A Cross-lingual Approach to the Discourse Automatic Annotation: Application to French and Bulgarian

Iana Atanassova, Antoine Blais, Jean-Pierre Desclés

LaLIC (LAngage, Logique, Informatique et Cognition) Laboratory
University of Paris-IV Sorbonne
Maison de la recherche, 28 rue Serpente 75006 Paris
iana.atanassova@gmail.com, antoine.blais@paris-sorbonne.fr, jean-pierre.descles@paris-sorbonne.fr

Abstract

In this paper we propose a cross-lingual approach to the discourse automatic annotation of scientific articles by the Contextual Exploration method. We present an application to French and Bulgarian as an illustration of the possibility to work with different languages that the method provides. We describe the methodology for the construction of the linguistic resources for Bulgarian based on the already existing resources for French and make an evaluation of the obtained annotations. We also discuss some of the applications of the discourse automatic annotation for automatic summarization and information retrieval.

Introduction

Presently text annotation is a widely used technique because it provides the basis for various applications, such as information retrieval, information extraction and electronic document management. The annotation consists in adding meta-data to parts of the text, by following rules and procedures defined in advance. Our objective will be the automatic annotation of texts according to a list of discourse categories.

The automatic semantic annotation and the automatic summarization by extraction have been for many years some of the major domains of research at the LaLIC laboratory at the University Paris–Sorbonne. Since 1995, different projects have been carried out, among which SERAPHIN (Berri, 1995), SAFIR (Berri et al., 1996), ContextO (Crispino et al., 2003) and EXCOM (Djioua et al., 2006). The EXCOM system, which is an implementation of the Contextual Exploration Method (Desclés, 1997, Desclés and Minel, 2005), allows the automatic annotation of texts on the semantic and discourse levels in view to further applications such as automatic summarization (Blais et al., 2006), the construction of flexible extracts, information retrieval, etc.

Unlike other systems for automatic semantic annotation, such as the GATE platform (Cunningham et al., 2002) and KIM (Popov et al., 2004), the EXCOM system is oriented above all towards the extraction of semantic relations in texts, rather than named entity recognition.

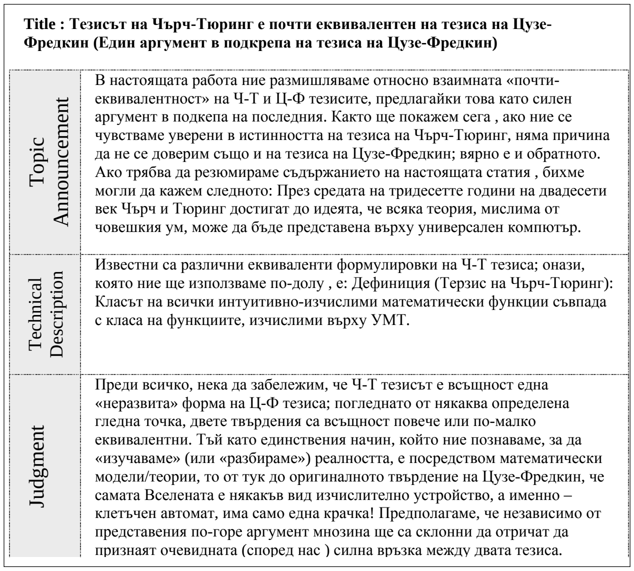
The Contextual Exploration (CE) method, developed in the LaLIC laboratory, is a linguistic tool which can access the semantic and discourse structure of a text without making any morphological or syntactic analysis. In working with French and Bulgarian we will show that the linguistic resources that are used by this method are relatively easy to transmit from one language to another. Our aim will be to illustrate the construction of linguistic resources for the Bulgarian language based on the existing resources for French. In the end, we will present an evaluation of the discourse annotations for the two languages.

