NOCANDO
A multilingual annotated corpus for the study of Information Structure

Lisa Brunetti\textsuperscript{1}  Stefan Bott\textsuperscript{2}  Joan Costa\textsuperscript{3}  Enric Vallduví\textsuperscript{3}

\textsuperscript{1}Université de Provence, Aix-Marseille I
\textsuperscript{2}Universitat Politècnica de Catalunya
\textsuperscript{3}Universitat Pompeu Fabra

Mannheim, 24. September 2009
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
4. Corpus exploitation
5. Conclusion and outlook
The NOCANDO project
Information Structure and NOCANS
The NOCANDO corpus
Corpus exploitation
Conclusion and outlook

NOCANDO: Construcciones no-canónicas en el discurso oral: un estudio transversal y comparativo
('Non-canonical constructions in oral speech: a crosslinguistic perspective')
Principal investigator: Enric Vallduví
Research group: Grup de Lingüística Formal (GLiF) http://parles.upf.edu/glif/
Dep. Traducció i Ciències de Llenguatge
Universitat Pompeu Fabra, Barcelona
Collaborators: Lisa Brunetti, Stefan Bott, Joan Costa, Estela Puig Waldmüller, Teresa Suñol, Louise McNally, Josep Maria Fontana, Alex Alsina
The project goal:

- NOCANDO seeks to establish a cross-linguistically valid taxonomy of non-canonical constructions (NOCANs) and study the relation between NOCANs and Information Structure (IS)

What is a NOCAN? A morphologically, syntactically and/or prosodically marked construction from the point of view of the properties of a language.

What is the function of a NOCAN? NOCANs optimize the way the informational content of a sentence is conveyed (Vallduví 1992). NOCANs are explicit marks of the INFORMATION STRUCTURE of the sentence.
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
4. Corpus exploitation
5. Conclusion and outlook
Information Structure

Information Structure (a.k.a theme-rheme, focus-background, topic-comment, ...) is like a poltergeist:

- We all know that poltergeists and information structure exist
- but they are only observable through evidence, which is
  - indirect
  and
  - sparse.

- Information-structural units (such as focus, topic, ...) cannot be directly observed.
- The exact inventory of these units cannot be determined by direct observation either (some categories are uncontroversial, others less clear).
What we know about Information Structure

- There is a long tradition of studies (which rely mainly on introspective data)
- Information-structural units (foci, topics, ...) are coded in natural language by means of prosody, syntax and morphology (NOCANs)
- IS highly depends on context
- Some registers (especially spoken text registers) are richer in IS marking than others
Information Structure

What we are interested in

- How do *different*, but related, languages mark IS units with different NOCANS (the crosslinguistic perspective)
- Study IS in spontaneously produced text
- IS marking in *context*
What we (at NOCANDO) assume about Information Structure

- Topic-comment distinction (Topic marks what the sentence is about)
- Focus-background distinction (The focus introduces new or contrastive information)
- There is also non-topic non-focus material (called *tails*, Vallduví 1992)
- Information-structural units are crosslinguistically stable, while their marking is variable
Examples of NOCANs and their function I

Clitic left dislocation

Al to-the hombre man
Ind. Object

"The man...

TOPIC
(Spanish)

drops the coffee”

COMMENT

se le cae el café
RFL to-him falls the coffee
obj.cl. Verb Subject
café

COMMENT
Focus Fronting

Pure

la LINGUACCIA, even

the tongue

Dir. Object

"Even the tongue did the frog put out to him”

FOCUS

(Italian)
Examples of NOCANs and their function III

Cleft sentences

I no se n’adona que la granota s’ha posat a davant
and not realizes that the frog herself has put in front

"And she does not realize that the frog placed himself before the baby"

i és ELLA
and is her

”And it is her

CONTRASTIVE
FOCUS

(Catalan)

que està a punt de prendre’s el biberó
who is about to take for-herself the baby-bottle

BACKGROUND
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
   - Annotated NOCANs
   - Corpus distribution
4. Corpus exploitation
5. Conclusion and outlook
Quantitative information

Spontaneous narrations in Catalan, Italian, Spanish, German, and English

Total number of speakers: **68**
Total number of narrations: **222**
Total duration: ca **16 h** (2'-10' per narration)

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Recording time</td>
<td>4:02:43 h</td>
<td>4:04:32 h</td>
<td>2:35:20 h</td>
<td>2:09:13</td>
<td>2:32:20 h</td>
</tr>
<tr>
<td>Word count (linux/cygwin)</td>
<td>37555 w</td>
<td>27392 w</td>
<td>25077 w</td>
<td>15944 w</td>
<td>21970 w (es)</td>
</tr>
<tr>
<td>Segment count</td>
<td>5856 seg</td>
<td>4306 seg</td>
<td>3801 seg</td>
<td>2154 seg</td>
<td>3140 seg (es)</td>
</tr>
</tbody>
</table>
Mostly students at the Universitat Pompeu Fabra in Barcelona. A smaller number from different working environments.

