A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement

Cristina Grisot\textsuperscript{a,b,1,2,*}

\textsuperscript{a} University of Geneva, Department of Linguistics, Switzerland
\textsuperscript{b} University of Neuchâtel, Cognitive Science Centre, Switzerland

Received 19 November 2016; received in revised form 28 June 2017; accepted 30 June 2017
Available online 28 July 2017

Abstract
This study deals with three key notions in the relevance-theoretic framework: conceptual and procedural linguistically encoded meaning on the one hand, and pragmatic meaning on the other hand. I argue that having objective and quantitative measures for distinguishing among these types of meaning is necessary. Concretely, a quantitative measure is proposed based on offline annotation experiments made by untrained native speakers. This is inter-annotator agreement measured with chance-corrected agreement coefficients, such as Cohen's kappa coefficient. In order to reliably use the three layered scale for interpreting the values of the kappa coefficient, a series of requirements regarding the building and the running of the experiment, as well as the analysis of results, must be adhered to. In this paper, the measure is applied to verbal tenses in order to identify and investigate their contextual usages. It is shown that when speakers are asked to consciously evaluate the contribution of verbal tenses to the interpretative process, three patterns emerge systematically. The first is the easiness of the task and the high rate of inter-annotator agreement when they deal with the distinction past/non-past. The second is a greater difficulty of the task and lower rates when they deal with temporal ordering eventualities. The third is the impossibility to have inter-annotator agreement beyond chance level when they have to consciously identify a subjective or non-subjective point of perspective. It is argued that this observed difference may be explained in terms of the different contents that the addressee deals with: conceptual, procedural, and respectively, purely pragmatic.

© 2017 Elsevier B.V. All rights reserved.

Keywords: Conceptual meaning; Procedural meaning; Pragmatic meaning; Empirical pragmatics; Inter-annotator agreement rates; Relevance theory

1. Introduction

For the last twenty years, we have experienced the rise of empirical and experimental linguistics and pragmatics, fields in which theoretical models are required to be tested and validated. Theory and testing are two essential components of reproducible science: while theoretical models increase in accuracy when they are tested, experimental testing of...
hypotheses and interpretation of results, including p-values and statistical coefficients, are meaningful only when they are theoretically informed. Experimental methods in linguistics were first used in the field of experimental syntax, which tests and develops syntactic models using psycholinguistic experimentation (Musolino, 2001; Musolino et al., 2010; Syrett et al., 2012). The field of experimental pragmatics (Noveck and Sperber, 2004) draws on pragmatics, psycholinguistics, psychology of reasoning, and neurolinguistics (cf. Bertuccelli Papi's 2010 discussion about how pragmatics fits with the brain). Other linguists turned towards empirical and probabilistic pragmatics, which draws on corpus linguistics, offline experiments with acceptability or linguistic judgement tasks, and web-based methods (Romero-Trillo, 2014 for the former; Degen, 2015 for the latter), and computational linguistics (Cartoni et al., 2013a,b; Zufferey and Cartoni, 2014; Zufferey and Popescu-Belis, 2017; Grisot et al., 2016).

One pragmatic topic that has resisted the experimental/empirical rise in pragmatics (with a few recent exceptions, such as Bolly and Degand, 2013; Zufferey, 2014; Grisot and Moeschler, 2014; Grisot, 2015; Bolly et al., 2015) is the relevance-theoretic conceptual/procedural distinction introduced by D. Blakemore (1987, 2002) to explain differences between words with a conceptual content, such as table, cat, think or walk on the one hand, and discourse connectives, such as but, so or also on the other hand. Blakemore's argument was that content words encode concepts that contribute to the proposition expressed by an utterance while the meaning of a discourse connective is better described in terms of constraints on the inferential phase of interpretation rather than in conceptual terms. Currently, our challenge is to identify the features of conceptual and procedural information and to experimentally and empirically find the traces which allow us to measure the unseen process they point to (i.e. building and manipulation of mental representations). The issue at stake in this article is twofold. Firstly, I argue that the existing work does not specify how we can know what kind of meaning (conceptually encoded, procedurally encoded or pragmatically inferred) is at play beyond researcher intuition. Secondly, I show how evidence from inter-annotator agreement is used to apprehend the underlying processes conceptual, procedural and purely pragmatic information point to and to draw a distinction among them.

In theoretical pragmatic studies, the application of the existing qualitative criteria depends on the linguists’ own judgements based on their intuition, with which other linguists might not agree. Research based on theoretical models that are not tested and validated empirically and experimentally is judged according to whether or not the reader found the explanation plausible. Consequently, linguistic expressions are understood and analysed differently from one study to another. One example is that of the connective but, which is analysed by Blakemore (1987, 2000, 2002) and Iten (2000) as encoding an instruction to process the clause that follows as contradicting and eliminating an assumption, by Bach (1999) as contributing to the truth-conditions of a proposition and therefore not encoding procedural information, and by Hall (2004) as encoding an instruction to suspend an inference that would result in a contradiction with what follows.

Additionally, the researchers’ subjective judgments are generally given for artificial examples built to serve their own purposes, rather than by non-trained (also called naïve) speakers on natural or on experimentally controlled data. In order to increase the reliability of studies targeting the conceptual/procedural distinction, we need to make use of real corpora and empirical methods, such as annotation experiments in which non-trained speakers annotate linguistic data according to a given set of annotation guidelines. This leaves us with the need for an objective and quantitative measure for evaluating the results of this type of experiment. I suggest to use inter-annotator agreement measured with chance-corrected agreement coefficients, such as Cohen's kappa coefficient (Carletta, 1996; Artstein and Poesio, 2008; Spooren and Degand, 2010) as an objective method for investigating the type of information judged: conceptual, procedural, or purely pragmatic recovered through non-demonstrative inferences (i.e. inferences which do not guarantee the truth of the conclusion from the truth of the two premises, as opposed to demonstrative inferences) (Sperber and Wilson, 1986, 1987). It is a tool permitting researchers to evaluate the results of experiments in which speakers are asked to judge linguistic expressions regarding their encoded or pragmatically determined meaning. The interpretation scale of this coefficient is directly dependent on the theoretical assumptions that underpin the experimental investigation, and it makes reference to the behaviour of conceptual and procedural types of information predicted by Wilson and Sperber's cognitive criteria (1993/2012).

This paper is organised as follows. In section 2, I review the qualitative criteria currently used by scholars for distinguishing between conceptual and procedural information, and I show that, while some of them are discriminatory criteria, others are too intuitive. Section 3 is dedicated to the proposal made in this paper, namely to use inter-annotator agreement for identifying the type of content dealt with in annotation experiments: conceptual, procedural, or pragmatic. The proposal is illustrated with a series of experiments targeting verbal tenses. Section 4 concludes this paper.

2. Current assessments of the conceptual/procedural distinction

In this paper, the conceptual/procedural distinction refers to types of encoded meaning, which make different contributions to the interpretative process. Conceptual meaning refers to concepts encoded by some linguistic expressions, that is, the concept for which that linguistic expression is the lexical entry. Procedural meaning points to encoded instructions about how to manipulate conceptual representations. Both the concept and the linguistic expression
are stored in the lexicon, where procedural information is embodied as rules written explicitly in the lexical entries of linguistic expressions (Curcó, 2011). Conceptual and procedural types of meaning point to two aspects of human cognition: representation and computation. A procedure acts upon a conceptual representation encoded either by the same linguistic expression (hence, a mixed expression as suggested by Moeschler, 2002, 2009, 2015, 2016; Fraser, 2006; Pons Borderia, 2008; Wilson, 2011) or by a different one. The prototypical expressions encoding a concept are lexical words, such as *door, bachelor, open*, etc. These lexical entries of the mental lexicon are used to refer to sets of entities (the sets of entities which are doors, bachelors, and actions of opening things, respectively). However, as Moeschler (2016, 126) points out, “a concept is not only a mental representation of different sets of individuals, it is the abstract or mental entity that allows the fixation of beliefs and knowledge: concepts are the locus of information construction, storage and retrieval.” According to this definition of a concept, both lexical and functional categories (such as verbal tenses and connectives) may encode conceptual meaning. This proposal defends a dual and mixed nature of linguistic expressions: conceptual and procedural encoding.

Most of the current research dedicated to this distinction lacks objective empirical corpus or experimental data, such as Blakemore (1987, 1988, 2000), Wilson and Sperber (1993/2012), Nicolle (1997), Moeschler et al. (1998), Moeschler (2000, 2003), Saussure (2003, 2011), Fretheim (2006) to name but a few. Some exceptions are Zufferey and Cartoni (2012) who use annotation experiments for identifying the meaning of discourse connectives in English and French, Zufferey (2014) who tests the procedural content of the French connective *puisque* with online and offline experimental methods, Bolly et al. (2015) who investigate discourse markers in French with respect to their procedural content, Bolly and Degand (2013) who consider the diachronic development of expressions encoding conceptual meaning into expressions encoding procedural meaning, Grisot and Moeschler (2014) and Grisot (2015) who use annotation experiments for investigating the categories of tense and aspect cross-linguistically. These studies have shown that the conceptual/procedural distinction exists and that it can be operationalised and traced by experiments.

