Chapter 4
Experimental Study Using Annotation Experiments

4.1 Dealing with Annotation Data: Inter-annotator Agreement and the K Coefficient

Inter-annotator agreement is widely used in corpus linguistics, computational linguistics, discourse studies and empirical pragmatics to evaluate 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. Inter-annotator agreement rates were needed because of scholars' worries about the subjectivity of the judgments required to create annotated resources, which may further serve as *gold-standard* data (i.e. trustworthy human-annotated data) for training, testing and evaluating the performance of automatic tools. As such, the main purpose was to assure *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).

As I ague in Grisot (2017a), following 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, of these three facets, reproducibility is 'the strongest realistic method by default' to assess reliability. This is the case because stability is directly dependent on the annotators' memory, while accuracy is not always achievable because, in some cases, known standards do not exist. Validity, on the other hand, may be established by 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

C. Grisot, *Cohesion, Coherence and Temporal Reference from an Experimental Corpus Pragmatics Perspective*, Yearbook of Corpus Linguistics and Pragmatics, https://doi.org/10.1007/978-3-319-96752-3_4

(1981, cited by Potter & Levine-Donnerstein), annotation guidelines are ‘a translation device that allows investigators to place utterances into theoretical categories’ (Poole and Folger 1981, 477). As such, 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 & Levine-Donnerstein 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 among annotators) should be used as a standard (p. 266). For them, inter-subjective judgements have the advantage in that they:

> give readers the sense that the patterns in the latent content[1] must be fairly robust and that if the readers themselves were to code the same content, they would make the same judgement.

So, Potter & Levine-Donnerstein point to five key elements which are essential for 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.

One of the first possible measures for inter-annotator agreement rate is *percentage agreement*. The percentage agreement is the ratio of observed agreements, either between two judges or in the majority of opinions among several judges. There is, though, a problem with inter-annotator agreement rate when it is measured by percentage agreement. This is *agreement due to chance*. 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 accounting for the annotation guidelines) depends on two conditions:

- The *number* of categories (e.g. a binary distinction, as with mutually exclusive antonyms such as *dead/alive*, or a distinction with more than two categories, as with other antonyms such as *beautiful/very beautiful/ugly/very ugly*).
- The *frequency* of the categories. When the categories are equally frequent, the data is normally distributed. When one category is much more frequent that the other(s), the data are not normally distributed, and are thus skewed.

Given two studies investigating the same phenomenon, the one with a smaller number of categories will have higher agreement rates simply by chance. For example, for two equally frequent categories, there is a 50% chance that, when one judge makes a decision, the second judge will make the same decision (a proportion based on the fact that there only two choices; for four categories, there is a 25% chance that the two judges will make the same judgment).

[1] Potter & Levine-Donnerstein distinguish between three types of content that can be dealt with in annotation experiments: manifest content (which is on the surface and easily observable, such as the presence or the appearance of a word); pattern content (objective patterns that all annotators should be able to uncover, such as lexical meaning); and projective content (contents for which the annotators’ content and world knowledge is required to judge meaning in context) (1999, 259).

In order to avoid the problem of agreement by chance, inter-annotator agreement can be measured with a series of chance-corrected coefficients, such as such as Cohen's *kappa* (Cohen 1960, Carletta 1996) or Aickin's *alpha* (1990). The most frequently used is Cohen's Kappa (Carletta 1996) (henceforth *Ϗ*), 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 option is measuring agreement separately for each pair of judges, and report the average (Artstein and Poesio 2008). Another option is measuring *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, calculating the majority of labels given by the annotators for each item.

In computational and corpus linguistics, the generally accepted threshold for trustworthy data is around 0.6–0.7. However, for pragmatics and discourse studies using this method, Spooren and Degand (2010) argued that *Ϗ* values lower than this threshold are frequent. According to them, there are two possible explanations for lower *Ϗ* values in linguistic studies. The first is that language is semantically underdetermined, redundant and economical, and so the addressees must interpret it in the context. The second is the potential for coding errors, which can be: (i) errors regarding the initial working hypotheses (the annotation guidelines do not entirely capture the considered phenomenon); and (ii) errors due to individual strategies for each judge.

They suggest three methods of reducing coding errors and increasing the reliability of the data. 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 method is *one-coder-does-all*, a method relying on systematic but probably subjective judgments. Spooren and Degand (2010, 254–255) explain lower *Ϗ* values with respect to the type of information encoded and its high context-dependence due to the fact that language is underdetermined. Their example is that of discourse relations, which can be marked explicitly or remain implicit. In their words,

> 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.

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 Ϗ value.

However, when annotation experiments are used to investigate naïve (i.e. untrained) speakers' intuitive behaviour when it comes to a linguistic or pragmatic phenomenon, the constraints mentioned above regarding annotator bias or methods of improving the value of κ are no longer relevant. As Spooren and Degand (2010, 254) say of 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 phenomenon of interest. […] So if our research question is whether judgements[2] occur more of often with *want* than with *omdat*¸ an overcoding of judgments will not impede answer to the research question.

This means that the annotator's strategy corresponds to his/her way of understanding the phenomenon of interest. In other words, one could expect that measuring inter-annotator agreement rates might be influenced by the type of information dealt with. In particular, based on Wilson & Sperber's cognitive foundations of the conceptual/procedural distinction (1993/2012) (cf. discussion in Sects. 2.3.1 and 2.3.2), one would expect to find systematically different behaviour among native speakers when they evaluate these two types of encoded information consciously. In other words, conceptual meaning is available to conscious thought. Consequently, judging conceptual information is a rather easy task, resulting in high inter-annotator agreement rates. Procedural meaning is more difficult to evaluate consciously than conceptual information. Consequently, procedural information is harder to judge than conceptual information, and it results in medium inter-annotator agreement rates.

4.2 Annotation Experiments with Tense and Its Description Using Reichenbachian Coordinates

4.2.1 Hypotheses and Predictions

The experiments presented in this chapter have three aims. The first is to assess whether comprehenders are able consciously to identify and categorize the configuration of Reichenbachian coordinates E, R and S and their interpretation at two levels. According to Reichenbach (1947) (cf. discussion in Sect. 1.2.1), the meanings of the target verbal tenses tested in this chapter should be described as in Table 4.1. In other words, the meaning of each verbal tense can be split into the three pairs of coordinates E/R, R/S and E/S. In this research, I make the assumption

[2] Here, the authors make reference to Sanders and Spooren's (2009) study, in which the meanings of two Dutch connectives were annotated: *omdat*, which is most frequently used in objective causal relations (that is, expressing 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 2013).

Table 4.1 The meaning of verbal tenses using E, R and S (following Reichenbach 1947)

Structure	English	French
E=R<S	Simple Past	Passé Simple
	He came.	*Il vint.*
E=R<S	Past Continuous	Imparfait
	When I saw, he was coming.	*Quand je l'ai vu, il venait.*
E<S=R	Present Perfect	Passé composé
	He has come.	*Il est venu.*
S=R=E	Simple Present	Préset
	He comes.	*Il vient.*
S<R=E	Simple Future	Future
	He will come.	*Il viendra.*

that the three pairs of coordinates do not act at the same level. The first level is the localization of eventualities E with respect to S. Two options are possible: $E < S$ (i.e. past); and $E \geq S$ (i.e. non-past). At this level, in English, the Simple Past and the Past Continuous both locate eventualities in the past, and therefore have the description $E < S$. The Simple Present and Future locate eventualities in the non-past, and therefore have the description $E \geq S$. As for French, the Passé Composé, Passé Simple and the Imparfait locate eventualities in the past, and therefore have the description $E < S$. As with English, the Présent and Future locate eventualities in the non-past, and therefore have the description $E \geq S$.

The second level is the localization of eventualities with respect to one another, making use of R. Two options are possible: the case of temporal progression from E_1 to E_2, thus $R_1 \rightarrow R_2$ (i.e. a narrative usage of the verbal tense corresponding to a sequential temporal relation); and the case of lack of temporal progression from E_1 to E_2, thus $R_1 = R_1$, or indeterminacy $E_1 ? E_2$ (i.e. a non-narrative usage of the verbal tense corresponding to simultaneous and undetermined temporal relations). This property has been operationalized as the [±narrativity] feature. In (448), the first three eventualities expressed by a Simple Past have a narrative usage, whereas the fourth and final is used non-narratively.

(448) John screamed [E_2]. His leg was broken [E_3]. Mary pushed him [E_1].
She felt betrayed [E_4].

The second aim is to test the existent theoretical assumptions about the link between verbal tenses and the temporal interpretation of the relations holding among eventualities. As discussed in Sects. 1.1.1, 1.1.3, 2.1 and 2.3.3, scholars have formulated a robust hypothesis according to which the Passé Simple instructs the comprehender to interpret sequentially the series of eventualities it expresses, the Imparfait is used when eventualities should be interpreted simultaneously, and the Passé Composé is undetermined with respect to this property. These assumptions are illustrated in example (449). The verbs *vint* 'came' and *monta* 'get in', expressed with the Passé Simple, have a narrative usage; the verb *s'asseyait* 'sit' has a non-narrative usage; and the verb *a regardé* 'looked at/has looked at', expressed with the Passé Composé, is undetermined with respect to this property.

(449) On raconte qu'un Anglais *vint* un jour à Genève avec l'intention de visiter le lac. Il *monta* dans l'une de ces vieilles voitures où l'on *s'asseyait* de côté comme dans les omnibus. Il *a regardé* le lac émerveillé.
It is said that un Englishman come.3SG.PS one day to Geneva with the intention visiting the lake. He get in.3SG.PS in one of these old cars where you sit.3SG.IMP along the sides as on a bus. He look.3SG.PC at the lake amazed.

The case of the Imparfait is slightly more complicated than it looks. French scholars have observed that the Imparfait may have two usages: non-narrative, and narrative. Its narrative usage, known as the narrative Imparfait ("imparfait de rupture"), is characterized by the presence of a subjectivity marker or a temporal adverbial or connective that encodes an immediate transition towards a resulting state. This information is inferential, and directs discourse computation towards temporal sequencing. Thus, both narrative and non-narrative occurrences of the Imparfait express reference to past time, and are viewed as continuous eventualities. The non-narrative Imparfait does not express temporal sequencing, and is not viewed as being completed, whereas the narrative Imparfait expresses temporal sequencing, and is viewed as being completed (the final boundary is expressed by a temporal adverbial, or the Imparfait is used with a punctual eventuality). The former is illustrated in example (450), and the latter in (451).

