Ana Fernández and Gloria Vázquez*

The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality

Abstract: In this paper, we present the annotation scheme used in the SenSem corpora (SSC), for Spanish and Catalan, to codify information regarding aspectuality, modality, polarity and factuality. As regards aspectuality, the most relevant contribution is the codification of information about dynamicity, telicity and iterativity. Regarding factuality, we present a more fine-grained annotation of uncertainty as applied to the identification of impossible events, completely uncertain events and neutral uncertain events. Although information about factuality in Spanish has been provided elsewhere, the Catalan SSC is the only corpus to do so for Catalan.

Keywords: corpus annotation, aspectuality, modality, polarity, factuality, assertivity, certainty, impossibility, dynamicity, telicity

*Corresponding author: Gloria Vázquez: Universitat de Lleida, English and Linguistics, Lleida 25003, Spain. E-mail: gvazquez@dal.udl.cat
Ana Fernández: Universitat Autònoma de Barcelona, English and German Barcelona, Spain.

1 Introduction

The Spanish and Catalan SenSem Corpora (SSC) are made up of approximately half a million annotated words: 455,905 and 391,267, respectively. The Spanish corpus (Fernández et al. 2006) contains 30,365 sentences (25,075 extracted from the journalistic register and 5,299 from the literary). The Catalan corpus (Vázquez et al. 2013) does not contain sentences from the literary register.

1 Acronym for Sentence Semantics.
2 These corpora are the result of the work carried out by the research team members for the last nine years in five different projects (2004–2012). The last is Standardization and Transference of Lexical and Textual Resources – Ministerio de Ciencia e Innovación – FFI2011-27774.
These two SSC have been annotated with linguistic information regarding the lexical, phrasal and sentential levels. First, words were annotated with morphosyntactical tags using FreeNet (Atseries et al. 2006). This was the only completely automatic annotation process applied. Verbs were also disambiguated with respect to the sense they exemplify. This was accomplished by assigning each verb a sense from our lexical database, the corresponding WordNet sense and information about their Aktionsart. The next level of annotation was the phrasal level. Each phrase was marked with respect to first its syntactic category, and then its semantic and syntactic function. At sentential level, taking into account the type and order of participants in each clause, we created a hierarchy of constructions related to argument structure. This hierarchy includes agentive, causative, passive, anticausative, reflexive and reciprocal constructions, among other constructions. At sentential level, we also annotated some other information related to aspectuality as well as modality, polarity and factuality. In this paper we will deal with the annotation scheme applied to these last four features.

During the years we have been working in this project, 5 trained linguists have manually annotated the sentences following the guidelines (Vázquez et al. 2005) proposed. In (Alonso et al. 2007) the score for the evaluation of inter-annotator agreement is presented.

The corpora themselves are available online. The online interface allows users to browse several linguistic phenomena at once in a significant number of sentences for just one verb or a set of verbs. The corpora can also be downloaded in XML format.

The main contributions of this proposal are, for aspectuality, the codification of information about dynamicity, telicity and iterativity, and regarding factuality, a more fine-grained annotation of uncertainty as regards the identification of impossible events, completely uncertain events and neutral uncertain events. The Catalan SSC is the only corpus covering information about factuality. Prior to this work, only the SIBILA corpus (Wonsever et al. 2008) covered factuality for Spanish.

Next, we present the state of the art with regard to corpora containing information on aspectuality, modality, polarity or factuality. In Section 3, we present a general description of how aspectual information has previously been annotated. In Section 4, we present the annotation of modality, polarity and factuality as applied to the SSC. Section 5 presents some results and Section 6 is devoted to

---

3 Currently, this information is not visualized in the user’s interface.
4 http://grial.uab.es/sensem/corpus
conclusions. Finally, in the appendix we present a summary of the tags used to annotate information regarding aspectuality, modality, polarity and factuality.

## 2 Related works

### 2.1 Corpora annotated with temporal and aspectual information

As highlighted in Feldman and Arshavskaya (2007), it is common to find corpora annotated with information concerning verb tense and mood in relation to morphological information. However, with the exception of English corpora, there are a few corpora annotated with more specific temporal and aspectual information.

