Social validation of collective annotations: Definition and experiment

Guillaume Cabanac [a,*], Max Chevalier [a,b], Claude Chrisment [a], Christine Julien [a]

[a] *Université de Toulouse, IRIT UMR 5505 CNRS, 118 route de Narbonne, F-31062 Toulouse cedex 9, France*

[b] *Université de Toulouse, LGC ÉA 2043, IUT Paul Sabatier, 129 avenue de Rangueil, BP 67701, F-31077 Toulouse cedex 4, France*

Abstract

People taking part in argumentative debates through collective annotations face a highly cognitive task when trying to estimate the group's global opinion. In order to reduce this effort, we propose in this paper to model such debates prior to evaluating their "social validation." Computing the degree of global confirmation (resp. refutation) enables the identification of consensual (resp. controversial) debates. Readers as well as prominent Information Systems may thus benefit from this information. The accuracy of the social validation measure was experimented through an online study conducted with 121 participants. We compared their human perception of consensus in argumentative debates with the results of the three proposed social validation algorithms. Their efficiency in synthesizing opinions is shown, as their accuracy is up to 84 %.

Key words: Collective Annotation, Consensus, Argumentation, Human Perception, Experiment

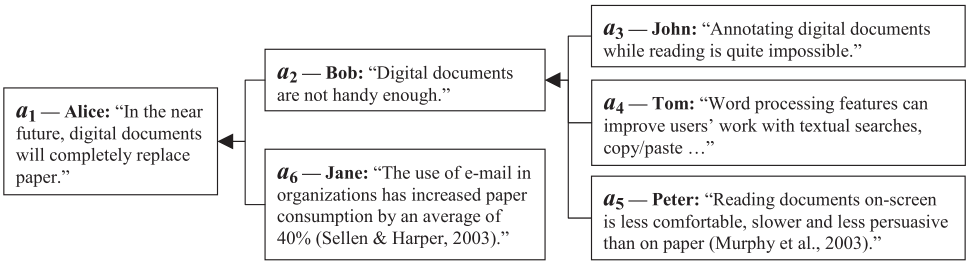
1. Introduction and motivations

Annotating paper documents is a common activity practiced since the early Middle Ages (Fraenkel and Klein, 1999). Field studies show that readers still make extensive use of annotations nowadays (Marshall, 1998; Wolfe and Neuwirth, 2001). Although seeming insignificant, they actually allow key purposes such as "active reading" by supporting critical thinking while reading (Adler and van Doren, 1972), learning by facilitating document appropriation, and proofreading, among many others.

With the widespread adoption of digital documents both in the workplace and at home, people felt frustrated at not being able to annotate them (Sellen and Harper, 2003, p. 96). Such a need led both research labs and companies to design a plethora of annotation systems—mostly targeting Web documents—since the 1990's (Wolfe, 2002).

[*] Corresponding author.
Email addresses: `Guillaume.Cabanac@irit.fr` (Guillaume Cabanac), `Max.Chevalier@irit.fr` (Max Chevalier), `Claude.Chrisment@irit.fr` (Claude Chrisment), `Christine.Julien@irit.fr` (Christine Julien).

At first, they implemented the usual paper-based annotation functions, mainly for a personal use. Then, taking advantage of modern computer storage and networking capabilities, annotation systems provided novel features for a collective use. In particular, they enabled annotation sharing through dedicated servers, so that users could view and access them in context, i.e. within the annotated document. Moreover, the subsequent readers may later reply to any annotation or any reply in turn, thus forming a debate anchored to the discussed passage. Such a debate is also called a "discussion thread" (e.g. Fig 1).



a_1 — **Alice:** "In the near future, digital documents will completely replace paper."

a_2 — **Bob:** "Digital documents are not handy enough."

a_3 — **John:** "Annotating digital documents while reading is quite impossible."

a_4 — **Tom:** "Word processing features can improve users' work with textual searches, copy/paste …"

a_5 — **Peter:** "Reading documents on-screen is less comfortable, slower and less persuasive than on paper (Murphy et al., 2003)."

a_6 — **Jane:** "The use of e-mail in organizations has increased paper consumption by an average of 40% (Sellen & Harper, 2003)."

Fig. 1. Example of an argumentative discussion thread comprising six arguments.

As an asynchronous way of communicating, collective annotations (i.e. annotations along with their discussion threads) are useful at two levels:

- From the *readers' point of view*, collective annotations enable them to discuss document passages in context. This is an advantage in comparison with Internet bulletin boards and forums where one needs to explain the context of his/her post to be understood. Moreover, feedback about a given annotated document is directly accessible to readers whereas it would be scattered over multiple sources otherwise. Besides being useful for readers, authors also benefit from collective annotations as they can improve their documents by taking into account the associated remarks.
- For *systems*, collective annotations are spontaneous contributions from the audience. They are processed for improving existing features and for providing new ones, as they are supposed to enhance original contents. Concerning relevance feedback for instance, the document passages annotated by a reader proved to be more effective than his/her explicit relevance judgments (Golovchinsky et al., 1999). Going by this result, XLibris displays links pointing to recommended documents next to the reader's annotations. In addition, Fraenkel and Klein (1999) envisioned that annotations may improve ad-hoc retrieval recall and precision, as the annotator generally explains or discusses a passage in his own words. Later, Agosti and Ferro's FAST system completed a search result by merging the documents whose discussion thread contents match the query (Agosti and Ferro, 2005, 2007). Finally, the POLAR system considered collective annotations for answering queries, intending to also reflect the polarity of arguments—positive or negative (Frommholz and Fuhr, 2006).

Although annotation systems have been developed for nearly two decades, annotating digital documents is not widespread whereas all the evidence points to the fact that users need such a feature (Sellen and Harper, 2003, p. 96). Thus, research works (Ovsiannikov et al., 1999; Vasudevan and Palmer, 1999; Wolfe and Neuwirth, 2001; Brust and Rothkugel, 2007) investigated the obstacles holding back annotation systems adoption. In this paper, we highlight three main usability issues regarding:

(i) How difficult it is for readers to understand discussion threads,

(ii) How the current in-context annotation visualization techniques are not scalable,

(iii) How Information Retrieval (IR) systems does not capitalize on the opinions from the audience.

The first issue concerns reading a discussion thread for acquiring other readers' feedback. One may need to synthesize the whole arguments: in the end, the social group may confirm or refute a disputed annotation's argument, especially in argumentative debates (e.g. Fig. 1). Synthesizing the group's global opinion from individuals' arguments requires a reader to achieve two tasks. He first needs to identify the opinion of each argument—ranging from a confirmation to a refutation, via a neutral opinion. Then, he has to aggregate the opinions in a recursive way, from the deepest arguments to the root annotation of the hierarchy corresponding to the debate. Both of these tasks require a substantial cognitive load from readers, especially for large forums that tend to be commonplace on the Web (Dave et al., 2004). Shirky (2008, p. 98) points how prominent this issue is: *"Even in a medium that allowed for perfect interactivity for all participants (something we have a reasonable approximation of today), the limits of human cognition will mean that scale alone will kill conversation."* The second issue refers to the fact that annotations clutter the display because each one is usually rendered as an icon, inserted into the document, next to the annotated passage. For instance, the W3C's Annotea/Amaya (Kahan et al., 2002) annotation system inserts the yellow pen icon "✏" to identify an annotated passage. Browsing some established Web sites with this annotation system[1] clearly illustrates this usability issue, i.e. the more a document is annotated, the more it is difficult to read. Thirdly, we identified a limit concerning how IR systems use annotations for improving search results by integrating the audience's contributions. Indeed, annotations are regarded as an objective piece of information, although they may be wrong and disputed in their associated discussion threads.

Regarding current systems, a common denominator of these issues is that nothing distinguishes one annotation from another. However, the social group that took part in each discussion thread expressed opinions about the annotation contents. These opinions may be worth considering to distinguish between sound annotations and nonsense. Nevertheless, no work to date addressed the aggregation of such opinions, as far as we know. Though Surowiecki (2005, p. 75) underlines how relevant and useful measuring the global opinion of the group may be: *"If*

[1] http://www.irit.fr/~Guillaume.Cabanac/annotation/demoAmaya.wmv

[a group] *has a means of aggregating all* [its] *different opinions, the group's collective solution may well be smarter than even the smartest person's solution."* For that purpose, this paper defines a measure called the "social validity" of a collective annotation. This measure exploits individuals' arguments gathered in a discussion thread to evaluate the social group's gradual degree of consensus or controversy.

The remainder of the paper is organized as follows: Sect. 2 defines the annotation social validation process, and details three alternatives, i.e. algorithms aiming to compute it. In order to assess if our proposals succeed in reflecting human perception of consensus in argumentative discussions, Sect. 3 details the protocol for evaluating the proposed algorithms. We report on the associated online experiment that rallied 121 people from 13 countries. Lastly, Sect. 5 discusses the results of this experiment before concluding the paper.

2. Measuring the "social validity" of collective annotations

We define the *social validation* of a collective annotation as the process of aggregating the opinions extracted from the arguments of its associated discussion thread. The resulting value, called *social validity*, represents the global opinion of the social group who took part in the debate. One may interpret this value as the degree of consensus (i.e. the group globally agrees with the annotation) or controversy (i.e. the group globally disagrees with the annotation) reached within the debate. For instance, annotation a_1 on Fig. 1 is refuted by a_2 and a_6, but a_2 is in turn refuted by one out of three of its replies. On the whole, the social validity of a_1 shows a strong controversy, as a_1 is nearly refuted completely by the social group. This section first underlines contexts of use for the proposed social validity measure. Then it formally defines the *annotation* and *discussion thread* concepts before introducing algorithms that compute their social validation.

2.1. *Contexts of use for the social validation process*

The potential use of the collective annotation practice ranges over various types of Information Systems (IS). A major advantage lies in the fact that it allows people to exchange ideas in context, in an asynchronous and structured way. We underline below how prominent IS may benefit from the proposed social validation process.

Annotation systems. The social validation process associated with a specific visualization, as proposed in (Cabanac et al., 2007a), enables readers to distinguish between consensual, controversial, and neutral annotations. This display adaptation intends to overcome the scalability issue stated beforehand. Indeed, the annotation display can be personalized regarding the reader's preferences, for either emphasizing consensual annotations (globally confirmed or refuted) or ongoing disputes (neutral social validity resulting from the coexistence of refutations and confirmations).

Decision support systems. Managers and knowledge workers using data warehouses achieve various kinds of analyses, which are represented as specific documents called "multidimensional tables." They commonly annotate hard copies of such tables by marking the cells and commenting on the data. Foshay et al. (2007) underline the great need for the transposition of this paper-based practice to digital counterparts. That is why we introduced the "decisional annotation" concept to allow analysts to annotate tables, to share comments between coworkers, and to foster opinion exchange within discussion threads (Cabanac et al., 2007b). From this perspective, the social validation process enables people to identify (dis)agreed analyses. Moreover, the resulting annotations and discussions are stored in a shared expertise memory in order to keep a record of the decision making process.

