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

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

(1) (2) LaLIC Université Paris Sorbonne France (3) Université féminine Séoul Corée
(1) motasem.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), InQuote[1], (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, EXCOM[4], 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]* / اليكم هذا الخبر *[Here is this news…]* /*누군가 에 따르면 [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]* /*누군가 에게서 – 라고 전해들었다 [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]* / أعلن فلان *[X declared]* /*누군가*

1 http://labs.google.com/inquotes/
2 http://press.jrc.it/NewsExplorer/home/fr/latest.html
3 http://www.ifao.egnet.net/kawakib
4 http://www.excom.fr/

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* / 고 / *ko*, 고도 / *koto*, 이라고 / *ilako*, etc.).
6 Examples in this paper are not identical from one language to another, but they belong to the same semantic categories.

말했다[X said]); an act of inter-locution (*X a informé Y [X informed Y]* / ...فلان أجاب فلانٌ *[X replied to]* /*누군가 누구에게 물었다 [X asked Y]*); an act of reception (*X a entendu [X heard]* / قرأ فلان في الجريدة *[X read in the newspaper]* /*누군가 들었다 [X heard]*) or finally an act of transmission (*X a rapporté [X reported]* / نقل فلان *[X forwarded]* /*누군가 전했다 [X conveyed to Y]*).

Sometimes, we can have one or more intermediates between utterer and speaker, we call this entity "transmitter" (*e.g. I heard from T that X said…/ According to T, X said…*).

This dialogical organization (reporting a locution or an inter-locution, transmission, reception) enables us first to know who deals with the reported utterance (utterer, speaker or transmitter) and to draw the first categorization below:

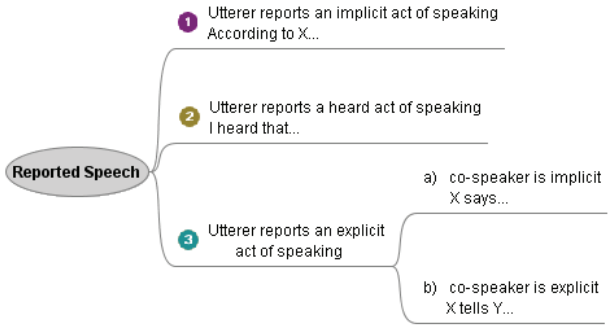

Figure 1: First organization for the semantic categories

The quotation introducers in our approach are 'enunciative' markers indicating an act of speaking. They can be verbs (*informer [to inform]* / سمع *[to hear]* / 주장하다 *[to assert]*), nominal groups (*déclaration [declaration]* / اشاعة *[rumor]* / *진술 [confess]*), adverbial phrases (*d'après* / بحسب, وفقاً لـ /*에 따르면 [According to]*) or present participles (*en affirmant [by affirming]* / مضيفاً *[by adding]* /*주장하면서 [by claming]*). These introducers are often combined with other modality clues (modalizers), as:

- Polysemic declarative markers (*se moquer de [to make fun of],* / تملّق *[to laughs]* / *비꼬다 [to give a sarcastic twist to one's word]*)

  She laughs at me by saying...

- Non declarative markers that denote the speaker's attitude (*interpeller [to hail]* / بادر *[to address]* /*말을 막았다 [stop someone's talking]*).

  He has addressed me and said...

- Other grammatical categories are also to be observed as modalities clues, such as adverbs (*enfin [finally]* / للأسف *[Unfortunately]* / *에둘러 [indirectly]*)

  Fortunately, he admitted that...

- Adjectives (*sincère [sincer]* / كاذب *[liar]* / *장황한 [long and boring]*).

  Sad, he added that...

The introducers and modalizers of quotations are of two types: indicators and clues. Indicators are the quotes, while clues help to disambiguate the indicators and to refine the categorization. Using all these markers, collected by corpora analysis, we will now refine the categorization seen above[7].

3. Organization of linguistic markers in a Semantic Map

In order to operate our categorization, we call upon the principles of the enunciative theory ((Bally 1932), (Benveniste 1966), (Culioli 1973), (Desclés 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* / زعم / *잡아떼다*), 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[8]:

$$\texttt{I-SAY}\ (\texttt{modus}_{\texttt{I}}\ (\texttt{X-SAYS}\ (\texttt{modus}_{\texttt{X}}\ (\lambda))))$$

where "I SAY" is a meta-linguistic operator which indicates that the utterer takes responsibility for the locution. 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 ($\texttt{modus}_{\texttt{I}}$) and/or the *modus* of the speaker ($\texttt{modus}_{\texttt{X}}$). 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:

7 Given the lack of space, we will describe only some sub parts of the map.

8 This expression can be defined inside the λ calculus in framework of appalicative grammar (Desclés 1976), (Desclés and Guentchéva 2000).