Numerous other attempts made in the literature to define and characterise conceptual vs. procedural information include qualitative features. I propose to divide them into two types: those that are cognitive (appealing to cognitive processes taking place when the speaker processes expressions encoding conceptual or procedural types of information) and those that are linguistic (referring to the linguistic system itself). The two types of criteria are summarised in Table 2.1.

<table>
<thead>
<tr>
<th>Cognitive criteria</th>
<th>Linguistic criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Representational status</td>
<td>1. Truth-value</td>
</tr>
<tr>
<td>2. Accessibility to consciousness</td>
<td>2. Behaviour with negation</td>
</tr>
<tr>
<td>3. Degree of availability to conscious thought</td>
<td>3. Compositionality</td>
</tr>
<tr>
<td>4. Degree of conceptualization</td>
<td>4. Rigidity</td>
</tr>
<tr>
<td>5. Degree of paraphrasability</td>
<td>5. Degree of paraphrasability</td>
</tr>
<tr>
<td>7. Type of inference triggered</td>
<td>7. Type of inference triggered</td>
</tr>
</tbody>
</table>

In the following paragraphs, I will discuss these criteria as they are currently assessed in the literature (cf. Carston, 2016 for a discussion of some of these tests, focusing on the heterogeneity of procedural meaning). This study does not question the conceptual/procedural distinction itself, its cognitive plausibility (which has been validated through several studies, as mentioned above),3 nor its great explanatory power for phenomena such as discourse connectives, referential expressions, tense, mood, modality, pronouns, etc. (cf. the studies in the volume edited by Escandell-Vidal et al., 2011). It questions the application of the above-mentioned criteria for discriminating between conceptual and procedural information because, as they are used, they do not allow reliable studies. The two categories of criteria have different drawbacks. Cognitive criteria are not available through introspection and they cannot be directly measured and operationalised for experimental testing, where operationalization refers to strictly defining variables into measurable

---

3 For example, it was suggested that connectives encode procedures that guide the hearer in the comprehension process (Blakemore, 1987, 1988, 2000, among many others). This was linked to a facilitation effect for processing the segment following the connective when the meaning of the connective correlates to the meaning of the discourse segment processed (Traxler et al., 1997; Canestrelli et al., 2013; Zufferey, 2014). In contrast, when the meaning of the connective does not correspond to the implicit discourse relation, no effect of facilitation is found (Cain and Nash, 2011; Grisot and Blochowiak, 2015). Furthermore, the erroneous use of a connective leads to a regression effect measured with eye tracking techniques at the end of the sentence when there is a mismatch between the meaning of the connective and the meaning of the discourse segment (Canestrelli et al., 2013; Zufferey et al., 2015).
factors (empirical and quantifiable), which allow replication and statistical analyses. In online experiments, measurements such as reaction and reading times are used for quantifying difficulty or easiness in processing. In offline experiments, measures such as inter-annotator agreement rates (cf. this proposal in section 3) could be used for measuring the degree of accessibility to conscious thought. Linguistic criteria have two problems. The first problem is that some of them are not discriminatory, as has been observed by several scholars (cf. Saussure, 2011). The second problem is that they are currently used as intuitive judgments made by the researchers themselves. One solution for those which can be proved to be discriminatory among conceptual, procedural, and purely pragmatic information, is to use them as tasks in annotation experiments with native speakers. These drawbacks will now be discussed in more detail.

The representational status criterion in Table 2.1 refers to the function of the encoded information: building conceptual representation vs. instructions for possible computations to be applied to mental representations (Sperber and Wilson, 1986/1995; Blakemore, 1987, 1992; Wilson and Sperber, 1993/2012). This criterion is based on the most basic assumptions defended in Relevance Theory: linguistic input provides input to the inferential phase of communication when the hearer builds and manipulates conceptual representations.

Criteria 2 to 4 in Table 2.1 come from Wilson and Sperber (1993/2012, 162), who characterise conceptual vs. procedural information as follows. Conceptual information is accessible to consciousness, capable of being reflected on and represents easily graspable concepts whereas procedural information is inaccessible to consciousness, not available through conscious thought and is resistant to conceptualization (Fodor, 1983; Wilson and Sperber, 1993/2012; Wilson, 2011). These features are based on the observation that we can consciously evaluate the meaning of conceptual words like chair, bachelor, murder and provide at least rough paraphrases for each of them whereas it is much harder to do this for words encoding procedural information, such as however, anyway, well etc. The linguistic criterion degree of paraphrasability (Blakemore, 1987, 2002) is based also on the same observation as these cognitive criteria: if we grasp a meaning intuitively and if the words are easy to paraphrase and to translate, then we are dealing with a conceptual expression. Similarly, if we cannot consciously come up with a meaning and if we cannot paraphrase or translate it, then we are dealing with procedural information. Of course, there are numerous counter-examples to this observation, as observed by Saussure (2011, 2017) and Carston (2016), such as the fact that evaluating the pronouns I and she seems easier than evaluating the meaning of words encoding a concept such as standard or serendipity. Because procedural information is inaccessible to consciousness, substitution tests have been used with success to identify the procedural content of an expression (Zufferey and Cartoni, 2012; Grisot and Moeschler, 2014).

Initially, following Blakemore’s work, it was assumed that conceptual information is truth-conditional whereas procedural information is non-truth-conditional. Truth-conditional content has logical properties: it can be contradicted, cancelled, used in entailment, and negated. Hence, according to the initial one-to-one mapping, conceptual and procedural information were characterised as follows:

- Conceptual information is truth-conditional, cannot be contradicted, cancelled, and negated, and is used in entailment.
- Procedural information is not truth-conditional, can be contradicted, cancelled, and negated, and it does not enter in entailment.

Based on Wilson and Sperber’s (1993/2012) convincing arguments against this one-to-one mapping showing that all four combinations are possible, the truth-evaluability criterion, as well as the other subsequent criteria (cancelability, contradiction, negation, and entailment) are no longer appropriate as discriminatory criteria.

The compositionality criterion (Rouchota, 1998; Blakemore, 2002) states that concepts combine with other concepts to build well-formed meaning (i.e. compositional semantics). Consequently, if two linguistic expressions combine to build well-formed meaning, it signals that they encode conceptual meaning (as suggested by Fretheim, 2006 for then). However, Assimakopoulos (2015) points out that discourse connectives do combine compositionally (e.g. but still, but after all), even if it is a minimal and less productive process than in the case of conceptual expressions. The same applies to logical connectives, that is and, or, if which can combine with discourse connectives (e.g. and so, but if). Similarly, Gosselin (2007) notes the existence in French of complex constituents that indicate temporal sequencing such as et puis alors ‘and so then,’ et alors ensuite ‘and so then’ and puis aussitôt après ‘then immediately after’. Hence, procedural

---

4 Moeschler (2009) makes a different proposal using negation. He suggests to distinguish between descriptive negation and metalinguistic negation, where a conceptual expression can only be negated through metalinguistic negation.

5 A very interesting observation regarding the lack of appropriateness of the truth-evaluability criterion is made by Assimakopoulos (2015, 67). He points out that this criterion “seems to directly clash with the pronounced aim of RT in relation to the investigation of semantics in cognitive processing terms”. More specifically, as argued by Carston (1999) and Blakemore (2000, 464), “the question that matters [for RT semantics] is not whether a linguistic expression contributes to something with truth conditions, but rather what kind of cognitive information an expression encodes” (Assimakopoulos, 2015, 67).
expressions tend to be less compositionally transparent and are more idiomatic (i.e. they can’t get the whole from the parts such as all of a sudden or after all) than conceptual expressions (i.e. they get more from the parts such as orange juice or newspaper, even though non-compositional conceptual words exist as well such as the French petit pois ‘green pea’). The problem seems to be that this difference can’t be used for an empirical measure. Hence, the compositionality criterion is not discriminatory since both conceptual and procedural expressions seem to display this feature.