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Spanish</th>
<th>Italian</th>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical origin</td>
<td>Catalonia (except one from Comunitat Valenciana)</td>
<td>Catalonia (except one from Castilla y León)</td>
<td>Different parts of Italy</td>
<td>Different parts of Germany</td>
<td>Different parts of USA and UK</td>
</tr>
<tr>
<td>Mean age</td>
<td><strong>22</strong> (18-30)</td>
<td><strong>20</strong> (17-29)</td>
<td><strong>29</strong> (20-56)</td>
<td><strong>34</strong> (22-67)</td>
<td><strong>27</strong> (20-41)</td>
</tr>
</tbody>
</table>
Corpus compilation and methodology

Speakers narrated the stories of three text-less picture story books by Mercer Meyer: *Frog goes to dinner*, *Frog on his own*, *One frog too many*.

a new frame

- Speakers told the experimenter the stories in a random order. Speakers could browse the book before starting the narration.
- A questionnaire was filled by speakers concerning age, geographical origin, personal language history.
Transcription and segmentation

Orthographic transcription based on the LIP corpus (De Mauro et al. 1993).

<table>
<thead>
<tr>
<th>Truncated phrase or sentence</th>
<th>la tortug-- la granota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pauses</td>
<td>#, ##, ###</td>
</tr>
<tr>
<td>Unintelligible words</td>
<td>[?], [?], [???]</td>
</tr>
<tr>
<td>Missing part</td>
<td>[...]</td>
</tr>
<tr>
<td>Reconstructed broken word</td>
<td>sta[te]</td>
</tr>
<tr>
<td>Truncated word</td>
<td>-pe-</td>
</tr>
<tr>
<td>Vocalic lengthening at the end of a word</td>
<td>bueno_</td>
</tr>
<tr>
<td>Extra-linguistic comment</td>
<td>[LAUGHS]</td>
</tr>
<tr>
<td>Phonetic symbol</td>
<td>[fff]</td>
</tr>
<tr>
<td>Hesitation</td>
<td>hm</td>
</tr>
<tr>
<td>Standard phrasing symbols:</td>
<td>(,) (.) (?) (!)</td>
</tr>
</tbody>
</table>

**Segmentation**

One clause - one line

- subordinate clauses included (cf. CHILDES).
- Temporal/aspectual and modal verb periphrases excluded. (Criteria to identify periphrases: Gavarró and Laca 2002).
Corpus annotation


Why these languages?

- Similar linguistic properties (relatively free word order, null sbj, SVO, Obj Cl Pro..)

- Similar strategies to express informational notions:
  - large use of syntax
  - limited use of prosody (as opposed to e.g. English, Vallduví and Engdhal 1996)

- Similar or identical NOCANs

However, NOCANs of these languages may vary in terms of:

- frequency (e.g. clrd in Catalan vs Spanish, Villalba 2007, Leonetti 2008)
- function (e.g. subject inversion in Spanish vs Italian)
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
   - Annotated NOCANs
   - Corpus distribution
4. Corpus exploitation
5. Conclusion and outlook
Corpus annotation: Subject oriented NOCANs

- **sbjinv** = subject inversion (the subject occurs after the verb)
  
  \[(1)\] Els va acompanyar el taxista.  
  them PAST take the taxi-driver  
  ’The TAXI-DRIVER drove them’ (Catalan)

- **sbjinv_deacc** = post-verbal deaccented subject in a declarative sentence

  \[(2)\] ...que està disfressat, aquest nen.  
  for is dressed-up this child  
  ’...for this child is dressed up’ (Catalan)
Corpus annotation: Subject oriented NOCANs

- nsbj = null subject (the subject is not expressed)

  (3) Invece il bambino è molto contento, perché ha salvato la sua rana.

