Annotating Information Structure in Italian: Characteristics and Cross-Linguistics Applicability of a QUD-Based Approach

Anonymous ACL submission

Abstract

We present a discourse annotation study, in which an annotation method based on Questions under Discussion (QuD) is applied to Italian data. The results of our inter-annotator agreement analysis show that the QUD-based approach, originally spelled out for English and German, can successfully be transferred cross-linguistically, supporting good agreement for the annotation of central information structure notions such as focus and non-at-issueness. Our annotation and inter-annotator agreement study on Italian authentic data confirms the cross-linguistic applicability of the QuD-based approach.

1 Introduction

In this paper, we present a discourse annotation study of Italian data, which uses the annotation scheme and discourse-analytic method, the QUD-tree framework, developed in Reyle and Riester (2016), Riester (to appear) and Riester et al. (2018). Its purpose is the cross-linguistic analysis of information structure and discourse structure of textual data. On the theoretical side, the QUD framework has been applied to a number of different languages, such as German, English and French in (Riester et al., 2018), and various Austronesian languages as discussed in Riester and Shiohara (2018) and Latrouite and Riester (2018). On the applied side, De Kuthy et al. (2018) showed that the QUD based method supports the successful annotation of discourse structure and information structure in German and English spoken language data. Here we want to broaden the cross-linguistic scope of the QUD framework and apply it to another Romance language, Italian. We will explore both the QUD annotation and the information structure annotation including all information structure labels that are part of the annotation scheme proposed in Riester et al. (2018), such as focus, background, contrastive topic, nai and topic. Topic is regarded as a notoriously difficult label in agreement studies (cf. Cook and Bildhauer (2013); Ritz et al. (2008)). While the results of our study show that the question-based annotation method supports the successful annotation of discourse structure and of information structure, in particular focus, we will also discuss, using the example of topic, some shortcomings of the QUD based annotation method.

2 The QUD framework

The QUD framework introduced in Riester et al. (2018) presents an explicit method for the reconstruction of QUDs which are usually only discussed as an abstract theoretical term. The center of the QUD framework is a compact representation format for QUD trees, in which the textual assertions (A) represent the terminal nodes of a discourse tree (preserving the linear order of the text from left to right) while (implicit or explicit) QUDs (Q) form the non-terminal nodes. An abstract QUD tree is shown in Figure 1.

![Figure 1: QUD tree](image)

The QUD-tree framework as spelled out in Riester et al. (2018) can be applied to any kind of written or spoken discourse or conversation. It is not language-specific and can, in principle, be
used in order to investigate data from any language. While the exact analysis procedure is described at great length within the guidelines document (Riester et al., 2018), we just briefly introduce some basic principles here.

2.1 Segmentation

Raw texts are segmented into atomic assertions. Apart from orthographic sentence boundaries, segmentation also applies at (1) (information-structurally relevant) coordinations and (2) before (optional) syntactic adjuncts. (Obligatory) sentential arguments (3) are not split off.

(1) A1: Ho appena terminato un romanzo
   I-have just finished a novel
   A1: e sono già al lavoro su un nuovo
   and am already at work on a new
   progetto, ...
   project

(2) A7=Di recente ho ripreso a leggere i
   of recently I-have re-started to read the
   romanzi di formazione.
   novels of coming-of-age
   A7=senza mai tralasciare la narrativa
   without ever neglecting the contemporaray
   contemporanea e i romanze.
   fiction and the romances

   Alex is result of-the my imagination
   A25=r:[nasce in relazione a Dave]R ,
   he-is-born in relation to Dave
   A26:=[[ho voluto che]N possa [“forte” ma non
   I-have wanted that he was strong but not
   invincibile]T].
   invincible

2.2 QUD principles

The actual identification of a QUD for each assertion is guided by a number of explicit principles adapted from the formal literature on information structure (Rooth, 1992; Schwarzschild, 1999; Büring, 2008, 2016), cf. Riester et al. (2018):

Q-A-CONGRUENCE: QUDs must be answerable by the assertion(s) that they immediately dominate.

Q-GIVENNESS: Implicit QUDs can only consist of given (or, at least, highly salient) material.

