A multilingual annotated corpus for the study of Information Structure 1

Lisa Brunetti, Stefan Bott, Joan Costa, Enric Vallduví

Lisa Brunetti, Stefan Bott, Joan Costa, Enric Vallduví

A multilingual annotated corpus for the study of Information Structure[1]

Abstract

This paper presents a corpus of spoken narrative texts in Catalan, Italian, Spanish, English, and German. The aim of this corpus compilation is to create an empirical resource for a comparative study of Information Structure. A total of 68 speakers were asked to tell a story in an acoustically isolated room by looking at the pictures of three textless books. A total of 222 narrations resulted in about 16 hours of speech. The recordings have been transcribed and an original annotation of non-canonical constructions for the Romance subgroup has been proposed, namely of morphosyntactically/prosodically marked constructions that relate informational categories such as topic, focus, and contrast. Transcriptions and annotations of some selected high quality recordings have been aligned to the acoustic signal stream. The corpus is available in audio and text format.

1. Introduction

In this paper we present a corpus that has been developed within the NOCANDO project, 'Non-canonical constructions in oral discourse: a cross-linguistic perspective' at the University of Pompeu Fabra in Barcelona (Spain). The main interest of the overall project was to study the cross-linguistic variation in the overt marking of Information Structure (from now on, IS) in general and more specifically in the spoken, narrative register.

Researchers largely agree on the fact that languages use syntactically, morphologically and/or prosodically marked constructions to represent informational categories such as 'topic', 'focus', 'contrast', 'background', etc. (cf. Vallduví 1992, Vallduví and Engdahl 1996, Lambrecht 1994, Erteschik-Shir 1997, Steedman 2000, among many others). A large part of the research on IS, in particular within the generative framework, has mostly or exclusively relied on introspective judgements on sentences in isolation (see e.g. the works by Belletti, Rizzi, Zubizarreta, among many others). Nevertheless, explicit marking of IS through non-canonical constructions is much more frequent in spontaneous speech than in written or controlled discourse. In addition, a written text does not represent intonation, which is extremely important for the marking of IS.[2] Furthermore, IS can only be truly understood if sentences are considered within their linguistic context. Sentences in isolation, such as those that are constructed for introspective judgements, are therefore suboptimal to understand the properties and function of informational categories.

A better source of data for the study of IS is therefore constituted by spontaneous speech corpora. Although speech corpora are available in literature, multilingual corpora that provide comparable data across languages are rather limited in number. Furthermore, access to speech corpora, in particular to the recordings, is often very restricted. These considerations led us to

[1] We wish to thank Estela Puig Waldmüller for collaborating in the recording, Teresa Suñol for her help with Catalan and Spanish transcriptions, Josep Maria Fontana, Louise McNally, Gemma Boleda, and Alex Alsina for their advice at different stages and on different aspects of the preparation of the corpus. We also thank the participants of the *Corpus Linguistics Conference* in Liverpool (July 20-23, 2009) and the *Corpus and Grammar 3* conference in Mannheim (Sept. 22-24, 2009) for their comments and questions. This research has been partially funded by the Spanish Ministry of Education and Science project OpenMT (TIN2006 15307-C03-02). The NOCANDO project was funded by the *Spanish Secretaria de Estado de Universidades e Investigació*n of the *Ministerio de Educación y Ciencia* (n. I+D HUM2004-04463).

[2] In fact, it is the most important resource for marking IS in certain languages, such as English.

compile a corpus of spontaneous spoken narrative texts in five different languages: Catalan, Italian, Spanish, German, and English. The audio recordings are freely accessible for consultation. A taxonomy of non-canonical constructions (from now on, NOCANs) was also established and an annotation of the relevant subset of the taxonomy was added to the Romance subset of the corpus. The annotation is meant to facilitate the search for IS markings in the text. The corpus is available in audio and text format. Some selected recordings were also aligned to the transcription and annotation using the PRAAT software for acoustic analysis (Boersma et al. 2009).[3] Our corpus is publicly available under a Creative Commons license which only excludes commercial use. Use for research is free as long as the work is properly cited and all derivates of the corpus are shared under the same conditions. A more detailed description of the corpus and its annotation is given in the following two sections.

2. The corpus

A total of 68 speakers were asked to tell a story by looking at the pictures of three textless picture story books. The result is 222 narrations of about 2-9 minutes each (a total of about 16 hours of speech). The quantitative information for each language is given in the table below.

	Catalan	**Italian**	**Spanish**	**German**	**English**
Speakers	19	16	13	9	11
Recording time	4:02:43 h	4:04:32 h	2:35:20 h	2:09:13	2:32:20 h
Word count	37555 w	27392 w	25077 w	15944 w	21970 w (estimated)
Segment count	5856 seg	4306 seg	3801 seg	2154 seg	3140 seg (estimated)

Table 1: Quantitative information on each language represented in the NOCANDO corpus.

2.1. Speakers

Participants were mostly university students. The Catalan and Spanish speakers were mostly undergraduate or graduate students with an average age of 22 for Catalan (ranging from 18 to 30) and 20 for Spanish (ranging from 17 to 29). The Catalan speakers were from Catalonia (except one speaker from Valencia). The Spanish speakers were also from Catalonia (except one from Castilla y León), but they spoke Spanish as their first language. The Italian and English speakers had recently arrived in Barcelona. The average age of the Italian speakers was 29 (ranging from 20 to 56). They spoke geographically different varieties of Italian. The English speakers' average age was 27, ranging from 20 to 41. They came from the United States and Great Britain. A large number of the German speakers were also short-term residents in Barcelona, but a smaller number exclusively resided in Germany. The German speakers' average age was 34, ranging from 22 to 67. They came from different parts of Germany.