Information retrieval systems. The social validation process may especially benefit three IR fields, as we suggested in (Cabanac et al., 2007a): contextual retrieval, user profiling, and interests/trends mining. Computing the global opinion expressed by the group who discussed a given passage may lead to improve contextual IR results. Indeed, such systems may give more importance to validated document granules than to disputed passages. Annotations may also be exploited as trails of users' activities. In particular, people's involvement within discussion threads may be mined for updating their user's profile so as to reflect a significant part of their social interactions. Finally, the analysis of ongoing debates may enable the identification of popular, controversial, as well as disputed documents and Web pages. Thus, this may be an additional indicator for finding an audience's current interests and trends.

Collaborative writing platforms. With the ever-growing popularity of the social web, wiki-based platforms turned worldwide collaborative writing into reality. Indeed, any reader can also become a writer by modifying any article published on a wiki. With such systems, requests for change and comments remain in a different page from the discussed article. This makes it hard for casual readers to consider them while reading the main article. Kittur et al. (2007) found a decrease in constructive contributions, for an increase in "revert wars" regarding the prominent Wikipedia encyclopedia. Their study suggests that such critical conflict points happen when too few people are involved in the discussion, leading to an insufficient diversity in point of views. However, the replacement of separate talk pages by annotations in context would enable readers to identify discussed passages clearly. As a beneficial side-effect, this may encourage people to get involved. Moreover, the conflict model for identifying controversial articles proposed in (Kittur et al., 2007) may be extended with the proposed social validation process. Lastly, it may also be exploited for Wikipedia's decision making-process, which is based on a "rough consensus" policy (Butler et al., 2008).

2.2. *Modeling collective annotations along with discussion threads*

An *annotation* (Def. 1) is formulated by a user on a resource, e.g. a text, a picture. Concerning its location, anchoring points may be specified within the resource or on an already existing annotation. In the latter case, the fact of *replying* to it forms a *discussion thread* (Def. 2). This section defines these concepts and illustrates them with a concrete example, see Fig. 2.

Definition 1. We define an *annotation* as a collective object that can be used for argumentative purposes. The "collective" adjective refers to the fact that an annotation is shared among multiple users of the annotation system. As a result, it may be consulted by any of them. The "argumentative" facet of an annotation comes from the ability to confirm or refute its contents within a discussion thread. In line with (Cabanac et al., 2005), we modeled such an annotation as a pair $\langle OD, SI \rangle$ where:

(i) The OD part represents Objective Data created by the annotation system. This consists of the following *mandatory* attributes for an annotation:

- Its *identification* by a unique identifier such as a Uniform Resource Identifier (URI).
- Its *author's identity* which enables to get information about him/her, e.g. name, email.
- Its *timestamp* which enables to organize arguments of discussion threads chronologically.
- Its *anchoring points* which unambiguously specify its locations within the annotated resource. Various approaches have been published for semi-structured documents, e.g. Annotea (Kahan et al., 2002) relies on XPointer (DeRose et al., 2002).

(ii) The SI part represents Subjective Information formulated by the user when he creates an annotation. This consists of the following *optional* attributes for an annotation:

- Its *contents*, a textual comment or audio recording for instance.
- Its *visibility* stored for privacy concerns, e.g. private, public, specified users only.
- Its author's *expertise* which may be useful for further readers, as people tend to trust experts more than novices (Marshall, 1998).
- The list of *references* provided by the annotator in order to justify its opinion. A book reference, a citation, a URL ... may be provided for that purpose.
- Various *annotation types* that extend the works of Kahan et al. (2002) by providing opinion types. We have grouped them into the two classes represented in Tab. 1. Annotation types reflect the semantics of the contents (modification, example, and question) as well as the semantics of the author's opinion (refute, neutral, or confirm). By combining these types, authors may express subjective point of views that are

gradual, e.g. consider a $\mathscr{R}$-typed *versus* $\mathscr{RE}$-typed annotation. Concretely, these types may be inferred from annotation contents by Natural Language Processing (Pang et al., 2002; Liu, 2007) and then validated by their author.

Table 1

Available annotation types for a collective annotation.

Class	Comment			Opinion (exclusive)		
Type	question	modification	example	confirmation	neutral	refutation
Notation	$\mathscr{Q}$	$\mathscr{M}$	$\mathscr{E}$	$\mathscr{C}$	$\mathscr{N}$	$\mathscr{R}$

Note that Agosti and Ferro (2007) recently published a "formal model of annotations of digital content" aiming to synthesize previous works on annotation models initiated since the 1990's, such as (Ovsiannikov et al., 1999; Agosti and Ferro, 2006; Frommholz and Fuhr, 2006). It includes the attributes of our model (Cabanac et al., 2005) except for the *expertise* attribute. For brevity concern and because of not being necessary in the remainder of this paper, we do present the formal notation and refer the reader to (Agosti and Ferro, 2007) for more details.

Definition 2. A *discussion thread* is a tree rooted on an annotation. This specific root annotation may be recursively annotated by *replies*. Note that replies are represented in a chronological order using their timestamps, i.e. their creation date and time.

Example 3. Figure 2 represents a discussion thread where the contents of the refuting a annotation and its replies r_i are given in Tab. 2. Note that $x \leftarrow y$ means that y confirms x. Conversely $x \not\leftarrow y$ means that y refutes x. Moreover, the hierarchy is constrained by annotation timestamps, e.g. $timestamp(r_1) \leqslant timestamp(r_2) \leqslant timestamp(r_3)$.

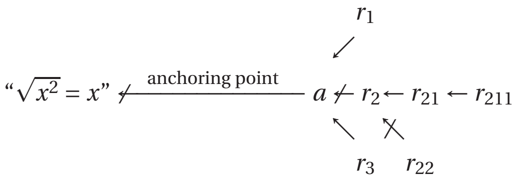


Fig. 2. Discussion thread about the expression "$\sqrt{x^2} = x$."

2.3. *Measuring the social validity of collective annotations*

This section describes three approaches and associated algorithms that compute the social validity $v(a) \in [-1, 1]$ of an annotation a. This continuous function reflects the intrinsic opinion of a as well as the global opinion expressed by its replies. Note that opinions are defined in Tab. 1. Regarding its interpretation, $v(a) \to 0$ means either that a has no reply or that replies expressing confirmation and refutation are balanced. Moreover $v(a) \to 1$ indi-

Table 2

Arguments associated with the mathematical discussion shown in Fig. 2.

Name	Type	Annotator's comment
a	$\mathscr{R}\mathscr{M}\mathscr{E}$	Wrong formula: check this counterexample: $\sqrt{(-2)^2} \neq -2$ Consider the following modification: $\sqrt{x^2} = \lvert x\rvert$.
r_1	$\mathscr{C}\mathscr{E}$	OK, for example $\sqrt{(-4)^2} = \lvert -4\rvert = 4$.
r_2	$\mathscr{R}$	This high school lesson only considers positive numbers ...
r_3	$\mathscr{C}$	More generally $\forall (x,n) \in \mathbb{R} \times \mathbb{R}^* \quad \sqrt[n]{x^n} = \lvert x\rvert$.
r_{21}	$\mathscr{C}\mathscr{M}$	Then you should precise $\forall x \in \mathbb{R}_+ \quad \sqrt{x^2} = x$.
r_{22}	$\mathscr{R}\mathscr{E}$	May be confusing when quickly and superficially read!
r_{211}	$\mathscr{C}\mathscr{M}$	$\mathbb{R}$ is unknown in high school, use "positive numbers" instead.

cates that a is totally confirmed by its replies; conversely $v(a) \to -1$ means that a is totally refuted by its replies. As a result a consensus (either refuting of confirming) is identified when $|v(a)| \to 1$. In order to define how $v(a)$ evolves, we consider the combination of opinion values for the annotation and its replies: Tab. 3 describes the four possible cases. For instance, Case 2 shows that replies which globally refute ($\mathscr{R}$) an annotation a (which is $\mathscr{C}$-typed) lowers its social validity: $v(a) \to 0$.

Table 3

Social validity of a parent annotation a regarding the opinion types of its replies.

	Case 1	Case 2	Case 3	Case 4
Opinion of annotation a	$\mathscr{C}$	$\mathscr{C}$	$\mathscr{R}$	$\mathscr{R}$
Global opinion of replies to a	$\mathscr{C}$	$\mathscr{R}$	$\mathscr{C}$	$\mathscr{R}$
Social validity $v(a)$	$v(a) \to 1$	$v(a) \to 0$	$v(a) \to -1$	$v(a) \to 0$

In order to compute $v(a)$, we have explored three distinct approaches. The first one considers the κ coefficient (Fleiss et al., 2003) provided by Social Sciences. Establishing that κ is not suitable for our concern, the second approach is based on an empirical recursive scoring algorithm (Cabanac et al., 2005). As a third approach, we intended to ground social validation into a more formal setting. Therefore, Cabanac et al. (2007a) extended the Bipolar Argumentation Framework proposed in (Cayrol and Lagasquie-Schiex, 2005a,b) in order to compute the social validity of collective annotations.

2.3.1. *First approach: statistical measure of inter-rater agreement*

Cohen's kappa coefficient $\kappa \in [0,1]$ measures the agreement among $n = 2$ "raters" (persons providing a rating) which split up N items in k mutually exclusive categories (Cohen, 1960). A generalization of this coefficient called Fleiss' κ may be applied for $n \geqslant 2$ raters (Fleiss, 1971). The value of the kappa coefficient $\kappa = \frac{P(A)-P(E)}{1-P(E)}$ varies according to $P(A)$, the agreement among the n raters as well as $P(E)$ representing chance agreement. Different ranges of values for κ have been characterized with respect to the degree of agreement they suggest. Concerning content analysis, a standard methodology of the Social Sciences, a value such that $\kappa > 0.8$ indicates a good degree of agreement (Krippendorff, 1980).

This coefficient does not take into account the fact that a rating may be disputed by other raters, thus forming a discussion tree. As a result, a κ-based approach did not seem to be suited for solving our concern. As a discussion thread is a tree of annotations, the following section proposes a recursive scoring algorithm as a second approach.

2.3.2. *Second approach: empirical recursive scoring algorithm*

In a second attempt, we defined a recursive algorithm that computes the social validation of an annotation (Cabanac et al., 2005). This algorithm is a twofold process working on the annotations of a discussion thread, i.e. the root annotation and its replies. The first step consists of evaluating the intrinsic *agreement* of each annotation using two parameters. *i*) The *confirmation* value $c(a) \in [-1,1]$ of an annotation a is based on its opinion types. Indeed, the confirmation $\mathscr{C}$-typed implies a positive confirmation value, whereas the refutation $\mathscr{R}$-typed implies a negative confirmation value. A gradual evaluation is obtained by considering the combination of the different types, see Fig. 3.

