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
Paris-Sorbonne University

FLAIRS-21, May 2008
1 Introduction: Cross-lingual Discourse Automatic Annotation

2 Linguistic resources for Bulgarian
   - Specifics of the Bulgarian Language
   - Creation of the linguistic resources

3 Evaluation
We will discuss the creation of the linguistic resources for the automatic annotation in Bulgarian based on the already existing resources for French.

**Method:** Contextual Exploration (J.-P. Desclés, 1997)  
This is a purely linguistic approach that permits a relatively easy implementation to different languages, which we will illustrate with French and Bulgarian.

**Task:** Creation of the linguistic resources for the automatic annotation of *scientific articles* according to a semantic map for the task of automatic summarization.

Implementation in the EXCOM system for automatic annotation.
We will discuss the creation of the linguistic resources for the automatic annotation in Bulgarian based on the already existing resources for French.

**Method: Contextual Exploration** (J.-P. Desclés, 1997)
This is a purely linguistic approach that permits a relatively easy implementation to different languages, which we will illustrate with French and Bulgarian.

**Task:** Creation of the linguistic resources for the automatic annotation of *scientific articles* according to a semantic map for the task of *automatic summarization*.

Implementation in the EXCOM system for automatic annotation.
We will discuss the creation of the linguistic resources for the automatic annotation in Bulgarian based on the already existing resources for French.

**Method:** Contextual Exploration (J.-P. Desclés, 1997)
This is a purely linguistic approach that permits a relatively easy implementation to different languages, which we will illustrate with French and Bulgarian.

**Task:** Creation of the linguistic resources for the automatic annotation of *scientific articles* according to a semantic map for the task of *automatic summarization*.

Implementation in the EXCOM system for automatic annotation.
We will discuss the creation of the linguistic resources for the automatic annotation in Bulgarian based on the already existing resources for French.

**Method: Contextual Exploration** (J.-P. Desclés, 1997)
This is a purely linguistic approach that permits a relatively easy implementation to different languages, which we will illustrate with French and Bulgarian.

**Task:** Creation of the linguistic resources for the automatic annotation of *scientific articles* according to a semantic map for the task of *automatic summarization*.

Implementation in the EXCOM system for automatic annotation.
The discourse categories are language independent. They are used by an author in order to articulate the discourse structure of a text.
The discourse categories are language independent. They are used by an author in order to articulate the discourse structure of a text.

The linguistic resources for Bulgarian cannot be obtained by a simple translation of the resources for French.

- The meaning of the indicators, which are lists of words and expressions, is in most cases context dependent and their usage vary from one language to another.
- An analysis of the variations in the expressions and the different possible ways to express the discourse categories in each language must be carried out.
1. The discourse categories are language independent. They are used by an author in order to articulate the discourse structure of a text.

2. The linguistic resources for Bulgarian cannot be obtained by a simple translation of the resources for French.
   - The meaning of the indicators, which are lists of words and expressions, is in most cases context dependent and their usage vary from one language to another.
   - An analysis of the variations in the expressions and the different possible ways to express the discourse categories in each language must be carried out.

3. However, once the resources are created for one language, the linguistic analysis needed to transmit them to another language is considerably faster and simpler than the initial one.
The Bulgarian is a member of the Slavic language family.

Differences from other Slavic languages:

- No cases
- Definite article in postposition (and a marker of the subject for masculine nouns)
- No verb infinitive (replaced by the subjunctive)
- Non-evidential mood
• The Bulgarian Language has a rich morphology (esp. the verbal paradigm)

• Free word order: SVO and OVS word orders are possible, as well as other word orders

• Complete grammaticalisation of the punctuation:
  - The use of punctuation signs is syntactically determined.
  - The presence or absence of a punctuation sign (for ex. comma) influences the meaning of the sentence.

• Homonymy: it seems that in Bulgarian there is less homonymy than in French.
The Bulgarian Language has a rich morphology (esp. the verbal paradigm)

Free word order: SVO and OVS word orders are possible, as well as other word orders

Complete grammaticalisation of the punctuation:
- The use of punctuation signs is syntactically determined.
- The presence or absence of a punctuation sign (for ex. comma) influences the meaning of the sentence.

Homonymy: it seems that in Bulgarian there is less homonymy than in French.
The Bulgarian Language has a rich morphology (esp. the verbal paradigm)

Free word order: SVO and OVS word orders are possible, as well as other word orders

Complete grammaticalisation of the punctuation:
- The use of punctuation signs is syntactically determined.
- The presence or absence of a punctuation sign (for ex. comma) influences the meaning of the sentence.

Homonymy: it seems that in Bulgarian there is less homonymy than in French.
The Bulgarian Language has a rich morphology (esp. the verbal paradigm)

Free word order: SVO and OVS word orders are possible, as well as other word orders

Complete grammaticalisation of the punctuation:
- The use of punctuation signs is syntactically determined.
- The presence or absence of a punctuation sign (for ex. comma) influences the meaning of the sentence.