Interest in annotating temporal information in corpus linguistics began in the last decade with the objective of improving results in the field of information extraction. Nowadays the most comprehensive and ambitious scheme annotation is TimeML (Saurí et al. 2006), which includes dates, times, temporal relations and some aspectual information such as phases, mode, progressivity or (im)perfectivity. However, it does not include information about dynamicity, telicity and iterativity. This annotation scheme also covers aspects related to the annotation of modality and polarity. Two corpora have been annotated with this language: Time Bank 1.2 and AQUAINT TimeML (Pustejovsky et al. 2006). Versions of TimeBank in both Spanish and Catalan are currently available (Saurí 2010). TimeBank covers a total of 68,000 words.

### 2.2 Corpora annotated with information about modality, polarity and factuality

Several projects in which aspectuality and temporality are accounted for also take an interest in questions related to the description of modality, since all these theoretical aspects of language description are closely related. Similarly, modality and polarity are also related, so it is also common that a corpus that seeks to address one of these issues also needs to address the other. By modality we refer to the point of view of the speaker in relation to the degree of certainty about the events described (i.e., factuality). This type of annotation is currently one of the most innovative issues in the field of corpus linguistics.
The distinction between factive and non-factive events is related to aspectuality, polarity and modality and it is crucial to extract “real” facts from texts (Saurí et al. 2006, Baker et al. 2010 and Hendrickx et al. 2012). Nevertheless, the direct annotation of factuality is not common.

In TimeBank, the annotation of polarity is very basic, being either positive or negative. Modality is not specified comprehensively either and, in many cases, factuality is not indicated at all. FactBank was therefore created to make up for these shortcomings (Saurí and Pustejovsky 2009). In FactBank, about 10,000 events (belonging to the same 208 documents annotated in TimeBank) were manually annotated with information about factuality such that an event can be true (or not), possible (or not) or probable (or not). In addition, the type of source is indicated, that is, whether the narrator was directly the source or instead the event is being reported indirectly by another source (another person, a newspaper, etc.). Another corpus project in which exceptionality factuality is directly tagged, also manually, is SIBILA.

In addition, Vincze et al. (2008) and Hendrickx et al. (2012) have recently put forward new proposals in the field of the annotation of polarity and modality. Thanks to the existence of corpora already annotated with this information, automatic annotation of these phenomena has recently started. Results in this area are still not very good but efforts are being made in this direction (Morante and Daelemans 2012).

3 Aktionsart, aspect and aspectuality in SenSem

3.1 Lexical and phrasal levels

With regard to lexical aspect (Aktionsart), the traditional categories (Vendler 1957) are state (for non-dynamic situations) and process, accomplishment and achievement (for dynamic situations). In our proposal (Coll 2007), we use 4 tags: we keep the labels for states and processes; accomplishments and achievements are grouped into just one general category, event; and finally, we have added a fourth type, process-event, to indicate those cases where telicity is not lexically defined, following the commonly accepted idea that telicity is a lexical property that may be modified by certain elements in the sentence, specifically the type of object (phrasal level) or the verb tense.

An example would be analizar (1) (‘analyze’): in (1a) the object (underlined) triggers a bounded interpretation of the event whereas (1b) is an unbounded process.
Thus, the codification of \textit{Aktionsart} is inherited from our lexical database and maintained through the other phases of annotation, except in the case of verbs annotated as \textit{process-event}. These verbs, like \textit{analizar} ‘analyze’ are disambiguated at phrasal level as being either bounded (1a) or unbounded (1b).

Other predicates that are also specified lexically as “process-event” are certain verbs of movement, such as \textit{viajar} (‘travel’) or \textit{perseguir} (‘follow’). When they are used with a PP that limits the end of the path, they are interpreted as events (2a). If they are used without this PP they are interpreted as processes (2b).

\textbf{3.2 Sentential level}

At sentential level we also codify three other pieces of information: a) \textit{im}perfectivity, b) the temporary/permanent nature of stativity and c) iterativity. The labels used are:

a) \textit{Perfective vs. Imperfective}
b) \textit{Permanent State vs. Temporary State}
c) \textit{Habitual}

First, the element that contributes to determine the (im)perfective nature of a construction in Spanish/Catalan is verb morphology, since, in these languages, verb tense, mode and aspect are expressed by verbal inflection (2b). Certain auxiliary verbs may express (im)perfectivity too.