(450) Il y a une heure Max *boudait* dans son coin, et ça n'est pas près de changer.
An hour ago Max sulk.3SG.IMP in a corner, and this is not about to change.
'For an hour, Max has been sulking in a corner, and this is not about to change.'

(451) Elle a fini par fuguer à Kaboul, où elle a été recueillie par une femme généreuse. Quelques mois plus tard, elle *épousait* un jeune cousin de sa bienfaitrice dont elle était tombée amoureuse.
She finally run.3SG.PC to Kabul, where receive3SG.PC.PSV by a kind woman. A few months later, she marry.3SG.IMP a younger cousin of her benefactor with whom she fall in love.3SG.PQP.
'Finally she run to Kabul, where she was taken in by a kind woman. A few months later, she married a younger cousin of her benefactor with whom she had fallen in love.'

The third aim is to test whether the [±narrativity] is cross-linguistically valid, and whether it can be used to predict the verbal tense used in a target language. For example, the analysis of translation corpora by the translation spotting method, discussed in Sects. 3.2.2 and 3.4.2, has shown that the English Simple Past translation

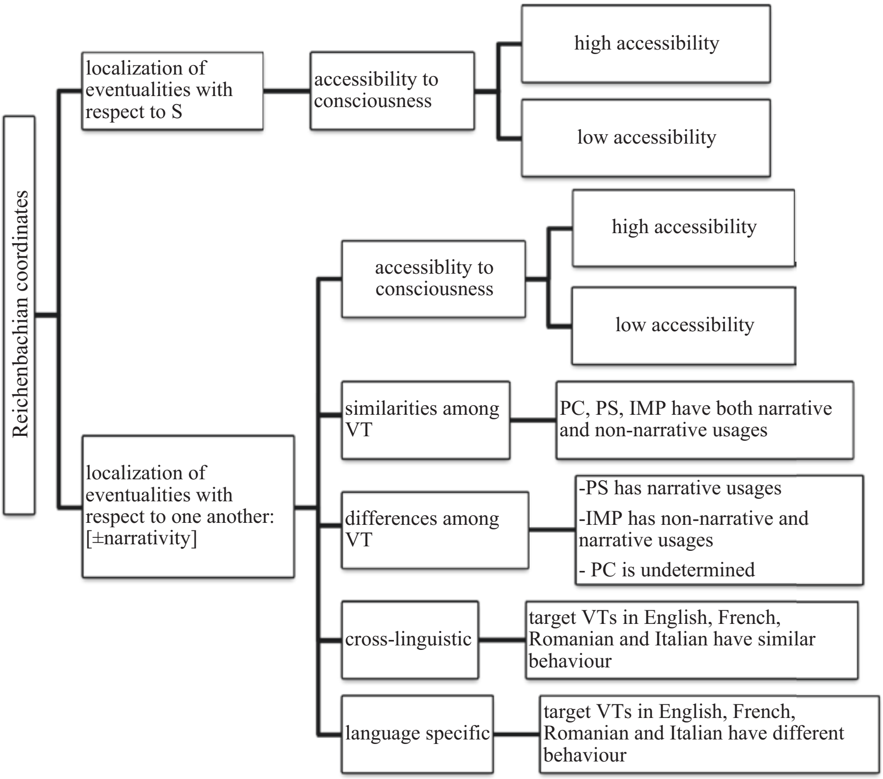

Reichenbachian coordinates
localization of eventualities with respect to S
accessibility to consciousness
high accessibility
low accessibility
localization of eventualities with respect to one another: [±narrativity]
accessiblity to consciousness
high accessibility
low accessibility
similarities among VT
PC, PS, IMP have both narrative and non-narrative usages
differences among VT
-PS has narrative usages
-IMP has non-narrative and narrative usages
- PC is undetermined
cross-linguistic
target VTs in English, French, Romanian and Italian have similar behaviour
language specific
target VTs in English, French, Romanian and Italian have different behaviour

Fig. 4.1 Possible scenarios and their predictions regarding the category of Tense and its encoding of Reichenbachian coordinates

paradigm consists of four verbal tenses in French, Italian and Romanian. Cross-linguistic analyses of the annotated corpora should show whether a correlation between the [±narrativity] features and the verbal tense used in a target language can be established.

Taking into account the semantic and pragmatic descriptions of the target verbal tenses tested in this chapter, two research questions can be formulated. The first is *how do comprehenders consciously deal with the encoded information from Reichenbachian coordinates and their possible configurations?* The second is *do the current theoretical studies of verbal tenses and their role in expressing temporal relations have empirical coverage?*

In order to answer these two research questions, a series of scenarios and their subsequent predictions was formulated. These scenarios are summarized in Fig. 4.1. Accessibility to consciousness is understood in terms of the ease with which participants consciously carry out the task in an accurate manner. With respect to encoded information, two types of degrees of accessibility are possible: (i) high accessibility, resulting in high κ values used to measure inter-annotator agreement; and (ii) low accessibility, resulting in low κ values. Based on the current theoretical descriptions of French verbal tenses carried out according to the procedural pragmatics approach (Nicolle 1997; de Saussure 2003; Escandell-Vidal and Leonetti 2011; Aménos-Pons 2011; cf. the discussion in Sect. 2.3.3), similar accessibility rates are expected for the localization of eventualities, with respect to S and to one another. This is broadly due to the fact that verbal tenses are considered to be procedural expressions, and their meaning is described using the three Reichenbachian coordinates. By contrast, if these two types of localizations have a different nature—in other words, take place at different levels of meaning—one would expect dissimilar degrees of accessibility to consciousness and, consequently, dissimilar inter-annotator agreement rates.

From a cross-linguistic perspective, if the [±narrativity] property is a cross-linguistically valid feature, then the target verbal tenses in the four languages studied in this research are comparable with respect to this feature. This means that we would expect to see strong correspondences between the narrative usages of these verbal tenses on the one hand, and their non-narrative usages on the other.

4.2.2 French Verbal Tenses and Reichenbachian Coordinates

Participants

Participants were 48 native speakers of French, Bachelor's students at the Faculty of Humanities of the University of Geneva and the University of Neuchâtel. Their participation in the experiment was voluntary and unpaid. They did not receive training before participating in the experiment.

Procedure and Material

The items used in this experiment were of two categories. The first category consists of 90 items randomly selected from the corpus (as described in Sect. 3.3), which represent *naturally occurring* items judged in their *original contexts*. The second category consists of 36 artificial sentences, built for the purpose of the experiment. Each item comprised a first sentence, to set the context, and a second sentence containing the targeted verb, as shown in examples (452), (453) and (454).

(452) 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. (Literature register)
'The young soldier who was accused behaved against his superior's orders, he (to be) imprisoned today and waiting to be judged for murder.'

(453) Marie a pris du poids. Avant de casser sa jambe, Marie (courir) tous les soirs pendant une heure. (Built example, the *past* condition)
'Mary gained weight. Before breaking her leg, Mary (to run) every evening for an hour.'

(454) Marie s'entraîne pour le marathon. Elle (courir) tous les soirs pendant une heure.
'Mary trains for the marathon. She (to run) every evening for an hour.'
(Built example, the *non-past* condition)

The role of the first sentence was to set a past or a non-past time context. All the experimental items were distributed into sets of 15 items (for the corpus data) and 18 items (for the artificial sentences), with a total of 8 sets. Each participant received either corpus (natural) or built (artificial) data. Each experimental item was judged by 6 participants.

Participants were asked to give the tensed form of a verb, provided by the infinitive, such that it corresponds to the surrounding context. They received annotation guidelines, in which the task of the experiment was explained, and had a training session on 3–5 items. Then, they received the set of items to annotate in an independent manner. Each participant received either corpus (natural) or built (artificial) data.

The results of this experiment were evaluated by counting the majority of answers for each item, since there were more than two participants. The number of concordant answers must exceed agreement by chance, which is 50%, given the binary choice (i.e. the past vs. non-past context). Where responses were equally distributed (3 out of 6 annotators), the item was evaluated as *inconclusive*. Inconclusive items were excluded from further analysis. Finally, for a given item, where under 50% of the judges (a maximum 2 out of 6 annotators) made the same judgment, it was considered to be a disagreement. Due to the reduced number of participants who saw each item—that is, 6 per item—the evaluation was made manually. Moreover, labels given by participants were compared to a baseline, established according to the translation corpus for the natural data, and defined by the experimenter for the artificial, built data.

Results

A total of 126 items were evaluated, according to the evaluation scheme described above. The judges agreed on their label for 119 items (94.4%), and disagreed on 3 items (2.4%). Four items were evaluated as inconclusive (3.2%). Disagreements and inconclusive items were excluded from further analysis. Table 4.2 provides the results of the comparison between the label provided by the annotators and the reference baseline (from the translation corpus) for all data. The correspondence between the judges' label and the reference of 111 items (93.3%) corresponds to a $\mathcal{K}$ of 0.86.

Regarding the two types of data (natural vs. artificial), all three disagreements and the four inconclusive items were natural data; annotators agreed on the label

Table 4.2 Annotators vs. Reference baseline for past/non-past distinction in all data

		Annotation		
		Past	Non-past	Total
Baseline	Past	57	2	59
	Non-past	6	54	60
Column Total		63	56	119

Table 4.3 Judges vs. Reference baseline for past/non-past distinction in natural data

		Annotation		
		Past	Non-past	Row total
Baseline	Past	39	2	41
	Non-past	6	36	42
Column Total		45	38	83

provided for all artificial items. When compared to the reference baseline, there is a one-to-one correspondence between the annotators' labels and the baseline. This corresponds to a K value of 1.

As for the natural data, the annotators agreed on 83 items (92%). Among the agreements, the items were judged as expressing reference to the non-past in 45 items (54.2%) and reference to the past in 38 items (45.8%). Table 4.3 provides the results of the comparison between the label provided by the annotators and the reference baseline for natural data only. The correspondence between the annotators' label and reference of 75 items (90%) corresponds to a K of 0.80.