Procedural and conceptual information are currently described in the literature in terms of rigidity vs. flexibility. Escandell-Vidal and Leonetti (2011) propose rigidity as the major feature of procedural information, which is opposed to the flexibility feature of conceptual information. In their words:\(^6\):

\[\ldots\text{Conceptual representations are flexible and malleable, which means that they can be enriched, elaborated on and adjusted in different ways to meet the expectations of relevance. (\ldots) We claim that instructions, in contrast, are rigid: they cannot enter into the mutual adjustment process, nor can they be modulated to comply with the requirements of conceptual representations, either linguistically communicated or not. The instructions encoded by an item must be satisfied at any cost for the interpretation to succeed.} \text{(Escandell-Vidal and Leonetti, 2011, 86)}\]

The main consequence of the rigidity feature displayed by procedural information is that cases of mismatch between the information coming from conceptual and procedural content will be solved following the procedural constraints on the interpretative path. Cúrcó (2011), too, describes procedural information as context-insensitive, that is, procedures must be executed regardless of contextual assumptions. However, Escandell-Vidal and Leonetti (2011) point out that despite the rigid character, instructions can give rise to a series of different interpretative effects. This results from the different contextual assumptions and the data varying from one context to another. In Grisot (2015) (also presented in section 3.3 for illustrative purposes), I present empirical evidence that the procedure encoded by the category of tense, namely to determine a temporal relation between two eventualities, is itself rigid and context-insensitive and it produces context-dependent results. Additionally, for Cúrcó (2011) procedures are arbitrary, in the sense that the link between a procedural item and the information it carries, or the process it triggers, is arbitrary and conventional. For her, procedural meaning is arbitrary and therefore, language-specific. Based on the empirical evidence provided in section 3.3, where the same procedure encoded by the category of tense was tested and validated for four languages (English, French, Italian, and Romanian), I disagree that procedures are necessarily language-specific.

Saussure (2011) proposes two linked criteria: the behaviour with pragmatic enrichment processes (i.e. loosening and narrowing) and the type of inference triggered. According to Saussure, an expression is procedural if it triggers inferences that cannot be predicted on the basis of a conceptual core to which the loosening and narrowing pragmatic inferences are applied, as in (1) where the interpretation of tired becomes too tired to go to the cinema through narrowing. Moreover, it was argued that conceptual information can be inferentially developed to obtain an explicature (Blakemore, 2002) whereas procedural information triggers specific inferences opposed to more general pragmatic inferences (Saussure, 2011; Cúrcó, 2011), as in (2) where puisque instructs the hearer that the information from the following is given and mutually manifested to both the speaker and the hearer (cf. Zufferey, 2014 for an experimental validation). More precisely, procedural expressions encode specific paths to follow for obtaining specific inferences (‘[procedures] do not instantiate general processing principles or vague guidelines for interpretation’ as noted by Cúrcó, 2011, 42).

(1) Peter: Do you want to go to the cinema? Mary: I am tired. (Sperber and Wilson, 1998/2012, 39)

(2) Jean est arrivé, puisque sa voiture est sur le parking. (Saussure, 2011, 61)

‘John has arrived, given that his car is in the parking lot.’

The first consequence of the existence of specific inferential paths for the hearer is that it is possible to make the intended inference in the absence of the procedural expression but this would happen (though it is not guaranteed) at a higher cognitive cost. The second consequence is that more general inferences that are not encoded specifically by linguistic expressions exist. These are non-demonstrative inferences, which make use of general world knowledge and which must be mutually manifest to the speaker and to the addressee. As Sperber and Wilson (1995, 65) put it “even under the best of circumstances, communication may fail.” This is the case for inferences obtained through general ways of pragmatic reasoning, as in (3), in which the negative answer may fail to be recovered if the addressee does not know that bread contains gluten and that people allergic to gluten should not consume it.

(3) - Do you want some bread? – I am gluten intolerant.

---

\(^6\) This idea was previously suggested for French markers expressing temporality by Moeschler (2000, 2003), arguing that procedural information, encoded by temporal connectives for example, is stronger than conceptual information, encoded by aspectual classes.
Behind these criteria there is the assumption that an expression can encode only one of the two types of information. However, numerous studies argued for a dual account: a linguistic expression can encode both procedural and conceptual information (see e.g. Nicolle, 1997; Moeschler, 2002, 2009; Fraser, 2006, and more recently Pons Borderia 2008; Moeschler, 2015, 2016; Wilson, 2011; to name but a few). For example, Moeschler (2016) argues that connectives are complex linguistic units conveying both conceptual and procedural information. He illustrates through an analysis of close connectives in French (parce que ‘because,’ donc ‘therefore’ and et ‘and’) the presence of conceptual and procedural contents, which trigger different levels of meaning. More precisely, they all share causal conceptual content, which plays a role for determining the entailment, the explicature, and respectively, the implicature of the proposition, but differ regarding the direction of the causal relation: cause-consequence for donc and et, and consequence-cause for parce que.

In conclusion, as they are, the existent qualitative features and criteria of conceptual and procedural types of information do not allow objective analyses and replicable studies. We can observe a great degree of heterogeneity with respect to the features of procedural and conceptual information, their semantic or pragmatic nature (cf. Cucicó, 2011), their language dependent vs. cross-linguistic status and finally, their dual vs. single application to a linguistic expression.

The following section is dedicated to introducing and discussing a quantitative measure that allows researchers to empirically test the content(s) encoded by a linguistic expression, to evaluate and to compare studies targeting the same or similar expressions.

3. A new proposal

The description of conceptual and procedural information as presented in section 2 seems to be accurate and to reflect the linguistic and cognitive reality of language users. However, despite these numerous qualitative features of conceptual/procedural information, no study has suggested a description in quantitative terms. Quantitative criteria would have one important advantage, namely that of permitting objective investigations and comparison among expressions and among studies. For example:

- Comparison in synchrony of linguistic expressions in order to know, for example, if one is more accessible to consciousness than another, and if it is, to what extent. In other words, is there a black-and-white distinction or a continuum between conceptual and procedural types of information at a certain stage in the evolution of a language?\(^7\)
- Comparison among expressions of the same type (procedural or conceptual), such as the French parce que and puisque (‘because’ or ‘since’) in order to find out if they activate similar or even identical procedures and to what degree of activation (i.e. low-medium-high).
- Comparison among studies that investigate the same linguistic expression in order to evaluate them on a basis other than if the reader found the explanation plausible or not.

For this purpose, I propose to use annotation experiments and inter-annotator agreement in order to investigate comprehenders’ intuitive behaviour regarding the type of meaning associated with a linguistic expression or pragmatic phenomenon. In section 3.1, I discuss the close relationship between reliability and validity, a series of technical aspects of using inter-annotator agreement, and I suggest a list of requirements that must be adhered to when annotation experiments are carried out.

3.1. Inter-annotator agreement, reliability and validity

Inter-annotator agreement is widely used in corpus linguistics, computational linguistics, discourse studies and empirical pragmatics for evaluating the agreement between two or more annotators when dealing with various types of linguistic information, ranging from semantic information to syntax, discourse phenomena (discourse relations, discourse connectives), figurative language, and pragmatic usages of linguistic expressions, to name but a few. The necessity to use inter-annotator agreement rates was triggered by scholars’ worries about the subjectivity of the judgments required to create annotated resources, which may further serve as gold-standard data\(^8\) for training, testing, and evaluating the

\(^7\) This is different from considering the diachronic evolution of the meaning of expressions from conceptual to procedural (cf. Nicolle, 1998 for a discussion of grammaticalization and the conceptual/procedural distinction). For example, Bolly and Degand (2013) empirically investigated two French discourse markers encoding procedural information, namely vu que ‘since’ and on a vu/nous avons vu que ‘we have seen that’ which have their roots in the verb voir ‘to see’.

\(^8\) Gold standard data refers to trustworthy and reliable annotated data, generally coming from human annotation experiments and used for training and meaningful evaluation of algorithms in machine learning (Wissler et al., 2014).
performances of automatic tools. Hence, the main purpose was to ensure reliability, defined as the adequate “consistency among independent measures intended as interchangeable” (Moss, 1994, 7) and validity, defined as the “consonance among multiples lines of evidence supporting the intended interpretation over alternative interpretations” (Moss, 1994, 7).

According to Krippendorff (1980), reliability has three facets: stability of the process over time, reproducibility of the process under varying circumstances, at different locations and using different annotators, and accuracy, which refers to the degree to which a process conforms to a known standard. Potter and Levine-Donnerstein (1999, 271) point out that among these three facets, reproducibility is “the strongest realistic method by default” to assess reliability. This is the case because stability is directly dependent of the annotators’ memory while accuracy is not always achievable since in some cases known standards do not exist. As for validity, it may be established through a two-step process. The first is to develop an annotation scheme, which guides the annotators in the analysis of the content submitted to them for judgement. According to Poole and Folger (1981, cited by Potter and Levine-Donnerstein), annotation guidelines are “a translation device that allows investigators to place utterances into theoretical categories” (Poole and Folger, 1981, 477).

Hence, when the annotation guidelines are anchored in a theory, their validity can be assessed against theoretical predictions. The second step for establishing validity is to assess the annotators’ judgement against a known standard. As Potter and Levine-Donnerstein (1999, 266) point out, this can be done when such a standard exists. When this is not the case, they suggest that the annotators’ intersubjective judgements (that is, judgements which are subjectively derived but shared across annotators) should be used as a standard.