(1) أما السيد ياسيك ساريوتز - فولسكي، الذي يفاوض بروكسل باسم بولونيا فيؤكد من جهته أن "بقاءنا على الهامش في محيط الاتحاد هو أمر لا يثير اهتمامنا... "

[As for Mr. Jacek Saryutz 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 ($modus_X$), 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]* / نددّ *[to condemn]* / *규탄하다 [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]* / وعد *[to promise]* / *격려하다 [to encourage]*) or to an appreciative relation expressed by the speaker towards the co-speaker (*louer [to praise]* / اعتذر من *[to apologize]* / *비난하다 [to criticise]*).

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

There are cases in which we are concerned with $modus_I$ rather than $modus_X$, such as in evaluative modalities where the utterer assesses the speaker's attitude (markers that indicate the quality of voice: *vociférer [to shoot],* تلعثم *[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 لفقتها على زوجها السابق أنه « كان يقطع الإعالة المالية 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 Guentchéva 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):

(3) نما الى علمي أن "ادوارد سعيد مصاب بسرطان الدم" وأنه "يقاوم ببسالة، يقعده المرض فيقوم رافعا قامته بكبرياء، كاتبا كلمته الشريفة بدون تغير ولا تبدي"

[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."]

(4) *Mme Royal aurait donc dit face à Sarkozy : "J'ai proposé à François Bayrou d'être mon Premier Ministre et il a accepté le poste".*

[Ms Royal would then have said to Sarkozy : "I suggested to François Bayrou to be my Prime Minister and he accepted the position"]

(5) 이를 본 사람들은 "역시 골프는 멘탈게임이야" 라며 이구동성으로 말했을 것이다.

["The golf is surely a mental game", could tell unanimously the people who saw this.]

The analysis of texts in a multilingual environment allowed us to better organize markers and identify around sixty semantic categories about reported speech. All these categories and their markers (introducers and modalizers) are organized into a semantic map. Each node of the map corresponds to an enunciative modality and is represented by a single metalinguistic formula. *Figure 2* shows a sub-part of the semantic map.

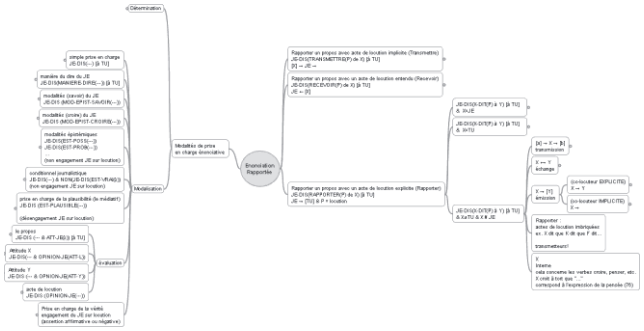

Figure 2: A sub-part of the semantic map of quotations

4. Computer Implementation

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 = {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*).

Figure 3: screenshot of the results of the annotation

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).

I-SAY(X-SAYS(is-true(λ)))

Examples:

(6) ورجح قائد الاركان : " أنا واثق ان الاسرائيليين سيتصرفون بشكل مغاير الاسبوع المقبل في حال تواصلت العمليات العسكرية "

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

(7) *"Je vis dans la peur, témoigne t elle. Quand j'allume le contact de ma voiture, je ferme les yeux. Et j'attends".*

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

(8) […]닛켄이 "역시 외출하였습니다"라고 자백해버렸으니 오바야시도 다카하시도 큰 창피를 당한 셈이지요.

[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).

I-SAY(X-SAYS(λ) & is-true(λ))

Examples:

(9) ولقد صدق الشيخ إذ قال : « ستكون حياتي أطول من حياة شانقي »

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

(10) *Ms Jin [...] a prononcé ces mots sincères : "Divine Performing Arts est l'espoir de l'humanité..."*

[Mme Jin [...] uttered these sincere words : "Divine Performing Arts is the hope of humanity…"]

(11) […] 이영훈 목사는 목회자 영성에 관해 "나 역시 40 년간 실패의 연속을 경험한 어려운 주제"라고 솔직히 말했다.

[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).

I-SAY(X-SAYS(λ) & is-correct(λ))

Examples:

(12) فالرئيس ساركوزي كان محقاً إذ قال : « إن الرهان في أفغانستان يتناول قيمنا الديمقراطية »

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

(13) *...le poète se trompait en disant : "Il y a plus de choses entre le ciel et la terre que notre philosophie n'en peut concevoir."*

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

(14) 맥 퓨처님은 최근 기획 문서를 보자는 걸 "문건 좀 보자"고 잘못 말했었다고 한다.

[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

[9] User can target corpus through an EXCOM module that enables to crawl texts from the Web.

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:

(15) *Lire " L'Aurore " et le dossier " Comment l'OMC fut vaincue ", Le Monde diplomatique, janvier 2000.*

[Read "L'Aurore" and the folder "Comment l'OMC fut vaincue", Le Monde diplomatique, january 2000.]

- 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 크게들 부르라고 주문을 할 때 "저는 크게 부르고 있어요" 라고 말했던 적이 있었다.

[…as teacher asked us to sing loudly, I remembered saying : "I do sing loudly."]

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

(18) ولسان حال الشاب يقول: « قد كنت أخشى أن يراني الناس فأسقط في نظرهم وها أنا قد سقطت فمم أخاف؟ »

[It's as if the young man said: "I was afraid to wane in people's eyes if they see me; but now that I am fallen, who would I fear?"]