Discussion

This experiment aimed to test whether speakers are able to categorize the configuration of two Reichenbachian coordinates (E with respect to S). The hypothesis defended in this research is that the relation between these two coordinates is of a conceptual nature, and the ad hoc concept is built contextually. According to the qualitative features proposed by Wilson and Sperber (1993) for conceptual and procedural information, it was argued that judging conceptual information results in high K values. This experiment provided evidence that the conceptual information encoded by verbal tenses—that is, past vs. non-past—is determined contextually, and that the agreement between the participants produced high K values: 1 for artificial data, 0.80 for natural data, and 0.86 for all the data.

With respect to natural vs. artificial data, the difference in results is that the natural data used in this research are much more complex and harder to understand than the artificial data built for the purposes of the experiment. This is partly due to the type of data, which originate in parliamentary discussions, legislation, journalistic and literature stylistic registers. The two types of data are exemplified in example (455) for the natural data, in which the baseline reference to past time was expressed by a Passé Simple, and example (456) for the artificial data, in which reference to past time was expressed by an Imparfait.

(455) De son côté, l'Eglise catholique avait organisé, en 1986, la Rencontre nationale ecclésiale cubaine (ENEC), qui - tout en rappelant que Cuba est une nation chrétienne - (prendre acte) de la société cubaine telle qu'elle était et non telle que l'Eglise l'aurait souhaitée. (Journalistic register)

'For its part, the Catholic church had organized, in 1986, the Cuban National Ecclesiastic Meeting, which – remember that Cuba is a Christian nation – (take cognizance of) Cuban society as it was and not as the Church would have wished it.'

(456) Après son accident, Marie était très triste. Elle ne pouvait plus faire ce qui la rendait si heureuse. Marie (jouer) du piano. (Built example)

‘After her accident, Mary was very sad. She could not do anymore what used to make her so happy. Mary (play) the piano.’

This experiment indicated that speakers have no difficulty consciously evaluating the localization of eventualities with respect to the moment of speech.

4.2.3 Passé Composé, Passé Simple, Imparfait and the [±Narrativity] Feature

Participants

Participants were 76 French native speakers, who were first year students at Faculty of Humanities from University of Geneva. Their participation in the experiment was organized during a linguistics class, but was unpaid and anonymous.

Procedure and Material

The materials used consisted of 300 items[3] randomly chosen from the French part of the parallel corpus, organized in 19 sets. Each participant received a set of 15 items. The data contained 127 occurrences of the Imparfait, described by the literature as most often non-narrative, 173 occurrences of the Passé Simple/Passé Composé (101 Passé Simple and 72 Passé Composé), described as most often narrative.

The annotation guidelines included two tasks. The first task was to read and understand the definitions of narrative and non-narrative usages, as follows:

- The eventualities are temporally linked. This means that E_1 happened before E_2. The relation may be explicitly expressed in the sentence, or may be implicit (it can be understood in the context).
- The eventualities are not temporally linked. This means that E_1 and E_2 either happened at the same time (simultaneously) or are not temporally linked (the opposite of the case above).

Each definition was accompanied by two explained examples, as given in (457), where the verbs *vint* ‘came’ and *monta* ‘get in’, expressed by the Passé Simple, have a narrative usage, the verb *s’asseyait* ‘sit’ has a non-narrative usage, and the verb *a regardé* ‘looked at/has looked at’, expressed by the Passé Composé, is undetermined with respect to this property.

[3] An item consists of a sentence where the verbal tense of interest occurs (for example, the Passé Simple, Passé Composé or Imparfait for Experiment 1) and another sentence, either preceding or following. This choice was made because of the need to have sufficient co(n)text for a pragmatic decision.

(457) On raconte qu'un Anglais *vint* un jour à Genève avec l'intention de visiter le lac. Il *monta* dans l'une de ces vieilles voitures où l'on *s'asseyait* de côté comme dans les omnibus. Il *a regardé* le lac émerveillé.
It is said that an Englishman come.PS one day to Geneva with the intention of visiting the lake. He get in.PS in one of these old cars when you sit.IMP on the sides as in a bus. He look.PC at the lake amazed.

Participants received training for 6 items, which was followed by a collective discussion. The evaluation was performed manually, according to the evaluation scheme which follows. The results were evaluated by counting the majority of answers for each item. The number of concordant answers must exceed agreement by chance, which is 50%, given the binary choice (i.e. narrative vs. non-narrative usage). When that was not the case, the item was evaluated as *inconclusive*. Inconclusive items were excluded from further analysis. Moreover, labels given by participants were compared to a baseline established according to theoretical descriptions of the verbal tenses considered.

Results

Table 4.4 provides the results of this annotation experiment, where 221 tokens of the Imparfait, Passé Composé and Passé Simple were considered. Of the 300 items annotated by four judges, 79 received showed no majority, and were thus inconclusive. These items were not considered in the analysis. In the clean data of 221 tokens, judges agreed with the theoretical reference for 182 items (82% of the data), with a κ value measuring inter-annotator agreement of 0.63.

The table shows that the narrative feature was identified for 86% of the annotated tokens according to the theoretical predictions (i.e. Passé Simple and Passé Composé together, 110 items labelled as narrative out of 128 in the corpus), and the non-narrative feature in 77% of cases (Imparfait, 72 items labelled as non-narrative out of 93 in the corpus). *A chi-square test performed on this result shows that the correlation between the annotator's judgment and the theoretical reference is statistically significant (*Chisq *86.96, df = 1, p < .0001).*

In particular, as shown in Table 4.5, judges clearly recognized a primary narrative usage for the Passé Simple (92%), but did not make the same clear judgment for the Passé Composé narrative (in 77% of cases) or the expected non-narrative primary usage of the Imparfait (77.5%).

Table 4.4 Narrativity for Passé Simple/Passé Composé and imparfait: majority of annotators and reference

		Majority of annotators		
		Narrative	Non-narrative	Total
Reference	Passé Simple/Passé Composé	110	18	128
	Imparfait	21	72	93
	Total	131	90	221

Table 4.5 Annotations for individual verbal tenses

Verbal tense/narrativity	Narrative	Non-narrative
Passé Simple	92%	8%
Passé Composé	77%	23%
Imparfait	22.5%	77.5%

Table 4.6 Narrativity for the Imparfait: Annotator 1 and Annotator 2

		Annotator 2		
		Narrative	Non-narrative	Total
Annotator 1	Narrative	17	35	52
	Non-narrative	19	159	178
Total		36	194	230

This results in about 23% of non-expected usages—that is, non-narrative usages—for the Passé Composé, and 22.5% of narrative usages for the Imparfait. This result opened the door to a further, finer-grained investigation: an annotation experiment of the Imparfait with the [±narrativity] feature.

4.2.4 *The Imparfait and the [±Narrativity] Feature*

Participants

The participants were 2 French native speakers, who were students at the Faculty of Humanities of the University of Geneva. They were paid for their participation in the experiment.

Procedure and Material

The material consisted of a total of 230 items containing Imparfait occurrences. 120 items were randomly selected from the French part of the parallel corpus, where French was the source language. 110 occurrences of the Imparfait were translations of Simple Past items into French, where French was the target language. The two annotators received annotation guidelines, consisting of the definition and examples for each type of usage. They received training for 6 items, which was followed by a group discussion. Evaluation was performed by calculating the inter-annotator agreement rate using the K coefficient.

Results

The results are presented in Table 4.6. Out of 230 annotated tokens, annotators agreed on the annotation of 179 tokens (77%), representing a K of 0.24. This very low K is explained by the fact that the two categories (narrative and non-narrative) are not equally distributed, and therefore the non-narrative category is the default case. The judges were not aware that there is a default case, and they assigned the categories by judging the sentences according to the annotation guidelines. If the analysis only considers the 179 cases of agreement, the Imparfait was categorized in 90% of cases as non-narrative, and in 10% of cases as narrative.

The annotation results have also been analysed according to the original language. For the 120 Imparfait tokens where French was the source language, judges agreed on 90 items (75%). In the cases of agreement, the Imparfait was labelled as non-narrative in 84% of cases, and narrative in 16%. As for the 110 Imparfait tokens where English was the source language, judges agreed on 86 items (78% of cases). In the cases of agreement, the Imparfait was labelled as non-narrative in 97% of cases, and narrative in 3%. The results of this experiment show that categorization of the Imparfait, in terms of narrative and non-narrative usages, presents different patterns regarding the source language. However, using Fisher's Exact Probability test, the difference in categorization between the two source languages is not shown to be statistically significant ($p > .05$).

4.2.5 Passato Prossimo, Passato Remoto, Imperfetto and the [±Narrativity] Feature

Participants
There were two participants, both Italian native speakers originating from the southern part of Italy (Naples). Their participation in the experiment was voluntary and unpaid.

Procedure and Material
84 items, containing 37 Passato Prossimo, 27 Passato Remoto and 21 Imperfetto, were randomly chosen from the Italian part of the multilingual translation corpus. These items were originally written in English, and the targeted Italian verbal tense corresponds to a Simple Past in the source language. Annotators received annotation guidelines and received a training session. The first task in the annotation guidelines was to read and understand the instructions, containing definitions of narrative and non-narrative usages. They also included two examples for each usage, as given in (458)–(460), where (458) is an example of non-narrative usage, whereas (459) and (460) are examples of narrative usage.

(458) V'erano porte tutt'intorno alla sala, ma *erano* [Imperfetto] tutte serrate.
(Literature Corpus)
'There were doors all around the hall, but they *were* all locked.'

(459) Ma, risalito dopo pranzo con tale proposito, appena varcata la soglia,
scorsi [Passato Remoto] lì dentro una ragazza che, inginocchiata
davanti al fuoco e circondata da scope e secchi di carbone.
(Literature Corpus)
'On coming up from dinner, however, and mounting the stairs with
this lazy intention, and stepping into the room, I saw a servant-girl
on her knees surrounded by brushes and coal-scuttles'.

(460) Malgrado le misure di controllo adottate dalle autorità delle isole Faroe, nel 2004 *sono stati segnalati* [Passato Prossimo] alla Commissione nuovi focolai della malattia. (EuroParl Corpus)
‘Despite the control measures undertaken by the Faroe Islands, further outbreaks of ISA occurred and were notified by that State to the Commission in 2004.’

The second task was to read each item and decide if the highlighted verb had a narrative or a non-narrative usage. Participants received training for 6 items, which was followed by a discussion.

Results

Annotators agreed on 64 items (76%), and disagreed on 21 items (33%). The value of the κ coefficient was 0.41. The disagreements were discussed in the second round of the experiment. The final results are provided in Table 4.7. Judges agreed on 76 items (89%), which represents a κ value of 0.74.