Therefore, Potter and Levine-Donnerstein point to five key elements that are essential for having a reliable and valid study: the annotation guidelines, the theory, the standard (if it exists), the inter-subjectivity of judgments (inter-annotator agreement), and the replicability of the results. In the following paragraphs, I will tackle several issues regarding inter-annotator agreement, specifically the measures for inter-annotator agreement, their limits and their advantages for creating human-annotated data and for quantifying the naïve speakers’ intuitions about linguistic phenomena.

One of the first possible measures for inter-annotator agreement rate is the percentage agreement. The percentage agreement is the ratio of observed agreements either between two judges or for the majority of opinions among several judges. The problem with percentage agreement is that it does not disentangle between agreement due to chance and annotators’ agreement. If we consider the case of two judges, the amount of agreement we would expect to occur by chance (if annotators took a decision without taking into account the annotation guidelines) depends on two conditions:

- The number of categories (e.g. a binary distinction as is the case for mutually exclusive antonyms dead/alive, or a distinction with more than two categories as is the case for other antonyms such as beautiful/very beautiful/ugly/very ugly).
- The frequency of the categories. When the categories are equally frequent then the data is normally distributed. When one category is much more frequent that the other(s), the data is skewed.

Given two studies investigating the same phenomenon, the one that will use a smaller number of categories will have higher agreement rates just by chance. For example, for two equally frequent categories there is a 50% chance for the two judges to make the same judgement. The proportion is based on the fact that there are only two choices.

In order to avoid the problem of agreement rate due to chance, inter-annotator agreement can be calculated with a series of chance-corrected coefficients, such as such as Cohen’s kappa (Cohen, 1960; Carletta, 1996) or Aickin’s alpha (1990) among others. The most frequently used is Cohen’s kappa (Carletta, 1996) (henceforth K), whose values range from 0 (signalling that there is no other agreement than that expected by chance) and 1 (signalling perfect agreement). In studies with more than two judges, several measures can be used to calculate inter-annotator agreement. One possibility is to measure agreement separately for each pair of judges and report the average (Artstein and Poesio, 2008). Another possibility is to measure pairwise agreement instead of percentage agreement. According to Artstein and Poesio (2008, 562), pairwise agreement for a certain item is the proportion of agreeing judgement pairs out of the total number of judgements for that item. In other words, it refers to calculating the majority of labels given by the annotators for each item. In section 3.3, I show how pairwise agreement was used for judging French verbal tenses expressing past tense by three and by seven judges.

When used for creating gold-standard data, the K coefficient has two main drawbacks: prevalence and annotator bias. Prevalence refers to the application of the K coefficient to severely skewed data, in which a majority of items fall under one of the two (or more) labels to be assigned. When a disproportionate amount of data falls under one category, the expected agreement is very high but the value of the chance-correcting coefficient is low. The low value of K in this case is due to its assumption that the observed judgments for each category may equally lead to agreement by chance. Hence, in order to

---

9 Such as Scott’s pi (Scott, 1955), Bennet et al.’s S (1954), and Gwet’s AC (2008) in studies with two annotators and two categories; and Krippendorff’s alpha (Krippendorff, 2004), multi-kappa (Davies and Fleiss, 1982) and multi-pi (Fleiss, 1971) in studies with more than two annotators.
ensure the reliability of the annotation, \( K \) is appropriate only for normally distributed data. In order to overcome this limitation, scholars have proposed additional statistics, such as the observed agreement, which avoids the difficulties encountered by \( K \) in skewed data but does not correct for chance agreement (Spooren and Degand, 2010). Annotator bias refers to the possibility of having individual judgement strategies. In order to steer clear of this downside, Artstein and Poesio (2008) show that the effect of the variance of the individual judges’ distributions of answers on the value of the chance-corrected coefficient becomes more similar to random noise as the number of judges increases. This also allows us to avoid the case when a judge has a deviant reading or interpretation of the annotation guidelines.

In computational and corpus linguistics, the generally accepted threshold of \( K \) for trustworthy categorization is around 0.6–0.7. However, for pragmatics and discourse studies using this method, Spooren and Degand (2010) argued that \( K \) values lower than this threshold are frequent. According to them, there are two possible explanations for lower \( K \) values in discourse studies. The first is that language is semantically underdetermined, redundant, and economical. Hence, the addressees must interpret it in context, which might trigger a different interpretation from a comprehender to another. The second is the possibility to have coding errors, which can be: (i) errors regarding the initial working hypotheses due to the fact that he annotation guidelines do not capture entirely the considered phenomenon, and (ii) errors due to individual strategies for each judge. They suggest three methods for reducing coding errors and increasing the reliability of the results. The first is double coding, which consists of a discussion of disagreements: individual strategies become cooperative strategies since this strategy requires making explicit the reasoning on which the judgement is based and convincing the other annotator of the quality of the reasoning (e.g. Sanders and Spooren, 2009 used double coding for their analysis of two connectives indicating causality in Dutch). The second one is one-coder-does-all method, which relies on probably subjective but systematic judgments. Spooren and Degand (2010) explain lower \( K \) values with respect to the type of information encoded and its high context-dependence due to the fact that language is underdetermined.

Using the example of discourse relations, which can be marked explicitly or remain implicit, they write:

> A coherence relation like cause-consequence can be marked explicitly (using a connective like because), or it can remain implicit (no connective), in which case the coherence has to be inferred; [...] This implies that establishing the coherence relation in a particular instance requires the use of contextual information, which in itself can be interpreted in multiple ways and hence is a source of disagreement. (Spooren and Degand, 2010, 254–255)

The third is the use of descriptive statistics, such as observed and specific agreement, and a discussion of the possible reasons for disagreements. These measures should complement the interpretation of the \( K \) value.

However, when annotation experiments are used to investigate naïve speakers’ intuitive behaviour about a linguistic or pragmatic phenomenon, the constraints mentioned above regarding annotator bias or methods to improve the value of \( K \) are no longer relevant. As Spooren and Degand (2010, 254) point out when speaking about the one-coder-does-all strategy,

> Of course the coding will be subject to individual strategies developed by the coder, but these strategies will presumably be systematic and there is no reason to assume that such strategies will be conflated with the phenomena of interest. [...] So if our research question is whether judgements occur more often with want than with omdat, an overcoding of judgments will not impede answer to the research question.\(^{10}\)

This means that the annotators’ strategy corresponds to their way of understanding the phenomenon of interest.

To increase the reliability (defined as having three facets: stability, replicability, and accuracy) of the results of annotation experiments, we can formulate a series of requirements that must be respected. These requirements are identified partly from the computational linguistics and corpus linguistics literature (Carletta, 1996; Artstein and Poesio, 2008; Spooren and Degand, 2010) and from pragmatic studies that used the \( K \) coefficient mainly for annotating discourse connectives in L1 and L2, discourse relations (such as Degand and Pander Maat, 2003; Sanders and Spooren, 2009; Cartoni et al., 2013a; Zufferey et al., 2012, 2015; Zufferey and Degand, 2013), verbal tenses and subjectivity (Grisot and Moeschler, 2014; Grisot, 2015, Grisot, 2017). These requirements are illustrated in section 3.3, in which I show how this quantitative measure can be used to study the contribution of verbal tenses to the interpretative process. The reliability requirements are as follows:

1. A large quantity of data, with randomly chosen samples, in which each category is equally represented, in order to avoid prevalence so that coefficients as \( K \) can be used to measure inter-annotator agreement.

\(^{10}\) Here the authors make reference to Sanders and Spooren (2009) study in which they annotated two Dutch connectives: omdat which is most frequently used in objective causal relations (that is, they express causality between events in the real world) and want which is considered to be a prototypical marker of subjective causal relations holding between the speaker’s conclusions on the basis of events in the world (Degand and Pander Maat, 2003; Pit, 2003; Canestrelli et al., 2013).
2. Data should be both naturalistic (from corpora with its original context) and artificial (built and carefully controlled items) in order to avoid having biased naturalistic (i.e. uncontrolled but natural) and biased built data (controlled but artificial). In section 3.3, I will show that dealing with corpus data was shown to be more difficult than dealing with artificial controlled data. High and, respectively, low $K$ values might be triggered by judging artificial data and, respectively, corpus data.

3. Clearly expressed guidelines based on a solid theoretical basis and written instructions for the judgement task fixed in advance and usable for other annotation experiments in order to allow replicability. The guidelines should include examples of each label. This requirement is necessary because the interpretation of the results depends on the falsifiable hypotheses previously formulated.

4. Testing the accurate comprehension of the guidelines and of the examples of all values of the categories in the training phase by having a debriefing session. This is necessary in order to avoid low $K$ values due to the misunderstanding of the task itself.

