes it harder to expect a high agreement on such a dimension. Also, the lack of clear definitions of sentimental terms can naturally lead to low values of agreement [Alm et al., 2005], specially when, as in our case, sentiment is related to evaluative judgement of appropriate behaviour, where there always is the problem that people may not agree on what counts as appropriate behaviour in a given situation [Watts, 2003].

**Table 3:** Alpha values for the annotation [Roman and Carvalho, 2010].

| *Dimension* | $\alpha$ | *Reliability* |
|---|---|---|
| Polarity | 0.843 | Reliable |
| Evaluated Party | 0.783 | Tentative |
| Type of Clause | 0.674 | Tentative |
| Intensity | 0.212 | Unreliable |
| Consequence | 0.085 | Unreliable |
| $\text{Consequence}_g$ | 0.175 | Unreliable |

With $\alpha < 0.67$, *Intensity* and *Consequence* turned out to be unreliable. Regarding *Consequence*, the results show that there is very low agreement on whether a clause signifies the consequence of an event reported in another clause ($\alpha = 0.175$, at $\text{Consequence}_g$), and even less whether a specific clause represents a cause ($\alpha = 0.085$). Actually, the very notion of cause/consequence between clauses involving sentiment in behaviour reports seems to be highly subjective. As for *Intensity*, its low score may be connected to the observation of [Ortony et al., 1988, pp. 34] that "[sentiments] vary a great deal in intensity both within

[13] Although, in their scheme, *Intensity* had five categories, whereas *Polarity* had seven.

and between people. [...] that the intensity of sentiments is influenced by a number of variables", which makes it improbable to achieve high agreement on this dimension.

Finally, another possible reason for *Intensity* and *Consequence* scoring so low[14] might be that most of the annotated items fall under one single category (in this case, Neutral Report, as shown on Figures 5 and 6). This prevalence problem (see [Eugenio and Glass, 2004]) increases the expected agreement by chance, making it harder to get a high figure for this coefficient of agreement [Artstein and Poesio, 2008].

When taking a closer look at the disagreement figures, our data also give us a clue about the importance of having more than just a couple of annotators carry out the task. In this case, an analysis of the agreement scores for each pair of annotators led us to the results shown in Figure 7. In this figure, the difference between the highest and lowest scores varies considerably, ranging from 0.180, for *Type of clause*, to 0.956 for *Intensity*. Also, it seems that the lower the overall agreement (see Table 3), the higher the difference between pairs. Interestingly, not a single pair of annotators was responsible for the lowest agreement in more than one dimension, that is all the worst scores for pairwise agreement come from different pairs of annotators. This, in turn, is an indicative that disagreement may have come mainly from the natural subjectivity of the task, instead of systematic deviant readings of the annotation instructions.

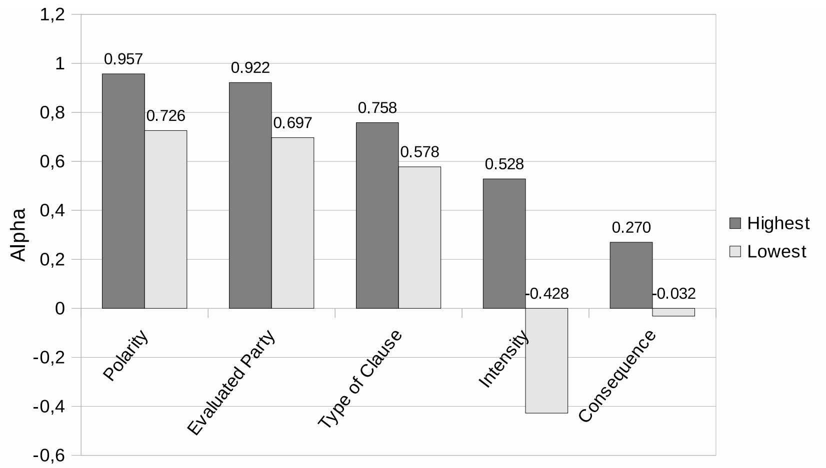


**Figure 7:** Highest and lowest pairwise alpha, for each category.

[14] And which may also have reduced the other dimensions' reliability to the same extent, but not enough to render them unreliable.