Up to now, the difficulty in the implementation of linguistic methods for different languages has been one of the important factors that have favored the development of text processing systems based mainly on statistical methods. Linguistic methods, as opposed to statistical methods, use language-specific resources that have to be created separately for each language. However, in this article we show that the Contextual Exploration method, although a purely linguistic approach, permits a relatively easy application to many languages. Once the linguistic resources for this method are created for a given language, their transmission to new languages presents a considerably smaller effort.

The CE method uses the identification of surface linguistic markers that allow the annotation of textual segments with discourse categories. These surface linguistic markers are generally domain independent. Furthermore, the linguistic resources for the method are relatively easily transmittable to different languages. These two reasons make this approach an important alternative to the statistical methods for automatic summarization. The advantage of statistical methods in general is that their algorithms, based on frequency calculations and machine learning, are essentially language independent. However, one significant weakness of these methods is that they consider only the physical structure of the text and usually do not take sufficiently into account its discourse organization. We note the approach of (Teufel and Moens,

1999) in which they try to differentiate sentences in scientific articles according to their rhetorical structure by machine learning.

The results of the discourse annotation obtained by the CE method can be used in a variety of applications related to automatic summarization and information retrieval. As an important example we consider the automatic production of text syntheses corresponding to one or several discourse categories (Fig. 1). It should be noted that this type of results is difficult to obtain by methods that do not make a linguistic analysis. Text syntheses are very useful if the user needs to go through a large number of scientific articles and is interested only in a specific type of information, for example the hypotheses used in these articles. In this case, a number of articles can be synthesized according to the discourse category in question, and the synthesis will present a list of textual segments containing this type of information.



Title : Тезисът на Чърч-Тюринг е почти еквивалентен на тезиса на Цузе-Фредкин (Един аргумент в подкрепа на тезиса на Цузе-Фредкин)

Topic Announcement	В настоящата работа ние размишляваме относно взаимната «почти-еквивалентност» на Ч-Т и Ц-Ф тезисите, предлагайки това като силен аргумент в подкепа на последния. Както ще покажем сега , ако ние се чувстваме уверени в истинността на тезиса на Чърч-Тюринг, няма причина да не се доверим също и на тезиса на Цузе-Фредкин; вярно е и обратното. Ако трябва да резюмираме съдържанието на настоящата статия , бихме могли да кажем следното: През средата на тридесетте години на двадесети век Чърч и Тюринг достигат до идеята, че всяка теория, мислима от човешкия ум, може да бъде представена върху универсален компютър.
Technical Description	Известни са различни еквиваленти формулировки на Ч-Т тезиса; онази, която ние ще използваме по-долу , е: Дефиниция (Терзис на Чърч-Тюринг): Класът на всички интуитивно-изчислими математически функции съвпада с класа на функциите, изчислими върху УМТ.
Judgment	Преди всичко, нека да забележим, че Ч-Т тезисът е всъщност една «неразвита» форма на Ц-Ф тезиса; погледнато от някаква определена гледна точка, двете твърдения са всъщност повече или по-малко еквивалентни. Тъй като единствения начин, който ние познаваме, за да «изучаваме» (или «разбираме») реалността, е посредством математически модели/теории, то от тук до оригиналното твърдение на Цузе-Фредкин, че самата Вселената е някакъв вид изчислително устройство, а именно – клетъчен автомат, има само една крачка! Предполагаме, че независимо от представения по-горе аргумент мнозина ще са склонни да отричат да признаят очевидната (според нас) силна връзка между двата тезиса.

Fig. 1. Visualization of a text synthesis according to discourse categories, obtained automatically

We consider that the main interest in the automatic processing of small languages, such as Bulgarian, lies above all in the development of systems for multilingual information retrieval. By exploiting the semantic annotations, obtained automatically, some information can be found in scientific texts in Bulgarian that have not yet been published in other languages.