5. Participants: two or more native speakers who pass through a training phase and who work independently of each other – providing thus independent samples of annotations of the same data. The amount of training is important because it influences the annotators’ results, as suggested by Bayerl and Paul (2011) and Scholman et al. (2016). The choice of naïve non-trained or highly trained annotators depends on the purpose of the study. Using highly trained annotators is better if one aims at producing annotated data (Scholman et al., 2016). However, if one’s purpose is to investigate how comprehenders deal with a certain type of content, using naïve non-trained annotators allows the researcher to capture their intuitive use of the language. Regarding the number of annotators, having several annotators has the advantage of reducing the influence of the? Annotator bias effect on the value of the chance-corrected coefficient.

6. Evaluation: inter-annotator agreement rate with a chance-corrected coefficient, supplemented by a critical discussion of the disagreements and their integration in a theoretical model, as well as descriptive agreement measures.

7. Test the statistical significance of the agreement with a Chi-squared ($\chi^2$) test\(^{11}\) applied to the frequency table (the Fisher exact test when the number of observations is less than 5 in one of the cells).

8. If possible, replication of the study with new items and new participants but with the identical guidelines and experimental setup.

These requirements ensure the reliability of the study, pointing to its three components: stability, reproducibility and accuracy. This is needed because in the current study, inter-annotator agreement measured with the $K$ coefficient is no longer used as a measure for reliable vs. unreliable annotations carried out to create gold-standard human annotated data but as an indicator of the type of information that is judged in the experiment. This is due to the fact that the purpose of these annotation experiments is to investigate the naïve speakers’ intuitions about encoded meanings and purely pragmatic interpretations of utterances. In other words, I am interested in the annotator's way of understanding and dealing with conceptual, procedural and purely pragmatic information.

In section 3.2, I show that the use of inter-annotator agreement as measure for conceptual, procedural, and pragmatic information is based on the cognitive foundations of the conceptual/procedural distinction proposed by Wilson and Sperber (1993/2012) and their definition of human comprehension as a non-demonstrative inferential process.

3.2. Hypotheses and interpretation

Similar to Spooren and Degand (2010), I show that the value of the chance-correcting coefficient measuring inter-annotator agreement rate primarily depends on the type of handled semantic, and respectively pragmatic, information. In other words, the $K$ value will reach an upper threshold, which is directly dependent on the type of information that is judged. This does not prevent the $K$ value from being improved through various techniques (such as those suggested by Spooren and Degand, 2010, Cartoni et al., 2013a, or Grisot and Moeschler, 2014).

As noted in section 2, the conceptual/procedural distinction was introduced in order to explain the difference between content words and discourse connectives. The fact that discourse connectives encode procedural information was rarely challenged in the literature (e.g. Bach's 1999 treatment of *but* or Moeschler's 2002 treatment of French *parce que* ‘because’). Cartoni et al. (2013a) showed through two offline experiments with linguistic judgement tasks, in which native speakers were asked to consciously evaluate the meaning of French *alors que* ‘while’ and English *while* that this task is difficult. For each experiment, in which two different judges participated, the $K$ value was 0.42.\(^{12}\) Cartoni et al. (2013a) explain the low $K$ value in the first experiment in terms of the procedural meaning of connectives. Using the same method,

\(^{11}\) The $\chi^2$ tests the degree of fit between the observed and the expected distribution of the categories. The null hypothesis postulates that the data are distributed randomly/evenly and that if there is a difference in the tested sample, it is due to random variation.

\(^{12}\) For other studies with an important level of inter-annotator disagreement when judging explicitly the meaning of connectives, see Prasad et al. (2004, 2008) for building the Penn Discourse Treebank.
Grisol (2015) and Grisol (2017) applied this measure to verbal tenses and to subjectivity in cross-linguistic corpus data (see section 3.3).

Based on Wilson and Sperber’s (1993/2012) cognitive foundations of the conceptual/procedural distinction, as well as on Sperber and Wilson’s (1986/1995) description of non-demonstrative inferences, we expect to find different behaviours for native speakers when they consciously evaluate the two types of encoded information and purely pragmatic inferential information. Hence, the hypotheses formulated based on Wilson and Sperber's descriptions are:

- Conceptual meaning is available to conscious thought. Consequently, judging conceptual information is a rather easy task and it results in high inter-annotator agreement rates.
- Procedural meaning is more difficult to consciously evaluate than conceptual information. Consequently, judging procedural information represents a more difficult task than judging conceptual information, and it results in moderate inter-annotator agreement rates.
- Pragmatic meaning is recovered through non-demonstrative inferences, which do not guarantee the success of the communication. Consequently, judging pragmatic meaning, which consists in consciously identifying and naming the reason why one has a certain interpretation of the sentence one reads, results in a low inter-annotator agreement rate.

As Sperber and Wilson (1986/1995) postulated, during communication the addressee's task is to identify the speaker's informative intention, i.e. the content they want to transmit, along with their communicative intention. For this, the addressee has to accomplish a series of sub-tasks, namely determining the logical form of the speaker's utterance (through decoding), the propositional form (through inference) called (first level) explicature, the speaker's propositional attitude called higher-level explicature, and the implicature(s) of the speaker's utterance. Conceptual, procedural and pragmatic meanings play a role in performing these sub-tasks. Procedural meaning encoded by specific linguistic items triggers specific inferences which constrain this interpretative inferential process, whereas conceptual information is treated at the level of the explicatures of the utterance through pragmatic enrichment processes as narrowing and broadening, linked to the concept they encode. Pragmatic interpretations are recovered through general inferences, they are not triggered by a linguistic expression and depend on the contextual hypotheses that the hearer formulates. Therefore, they can be located at the level of implicatures (Wilson and Sperber, 1993/2012; Saussure, 2011). The interpretation of inter-annotator agreement regarding conceptual, procedural, and pragmatic types of meaning is given in Table 3.1.

<table>
<thead>
<tr>
<th>Inter-annotator agreement</th>
<th>Type of meaning</th>
<th>Type of operation</th>
<th>Level of meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (&gt;0.7)</td>
<td>Conceptual meaning</td>
<td>Narrowing, loosening</td>
<td>Explicature</td>
</tr>
<tr>
<td>Moderate (0.4–0.7)</td>
<td>Procedural meaning</td>
<td>Specific inference</td>
<td>Constraints on explicature or on implicatures</td>
</tr>
<tr>
<td>Low (&lt;0.4)</td>
<td>Pragmatic meaning</td>
<td>General non-demonstrative inference</td>
<td>Implicature</td>
</tr>
</tbody>
</table>

Examples (4), (6), and (8) illustrate these three levels (repeated below from section 2). For (4), the hearer builds the complete propositional form provided in (5) (its explicature), through narrowing the pro-concept13 TIRED and building the ad hoc concept “TIRED, that means here ‘too tired to go to the cinema.’” In (6), the connective puisque gives a specific inferential path to follow, which is interpreting the segment following it as given information (cf. Zufferey, 2014). The explicit interpretation of (6) is provided in (7), where the procedural content of puisque is rendered explicit. Finally, in (8), Mary's interpretation of John's answer is dependent on the contextual hypothesis she formulates. If she makes the hypothesis that a gluten intolerant person will never eat bread, then she will infer a negative answer. However, if she makes the hypothesis that a gluten intolerant person might, from time to time, eat bread, she might infer a positive answer and/or ask for confirmation.

(4) Peter: Do you want to go the cinema? Mary: I am tired.
(5) Mary is too tired to go to the cinema.
(6) Jean est arrivé, puisque sa voiture est sur le parking.
   ‘John has arrived, since his car is in the parking lot.’

---

13 Relevance Theory adopts the view that many words seem to encode not a full-fledged concept but a very broad concept, called a pro-concept, which is either narrowed or loosened in a given utterance (Sperber and Wilson, 1998/2012; Carston, 2004). For example, the general concept TIRED encoded by the lexical item tired could give access depending on the utterance to a large number of ad hoc built concepts of tiredness, such as too tired to go to the cinema, too tired to go the gym, too tired to go to work, tired of listening to someone, exhausted, etc.
Jean est arrivé, comme tu le sais déjà, sa voiture est sur le parking.
‘John has arrived, as you already know, his car is in the parking lot.’

Mary: Do you want some bread? John: I am gluten intolerant.

There are three issues at stake regarding this interpretative scale. The first is hypothesis formulation. As in all type of experiments, the interpretation of the results and their significance are dependent on the hypotheses formulated based on the theory. In other words, the application of the linguistic criteria discussed in section 2 and the interpretative scale from Table 3.1 are complementary methods. To determine whether a linguistic expression encodes conceptual and/or procedural meaning, annotation experiments indicate what kind of meaning is at play beyond the researcher’s own intuition. The values suggested are indicative thresholds. They recall Landis and Koch’s (1977) conventions concerning the K-like coefficients, where the 0.4 value distinguishes between fair and moderate agreement. Further experimental investigations are required to validate them.