MAXIMIZE-Q-ANAPHORICITY: Implicit QUDs should contain as much given (or salient) material as possible.

Example (4) shows that from these principles we can derive QUD Q32 for assertion A32 in the context of A31, whereas any of the questions in (5), used in place of Q32, would violate at least one of the QUD constraints in the same context.

(4) A31: Anche tra i bilingui precoci che
   even among the early bilinguals who
   parlano due lingue quasi mai le due
   speak two languages almost never the two
   lingue sono del tutto equivalenti,
   languages are completely equivalent
   Q32: {What about the two languages
   instead?}
   > A32: and [normalmente]N [ogni lingua]T and
   usually each language
   [si sviluppa in un contesto specifico]R ~
   itself develops in a context specific
   (5) a: [What about speaking two languages?]
   (#Q-A-CONGRUENCE)
   b. {What about the specific context?}
   (#Q-GIVENNESS)
   c. {What happens next?}
   (#MAXIMIZE-Q-ANAPHORICITY)

Two or more assertions are defined as parallel if and only if they share some semantically identical content and represent partial answers to the same QUD, see Example (6), where the semantically shared content is Alek (omitted in the second assertion).

PARALLELISM: The background of a QUD with two or more parallel answers consists of the (semantically) common material of the answers.

(6) Q25: {What about the connection with re-
   ality in Alek?}
   > A25: [[Alek]T [è frutto della mia
   Alex is result of-the my
   fantasia]T] ,
   imagination
   > A25=r: [nasce in relazione a Dave]R ,
   is-born in relation to Dave

The resulting tree structure is shown in Figure 2.

Figure 2: Two coordinated (parallel) assertions.

3 QUDs and information structure

The basis of our annotation approach is an alternative-based definition of information structural categories, in line with e.g. Rooth (1992), Büring (2003), Beaver and Clark (2008), Krifka
empirical study, in which annotators followed the QUD guidelines described in Riester et al. (2018) to annotate two Italian blog interviews.

For the QUD-based annotation we use the tool TreeAnno introduced by De Kuthy et al. (2018), which enables the analyst to semi-automatically segment texts, systematically enhance them with implicit Questions under Discussion (QUDs), and transform the data into a discourse tree called QUD tree, as described in De Kuthy et al. (2018).

4 Evaluation: Discourse structure

In the present annotation study based on the above described QUD framework, our goal is to show that the discourse annotation in terms of QUDs can be applied reliably to naturally occurring data - in this particular case, Italian data. We conducted an empirical study, in which annotators followed the QUD guidelines described in Riester et al. (2018) to annotate two Italian blog interviews.

For the QUD-based annotation we use the tool TreeAnno introduced by De Kuthy et al. (2018), which enables the analyst to semi-automatically segment texts, systematically enhance them with implicit Questions under Discussion (QUDs), and transform the data into a discourse tree called QUD tree, as described in De Kuthy et al. (2018).

4.1 Evaluation setup

Two trained annotators (and also native speakers of Italian) analyzed and annotated two short Italian blog interviews downloaded from the internet¹. The first blog interview consists of 95 text segments, the second one of 113 segments. The QUD discourse tree for Blog 1 resulting from the first annotator is shown in Figure 3, the other three discourse trees are uploaded as separate files as part of the Appendix.

4.2 Method and results

For the comparison of the two annotated documents, we follow the method described in De Kuthy et al. (2018). The basic idea is that for the comparison of two QUD annotations one needs to calculate an inter-annotator agreement score that takes into account, for every segment and every possible span of segments, whether a QUD is present or not. In order to compute a \( \kappa \) statistics (Cohen, 1960) based on our QUD annotations, De Kuthy et al. (2018) propose to follow the method described in Marcu et al. (1999), which was developed for measuring agreement in the labelling of rhetorical structure categories in texts. The method is based on the idea of mapping the hierarchical structure of a discourse tree onto a matrix or chart filled with categorical values. In our case, the values are whether there exists a (Q)uestion spanning the respective segments – start to end – or (n)ot.