As far as the analysis of individual verbal tenses is concerned, only the data containing agreements were considered (76 items). 16 Imperfetto were judged to be non-narrative (84%), 30 Passato Prossimo were judged to be narrative (88%), and 22 Passato Remoto were judged to be narrative (96%)(Table 4.8).

The results of this experiment indicate that the [±narrativity] feature is identifiable by native speakers, with reliable κ values. Regarding this information, most often narrative values are attributed to the Passato Remoto and the Passato Prossimo, and non-narrative values to the Imperfetto. Like English and French speakers, Italian speakers have little ability to evaluate the temporal relations triggered by verbal tenses consciously. They do better when asked to insert connectives, which explicitly express the same implicit content. These findings provide a solid empirical basis to argue that the [±narrativity] feature is procedural, and that it is a cross-linguistically valid feature.

Table 4.7 Narrativity for Italian verbal tenses: Annotator 1 vs. Annotator 2

		Annotator 2		
		Narrative	Non-narrative	Total
Annotator 1	Narrative	55	4	59
	Non-narrative	5	21	26
Total		60	25	85

Table 4.8 Narrativity for Passato Remoto, Passato Prossimo and Imperfetto

	Narrative	Non-narrative	Total
Imperfetto	3	16	19
Passato Prossimo	30	4	34
Passato Remoto	22	1	23
Total	55	21	76

4.2.6 Perfectul Compus, Perfectul Simplu, Imperfectul and the [±Narrativity] Feature

Participants and Material
There were two participants, both Romanian native speakers. One of the judges is a research peer, and the other is a Bachelor's student from University of Geneva, Faculty of Humanities. Their participation in the experiment was unpaid.

Procedure
85 items, containing 50 Perfectul Compus, 14 Perfectul Simplu and 21 Imperfectul, were randomly chosen from the Romanian part of the multilingual translation corpus. These items were originally written in English, and the targeted Romanian verbal tense corresponds to a Simple Past in the source language. Annotators received annotation guidelines and received a training session. The first task in the annotation guidelines was to read and understand the instructions, containing definitions of narrative and non-narrative usages. They also included two examples for each usage, as given in (461)–(463), where (461) is an example of non-narrative usage and (462) and (463) are examples of narrative usage.

(461) Erau uşi de jur împrejurul holului dar toate *erau* [Imperfectul] încuiate. (Literature Corpus)
'There were doors all around the hall, but they were all locked.'

(462) Aşa că, întorcându-mă de la masă, urcai scările cu intenţia de a-mi petrece după-amiaza lenevind. Când să intru în odaia mea, *văzui* [Perfectul Simplu] o tânără servitoare, îngenuncheată lângă sobă, înconjurată de perii şi găleţi cu cărbuni. (Literature Corpus)
'On coming up from dinner, however, and mounting the stairs with this lazy intention, and stepping into the room, I saw a servant-girl on her knees surrounded by brushes and coal-scuttles'.

(463) Cu toate că autorităţile din insulele Feroe au pus în aplicare măsuri de combatere au apărut alte focare de AIS, care *au fost notificate* [Perfectul Compus] Comisiei de această ţară în 2004. (EuroParl Corpus)
'Despite the control measures undertaken by the Faroe Islands, further outbreaks of ISA occurred and were notified by that State to the Commission in 2004.'

The second task was to read each item and decide if the highlighted verb had a narrative or a non-narrative usage. Participants received training for 6 items, which was followed by a discussion.

Results
The results are provided in Table 4.9. Judges agreed on 64 items (75%), and disagreed on 21 items (25%). The value of K coefficient was 0.42.[4]

[4] This experiment was carried out in two rounds. 42 items were judged in the first round, and 43 items in the second. Due to the two judges' unfortunate lack of availability, only the data from the

Table 4.9 Narrativity for Romanian verbal tenses: Annotator 1 vs. Annotator 2

		Annotator 2		Total
		Narrative	Non-narrative	
Annotator 1	Narrative	47	0	47
	Non-narrative	21	17	38
Total		68	17	85

Table 4.10 Narrativity for Perfectul Simplu, Perfectul Compus and Imperfectul

	Narrative	Non-narrative	Total
Imperfectul	4	10	14
Perfectul Compus	30	6	36
Perfectul Simplu	13	1	14
Total	47	17	64

As far as the analysis of individual verbal tenses is concerned, only the data containing agreements were considered (64 items). The Imperfectul was judged to be non-narrative in 10 cases (71%), the Perfectul Compus was judged to be narrative in 30 cases (83%), and the Perfectul Simplu was judged to be narrative in 13 cases (93%) (Table 4.10).

As with Italian, this experiment shows that the [±narrativity] feature is identifiable by Romanian native speakers with reliable K values. Regarding this information, most often narrative values are attributed to the Perfectul Simplu and the Perfectul Compus, and non-narrative values to the Imperfectul. Moreover, native Romanian speakers have little ability to evaluate temporal relations triggered by verbal tenses consciously. They do better when asked to insert connectives, which explicitly express the same implicit content.

4.2.7 The Simple Past and the [±Narrativity] Feature

Participants

There were two participants, both English native speakers from the United Kingdom, who were studying Bachelor's level linguistics at the Faculty of Humanities of the University of Geneva. Their participation in the experiment was paid.

Procedure and Material

The material used consisted of 458 Simple Past tokens randomly chosen from the English part of the parallel corpus. As in the first two experiments, judges received

first round were judged a second time, to resolve the disagreements. For the first 42 items, the K value improved from 0.23 (agreement in 62% of cases) to 0.75 (agreement in 88% of cases). The results provided in Table 4.9 represent the data obtained after the second round, with the first 42 items, and the sole round, with the other 43 items. The low K value of the entire data is due to the fact that disagreements on the 43 items were not resolved.

annotation guidelines and received a training session. The first task from the annotation guidelines was to read and understand the instructions containing definitions of narrative and non-narrative usages. They also included two examples for each usage, as given in examples (464) and (465). The second task was to read each item and decide if the highlighted verb had a narrative or non-narrative usage. Participants received training on 10 items, which was followed by a discussion where each annotator had to "think aloud" his/her answers.

In the first example below, there are two events: 'the marriage that happened', and 'the wealth which was added'. The second event is presented in relation to the first (first he got married, and then he added to his wealth), which is why the Simple Past verbs happened and added are in narrative usage. In the second example, there are three states (was a single man, lived and had a companion) that describe the owner of the estate. States are not temporally ordered, which is why this example illustrates the non-narrative usage of the Simple Past.

(464) By his own marriage, likewise, which happened soon afterwards, he *added* to his wealth. (Literature Corpus)

(465) 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. (Literature Corpus)

Evaluation of inter-annotator agreement rate was performed with the $\mathcal{K}$ coefficient. In terms of cross-linguistic evaluation, the judged items were compared to a reference baseline containing the verbal tenses used for the translation of the Simple Past into French, from the French part of the parallel corpus.

Results

The results are provided in Table 4.11. Annotators agreed on 325 items (71%) and disagreed on 133 items (29%). The value of $\mathcal{K}$ coefficient was 0.42. This value is higher than chance, but not high enough to point to entirely reliable linguistic decisions. Of the 113 items of disagreement, 19 items were signalled as having insufficient context for a pragmatic decision. They were excluded from further analysis.

Error analysis showed that the main source of errors was the length of the temporal interval between two eventualities, which was perceived differently by the two annotators. This led to ambiguity between temporal sequence or simultaneity, each of them corresponding to narrative and non-narrative usage respectively, as in example (466), where the eventualities "qualify" and "enable" were perceived as simultaneous by one judge but successive by the other.

Table 4.11 Narrativity for Simple Past: Annotator 1 and Annotator 2

		Annotator 2		
		Narrative	Non-narrative	Total
Annotator 1	Narrative	180	83	263
	Non-narrative	50	145	195
Total		230	228	458

(466) Elinor, this eldest daughter, whose advice was so effectual, possessed a strength of understanding, and coolness of judgment, which qualified her, though only nineteen, to be the counsellor of her mother,
and *enabled* her frequently to counteract, to the advantage of them all, that eagerness of mind in Mrs. Dashwood which must generally have led to imprudence. (Literature Corpus)

A possible explanation is the fact that personal world knowledge is used to infer temporal information, such as the length of the temporal interval between two eventualities—i.e. information which allows the judge to decide whether or not the eventualities are temporally ordered. Cases where the length of the temporal interval between two eventualities was greatly reduced were ambiguous for the judges, so each of them decided differently whether it was long enough for temporal sequencing or too short, in which case the simultaneity meaning was preferred.

Disagreements (114 items) were resolved in a second round of the annotation experiment, where the narrativity feature was identified with a new linguistic test that was explained to two new participants[5] (as suggested by Spooren and Degand 2010). Judges were asked to insert a connective, such as *and* and *and then* when possible, in order to make explicit the 'meaning' of the excerpt—that is, the temporal relation existent between the two eventualities considered. The connective *because* (for a causal relation) has also been proposed by annotators under the [+narrative] label, showing that causal relations should also be considered. The inter-annotator agreement rate in this second round of the experiment was corresponds to a K of 0.91, signalling very strong and reliable agreement.

In the data containing agreements, the Simple Past was judged as having narrative usages in 59% of cases and non-narrative usages in 41% of cases. This finding suggests that the Simple Past is not specialized for either of the possible values of the [±narrativity] feature. The cross-linguistic application of these findings consists of the observation of a pattern in the parallel corpus. The data containing agreements from both annotation rounds (435 items) were investigated and analyzed in relation to the reference baseline, defined according to the parallel corpus. The two alternative hypotheses are:

- The non-narrative Simple Past is more often translated with an imperfective.
- The narrative Simple Past is more often translated with a simple past or a compound past.

The results are provided in Table 4.12. They show that the narrative usages of the Simple Past correspond to narrative usages in the French part of the corpus (translation by a Passé Composé or Passé Simple) and the non-narrative usages of the Simple Past correspond to the non-narrative usages in the French text (translation with an Imparfait) in 338 items (78%). Using a chi-square significance test, this cor-

[5] The new participants are the author and a research peer, who was not aware of the purpose of the research. They are fluent in spoken and written English, and use it as professional language.