The second issue is the necessity for a study using this interpretative scale to adhere to the requirements provided in 3.1 in order to ensure the reliability of the data. This is as crucial as respecting a series of postulates in statistics in order to correctly use inferential statistics (such as the independence of observations, controlling of variables, randomisation in choosing a sample and in assigning items to participants, replication of the observation using a different sample), or in order to be able to apply a certain type of test and accurately interpret it (e.g. having a normal vs. non-normal distribution of the data, having independent groups vs. repeated measures, etc.).

The third issue is the risk of making erroneous interpretations. These risks are similar to Type I and Type II errors in statistics. For example, high rates of the inter-annotator agreement might erroneously be interpreted as conceptual meaning instead of procedural meaning in case of problematic annotation guidelines (for example, if they are unrealistic they render trivial the judgement task), or in case of non-respect of the independence requirement regarding the annotators’ performance. Similarly, low rates might erroneously be interpreted as procedural meaning instead of conceptual meaning in case overly complex annotation guidelines, which render the task excessively complicated. Replicating the study using a different sample is one of the options to confirm the initial interpretation of inter-annotator agreement rates.

In section 3.3, I will illustrate how the methodology presented above is applied to verbal tenses. More precisely, I will point to three types of content linked to the contribution of verbal tenses to the interpretative process. These are the localisation of an eventuality (event or state) with respect to the moment of speech (section 3.3.1), the localisation of an eventuality with respect to another eventuality (section 3.3.2), and the capacity of certain verbal tenses to make available a subjective point of perspective (section 3.3.3). In section 3.3.4, I will discuss how these annotation experiments provide evidence in favour of the interpretative scale provided in Table 3.1.

3.3. Illustration with verbal tenses

In Grisot (2015, and previous work), I investigate the type of meaning encoded by verbal tenses in a cross-linguistic approach. In discourse, verbal tenses establish temporal reference, which is situated at two levels: they locate eventualities with respect to the moment of speech (S) and with respect to one another (when several eventualities are involved). Furthermore, it was suggested that certain verbal tenses make available the speaker’s subjective viewpoint on the situation described (for example, Genette, 1972; Fleischman, 1990; Traugott, 1989, 1995, 1999; Stioul, 2000; Tahara, 2000; Saussure, 2013; Moeschler, 2014, among others). These three functions of verbal tenses were investigated experimentally in annotation experiments with linguistic judgement tasks (Grisot, 2015, Grisot, 2017). These experiments have revealed three systematic patterns. When participants deal with the localisation of eventualities with respect to S, that is in the past or non-past (present or future), they point out the easiness of the task and have high rates of inter-annotator agreement. When they deal with the localisation of an eventuality with respect to another, they manifest greater difficulty for the task and have lower rates of inter-annotator agreement. When they have to consciously identify a subjective or non-subjective perspective, they point out the great difficulty of the task and have rates of agreement similar to chance. These three patterns are interpreted in terms of different cognitive costs required for accomplishing these tasks: reduced cost for the first, higher costs for the second, and very high for the third. I argue that this observed difference may be explained in terms of different contents the addressee deals with: conceptual, procedural, and purely pragmatic, respectively.

Therefore, the results of the experiments carried out challenge the current procedural model of verbal tenses (e.g. Moeschler et al., 1998; Moeschler, 2002; de Saussure, 2003, 2011; Amenós-Pons, 2011) and confirm the current assessments of grammatical and lexical aspect (Escandell-Vidal and Leonetti, 2011; cf. Magliano and Schlech, 2000 for an experimental study). In Grisot (2015), I propose a revisited theoretical model of temporal reference according to which the categories of tense and aspect (lexical and grammatical aspect) encode conceptual and procedural types of meaning. Distinguishing among these types of meaning is crucial for having a fine-grained understanding of the usages of verbal
tenses and of their contribution to the interpretative process. The proposal is the following. A verbal tense encodes conceptual information, which refers to the localisation of an eventuality with respect to S, and procedural information, which refers to the localisation of an eventuality with respect to another eventuality (phenomenon classically treated as temporal sequencing). As for aspectual information, grammatical aspect encodes procedural information (in the form of constraints that apply to eventualities) whereas lexical aspect encodes conceptual information (i.e. the eventualities that hearers manipulate during the interpretative process). Accordingly, it is suggested that a verbal tense has contextual usages associated with the various combinations of conceptual and procedural contents of the subsequent categories: tense and grammatical aspect applied to lexical aspect. In Grisot (2017), I show that subjectivity, roughly the expression of the speaker's subjective viewpoint, is not triggered by linguistic expressions, such as verbal tenses or grammatical aspect (when it is morphologically expressed). On the contrary, it is a global phenomenon and it is the result of a purely pragmatic interpretation using general pragmatic inferences.

In what follows, I illustrate how annotation experiments and inter-annotator agreement rates can be used for testing the type of meaning of a linguistic expression, be it lexical or functional. In the experiments from Grisot (2015), the information regarding localisation with respect to S was operationalised as the past/non-past distinction and the information regarding temporal relations was operationalised as the \([\pm\text{narrativity}]\) feature. All the experiments carried out follow the annotation protocol and the list of requirements provided in section 3.1.

3.3.1. Annotation experiment targeting the past/non-past distinction

The material used in this experiment consisted of 90 items of naturally occurring data from corpora judged in their original contexts as in (9) and 36 artificial sentences built for the purpose of the experiment as in (10). The target verb was provided in the infinitive form and the participants' task was to provide its inflected form based on the context. For the built example, the past or non-past time context was set by the first sentence.

(9) Le jeune soldat mis en cause a agi contre les ordres de ses supérieurs, il (être) aujourd'hui incarcéré et en attente d'être jugé pour meurtre.
   ‘The suspected young soldier behaved against his superior's orders, he (to be) today imprisoned and waiting to be sentenced for murder.’

(10) Jean a pris du poids. Avant de casser sa jambe, Jean (courir) tous les soirs pendant une heure.
   ‘John gained weight. Before breaking his leg, John (to run) every evening for an hour.’

All experimental items were distributed in sets of 15 items for the corpus data and 18 items for the artificial sentences (a total of 8 sets). Each participant received either corpus (natural) or built (artificial) data. Each experimental item was judged by 6 participants. The participants were 48 native speakers of French (students at the Faculty of Humanities). The procedure was as follows: Participants received the annotation guidelines, in which the task of the experiment was explained, and went through a training set of 3–5 items. Then, they received the set of items to annotate in an independent manner. The results were evaluated by calculating pairwise agreement.

The results are as follows: when judging corpus data, the judges agreed on their label for 119 items and disagreed on 3 items. 4 items were evaluated as inconclusive (no majority). Disagreements and inconclusive items were disregarded from further analysis. A correspondence between the annotators’ label and the correct inflected form was found for 111 items, which corresponds to a \(K\) of 0.86. Using a Chi-square test, this correspondence was shown to be statistically significant (\(\chi^2 = 51.41, df = 1, p < .0001\)). In other words, the annotators' labels are significantly different than the 50% rate of expected distribution of labels. When judging artificial data, the judges' label was identical to the correct inflected form for the totality of 36 experimental items, which corresponds to a \(K\) value of 1 signalling perfect agreement.

This experiment indicated that speakers have no difficulty in consciously evaluating the localisation of eventualities with respect to the moment of speech. The high \(K\) value of 0.86 suggests that this information is highly accessible to conscious thought and easily conceptualised. Based on Wilson and Sperber (1993/2012), this type of behaviour corresponds to conceptual meaning. Dealing with this kind of information is not cognitively costly because it points to concepts that speakers have and deal with in almost every utterance: the localisation of eventualities in the past or in the present.

3.3.2. The \([\pm\text{narrativity}]\) feature

Technically speaking, the \([\pm\text{narrativity}]\) feature refers to temporal relations holding between the eventualities referred to in two discourse segments, containing both forward and backward temporal directional inferences. It has two contextually determined possible values: narrative as in (11) and non-narrative as in (12) (here illustrated with the English Simple Past), where the former indicates that the two eventualities expressed are temporally ordered whereas the latter points to the fact that the eventualities take place simultaneously.
By his own marriage, likewise, which happened soon afterwards, he added to his wealth. (literary register)

The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister. (literary register)

The purpose of the following experiments was to test whether native speakers are able to identify this feature for the following verbal tenses: the English Simple Past (SP), the compound past (PC), the simple past (PS) and the imperfect (IMP) in French, Italian, and Romanian, which may be used to translate the SP in these languages. Scholars have suggested that the PS has narrative usages, the PC is undetermined with respect to this information whereas the IMP has non-narrative usages. The material used was corpus data. They were similar to the sentences from examples (11) and (12). The experimental items were randomly chosen from the bilingual and multilingual corpora built by Grisot (2015). The experiments carried out are summarised in Table 3.2, where French$_2$ represents the replication of the first experiment carried out on French verbal tenses (French$_1$).