Discourse Automatic Annotation by Contextual Exploration

The CE method uses the presence or absence of surface linguistic markers for the automatic annotation of textual segments according to a set of discourse categories. The linguistic markers are some specific words, expressions, typographic marks or more complex textual elements. The lists of markers are obtained after a linguistic analysis of a large number of texts. In this method we distinguish two main types of linguistic markers: indicators and contextual clues. The principles of the method have been described in detail in (Desclés, 1997, Djioua et al., 2006).

Before the actual annotation, the text is automatically segmented by the SegaTex module, designed by B. Djioua and Fl. Le Priol, by using an analysis of the textual typography (Mourad, 2001). The annotation itself is carried out by Contextual Exploration rules, implemented in the EXCOM system (Djioua et al., 2006).

We consider as one of the major advantages of the CE method the fact that it is based entirely on the localization of surface linguistic markers and therefore it does not need any preliminary syntactic or morphological analysis, which renders the text processing faster and more reliable. However, this approach is compatible with other text processing tools, such as named entity recognition, domain ontology, statistical methods, etc. As the EXCOM system is conceived for the annotation of semantic relations in texts, rather than named entity recognition, its aim is not to create an alternative to the actual methods for NLP, but rather to propose an approach that is complementary and more appropriate for some specific tasks, such as the construction of text syntheses according to discourse categories.

Discourse Categories

In our work we have focused our attention on a specific type of texts – scientific articles, because they have two important characteristics. The first one is their underlying communicative role. The general aim of scientific articles is to convince the reader in the validity of some scientific argumentation (Swales, 1990) and also to transmit information about a research work, theory, etc. Secondly, scientific articles are characterized with a number of argumentation techniques (expressed by surface linguistic forms) and discourse forms that are proper to this kind of texts. In fact, scientific texts can be distinguished from other types of texts by the presence/absence (or very high/low frequency) of some classes of linguistic markers (Bronckart et al., 1985), such as the expressions of argumentations and the meta-discourse elements, and the absence of other linguistic markers, such as temporal organizers, use of past tenses, etc.

Our approach is based on the hypothesis that in scientific texts there exist certain linguistic units whose main function is to reveal the discourse structure and organization of the text. Their role is to help the reader navigate in the text and understand the discourse function of the textual segments in this structure. These linguistic units appear on the surface level of the text, just like the lexical and grammatical units, but they are for the most part more complex in nature and are composed of lexical and grammatical units. Examples of discourse markers are: *Our hypothesis is that…; As a conclusion… It should be noted that…*

These discourse markers represent the author's attitude towards the information expressed and determine the discourse function of the textual segment (for example: hypothesis, conclusive remark, emphasis). In some cases these markers can be discontinued.

The goal of our work is to locate automatically the textual segments (in French and Bulgarian) belonging to a number of discourse categories by using the presence of discourse linguistic markers. The discourse categories we have chosen to consider are issued of a linguistic analysis of the types of textual segments that occur in scientific articles. Our supposition is that these discourse categories are explicitly marked in texts by the author by using surface linguistic units. Fig. 2 presents a semantic map of the discourse categories on which we focus our attention.