Table 4.12 Narrativity for the Simple Past: Annotators vs. Baseline

		Baseline		
		PC/PS	IMP	Total
Annotators	Narrative	208	49	257
	Non-narrative	48	130	178
Total		256	179	435

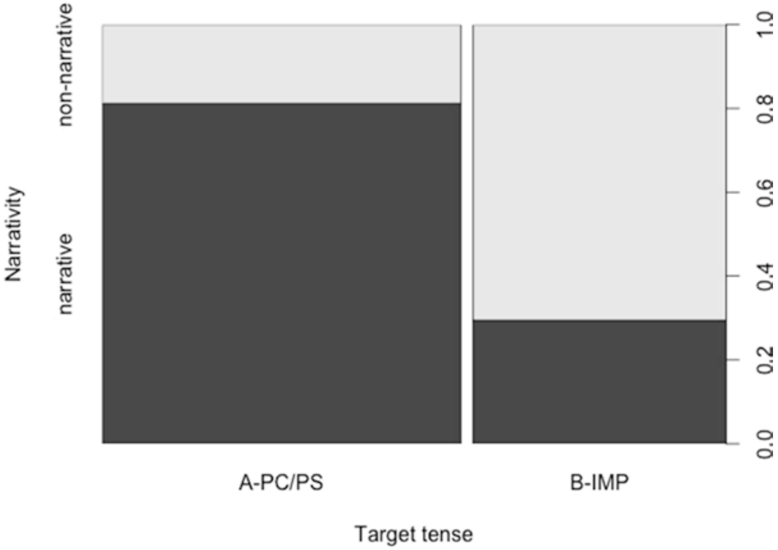


Fig. 4.2 Correlation between narrativity and target tense

respondence is shown to be statistically significant (Chisq 124.26, df = 1, $p < .001$). This correlation, shown in Fig. 4.2, is intermediately strong, having a Phi-coefficient of 0.52. The remaining 22%—for which annotators agreed on the narrativity label but which are not consistent with the verbal tense used in French—point to narrative usages of the Imparfait and to non-narrative usages of the Passé Composé.

The association plot in Fig. 4.3 shows the contribution to the overall significative chi-square of every cell (levels of the dependent and independent variable). In this plot, the area of the box is proportional to the difference in observed and expected frequencies. The black rectangles above the dashed line, indicating observed frequencies exceeding expected frequencies, correspond to narrative usage of the Simple Past positively correlated with the Passé Composé/Passé Simple value of the Target tense dependent variable, and to the non-narrative usage of the Simple Past positively correlated with the Imparfait value of the dependent variable. The grey rectangles below the dashed line, indicating observed frequencies smaller than expected frequencies, correspond to the lack of correlation between non-narrative

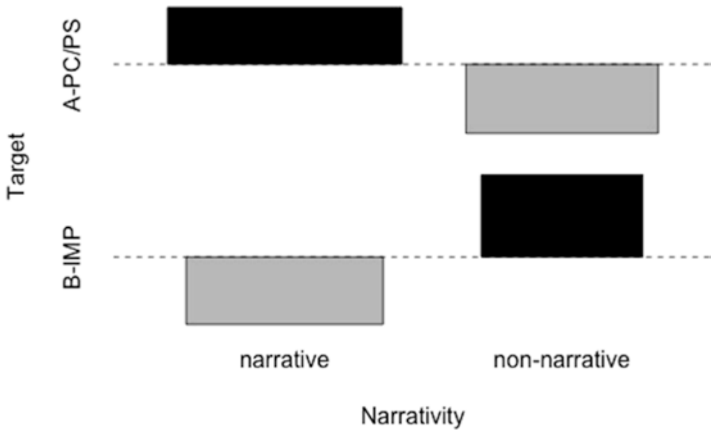


Fig. 4.3 Association plot for narrativity and target tense: residuals

usage of the Simple Past and the Passé Composé/Passé Simple, and narrative usage of the Simple Past viewpoint and the Imparfait value of the dependent variable.

The experiment described in this section showed that native speakers of English have little ability to consciously evaluate temporal interpretations triggered by Tense, operationalized as the [±narrativity] feature. The difficulty in consciously evaluating this type of information provides strong empirical evidence for the procedural nature of this feature, which is described as not easily accessible to consciousness. When speakers do not have conscious access to the instructions encoded by linguistic items, this information can be uncovered by other means. Participants were asked to propose a connective that would render explicit the implicit temporal relation (such as *and then*) or the implicit lack of temporal relation (such as *and at the same time*). The results showed that explicitating the implicit relation is an easier task for speakers than consciously evaluating these temporal relations. This represents strong empirical evidence for the procedural nature of this feature.

The results of experiments from Sects. 4.2.3 and 4.2.4 indicated that, for both French and English verbal tenses, the narrativity feature is identifiable after the second phase, when the judges inserted temporal connectives in order to render explicit the implicit temporal relation existing between the eventualities expressed. From a cross-linguistic perspective, the narrative usage of the Simple Past is translated with Passé Composé or Passé Simple (which themselves have a narrative usage), while an Imparfait is used to translate the non-narrative usage of the Simple Past. Moreover, when investigated in translation corpora, narrative usages of the Simple Past also point to narrative usages of the Imparfait (known as the *historical/breaking/narrative* Imparfait). These findings confirm the scenario according to which the [± narrativity] feature is procedural, and that it is a cross-linguistically valid feature.

The experiments presented in this section have shown two systematic patterns. When participants deal with the localization of eventualities with respect to S—that is, in the past or non-past (present or future)—they point out the ease of the task, and have high rates of inter-annotator agreement. When they deal with the localization of one eventuality in respect to another, they express the greater difficulty of the task, and have lower rates of inter-annotator agreement. These patterns are interpreted in terms of the different cognitive costs required to accomplish these tasks: a reduced cost for the first; and a higher cost for the second. I argue that this observed difference may be explained in terms of the different content which the comprehender is dealing with: conceptual for the former, and procedural for the latter. The results of these experiments support the interpretation according to which the category Tense encodes conceptual information, which refers to the localization of an eventuality with respect to S, as well as procedural information, which refers to the localization of an eventuality with respect to another eventuality (the phenomenon classically treated as temporal sequencing). These two localizations are contextually determined.

4.3 Annotation Experiments with Aspect and Aktionsart

4.3.1 Hypotheses and Predictions

This section has two aims. The first aim is to assess whether comprehenders are able consciously to identify and categorize the inherent aspectual properties of verbal phrases, and the completion/entirety vs. ongoing status of an eventuality. The former property was operationalized as the [±boundedness] feature, and the latter as the [±perfectivity] feature.

Eventualities are theoretically distinguished between *bounded* (generally, achievements and accomplishments) and *unbounded* (generally, states and activities). Dowty (1986) suggested the link between eventuality type, temporal progression and verbal tense. He argued that bounded eventualities trigger temporal progression, as in examples (467) and (468), whereas unbounded eventualities express lack of temporal progression, as in examples (469) and (470).

(467) John entered the president’s office. The president walked toward him.
(468) John entered the president’s office. The president stood up.
(469) John entered the president’s office. The president sat behind a huge desk.
(470) John entered the president’s office. The clock on the wall ticked loudly.

Eventualities can be presented with a perfective or an imperfective point of view. The imperfective aspect restraints temporal progression, by presenting the situation as ongoing, or by setting a focus on an internal phase, as in (471). The perfective aspect favours temporal expression by presenting the situation as a completed whole (Comrie 1976; Dowty 1986) as in Sect. 4.3.3.

(471) John entered the president's office. The president was writing a letter.
(472) John entered the president's office. The president wrote a letter.

The second aim is to investigate the relation between the type of eventuality and the verbal tense used in a target language, as well as the relation between the speaker's viewpoint of that eventuality and the verbal tense used in a target language. From a bilingual perspective, Kozlowska (1998b) argued that there is temporal progression in French with bounded eventualities expressed with the Passé Simple, as in (473) and (474), but no temporal progression with unbounded eventualities expressed with the Imparfait, as in (475) and (476), where examples (473)–(476) are the French translation of examples (467)–(470).

(473) Jean entra dans le bureau du président. Le président *s'avança* vers lui.
(474) Jean entra dans le bureau du président. Le président *se leva*.
(475) Jean entra dans le bureau du président. Le président *était* assis derrière un énorme bureau.
(476) Jean entra dans le bureau du président. L'horloge murale *marchait* bruyamment.

From a bilingual perspective, the French Passé Simple and Passé Composé are described as expressing the perfective aspect, whereas the Imparfait is associated with the imperfective aspect in its non-narrative usages. However, the Imparfait has also narrative usages that present the situation as a completed whole (like the perfective aspect), in particular in its narrative usages. Narrative and non-narrative usages of the Imparfait were confirmed in the annotation experiment described in Sect. 4.2.4.

Taking into account these semantic and pragmatic correspondences which scholars have proposed to hold between both of these aspectual properties of eventualities, expressed with the Simple Past in English and the verbal tense used in French, two research questions can be formulated. The first is *are comprehenders able consciously to identify the boundedness and perfectivity status of eventualities?* The second is *can these pieces of aspectual information be used to predict the verbal tense used in French as the target language?*

In order to answer these two research questions, a series of scenarios and their subsequent predictions can be formulated, as given in Fig. 4.4. As with localization of eventualities with respect to S and to one another (Sect. 4.2), accessibility to consciousness points to the participants' ability consciously to carry out the task in an accurate manner. Hence, the degree of accessibility to consciousness of [±boundedness] and [±perfectivity] will be inferred from inter-annotator agreement rates. Previous studies have suggested that Aktionsart and Aspect differ with respect to their nature of encoding: conceptual for the former, and procedural for the latter. High rates of inter-annotator agreement, signalling high accessibility, are expected for the former, and low rates of inter-annotator agreement, signalling low accessibility, are expected for the latter.