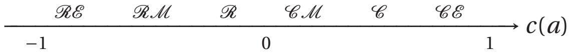


Fig. 3. *Confirmation* value of an annotation a regarding its types.

ii) We assumed that the *agreement* of an annotation has to be increased considering the involvement of its author. Indeed, when someone provides a comment as well as some references (i.e. optional information) then he makes an extra cognitive effort in order to justify his opinion. Such information brings an additional value: compare a simple refutation (e.g. "This is wrong!") with commented references that bring proof of the refutation (e.g. "Regarding X and as it can be seen on Y, I think that it is false because ..."). Concretely, if A denotes the set of annotations, the $i_c : A \rightarrow [0,1]$ function (1) reflects the presence of a comment. Moreover, the $i_r : A \rightarrow [0,1]$ function (2) varies according to the number of provided references. Note that the dot notation refers to the annotation model of Def. 1, e.g. $a.contents$ means "the contents of annotation a."

$$i_c(a) = \begin{cases} 0 & \text{if } ||a.contents|| = 0 \\ 1 & \text{else} \end{cases} \qquad (1)$$

$$i_r(a) = \frac{|a.references|}{1 + \max_{x \in A} |x.references|} \qquad (2)$$

The *agreement* of an annotation is evaluated by the $a : A \rightarrow [0,1]$ function (3) that takes into account the types as well as the contents (comment and references) and the confirmation value $c(a)$ of an annotation. The $\alpha, \beta \in [0,1]$ parameters allow the adjustment of the relative weights of these two functions.

$$a(a) = \frac{c(a)\ (1 + \alpha \cdot i_c(a))\ (1 + \beta \cdot i_r(a))}{(1 + \alpha)\ (1 + \beta)} \qquad (3)$$

The second step consists in combining the intrinsic agreement of the annotation with the global agreement expressed by its replies. This combination called *social validation* is computed by the $v : A \rightarrow [-1,1]$ function (4). It is interpreted as follows: $v(a) = 0$ if a is the root ($a.ancestor = \lambda$) of an empty discussion thread ($|a.replies| = 0$) which means that a is neither confirmed nor refuted. Otherwise, $v(a)$ is computed according to its *agreement* value and a *synthesis* of its replies. To do that, the s function returns a negative value when the replies globally refute the parent item, or a positive value otherwise. Finally, the $\gamma \in [0,1]$ parameter allows the adjustment of the discussion thread impact on the measure of the social validity of a.

$$v(a) = \begin{cases} 0 & \text{if } a.ancestor = \lambda \wedge |a.replies| = 0 \\ \frac{1}{2} \cdot a(a) \cdot \big(1 + \gamma \cdot s(a)\big) & \text{else} \end{cases} \qquad (4)$$

The *synthesis* $s : A \rightarrow [-1,1]$ function (5) is important because it takes into account the replies of an annotation as a whole. We based this function on a weighted mean that gives a prominent role to replies qualified with a greater expertise (the range of the expertise attribute being strictly positive). We also increased the value of the *synthesis* considering the number of replies: we give more importance to annotations as they sparked off numerous replies.

$$s(a) = \begin{cases} 1/\gamma & \text{if } |a.replies| = 0 \\ \frac{\sum\limits_{r \in a.replies} v(r) \cdot r.expertise}{\sum\limits_{r \in a.replies} r.expertise} \left[1 + \ln\left(1 + \frac{|a.replies|}{m(1+l(a))}\right) - \ln(2)\right] & \text{else} \end{cases} \qquad (5)$$

In (5), $a.replies$ denotes the set of replies associated with the a annotation. Moreover, the $l : A \rightarrow \mathbb{N}_+$ function returns the level of a given annotation in the discussion thread, the root annotation being at level 0. Finally, the $m : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ function returns for $m(l)$ the maximum number of replies at the l^{th} level of the tree corresponding to the discussion thread. To sum up, if an annotation a belongs to the discussion thread and does not have any reply, its *social validity* corresponds to the *agreement* value: $|a.replies| = 0 \Rightarrow v(a) = a(a)$. On the contrary, if a

has replies, the *synthesis* value relies on the weighted mean of their social validity (A) multiplied by an expression (B) increasing with the number of replies. B takes into account the maximum number of replies existing at the same level than the level of a. For instance, if a (level l) has n replies and if the maximum number of replies for the annotations at the l^{th} level is denoted $N = m(l)$, then $B = 1 + \ln\left(1 + \frac{n}{N}\right) - \ln(2)$. Note that the upper bound of B is reached when $n = N \Rightarrow B = 1$. Therefore $B \in \left[\ln\left(\frac{e\cdot(N+1)}{2N}\right), 1\right]$, we used a natural logarithm for reducing differences between small and large values of N. As $B \leqslant 1$, it cannot increase the value of A.

The $v(a)$ value of the *social validation* associated with the a annotation allows to evaluate it according to the *agreement* and *synthesis* functions. Indeed, $|v(a)| \to 1$ evidences an annotation a that contains replies which globally agree with it. According to the sign of $v(a)$, one can conclude that people validate a confirm $\mathscr{C}$-typed (resp. a refute $\mathscr{R}$-typed) annotation when $v(a) \to 1$ (resp. $v(a) \to -1$). The social validation algorithm presented in this section is mostly based on heuristics and empirically instantiated parameters. In order to go beyond this first empirical approach, the following section describes another proposal for computing social validation. This approach is grounded on formal research from the Artificial Intelligence domain (Cayrol and Lagasquie-Schiex, 2005a,b). Its main advantage lies in the fact that it builds on the theoretical principles of argumentation; moreover it has been used in numerous contexts, e.g. medical decision making.

2.3.3. *Third approach: formal bipolar argumentation framework extension*

Dung (1995) models an argumentation as the pair $\langle A, R\rangle$ where A is a set of arguments and R is a binary relation on A^2 called a defeat relation. The so-called resulting "argumentation framework" may be represented as a directed graph whose vertices are arguments and edges represent their targets. Then, the identification of attack and defense branches enables one to decide on the acceptability of an argument—which is a binary value: acceptable or not acceptable—regarding conflict-free and collective defense sets. Later, Cayrol and Lagasquie-Schiex (2005a) noticed that most work to date on argumentation considered a single type of interaction between arguments: the *attack*. However, previous works (Karacapilidis and Papadias, 2001) suggested that, for numerous concrete contexts, another type of interaction must be considered in order to properly represent knowledge. This other type of interaction is *defense*. Thus, regarding attack and defense arguments, Cayrol and Lagasquie-Schiex (2005a) extend Dung's argumentation framework (Dung, 1995) and define a Bipolar Argumentation Framework (Def. 4).

Definition 4. A *Bipolar Argumentation Framework* (BAF) is represented as a triple $\langle A, R_{app}, R_{att}\rangle$ where:

– $A = \{a_1, \ldots, a_n\}$ is a set of arguments,

– R_{app} is a binary *support* relation on A^2. The pair $(a_i, a_j) \in R_{app}$ is denoted by $a_i \to a_j$,

– R_{att} is a binary *defeat* relation on A^2. The pair $(a_i, a_j) \in R_{att}$ is denoted by $a_i \not\rightarrow a_j$.

Besides Def. 4, (Cayrol and Lagasquie-Schiex, 2005a) defined a gradual valuation v on their framework, following 3 principles. **P1**: the valuation of an argument depends on the values of its direct defeaters and of its direct supporters. **P2**: if the quality of the support (resp. defeat) increases then the value of the argument increases (resp. decreases). **P3**: if the quality of the supports (resp. defeats) increases then the quality of the support (resp. defeat) increases. Regarding these principles, the authors set $a \in A$ with $R^-_{app}(a) = \{b_1, \ldots, b_p\}$ and $R^-_{att}(a) = \{c_1, \ldots, c_q\}$.[2] Finally, they defined a gradual valuation as the application (6) $v : A \rightarrow V$.[3]

$$v(a) = g\left(h_{app}\left(v(b_1), \ldots, v(b_p)\right), h_{att}\left(v(c_1), \ldots, v(c_q)\right)\right) \tag{6}$$

In (6), the function h_{app} (resp. h_{att}) : $V^* \rightarrow H_{app}$ (resp. $V^* \rightarrow H_{att}$) evaluates support (resp. attack) upon an argument. The function $g : H_{app} \times H_{att} \rightarrow V$ with $g(x, y)$ is increasing on x and decreasing on y. In addition, the function h (either $h = h_{app}$ or h_{att}) must satisfy the three following conditions:

C1 if $x_i \geqslant x'_i$ then $h(x_1, \ldots, x_i, \ldots, x_n) \geqslant h(x_1, \ldots, x'_i, \ldots, x_n)$,

C2 $h(x_1, \ldots, x_i, \ldots, x_n, x_{n+1}) \geqslant h(x_1, \ldots, x_i, \ldots, x_n)$,

C3 $h() = \alpha \leqslant h(x_1, \ldots, x_i, \ldots, x_n) \leqslant \beta$ for all $x_1, \ldots, x_i, \ldots, x_n$.

Cayrol and Lagasquie-Schiex (2005a) proposed two instances for this generic evaluation. In the first one, argument values are aggregated by retaining the maximum of *direct* attacks and supports, i.e. $h_{att} = h_{app} = \max$. This first approach is not acceptable for our context because it does not take into account the whole hierarchy of expressed arguments, i.e. replies to replies. As a result, we rejected this first approach and opted for the second one, where an instance is characterized by the following parameters: $V = [-1, 1]$, $\alpha = 0$ and $\beta = +\infty$ thus leading to $H_{app} = H_{att} = [0, +\infty]$. Moreover, $h_{app} = h_{att} = \sum_{i=1}^{n} \frac{x_i+1}{2}$ and $g(x, y) = \frac{1}{1+y} - \frac{1}{1+x}$.

As a direct application, we compute the social validity of an annotation thanks to (6), which is applied on the BAF created from the annotation's discussion thread (Ex. 5). Its A set contains nodes of the discussion thread; pairs of the R_{app} (resp. R_{att}) set are defined by annotations and reactions of the confirm $\mathscr{C}$ (resp. refute $\mathscr{R}$) type along with their parent targets.

Example 5. The discussion shown in Fig. 2 is modeled by $\langle A, R_{app}, R_{att} \rangle$, a Bipolar Argumentation Framework where $A = \{a, r_1, r_2, r_3, r_{21}, r_{22}, r_{211}\}$ figures the annotation a and its replies r_i. Relations between annotations are expressed by $r_{app} = \{(r_1, a), (r_3, a), (r_{21}, r_2), (r_{211}, r_{21})\}$ for the $\mathscr{C}$ type whereas $r_{att} = \{(r_2, a), (r_{22}, r_2)\}$ reflects the $\mathscr{R}$ type. The computed social validity of this discussion thread is $v(A) = 0.152$.

[2] $R^-_{app}(a)$ (resp. $R^-_{att}(a)$) are direct supports (resp. defeats) of the a argument.

[3] V denotes a completely ordered set, and V^* denotes the set of finite sequences of element V including the empty sequence.