2.2. Methodology

Speakers were asked to narrate three stories to an experimenter while looking at the pictures of three textless books by Mercer Meyer. The books were given to each speaker in random order. Speakers were allowed to browse through the book before they started the narration. Most speakers were recorded in an acoustically isolated room while some were recorded with a

[3] Praat is available at http://www.praat.org/.

portable recording device in a silent room. In both situations speakers were sitting in front of the experimenter.

The books are entitled *Frog goes to dinner*, *A frog on his own*, and *One frog too many* and are about the adventures of a boy and his pet frog. Mercer Meyer's books have already been used in literature for the study of narration strategies in monolingual or bilingual children, adults, and second language learners (cf. Berman & Slobin 1994, Strömqvist & Verhoven 2004 and references quoted therein). The book used for those studies is *Frog where are you?*. We made a first set of pilot recordings with this book as well as with four other books. The three aforementioned books gave better results in terms of variety of NOCANs to be used by the speakers.[4] The story entitled *Frog goes to dinner* is about the disastrous effects of the presence of the frog in a very elegant restaurant. The advantage of this story is that it includes many different characters interacting with the frog in different ways so we could expect many topic changes. The story entitled *A frog on his own* tells of the adventures of a frog taking a walk by himself in the park. The frog interacts with other characters, but unlike in the dinner story, where the other characters may temporarily have a prominent role, in this story the frog always remains the central character in the narration. The story entitled *One frog too many* tells of the frog's jealousy of a younger frog who has become the boy's new pet. This story was chosen because it presents situations in which the speaker has to distinguish between the two frogs. These situations are interesting because they induce the speaker to adopt constructions that explicitly mark the informational category of *contrast*.

2.3. Transcription and segmentation

The recordings have been transcribed according to the guidelines for the transcription of the LIP corpus (*Lessico di Frequenza dell'Italiano Parlato* 'Frequency lexicon of spoken Italian', De Mauro et al. 1993). We represented the phenomena that are typical of spoken text: pauses, false starts, truncated words, laughs, hesitations, and vowel lengthening, among others. The transcription also followed orthographic standards.

Since the notion of 'sentence' is often not clear in spoken text, the segmentation was carried out by separating clauses: each segment generally contains one main verb, except for modal, temporal, and aspectual periphrases and verbless clauses. This criterion is very similar to that used in the transcription of CHILDES data (MacWhinney 2000). In order to recognize verb periphrases, we adopted the criteria proposed by Gavarró and Laca (2002). Segments were separated as different XML-marked units; each unit was given a unique id and NOCANs were treated as attributes of segments.

Some selected recordings were aligned to the transcription and annotation using PRAAT, as exemplified in figure 1. The alignment between text and speech signal makes it possible to quickly identify the relevant segments and locate them in the audio recording. Subsequently, all options for acoustic analysis offered by PRAAT may be exploited.

[4] It must also be noted that *Frog where are you?* has been preferred in past literature because it did not presuppose a specific socio-cultural background so it could be used with speakers of different origins. Since our purpose was not to study the sociolinguistic aspects of narration, we did not consider the socio-cultural implications of a story as a relevant factor for choosing it or not.

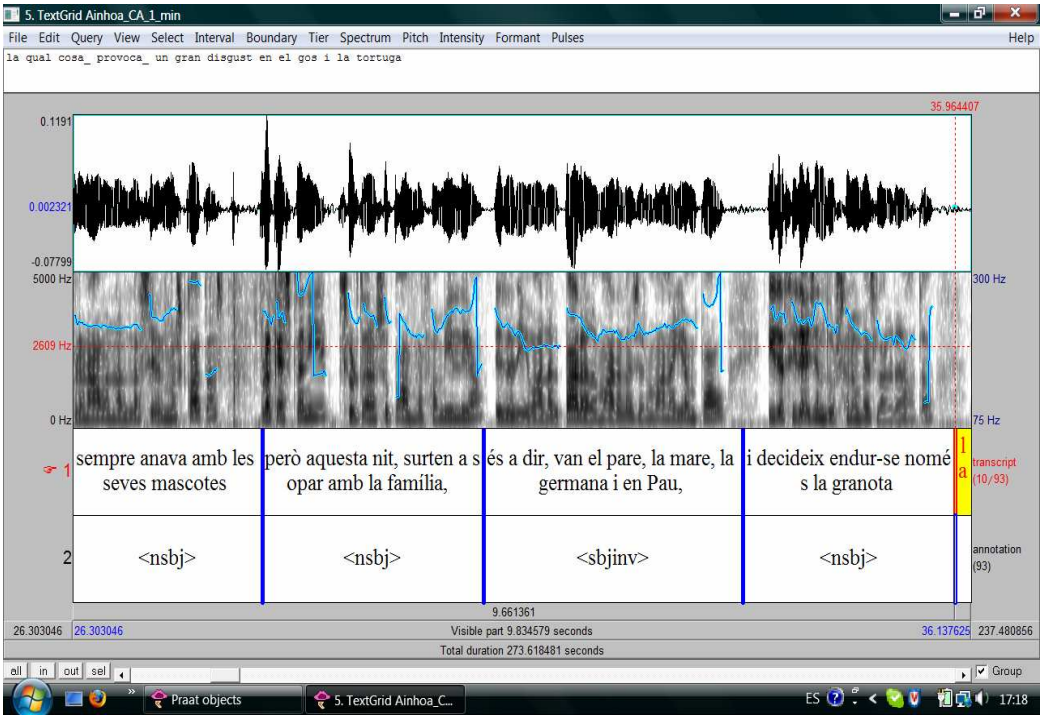

Figure 1: Audio-transcription alignment and phonetic analysis with PRAAT.