Second, we distinguish those states that are permanent (3a) from those that are temporary and, therefore, denote a reversible state (3b).
(3) a. Entre 40 y 45 personas (…) caben en una patera tradicional.  
`Between 40 and 45 people (…) fit in a traditional fishing boat.’

    b. La señora se encontraba mejor (…)  
`The lady of the house was feeling better (…)’

As regards the third element, iterativity, we annotate as habitual those sentences that denote repeated actions and are, therefore, not related to specific time-space coordinates. There are numerous linguistic elements that may contribute to this interpretation. For instance, in (4) there are two habitual events (gastar ‘consume’ and crecer ‘grow’). The key elements that point to a habitual interpretation for this sentence are the subject (Cada catalán ‘Every Catalan’, with distributive interpretation), the use of an imperfective tense (present), and the reference to the repetition of years (al año ‘annual’).

(4) Cada catalán gasta 3,3 toneladas de petróleo al año y la demanda crece un 3 % anual.  
`Every Catalan consumes 3.3 tons of petroleum a year and the demand is growing by 3% every year’.

4 Modality, polarity and factuality in SenSem

There are four basic types of epistemic modality described in the literature: (a) certainty, (b) non-certainty or counterfactuality, (c) possibility or probability, and (d) impossibility or improbability. The former two categories refer to what is certain (what really happens/has happened or does not happen/has not happened, 4.1), while the latter two refer to what is uncertain (facts are presented as uncertain, 4.2). In SenSem we identify these two categories with the tags assertive and non-assertive. Deontic modality has also been tagged as non-assertive. Sentences have also been annotated with information about polarity, positive or negative. The annotation of factuality derives from the information about modality and polarity, as we will explain below.

4.1 Factuality vs. non factuality/counterfactuality

Within the group of assertive sentences we further differentiate positive assertive constructions that express certain or factual events (5) from assertive events with negative polarity, that is, counterfacts (6):
(5) (…) Oftáler (…) ha invertido 110 millones en un nuevo servicio de cirugía refractiva.
   ‘(…) Oftáler (…) has invested 110 million [euros] in a new refractive surgery service.’

(6) El Athletic quedó noqueado y no se recuperó tras el descanso.
   ‘Athletic [Football Club] was knocked down and did not recover after the break.’

4.2 Annotation of non-factuality

In contrast to assertivity, polarity does not affect non-assertive sentences. In the literature, we find different subcategories within the field of non-assertivity that indicate degrees of uncertainty. In SenSem we have made three distinctions, some of which are new in the field of corpora annotation.

First, following proposals such as Ahern (2008) and Wonsever et al. (2008), we further differentiate two sub-categories within the field of doubt (i.e., non-assertivity). We have annotated the cases of non-assertivity when it refers to past or present events (7) differently than when it refers to future events (8):

(7) Es posible que, inicialmente, hayan beneficiado a los países desarrollados.
   ‘They may initially have benefited the developed countries.’

(8) Quizá me destinarán fuera de aquí.
   ‘Maybe I will be sent away.’

This distinction, not present in FactBank but present in SIBILA, is important in order to make inferences. In (7), uncertainty is based on ignorance regarding whether the facts are certain or not; that is, facts are presented as uncertain, regardless of whether they might have truly occurred or might not have, so that their degree of certainty is actually unknown. By contrast, the second type of uncertainty (8) might be called absolute uncertainty because we are describing something that belongs to the future.

Second, we have annotated impossible events, a tag used in neither FactBank nor SIBILA, to label past or present situations presented as completely unreal (9), because the interpretation is that the situation described never happened and therefore we consider them intrinsically impossible (Morante and Daelemans 2012).

5 In terms of factitivity, these are counterfacts.
This is not the case for future events, because, even when they are presented as virtually impossible (10), they are not intrinsically impossible since they are still outside the scope of what is real. Again, this distinction is important when it comes to making inferences.

(9) Si el Barcelona hubiera tenido diez palcos, los habría llenado hasta los topes

‘If Barcelona [Football Club] had had ten boxes, they would have been completely full.’

(10) Ocasiones que no se consiguen, pero que de conseguirse nos transformarían en ángeles, evitarían el que siguiéramos enangados en el crimen y el pecado (. . .).

‘Opportunities which do not materialize but which, if they did, would transform us into angels, would prevent us from becoming mired down in crime and sin (. . .).’