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of experimental items</th>
<th>Verbal tense judged</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>458</td>
<td>SP</td>
</tr>
<tr>
<td>French$_1$</td>
<td>300</td>
<td>PC, PS, IMP</td>
</tr>
<tr>
<td>Italian</td>
<td>85</td>
<td>PC, PS, IMP</td>
</tr>
<tr>
<td>Romanian</td>
<td>85</td>
<td>PC, PS, IMP</td>
</tr>
<tr>
<td>French$_2$</td>
<td>84</td>
<td>PC, PS, IMP</td>
</tr>
</tbody>
</table>

The participants were how many native speakers of the language tested in the experiment in which they participated. The way in which they were recruited, the number of annotators, and their age varied how from one experiment to another. The procedure of the experiments was as follows: the participants received the annotation guidelines, in which the [±narrativity] feature and its two values (narrative and non-narrative) were defined and exemplified. They also received training on 6–10 items, which they annotated independently. The training was followed by a debriefing session, in which each participant explained how they judged the items. The purpose of the training was to be sure that the annotators understood the annotation guidelines. The participants independently annotated the experimental items. The results were evaluated using inter-annotator agreement rates measured with the K coefficient or using pairwise agreement when there were more than two participants. When a generally accepted standard existed (cf. discussion in section 3.1), the participants’ judgments were compared to that standard. Disagreements were discussed in the second round of the experiment with the application of the double coding technique (Spooren and Degand, 2010), in which the same participants were asked to explain their judgement and to try to find agreement for each item. In the experiment on the English SP, the participants were asked to suggest a connective that would render the existing implicit temporal relation explicit.

The results of each experiment are summarised below.

The two English judges agreed on 325 items and disagreed on 133 items, corresponding to a K value of 0.42. Using a Chi-square test, this distribution was shown to be statistically significant ($\chi^2 = 80.53, df = 1, p < .0001$). After applying the test of suggesting a connective, the inter-annotator agreement rate went up to 0.91. In the data containing agreements, the SP was judged as having narrative usages in 59% of the cases and non-narrative usages in 41% of the cases. In the French$_1$ experiment, 76 French speakers (students at the Faculty of Humanities) participated. Each of them saw 15 items and each experimental item was judged by 4 judges. The evaluation was done through calculating pairwise agreement: 79 items received equal judgement and for 221 items a majority was calculated. Among these 221 items, the PS was judged as narrative in 92% of the cases, the PC was judged as narrative in 77% of the cases and the IMP was judged as non-narrative in 77.5% of the cases. The annotators’ judgement was compared to a standard, which corresponds to the expected usages of these three verbal tenses: 100% non-narrative for IMP, 100% narrative for PS and 50% narrative for the PC. The annotators agreed with the standard in 82% of the cases, which correspond to a K value of 0.63. Using a Chi-square test, this correlation between the annotators’ judgement and standard is statistically significant ($\chi^2 = 86.96, df = 1, p < .0001$).

---

14 In future research, these experiments could be repeated using both corpus and built data in order to strengthen the current interpretation of the results.

15 Scholman et al. (2016) propose a step-wise annotation procedure applied to discourse relations, in which substitution and paraphrase tests are introduced in order to guide judges (naive native speakers) during the experiment.
The two Italian judges agreed on 64 items and disagreed on 21 items, corresponding to a $\kappa$ value of 0.41. After the discussion, the judges agreed on 76 items, which represents a $\kappa$ value of 0.74. As far as the analysis of individual verbal tenses is concerned, the data containing only agreements was considered (76 items). Specifically, the IMP was judged as non-narrative in 84% of the cases, the PC was judged as narrative in 88% of the cases, and the PS was judged as narrative in 96 of the cases. Using a Chi-square test, this distribution was shown to be statistically significant ($\chi^2 = 44.09$, df = 1, $p < .0001$).

The two Romanian judges agreed on 64 items and disagreed on 21 items, corresponding to a $\kappa$ value of 0.42. Using a Chi-square test, this distribution was shown to be statistically significant ($\chi^2 = 28.09$, df = 1, $p < .0001$). The two judges discussed disagreements, which resulted in a $\kappa$ value of 0.75. The analysis of individual verbal tenses was done on the data containing only agreements (64 items). Specifically, the IMP was mainly judged as non-narrative (71%), the PC (83%) and the PS (93%) as narrative.

In the French$_2$ experiment, 49 French speakers participated (students at the Faculty of Humanities, who had not taken part in the French$_1$ experiment). The data was organised in 7 sets of 12 items. The evaluation of each item was done through calculating pairwise agreement (cf. section 3.1). Table 3.3 is the agreement table indicating pairwise agreement for the first 7 items of the data and the French verbal tense that was judged (Di Eugenio and Glass, 2004; Artstein and Poesio, 2008). The judgements of the 7 judges are provided in the left side of the table and the results of only the first 3 judges are provided in the right side of the table. Table 3.3 shows that the judgments are exactly the same in the two cases indicating that, at least in this case, having more than 3 judges only confirms their judgments. In the clean data of 81 items, the label given by the annotators correlates with the verbal tense used in the corpus for 73% of the data. Judges clearly recognised a primary narrative usage for the PS (94%) but did not make the same clear judgement for the PC (65%) or for the IMP (69%) as being primarily non-narrative. This experiment replicated the findings of the French$_1$ experiment.

<table>
<thead>
<tr>
<th>Item</th>
<th>Narrative</th>
<th>Non-narrative</th>
<th>Verbal tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item1</td>
<td>5</td>
<td>2</td>
<td>IMP</td>
</tr>
<tr>
<td>Item2</td>
<td>6</td>
<td>1</td>
<td>PS</td>
</tr>
<tr>
<td>Item3</td>
<td>2</td>
<td>5</td>
<td>IMP</td>
</tr>
<tr>
<td>Item4</td>
<td>5</td>
<td>2</td>
<td>PC</td>
</tr>
<tr>
<td>Item5</td>
<td>4</td>
<td>3</td>
<td>IMP</td>
</tr>
<tr>
<td>Item6</td>
<td>4</td>
<td>3</td>
<td>PC</td>
</tr>
<tr>
<td>Item7</td>
<td>0</td>
<td>7</td>
<td>PC</td>
</tr>
</tbody>
</table>

The results of these experiments indicate that the [+narrativity] feature is identifiable by native speakers with medium $\kappa$ values (situated between the 0.4--0.7). This information receives narrative values most frequent for the PS and the PC, and non-narrative values for the IMP. Similar to English speakers, French, Italian, and Romanian speakers have poor abilities to consciously evaluate temporal relations triggered by verbal tenses. They have better results when they are asked to insert connectives, which express explicitly the same implicit content. This may be explained by the fact that a direct and conscious evaluation of the meaning of a procedural expression results in medium $\kappa$ values, whereas suggesting a different expression that renders explicit the implicit meaning of the initial procedural expression results in high agreement rates, therefore it increases the $\kappa$ value. These findings form a good empirical basis for arguing that the [+narrativity] feature is procedural and that it is a cross-linguistically valid feature whose result is context-dependent. In order to confirm this, future research should propose alternative methods for tracing this feature.

### 3.3.3. Subjectivity

In Grisot (2017), I present three experiments in which the notion of subjectivity is discussed and tested experimentally in three languages: English, French, and Serbian. Subjectivity was defined as *the speaker's viewpoint, psychological perspective and perceptions, which might or might not be included into the description of a situation or series of situations*. The definition I propose in Grisot (2017) provides two elements. Firstly, the locus of subjectivity is the speaker. Secondly, the linguistic sources of subjectivity are verbal tenses (the notion of perspective) and grammatical aspect (the notion of viewpoint) and perception verbs (following C. Smith 2003). The *purpose* of these experiments was to test whether or not native speakers of these languages are able to identify in naturally occurring data the notion of subjectivity, and whether or not verbal tenses can be linked to this notion (as it was suggested by Benveniste, 1966; Genette, 1972; Fleischman, 1990;...
Traugott, 1989, 1995, 1999; Sthioul, 2000; Tahara, 2000; Saussure, 2013; Moeschler, 2014, among others). For example, it was suggested that using IMP tense or imperfective grammatical aspect allows the hearer to include their viewpoint, perspective, and perception into the description of a situation, that is in a subjective manner, whereas using SP, PS, PC or perfective grammatical aspect allows the speaker to present a situation without including her viewpoint and perspective, that is, in a non-subjective manner.

For experimental testing, subjectivity defined as above was operationalised as the [±subjectivity] feature. With respect to this feature, sentences can be judged as subjective or not subjective (objective). A sentence is subjective when the description of a situation or a series of situations is centred on the speaker’s psychological perspective, as in (13). A sentence is not subjective when the speaker merely reports a situation or a series of situations that are related in the world, as in (14).