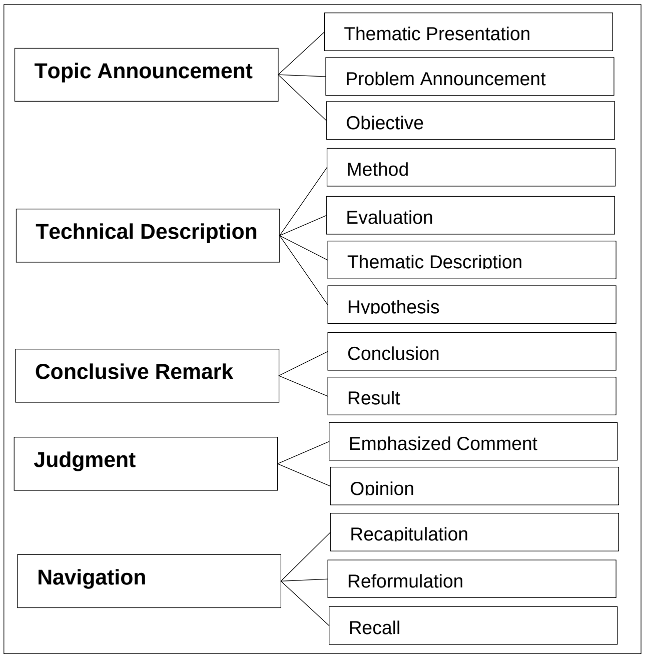


Fig. 2. Semantic map of the discourse categories for summarization

These categories are specific to and essentially present in scientific articles. We consider their extraction as important for the task of automatic summarization, especially the extraction of Topic Announcements and Conclusive Remarks, because the meanings they carry are indicative of the contents of the text. In fact, is has been shown that human professional summarizers tend to extract segments belonging to similar categories in the case of scientific articles (for details see Endres-Niggemeyer, 1998, Liddy, 1991). (Blais et al., 2006 and 2007) describes an automatic summarization strategy that uses the discourse annotation of the discourse categories on fig. 2.

We consider as important the fact that the usage of these discourse markers is not in general restricted to a given domain, unlike named entities. They can be found in texts in all scientific domains (for ex. linguistics, geography, biology ...), their discourse role remaining the same and independent of the subject of the article. This means also that the linguistic resources we create are generally domain independent.

Cross-lingual Approach

The discourse categories we consider in our approach represent the means that the author uses in order to express the discourse structure of the text. The same discourse categories can be used in different languages and are language independent, because they are proper to the discourse structure of texts. In particular, in the case of scientific articles it is clear that the same discourse categories (conclusions, hypotheses, methods …) will be used, regardless of the language of the text.

Therefore, we consider that the method we propose for the discourse annotation of scientific texts is valid and applicable to any natural language. Our experience in the LaLIC laboratory in working with different natural languages confirms this. Up to now, the method has been applied successfully for the automatic annotation of texts in French (Blais et al., 2007), English, Bulgarian (Atanassova, 2006), Korean (Kang, 2004), Arabic (Alrahabi et al., 2006) and Chinese (Zhang, 2006).

We note, of course, that the linguistic resources used for the CE method vary from one language to another. In fact, the indicators and clues consist of lists of words and expressions that are language specific. It is true that they cannot be obtained by a simple lexical translation of the linguistic markers for another language. However, once the resources are constructed for one language, the linguistic analysis needed to rebuild them for another language is considerably faster and simpler than the initial one.

Construction of Linguistic Resources

The linguistic resources are created after a linguistic study of a corpus of scientific articles. After a first stage of manual annotation, an analysis is carried out on the possible ways that each discourse category can be presented in texts. The lists of linguistic markers and the Contextual Exploration Rules are at first inferred from the examples present in the corpus, then continually accumulated and generalized through linguistic observation and reflection on the possible variations, as well as working with dictionaries.

There are several important principles that must be followed during the creation of the linguistic resources:

- All the linguistic markers are relatively independent of the subject and the domain of the article.
- The presence of the indicator and the contextual clues in a segment must be a sufficient condition for the annotation of the segment with the discourse category.
- The signification of the indicator must be suggestive of the discourse category. The role of the contextual clues is complementary: they determine the concrete sense of the indicator in the cases of polysemy.
- The indicators are words or complex expressions that should not occur very frequently in texts, because from a

linguistic point of view the more frequent an expression is, the less is its signification. From programming point of view, the usage of indicators that occur less frequently in texts makes the algorithm execution much faster.

The construction of the linguistic resources for French has been discussed in (Blais et al., 2007). Here we will describe the methodology that we have used to transmit these resources to another language, namely Bulgarian.

As first, we have carried out the manual annotation of a corpus of 15 scientific articles in Bulgarian in order to get a general idea of the types of linguistic markers we will be looking for. After that, we have studied the resources for French by focusing on each CE Rule in order to determine if it is applicable to Bulgarian and in what circumstances.