3. The annotation

An annotation of non-canonical constructions (NOCANs) was carried out for the three Romance languages of the recording: Catalan, Spanish, and Italian. The reason for choosing these languages is that they are typologically similar. For instance, they all have a relatively free word order with SVO being the canonical one; they allow for null subjects, and they display left and right dislocations, including clitic dislocations. Their similar linguistic properties are parallel to their similar strategies for expressing informational notions. All these languages largely use syntax for this purpose, as opposed to languages like English, which nearly exclusively use prosody (see Vallduví and Engdahl 1996). The NOCANs found in these languages are indeed very similar or identical cross-linguistically. Despite these similarities, however, it has been shown in literature (cf. Villalba 2007, Leonetti 2008) that the use of individual NOCANs within Romance languages can vary considerably in frequency and function. A comparison between these languages is therefore important in order to show what such (quantitative or qualitative) differences are.

We must stress that an annotation of NOCANs is extremely rare in existing corpora, even those created with the explicit purpose of studying IS. The only example we are aware of is the MULI corpus ('MUltiLingual Information structure') of read German and English newspaper texts (Baumann 2006). Its annotation, which only concerns a small part of the corpus (only 250 sentences for the German part), includes NOCANs such as clefts, pseudo-clefts, extraposition, fronting, and passive sentences. Informational categories themselves are also annotated. Apart from its reduced extension, the main limit of this corpus is that it is not made of spontaneous speech, which is a crucial aspect for the study of IS, as argued above. Note further that *read* intonation is known to be different from intonation in spontaneous speech, in particular concerning the marking of informational categories (Hirschberg 1995). We therefore think that our corpus provides an empirical resource which had been missing up to now and which can enhance the research on IS and on its interaction with the other parts of grammar.

Finally, it is important to clarify the difference between NOCANs as objectively observable events in language production, and IS proper. Studying IS can be compared to hunting for ghosts in that both phenomena are not directly and objectively observable. While ghosts manifest themselves through knocks and falling objects, IS units such as topics and foci manifest themselves through linguistic marking (i.e. NOCANs). In both cases we have to deduce the existence and presence of the object we are primarily interested in from objectively observed events. That explains why in the creation and annotation of the corpus we decided not to annotate directly IS units but the NOCANs which mark them. Although annotating IS categories is possible, in principle, the resulting annotation is less reliable because extra-sentential factors such as the linguistic context or even extra-*linguistic* factors such as the speaker's intentions have to be taken into account, and that often induces the annotator to make subjective choices. Furthermore, the annotation of IS categories is often hardly free from the bias of a specific theory on IS. In fact, a considerable amount of disagreement is still found among different schools concerning the identification and the definition of IS categories. The annotation of NOCANs is, thus, more neutral and objective.

4. Non-canonical constructions (NOCANs)

As we said above, NOCANs are marked constructions from a syntactic, morphological and/or prosodic point of view. Given the freedom of syntactic possibilities existing in the three Romance languages, the NOCANs we annotated where mostly morpho-syntactically marked constructions. The only exceptions are clear cases of deaccenting and focus fronting. In the latter construction, syntactic fronting is accompanied by prosodic fronting of the focal accent, and when the fronted focus is a preverbal element, fronting is only revealed by prosodic marking. It must be made clear that our preference for syntactic NOCANs does not mean that prosody plays no role in the representation of IS in the three Romance languages under study. Nevertheless, a fine-grained phonological annotation (e.g. of different types of pitch accents) is a very complex task that goes beyond the scope and the expertise of our team. It must be pointed out, however, that although phonological properties are not annotated, the relevant information is stored in the audio recordings and a subset of the recordings is aligned with the transcription and thus prepared for in-depth phonetic and phonological analysis. Our corpus is therefore potentially ready to be enriched with this kind of annotation in the future.

As we said above, our interest in NOCANs is in the fact that these constructions reveal the IS properties of a sentence. More precisely, NOCANs usually single out a specific information structural unit, such as a *topic* or a *focus*. Some constructions are used for instance to distinguish a topic from its *comment*. A topic expression indicates the referent, about which the sentence conveys some information, which is represented by the comment (Strawson 1964, Reinhart 1982, Vallduví 1992). The construction called 'clitic left dislocation' (Cinque 1990) is used in many Romance languages precisely to make such a distinction. An example is given by (1), from the Spanish sub-corpus:

(1) Al hombre se le cae el café
to-the man RFL to-him falls the coffee
Ind. object obj.cl. Verb Subject
"The man drops the coffee"

The indirect object *al hombre* is left dislocated; the canonical position would be the post-verbal one. The left dislocation leaves a remnant clitic pronoun (*le*) in verb adjacent position and the subject occurs post-verbally. By occupying a sentence initial, pre-verbal position, the indirect

object is easily identified as the sentence topic, while the rest of the sentence (verb + subject) constitutes the corresponding comment.