Finally, the distinction between probable and possible events, on the one hand, and improbable and impossible events, on the other, is represented in FactBank, thus allowing us to express the full range of degrees of uncertainty. In SSC we feel that it is important to differentiate between when an statement is presented neutrally as uncertain (10) and when it is presented as uncertain by the use of an epistemic expression denoting probability -*por supuesto* 'of course'- or possibility -*quizás* 'maybe' (8).

(11) Banderas aseguró que ‘en España nunca se hubiera concedido a un actor’ un reconocimiento de este tipo.

‘Banderas declared that ‘recognition of this kind would never have been granted to an actor in Spain’.

Had we used the adverb *probablemente* ‘probably’ with *llenar* ‘be full’ in sentence (9), the interpretation of uncertainty would be more clear-cut, as is the case of example (8). In this respect, the uncertainty is more ‘neutral’ since a subjective opinion is not directly identified in (9). Furthermore, in SSC, whenever we come across an uncertainty marker, we further annotate the sentence as epistemic.

The tagging of modality, polarity and factuality takes place at the sentential level. Occasionally, we have been able to automatically pre-assign some of these values. For example, it is simple to detect the presence of words such as adverbs, determiners and pronouns that make statement negative. All remaining statements are therefore labeled positive. Thus, the annotation of polarity becomes a trivial task that can be automatized. Regarding modality, the pre-assignment of
values has sometimes been possible for non-assertive sentences when formal marks, such as specific verb tenses, like subjunctive, future or conditional (non-assertive general tag), are used. However, this automatic annotation must be double-checked manually given the fact that we can find instances like (12) where the simple past conditional expresses assertivity:

(12) A lo largo de esa angustiosa madrugada, Juan me aclararía que no había ningún secreto (...).

‘In the course of those distressing early morning hours, Juan would explain that there was no secret (...)’

In appendix 1 we present a summary of the tags used in relation to polarity and assertivity (scheme 2) and also show how these tags are combined to express factuality (scheme 3).

5 Results

We present a graphical example of the annotation (Figure 1), in this case taken from the Spanish SSC. It corresponds to the annotation of sentence (4), which is associated with sense 1 of the verb gastar (‘consume’). The sentence excerpted from the corpus (‘Cada Catalán gasta 3,3 toneladas de petróleo al año’) is given in the middle bar. The three bands 2a–2c above the sentence are used to represent information about sentence semantics: semantic construction at argument-structure level is labeled in (2a), factivity in (2b) and aspectuality in (2c). For factuality, the combination of assertive and positive polarity indicates that we are describing a “real” event. For aspectuality, on the one hand, the sentence is annotated as an event, in this case denoting an iterative situation (“habitual” tag). The tag “event” has been inherited from the lexicon. Information about argument structure is provided in bands 1a–1c under the sentence (orange-shaded boxes) with syntactic functions appearing in (1a), syntagmatic categories in (1b) and semantic roles in (1c).

In Figure 2 we present some information in the lexical entry for gastar 1, where we see, among other information, that this verb is classified lexically as bounded (event).

6 We only present data for Spanish because, except for the literary subcorpus, Catalan sentences are the same as Spanish sentences and, consequently, the annotation of aspectuality, modality, polarity and factuality is coincident. Furthermore, we do not differentiate according to register for Spanish SSC because differences do not appear to be significant.
**gastar (1):**

| 1a | Nominal phrase (common noun) |
| 1b | Agent |
| 1c | Nominal phrase (common noun) |
| 2a | Cada catalán | gasta | 3,3 coneladas de peróleo | al año | Industria aconseja el ahorro domés |
| 2b | Modality: assertive · Positive polarity |
| 2c | Event - Habitual |

Fig. 1: Example of annotation

**Definition:**
Hacer uso de dinero u otro bien o producto que pueda consumirse.

**Semantic roles:**
Agent,
Theme,
Purpose

**Aspectual class:**
Event

**Wordnet:**
00661955v

Fig. 2: Lexical entry for *gastar_1* in SenSem lexicon
We present the figures for modality and polarity in the Spanish corpus in Tables 1 and 2. As can be seen, the vast majority of sentences are assertive and positive. The difference between positive and negative statements is highly significant. Negative statements represent merely 7.6% of the sentences in the corpus.