(13) She had an excellent heart – her disposition was affectionate, and her feelings were strong – but she knew how to govern them: it was a knowledge which her mother had yet to learn; and which one of her sisters had resolved never to be taught.

(14) The military recently blocked a government move to place Pakistan's infamous intelligence service, the ISI, under the control of the interior minister rather than the prime minister.

The material used is described in Table 3.4. The experimental items were similar to examples (13) and (14). The data in English and half of the data in French was randomly selected from texts originally written in English and in French from the corpus built by Grisot (2015). The other half of the data in French and the Serbian data was the translation of the English data into French and Serbian. As it will be shown below, no statistical significant difference was found for the results on French as original or translated language.

<table>
<thead>
<tr>
<th>Language</th>
<th>Total no. of experimental items</th>
<th>No. of items per verbal tense judged</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>99</td>
<td>SP</td>
</tr>
<tr>
<td>French</td>
<td>214</td>
<td>PC (72), PS (65), IMP (77)</td>
</tr>
<tr>
<td>Serbian</td>
<td>109</td>
<td>Verbs with perfective (54) and imperfective (55) morphemes</td>
</tr>
</tbody>
</table>

There were three participants for each experiment. They were native speakers of the languages tested in the experiment in which they participated. The procedure was identical to the procedure followed in the experiments presented in sections 3.3.1 and 3.3.2. The results are the following.

The three English native speakers, who judged occurrences of the SP with respect to the [±subjectivity] feature, had an agreement rate corresponding to a mean K value of 0.0508 (it corresponds to the mean of K value for each pair of the three annotators). This K value shows that the annotators did not have other agreement than that occurring by chance.

In the experiment on French, 105 native speakers participated. The 214 items were distributed in 21 sets. Each participant received only one set. Each set was judged by 5 participants. The clean data consisted of 203 items. There were 11 missing data because some participants did not judge all the items from the set they received. The mean K value for the inter-annotator agreement rate among the five judges was 0.29. 65 items received an identical judgement by all the judges and 138 items received an identical judgement by a majority of judges (half+1 or half+2). All three verbal tenses analysed were judged as having subjective and non-subjective usages. When looking at the 65 items, for which all the 5 judges agreed, the IMP was judged as subjective in 64% of the case, the PC in 33% of the cases, and the PS in 56% of the cases. The three Serbian annotators had a mean K value of 0.40. In the data of agreements, the imperfective viewpoint is most frequently judged as subjective (76%) than non-subjective (24%), whereas the perfective viewpoint is more frequently judged as non-subjective (54%) than subjective (41%).

These low values of K for English and French verbal tenses indicate that the information about the speaker's perspective is not encoded by verbal tenses, and that it depends on the contextual assumptions that the hearer may formulate. As for Serbian, subjectivity seems easier to identify when grammatical aspect is expressed morphologically than when it is not. However, these three experiments show that subjectivity seems to be a heterogeneous phenomenon, which is interpreted at the global level and which is not directly triggered by the categories of tense and aspect.

3.3.4. Discussion

These experiments were carried out following the same protocol and they respect the conditions provided in section 3.1. This indicates that the systematic differences in the participants' behaviour found in experiments from sections 3.3.1,
3.3.2 and 3.3.3, respectively, are not due to differences in the experimental protocol. As suggested by Spooren and Degand (2010) and Bayerl and Paul (2011), one factor that might influence inter-annotator agreement rates is the type of content dealt with in the experiment.\textsuperscript{16} My proposal is that the type of content can be conceptual, procedural, or purely pragmatic inferential meaning. As predicted by Wilson and Sperber (1993/2012), when participants deal with conceptual meaning of a linguistic expression, they have no difficulties to consciously evaluate it, as it is available to conscious thought and it is easily expressed in conceptual terms. In contrast, when participants deal with procedural meaning encoded by a linguistic expression, the procedure is automatically executed regardless of contextual assumptions. This procedure leads to a specific pragmatic inference, whose result depends on the contextual assumptions formulated. Consciously evaluating this type of meaning, which is not available to consciousness, is a rather difficult task for annotators. This is shown by their systematic behaviour when judging procedural information: the inter-annotator agreement rates are moderate. Finally, when annotators deal with subjectivity and its triggers by the perception, perspective and viewpoint categories of linguistic sources, as defined by Grisot (2017), they are not able to identify this pragmatic feature using these linguistic cues. They have inter-annotator agreement rates close to those expected to occur by chance. Based on this evidence, my suggestion is that comprehenders identify subjectivity defined as the speaker’s viewpoint, psychological perspective, and perceptions using a general pragmatic inference.

Inter-annotator agreement rates measured with chance-corrected coefficients allow us to capture the fine-grained differences between interpretations observed in the literature regarding verbal tenses: the localisation of situations in the past or non-past, the temporal sequencing phenomenon and the expression of the speaker’s subjective perspective. More precisely, comprehenders do not have similar performances when they are asked to judge these three types of interpretations. Their performance is significantly better when they judge the localisation of situations in the past or non-past, than when they judge temporal sequencing phenomenon (inter-annotator agreement rates are much higher than the chance level), or when they judge the subjectivity phenomenon (inter-annotator agreement rates close to chance agreement).

4. Conclusion

In this paper, I have discussed current assessments of the conceptual/procedural distinction, challenged the application of current cognitive and linguistic criteria used to distinguish between the two types of information, and proposed a quantitative measure that would allow objective and reproducible findings. This new measure will guarantee the validity and reliability of investigations of the different levels of the inferential interpretative process: pragmatic adjustment of conceptual meaning through narrowing and broadening, specific inferential paths signalled by expressions encoding procedural information, and general inferences (i.e. implicated conclusions formulated based on implicated premises and the propositional form of the utterance).

I argued that existing qualitative features, used as they are, unfortunately result in subjective judgments made by the researchers themselves. However, linguistic features, such as the behaviour with loosening and narrowing, can be used in a preliminary step of a research, e.g. during the formulation of theoretical hypotheses regarding the procedural and/or conceptual contents of an expression and/or as a task in offline experiments. Finally, the results of offline experiments in terms of inter-annotator agreement rate should be evaluated with a chance-corrected coefficient, such as the $\kappa$ statistic. Crucially, the interpretation of the values of inter-annotator agreement is dependent on the hypotheses formulated before the experiment, therefore any experiment must be based on a rigorous and solid preliminary theoretical investigation, as well as on the respect of a series of requirements regarding the choice of data and the protocol.

From a more conservative perspective, inter-annotator agreement rates can also be interpreted purely as an indicator signalling the presence of conceptual or procedural information. This indicator may be used to test theoretical working hypotheses resulting from the application of the cognitive and linguistic criteria discussed in section 2. For example, Relevance Theory predicts a difference in the cognitive treatment of the two types on information since they play different roles: conceptual information contributes to the construction of mental representations and procedural information provides instruction about how to manipulate them. This difference in cognitive processing was observed both in offline and in online experiments for connectives, as discussed in Cartoni et al. (2013a,b) and Zufferey (2012, 2014), and respectively, Zufferey (2014) for the connective puisque.

\textsuperscript{16}As one of the anonymous reviewers pointed out, there are other factors than the type of information judged that might inter-annotator agreement rates found in the experiments provided in section 3.3.1 compared to section 3.3.2 and to section 3.3.3. I think that since the same data selected from the bilingual and multilingual translation corpus was used in these experiments (even though there is only a partial overlap because a different number of items was used from one experiment to another), linguistic and semantics factors were kept constant. Therefore, one can conclude that the differences observed for inter-annotator agreement rates are due to the type of information judged. However, further research would be necessary in order to control for other factors, such as inter-individual variability, the order of items, the translation, the language, or the instructions.
Further work needs to be done to control for other factors that might influence inter-annotator agreement rates, such as inter-individual variation, the formulation of annotation guidelines, the order of the items, the length of the items, etc. Additionally, further research is needed to experimentally validate the indicative thresholds for the K-like coefficients and to complement this offline investigation with the online study of the cognitive operations involved when dealing with conceptual, procedural and pragmatic types of meaning.

References


Cristina Grisot carried out her doctoral research on temporal reference in English and Romance languages (French, Italian and Romanian), in which she adopted an empirical perspective (corpus work, contrastive analysis and offline experiments). She is currently working on off-line and on-line experimental investigations of human processing of verbal tenses, temporal connectives and temporal adverbials, as well as on their rich inter-relations. She is interested in the expression of temporal reference in natural language by looking at typologically different languages, such as tense-prominent, aspect-prominent and tenseless languages. Additionally, her research targeted new methods for improving the results of machine translation systems by adding pragmatic information (research carried out in two Swiss inter-disciplinary projects). More generally, she is interested in the human cognitive processing of temporality in natural language and its tight relations with causality. Currently, she is lecturer at University of Neuchâtel and a post-doctoral researcher at University of Geneva.