Direct Reported Speech in Multilingual Texts: Automatic Annotation and Semantic Categorization

Motasef Alrahabi (1), Jean-Pierre Desclés (2), Jungyeon Suh (3)

(1) (2) LaLIC  Université Paris Sorbonne  France (3) Université féminine Séoul  Corée
(1) motasef.alrahabi@gmail.com
(2) jean pierre.descles@paris sorbonne.fr
(3) alasseur13@yahoo.fr

Abstract
We propose an application for the automatic identification and categorization of quotations. The categorization is based on a semantic map of enunciative modalities. The texts are treated in three languages: Arabic, Korean and French.

1. General presentation and related works
Automatic identification of quotations using natural language processing (NLP) is now significantly growing in recent studies (Mourad 2001), (Krestel, Bergler, and Witte 2008), InQuote1, (Pouliquen, Steinberger, and Best 2008)2, (Audebert, Gaubert, and Jaccarini 2009)3 and (De la Clergerie et al. 2009).

We propose in this study an application for the automatic identification and categorization of quotations. This work can be distinguished from the previous ones in many aspects. First of all, our concerns are not to detect the source (holder) of the quotation, neither its anaphoric analysis, but we aim to identify all forms of quotation in texts by taking into consideration of its potential constructions. In addition, by using the theory of enunciation, we aim to automatically categorize the quotations in terms of various semantic criteria (commitment, opinion, judgment…), in a multilingual context (Arabic, French and Korean). Finally, the tool we use for automatic annotation, EXCOM4, is a rule-based system that does not deal with any morpho-syntactic analysis or named entities recognition (Alrahabi and Desclés 2009b). EXCOM, implementing the method of Contextual Exploration (Desclés 2006), automatically performs the annotations using the surface forms of certain linguistic markers.

In the following sections, we begin by presenting the linguistic analysis of quotations, and then we explain how the linguistic markers can be organized in a semantic map. We finish the article by showing the result of the evaluation, and the perspectives.

2. Quotation analysis
First, let’s introduce this important distinction between “utterer” (énonciateur) and “speaker” (locuteur). The utterer is the entity that reports the speech, whereas the speaker is the source (holder) of the speech.

We consider, on the formal level, that a quotation is any kind of speech delimited by meta-characters (the typographical signs of quotation) and introduced by, at least, one linguistic marker referring to an act of speaking, whether the speaker is explicitly defined or not. We take into consideration any form of direct reported speech, as long as these rules are observed, i.e. the canonical forms and hybrids or mixed forms (such as the direct style introduced by “that”, see (Tuomarla 2000))5.

In general, we consider that an utterer can report a speaker’s discourse in, at least, three ways 6:

- By attributing to a speaker an implicit act of locution (Pour X [As for X] / لا يوجد هذا الخبر [Here is this news…] / [According to X]). This reflects the distance that the utterer takes in relation to the reported content.

- By attributing to a speaker a speech as an act of “hearing” (Je me suis laissé entendre [It was intimated to me…] / بلغنا ما يلي [This news has reached us] / [heard from X]). This often indicates the spread of information (or rumors).

- By attributing to a speaker an explicit act of locution (X a décidé [X decided] / أعلنت فلان [X declared] / [X declared]).

5 In Korean (Pak et al. 2009), a set of linguistic markers following quotation marks often indicate a real quotation, such as (라고 / lako, 라고도 / lakoto / 큰 / 과도, 과도 / koto, 아라고 / ilako, etc.).
6 Examples in this paper are not identical from one language to another, but they belong to the same semantic categories.
The introduction of linguistic markers in a Semantic Map

In order to operate our categorization, we call upon the principles of the enunciative theory ((Bally 1932), (Beveniste 1966), (Culioli 1973), (Descles 1976)), in particular, the logical distinction within an utterance, between modus and dictum as in this example: “I think it’s raining”, where the modus corresponds to “I think” and the dictum to “it’s raining”. We notice that this distinction is not always easy to make at the surface level (see for instance the verb to claim (prétendre / ژاپا / 짤아 맘다), but it can be made on an abstract level where modus and dictum are represented by operations. This distinction is not concerned with separating the subjective from the objective in an utterance, because we consider that both dictum and modus are subjective representations of reality, as it is perceived by the utterer. Finally, in a reported speech, we can distinguish two modus and two dictums, depending on whether we are on the main plan (that of the utterer) or on the reported dialogic plan (that of the speaker) (ex. I assure you that she has confirmed...). Here is is the standard meta-linguistic formula of a direct reported speech (we ignore the aspecto-temporal parameters in this analysis), expressed by operators acting onto operands:

$$I\text{-SAY } \text{(modus}_1 \text{(X-SAYS (modus}_2 \text{(λ) ))})$$

where “I SAY” is a meta-linguistic operator which indicates that the utterer takes responsibility for the location. The latter, in a reported speech, is represented by the operator “X SAYS” which indicates the speaker’s commitment to the reported utterance “λ”. Enunciative modalities can then be analyzed as operators that participate in the construction of the modus of the utterer (modus$_1$) and/or the modus of the speaker (modus$_2$). These operators concern enunciative relations developed between the utterer or the speaker and their utterance (commitment, disengagement, distancing, opinion...), they concern also the relations between actors in a reported speech (control, assessment, judgments, attitudes...). These different relationships are embedded in spatio-temporal and thematic referential (see (Alrahabi and Desclés 2008), (Alrahabi and Desclés 2009a)).

Using this analysis, Figure 1 will now be refined by other semantic relations, such as the speaker’s commitment in relation to the content:
As for Mr. Jacek Saryucz Wolski, who is negotiating Brussels in the name of Poland, affirms that "Staying at the margin in the periphery of the Union does not interest us..."

In this example, the introducer of quotation (يؤكد / affirms) participates in the construction of (modusX), and can be represented by the operator of commitment "is true":

I-SAY (X-SAYS (is-true(λ)))

Another example is the opinion of speaker about the reported utterance "λ" (applaudir to applaud / تدعم [to denounce] / uggy 타다 [to denounce]).

The relation between the speaker and the co-speaker (the branch b in the figure 1) can relate to a "will relationship" (ordonner to order / معدل [to promise] / 격려하다 [to encourage]) or to an appreciative relation expressed by the speaker towards the co-speaker (louer to praise / اعتذر من / 비난하다 [to criticise]).

I-SAY(X-SAYS(λ) to Y & EVALUATION-RELATIONSHIP(X-Y))

There are cases in which we are concerned with modusY rather than modusX, such as in evaluative modalities where the utterer assesses the speaker's attitude (markers that indicate the quality of voice: vociférer [to shoot], تعلم [to stammer] / 소리치다 [to cry]), shows his own, evaluates the act of locution as a whole, or the content of the reported speech in relation to the truth value (and therefore the sincerity of the speaker)

(2) ومن التهم التي لفتت أن، "لقد علق على زوجه السابق أن، "كان يقطع الإعلاما "بشكل مستمر..

[Among the charges she has leveled at her former husband is that "he used to stop paying alimony"

We can finally mention other types of modalities, such as evidentiality (Desclés and Guéntchéva 2000). In this mode of communication, the access to the presented information is done by a median way, and the utterer presents the locution as "plausible" (so no relation with true or false values):

It came to my Knowledge that "Edward Said has blood cancer" and that he is "resisting valiantly. Even when disease strikes him down, he raises proudly to write his noble words, unchanged, and without boastfulness."

We used the platform EXCOM (Djioua et al. 2006) (Alrahabi and Desclés 2009b) which is based on the Contextual Exploration (CE) rules. We created rules for the identification and the categorization of quotations in Arabic, French and Korean, and we tested and validated them on new corpus in these different languages.

To make annotations, EXCOM needs only one pre-treatment phase of segmentation according to a specific model using also CE rules. It helps in determining the search fields for linguistic markers, and the textual segments which are to be annotated. This consists in defining the boundaries of sections, titles, paragraphs and sentences.

The presence of indicators in the text triggers the CE rules, and then, additional clues are found in a context defined by the rules, which leads to the annotation of the segment considered. Different types of rules exist, depending on the research space or the nature of linguistic markers. EXCOM allows to use the already annotated segments as markers, to order the rules and to use negative clues that cancel certain rules. A rule (R) is formally defined by a set of arguments:

\[ R = \{\text{indicator, clues, context of clues, order of clues, research space, annotating space, priority of rule, annotation}\} \]

Annotated segments are collected in separate files corresponding to the nodes of the semantic map. They are then exploited by final users, with dedicated interfaces. We have for some 800 French markers, 900 for Arabic and 600 for Korean. The core of the semantic map uses approximately forty CE rules by language.
5. Scenario for the application’s use
The typical use of our application by a final user consists in submitting a new corpus of his choice to the system, in one of the offered languages (Arabic, French or Korean)\(^9\). The semantic map is then visualized and the user is offered the possibility of choosing the categories to be used for annotating his corpus.