The gradual evaluation $v(A)$ of this example does not take into account some available argument data. Indeed some subjective information (SI, cf. Def. 1) associated with the nodes of the discussion thread are not considered, e.g. “comment” class types, expertise, comment and references. This is the reason why we extended the BAF in (Cabanac et al., 2007a), by redefining the v application (6). Thus, we provided the $v' : A \to V$ application (7) such that:

$$v'(a) = g\big(h_{app}\big(i(b_1)\cdot v'(b_1),\ldots,i(b_p)\cdot v'(b_p)\big), h_{att}\big(i(c_1)\cdot v'(c_1),\ldots,i(c_q)\cdot v'(c_q)\big)\big) \tag{7}$$

We introduced the $i : A \to I$ function (8) as a parameter of $h_{app} : V^* \to H_{app}$ and $h_{att} : V^* \to H_{att}$ in order to measure the intrinsic value of an argument by taking into account n criteria. Our choice for the second evaluation instance required that $V = [-1,1]$. Therefore, we defined $I = [0,1]$ in order to comply with $\forall a \in A \quad i(a)\cdot v'(a) \in V$. Moreover, the n coefficients $\pi_i \in [0,1]$ defined such that $\sum_{i=1}^{n} \pi_i = 1$ allow the adjustment of the relative importance of the n criteria evaluated by the $f_i : A \to F_i \subseteq \mathbb{R}_+$ functions. Note that the sup function returns the least upper bound of the f_i functions domain of definition. The $\delta \in [0,1]$ coefficient allows the adjustment of the n criteria global impact on the v' evaluation, note that $\delta = 0 \Longrightarrow v'(A) = v(A)$.

$$i(x) = \delta \cdot \sum_{i=1}^{n} \frac{\pi_i \cdot f_i(x)}{\sup(F_i)} \tag{8}$$

Regarding the collective annotation model (Def. 1) we identified $n = 4$ criteria that may be taken into account to evaluate an argument: *expertise* and *agreement* of its supporters and defeaters, as well as its author’s implication in terms of *comments* and *references*. Thus, the f_1 function increases according to the *expertise* associated with the evaluated argument; we proposed a five-point expertise scale: novice ≺ beginner ≺ intermediary ≺ confirmed ≺ expert. Then, the f_2 function associates a real value representing the annotator’s *agreement* with each combination of types from the “Comment” class. For example, someone giving an example ($\mathscr{E}$) agrees more than someone who proposes a modification ($\mathscr{M}$).

$$\xrightarrow[F_2]{\mathscr{M}\mathscr{Q}\mathscr{E}\qquad \mathscr{M}\mathscr{Q}\qquad \mathscr{M}\mathscr{E}\qquad \mathscr{Q}\mathscr{E}\qquad \mathscr{M}\qquad \mathscr{Q}\qquad \mathscr{E}} f_2(a)$$

The f_3 function evaluates the annotator’s implication regarding the existence of a *comment* on a given annotation. Finally, the f_4 function increases with the number of *references* cited by an annotation: $f_4(x)$ is the ratio between the number of references contained by x and the maximum number of references by annotation. Regarding the implementation of the proposed approach, we set $\delta = 1$ in order to maximize the impact of the n criteria on the evaluation of arguments. While foreseeing to experiment other weightings, we intuitively defined in (Cabanac et al., 2007a) $\pi_2 = \pi_3 = \frac{1}{3}$ and $\pi_1 = \pi_4 = \frac{1}{6}$ so that it mostly takes into account comments as well as agreement of attacks and supports. In conclusion, an annotation is the subject of social consensus, i.e. it is *socially validated* when its reactions globally express a concordant opinion, either refutation or confirmation. This is obtained when $\left(v'(A) \to 1\right) \vee \left(v'(A) \to -1\right) \Longleftrightarrow \left|v'(A)\right| \to 1$. Such results really represent a semantical-based, in-context synthesis of social groups' opinions.

3. Methodology for experimenting with the social validation algorithms

We designed an original experiment for validating the proposed algorithms. It consists in checking human perception of consensus against social validity values computed by the algorithms, in order to assess their relevance. For that purpose, we first comment on the inadequacy of existing corpora stemming from prominent IR evaluation campaigns in Sect. 3.1, thus justifying the need to build an appropriate corpus. Then, Sect. 3.2 details the tasks achieved by the participants thanks to the software platform (Sect. 3.3) we designed for that purpose. Finally, Sect. 3.4 discusses the recruitment of the 121 participants who voluntarily took part to the experiment.

3.1. *Building a corpus of argumentative debates*

The proposed experiment intends to compare the social validation algorithms to human perception. This requires ① a corpus of discussion threads whose arguments are labeled with opinions, as well as ② the "standard" observed human perception value to be compared to algorithmic results. A typical example would be ① the debate in Fig. 1 along with ② the perceived social validity $v = 0.82$, for instance. Unfortunately, previous works noticed the lack of such a discussion thread corpus for validation purposes (Agosti and Ferro, 2006, 2007; Frommholz and Fuhr, 2006). Moreover, resources stemming from Opinion Mining evaluation (e.g. Blog Track at TREC, Multilingual Opinion Analysis at NTCIR) are not suitable as they only provide argumentative sentences, although structured debates are needed. This led us to constitute an adequate discussion thread corpus. In order to get representative data we planed to ask participants to evaluate a dozen debates, hence we needed self-contained discussion threads about varied topics with short arguments (at most two sentences) in order to limit the efforts required for understanding them. At first, we considered public debates hosted at `slashdot.com` or `agoravox.com`, and discussion pages associated with any Wikipedia article. Then, we noticed that argument maps used in philosophy (Twardy, 2004) better fit the aforementioned requirements, being self-contained debates with short arguments in essence. Therefore, we used three sources (Tab. 4) comprising realistic and diverse discussions published by scholars (groups A and B), as well as the argumentative dialogues in French (group C) from (Cayrol and Lagasquie-Schiex, 2004).

Table 4

Origin of the 13 argumentative debates constituting the corpus for the experiment.

Group	Debate #	Description	Authors
A	1, 3–5, 7, 10	Argument maps [4]	Melbourne University (AU)
B	2, 9, 11, 12	Argument maps [5]	Ohio University (US)
C	6, 8, 13	Argumentative dialogues	Toulouse University (FR)

After translating the French resources into English, we obtained 13 debates in English (i.e. 222 arguments) available online. [6] Each argument is a short phrase that expresses a single opinion. As shown in Tab. 5, their characteristics are varied. The same is true of their topics, as they contain disputes about tobacco consumption, global warming, alternatives for a surgical operation, curriculum counseling, interpretations of the Bible . . .

Table 5

Characteristics of the constituted corpus comprising 13 debates.

Characteristic	Minimum	Maximum	Mean	Standard Deviation
Number of arguments	5	34	17.1	8.1
Debate depth	3	7	4.2	1.3
Debate length	3	15	7.9	3.2

Once the evaluation corpus constituted ①, we had to provide ② the "standard" social validity that an individual would mentally synthesize. As this value is rather subjective, we felt that assigning a value by ourselves would bias the experiment. Therefore, as explained in the next section, we let each participant evaluate the social validity values, which were compared to the results of the algorithms afterwards.

3.2. *Tasks of a participant: arguments' opinion identification and synthesis*

Participants involved in the experiment had to achieve two consecutive tasks (denoted ❶ and ❷) for each of the 13 debates which comprise 222 arguments overall. In order to limit a potential dropout from the beginning of the experiment, we ordered the debates according to an increasing difficulty, which we subjectively estimated from their characteristics and topics. Table 4 shows the order of the debates regarding their source. The two tasks assigned to the participants are defined below:

[4] http://austhink.com/reason/tutorials

[5] http://jostwald.com/ArgumentMapping

[6] http://www.irit.fr/~Guillaume.Cabanac/expe/corpus

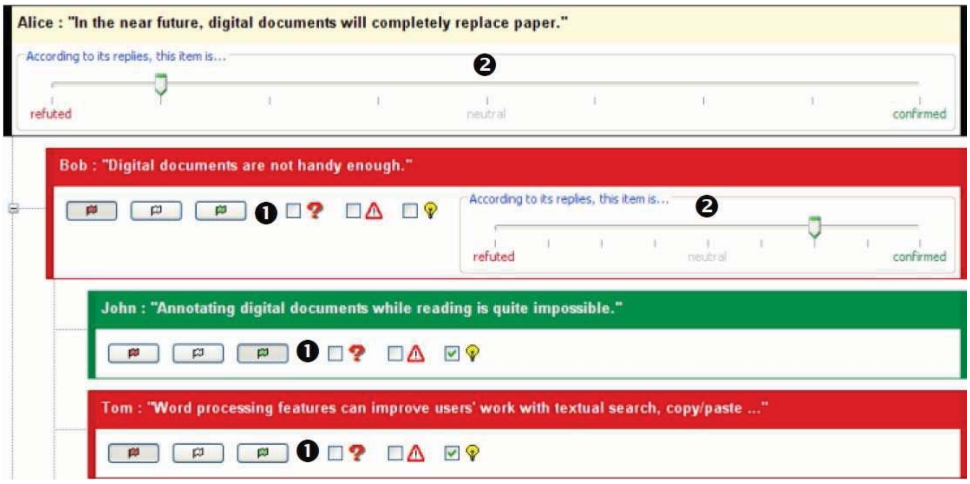
- For task ❶, participants evaluated the semantics of the links that tie arguments in each debate, represented as $a_i \leftarrow a_j$ in Fig. 1. This evaluation consisted of identifying one opinion type and the comment type(s) existing between an argument a_j and its direct parent a_i in the debate. Referring to Tab. 1, three opinion types (refutation, neutral, and confirmation) and three comment types (modification, question, and example) were available. Combining for instance opinion and comment types enabled participants to adjust the strength of an argument: a confirmation along with an example being stronger than a confirmation alone. Opinion identification must be objective: participants were asked to take into account argument contents only, as their personal judgment is unwanted. In Fig. 1, one may identify a confirmation for $a_2 \leftarrow a_5$, as well as a refutation with example for $a_2 \leftarrow a_4$.
- Participants proceeded with task ❷, once each link tying arguments of the current debate was valued. This task consisted of synthesizing the opinions of the debate mentally. For any argument a_i having a set of children $\{a_j\}$, participants had to associate with a_i a "synthesis" value on a 10-point scale. It ranged from "refuted" to "confirmed" through "neutral" and enabled participants to specify the global opinion of a_i's children arguments. Achieving this task in a recursive way led participants to associate the synthesized group's global opinion with the root argument. For instance, in Fig. 1, one may estimate that a_2 is 66% confirmed by synthesizing the opinions of a_3, a_4, and a_5. Such a valuation would be rational as it corresponds to the $\frac{\#\text{ confirmations}}{\#\text{ refutations}}$ proportion. Note that participants were not taught about this specific way of synthesizing, in order not to influence them.

Participants took part in the experiment through a specific software which allows the displaying of debates, and the storage of people's contributions, i.e. identified opinions ❶ and estimated syntheses ❷. The next section details the software platform we designed for that purpose.