Regardless of the dimension, the overall $\alpha$ value is substantially higher than that of the *lowest* pairwise agreement score. Similarly, this value is also substantially lower than that of the *highest* pairwise agreement score. As such, if we compare pairwise agreement with that of the whole set of annotators, in the worst case we might end up with an agreement level that is either too high or too low, should we rely on only two annotators (see Table 3 and Figure 7). This suggests that a single couple of annotators may not be enough to make sure that no undue weight is given to their occasional alignment or misalignment.

Take, for example, *Evaluated Party.* If we were to rely on two annotators only, and these happened to be the ones with the highest agreement, this dimension would come out as reliable, whereas our overall score for nine annotators indicates that we should go no further than making tentative conclusions. At the other end, *Type of Clause* could come out as unreliable, if we were to use only the pair with the lowest agreement. Similarly, *Polarity* – the single reliable dimension – could come out as only allowing tentative conclusions if we depend on no more than two annotators.

As for the question on at what number of annotators we can reach the same conclusion as we do with the full set of annotators, we found no pattern in the dataset, as shown in Tables 4 and 5. In these tables, we see that each dimension requires a different minimum number of annotators to reach some stability in their classification as reliable, tentative or unreliable. Conclusions reached with nine annotators can already be reached for *Intensity* and *Consequence*, for instance, from only two annotators (*i.e.*, their classification as unreliable would not change, should we take the best or worst pairwise agreement), whereas *Polarity* demands at least six. Interestingly, both tentative dimensions (*i.e. Type of Clause* and *Evaluated Party*) required the full set of annotators in their calculation for alpha. This could be an indicative that, for these dimensions, even nine annotators may not be enough.

**Table 4:** Best and worst alpha values for subsets from 2-5 annotators.

| *Dimension* | *2 annotators* | | *3 annotators* | | *4 annotators* | | *5 annotators* | |
|---|---|---|---|---|---|---|---|---|
| | *Best* | *Worst* | *Best* | *Worst* | *Best* | *Worst* | *Best* | *Worst* |
| Polarity | 0.957 | 0.726 | 0.938 | 0.762 | 0.919 | 0.780 | 0.906 | 0.796 |
| Evaluated Party | 0.922 | 0.697 | 0.907 | 0.705 | 0.888 | 0.715 | 0.864 | 0.723 |
| Type of Clause | 0.758 | 0.578 | 0.756 | 0.607 | 0.748 | 0.617 | 0.744 | 0.634 |
| Intensity | 0.528 | -0.428 | 0.389 | -0.146 | 0.355 | -0.008 | 0.329 | 0.066 |
| Consequence | 0.270 | -0.032 | 0.211 | -0.010 | 0.182 | -0.009 | 0.161 | 0.012 |

**Table 5:** Best and worst alpha values for subsets from 6-9 annotators.

| *Dimension* | *6 annotators* | | *7 annotators* | | *8 annotators* | | *9 annotators* |
|---|---|---|---|---|---|---|---|
| | *Best* | *Worst* | *Best* | *Worst* | *Best* | *Worst* | |
| Polarity | 0.887 | 0.811 | 0.871 | 0.821 | 0.857 | 0.830 | 0.843 |
| Evaluated Party | 0.846 | 0.744 | 0.828 | 0.757 | 0.807 | 0.771 | 0.783 |
| Type of Clause | 0.717 | 0.644 | 0.699 | 0.654 | 0.687 | 0.664 | 0.674 |
| Intensity | 0.295 | 0.111 | 0.266 | 0.149 | 0.239 | 0.178 | 0.212 |
| Consequence | 0.141 | 0.021 | 0.124 | 0.044 | 0.108 | 0.065 | 0.085 |

## 7 Conclusions

In this article we introduced an annotated corpus of 240 human produced summaries, along with its corresponding multidimensional annotation scheme. Primarily designed for sentiment and behaviour assessment in dialogue summaries, this scheme can serve a variety of purposes, from identifying polite/im-polite behaviour to detecting bias in sentiment and behaviour reports (*cf.* [Roman et al., 2006]).