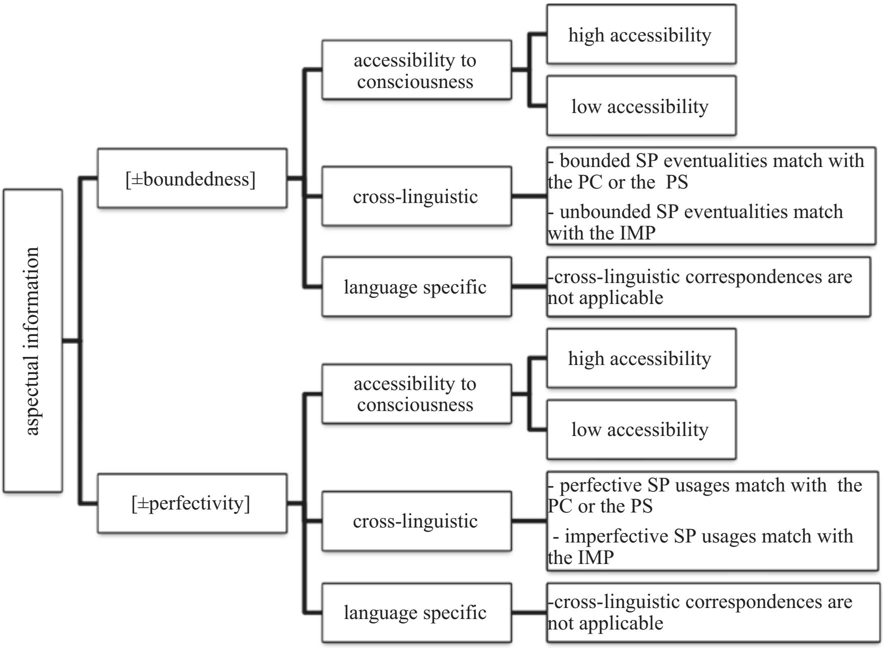

aspectual information
[±boundedness]
accessibility to consciousness
high accessibility
low accessibility
cross-linguistic
- bounded SP eventualities match with the PC or the PS
- unbounded SP eventualities match with the IMP
language specific
-cross-linguistic correspondences are not applicable
[±perfectivity]
accessibility to consciousness
high accessibility
low accessibility
cross-linguistic
- perfective SP usages match with the PC or the PS
- imperfective SP usages match with the IMP
language specific
-cross-linguistic correspondences are not applicable

Fig. 4.4 Possible scenarios and their predictions regarding aspectual information from verbal tenses

From a cross-linguistic perspective, robust cross-linguistic correspondences are expected for each feature. This means that the data will indicate that, in the French data, frequently bounded Simple Past eventualities match with the Passé Composé or Passé Simple, and unbounded Simple Past eventualities match with the Imparfait. Similarly, in the French data, perfective Simple Past usages match with the Passé Composé or Passé Simple, and imperfective Simple Past usages match with the Imparfait.

4.3.2 The Simple Past and the [±Boundedness] Feature

Participants
A previous pilot experiment with the same feature showed that judging lexical aspect required a certain level of theoretical knowledge, and that training did not manage to improve their results. In order to have reliable data annotated with the [± boundedness] feature, two research peers were asked to participate in this experiment. They were not native speakers, but were fluent in spoken and written English,

Table 4.13 Linguistic tests for the [±boundedness] feature

Test	Bounded eventualities	Unbounded eventualities
in/for adverbials	*in* adverbials	*for* adverbials
Homogeneity	–	+
Entailment with progressive	–	+

and used English as professional language.[6] They were not paid for their participation in the experiment.

Procedure and Material

The material used is the clean data resulting from the experiment presented in Sect. 4.2.7: that is, 435 items containing Simple Past tokens. Participants received annotation guidelines, consisting of the definition of the bounded and unbounded eventualities, their descriptions according to their behaviour in the three linguistic tests provided in Table 4.13, as well as two examples for each category. Bounded situations are situations which have attained their natural endpoint, as in example (477), where the running of the one-mile race is finished. The same true of situations which do not have a natural endpoint, but which are viewed as finished, as in example (478). Unbounded situations are situations which have not attained their natural endpoint, as in example (479), where the running of the one-mile race is not finished. The same is true of situations like example (480), where living in Paris does not have a natural endpoint.

(477) Max ran the one-mile race.
(478) I have lived in Paris from June to December 1998.
(479) Max is running the one-mile race.
(480) I have lived in Paris.

Evaluation of inter-annotator agreement rate was performed with the κ coefficient. In terms of cross-linguistic evaluation, the labelled items were compared to a reference baseline, containing the tenses used for the translation of the Simple Past into French, from the French part of the parallel corpus.

Results

The results are provided in Table 4.14. Judges agreed on the label for 401 items (92%) and disagreed on 34 items (8%). The agreement rate corresponds to a κ value of 0.84. All 34 disagreements were resolved in the experiment's second phase, consisting of a discussion between the two judges, corresponding to a κ value of 1. The κ values of both phases of annotation indicate that the judges understood the annotation guidelines and that their judgments were reliable. The data contains 236 Simple Past tokens, judged to be bounded, and 199 judged to be unbounded: that is, 54% and 46% respectively.

[6] For more accurate results, this experiment could be carried out with native speakers in further research.

Table 4.14 Boundedness for Simple Past: Annotator 1 and Annotator 2

		Annotator 2		Total
		Bounded	Unbounded	
Annotator 1	Bounded	210	8	218
	Unbounded	26	191	217
Total		236	199	435

Table 4.15 Boundedness for the Simple Past: annotators and reference

		Annotators		Total
		Bounded	Unbounded	
Reference	Passé Composé/Passé Simple	208	28	236
	Imparfait	47	152	199
Total		255	180	435

From a cross-linguistic perspective, the data containing agreements from both annotation rounds (435 items) were investigated and analysed in relation to the reference translation, defined according to the parallel corpus. The results are provided in Table 4.15. They show that bounded eventualities expressed with a Simple Past correspond to translation by a Passé Composé or Passé Simple, and unbounded eventualities expressed with a Simple Past correspond to translation by an Imparfait, for 360 items (82%). Using a chi-square test, this correspondence is shown to be statistically significant (Chisq 182.62, df = 1, $p < .001$). This correlation, shown in Fig. 4.5, is intermediately strong having a Phi-coefficient of 0.661.

The association plot in Fig. 4.6 shows the contribution to the overall significative chi-square of every cell. The black rectangles above the dashed line, indicating observed frequencies exceeding expected frequencies, correspond to the bounded type of situations positively correlated with the Passé Composé/Passé Simple value of the Target tense dependent variable, and to the unbounded type positively correlated with the Imparfait value of the dependent variable. The grey rectangles below the dashed line, indicating observed frequencies smaller than expected frequencies, correspond to the lack of correlation between unbounded situations and the Passé Composé/Passé Simple, and between bounded situations and the Imparfait.

To sum up, this experiment showed that the Simple Past is compatible with both bounded and unbounded eventualities, and that this is observable in natural data. In this experiment, the two judges had a very high agreement rate. According to Wilson and Sperber (1993) description of the cognitive foundations of the conceptual/procedural distinction, the information dealt with in this experiment is conceptual. From a cross-linguistic point of view, unbounded situations are most frequently correlated with an imperfective form, whereas bounded situations correlate with a simple past/compound past form in the target language. This correlation is statistically significative. Therefore, one could expect that the [± boundedness] feature is a

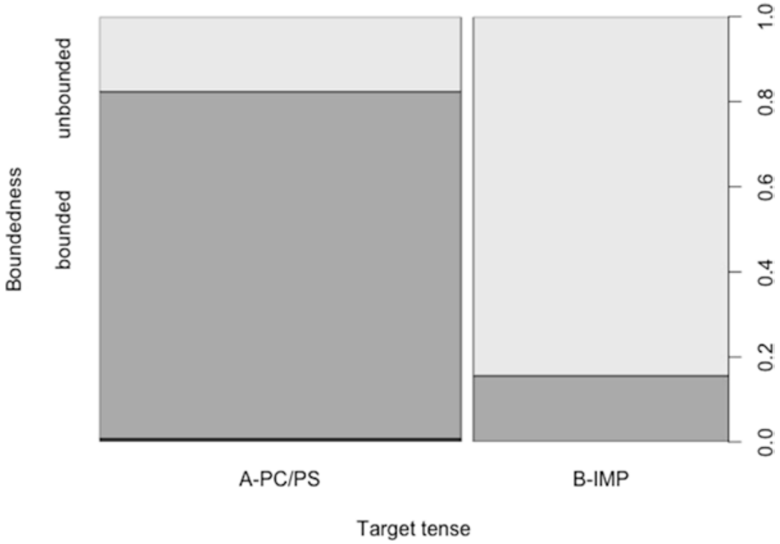
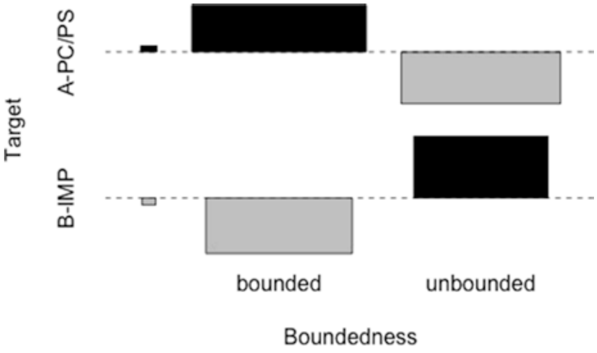


Fig. 4.5 Correlation between boundedness and target tense

Fig. 4.6 Association plot for boundedness and target tense: residuals



significant factor in predicting the verbal tense used in the target language. This will be investigated in a multifactorial analysis (see Sect. 4.4).

4.3.3 The Simple Past and the [±Perfectivity] Feature

The [±perfectivity] feature was assessed in two ways. The first was to carry out an annotation experiment, in which participants were asked consciously to identify the perfective and imperfective usages of the English Simple Past. The second was to make use of the translation of the English data into Serbian in order to identify the two aspectual categories in Serbian and totransfer to the English initial source data.

Participants

The participants in the annotation experiment were two English native speakers from the United Kingdom. They were the same participants from Experiment 3, in which Simple Past tokens were annotated with the [±narrativity] feature. Their participation in the experiment was paid.

Procedure and Material

The material used consisted of 62 items containing Simple Past tokens, chosen randomly from the data annotated in Experiment 3—more specifically, from the 22% of cases where the judges' label did not correspond to the verbal tenses used in the target language in the translation corpus. The participants received annotation guidelines, consisting of the definition of the perfective and imperfective viewpoints, as well as two examples for each category. Perfective situations are viewed as finished, and the situation as a completed whole, as in example (481), where the letter was finished when John entered the president's office. Imperfective situations are viewed as being in progress, and the situation is not completed, as in example (482), where the letter was not finished when John entered the president's office.

(481) John entered the president's office. The president *wrote* a letter.
(482) John entered the president's office. The president *was writing* a letter

A training session was carried out using 13 items, followed by a collective discussion, where each judge had to 'think aloud' his/her decisions.