Each CE Rule annotates a set of textual segments that belong to the same discourse category and have similar syntactic and semantic structures. A discourse category is represented by a number of rules, each one treating a different class of segments. In the transmission of the linguistic resources to Bulgarian, we analyze each of the rules and determine if the type of textual segments it annotates is acceptable and can occur in Bulgarian texts. Then, we proceed to the modification, if necessary, of the rule and the reconstitution of the lists of linguistic markers for Bulgarian. In doing so, our aim is that the new rule annotates those textual segments in Bulgarian that more or less correspond to a translation and sometimes reformulation of the segments annotated by the similar rule in French.

It is important to note that the new linguistic markers for Bulgarian cannot be obtained by a simple translation from those in French. In fact, the indicators consist of lists of words and expressions, whose meaning in most cases is context dependent and their usage may vary from one language to another. Similarly, the contextual clues, which confirm the concrete meaning of the indicator, are used in texts in combination with it and therefore are dependent on the different possible usages of the indicator in the respective language. Moreover, the lists of linguistic markers often consist of regular expressions that can take account of some lexical and form variations and also locate more complex discontinued expressions.

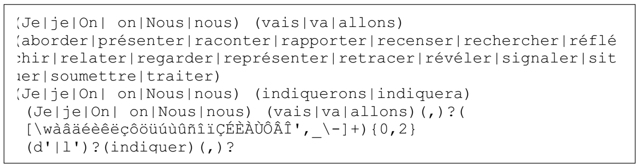
It must be considered also that in the analysis of a given Contextual Exploration Rule it is necessary to modify it in order to make it applicable to Bulgarian, because of the intrinsic differences between the two languages. Most often we have encountered two different reasons for the modification of a rule. The first reason is the slight difference in the word order in the two languages, which obliges us to modify the possible contexts of the contextual clues (to the left or to the right of the indicator). The second reason is that sometimes the indicators in the two languages have a different degree of polysemy, i.e. certain expressions that are relatively non-ambiguous in French can have several possible semantic values in Bulgarian and vice versa. In this case, the role of the contextual clues in the CE Rule must be reexamined and the rule has to be modified by removing or introducing new contextual clues.

To illustrate this approach, we will take an example of a CE Rule in French for the discourse category of Topic Announcement (Table 1):

Rule name:	R3_Presentation_Thematique
Indicator:	List "*forme-presentation-auteur*"
Contextual clues:	List "*document-en-cours*"
Context:	Sentence
Annotation:	Topic Announcement

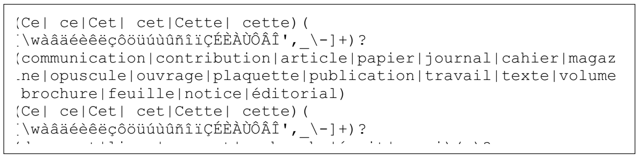
Table 1. A Contextual Exploration Rule

This rule annotates textual segments that represent a Topic Announcement of the author concerning the document in which they occur. The indicator is a list of regular expressions that recognize patterns in French such as *"je presenterai" ("I will present"), "nous allons discuter" ("we will discuss")…* In this case the presence of the indicator is not enough to attribute the discourse role of Topic Announcement to the segment, because it is possible that this indicator occurs in other segments that are not Topic Announcements. The contextual clue is a list of regular expressions that recognize patterns such as *"cet article" ("this article"), "le présent papier" ("the present paper")…* Figures 3 and 4 present simplified versions of the two lists.



```
(Je|je|On| on|Nous|nous) (vais|va|allons)
(aborder|présenter|raconter|rapporter|recenser|rechercher|réflé
chir|relater|regarder|représenter|retracer|révéler|signaler|sit
uer|soumettre|traiter)
(Je|je|On| on|Nous|nous) (indiquerons|indiquera)
 (Je|je|On| on|Nous|nous) (vais|va|allons)(,)?(
 [\wàâäéèêëçôöüúùûñîïÇÉÈÀÙÔÂÎ',_\-]+){0,2}
 (d'|l')?(indiquer)(,)?
```