A \( \kappa \) statistics can then be computed between two charts that represent two different QUD annotations for the same text; more precisely between the two resulting sets of cells in the upper half of each chart. For our two annotated documents we calculated \( \kappa \) values for the annotation charts derived from our QUD annotations, based on the

above described method. For the text Italian Blog 1, consisting of 95 segments, we calculated the $\kappa$ statistics based on 4,256 items, for Italian Blog 2 with 113 segments based on 5,350 items. The results are shown in Table 2.

<table>
<thead>
<tr>
<th>Text</th>
<th>Segments</th>
<th>Cells</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian Blog 1</td>
<td>95</td>
<td>4,256</td>
<td>.61</td>
</tr>
<tr>
<td>Italian Blog 2</td>
<td>113</td>
<td>6,187</td>
<td>.51</td>
</tr>
</tbody>
</table>

Table 2: Kappa values for QUD-annotated Italian dialogues

The values show moderate agreement between the annotator pairs. For Blog 1, the $\kappa$ value is at .61, which is substantially higher than what (De Kuthy et al., 2018) report for the QUD annotations of their German and English texts: their $\kappa$ values are around .5. For our Blog 2, the $\kappa$ value is at .51, which is thus very similar to the scores reported in (De Kuthy et al., 2018) for texts of similar length.

Our two annotated Italian texts are relatively short, only around 100 sentences each, so it is perhaps too early to interpret the results, in particular since this is a rather complex task. However, since the results are comparable to those reported in (De Kuthy et al., 2018), we take this as a further proof that the QUD-based annotation of discourse can successfully be applied cross-linguistically.

5 Evaluation: Information structure

The second major issue we are interested in is to evaluate the reliability of information-structure annotation based on the previous identification of QUDs.

5.1 Evaluation setup

For the evaluation of the information structural annotation, the same two Italian blog texts were annotated by the same two trained annotators, who still followed the guidelines of Riester et al. (2018). We aimed at annotating all five categories that are mentioned in Riester et al. (2018): focus (F), background (BG), non-at-issue material (NAI), contrastive topic (CT) and topic (T). Focus domain labels (~) were not annotated, since each text segment (assertion) already corresponds to one focus domain. The annotators based their annotations on the previously performed QUD analysis in the TreeAnno tool. As an annotation tool for the token-based information-structure annotation, WebAnno (Yimam et al., 2013) was chosen. Figure 4 shows a screenshot of the information-structure annotation of the beginning of Blog 1.

5.2 Method and results

As agreement measure for the evaluation of the information structure annotation, we calculated $\kappa$ values on the annotated texts based on tokens, following previous work (Ritz et al., 2008; Calhoun et al., 2010; De Kuthy et al., 2018). In addition to the specifications in Riester et al. (2018), in particular the QUD-to-information-structure mapping from Table 1, we defined a number of heuristic (but potentially debatable) rules in order to prevent disagreement due to theoretically unclear issues, such as:

- Discourse connectors (but, and, although, because, therefore etc.) at the beginning of discourse segments are not annotated.
• Punctuation: Quotation marks around an expression, commas within and at the right edge of an expression are part of the markable. Periods, colons, semicolons, exclamation marks are not.

Results are shown in Table 3, divided into scores for all labels taken together, and individual scores for each of the four labels.

<table>
<thead>
<tr>
<th>Text</th>
<th>Label</th>
<th>Tokens</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian Blog 1</td>
<td>all</td>
<td>847</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>BG</td>
<td></td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td></td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>NAI</td>
<td></td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td>.45</td>
</tr>
<tr>
<td>Italian Blog 2</td>
<td>all</td>
<td>1243</td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>BG</td>
<td></td>
<td>.1</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td></td>
<td>.1</td>
</tr>
<tr>
<td></td>
<td>NAI</td>
<td></td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td>.35</td>
</tr>
</tbody>
</table>