Results

The two judges agreed on the label for 41 items (66%), and disagreed on 21 items (33%). The agreement rate corresponds to a *Ķ* value of 0.32. Disagreements were not resolved after the discussion between the two judges. The results of this experiment show that the data annotated with the [±perfectivity] feature is not reliable. In order to have reliable data annotated with this feature, another method was used.

Translation and Cross-Linguistic Transfer of Properties

A native speaker translated the data, consisting of 435 items containing Simple Past tokens, into Serbian. The translator was a linguistics student from the University of Geneva, and a native speaker of Serbian. Participation in the experiment was paid. Grammatical aspect was identified in Serbian for each item, and transferred to the initial English source according to the cross-linguistic transfer of properties method. The Simple Past was labelled as perfective for 204 items (47%), and as imperfective[7] for 231items (53%).

[7] For seven items, the translator was free to choose between perfective and imperfective, both aspects being possible. The verbs which occurred in these sentences are *to promise, to spend, to reproach, to organize, to despise, to stay* and *to try.* All these verbs express atelic situations.

Table 4.16 Perfectivity for the Simple Past: annotation by translation and baseline

		Annotation through translation		
		Perfective	Imperfective	Row total
Baseline	PC/PS	144	36	180
	IMP	60	195	255
Column Total		204	231	435

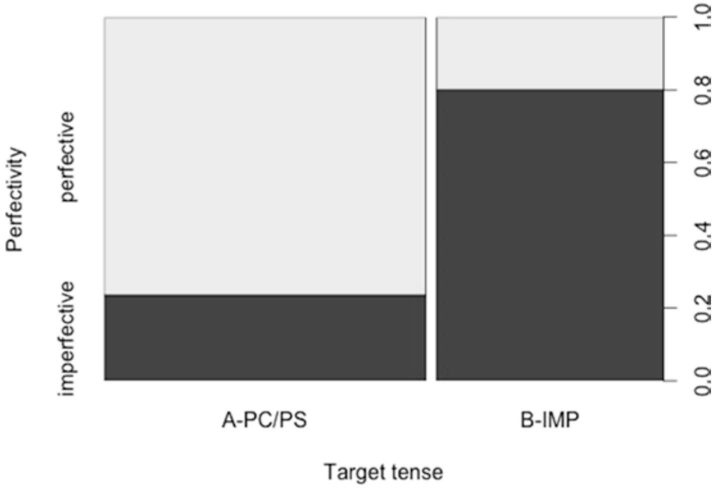


Fig. 4.7 Correlation between perfectivity and target tense

Table 4.16 presents the results of the contrastive analysis between the value of Aspect and the verbal tense used in French. It shows that the perfective viewpoints expressed with a Simple Past correspond to a translation by a Passé Composé/Passé Simple, and imperfective viewpoints expressed with a Simple Past correspond to a translation by an Imparfait for 339 items (78%). Using a chi-square test for independence, this correspondence is shown to be statistically significant (Chisq 132.86, df = 1, $p < .0001$). This correlation, shown in Fig. 4.7, is intermediately strong having a Phi-coefficient of 0.557.

The association plot in Fig. 4.8 shows the contribution to the overall significative chi-square of every cell. The black rectangles above the dashed line, indicating observed frequencies exceeding expected frequencies, correspond to the perfective viewpoint positively correlated with the Passé Composé/Passé Simple, and to the imperfective viewpoint positively correlated with the Imparfait. The grey rectangles below the dashed line, indicating observed frequencies smaller than expected frequencies, correspond to the lack of correlation between the imperfective viewpoint and the Passé Composé/Passé Simple, and between the perfective viewpoint and the Imparfait.

Firstly, the experiment described in this chapter has shown that native speakers of English have little ability consciously to evaluate the meaning of Aspect, opera-

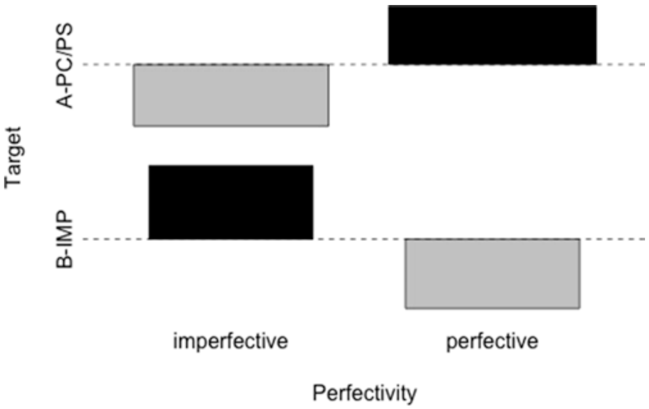


Fig. 4.8 Association plot for perfectivity and target tense: residuals

tionalized in this research as the [±perfectivity] feature. The difficulty of consciously evaluating the type of viewpoint provides strong empirical evidence for the procedural nature of this feature, which is described as not easily accessible to consciousness. When speakers do not have conscious access to the instructions encoded by linguistic items, this information can be found elsewhere. Based on parallel corpora, the novel cross-linguistic transfer of properties technique was used in order to reveal procedural information for English verbs, which is expressed morphologically in Slavic languages.

Secondly, translation data annotated with the [±perfectivity] feature were analysed cross-linguistically. The results pointed to the strong correlation between perfective usages of the Simple Past and the Passé Composé/Passé Simple, and between imperfective usages of the Simple Past and the Imparfait. Another finding is the existence of less frequent cases, such as imperfective usages of the Simple Past and the Passé Composé/Passé Simple, and perfective usages of the Simple Past and the Imparfait.

4.4 A Generalized Mixed Model with Tense, Aspect and Aktionsart

The results of the experiments from this chapter showed that the English Simple Past, on the one hand, and the French Passé Composé/Passé Simple and Imparfait, on the other, are correlated when it comes to three types of encoded information: the narrativity feature (i.e. temporal and causal relations); Aspect; and Aktionsart. As such, the Simple Past is used both for bounded and unbounded situations, presenting them from a perfective or an imperfective viewpoint, having narrative or non-narrative interpretations. Cross-linguistic analysis of translation corpora revealed that different combinations of these features correspond to translations into French either by an Imparfait or a Passé Compsé/Passé Simple.

Multifactorial statistical analyses were performed to investigate the relationships between the [±narrativity], [±boundedness] and [±perfectivity] features in predicting the verbal tenses used in the target language. In this section, I provide the results of the multifactorial analyses, performed with the statistical program R, and their interpretation.

The data used in multifactorial analyses consists of 435 items containing annotated Simple Past tokens for which the following information is known:

a. the verbal tense used in the target language
b. the verb in the source language in the infinitive
c. the stylistic register
d. for each item in the source language, the value of the [±narrativity], [±boundedness] and [±perfectivity] features

The dependent variable is a binary categorical variable—i.e. the verbal tense used in the target language, comprising 255 occurrences of the Passé Composé/Passé Simple and 180 occurrences of the Imparfait. The independent variables were classified as fixed predictors (the [±narrativity], [±boundedness] and [±perfectivity] features) and random predictors (the verb and the stylistic register). The three fixed predictors are correlated as shown by the two-by-two figures below (Figs. 4.9, 4.10 and 4.11). The Perfectivity and Boundedness correlation is statistically significant (Chisq 224.57, df = 2, $p < .05$), corresponding to a Cramer's V value of 0.469. The Perfectivity and Narrativity correlation is statistically significant (Chisq 95.71, df = 1, $p < .05$), corresponding to a Cramer's V value of 0.469. Finally, the Narrativity and Boundedness correlation is statistically significant (Chisq 147.28, df = 2, $p < .05$), corresponding to a Cramer's V value of 0.582.

Figure 4.12 presents the distribution of the data regarding the three fixed predictors established. It shows that there are two main tendencies, and that all combinations are possible for the Simple Past. The first main tendency is that the perfective viewpoint is associated with bounded situations in narrative contexts, and the second is that the imperfective viewpoint is associated with unbounded situations in non-narrative contexts.