The process pipeline starts with segmentation and the annotation process is then called, the results are directly displayed in a base of annotated segments, according to their classification in the semantic map. The user can then navigate between the base and the original sources, or carry out a search by keywords on various spaces defined by the segmentation or by the places of the markers in segments, i.e. the content of quotation, the place of speaker, the theme of quotation...In figure 3, the base of annotations contains quotations annotated in Arabic, the user filters only those having the annotation of “opinion” of speaker, on which he carries out a request with the keyword “رسوم”/“drawings”.

6. Evaluation
We set up an evaluation for testing the capacity of EXCOM to identify and categorize the quotations according to the semantic map. To this end, we chose three rather representative categories from the map, in the sense that, on the one hand, they have complex dialogical relations (between utterer and speaker), and on the other hand, they concern important modalities which are commitment and evaluation. Here is a short description of the three selected categories:

- **Category 1:** the commitment of the speaker in relation to the reported speech (assertion).

  \[
  \text{I-SAY (X-SAYS(is-true(\lambda)))}
  \]

  Examples:

  - [The Chief of Staff favored: “I am confident that the Israelis would act differently next week in the event of continued military operations”]

  - [“I live in fear, she testifies. When I switch on the ignition in my car, I close my eyes. And I wait”]

  - [Because Niken confessed: “I went out too”, Obayashi and Takashi are embarrassed.]

- **Category 2:** the judgment of the utterer on the truth value (true or false) of the speaker’s reported speech (the speaker is presented by the utterer as sincere or liar).

  \[
  \text{I-SAY (X-SAYS(\lambda) \& is-true(\lambda))}
  \]

  Examples:

  - [The Sheikh was right when he said: “My life will be longer than the life of my hangman...”]

  - [Mme Jin [...] uttered these sincere words: “Divine Performing Arts is the hope of humanity...”]

  - [The pastor said sincerely about the preaching: “Me too, that is so difficult that I have experienced a couple of failure for 40 years.”]

- **Category 3:** the judgment of the utterer as to the “correctness value” (correct or not) of the reported speech (the speaker is presented as being right or wrong).

  \[
  \text{I-SAY (X-SAYS(\lambda) \& is-correct(\lambda))}
  \]

  Examples:

  - [The president Sarkozy had it right when he said: “what is at stake in Afghanistan is the fate of our democratic values”]

  - […the poet was mistaken in saying: “There is more in heaven and earth than is dreamt of in our philosophy.”]

  - [Mac future had said wrongly let’s look at the document instead of seeing the document of the plan.]

Starting from a large set of new texts in the three languages, we began to manually annotate quotations according to the selected categories. We stopped when we got 45 quotations for each language (15 quotations by category). Our choices of these quotations were motivated...
by the concern for covering the maximum of difficult and ambiguous cases, so as to test to the best, the capacity of the system to annotate. Thus, were taken into account the following criteria:

- the use of all quotation constructions (the introducer is before, inside or after the quotes);
- the use of all the lexical categories of introducer or modalizer markers (verbs, nouns, gerunds, adverbs, adjectives and adverbials);

We also added 6 more quotations that contain:

- fake quotations. Ex. Quotation marks which do not delineate a real quotation, as in:


- quotations not introduced by enunciative introducers. Ex:

  [(16) *L'avocat, ravi de son effet : « Et c'est signé Nicolas Sarkozy, sous l'en tête » [The lawyer, delighted with his effect : “ And it is signed Nicolas Sarkozy, under the header”]


- self-quotations (when the utterer mentions his own words). Ex:

  [(17) … 선생님이 좀 크게 턱 부르르하고 주문을 할 때 ”저는 크게 뭉크고 있어요” 라고 말했던 적이 있었다.