3.3. *Designing the software platform supporting the experiment*

The methodology described in the previous section was implemented as an online platform for experimenting, which conforms to the standards of Internet experimenting (Reips, 2002, 2007). This platform federates three components: a Web page[7] for explaining the purpose of the experiment and how to take part, a program for displaying the debates to the participants, and a database for storing their contribution. The Web page enables a participant to launch the Java-Swing application on his/her computer thanks to the Java Web Start deployment technology. Following the registration of the participant, a tutorial describes the work to achieve for each of the 13 debates, i.e. tasks ❶ and ❷. Lastly, the application displays the Graphical User Interface (GUI) of the first debate. Figure 4 illus-

[7] http://www.irit.fr/~Guillaume.Cabanac/expe



Alice : "In the near future, digital documents will completely replace paper."

Bob : "Digital documents are not handy enough."

John : "Annotating digital documents while reading is quite impossible."

Tom : "Word processing features can improve users' work with textual search, copy/paste ..."

Fig. 4. Screen shot of the application GUI displaying the first debate to evaluate.

trates how the debate shown in Fig. 1 is presented to the participant. Note that we only represented arguments a_1 to a_4 due to space restrictions. In order not to confuse participants, we opted for a common hierarchical visualization they would be used to, as it is widely used for displaying file systems and newsgroups. Each argument is displayed in an horizontal box, initially bordered with light yellow. Participants achieve task ❶ using the exclusive buttons representing colored flags, and through the check boxes. Indeed, each opinion type is associated with a metaphorical color: refutation/red, neutral/white, confirmation/green. For readability concerns, selecting a button turns the arguments' border into the associated color, as shown in Fig. 4 (John's argument is in green whereas the other ones are in red). The comment types are displayed as check boxes along with icons: question/question mark, modification/warning sign, example/bulb. Once the mandatory opinion types are selected, participants proceed with task ❷ by synthesizing children's opinions (e.g. John's, Tom's, and Peter's) on their father's (Bob's) slider widget. Indeed, this widget is associated with arguments that have replies (children arguments). It consists of a 10-point scale varying from refuted to neutral, and to confirmed (Fig. 8). The slider associated with the root argument intends to reflect the global opinion of the group, i.e. its mentally estimated social validity. As long as any of the two tasks are not completed for the current debate, the program restrains participants from evaluating the next debate.

A dedicated database stores the contents of the 13 debates, as well as data concerning participants and their participations. Figure 5 shows the conceptual model of this database, which is implemented with the Oracle 10g2 RDBMS. The corpus consists of the aforementioned Resources that served as the basis of the 13 Debates we extracted and ordered. Each debate is related to a root Argument, which is disputed by reply arguments. In order to store information about the participants, we modeled the Participant class that keeps track of their identity, email

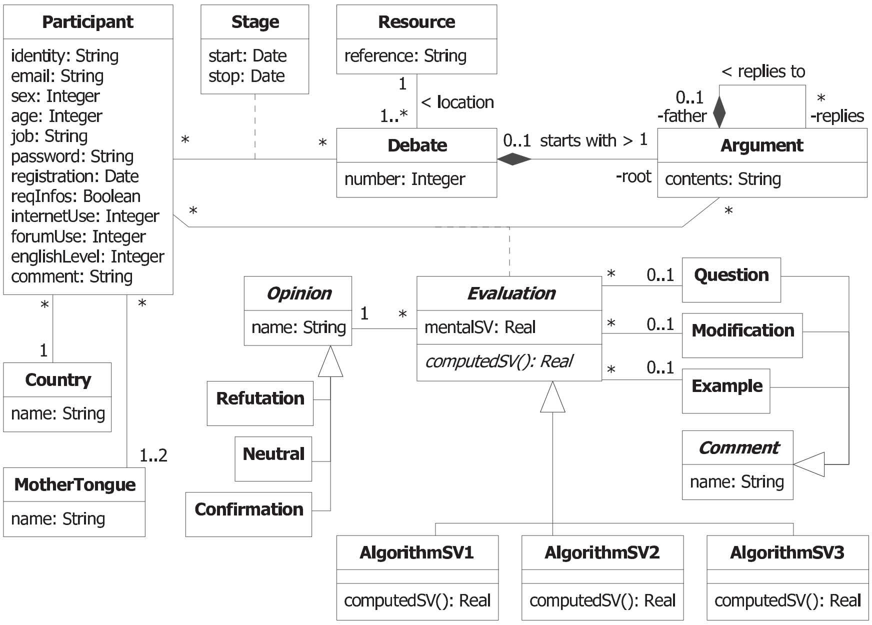
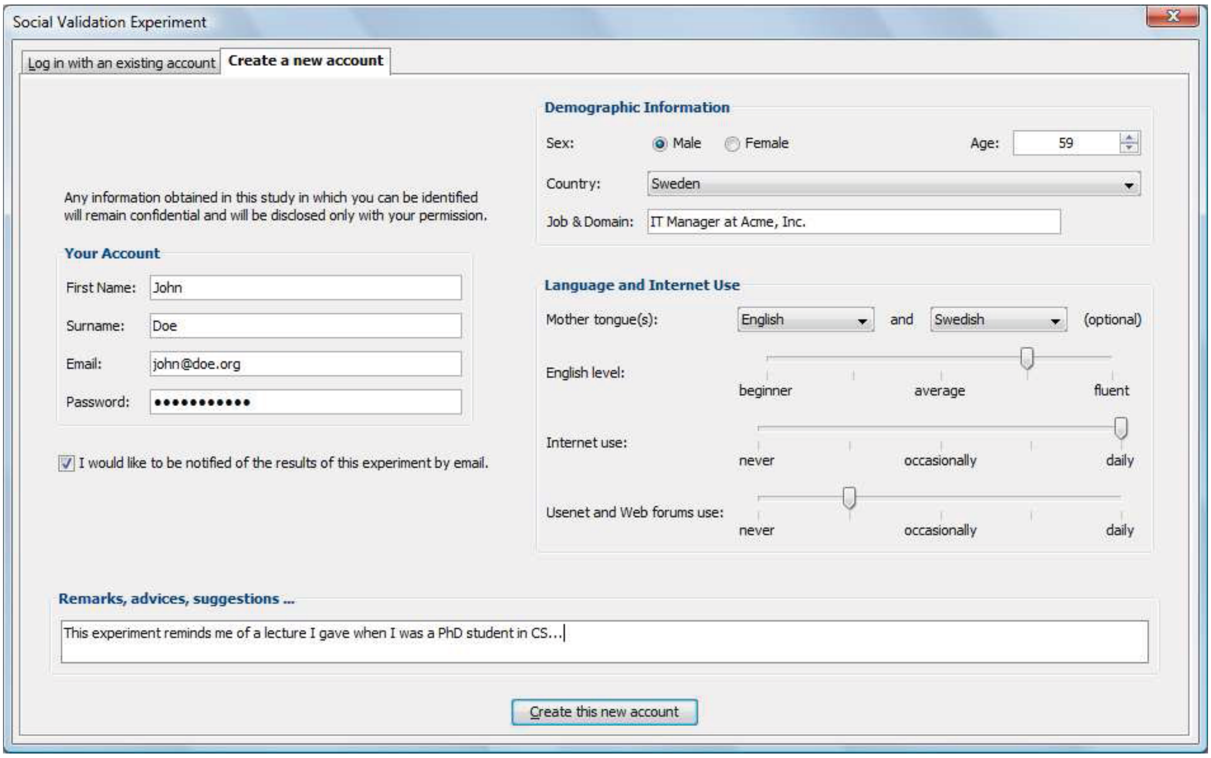


Fig. 5. UML class diagram for the experimentation platform.

contact, sex, age, job, registration date, country, and mother tongues—only one being mandatory. In addition to this data, participants were asked for their English level, and for their familiarity with the Internet and forums. Finally, they could ask to receive information related to the experiment (reqInfos) and leave a comment. All these information was obtained through the registration form (Fig. 6), only identity and email were mandatory.

The experiment consists of evaluating the 13 debates independently. Indeed, for practical reasons, participants were able to stop the experiment whenever they wanted, and could resume it later. We call a "Stage" the evaluation of a debate; it consists of providing an *Evaluation* for any argument of the current debate, i.e. achieving task ❶ (*Opinion* and *Comment* classes) and task ❷ (mentalSV attribute, for parent arguments only). As opposed to mentalSV which represents human perception, the *computedSV* function computes the social validity according to an algorithm under study (AlgorithmSVi concrete classes, detailed in Sect. 4.3). It is worth noting that these algorithms take as input the opinion and comment types identified by the participant. Therefore, human perception is compared with results from the algorithms, on the basis of the same p participant's types. For example, we may compare the value of the slider that the p participant set for Bob's argument, in Fig. 4, with the result of a social validation algorithm run with comment and opinion types identified by the same p participant.



Any information obtained in this study in which you can be identified will remain confidential and will be disclosed only with your permission.

Your Account

First Name: John

Surname: Doe

Email: john@doe.org

Password: ●●●●●●●●●●●

☑ I would like to be notified of the results of this experiment by email.

Demographic Information

Sex: ● Male ○ Female Age: 59

Country: Sweden

Job & Domain: IT Manager at Acme, Inc.

Language and Internet Use

Mother tongue(s): English and Swedish (optional)

English level: beginner average fluent

Internet use: never occasionally daily

Usenet and Web forums use: never occasionally daily

Remarks, advices, suggestions ...

This experiment reminds me of a lecture I gave when I was a PhD student in CS...

Fig. 6. Screen shot of the GUI displaying the registration form.

3.4. *Recruiting worldwide, independent, and diverse participants*

The recruitment of participants is a key issue for any experiment. Among many others, Wolfe and Neuwirth (2001) denounced the recruitment of dependent and like-minded, e.g. the experimenters' colleagues or students, as behavior biases due to such relationships were reported. As we intended to foster participants' independence and diversity, we sent out a call for participation on the Web. Since April 2007, we progressively sent it over French and international mailing lists related to computer science: the ACM's chi-student and chi-web, the W3C's www-annotation, semanticweb, webir ... In September 2007, we registered our experiment with dedicated psychological research sites to widen the potential participants range: the *Web Experiment List*[8] (Reips and Lengler, 2005) and *Psychological Research on the Net.*[9] Note that people took part on the basis of volunteering, as no reward was offered.

[8] http://genpsylab-wexlist.unizh.ch

[9] http://psych.hanover.edu/research/exponnet.html

4. Results: analyzing the evaluations from the 121 participants

This section reports on the data acquired during the 15-months online experiment we conducted. [10] We performed statistical analyses to measure how close social validation algorithms are from human perception of consensus.

4.1. *Quantitative analysis of the 121 participations*

As of July 2008, 181 people launched the experiment and registered with it. In the remainder of this paper, we call "participants" the 121 who actually started it by evaluating one debate at least. For privacy concerns, completing the form shown in Fig. 6 was not mandatory. As a result, only 56 participants among 121 provided personal data. On the basis of those 56 completed forms, participants come from 13 countries. Table 6 shows the proportion of their geographic origins. French people represent the majority of the participants, partly because the call for participation was first sent on French mailing lists.

Table 6

Origins of the 56 participants who completed the registration form.

Origin	France	Americas	Europe	Other
Proportion	67%	13%	10%	10%

Data about participants' origins are consistent with Tab. 7 which shows that French is the most common mother tongue among the participants.

Table 7

Mother tongues of the 56 participants who completed the registration form.