Table 3: Kappa for information structure annotation

The results are rather heterogeneous in both texts but overall they show that the QUD-based method does contribute to a successful annotation of information structure in Italian for a range of labels. For the first text Blog 1, the overall agreement score for all annotated categories taken together is at .7, which shows substantial agreement, the score for focus annotation alone being at .72. The agreements scores for the second blog are overall lower, but with .58 for the overall agreement and .51 with agreement for focus they are still at a relatively high level and still comparable to the scores that (De Kuthy et al., 2018) report for the annotation of German and English data (which are at around .65). The category NAI, the classification of non-at-issue material, also received reasonable agreement scores at .53 in Blog 1 and .62 for Blog 2. The agreement scores for the other three categories, BG and CT, differ a lot between the two texts. In Blog 1, the score for contrastive topic is very high with .85, in Blog 2 the score .1 shows that there was hardly any agreement between the two annotators. This might be due to the fact that there were only very few cases for which the label CT was used. In Blog 1, the label CT was used for 9 and 12 tokens in the two annotations, in Blog 2 it was assigned to 13 and 14 tokens (out of 1243 tokens). The case is similar with respect to background: in the two annotated documents, the label BG was only assigned for around 40 tokens in Blog 1 and 30 tokens in Blog 2. This means that, if the annotators disagreed in only one token when assigning the label CT or BG, this had a much greater impact on the agreement scores for these labels than in the cases of disagreement for assigning focus labels. The category topic (T) received relatively low agreement scores at .45 and .35, but still at a level which other studies report for categories like focus (cf. Ritz et al. 2008 report a \(\kappa\) of .44 for focus). In the following section we will qualitatively evaluate why the annotation scheme seems to better support the successful annotation of a category like focus, whereas there seems to be much more disagreement when annotating topic.

6 Qualitative Evaluation: The Case of Topics

In the question-based definitions of our information structure labels, the focus corresponds to those parts of an assertion that answer the current QUD. Especially in case of overt questions, but with implicit QUDs, the annotators agree on focus.

The definition of topic in the QUD framework, however, is the only one that does not take the current QUD into account. As remarked by Riester et al. (2018), while potentially all referential expressions inside the background could be labelled as topic, one might argue that not all referential expressions inside the background are actually aboutness topics. But unfortunately, the QUD method is not meant to single out the best topic candidate. And Riester et al. (2018) do not provide any rules that help to distinguish between better and worse topic candidates. The only cue that is given through the current QUD is that all focal expressions are excluded as topic candidates.

A typical topic expression in Italian would be a clitic left or right dislocated phrase (see quel libro below), but no dislocation was present in our data, probably due to the fact that a blog interview is less interactive than an spoken conversation, and these construction are typically used in interaction.

(8) a. Quel libro, l’ho dato a Giorgio.

b. L’ho dato a Giorgio, quel libro.
Clitic personal pronouns, such as le in A2 in (9), are also typical candidates for (continuing) topics.

(9) A1: Abbiamo fatto quattro chiacchiere con Maria Verdiana Rigoglioso per parlare di Senza Etichette, il romanzo che ha pubblicato con Libromania. 'We had a chat with Maria Verdiana Rigoglioso to talk about Senza Etichette, the novel she published with Libromania.'

Q3: {What did you do with her exactly?}

> A2: [[Le]T [abbiamo fatto un po’ di domande] to-her we-have made a little of questions

> Q3: {What for?}

> A3: [per [conoscere retroscena e curiosità del to know ins-and-outs and trivia of-the romanzo]T].

What about cases where the topic is neither a dislocated expression, nor a clitic? Our annotation method should be able to single out such cases, but this is not always true. The example above nicely illustrates a case where our annotators disagreed about labelling a given referential expression as topic: the PP del romanzo in A3, which is already introduced in the previous sentence, A1. One annotator chose to nevertheless include it in the focus and label A3 as an all-focus assertion. The other annotator, while annotating a similar QUD, chose to label the PP as a topic. Indeed, strictly speaking, this given PP should then also be part of the QUD ("What for, with respect to the novel?'").

(10) Q3: What for with respect to the novel?

> A3: [[[conoscere retroscena e curiosità del to know ins-and-outs and trivia of-the romanzo]T].

It may be observed that the PP del romanzo is embedded inside the verb’s direct object NP. Although informational categories are in principle independent from the syntactic structure, we may suppose that it was the syntactically embedded position of this expression that led one annotator to consider it as part of the focus (which is the first part of the NP retroscena e curiosità del romanzo). The relationship between the given-new structure and the syntactic structure has not been discussed by Riester et al. (2018), but it is something that might be worth addressing in the future.

In other cases, the topic was well identified by both annotators, such as le due lingue in (11).