Another informational partition that can be marked by specific NOCANs is the *focus-background* partition. The focus is the informationally most relevant part of a sentence in a particular context. In the construction exemplified in (2), the focused direct object (capital letters indicate focal accent) occupies a non-canonical preverbal position (*without* clitic remnant, in this case).

(2) Pure la LINGUACCIA, gli fa, la rana.
even the tongue to-him$_{cl}$ he-puts-out the frog
Dir. Object Verb Subject
"Even the tongue did the frog put out to him" Italian

The construction hence explicitly and unambiguously separates the focus element from the rest of the sentence (the background). A *focus-background* construction can also be represented by a cleft sentence, as in (3):

(3) i és ELLA que està a punt de prendre's el biberó
and is her who is about to take for-herself the baby-bottle
"and it's him who is going to drink from the bottle" Catalan

A cleft is made of two clauses: one introduced by the verb 'to be', and the other introduced by a complementizer. The two clauses allow for a clear separation of the focus from the background: the copular clause is occupied by the sentence focus (the subject *ella* in 3), and the other clause represents the remaining background.

5. Taxonomy of NOCANs

A full list of NOCANS that are represented in our taxonomy is given in Table 2.

Label	**Description**		**Label**	**Description**
sbjinv	Subject inversion		*cldbl*	Clitic doubling
sbjinv_deacc	Subject inversion with deaccenting		*obj-sep*	Object separation
nsbj	Null subject		*narg*	Null argument
nsbj_c	Null subject in a coordinate clause		*focfr*	Focus fronting
arbnsbj	Arbitrary subject		*deacc*	De-accenting
sbj-sep	Subject separation		*pres*	Presentational sentence
clld	Clitic left dislocation		*pass*	Passive construction
ld	Left dislocation		*impers*	Impersonal construction
ht	Hanging topic		*cleft*	Cleft sentence
clrd	Clitic right dislocation		*pscleft*	Pseudo-cleft sentence
rd	Right dislocation		*inv-pscleft*	Inverted pseudo-cleft sentence

Table 2: The taxonomy of NOCANs for the Romance languages in the NOCANDO corpus

Whenever a certain construction represents a particular case of a more general construction, we assigned the former a label that contains the label of the latter. For instance, the label *arbnsbj* for arbitrary subjects, which are a particular case of null subjects, contains the label of null subjects: *nsbj*. We subdivided our labels into three groups: NOCANs that are specific to the

subject, NOCANs that concern all arguments, and sentential NOCANs. We will describe them in details in the following subsections.

5.1. Labels specific to subjects

A subset of labels that we have proposed is specific to subjects. The three Romance languages have a default SVO order, so all constructions in which the subject did not occupy a pre-verbal position are marked as NOCANs. The label for post-verbal subjects is s*bjinv* ('subject inversion'). A Catalan example is given below:

(4) Els va acompanyar el taxista
them$_{cl}$ PAST take the taxi-driver
"The taxi-driver drove them" Catalan

The subject occurs post-verbally, while the direct object is moved to a pre-verbal position, thus rendering an OVS order. Inversion usually leaves the subject in focus or part of the focus. Note however that the informational role of the subject does not only depend on its postverbal position. If the postverbal subject is deaccented, it will be part of the background. For this reason, a deaccented post-verbal subject is marked with an additional label: *sbjinv_deacc*. In (5), *aquest nen* (orthographically separated from the verb by a comma) is the deaccented subject.

(5) ...que està disfressat, aquest nen
for is dressed-up this child
"...for this child is dressed up" Catalan

Since all of the Romance languages we examined may avoid expressing the subject overtly, we frequently find null subjects (*nsbj*). We only annotated *nsbj*s in finite clauses, as infinitive ones are canonically subject-less. Null subjects generally refer to an entity that is salient in the context. They function largely in the same way as unaccented pronouns in languages like English. They can neither be focal nor topical. In (6), the *nsbj* in the *perché*-clause refers to the boy, namely an entity that is already salient in the context (it is the topic of the preceding clause).

(6) Invece il bambino è molto contento, perché ha salvato la sua rana
instead the boy is very happy because has saved the his frog
"The boy on the contrary is very happy, because he saved his frog" Italian

A special case of *nsbj*s are those which occur in coordinated clauses. We annotated them apart with the label *nsbj_c*. The reason is that these constructions can be interpreted as VP coordination, in which case subject omission in the second conjunct is expected.

(7) Entonces la tortuga lo ve y hm se lo dice al niño
Then the turtle it sees and uh him it tells to-the boy
"Then the turtle sees it and she tells the boy" Spanish

A null subject may not refer to a definable entity but receive an arbitrary interpretation. While in English a plural pronoun is used in these cases (e.g. *they killed Kenny*), in Romance these subjects are necessarily non-overt. We assigned them the label *arbnsbj* for 'arbitrary (null) subjects'.

(8) Y un día a este niño le regalaron pues una caja muy grande
and one day to this boy to-him$_{cl}$ they-gave well a box very big
"And one day this boy received a large box" Spanish

The importance of marking *arbnsbj*s is that they seem to play a role in sentence topic selection: the fact that the subject is arbitrary in reference makes the object a potentially better topic (cf. Brunetti 2009a).