Mother tongue (ISO code)	fr	en	ar	de	pt	el	oc	ro	tr
Proportion (%)	61.8	19.1	7.4	2.9	2.9	1.5	1.5	1.5	1.5

Table 8 illustrates the participants' characteristics obtained through the registration form (rows 1–4) or computed from the acquired data (last row). The ordinal 5-point scales were coded as an integer in the $[\![1,5]\!]$ range. The typical participant is a male (60%) in his thirties who declared a good English level. Participants are very familiar with the Internet as they use it daily, and less knowledgeable about Usenet or Web forums, as they declared to use them occasionally.

[10] The data acquired is available on `http://www.irit.fr/~Guillaume.Cabanac/expe/data.xml` for reproducibility concerns.

[11] Statistics concerning the 53 participants who completed the experiment.

Table 8

Quantitative data about the 56 participants who took part in the experiment.

Variable	Minimum	Maximum	Average	Std Deviation
Age	20	61	32.1	9.4
English level	1	5	3.8	1.1
Internet use	3	5	4.9	0.4
Usenet and Web forums use	1	5	3.2	1.1
Number of interruptions [11]	0	12	4.5	3.9

We designed the experiment so that it may be stopped and resumed at any time; on average a participant completes the experiment with 4.5 interruptions (Tab. 8). As only 53 participants over 121 completed the experiment, the number of participants by stage decreases according to the stage number (Fig. 7). This corresponds to a 56% dropout, which is slightly over the reported 45% dropout for unrewarded experiments (Reips, 2007, p. 387). It is worth noting that, even when not completed, people's participations are analyzable, as they actually completed from one to twelve stages.

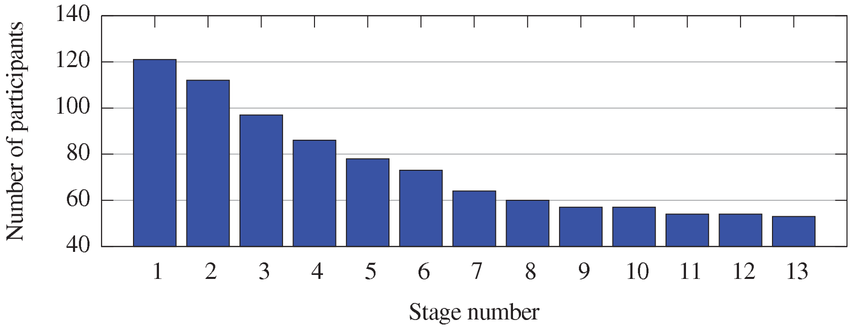


Fig. 7. Dropout curve showing the number of participants who completed each stage.

On the basis of the 966 completed stages, evaluating a debate is 6.5 minutes long on average, corresponding to 84.5 minutes for the whole experiment completion. Note that the average debate evaluation length is measured from the moment it is displayed until the moment the participant sees the following one. As a natural experiment, i.e. in the field, participants were not under the experimenters' observation, so we were unaware of participants' breaks and other activities, e.g. answering the phone. This may contribute to explain the high standard deviation which indicates an important variability between participants ($\sigma = 10.6$ minutes). Length-related estimates are pessimistic as they overestimate the time that the participant spent evaluating a debate. When ignoring the longest evaluation of each participant, the average debate evaluation is 5.2 minutes long, with a much lower variability

(σ = 4.6 minutes). This situation is closer to the durations reported by the participants that we could ask for *a posteriori*.

4.2. *Qualitative analysis of the 121 participations*

Conducting a natural experiment online brings several key advantages: participants can be worldwide, diverse, and independent. On the other hand, this way of experimenting presents some drawbacks, compared to lab experiments under the experimenters' supervision: participants may not read the instructions, misunderstand them, get bored and not reply conscientiously or even make deliberate mistakes. In order to deal with this issue, we defined indicators for assessing the quality of the acquired data. During task ❶, participants identified argument opinions. This is a rather subjective task, especially when it comes to choosing between $\mathscr{R}$ and $\mathscr{N}$, or between $\mathscr{C}$ and $\mathscr{N}$. Indeed, two people may have identified completely opposed opinion types. However, this is not an issue in the context of this experiment, since we compare each participant's mental synthesis (based on his/her identified types) with the results of the algorithms, which took as input the same identified types. To sum up: interpersonal agreement for type identification is not required for comparing human perception with social validation algorithms. As a result, no quality indicator is required regarding task ❶.

On the other hand, opinion synthesis values acquired during task ❷ must be checked for errors. Indeed, taking into account participants who did not understand the instructions or the interface, or set the sliders almost randomly would bias the results. That is why we defined the four error indicators shown in Tab. 9. We observed such evidence of irrational opinion syntheses for 21% of the 5,647 evaluations (Evaluation class, Fig. 5). Table 9 details the proportion of each indicator, e.g. indicator ① is an example of the case of an argument marked as refuted although none of its children is refuting it ($\neg\mathscr{R}$). Indicator ② is the counterpart of indicator ①. The most frequent irrational situation (35%) is reflected by indicator ③ when confirmations only ($\neg\mathscr{N} \wedge \neg\mathscr{R}$) or refutations only ($\neg\mathscr{N} \wedge \neg\mathscr{C}$) are synthesized as neutral. Finally, indicator ④ points neutral arguments ($\mathscr{N} \wedge \neg(\mathscr{R} \vee \mathscr{C})$) synthesized as a confirmation or a refutation.

Table 9

Irrational opinion syntheses observed for 21% of the 5,647 evaluations.

Error indicator	①	②	③	④
Father's synthesis	refuted	confirmed	neutral	confirmed or refuted
Children's opinions	$\neg\mathscr{R}$	$\neg\mathscr{C}$	$\neg(\mathscr{R} \wedge \mathscr{C} \vee \mathscr{N})$	$\mathscr{N} \wedge \neg(\mathscr{R} \vee \mathscr{C})$
Observed rate	28%	24%	35%	12%

The average error rate for a participant is 7% ($\sigma = 7\%$), the most mistaken participant having a 26% error rate. Hypothesizing that an individual error rate over 20% is a sign of incomprehension, we discarded the contributions of the corresponding 7 participants—some of whom had completed the experiment though—from further analyses. On the basis of the remaining participations, the next section draws a comparison between the human perception of consensus and the three social validation algorithms.

4.3. *Do social validation algorithms approximate human perception?*

The analysis of the 121 participations, excluding the 7 discarded ones, enabled us to evaluate to what extent social validation algorithms approximate the human perception of consensus. This section compares the three aforementioned algorithms with opinion synthesis values provided by the participants. In the remaining of this paper, hp denotes human perception, i.e. opinion synthesis values provided during task ❷. In addition, we refer to the first algorithm (Cabanac et al., 2005) as sv_1 (Sect. 2.3.2). Cayrol and Lagasquie-Schiex's algorithm (Cayrol and Lagasquie-Schiex, 2005a) is referred to as sv_2, and our extension to this algorithm (Cabanac et al., 2007a) is referred to as sv_3 (Sect. 2.3.3). Note that sv_1 and sv_3 consider information that does not exist in the corpus, e.g. annotator's expertise, number of references. As a result, these aspects were not evaluated in this experiment.

In order to compare hp with sv_i, we had to choose between taking into account every argument having children (e.g. a_1 and a_2 in Fig. 1) or only root arguments (e.g. a_1). With the first rationale, non-root arguments clearly outnumber root arguments. However, the mental evaluation of the opinion synthesis for a root is harder than for a non-root, as it requires to synthesize up to 34 arguments (Table 5). Thus, the resulting comparison would be biased because of not strictly reflecting the ability of algorithms to synthesize a whole debate. As a result, we opted for the second—stricter—rationale. Excluding the previously discarded participations, we extracted from the database 887 tuples such as $\langle p, a, hp(p,a), sv_1(p,a), sv_2(p,a), sv_3(p,a)\rangle$ where p is a participant, a is an argument, $hp(p,a)$ is the opinion synthesis value assigned by p to a, and $sv_i(p,a)$ is the result of the algorithm sv_i run with the opinions identified by p for the argument a. We had recourse to model-fitting for estimating the parameters related to each sv_i algorithm. The hp and sv_i values were defined in the range $[-1,1]$ as represented by the 10-point scale in Fig 8. Note that, for practical reasons, the slider widget restrained participants' selection to one of the 10 points, hence the discontinuous values for hp. On the contrary, the sv_i algorithms compute continuous values. This difference between the hp and sv_i values is considered in the remainder of this section.

In order to compare algorithms with human perception, we first computed their difference $sv_i - hp$ and plotted the resulting distribution of differences (Fig. 9). We hypothesized that sv_i exactly matches (error $x = 0$) a given

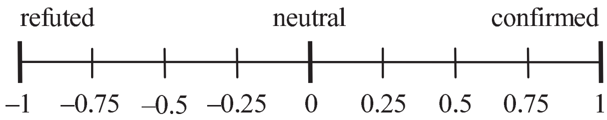


Fig. 8. Encoding of opinion synthesis values on a 10-point scale.

hp position if sv_i is closer to the hp position than to any other one. As there is 0.25 between two neighboring positions, we report a $x = 0$ error when $|sv_i - hp| \leqslant \frac{0.25}{2} = 0.125$. Indeed, Fig. 9 represents the y proportion of a given $x \pm 0.125$ difference for the three algorithms: $y = p(x - 0.125 \leqslant sv_i - hp \leqslant x + 0.125)$. Ideally, the algorithms would equal human perception, i.e. $y(0) = 100\%$. In reality, they exactly match human perception in around $y(0) = 25\%$ of the cases. Moreover, Fig. 9 shows that the three algorithms perform similarly.

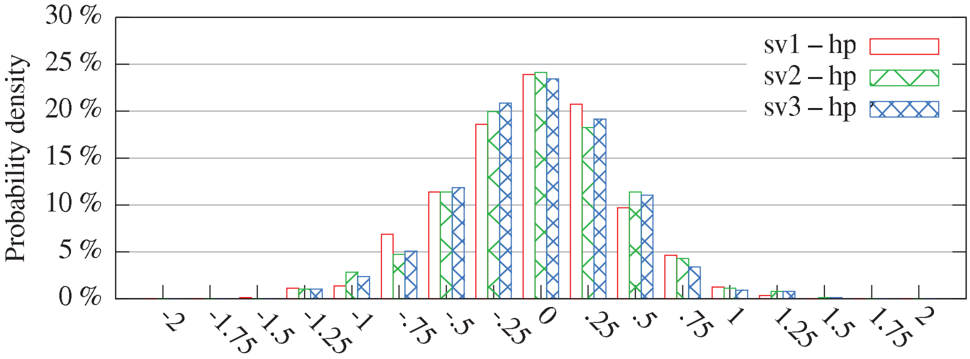


Fig. 9. Difference between the three algorithms sv_i and human perception hp.