  'The boy on the contrary is very happy, because he saved his frog' (Italian)

- nsbj_c = null subject in a coordinate clause
Corpus annotation: Subject oriented NOCANs

- arbnsbj = arbitrary subject (which is null in these languages, cf. Jaeggli 1986)

(4) Y un día a este niño le regalaron pues una caja muy grande
’And one day this boy received a large box’
literally: ’And one day they gave the boy a large box as a present’ (Spanish)
Corpus annotation: Subject oriented NOCANs

• sbj-sep = separation of the (preverbal) subject from the verb by sentential adverbials or other material that has scope over the entire clause

(5)  però la rana come al solito riesce a infilarsi nella _ nella situazione
  'but the frog, as usual, manages to sneak herself into-the into-the situation'

  (Italian)
Corpus annotation: Subject oriented NOCANs


  \[(6) \quad \text{Al hombre se le cae el café.} \quad \text{to-the man RFL to-him drop the coffee} \]

  "The man drops the coffee" (Spanish)

- **ld = left dislocation.** Dislocation to the left WITHOUT clitic resumption

  \[(7) \quad \text{A un bambino un giorno arriva un regalo.} \quad \text{to a boy one day arrives a present} \]

  'One day a boy receives a present'
Corpus annotation: Subject oriented NOCANs


\[8\] La rana grande, la situación no le gusta mucho.

’As for the big frog, she didn’t like the situation at all’ (Spanish)
Corpus annotation: Subject oriented NOCANs


(9) el gat ja l’ha vist, a la granota
    the cat already it-has seen to the frog
    ’The cat already SAW the frog’ (Catalan)
Corpus annotation: Subject oriented NOCANs

- $rd =$ right dislocation. Dislocation to the right WITHOUT clitic resumption

$y \ le_\_ \ muerde_\_ [\ldots] el \ anca \ a \ la \ otra \ and \ to-him \ he-bites \ the \ hip \ to \ the \ other \ ranita, \ la \ rana \ grande. \ little-frog, \ the \ frog \ big \ ' \ and \ the \ big \ frog \ bites \ the \ other \ frog's \ hip' \ (Spanish)$
Corpus annotation: Subject oriented NOCANs

- cldbl = clitic doubling. It differs from clrd in that the doubled argument is in the same intonational contour as the verb (focus domain).

(11) Entonces la tortuga lo ve y se lo dice al niño.

'then the turtle it sees and to-him it says to-the boy
'So the turtle sees what happened and tells the boy everything’ (Spanish)
Corpus annotation: NOCANS concerning all arguments

- obj-sep = separation of the (postverbal) direct object from the verb

(12) Y cogió en su mano a la ranita pequeña and he-took in his hand to the frog little
’And he took the little frog into his hands’ (Spanish)
Corpus annotation: NOCANS concerning all arguments

- narg = null argument

(13) i allavors en Jaume es va adonar que and then the Jaume RFL PAST realizes that que, home, era la seva granota that well it-was the his frog
Áand therefore Jaume realizes that that, well, it was his frog’ (Catalan)
Corpus annotation: NOCANS concerning all arguments

- **focfr = focus fronting:** A left peripheral element with focal stress (Benincà 2001, Rizzi 1997, Zubizarreta 1998)

  (14) veu una dona que està amb un cotxet; sees a woman who is with a baby-carriage així de LLUNY la veu. like-that from far her he-sees 'He sees a woman with a baby carriage; she sees it from far away’ (Catalan)
Corpus annotation: NOCANS concerning all arguments

- deacc = de-accenting

(15) 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)
Corpus annotation: NOCANS concerning sentence types

- pres = presentational sentences

(16) C’era una volta un bambino
    there was one time a boy
    ’Once upon a time there was a boy’ (Italian)
pass = passive construction

(17) la familia de William es expulsada del restaurante.

’William’s family is expelled from the restaurant’ (Spanish)
Corpus annotation: NOCANS concerning sentence types

- impers = impersonal construction

(18) e lui continua hm a indicare non si sa and he keeps hm to point not IMP knows dove.
where
'And he keeps pointing who knows where’ (Italian)
Corpus annotation: NOCANS concerning sentence types

- cleft = cleft sentences. Construction: Verb 'to be' + XP + 'that' S without XP

  (19) Era ese SAPO que les había querido hacer algún susto
       was this frog that them had wanted make some scare
       'It was the frog who wanted to scare them' (Spanish)
Corpus annotation: NOCANS concerning sentence types