To test the soundness both of the developed scheme and its corresponding annotation guidelines, we carried out reliability studies with nine independent annotators. With current research relying on sets that range from a single annotator [Litman et al., 2003], to as many as 11 [Craggs and Wood, 2004], with two seeming to be the commonest choice (*e.g.* [Devillers et al., 2002, Lee et al., 2002, Forbes-Riley and Litman, 2004, Abdul-Mageed and Diab, 2011, Boldrini et al., 2012, Maks and Vossen, 2012, Mouka et al., 2012]), this seems to be an appropriate number of annotators, being higher than the minimum of five established by [Bayerl and Paul, 2011] for very critical tasks.

Results show that out of the five original dimensions, three were sufficiently reliable (*i.e.* $\alpha \geq 0.67$). These were *Type of Clause*, *Evaluated Party* and *Polarity.* For the two remaining dimensions (*Intensity* and *Consequence*) reliability could not be established. Another interesting feature of our corpus lies in the amount of neutral units and, amongst the non-neutral ones, the prevalence of negative reports, even though the set of source dialogues was balanced. This is not only in line with related research on annotating emotional features in corpora, but also with some findings in psychological research about emotion and sentiment.

By determining the highest and lowest reliability amongst subsets of two to nine annotators, we presented new empirical evidence for the use of more than two, even though we were not able to find a standard "minimum number of annotators" required to reach a conclusion about the reliability of the proposed dimensions. Furthermore, the fact that two of the dimensions (*Type of Clause* and *Evaluated Party*, *i.e.* those allowing for only tentative conclusions) required all nine annotators might be an indicative that this number should be raised. This, however, is an issue to be addressed in future research.

Finally, it is worth noticing that, despite the fact that our annotation scheme was primarily developed for use in dialogue summaries, it can still be adapted to other tasks with relative ease. Firstly, Polarity, Intensity, and the main dichotomy between BSentimentReport and BNeutralReport are categories of general use, along with Consequence, which deals with bounding units of annotation, whatever they are. Secondly, Evaluated Party can be adapted to the situation at hand. In this research, we had this category represent the dialogue participant whose behaviour was reported in the unit under consideration. In alternative setups, however, it could represent, for example, the political party that is evaluated, should one be interested in studying bias in political reports.

As for avenues for future research, we intend to apply this annotation scheme to a bigger corpus (*cf.* [Roman et al., 2013]), following the procedure already adopted in [Roman and Carvalho, 2010], so as to verify whether the results described in [Roman et al., 2006] still hold. Also, it would be interesting carrying out a deeper analysis of the characteristics and possible causes for disagreement. That, however, would demand a higher number of annotators, so as to increase the statistical power of the applied tests, thereby allowing for meaningful conclusions to be drawn. Finally, it would be useful to identify a way to reliably measure intensity of reports, since this is a feature that plays a prominent role in many theories of sentiment and evaluation (*e.g.* [Ortony et al., 1988]).

## Acknowledgements

We would like to thank Ivandré Paraboni, Svetlana Stoyanchev, Tu Anh Nguyen, Richard Doust, Luciano Antonio Digiampietri, Tomasz Kowaltowski and Arnaldo Mandel for their invaluable comments on a previous version of this article.

This research was sponsored by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Part of it was also supported by the EC Project NECA IST-2000-28580 and the Programa de Educação Tutorial (PET) – MEC/SESu.

## References

[Abdul-Mageed and Diab, 2011] Abdul-Mageed, M. and Diab, M. T. (2011). Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*, pages 110–118, Portland, Oregon, USA.

[Alexandersson et al., 2000] Alexandersson, J., Poller, P., Kipp, M., and Engel, R. (2000). Multilingual summary generation in a speech-to-speech translation system for multilingual dialogues. In *Proceedings of the First International Conference on Natural Language Generation*, pages 148–155, Mitzpe Ramon, Israel.

[Alm et al., 2005] Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada.

[Amini, 2000] Amini, M.-R. (2000). Interactive learning for text summarization. In *Proceedings of the PKDD'2000 Workshop on Machine Learning and Textual Information Access*, Lyon, France.

[Artstein and Poesio, 2005] Artstein, R. and Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language (FG-MoL 2005)*, Edinburgh, Scotlan.