4.3.1. *Algorithms* versus *human perception: statistical hypothesis testing*

In order to compare the (hp, sv_i) pairs by means of statistical hypothesis testing, we first checked the normality of the errors distribution $(sv_i - hp)$. Indeed, selection of the appropriate statistical (parametric or non-parametric) test depends on the error distribution. We used the Shapiro-Wilk's test (Shapiro and Wilk, 1965) to check whether the three series (hp, sv_i) conformed to a Normal distribution. Table 10 shows the significance p-values corresponding to the hypothesis that (hp, sv_i) follow a Normal distribution. Values $p(hp, sv_{\{2,3\}}) < \alpha$ reject this hypothesis of normality, where $\alpha = 0.05$ is the common agreed threshold (Hull, 1993). As a result, the following analyses must use non-parametric tests.

Table 10

Significance of the Shapiro-Wilk's normality test for the paired (hp, sv_i).

	(hp, sv_1)	(hp, sv_2)	(hp, sv_3)
Significance p-value for the normality test	0.0605	0.0007	0.0150

To go into detail, we used the Wilcoxon signed-rank (WSR) test of paired samples (Wilcoxon, 1945). This test is the non-parametric counterpart of the well-known parametric Student's t-test. It enabled us to compute the significance p-value in $[0,1]$ which "can merely be viewed as an estimate of the likelihood that two methods are different" (Hull, 1993). Indeed, the two methods, compared on the basis of their resulting p-values, are said statistically different when $p < \alpha$, where $\alpha = 0.05$ commonly used (Hull, 1993). In a word: the more $p \to 0$, the more the two methods are different. In addition, we used a second statistical coefficient, Pearson's product-moment correlation coefficient $r \in [-1,1]$, which evaluates how close two methods are: a $r \to 1$ value evidences a positive linear relationship, i.e. $y = x$. On the other hand, $r \to -1$ supports a negative linear relationship, i.e. $y = x^{-1}$. As a result, the more $r \to 1$, the more the two methods are similar.

Table 11 shows the resulting significance p-values and correlation r-values. The first row concerns the whole debates (1–13). The two values $p\left(hp, sv_{\{2,3\}}\right) < \alpha$ show that the differences between the sv_i algorithms and hp are statistically significant. Moreover, because $p(sv_1) > p(sv_{\{2,3\}})$, the algorithm sv_1 is less different from hp than sv_2 and sv_3 are. Then, we measured their correlation $r \approx 0.5$ which shows a medium correlation between sv_i and hp. Note that differences between the three algorithms do not seem significant, as they are very close to each other.

Table 11

Significance p of the Wilcoxon test and Pearson's r-values for the paired (hp, sv_i).

Debates	WSR test significance p			Pearson's r-values		
	(hp, sv_1)	(hp, sv_2)	(hp, sv_3)	(hp, sv_1)	(hp, sv_2)	(hp, sv_3)
1–13	0.0667	0.0475	0.0011	0.4859	0.4714	0.4905
2–13	0.0288	0.0626	0.0021	0.4954	0.4766	0.4971
3–13	0.0381	0.1970	0.0083	0.5161	0.4912	0.5146
4–13	0.1610	0.7319	0.0732	0.5316	0.5032	0.5281

Some participants warned us that their evaluations at the beginning of the experiment should be discarded, as they afterwards realized they were mistaken. In fact, the program does not allow any modification on completed past stages, so they were not able to correct the associated evaluations. As a result, we wanted to check the impact of the "warming-up" first stages. That is why rows 2–4 of Tab. 11 progressively get rid of stages 1–3. It appears that p-values for debates 4–13 are greater than the other p-values (for debates 1–13, 2–13, and 3–13). This seems to confirm the fact that the first evaluations were erroneous.

Regarding debates 1–13, Fig. 10 shows the plotted couples (sv_i, hp). Note that the data points correspond to the horizontal positions on the hp scale, as we obtained them from the scale in Fig. 8. For each algorithm, we performed a linear regression and obtained the associated estimated lines $\hat{y} = ax + b$. With an algorithm performing exactly the same way as human perception, we would obtain $\hat{y} = x$. In Fig. 10, one may notice that algorithms sv_2 and sv_3 have similar lines, moreover they are closer to the ideal $\hat{y} = x$ line than sv_1 is.

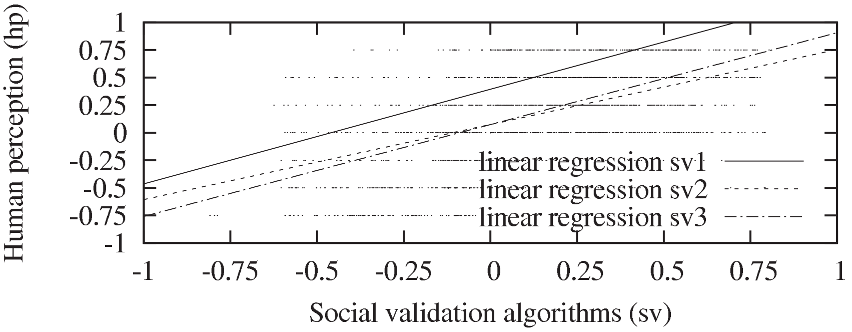


Fig. 10. Values (sv_i, hp) along with the corresponding linear regression lines.

As sv_2 and sv_3 seem very close to each other in Fig. 10, we tested whether each algorithm is statistically different from the two others (Tab. 12). For that purpose, we used the WSR test because the distribution of errors $(sv_i - sv_{j\neq i})$ did not fit the Normal distribution, as reported in the first row of Tab. 12. The significance values $p \ll \alpha$ for the WSR test point a significant difference between the pairwise algorithms.

Table 12

Pairwise comparison of the three sv_i algorithms.

Test	(sv_1, sv_2)	(sv_1, sv_3)	(sv_2, sv_3)
Shapiro-Wilk's normality test significance	$2.7 \cdot 10^{-13}$	$8.6 \cdot 10^{-5}$	$1.1 \cdot 10^{-23}$
Wilcoxon's signed-rank test significance	$1.5 \cdot 10^{-1}$	$2.3 \cdot 10^{-8}$	$1.2 \cdot 10^{-7}$

In this section, the different results obtained with statistical tests indicate that the three sv_i algorithms perform in a similar way. In statistical terms, they partly approximate human perception ($r \approx 0.5$). The next section investigates more user-centered indicators to evaluate the degree of approximation of the human perception hp by the social validation algorithms sv_i.

4.3.2. *Algorithms* versus *human perception: user-centered indicators*

Besides the common statistical tests reported in the previous section, we defined user-centered indicators for comparing sv_i with hp. These indicators allow the evaluation of the sv_i algorithms' ability to match hp in terms of distance, polarity, and strength. Table 13 reports on the results regarding these three indicators; it shows the corresponding predicates [12] along with the associated proportions, for each sv_i. Each indicator is available in two

[12] The sgn : $x \mapsto \{-1, 0, 1\}$ function returns a value corresponding to the sign of x.

versions (conservative and relaxed). They are explained as follows:

- Distance (D_i) measures the difference between sv_i and hp according to the number of positions separating them on the scale shown in Fig. 8. Indicator D_1 (resp. D_2) means a gap of one (resp. two) position(s) between the result of the algorithm and the human perception.
- Polarity (P_i) measures the difference between sv_i and hp according to their sign, i.e. opinion type ($\mathscr{R}$, $\mathscr{N}$, and $\mathscr{C}$ in Tab. 1). Indeed, P_1 is true when sv_i and hp share exactly the same type. Indicator P_2 is less conservative than P_1 as sv_i, hp, or both may be neutral.
- Strength (S_i) measures the ability of the algorithms sv_i to match the same “area” as hp on the scale represented in Fig. 8. We divided this scale into three areas whose boundaries are expressed in the S_i predicates. These areas correspond to the negative consensus, controversy, and positive consensus.

Table 13

Comparison between human perception (hp) and social validation algorithms (sv_i).

Indicator	Predicate	sv_1 (%)	sv_2 (%)	sv_3 (%)
D_1	$\lvert hp - sv_i \rvert \leqslant 0.25$	48.0	46.4	47.0
D_2	$\lvert hp - sv_i \rvert \leqslant 0.50$	77.1	76.3	77.3
P_1	$\operatorname{sgn}(hp) = \operatorname{sgn}(sv_i)$	66.1	65.4	65.4
P_2	$\operatorname{sgn}(hp) = \operatorname{sgn}(sv_i) \vee hp \cdot sv_i = 0$	84.8	84.7	84.6
S_1	$(hp, sv_i) \in \left\{\left[-1, -\frac{1}{3}\right[, \left[-\frac{1}{3}, \frac{1}{3}\right], \left]\frac{1}{3}, 1\right[\right\}$	54.7	55.7	57.5
S_2	$(hp, sv_i) \in \left\{\left[-1, -\frac{2}{3}\right[, \left[-\frac{2}{3}, \frac{2}{3}\right], \left]\frac{2}{3}, 1\right[\right\}$	73.5	73.7	74.2

Summing up the analyses reported in this fourth section, the comparison of the algorithms sv_i with the human perception hp of consensus in argumentative debates showed that the three algorithms perform very similarly regarding statistical testing (Tab. 11), as well as user-centered indicators (Tab. 13). Overall, there is a medium correlation between sv_i and hp. The relaxed versions of user-centered indicators indicate that sv_i matches hp in terms of distance, polarity, and strength in about 80% of the cases. Although the algorithms perform similarly, they may be distinguished according to their performance. For that purpose, Tab. 14 reports on the relative processing time of each algorithm, by comparing execution times over a sample of 5,647 arguments.

The fastest algorithm is sv_2. Algorithm sv_3 performs quite similarly with an increase of 30% in time. Finally sv_1 is nearly three times longer than sv_1. As a result, sv_2 and sv_3 may be the best alternatives regarding both accuracy and time performance.

Table 14

Execution time of the sv_i algorithms, for a sample comprising 5,647 arguments.

	sv_1	sv_2	sv_3
Processing time (in seconds)	66	13	17
Multiplicative factor (baseline sv_2)	3.9	1.0	1.3

5. Discussion

The proposed social validation process only aims to reflect the social group's global opinion. Definitively, it does not intended to evaluate how *true* an annotation is, since social validity is biased by the so-called tyranny of the majority. For instance, the evaluation of Galileo's fictitious annotation "Sun is the center of the universe, not the Earth" along with his detractors' arguments at that time would lead to a strong negative consensus, i.e. the group denied Galileo outright. This evaluation is faithful to the group's opinion, even if those people were wrong: one has to remember this case when interpreting the social validity of an annotation. Another limit related to our approach concerns "discussion drifts," e.g. a discussion starting about the "$E = mc^2$" equation that drifts into discussing Albert Einstein's personal life although this was not the original subject of the discussion. The social validity computed for the root statement would be biased by the unrelated arguments. As a first solution to this issue, topic shift detection techniques (Radev, 1999) may be exploited to cut the discussion thread when detecting the discussion drift.