Fig. 3. Part of the list "*forme-presentation-auteur*" (French)



```
(Ce| ce|Cet| cet|Cette| cette)(
[\wàâäéèêëçôöüúùûñîïÇÉÈÀÙÔÂÎ',_\-]+)?
(communication|contribution|article|papier|journal|cahier|magaz
ine|opuscule|ouvrage|plaquette|publication|travail|texte|volume
 brochure|feuille|notice|éditorial)
(Ce| ce|Cet| cet|Cette| cette)(
[\wàâäéèêëçôöüúùûñîïÇÉÈÀÙÔÂÎ',_\-]+)?
```

Fig. 4. Part of the list "*document-en-cours*" (French)

This CE Rule annotates segments such as:

« **Je présenterai** dans <u>cet article</u> un modèle automobile qui vient de sortir des usines de France. » *("**I will present** in <u>this article</u> an automobile model that has just left the factories in France.")*

« Dans <u>le présent document</u> **nous discuterons** les raisons de la disparition des dinosaurs. » *("In <u>the present document</u> **we discuss** the reasons for the disappearance of the dinosaurs.")*

In Bulgarian, a Topic Announcement can be expressed in a similar way. For example:

"В <u>настоящата статия</u> **ще предложим** два метода за оценка на автоматично резюме."*("In <u>the present article</u> **we will propose** two methods for automatic summarization.")*

We consider that in this case the indicator and the contextual clues operate in a similar manner in French and in Bulgarian. The corresponding indicator in Bulgarian (expressions such as "*we will propose*") is characteristic of segments that contain Topic Announcements. On the other hand, the indicator alone can occur also in other types of segments that do not contain the contextual clues, for example:

"За да поясним тази идея, **ще предложим** на читателя да разгледа следната диаграма." *("To clarify this idea, **we will propose** to the reader to look at the following diagram.")*

For this reason, in Bulgarian, as in French, we also need the presence of the contextual clues (expressions like "*in this article*") in the segment in order to confirm that it is a Topic Announcement.

Therefore, in this case the lists of linguistic markers for Bulgarian can safely be created by analogy with the corresponding lists in French and the CE Rule can be transmitted to Bulgarian without significant modification. Figures 5 and 6 present simplified versions of the two lists for Bulgarian.

```
(ще|Ще|няма да|Няма да|смятам да|Смятам да|смятаме да|Смятаме
да)?
(представим|представя|изложа|изложим|разгледаме|разгледам|разгл
еждаме|разглеждам|дадем|дадеме|дам)
…
```

Fig. 5. Part of the list "*forme-presentation-auteur*" (Bulgarian)

```
(този|тази|тези|настоящият|настоящата|настоящите|текущият|текущ
ата|текущите) ([а-я]+)? (документ|статия|записки|доклад|лекция
|студия|текст|документи|статии|записка|доклади|лекции|студии|те
кстове)
(този|тази|тези|настоящият|настоящата|настоящите|текущият|текущ
ата|текущите) (глава|статия|лекция|увод|изложение|заключение
|част|секция) …
```

Fig. 6. Part of the list "*document-en-cours*" (Bulgarian)

It is important to note that, both French and Bulgarian being highly morphological languages, the lists of linguistic markers can contain a number of lexical forms for each lexeme but rarely whole paradigms. Indeed, the indicators and clues convey the meaning of the discourse category through their semantic features and also grammatical properties. For example, the list *"forme-presentation-auteur"* includes some verb forms like *"we will present"* and *"I propose"*, but does not include the past forms of these verbs.