- fictitious quotations that are “attributed” by the utterer to the speaker. Ex:

  [(18) ولسان حال الشاب يقول: "قد كنت أخشى أن يرأني الناس فاستسلم في نظرهم و هل أنا قد سقطت أمام آخرين؟"


- and 18 for Korean (Their level is between the third and the fifth year of the university)

- the use of all the lexical categories of introducer or modalizer markers (verbs, nouns, gerunds, adverbs, adjectives and adverbials);

First, we annotated by EXCOM the texts that contain these quotations, according to the three categories cited above. It allowed us to estimate the capacity of EXCOM to identifying quotations. We then obtained the following results:

<table>
<thead>
<tr>
<th></th>
<th>Arabic</th>
<th>French</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>7%</td>
<td>5%</td>
<td>3%</td>
</tr>
<tr>
<td>Silence</td>
<td>10%</td>
<td>11%</td>
<td>6%</td>
</tr>
</tbody>
</table>

The next step of the evaluation is to compare the results (excluding the results of noise) with human judgments, both in terms of identification and categorization of quotations. Then we asked the evaluators, first, to distinguish, within a limited time span, between quotations and non-quotations (see §2), and then to categorize the retained quotations according to one of the three previously cited categories. Finally, the manual results were compared with those obtained automatically, and computed according to recall and precision measures.

7. Comments

The value of silence in the first test is due to the fact that some markers are not yet added to our resource base. The noise in French and Arabic is usually caused by the presence of fake quotation marks in the context of quotation introducers, as in the following example:

[(19) *Et on conclut que le prix de la viande "consommée" n’a pas augmenté... [And we conclude that the price of meat "consumption" has not increased...]

In this example, the emphatic quotation marks are preceded by an introducer (on conclut que). We can cite another problem with quotation marks, even if we have not faced it in this evaluation, which is the nested english quotation marks, where a quotation can contains another one, generally used as an emphatic quote:

\[ X \text{ says: } " \ldots \text{ " } \ldots \ldots \] \]

In this example, the emphatic quotation marks are preceded by an introducer (on conclut que). We can cite another problem with quotation marks, even if we have not faced it in this evaluation, which is the nested english quotation marks, where a quotation can contains another one, generally used as an emphatic quote:

\[ X \text{ says: } " \ldots \text{ " } \ldots \ldots \] \]

The disagreement between annotators in the results of the second test (ex. 36% for the category 3, in all three languages) shows that the semantic categorization that we have made is quite difficult for some evaluators. This categorization could be revised to collect several sub-categories in categories less fine.

These tests have allowed us to draw comparisons between French, Arabic and Korean on several levels. Firstly, we have noticed that in Arabic the surface forms are generally more polysemous than in French and Korean, especially the forms that have a three-letter root. This difficulty, already well known (Roth et al. 2008), (Dichy 2001), is due to the morphological ambiguity in Arabic, caused, above all, by the absence of vocalisation, the
agglutination and the relatively free word order in a sentence. To resolve this problem, we have used clues for the disambiguation of certain markers, in order to validate or not their correspondence to the researched forms. Secondly, we remark that the occurrences of direct speech in French texts and the use of enunciative modalities are richer than in texts in Arabic as well as in Korean.

In Korean, it seems easier to recognize the quotations than in French and in Arabic because of the specific markers of quotations in Korean (ko, lako...), etc. Introducers always occur after the quotation marks in Korean; in the beginning and the end in Arabic; and in the beginning, the middle and the end in French.

Finally, we mention that our analysis of reported speech was performed first on Arabic and French languages. We expanded it in this study to Korean. The transition to Korean was easy and fast: linguistic resources have been transposed into Korean by adapting French markers and by working on Korean corpus; the CE rules have been adapted or re-written using, always, the same tool, EXCOM. On the other hand, the semantic categorization was confirmed by the analysis of Korean. Indeed, there are categories where we can not have specific markers. There are also markers that have necessitated the creation of new categories in the map. But we have not encountered any conflicts or cases of misinterpretation between the three languages.

8. Perspectives

The results allow us to say that our application using EXCOM is robust and the adaptation of our analysis to the multilingualism is quick and operational. The ongoing task consists in testing resources (markers and rules) of appreciative modalities (opinion, position, attitude...) in the three languages in question, and to expand it to English.

Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful feedback, and everyone who participated in the evaluation tests.

References