Regarding the methodology of this experiment, one may wonder if participants with limited skills in English were confused by the debates in English. Indeed, most of the participants are non native speakers (Tab. 7), while they declared a good command of the language (Tab. 8). As a result, some participants were certainly puzzled during task ➊, which consisted of identifying the opinions expressed in arguments. However, this is not a critical issue for two reasons. The first reason is stated in Sect. 4.2: even if people identify completely different opinions, it is not a problem since we compare their human perception with the results of the algorithms, run on the basis of the *same* opinions. Secondly, this language difficulty reflects the current situation where non English people with a poor level in English are confronted with a predominantly English-speaking Web. Still concerning the methodology, we ordered the debates such that participants follow a progressive course with an increasing difficulty. Another alternative would have consisted in randomizing the debates for each participant. This rationale would have led to a constant number of evaluations by debate, instead of 121 evaluations for the first debate and 53 evaluations for the thirteenth. However, the randomization alternative would have ruled out the progressive course, which we thought to be a strong requirement for people not to dropout at the early stages of the experiment. As such, we went against Reips' "high entrance barrier" recommendation (Reips, 2002, 2007), which aims to "provoke dropout to happen early and ensure continued participation after someone makes the decision to stay." Nevertheless, a dropout in our case is not pure loss as the evaluated stages from partial participations were analyzed.

Concerning the comparison of the social validation algorithms results with the human perception, two points may be discussed. Firstly, we used a model-fitting technique to estimate the sv_i algorithm parameters leading to $sv_i \rightarrow hp$. Thus, we reported on a comparison between hp and sv_i in the better case. Secondly, the sliders in the GUI (Fig. 4) correspond to ordinal scales, whose labels are *refuted*, *neutral*, and *confirmed*. We had to encode their values in order to achieve the comparison reported in Sect. 4; the real values that we currently assigned appear in Fig. 8. We wonder whether these values correspond to the participants' perception: is the 25% gap between any two given positions coherent with human perception? Indeed, one may estimate that the position before *confirmed* represents a 90% agreement, whereas the encoded value is 75%. Thus, an alternative would be to encode the scale following a logarithmic distribution. Nevertheless, we found no study supporting this alternative to date.

6. Conclusion and perspectives of work

On paper documents, the long-lived annotation practice remains very common still nowadays. Besides satisfying the need to annotate digital documents the same way, the so-called annotation systems capitalize on the ever-astonishing capabilities of networked modern computers in order to improve users' experiences. One significant feature allowing the creation of "discussion thread" consists in enabling worldwide readers to annotate and discuss in context any part of digital documents. This paper introduced a model for "collective annotations," i.e. annotations along with discussion threads. On the basis of this model, we proposed three algorithms aiming to compute the "social validity" of a collective annotation. This value corresponds to the global opinion of the group that took part in the associated discussion. We underlined how effective the social validation process may be, regarding prominent IS such as annotation systems, decision support systems, information retrieval systems, and collaborative writing platforms. In order to evaluate the proposed social validation algorithms, we defined a natural experiment, i.e. on the field as opposed to a laboratory experiment under the supervision of experimenters. Its purpose is to compare the results of the algorithms with the human perception of consensus in argumentative debates, hence the recourse to participants. Due to the reported lack of suitable debate corpus, we had to constitute one[13] from various academic resources. Following a call for participation we sent on international mailing-lists, 181 worldwide people registered with the online experiment we set up, 121 of whom effectively contributed their

[13] http://www.irit.fr/~Guillaume.Cabanac/expe/corpus

evaluations. The analysis of these evaluations, thanks to both statistical and user-centered indicators, shows that the three algorithms perform similarly. They approximate human perception up to 84% of the cases. Nevertheless, it is worth noting that two of the algorithms are much better than the other one, regarding execution times.

Many perspectives may be explored in order to extend and improve the social validation process. Firstly, the proposed algorithms evaluate the debates in a democratic way. Indeed, each argument is equally considered, which makes our approach subject to the "tyranny of the majority" issue. Although being accurate for specific contexts such as participative democracy and e-government, this characteristic is certainly not suitable for all contexts. Regarding corporate decision support systems for instance, each analyst has specific abilities and roles which matter when it comes to making a decision, e.g. a supervisor's decision may overcome his subordinates' ones. This kind of reasoning may be taken into account when computing the social validity of an argument in corporate contexts. Another perspective refers to providing a *personalized* social validation process. Indeed, the readers' preferences are not considered to date. However, individuals prove to better trust some people than others—especially regarding their annotations (Marshall, 1998)— and prove not to question the ones they call experts, etc. Such requirements may imply modeling users' evolving preferences from the adequate social networks, in order to adapt the social validation process consequently.

Acknowledgments

The authors thank the 181 volunteer participants who registered with the evaluation for their time and effort. Special thanks go to Gaëlle Lortal, Aurélie Névéol, and Anne-Marie Vercoustre for their insightful feedback and remarks.

References

Adler, M. J. and van Doren, C. (1972). *How to read a book.* Simon & Shuster, NY.

Agosti, M. and Ferro, N. (2005). Annotations as Context for Searching Documents. In *CoLIS'05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences*, volume 3507 of *LNCS*, pages 155–170. Springer.

Agosti, M. and Ferro, N. (2006). Search Strategies for Finding Annotations and Annotated Documents: The FAST Service. In *FQAS'06: Proceedings of the 7th International Conference on Flexible Query Answering Systems*, volume 4027 of *LNCS*, pages 270–281. Springer.

Agosti, M. and Ferro, N. (2007). A Formal Model of Annotations of Digital Content. *ACM Trans. Inf. Syst.*, 26(1):3.

Brust, M. R. and Rothkugel, S. (2007). On Anomalies in Annotation Systems. In *AICT'07: Proceedings of the 3rd International Conference on Telecommunications*, page 3 (8 pp). IEEE.

Butler, B., Joyce, E., and Pike, J. (2008). Don't Look Now, But We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *CHI'08: Proceeding of the 26th annual SIGCHI conference on Human factors in computing systems*, pages 1101–1110, New York, NY, USA. ACM.

Cabanac, G., Chevalier, M., Chrisment, C., and Julien, C. (2005). A Social Validation of Collaborative Annotations on Digital Documents. In Boujut, J.-F., editor, *International Workshop on Annotation for Collaboration*, pages 31–40. CNRS.

Cabanac, G., Chevalier, M., Chrisment, C., and Julien, C. (2007a). Collective Annotation: Perspectives for Information Retrieval Improvement. In *RIAO'07: Proceedings of the 8th conference on Information Retrieval and its Applications*. CID.

Cabanac, G., Chevalier, M., Ravat, F., and Teste, O. (2007b). An Annotation Management System for Multidimensional Databases. In Song, I.-Y., Eder, J., and Nguyen, T. M., editors, *DaWaK'07: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery*, volume 4654 of *LNCS*, pages 89–98. Springer.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2004). Bipolarité en argumentation. Report 2004-07-R, IRIT, Toulouse, France.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005a). Gradual Valuation for Bipolar Argumentation Frameworks. In Godo, L., editor, *ECSQARU'05: Proceedings of the 8th European Conference on Symbolic and Qualitative Approaches to Reasoning and Uncertainty*, volume 3571 of *LNCS*, pages 366–377. Springer.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005b). Graduality in Argumentation. *J. Artif. Intell. Res.*, 23:245–297.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20:37–46.

Dave, K., Wattenberg, M., and Muller, M. (2004). Flash Forums and ForumReader: Navigating a New Kind of Large-scale Online Discussion. In *CSCW'04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 232–241, New York, NY, USA. ACM Press.

DeRose, S., Daniel, R., Grosso, P., Maler, E., Marsh, J., and Walsh, N. (2002). *XML Pointer Language (XPointer)*. W3C.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in a nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77:321–357.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull.*, 76(5):378–382.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). The Measurement of Interrater Agreement. In Fleiss, J. L., Levin, B., and Paik, M. C., editors, *Statistical Methods for Rates and Proportions*, chapter 18, pages 598–626. John Wiley & Sons, Inc., 3 edition.

Foshay, N., Mukherjee, A., and Taylor, A. (2007). Does data warehouse end-user metadata add value? *Commun. ACM*, 50(11):70–77.

Fraenkel, A. S. and Klein, S. T. (1999). Information Retrieval from Annotated Texts. *J. Am. Soc. Inf. Sci.*, 50(10):845–854.

Frommholz, I. and Fuhr, N. (2006). Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. In *JCDL'06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 55–64, New York, NY, USA. ACM Press.

Golovchinsky, G., Price, M. N., and Schilit, B. N. (1999). From Reading to Retrieval: Freeform Ink Annotations as Queries. In *SIGIR'99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA. ACM Press.

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, New York, NY, USA. ACM Press.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., and Swick, R. R. (2002). Annotea: an open RDF infrastructure for shared Web annotations. *Comp. Netw.*, 32(5):589–608.

Karacapilidis, N. and Papadias, D. (2001). Computer supported argumentation and collaborative decision making : the HERMES system. *Inf. Syst.*, 26(4):259–277.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He Says, She Says: Conflict and Coordination in Wikipedia. In *CHI'07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA. ACM.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Sage Publications, Thousand Oaks, CA.

Liu, B. (2007). Opinion Mining. In Liu, B., editor, *Web data mining: Exploring hyperlinks, contents, and usage data*, Data-Centric Systems and Applications, chapter 11, pages 411–447. Springer.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Hypertext'98: Proceedings of the 9th conference on Hypertext and hypermedia*, pages 40–49, New York, NY, USA. ACM Press.

Ovsiannikov, I. A., Arbib, M. A., and McNeill, T. H. (1999). Annotation technology. *Int. J. Hum.-Comput. Stud.*, 50(4):329–362.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment Classification using Machine Learning Techniques. In *EMNLP'02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. ACL.

Radev, D. R. (1999). Topic shift detection – finding new information in threaded news. Technical Report CUCS-026-99, Columbia University, Manhattan, NY, USA.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Exp. Psychol.*, 49(4):243–256.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In Joinson, A. N., McKenna, K. Y. A., Postmes, T., and Reips, U.-D., editors, *The Oxford Handbook of Internet Psychology*, chapter 24, pages 373–390. Oxford University Press, New York, NY, USA.

Reips, U.-D. and Lengler, R. (2005). The *Web Experiment List*: A Web service for the recruitment of participants and archiving of Internet-based experiments. *Behav. Res. Meth.*, 37(2):287–292.

Sellen, A. J. and Harper, R. H. (2003). *The myth of the paperless office*. MIT Press, Cambridge, MA, USA.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shirky, C. (2008). *Here comes everybody: The power of organizing without organization*. The Penguin Press, London, UK.

Surowiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Anchor Books, New York.

Twardy, C. (2004). Argument Maps Improve Critical Thinking. *Teaching Philosophy*, 27(2):95–116.

Vasudevan, V. and Palmer, M. (1999). On Web Annotations: Promises and Pitfalls of Current Web Infrastructure. In *HICSS'99: Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, volume 2, page 2012 (9 pages), Washington, DC, USA. IEEE Computer Society.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biom. Bull.*, 1(6):80–83.

Wolfe, J. (2002). Annotation technologies: A software and research review. *Computers and Composition*, 19(4):471–497.

Wolfe, J. L. and Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *J. Bus. Tech. Commun.*, 15(3):333–371.