While rewriting the CE Rules in this manner, our aim is to reconstitute a set of rules for Bulgarian that annotates all the occurrences of the respective discourse category. Because of the particularities of each language, it may arrive that some of the rules are highly language specific and cannot be obtained from or transmitted to similar rules for another language. That is why the possible expressions of the discourse category must be carefully analyzed for each new language through the linguistic study of corpora and also personal reflection of the linguist. Eventually, new rules can be added that do not have direct analogues in other languages.

For example, in French there exists an impersonal form of verbs, e.g. *"on présente" ("I/we/one present(s)")*, that is sometimes used in scientific articles and is an indicator of a Topic Announcement. As this form is part of the general verb paradigm, it is well recognized by the rule we have discussed above. Such a form does not exist in Bulgarian. The same meaning in Bulgarian can be conveyed by a form in the middle voice: *"се представя" ("is presented / presents")*. A French equivalent of this form exists *("se présente")*, but it cannot be used in scientific articles in the same manner. Therefore, in order to take into account this difference, we have to create a new CE Rule for Bulgarian, which does not exist for French and takes as indicators some similar verb forms in the middle voice.

Evaluation

We have carried out an evaluation of the discourse annotation for each separate discourse category. For French the evaluation was based on a corpus of 20 scientific articles, in the domains of Linguistics, Natural Language Processing and Artificial Intelligence. At first, we have annotated manually the corpus for each discourse category and executed the automatic annotation of the same corpus. Then we have calculated the measures of recall and precision for the different discourse categories. Table 2 presents the results of this evaluation for French:

	Precision	**Recall**
Topic Announcement	77.81 %	62.83 %
Technical Description	80.59 %	60.75 %
Conclusive Remark	70.20 %	81.25 %
Judgment	74.37 %	82.00 %
Navigation	80.28 %	88.97 %

Table 2. Results of the evaluation for French

In this table we present only the results for the main categories. In fact, each one of them is divided into several sub-categories (see the semantic map on fig. 2). The values of precision and recall can vary for the different sub-categories of the same category. For example, the precision for the sub-category of Objective is 73.81% and that of Problem Announcement is 81.48%, both being sub-categories of the Topic Announcements.

For Bulgarian, we have worked on a corpus of 20 scientific articles, in the domains of Physics, Chemistry, Linguistics, Informatics and Economy. The evaluation for Bulgarian was carried out in the same manner as the evaluation in French.

	Precision	**Recall**
Topic Announcement	69.70 %	64.38 %
Technical Description	73.08 %	55.56 %
Conclusive Remark	83.33 %	88.57 %
Judgment	88.37 %	68.18 %
Navigation	80.00 %	65.00 %

Table 3. Results of the evaluation for Bulgarian

The results of the evaluation, on table 3, show that in most cases the values for Bulgarian are close to the respective values for French.

It can be seen that the values of the recall in Bulgarian are lower than the precision. This means that most of the errors consist in the non localization of segments that contain the discourse categories, rather than the attribution of the categories to segments that do not contain them. The values of the precision, indicating the noise in the system, are due to the presence of some polysemic markers that have not been taken into account by the CE rules. These values can be improved by adding more specific contextual clues to the rules containing the polysemic markers.

On the other hand, there are two possible reasons for the lower values of the recall. Firstly, the lists of indicators and contextual clues do not contain all the synonym expressions that can occur in texts and because of this, some of the occurrences of the discourse categories in the corpus have not been localized by the system. In this case to improve the results we have to elaborate the lists of linguistic markers to take into account more surface forms. Secondly, some the segments may have not been annotated in Bulgarian because they contain structures that are specific for this language and that do not have similar equivalents in French. Therefore to improve the recall we need also to introduce new CE rules for Bulgarian that will annotate these segments.