Homonymy: it seems that in Bulgarian there is less homonymy than in French.
Rewriting the linguistic resources for a second language

**Language 1 (French)**
- Linguistic study of corpora, manual annotation
- Analysis of the possible ways to express a category
- Lists of markers & CE rules (Language 1)
  - Generalisation

**Language 2 (Bulgarian)**
- Analysis of the lists and rules for Language 1
- Analysis of the differences between the two languages
- Variations in the expressions
- Lists of markers & CE rules (Language 2)
Example: A CE rule for Topic Announcements (French)

- **Indicator:** list *author-presentation*

  ```
  (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|sommettre|traiter)
  (Je|je|On| on|Nous|nous) (indiquerons|indiquera)
  (Je|je|On| on|Nous|nous) (vais|va|allons)[,]?(
  [\wàâëéêêëèêöûûûûûûïïÇÉÈÀÜÔÂÏ',-_]+){0,2}
  ```

- **Contextual clues:** list *current-document*

  ```
  (Ce| ce|Cet| cet|Cette| cette)
  ([\wàâëéêêëèêöûûûûûûïïÇÉÈÀÜÔÂÏ',-_]+)?
  (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)
  ([\wàâëéêêëèêöûûûûûûïïÇÉÈÀÜÔÂÏ',-_]+)?
  ```
Example: A CE rule for Topic Announcements (French)

- **Indicator**: list *author-presentation*

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

- **Contextual clues**: list *current-document*

  (Ce| ce|Cet| cet|Cette| cette)(
  ["\wáæééêêçöûûñüñíïçÉÉÀÜÔÀÏ', _,\-]+) ?
  (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)(
  ["\wáæééêêçöûûñüñíïçÉÉÀÜÔÀÏ', _,\-]+) ?

- **Context**: left / right   **Annotation**: Topic Announcement
Target sentences (French)

Example

“Je présenterai dans cet article un modèle automobile qui vient de sortir des usines de France”.
“I will present in this article an automobile model that has just left the factories in France.”

Example

“Dans le présent document nous discuterons les raisons de la disparition des dinosaures.”
“In the present document we discuss the reasons for the disappearance of the dinosaurs.”
In the present article we will propose two methods for automatic summarization.

"В настоящата статия ще предложим два метода за оценка на автоматично резюме."
Example

“In the present article we will propose two methods for automatic summarization.”

Absence of the contextual clue:

Example

“To clarify this idea we will propose to the reader to look at the following diagram.”
We could write the corresponding CE rule for Bulgarian, without any major changes:

- **Indicator**: list `author-presentation-bg`
  
  (ще|Ще|няма да|Няма да|смятам да|Смятам да|смятаме да|Смятаме да)?
  (представим|представя|изложа|изложим|разгледаме|разгледам|разглаждам|дали|далиме|дадем|дадеме|дал)
  ...

- **Contextual clues**: list `current-document-bg`
  
  (този|тази|тези|настоящият|настоящата|настоящите|текущият|текущата|текущите) ([а-я]+)? (документ|статья|записки|доклад|лекция|студия|текст|документи|статьи|записка|доклади|лекции|студии|текстове)
  (този|тази|тези|настоящият|настоящата|настоящите|текущият|текущата|текущите) (глава|статья|лекция|увод|изложение|заключение|част|секция) ...

Atanassova et al. Automatic Discourse Annotation of French and Bulgarian
We could write the corresponding CE rule for Bulgarian, without any major changes:

- **Indicator:** list *author-presentation-bg*
  
  (ще|ще|няма да|няма да|смятам да|смятам да|смятаме да|смятаме да)  
  (представим|представя|излага|изложим|разгледаме|разгледам|разгледаме|разгледам|дадем|дадеме|дам)
  ...

- **Contextual clues:** list *current-document-bg*
  
  (този|тази|тези|настоящият|настоящата|настоящите|текущият|текущата|текущите)  ([a-я]+)? (документ|статья|записки|доклад|лекция|студия|текст|документи|статьи|записка|доклади|лекции|студии|теккстове)
  (този|тази|тези|настоящият|настоящата|настоящите|текущият|текущата|текущите) (глава|статья|лекция|введение|изложение|заключение|часть|секция) ...

- **Context:** left / right  
  **Annotation:** Topic Announcement

Atanassova et al.  
Automatic Discourse Annotation of French and Bulgarian
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. We have to consider the different means that each language provides for the expression of the discourse categories.
Linguistic Differences: Example

- In French: “on présente” ("I/we/one present(s)") is sometimes used in scientific articles as an indicator of a Topic Announcement.
In French: “on présente” ("I/we/one present(s)") is sometimes used in scientific articles as an indicator of a Topic Announcement.

This form is well recognized by the rule for French, but it does not exist in Bulgarian.

The same meaning in Bulgarian can be expressed by a form in the middle voice: "се представя" ("is presented / presents").
Linguistic Differences: Example

- In French: "on présente" ("I/we/one present(s)") is sometimes used in scientific articles as an indicator of a Topic Announcement.