**Table 1: Modality in the Spanish SSC**

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assertive [5/6]</td>
<td>23,451</td>
<td>77.21%</td>
</tr>
<tr>
<td>Non-assertive sentences [14]</td>
<td>6,923</td>
<td>22.79%</td>
</tr>
</tbody>
</table>

**Table 2: Polarity in the Spanish SSC**

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive [5/7a]</td>
<td>27,987</td>
<td>92.14%</td>
</tr>
<tr>
<td>Negative [6/7b]</td>
<td>2,387</td>
<td>7.85%</td>
</tr>
</tbody>
</table>

In Table 3 we present the figures for each subtype of non-assertivity in the Spanish corpus and note that future non-assertive events clearly predominate (they constitute more than half). Impossible sentences are poorly represented (less than 1%). Also, there are very few sentences in which a direct mark for epistemicity is found (1.40%).

**Table 3: Non-assertivity in the Spanish SSC**

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future [8]</td>
<td>4,521</td>
<td>65.30%</td>
</tr>
<tr>
<td>Past and present [11]</td>
<td>1,070</td>
<td>15.46%</td>
</tr>
<tr>
<td>Impossible [10]</td>
<td>65</td>
<td>0.93%</td>
</tr>
<tr>
<td>Epistemic [8]</td>
<td>97</td>
<td>1.40%</td>
</tr>
</tbody>
</table>

Dynamic situations are the most common type of action found in the corpus (86.34%) (Table 4) and events practically double the number of processes. As for

7 Numbers between square brackets in this section refer to the sentences in the paper used to exemplify this kind of meaning.
states, a small number of temporary stative constructions have been found in the corpus. At the other extreme, the difference between the number of perfective and imperfective situations is not significant. Finally, only 10% of the actions are habitual.

6 Conclusions

The SSC are quite large in terms of the diversity of information reported. Furthermore, they constitute important resources in that they incorporate information about high-level sentence semantics, which has been the focus of this paper, namely aspectuality, modality and polarity, which together provide information concerning factuality. It should be emphasized that the Spanish SSC is one of only two corpora with information regarding factuality in Spanish. As for Catalan, this is the first corpus to annotate this kind of information.

As indicated by Morante and Daelemans (2012), “factuality involves polarity, epistemic modality, evidentiality and mood” (p. 3). We would add that aspectuality also plays a central role, as in the case of habitual interpretation. This is one key contribution of this project, since there are no other corpora where habitual actions are annotated. Also, the SSC are the first corpora in which reversibility of states is represented. Another contribution is the annotation of the degrees of uncertainty, which is not considered in similar projects. The identification of

8 States and future or imperative actions (30% of the total) have not been annotated regarding (im)perfectivity.
these nuances in texts are very important when it comes to extracting inferences from them, which can be very useful in specific tasks of natural language processing (NLP).

These two SSC are also the first corpora where dynamicity and telicity is represented. Though this kind of information has no direct applicability in NLP, it does hold a certain value from the point of view of descriptive linguistics.

As regards results, we can claim that, both in Spanish and Catalan, the most frequent situations found are positive assertive and dynamic actions. Imperfective and perfective situations are almost equally represented. As for non-assertive situations, they are clearly more represented when dealing with the future than with the present or the past.

Appendix

Scheme I. Aspectuality tags

I. Lexical level

- Inheritance of tags from the lexical database:

Dynamicity:
- Neutral with respect to telicity:
  - State (Non dynamic)
  - Process-event (Dynamic)

Defined with respect to telicity:
- Event (Dynamic)
- Process (Dynamic)

II. Phrasal level

- Modification of process-event tag:

Telicity:
- Event
- Process

III. Sentential level

- Subspecification of state tag:

Duration:
- Permanent state
- Temporary state
Addition of tags:
Perfectivity:
  ➢ Imperfective
  ➢ Perfective
Iterativity:
  ➢ Habitual

Scheme 2. Combination of polarity and modality tags to express factuality

Sentential level
- **Factuality/Certainty**: Positive polarity + Assertive
- **Counter-factuality/Non certainty**:
  ◦ Negative polarity + Assertive
  ◦ Positive/Negative polarity + Non-assertive + Past-present + Impossible
- **Non-factuality/ Uncertainty**: Positive/Negative polarity + Non-assertive
  ◦ Scope of uncertainty:
    ■ Unknown uncertainty: Past/present
    ■ Absolute uncertainty: Future
  ◦ Degrees of uncertainty:
    ■ Epistemicity: Epistemic

References


Bionotes

Gloria Vázquez ■■■

Ana Fernández ■■■