- pscleft = pseudo-cleft sentences. Construction: Dem. pron. + Rel. clause + verb 'to be' + 'that' S / NP

\[(20) \quad y \quad lo \quad que \quad pasa \quad es \quad que \quad el \quad barquito \quad se \quad hunde\]

\[\text{'and what happens is that the boat sinks'} \quad (\text{Spanish})\]
inv-pscleft = inverted pseudo-cleft sentences. Construction:
NP + verb 'to be' + Dem. pron. + Rel. clause

(21) Y bueno, el niño es el que dirige la balsa
and well the boy is he that directs the raft
’and well the boy is the-one who leads the raft’ (Spanish)
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
   - Annotated NOCANs
   - Corpus distribution
4. Corpus exploitation
5. Conclusion and outlook
Erano stanchi della città,
avevano bisogno di un poco di natura, di aria fresca,
di camminare in mezzo agli alberi...
E così, tutti contenti, uscirono di casa  
e andarono verso il bosco.
Michelino aveva messo la rana e la tartaruga in un secchiello,  
il cane nel secchiello non c’entrava, ovviamente.
Cammina cammina,  
a un certo punto la rana ne approfitta di un momento di distrazione, di Michelino,
Alignment audio-transcription-annotation
The corpus distribution

The NOCANDO corpus is available online under a Creative Commons license (http://nocando.barcelonamedia.es)

The project

NOCANDO seeks to establish a crosslinguistically taxonomy of noncanonical constructions (NOCANs). The languages studied and compared are Catalan, Spanish, Italian, English, and German.

The NOCANDO Corpus is a corpus of spoken narrative text. It was created by recording free picture-based narrations of native speakers in the languages mentioned above. The texts were transcribed, annotated and aligned, using the Prat software.

At the moment only parts of the corpus are available in annotated and aligned form.

The corpus can be found here.

If you want to use our data, please quote this webpage, and respect the constraints indicated in the link below:

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Spain License.
The distribution consists of audio files (mp3), transcription, annotation and alignment files (when available).
Outline

1. The NOCANDO project
2. Information Structure and NOCANS
3. The NOCANDO corpus
4. Corpus exploitation
5. Conclusion and outlook
Corpus exploitation

- Naturally occurring data for theoretical studies on information structure and discourse (cf. Bott 2007, Brunetti 2009a,b, Mayol 2009).
- Cooccurrence of NOCANs with particular linguistic environments or with other NOCANs (cf. Brunetti 2009a).
- Comparison among Romance languages.
## Corpus exploitation

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>overt sbj</td>
<td>1561 35,7 %</td>
<td>1262 38,9 %</td>
<td>1027 35,5 %</td>
</tr>
<tr>
<td>nsbj</td>
<td>1665 38,1 %</td>
<td>1173 36,1 %</td>
<td>1084 37,5 %</td>
</tr>
<tr>
<td>arbnsbj</td>
<td>22 0,5 %</td>
<td>7 0,2 %</td>
<td>32 1,1 %</td>
</tr>
<tr>
<td>sbjinv</td>
<td>332 7,6 %</td>
<td>215 6,6 %</td>
<td>265 9,1 %</td>
</tr>
<tr>
<td>clld+ld</td>
<td>62 1,4%</td>
<td>44 1,35%</td>
<td>39 1,35%</td>
</tr>
<tr>
<td>clrd+rd</td>
<td>22 0,5 %</td>
<td>21 0,64 %</td>
<td>11 0,38 %</td>
</tr>
<tr>
<td>ht</td>
<td>10 0,2%</td>
<td>2 0,06%</td>
<td>9 0,3 %</td>
</tr>
<tr>
<td>cldbl</td>
<td>92 2,1 %</td>
<td>7 0,2 %</td>
<td>61 2,1 %</td>
</tr>
<tr>
<td>cleft</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>pscleft</td>
<td>40 0,9 %</td>
<td>10 0,3 %</td>
<td>37 1,28 %</td>
</tr>
<tr>
<td>+ inv-pscleft</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pass</td>
<td>5 0,1 %</td>
<td>67 2 %</td>
<td>7 0,24 %</td>
</tr>
</tbody>
</table>
53.89% of all segments contain a NOCAN

The overall frequencies for NOCANs in Romance languages are *surprisingly* similar.
There is a considerable variation in the use of different NOCANS among speakers of the same language. Some examples:

- 3 out of 7 passives in Spanish (a marked construction) were produced by only 1 speaker.
- Some NOCANs, like pseudo-clefts and inverted pseudo-clefts were only produced by less than half of the speakers, but some speakers use them relatively frequent (in up to 0.65% of the segments in the case of pscleft and inv-pscleft).
- The use of impersonal constructions ranges from 0 to 4.95% (all Romance languages).
- Deaccenting in Italian is very rare in Italian, but one speaker used it very commonly (13 out of 38 deaccentuations were produced by this speaker).