First, we annotated by EXCOM the texts that contain theses 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:

	Noise	Silence
Arabic	7%	10%
French	5%	11%
Korean	3%	6%

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 [10], first, to distinguish, within a limited time span [11], 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. We considered then, that the "correct" annotation is the most frequently chosen by the evaluators. The annotation protocol as well as the corpus of the segments submitted to the evaluators will be soon online. The results are the following:

	Category 1		Category 2		Category 3	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Arabic	86%	83%	85%	83%	85%	78%
French	88%	82%	85%	85%	84%	80%
Korean	93%	85%	88%	83%	82%	87%

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 says: " …… " … "…… "
1 a b 2

The error can occur in this case by considering the quotes *1* and *a* as the real quotation marks surrounding the reported speech. Finally, in this example:

(20) *Ainsi, de 1900 à 1996, on constate une dérive d'environ 0,003" par an, approximativement le long du 80e méridien Ouest […] par rapport au point central restent inférieures à 0,3" sur une année.*

the quotation marks indicating *seconds* units (*0,003" / 0,3"*) were selected as real quotation marks, in the presence of the introducer *constate (to notice)*.

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

10 The evaluators, all native language, were 11 for Arabic, 23 for French and 18 for Korean (Their level is between the third and the fifth year of the university)

11 Two hours.

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

Alrahabi M., Descles J. P. 2008. Automatic annotation of direct reported speech in Arabic and French, according to semantic map of enunciative modalities. In *Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*. August 25 27, 2008, Gothenburg, Sweden.

Alrahabi M., Descles J. P. 2009a. Opérations de prise en charge énonciative : assertion, médiatif et modalités dans le discours rapporté direct, en arabe et en français. *Methods of lexical analysis, theoretical assumptions and practical applications*, (Bogacki, K., Cholewa, J., Rozumko, A. eds.), Bialystok: Wydawnictwo Uniwersytetu w Bialymstoku, 17 26.

Alrahabi M., Desclés J. P. 2009b. EXCOM : Plate forme d'annotation sémantique de textes multilingues, *TALN* 2009, Senlis, 24 26 juin 2009.

Audebert C., Gaubert C. and Jaccarini A. 2009. Minimal Resources for Arabic Parsing: an Interactive Method for the Construction of Evolutive Automata, article en ligne, MEDAR 2009 conference, Le Caire.

Bally Ch. 1965 (first edition 1932). *Linguistique générale et linguistique française*, Bern, Francke.

Benveniste E. 1966. *Problèmes de linguistique générale*, Tomes 1, Paris, Gallimard,

Culioli, A. 1973. *Sur quelques contradictions en linguistique*. *Communications* 20, 83 91, repris dans PLE (1999 : 43 52).

De La Clergerie E., Sagot B., Stern R., Denis P., Recourcé G. and Mignot V. 2009. Extracting and Visualizing Quotations from News Wires, LTC 2009, Poznan,

Desclés J. P. 2006. Contextual Exploration Processing for Discourse Automatic Annotations of Texts. In *Proceedings of the FLAIRS 2006, invited speaker*, Melbourne, Florida, pp 281 284

Desclés J. P. 1976. Quelques opérations énonciatives élémentaires. *Logique et niveaux d'analyse linguistique*. Paris : Klincksieck, (ed. R. Martin & G. David). 213 242.

Desclés, J. P. Guentchéva Z. 2000. Enonciateur, locuteur, médiateur. *Les Rituels du dialogue*, Nanterre: Société d'ethnologie, *Recherches thématiques* 6. 79 112.

Dichy, J.2001. On lemmatization in Arabic. In *Proceedings of the Workshop on Arabic Language Processing*, Toulouse.

Djioua B., Flores J. G., Blais A., Desclés J. P., Guibert G., Jackiewicz A., Le Priol F., Leila N. B., Sauzay B. 2006. EXCOM : an automatic annotation engine for semantic information. *Actes de FLAIRS 2006*, Floride.

Krestel R., Bergler S., and Witte R.. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, May 28 30, 2008, Marrakech, Morocco.

Mourad G. 2001 Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations. Réalisation des applications informatiques SegAtex et CitaRE. Ph. D. Diss., Dept. of Mathematics, Linguistics and Logic. Paris IV Sorbonne University.

Pak S.H., Soh J.S., Suh J.Y. 2009. Automatic identification of quotations in Korean newspapers. *Annotations automatique et recherche d'informations* (ed. Desclés & Le Priol), Paris, Hermès. 123 147.

Pouliquen B., Steinberger R., Best C. 2008. Automatic Detection of Quotations in Multilingual News. In *Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*. August 25 27, 2008, Gothenburg, Sweden.

Roth, Ryan, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of Association for Computational Linguistics (ACL)*, Columbus, Ohio.

Tuomarla U. 2000. *La citation mode d'emploi. Sur le fonctionnement discursif du discours rapporté direct*. Academia Scientifiarum Fennica. Ser. Humaniora. 308. Saarijärvi, Finland.