(11) A1: [Spesso]NAI [si pensa]NAI [che sia bilingue solo [chi è stato esposto a due lingue only who has been exposed to two languages since earliest infancy]

Q1.1: {One thinks that bilinguals are those who do what, with such two languages?}


perfect and equivalent

In this example syntax does not help to identify the topic status of the direct object le due lingue. Such expression is mentioned in A1 as part of the focus, but instead of being promoted to topic in the subsequent utterance by some syntactic device for topic shift (such as left dislocation, cf. Lambrecht 1994), it is left in situ. One reason for the speaker’s choice may be the fact that the topic expression is inside a free relative, a construction that seems to be incompatible with dislocations, as the unacceptability of examples below shows:

(12)a. ??Chi l’italiano, lo conosce sa bene who the italian it knows dove sta l’errore. well where is the mistake

b. ??Ho dato un bel voto a chi il primo I-have given a good note to whom the first esercizio, lo ha fatto bene. exercise it they-have done well

Since due lingue is mentioned in the previous sentence, the context tells us that this expression is clearly background. Since it’s a referential expression, it has all that is required to be identified as topic. Note that a clitic pronoun might have been acceptable here (see example (13)), but this option is not chosen by the speaker/writer.


equivalent

The mechanism of identifying parallel structures (multiple answers to the same question) is a strategy that our annotation tool provides to help recognizing ‘hidden’ topics.

(14) A53: I genitori dovrebbero lasciare spazio the parents should leave space al bambino o bambina che to-the boy or girl which c’è in loro there is in them

Q54: {To do what?}
Cases of topic shift were easily recognized by the two annotators. One example is given below in (15). The referent la mamma che parla la lingua minoritaria per crescere i suoi bambini bilinguali is introduced in the overt question Q24.1 and then it continues as topic in the answer A24.1. Then the topic changes and becomes i bambini in A25. In A26, the topic changes back to la mamma madrelengua.

(15)Q24.1: La mamma che parla la lingua the mother who speaks the minority
minoritaria per crescere i suoi language to raise the her
bambini bilinguali, cosa fa? bilingual children, what she does

> A24.1: [Parla la propria lingua ai she-speaks the-her own language to-the
figli, lei.] children'

> Q25: {What do the children do?}

> A25: Solo che [molto spesso] [i only that very often the
bambini] [pur capendola children even understanding-her
perfettamente] [non parlo perfectly not speak
attivamente la sua lingua] [actively the her language

> Q26: {What can the mother do then?}

> A26: Eccoti quindi che [la mamma there then that the mother
madrelengua] puo' cominciare ad mother tongue can start to
usare la creativita'. use the creativity

The fact that the topic is a preverbal subject also helped the annotators to recognize it. As discussed in (Brunetti, 2009), preverbal subjects are typical sentence topics, and our two annotators agreed more often when the topic was in that position. The so-called hidden topics were more challenging.

And even if an expression was correctly included within the background, the two annotators still had to decide for every referential item that was part of the background whether to label it as a topic or not. Not surprisingly, they sometimes agreed, as in (11), and they sometimes picked different elements. Since there are several characteristics of the text and the preceding discourse that have to be taken into account for the identification of possible topics, we hypothesise that this category will probably always be annotated with less accuracy than the other information structure categories such as focus or non-at-issue material.

7 Conclusion

We have presented a novel method for the annotation of information structure which achieves good inter-annotator scores. In particular the agreement scores for focus are much higher than the results reported in other similar annotation studies on naturally occurring data (cf. Ritz et al. 2008). The method is based on the reconstruction of QUDs, from which the annotation of IS categories is then derived. The results of our inter-annotator agreement analysis show that the QUD-based approach, originally spelled out for English and German, can successfully be transferred cross-linguistically, supporting good agreement for the annotation of central information structure notions such as focus and non-at-issueness, with (contrastive) topic and background showing lower levels of agreement for some texts due to underrepresentation of those information structural categories in some of the data analyzed. Thanks to the QUD-based method, attention was drawn to some interesting aspects of Italian information structure, and in particular of Italian topics. Some difficulties of topic identification were shown to be reduced by the adopted annotation procedure.

References


Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation, 44:387–419.