An example of omitted argument that cannot without controversy be called subject is the argument of a copula sentence. In the subordinate clause of (9), the copula verb occurs in initial position and there is only one argument in the sentence (*la seva granota*). Since copula constructions always need two arguments, evidently one is missing here. However, it is not entirely clear that the missing argument is really the subject (Alsina 2004). In order to keep such cases apart from uncontroversial cases of *nsbj*s, we label them *narg* ('null arguments').

(9) i llavors en Jaume es va adonar que que, home, era la seva granota
and then the Jaume RFL PAST realizes that that well was the his frog
"and therefore Jaume realizes that that, well, it was his frog" Catalan

Finally, constructions are annotated where the subject is separated from the verb by intervening material. If the separating material is another argument, then the subject is dislocated even though no subject clitics exist in these languages to mark the dislocation explicitly. We will discuss this case below when we introduce dislocation constructions. When the subject is separated from the verb by adjunct material, its syntactic position and its informational status are less clear. That is why we give these constructions a label apart: *sbj-sep* ('subject separation').

5.2. Labels for all arguments

We have marked different kinds of argument detachments. Since object clitics exist in these languages, object dislocations are marked with a clitic remnant adjacent to the verb. Clitic left dislocations have been annotated with the label *clld*. As we said in section 2, ex. (1), the dislocated element is generally recognized as the sentence topic (cf. Vallduví 1992, Benincà 1988[2001], Zubizarreta 1998, 1999, among many others).

There is a certain variation among Romance languages with respect to the use of clitics in object dislocations. When the dislocated element is not accompanied by a clitic remnant, the construction is labelled *ld* ('left dislocation'). In Italian, the remnant clitic of an indirect object is not obligatory (cf. 10). Its presence is associated with register, namely it is more common in colloquial speech.

(10) A un bambino un giorno arriva un regalo
to a boy one day arrives a present
"One day a boy receives a present" Italian

Given that there are no subject clitics in the three Romance languages, a preverbal subject can be demonstrated to be left dislocated only if it is separated from the verb by another argument. In that case, the subject will be assigned the same label *ld*.

(11) Esta a mí no me quiere nada bien
this-one to me not me$_{cl}$ loves no good

"this one doesn't love me at all" Spanish

Following Vallduví (1992), among others, we assume that the informational function of an *ld* is not different from that of a *clld* (it marks the sentence topic).

Another left dislocation construction existing in these languages is 'hanging topic' (cf. Cinque 1977, 1990). We assigned it the label *ht*.

(12) La rana grande, la situación no le gustaba mucho
the frog big the situation not to-her$_{cl}$ pleased much
"As for the big frog, she didn't like the situation at all" Spanish

An *ht* is a detached element that has no marking of grammatical function (it is never a Prepositional Phrase, always a Noun Phrase) and is obligatorily resumed by a pronoun expressing its grammatical function. Unlike a regular *clld*, an *ht* can also be resumed by a strong pronoun or a demonstrative. If the *ht* is a subject or an object, the construction can only be distinguished from a *clld* by the presence of a strong pronoun or a demonstrative.[5] As the name itself makes clear, this construction marks topic material as well.

Clitic right dislocations (*clrd*) are dislocations of an argument to the right, with resumption of a clitic inside the clause. Unlike *clld*s, the clitic is always optional. Prosodically, a *clrd* is deaccented or has a reduced accent. Indeed, *clrd*s must always be old and salient in the discourse context (Vallduví 1992, Bott 2007, Brunetti 2009c), and prosodic weakness is precisely a marking of this constraint.[6] In (13), for instance, the frog has been mentioned in the immediately preceding discourse context.

(13) el gat ja l'ha vist, a la granota.
the cat already it$_{cl}$ has seen to the frog
"The cat already SAW the frog" Catalan

When the right dislocated argument has no resumptive clitic, the construction is simply called 'right dislocation' (*rd*). It mostly concerns cases in which the dislocated element is a subject.

A further related construction is 'clitic doubling' (*cldbl*). In terms of word order, *cldbl* is similar to *clrd*. The difference between the two lies in their intonation: while a right dislocated argument is deaccented, with *cldbl* the verb and the argument are in the same intonational unit, the nuclear accent falls on the doubled argument, and either the whole VP or the argument are in focus.

(14) Entonces la tortuga lo ve y se lo dice al NIÑO.
so the turtle it$_{cl}$ sees and to-him$_{cl}$ it$_{cl}$ says to-the boy
"So the turtle sees what happened and tells the boy everything" Spanish

Sometimes the object is separated from the verb by non-argument material, although it still occurs postverbally. We label this construction *obj-sep* ('object separation'). Prosodically, the object is *not* deaccented, which means that it represents the focus or part of the focus, together with the verb.

Dislocations to the left single out topic material. An exception is focus fronting (*focfr*), which we already mentioned in section 4, ex. (2). The distinctive feature of this NOCAN is the prosodic marking of the affected phrase, which is assigned focal accent. In these languages, the

[5] In fact, prosody also contributes to distinguishing between the two constructions.

[6] Within Vallduví's (1992) model of IS, *clrd* and *clld* identify different informational units, called *tail* and *link* respectively.

focal accent canonically falls at the end of the clause, while in this construction it is in sentence initial position. Prosodic marking may not be accompanied by syntactic fronting. That happens for instance when the expression is the subject, as in (15).