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[Balahur et al., 2012] Balahur, A., Hermida, J. M., and Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53:742–753.

[Balahur et al., 2009] Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., and Martínez-Barco, P. (2009). Summarizing threads in blogs using opinion polarity. In *Proceedings of the Events in Emerging Text Types Workshop of the RANLP*, Borovets, Bulgaria.

[Balahur and Montoyo, 2008] Balahur, A. and Montoyo, A. (2008). Determining the semantic orientation of opinions on products - a comparative analysis. *Procesamiento del lenguaje natural*, (41):201–208.

[Balahur and Steinberger, 2009] Balahur, A. and Steinberger, R. (2009). Rethinking sentiment analysis in the news: from theory to practice and back. In *Proceedings of the first Workshop on Opinion Mining and Sentiment Analysis (WOMSA-2009)*, pages 1–12, Sevilla, Spain.

[Barrett et al., 2007] Barrett, L. F., Lindquist, K. A., and Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, 11(8):327–332.

[Batliner et al., 2000] Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., and Fischer, K. (2000). The recognition of emotion. In (Ed.), W. W., editor, *Foundations of Speech-to-Speech Translation*, pages 122–130. Springer.

[Bayerl and Paul, 2011] Bayerl, P. S. and Paul, K. I. (2011). What determines intercoder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

[Beineke et al., 2004] Beineke, P., Hastie, T., Manning, C., and Vaithyanathan, S. (2004). An exploration of sentiment summarization. In *AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, USA. Technical Report SS-04-07.

[Birnbaum, 2004] Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, 55:803–832.

[Blackwood et al., 2003] Blackwood, N., Bentall, R., Ffytche, D., Simmons, A., Murray, R., and Howard, R. (2003). Self-responsibility and the self-serving bias: an FMRI investigation of causal attributions. *Neuroimage*, 20(2):1076–1085.

[Boldrini et al., 2012] Boldrini, E., Balahur, A., Martnez-Barco, P., and Montoyo, A. (2012). Using emotiblog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery*, 25(3):603–634.

[Callejas and López-Cózar, 2008] Callejas, Z. and López-Cózar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50(5):416–433.

[Cavicchio and Poesio, 2012] Cavicchio, F. and Poesio, M. (2012). The rovereto emotion and cooperation corpus: a new resource to investigate cooperation and emotions. *Language Resources and Evaluation*, 46(1):117–130.

[Clematide et al., 2012] Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., and Wiegand, M. (2012). Mlsa – a multi-layered reference corpus for german sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Cohn et al., 2008] Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–914.

[Craggs and Wood, 2004] Craggs, R. and Wood, M. M. (2004). A two dimensional annotation scheme for emotion in dialogue. In *AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications*, pages 44 – 49, Stanford, USA. Technical Report SS-04-07.

[Devillers et al., 2002] Devillers, L., Vasilescu, I., and Lamel, L. (2002). Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *ISLE Workshop on Dialogue Tagging*, Edinburgh, Scotland.

[Dyer, 1983] Dyer, M. (1983). The role of affect in narratives. *Cognitive Science*, 7(3):211–242.

[Eelen, 2001] Eelen, G. (2001). *A Critique of Politeness Theories*. St. Jerome.

[Eugenio, 2000] Eugenio, B. D. (2000). On the usage of kappa to evaluate agreement on coding tasks. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-00)*, Athens, Greece.

[Eugenio and Glass, 2004] Eugenio, B. D. and Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

[Forbes-Riley and Litman, 2004] Forbes-Riley, K. and Litman, D. J. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 201–208, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Gianti et al., 2012] Gianti, A., Bosco, C., Patti, V., Bolioli, A., and Caro, L. D. (2012). Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ($ES^3$ 2012)*, Istanbul, Turkey.

[Gunes et al., 2011] Gunes, H., Schuller, B., Pantic, M., and Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of the 1st International Workshop on Emotion Synthesis, Representation, and Analysis in Continuous Space (EmoSPACE 2011)*, pages 827–834, Santa Barbara, California, USA.

[Gut and Bayerl, 2004] Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proceedings of the Second International Conference for Speech Prosody 2004 (SP2004)*, Nara, Japan. Poster.