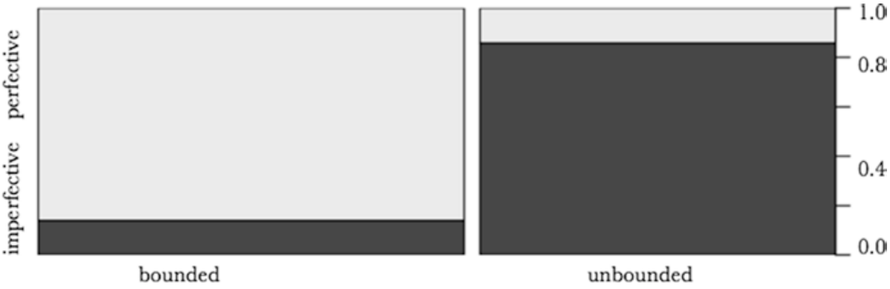


Fig. 4.9 Correlation between perfectivity and boundedness

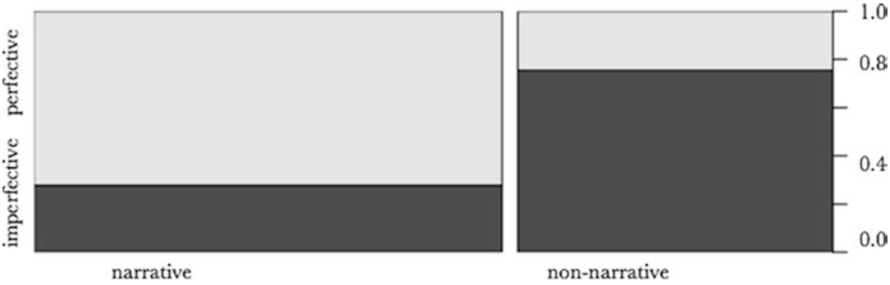


Fig. 4.10 Correlation between perfectivity and narrativity

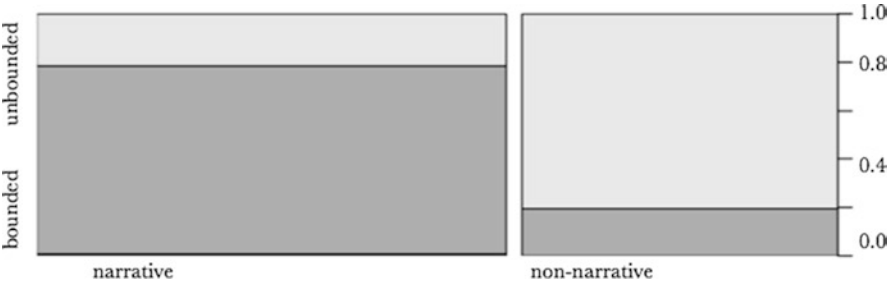


Fig. 4.11 Correlation between boundedness and narrativity

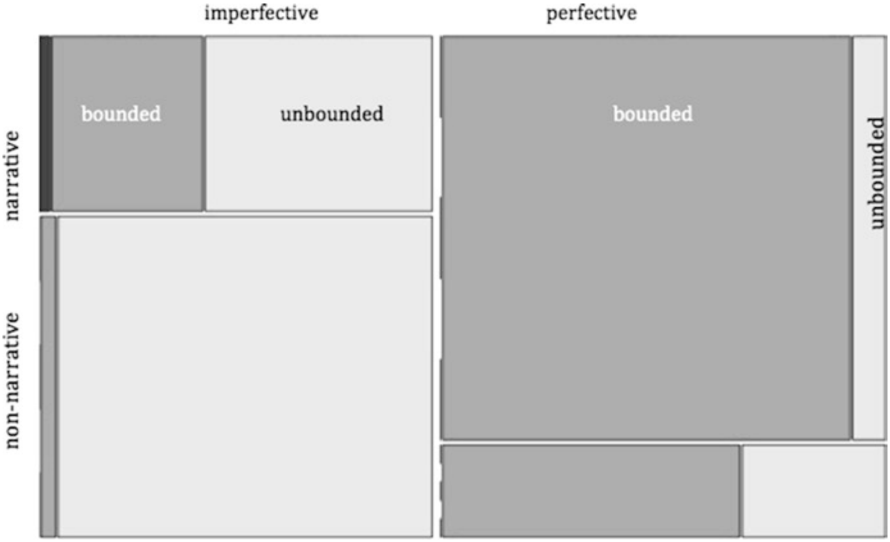


Fig. 4.12 Mosaic plot of the data with three fixed predictors: narrativity, perfectivity and boundedness

The order of the predictors for finding the best model (i.e. the balance between high within-dataset accuracy and high predictive accuracy for new data) was calculated with the Step function. An ANOVA performed on the results of the Step function is provided in Table 4.17. It can be seen that there are four significative predictors, one significative interaction (indicated by the ':' colon symbol) between Aktionsart and narrativity, and one almost significative interaction (between Aktionsart and Aspect).

Following the standard stepwise procedure which aims to adhere to Occam's razor, a maximal model was built—i.e. the model which includes all fixed and random predictors and their interactions. Secondly, other models were built by iteratively deleting the least relevant predictor. Finally, an ANOVA was performed on all the models, and the most effective model with the highest number of degrees of freedom was retained. The model best fitting the data is the model that considers the three fixed predictors and the interaction between boundedness and narrativity, as well as one random predictor, the verb. Table 4.18 provides the results of the best fitting model, and shows that narrativity and perfectivity, as well as the interaction between lexical aspect and narrativity, are statistically significant factors when predicting the verbal tense used in the target language.

Moreover, perfective viewpoint is negatively correlated with the Imparfait, whereas narrative usages of the Simple Past are positively correlated with the Imparfait. Moreover, bounded situations in non-narrative contexts are also negatively correlated with the Imparfait. This interaction is seen in Fig. 4.13. This model's predictive force when applied to new data is 0.83.

The results of the multifactorial analyses described in this section point to the cross-linguistic correlations between contextual usages of a verbal tense in the

Table 4.17 Order of predictors and their *p* value

Predictor	Df	Chi-square *p*
Boundedness	2	<.0001
Narrativity	1	<.0001
Perfectivity	1	0.001
Boundedness:Narrativity	1	0.03
Boundedness:Perfectivity	1	0.08

Table 4.18 Results of the mixed model

Fixed factors	P value
Boundedness	0.968
Narrativity	<.0001 ***
Perfectivity	0.004 **
Boundedness:Narrativity	0.04 *

The number of * signals the level of significance: *** highly significant, ** very significant, * significant

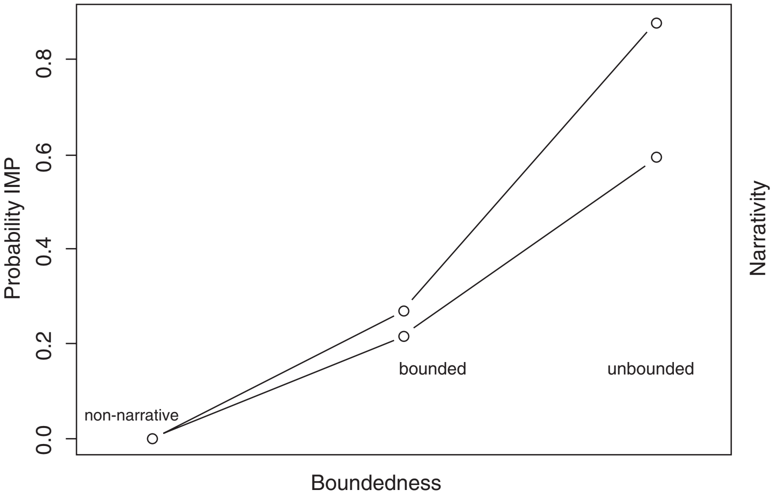


Fig. 4.13 Interaction boundedness*narrativity

source language and the corresponding verbal tenses used in a target language. A mixed model fitting the data indicated that there are three significative factors for predicting the verbal tense in a target language. The Imparfait can be predicted according to the procedural feature encoded by Tense, operationalized as the [±narrativity] feature, the procedural feature encoded by Aspect, operationalized as the [±perfectivity] feature, and, thirdly, the interrelation between the procedural feature [±narrativity], which constrains the interpretation of conceptual information encoded by Aktionsart, operationalized as the [±boundedness] feature. My suggestion is that humans treat temporal information from these three sources in a coherent manner. In particular, these linguistic data point to *temporal cohesion*, established at the level of the discourse. I will tackle this matter in Chap. 5. With respect to the addressee's cognitive faculties involved in the interpretation process, my suggestion is that comprehenders treat this temporal interpretation in a coherent manner, and that one can therefore speak about *cognitive temporal coherence.* This notion will be discussed in more detail in Chap. 6.

The predictive force of the model when applied to new data, at 0.83, illustrates that there is a share of the variability, when dealing with human language, which can neither be predicted nor modelized.[8] This share may be explained by the speaker's personal choices, as well as the translator's personal choices. When it comes to the variability that can be predicted, some specifications can be made. Four fixed factors and two random factors (i.e. stylistic register and the verb itself) have been considered in this mixed model. Other factors that might be studied are the conceptual difference between past and non-past, the speaker's subjective viewpoint, and the

[8] This indicates the lack of expectation of a deterministic linguistic model, and of the suggestion that there might be a share of variability due to the speaker's personal preference regarding, for example, the choice between a Passé Composé or a Passé Simple.

usage of the English progressive. The first was not included in this model, because all the verbal tenses from the target language are past time verbal tenses. The second one—to be precise, subjectivity—does not seem to be a type of information to which comprehenders have conscious access (Grisot 2017c). Finally, the third factor should be considered in future research, since it partly shares the same semantic and pragmatic domain as the Imparfait.

4.5 Summary

This chapter was dedicated to describing annotation experiments carried out in order to investigate how comprehenders consciously judge a series of characteristics linked to the encoded and inferred meanings of Tense, Aspect and Aktionsart. I have suggested that dealing with annotation data raises a certain number of issues, such as how to measure inter-annotator agreement rates, how to ensure the reliability of the data, and how to interpret the results. Following the proposal made in Grisot (2017a), in this chapter I used a chance-corrected statistical notion, the K coefficient, to measure inter-annotator agreement rate, and interpret high vs. low rates as indicative of high vs. low degrees of the experimental information's accessibility to consciousness. Additionally, according to Wilson & Sperber's cognitive foundations of the conceptual/procedural distinction (1993/2012), I expected to find systematically different behaviour among native speakers when they consciously evaluated these two types of encoded information—therefore, that conceptual meaning is available to conscious thought. For this reason, annotating conceptual information is expected to be a rather easy task, resulting in high inter-annotator agreement rates. Procedural meaning is more difficult to evaluate consciously than conceptual information is; as such, annotating procedural information is expected to be a more difficult task than judging conceptual information, resulting in medium inter-annotator agreement rates.

For these experiments, I have formulated a series of hypotheses based on previous research, and I have discussed their predictions in terms of accessibility to conscious thought and their cross-linguistic vs. language-dependent status. Two series of experiments were carried out. The first series targeted the category of Tense in English, French, Romanian and Italian, and the description of its meaning using Reichenbachian coordinates. The second series focused on temporal information, as conveyed by Aktionsart on one hand and Aspect on the other.

The experiments in Sect. 4.2 showed two systematic patterns. When participants deal with the localization of eventualities with respect to S—that is, in the past or non-past (present or future)—they indicate the ease of the task, and have high rates of inter-annotator agreement. When they deal with the localization of one eventuality with respect to another, they express the greater difficulty of the task, and have lower rates of inter-annotator agreement. Similarly, the experiments from Sect. 4.3 revealed the same patterns with respect to Aktionsart and Aspect. Again, these results were interpreted as indicating the conceptual nature of Aktionsart and the procedural nature of Aspect.

Finally, in Sect. 4.4, I reported the results of a generalized mixed model, built on the English-French data previously annotated with the [±narrativity], [±boundedness] and [±perfectivity] features. This analysis aimed to investigate the relation between these features when predicting the verbal tenses used in the target language. This mixed model indicated that the pieces of information from Tense (that is, the [±narrativity] feature) and Aspect (that is, the [±perfectivity] feature), as well as the interaction between Aktionsart (that is, the [±boundedness] feature) and Tense (again the [±narrativity] feature), are statistically significant factors when predicting the verbal tense used in the target language. In other words, cross-linguistically speaking, the three cohesive ties which this research considers model 83% of the temporal information expressed in a discourse, and allow the prediction of the verbal tense form to be used in a target language. Based on these results, in the next chapter I will propose a pragmatic model of temporal cohesive ties, and a cross-linguistically valid reanalysis of verbal tenses built on this model.