- This form is well recognized by the rule for French, but it does not exist in Bulgarian.

- The same meaning in Bulgarian can be expressed by a form in the middle voice: "се представя" ("is presented / presents").

- In order to take into account this difference we have to create a new CE rule for Bulgarian, which does not exist in French, and takes as indicators some similar verb forms in the middle voice.
Outline
Introduction
Linguistic Resources for Bulgarian Evaluation
Specifics of the Bulgarian Language
Creation of the Linguistic Resources
Accounting for the Linguistic Differences
Visualizations

Atanassova et al. Automatic Discourse Annotation of French and Bulgarian
Title: ЗА МИСЛОВНИТЕ ЕКСПЕРИМЕНТИ В НАУКАТА | Author: В. Грозданов

Sentences: 124 | Annotated sentences: 22

<table>
<thead>
<tr>
<th>Category</th>
<th>Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Announcement</td>
<td>20 Извънът преди метода на мисловните експерименти се състои в специфична комбинация на късна и логическо мислене, което води по картинно но убедителен начин до дефиниране на проблем, разкриване на парадокс или просто поставя не на интерен вопрос.</td>
</tr>
<tr>
<td></td>
<td>37 Целта е да се получат събития (физически) данни, които да бъдат интерпретирани и които не биха могли да се получат освен чрез физическата манипулация, т.е. физически реален експеримент.</td>
</tr>
<tr>
<td></td>
<td>89 Въпросът, който Айнштейн задава, е: “Каква е скоростта на светлината лъч спрямо влака?”.</td>
</tr>
<tr>
<td></td>
<td>117 Проблемите, породени от парадоксите на Зенон, емпирически ограничени възможности да се тестят проблеми като тези, изследван от Лукреций в мисловния експеримент за безкрайността на пространството, въпросът в дилемата, която поставя т АПР и Котка на Шрьодингер биха могли да илюстрират подобна теза.</td>
</tr>
<tr>
<td>Technical Description</td>
<td>17 Една от основните цели на мисловните експерименти е именно често залагащият резултат на тяхното заключение.</td>
</tr>
<tr>
<td></td>
<td>61 Ако една теория не издава хипотези, които са експериментално проверяеми, тя не се приема за научна.</td>
</tr>
<tr>
<td>Conclusivo Remark</td>
<td>04 В епистемологично спониране мисловните експерименти много наподобяват в ригор тип познание, т.е. оправдаване на предпоставките и заключението на експеримента без участие на спита.</td>
</tr>
<tr>
<td></td>
<td>76 Същността на централния дебат се свежда до произход на причините, които оправдават предпоставките и заключението на мисловните експерименти.</td>
</tr>
<tr>
<td></td>
<td>08 Получените резултати са логично противоречиво, което представлява класическо reductio ad absurdum на първоначалния тезис, че тежкият тела падат по бързо от леките.</td>
</tr>
<tr>
<td></td>
<td>111 Простата смъртта показа, че мисловнияят АПР, макар и със сега прието като проблематично заключение в оригиналната статия, спектира на науката бециени години като формулира дилема, над чието решение физиците все още работят усилено.</td>
</tr>
</tbody>
</table>
Recall

![Bar chart showing recall for different tasks and languages: Topic Announcement, Technical Description, Conclusive Remark, Judgement, and Navigation. The chart compares BG (blue) and FR (orange) languages.]
Precision

![Bar chart showing precision for different categories: Topic Announcement, Technical Description, Conclusive Remark, Judgement, Navigation. The chart compares results for Bulgarian (BG) and French (FR).]
Remarks

- The values of the recall in Bulgarian are a little lower than the values of the precision.

- The values of the precision indicate the noise in the system, which is due to some polysemic markers that have not been taken into account in the rules. These values can be improved by adding more specific contextual clues for the disambiguation of the polysemic markers.

- The values of the recall can be improved on one hand by the elaboration of the linguistic markers for the identification of more surface linguistic forms, and on the other hand by the introduction of new CE rules.
The values of the recall in Bulgarian are a little lower than the values of the precision.

The values of the precision indicate the noise in the system, which is due to some polysemic markers that have not been taken into account in the rules. These values can be improved by adding more specific contextual clues for the disambiguation of the polysemic markers.

The values of the recall can be improved on one hand by the elaboration of the linguistic markers for the identification of more surface linguistic forms, and on the other hand by the introduction of new CE rules.
**Remarks**

- The values of the recall in Bulgarian are a little lower than the values of the precision.
- The values of the precision indicate the noise in the system, which is due to some polysemic markers that have not been taken into account in the rules. These values can be improved by adding more specific contextual clues for the disambiguation of the polysemic markers.
- The values of the recall can be improved on one hand by the elaboration of the linguistic markers for the identification of more surface linguistic forms, and on the other hand by the introduction of new CE rules.
Thank you!