[Hasler, 2007] Hasler, L. (2007). From extracts to abstracts: Human summary production operations for computer-aided summarisation. In *Proceedings of the RANLP 2007 Workshop on Computer-Aided Language Processing (CALP)*, pages 11–18, Borovets, Bulgaria.

[Herzig et al., 2011] Herzig, L., Nunes, A., and Snir, B. (2011). An annotation scheme for automated bias detection in wikipedia. In *Proceedings of the Fifth Law Workshop (LAW V)*, pages 47–55, Portland, USA.

[Higgins and Bhatt, 2001] Higgins, N. C. and Bhatt, G. (2001). Culture moderates the self-serving bias: Etic and emic features of causal attributions in India and in Canada. *Social Behavior and Personality*, 29(1):49–62.

[Hijikata et al., 2007] Hijikata, Y., Ohno, H., Kusumura, Y., and Nishida, S. (2007). Social summarization of text feedback for online auctions and interactive presentation of the summary. *Knowledge-Based Systems*, 20(6):527–541.

[Hovy, 1990] Hovy, E. (1990). Pragmatics and natural language generation. artificial intelligence. *Artificial Intelligence*, 43(2):153–198.

[Hu and Liu, 2004a] Hu, M. and Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle USA.

[Hu and Liu, 2004b] Hu, M. and Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pages 755–760, San Jose, USA.

[Hunston, 1999] Hunston, S. (1999). Evaluation and the planes of discourse: Status and value in persuasive texts. In Hunston, S. and Thompson, G., editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*, pages 176–207. Oxford University Press.

[Jing and McKeown, 1999] Jing, H. and McKeown, K. R. (1999). The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, Berkeley, USA.

[Kameyama et al., 1996] Kameyama, M., Kawai, G., and Arima, I. (1996). A real-time system for summarizing human-human spontaneous spoken dialogues. In *Proceedings of the 4th International Conference on Spoken Language (ICSLP 96)*, volume 2, pages 681–684, Philadelphia, USA.

[Kaplan and Ruffle, 2004] Kaplan, T. and Ruffle, B. (2004). The self-serving bias and beliefs about rationality. *Economic Inquiry*, 42(2):237–246.

[Katagiri and Takahashi, 2003] Katagiri, Y. and Takahashi, T. (2003). Social summarization for semantic society. In *JSAI2003 Workshop "From Semantic Web to Semantic World"*.

[Keenan et al., 1977] Keenan, J., MacWhinney, B., and Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior*, 16(5):549–560.

[Knee and Zuckerman, 1996] Knee, C. R. and Zuckerman, M. (1996). Causality orientations and the disappearance of the self-serving bias. *Journal of Research in Personality*, 30(1):76–87.

[Krippendorff, 2004] Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. SAGE, 2nd edition.

[Lang, 1995] Lang, P. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5):372–385.

[Lee et al., 2002] Lee, C. M., Narayanan, S. S., and Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, Denver, Colorado, USA.

[Lehnert, 1982] Lehnert, W. (1982). Plot units: A narrative summarization strategy. In Lehnert, W. G. and Ringle, M. H., editors, *Strategies for Natural Language Processing*, pages 375–412. Erlbaum.

[Litman et al., 2003] Litman, D., Forbes, K., and Silliman, S. (2003). Towards emotion prediction in spoken tutoring dialogues. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*, pages 52–54, Edmonton, Canada. Short papers.

[Lloret et al., 2009] Lloret, E., Balahur, A., Palomar, M., and Montoyo, A. (2009). Towards building a competitive opinion summarization system: challenges and keys. In *Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium*, pages 72–77, Boulder, USA.

[Maks and Vossen, 2012] Maks, I. and Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688.

[Mann and Thompson, 1988] Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[Martin, 1999] Martin, J. R. (1999). Beyond exchange: Appraisal systems in english. In Hunston, S. and Thompson, G., editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*, pages 142–175. Oxford University Press.

[Miller, 2002] Miller, J. (2002). *An Introduction to English Syntax*. Edinburgh University Press Ltd, Edinburgh, Scotland.

[Mills, 2003] Mills, S. (2003). *Gender and Politeness*. Cambridge University Press.