Conclusions and Future Work

We have presented in this article a methodology for the creation of resources for the automatic discourse annotation of texts based on the existing resources for another language using the same semantic map. As we have shown, the creation of the resources for new languages does not require any extensive corpus analysis but rather a thorough analysis of the existing resources. The results we have obtained after the evaluation for French and Bulgarian show that the transmission of the linguistic resources from one language to another does not reduce significantly the performance of the system.

In our future work we will consider the implementation and evaluation of this method for other languages in view to the development of new applications based on the discourse annotation in different languages, notably systems for multilingual information retrieval.

References

Alrahabi, M., Ibrahim, A. H., and Desclés, J.-P., 2006, Semantic Annotation of Reported Information in Arabic. *FLAIRS 2006*, Melbourne, Florida.

Atanassova, I., 2006, *Annotations sémantiques automatiques de textes bulgares intégrées dans la plate-forme EXCOM. Fiches de résumé*. Master's thesis under the direction of J.-P. Desclés, Paris-Sorbonne Univ.

Berri, J., 1995, *Contribution à la méthode d'exploration contextuelle. Applications au résumé automatique et aux représentations temporelles*. Ph. D. Thesis, Paris-Sorbonne Univ.

Berri, J., Cartier, E., Desclés, J-P., Jackiewicz, A. and Minel, J-L., 1996, SAFIR, système automatique de filtrages de textes. *TALN'96*, Marseille.

Blais, A., Atanassova, I., Desclés, J.-P., Zighem, L., and Zhang, M., 2007, Discourse Automatic Annotation of Texts: an Application to Summarization. *FLAIRS 2007*, Key West, Florida.

Blais, A., Desclés, J-P. and Djioua, B., 2006, Le résumé automatique dans la plate-forme EXCOM. *Digital Humanities 2006*, Paris.

Bronckart, J.-P. et al., 1985, *Le fonctionnement des discours. Un modèle psychologique et une méthode d'analyse*. Paris, Delachaux & Niestlé.

Crispino, G., 2003, Une plate-forme informatique de l'exploration contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes, Ph. D. Thesis, Paris-Sorbonne Univ.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V., 2002, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia.

Desclés, J-P. and Minel, J.-P., 2005, Interpréter par exploration contextuelle. *Interpréter en contexte*, pp 305–328, Paris, Hermes

Desclés, J-P., 1997, Systèmes d'exploration contextuelle. *Co-texte et calcul du sens*, (Claude Guimier). Presses Universitaires de Caen, 215-232.

Djioua, B., Garcia Flores, J., Blais, A., Desclés, J-P., Guibert, G., Jackiewicz, A., Le Priol, F., Nait-Baha, L., Sauzay, B., 2006, EXCOM: an Automatic Annotation Engine for Semantic Information. *FLAIRS 2006*, Melbourne, Florida.

Endres-Niggemeyer, B., 1998, *Summarizing Information*. Springer, Berlin

Kang, J. Y., 2004, *Le résumé automatique par le filtrage sémantique. Contribution à la méthode d'Exploration Contextuelle,* Master's thesis under the direction of J.-P. Desclés, Paris-Sorbonne Univ.

Liddy, E. D., 1991. The discourse level structure of empirical abstracts; An Exploratory Study. *Information Processing and Management*, Vol. 35, No. 2

Mourad, G., 2001, *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE*, Ph. D. Thesis, Paris-Sorbonne Univ.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D. and Kirilov, A., 2004, KIM - a semantic platform for information extraction and retrieval. *Natural Language Engineering (2004),* Cambridge University Press

Swales, J., 1990, *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Teufel, S. and Moens, M., 1999, Argumentative Classification of Extracted Sentences as a First Step towards Flexible Abstracting. In: Mani, I., Maybury, M. (eds.), *Advances in Automatic Text* Summarization, MIT Press.

Zhang, M., 2006, *Résumé automatique de texte en chinois-mandarin par la méthode d'exploration contextuelle avec intégration dans la plate-forme informatique EXCOM*. Master's thesis under the direction of J.-P. Desclés, Paris-Sorbonne Univ.