(15) A questo punto anche LARA è dispiaciuta
at this point even Lara is sorry
"At this point even Lara is sorry about that" Italian

Finally, we marked deaccented material (*deacc*), namely absence of a pitch accent on a word that would otherwise be expected to be accented (Swerts et al. 2002). Romance languages do not usually recur to *deacc*, although this is very common in other European languages (e.g. English). Nevertheless, we can sometimes find deaccented material in Romance, as in (16). The main accent would be expected to be on *rana*. On the contrary, the accent falls on *simpatica*, and the adjunct is deaccented. *Deacc* cannot be focal material, and it is not usually topic either.

(16) ma Lara non è molto SIMPATICA, con questa rana
but Lara not is very nice with this frog
"But Lara is NOT very nice, towards this frog" Italian

Focfr and *deacc* (including *sbjinv_deacc*, see ex. 5) are the only NOCANs that explicitly (and in the case of *deacc*, exclusively) make reference to *prosodic*, rather than syntactic, non-canonicity.

5.3 Labels marking non-canonical types of sentences

The NOCANs described so far all affect isolated parts of a sentence. However, there is also a series of NOCANs that affect the entire sentence. We will present them in this section.

Presentational sentences (*pres*) are used to introduce new referents and states of affairs. They usually only contain new information. In Italian their most common form is *Locative clitic + verb 'to be' + NP* (see 17); in Catalan the corresponding form is *Locative clitic + verb 'to have' + NP*, and in Spanish they are typically introduced by the impersonal form of *haber* 'to have' (*hay, había, etc.*).

(17) C'era una volta un bambino
there was one time a boy
"Once upon a time there was a boy" Italian

Passive constructions (*pass*) also have a relation to IS. On one hand, the direct object of the active form becomes the subject of the corresponding passive form and therefore occupies a (canonical) preverbal position. Since such position is typically occupied by the sentence topic, passives may favour a topic interpretation of the direct object. Furthermore, in passives the subject of the active form corresponds to an adjunct *by*-phrase, which can be omitted. The main function of passives is in fact to omit or hide the agent of the event, which in the active form is always the subject. The agent may be hidden for various reasons, one being for instance that the speaker ignores its referent. In this sense, passives carry out a similar function as arbitrary subjects (see 9), in that they favour the presence of an indirect object (when given) as sentence topic (Brunetti 2009a).

Impersonal constructions (*impers*) were also annotated. Not even these verbs select an agent, so they are supposed to have similar effects to *arbnsbj*s and passives with respect to topic selection, as argued by Brunetti (2009a).

(18) e lui continua hm a indicare non si sa dove
and he keeps hum to point not IMP knows where
"And he keeps pointing who knows where" Italian

Finally, we annotated three constructions where the sentence is divided into two separated clauses, and each clause typically represents a particular informational unit: cleft sentences (*cleft*), pseudo-cleft sentences (*pscleft*), and inverted pseudo-clefts (*inv-pscleft*). A cleft sentence has the following syntactic form: *Copula verb + XP + Comp + S missing XP*. As already seen in section 4, ex. (3), the XP is the focus (typically a contrastive one), and the remaining clause represents the background. *Pscleft*s are related to cleft sentences. They have the form: *NP + relative clause + Copula verb + NP or S*, but unlike ordinary clefts, they do not mark a focus-background structure: it is the second part of the construction that is in focus instead (cf. *que el barquito se hunde* in 19).

(19) y lo que pasa es que el barquito se hunde
and it that happens is that the little-ship IMPERS sinks
"and what happens is that the boat sinks" Spanish

Finally, *inv-pscleft*s are *pscleft*s that occur in a reversed order, namely the NP follows the copula verb.

6. Corpus exploitation

The corpus and the annotation of NOCANs provide a valuable source of *qualitative* data to be used as examples in theoretical studies on IS (see for instance Brunetti 2009a, Bott 2007). The corpus can obviously also be exploited for *quantitative* analyses of specific informational phenomena (see Mayol 2009, Mayol and Clark in press, Brunetti 2009b). In the following section we present a general overview of the annotation results on the three Romance sub-corpora, and we will propose some possible lines of research that may stem from them.

6.1 An overview of the annotation results

In Table 3 we report the most interesting results concerning the differences in frequency of NOCANs in the three Romance languages. The relative frequencies are given with respect to the total number of finite clauses.[7]

	Catalan		**Italian**		**Spanish**	
overt sbj	1561	**35.7 %**	1262	**38. 9 %**	1027	**35.5 %**
nsbj	1665	**38.1 %**	1173	**36.1 %**	1084	**37.5 %**
arbnsbj	22	**0.5 %**	7	**0.2 %**	32	**1.1 %**
sbjinv	332	**7.6 %**	215	**6.6 %**	265	**9.1 %**
clld+ld	62	**1.4%**	44	**1.35%**	39	**1.,35 %**
clrd+rd	22	**0.5 %**	21	**0.64 %**	11	**0.38 %**
ht	10	**0.2%**	2	**0.06%**	9	**0.3 %**

[7] More precisely, we counted all main clauses and adjunct clauses, and we excluded all non-finite clauses (infinitives, gerunds) and relative clauses.

cldbl	92	**2.1 %**	7	**0.2 %**	61	**2.1 %**
cleft	3		4		2	
pscleft + inv-pscleft	40	**0.9 %**	10	**0.3 %**	37	**1.28 %**
pass	5	**0.1 %**	67	**2 %**	7	**0.24 %**

Table3: Absolute and relative frequencies of some NOCANs with respect to segments.