⇒ Some NOCANs are subject to strong rhetorical variation. Others are more evenly distributed.
Null subjects (nsbj) are nearly as frequent as overt subjects. → Null subjects are probably not non-canonical.

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>overt sbj</td>
<td>1561</td>
<td>1262</td>
<td>1027</td>
</tr>
<tr>
<td>nsbj</td>
<td>1665</td>
<td>1173</td>
<td>1084</td>
</tr>
</tbody>
</table>

- 35.7%  - 38.9%  - 35.5%
- 38.1%  - 36.1%  - 37.5%
Null subjects (nsbj) are nearly as frequent as overt subjects. → Null subjects are probably not non-canonical.

Mayol (2009) concludes (on the basis of Catalan NOCANDo data) that the choice between null subjects and subject pronouns is partly dependent on IS, but partly the choice is dependent on purely syntactic factors (antecedent is subject or direct object).
Subject inversion has a low percentage, but is still one of the most frequent NOCANs. The high frequency of subjinv in Spanish is partly due to syntactic effects. A postverbal subject cannot serve as a sentence topic (and is usually focal).

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>subjinv</td>
<td>332</td>
<td>215</td>
<td>265</td>
</tr>
<tr>
<td>%</td>
<td>7.6 %</td>
<td>6.6 %</td>
<td>9.1 %</td>
</tr>
</tbody>
</table>
Subject inversion has a low percentage, but is still one of the most frequent NOCANs. The high frequency of subjinv in Spanish is partly due to syntactic effects. A postverbal subject can not serve as a topic and are usually focal.

Mayol (2009) concludes that subjects are preferred as background material (mostly topics). She found that only 1,66% of the segments in the Catalan corpus are either postverbal or directly marked as focused. → There is a strong relation between subjects and topichood.

74,72% of the focal subjects in the Catalan corpus were discourse new (Mayol 2009).
Surprisingly, and contra previous assumptions (e.g., Villalba 2007), Catalan and Spanish behave very similar with respect to right dislocations. This may be a consequence of the text genre.

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clld + ld</td>
<td>62</td>
<td>1,4%</td>
<td>44</td>
</tr>
<tr>
<td>Clrd + rd</td>
<td>22</td>
<td>0,5%</td>
<td>21</td>
</tr>
<tr>
<td>ht</td>
<td>10</td>
<td>0,2%</td>
<td>2</td>
</tr>
</tbody>
</table>
There is a clear difference between Catalan/Spanish vs Italian concerning clitic doubling: cldbl is considered as highly marked for colloquial speech in Italian. The difference between Italian and the other two languages is presumably more syntactic in nature than informational.

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>cldbl</td>
<td>92</td>
<td>7</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>2.1 %</td>
<td>0.2 %</td>
<td>2.1 %</td>
</tr>
</tbody>
</table>
Passives are extremely more common in Italian than in Spanish or Catalan. In Spanish/Catalan passives are traditionally considered as a marked construction (recommended by some normative grammars only in restricted contexts). These languages use other constructions instead, like e.g. arbitrary subjects.

<table>
<thead>
<tr>
<th></th>
<th>Catalan</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>pass</td>
<td>5</td>
<td>0,1 %</td>
<td>67</td>
</tr>
</tbody>
</table>
Conclusions

- The NOCANDO corpus is a valuable resource for the study of information structure in spoken register.
- It is potentially valuable for a wider study of spoken register in the covered languages.
- Some previous assumptions about IS and IS marking can be confirmed with quantitative data, some findings raise interesting questions.
- The NOCANDO corpus is also a good source for spontaneous data for qualitative studies.
- It allows to study IS marking in linguistic context.
- The NOCANDO corpus can be both exploited and extended by third parties.
Future work

- Extension of the corpus
- Collection and annotation of a corpus of dialogues in the same languages
- Extension of the corpus to other languages
- Annotation of NOCANs in Germanic languages
We would like to thank: Estela Puig Waldmüller, Teresa Suñol, Josep Maria Fontana, Louise McNally, Gemma Boleda, Alex Alsina
Thanks

Thank you very much!
Selected references


- Bott. S. 2007 Information Structure and Discourse Modelling. PhD.Diss, UPF.


Selected references