[Moens, 2002] Moens, S. T. M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

[Morrison et al., 2007] Morrison, D., Wang, R., and Silva, L. C. D. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, pages 98–112.

[Mouka et al., 2012] Mouka, E., Giouli, V., Fotopoulou, A., and Saridakis, I. E. (2012). Opinion and emotion in movies: a modular perspective to annotation. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES$^3$ 2012)*, Istanbul, Turkey.

[Murray et al., 2005] Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal.

[Ortony et al., 1988] Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.

[Pápay et al., 2011] Pápay, K., Szeghalmy, S., and Szekrényes, I. (2011). Hucomtech multimodal corpus annotation. *Argumentum*, 7:330–347. Working paper.

[Park et al., 2011] Park, S.-B., Yoo, E., Kim, H., and Jo, G.-S. (2011). Automatic emotion annotation of movie dialogue using wordnet. In *Proceedings of the Third international conference on Intelligent information and database systems (ACIIDS'11)*, pages 130–139.

[Petukhova and Bunt, 2012] Petukhova, V. and Bunt, H. (2012). The coding and annotation of multimodal dialogue acts. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Picard, 1995] Picard, R. (1995). Affective computing. Technical Report 321, MIT Media Laboratory, Perceptual Computing Section, Cambridge, USA.

[Reithinger et al., 2000] Reithinger, N., Kipp, M., Engel, R., and Alexandersson, J. (2000). Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'2000)*, pages 310–317, Hong Kong, China.

[Roman, 2013] Roman, N. T. (2013). Resdial – coding description (v.1.0). Technical Report PPgSI-001/2012, School of Arts, Sciences and Humanities – University of São Paulo, São Paulo, SP – Brazil.

[Roman and Carvalho, 2010] Roman, N. T. and Carvalho, A. M. B. R. (2010). A multi-dimensional annotation scheme for behaviour in dialogues. In Kuri-Morales, A. and Simari, G. R., editors, *Proceedings of the 12th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2010)*, volume 6433 of *Advances in Artificial Intelligence*, pages 386–395, Baha Blanca, Argentina. Springer.

[Roman et al., 2006] Roman, N. T., Piwek, P., and Carvalho, A. M. B. R. (2006). Politeness and bias in dialogue summarization: Two exploratory studies. In Shanahan, J. G., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 171–185. Springer, Dordrecht, The Netherlands, 1st edition.

[Roman et al., 2013] Roman, N. T., Piwek, P., Carvalho, A. M. B. R., and Alvares, A. R. (2013). Introducing a corpus of human-authored dialogue summaries in portuguese. In *Proceedings of the 2013 International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, pages 692–701, Hissar, Bulgaria. ISSN 1313-8502.

[Rozin and Royzman, 2001] Rozin, P. and Royzman, E. (2001). Negativity bias, negativity dominance and contagion. *Personality and Social Psychology Review*, 5(4):296–320.

[Spertus, 1997] Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Innovative Applications of Artificial Intelligence (IAAI 97)*, pages 1058–1065.

[Taboada and Das, 2013] Taboada, M. and Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2). ISSN: 2152-9620.

[Takahashi and Katagiri, 2003] Takahashi, T. and Katagiri, Y. (2003). Telmea2003: Social summarization in online communities. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 03)*, pages 928–929, Fort Lauderdale, USA.

[Teufel and Moens, 1997] Teufel, S. and Moens, M. (1997). Sentence extraction as a classification task. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain.

[Turney and Littman, 2003] Turney, P. and Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

[van Deemter et al., 2008] van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., and Baumann, S. (2008). Fully generated scripted dialogue for embodied agents. *Artificial Intelligence*, 172(10):1219–1244.

[Vieira and Poesio, 2000] Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

[Watts, 2003] Watts, R. (2003). *Politeness*. Cambridge University Press.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

[Wilson et al., 2009] Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

[Yeh et al., 2005] Yeh, J.-Y., Ke, H.-R., Yan, W.-P., and Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management*, 41(1):75–95.

[Zechner and Waibel, 2000] Zechner, K. and Waibel, A. (2000). Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000*, pages 968–974, Saarbruecken, Germany.