A first general observation to be made is that the overall number of NOCANs is relatively low. This is indeed expected as NOCANs are *marked* constructions with respect to the linguistic properties of a language. Although spontaneous speech is assumed to have more NOCANs than controlled speech or written text, this does not mean that the number of NOCANs in the former is very high.[8] An exception to low frequency is *nsbj*. Catalan, Italian and Spanish all allow for subject omission in finite clauses. In the corpus, however, *nsbj*s are in practice nearly as frequently as overt ones. Therefore, an empirical support for assuming that these languages *canonically* have overt subjects is rather weak. Although the corpus annotates *nsbj*s as NOCANs, we can actually conclude that they are at least as canonical as overt subjects. The percentages of all NOCANs including *nsbj* is 72.5% (uniformily distributed in the three languages: 72.2% in Catalan, 74.4% in Spanish, 71.1% in Italian). Without *nsbj*, the percentage drops consistently, but not dramatically, to 24.1% (23.57% in Catalan, 26.3% in Spanish, 22.8% in Italian).

Another NOCAN that has a higher frequency than the others, but low enough for the phenomenon to be still undoubtedly considered as non-canonical, is *sbjinv*. Among the factors that determine *sbjinv*, our data show that an important role is played by the type of subject. We found that a high percentage of inverted subjects correspond to the pronoun *tots* (Cat.) / *todos* (Sp.) / *tutti* (It.) 'all'. We also found a strong correlation between certain situations in the storyline and inversion. There are two points in the narrations where nearly all speakers use *sbjinv*. In both cases the character referred to by the subject had been absent from the story for some time and makes a sudden re-appearance. Indeed, in these contexts the subject is in focus and we know that a focused subject tends to occupy a postverbal position whenever possible. The data also confirm the well known relation between *sbjinv* and type of verb. In general, unaccusative verbs (e.g. 'come in', 'fall down', 'appear', etc.) accompany inversion in our data.

Another interesting observation to be made on the annotation results is the rather even distribution of NOCANs among the three languages. This is in itself an interesting fact. The variation among individual speakers is higher than the variation across languages. For example, we observed that 3 out of the total of 5 passives in the Spanish subcorpus were produced by only one speaker. Another good example is impersonals. Their overall frequency in Romance is 2.1%, but we find speakers who do not use impersonals at all, while a small set use them with a high frequency of approximately 5% and one speaker even used it with a frequency as high as 10.3%. We conclude from this that NOCANs are also subject to personal style and variation.

Dislocations also behave in a very similar way in the three languages. This result is rather unexpected, as it has been claimed in literature that these three languages vary with respect to the frequency and use of dislocations. For instance, it has been argued that *clrd*s and *rd*s are much more common in Catalan than they are in Spanish (see e.g. Villalba 2007, who studied right dislocations in a Catalan theatre play and its Spanish translation). The frequencies we obtained from the corpus seem to contradict these conclusions. Upon closer inspection,

[8] Cf. e.g. the work by Carter-Thomas and Rowley-Jolivet (2001) on English written and spoken scientific discourse. These scholars have shown that the frequency of certain NOCANs in a written article is much lower than in the corresponding spoken presentation. What can be deduced from their data, however, is that there is a rather low number of these constructions in both kinds of discourse.

however, we found that 5 out of the 7 *clrd*s and all 4 *rd*s in the whole corpus were produced by only one speaker. In addition to that, the Spanish native speakers we consulted confirmed that all these instances are grammatical but highly marked. These cases in fact look like literal translations from Catalan. So we considered as plausible that the speaker in question, who was born and raised in Catalonia, showed a strong interference from Catalan. If we exclude that speaker, we find that the total of *clrd*s and *rd*s in Spanish only have a frequency of 0.2% (which would confirm Villalba's findings).

Finally, clitic doubling, although syntactically similar to dislocations, constitutes a case apart. We observe a clear difference between Catalan and Spanish on one hand and Italian on the other: the frequency of *cldbl* in the former languages is much higher than in the latter. This result is not surprising, as *cldbl* in Italian has morpho-syntactic restrictions that are totally absent in the other two languages. For instance, in Italian the clitic is only fully accepted if followed by another clitic (see Benincà 1988 [2001]:151).

7. Conclusion and future work

The corpus we have presented in this article is an important resource for the study of information structure and potentially for the study of all phenomena found in spoken narration. The availability of high quality recordings allows the study of phonetic and phonological phenomena. The annotation of NOCANs identifies IS-related constructions within their context of appearance, and allows a quantitative analysis of them.

The developments and extensions of the corpus that we foresee in the near future take several directions. With respect to corpus compilation, we are interested in extending the corpus to other types of discourse, in particular spoken *dialogue*. Such an extension would allow a quantitative and cross-linguistic study of the differences between monologue and dialogue with respect to the use of NOCANs and the organization of information. The cross-linguistic side of the corpus can also be improved by collecting recordings of other languages, both within the Romance family (Portuguese, Romanian) and different language families. With respect to the annotation, an obvious extension concerns the annotation of NOCANs in the Germanic sub-corpora already available (English and German). Further annotations that may contribute to a better understanding of IS-related phenomena are conceivable. A phonological annotation of prosodic grouping and types of accents would offer a more detailed description of prosody-related IS phenomena. Semantic annotations of various kinds – thematic roles, animacy, degree of saliency of a referent in the discourse, etc. – would also contribute to a better understanding of certain phenomena.

As a last remark, it is important to stress that the NOCANDO corpus is publicly available and third parties are allowed (and encouraged) to enlarge and enrich the corpus both in terms of further annotations and of further languages.

References

Alsina, A. (2004): La inversió copulativa en català. In: Anuari de Filologia 26, 9-44.

Baumann, S. (2006): Information Structure and Prosody: Linguistic Categories for Spoken Language Annotation. In: Sudhoff, S. et al. (eds.): Methods in Empirical Prosody Research. Language, Context and Cognition 3. Berlin: Walter de Gruyter.

Benincà, P. (1988 [2001]): L'ordine delle parole e le costruzioni marcate. In: Renzi, L./Salvi, G./Cardinaletti, A. (eds.): Grande grammatica italiana di consultazione. Vol. 1. Bologna: Il Mulino, 129-239.

Berman, R.A./Slobin, D. I. (eds.) (1994): Relating events in narrative: A crosslinguistic developmental study. Lawrence Erlbaum Associates.

Boersma, P./Weenink, D. (2009): Praat: doing phonetics by computer (Version 5.0.47) [Computer program]. Available at: http://www.praat.org/.
Bott. S. (2007): Information Structure and Discourse Modelling. PhD Diss, Universitat Pompeu Fabra.
Brunetti, L. (2009a): On the semantic and contextual factors that determine topic selection in Italian and Spanish. In: van Bergen, G./de Hoop, H. (eds.): Special issue on Topics Cross-linguistically. The Linguistic Review 26, 2/3.
Brunetti, L. (2009b): Discourse Functions of Fronted Foci in Italian and Spanish. In: Dufter, A./Jacob, D. (eds.): Focus and Background in Romance Languages. Studies in Language Companoion Series, Benjamins.
Brunetti, L. (2009c): On links and tails in Italian. In: Lingua 119, 5, 756-781.
Carter-Thomas, S./Rowley-Jolivet, E. (2001): Syntactic differences in oral and written scientific discourse: the role of information structure. In: ASp, la revue du Geras, vol 31, 19-37.
Cinque, G. (1977): The Movement Nature of Left Dislocation. In: Linguistic Inquiry 8, 397-411.
Cinque, G. (1990): Types of A'-Dependencies. Cambridge: MIT Press.
De Mauro, T./Mancini, F./Vedovelli, M./Voghera, M. (1993): Lessico di frequenza dell'italiano parlato. Milano: Etas.
Erteshik-Shir, N. (1997): The Dynamics of Focus Structure. Cambridge: CUP.
Gavarró, A./Laca, B. (2002): Les perífrasis temporals, aspectuals i modals. In: Solà, J. et al. (eds.): Gramàtica del català contemporani. Vol. 3. Barcelona: Empúries, 2665-2774.
Hirschberg, J. (1995): Prosodic and other acoustic cues to speaking style in spontaneous and read speech. In: Proceedings of International Congress on Phonetic Sciences. Vol. 2, 36-43.
Lambrecht, K., (1994): Information structure and sentence form: Topic focus, and the mental representations of discourse referents. New York: Cambridge University Press.
Leonetti, M. (2008): Alcune differenze tra spagnolo e italiano relative alla struttura informativa. In: XVIII Congresso A.I.P.I. Associazione Internazionale Professori di Italiano, Universidad de Oviedo.
MacWhinney, B. (2000): The CHILDES Project: Tools for Analyzing Talk. Voll. 1-2. Mahwah, NJ: Lawrence Erlbaum Associates.
Mayol, L. (2009): Pronouns in Catalan: information, discourse and strategy. Ph.D. Dissertation, University of Pennsylvania.
Mayol, L./clark, R. (in press): Pronouns in Catalan: Games of Partial Information and the Use of Linguistic Resources. In: Journal of Pragmatics.
Reinhart, T. (1981): Pragmatics and Linguistics: An Analysis of Sentence Topics. In: Philosophica, 27, 53-94.
Steedman, M. (2000): Information structure and the syntax-phonology interface. In: Linguistic Inquiry 34, 649–689.
Strawson, P. (1964): Identifying reference and truth-value. Theoria 30, 96-118. Reprinted in: Strawson, P. (1971): Logico-linguistic papers, Methuen, 75-95.
Strömqvist, S./Verhoven, L.T. (eds.) (2004): Relating events in narrative. Vol.2: Typological and contextual perspectives. Lawrence Erlbaum Associate, Mahwah, NJ.
Swerts, M./Krahmer, E./Avesani, C. (2002): Prosodic Marking of Information Status in Dutch and Italian: A comparative analysis. In: Journal of Phonetics, 30(4), 629-654.
Vallduví, E. (1992): The informational component. New York: Garland.
Vallduví, E./Engdahl, E. (1996): The linguistic realization of information packaging. In: Linguistics, 34.
Villalba, X. (2007): La dislocació a la dreta en català i castellà, microvariació en la interfície sintaxi/pragmàtica. In: Caplletra: revista internacional de filología, 42, 273-302.
Zubizarreta, M.L. (1998): Topic, Focus and Word Order. Cambridge, MA: MIT Press.

Zubizarreta, M. L. (1999): Las funciones informativas: tema y foco. In: Bosque, I./Demonte, V. (eds.): Gramática descriptiva de la lengua española. Vol 3, Madrid: Espasa Calpe, 4215-4244.