Discourse Automatic Annotation of Texts: an Application to Summarization

Antoine Blais, Iana Atanassova, Jean-Pierre Desclés, Mimi Zhang, Leila Zighem

LaLIC (LAngage, Logique, Informatique et Cognition) Laboratory
University of Paris-IV Sorbonne
Maison de la recherche, 28 rue Serpente 75006 Paris
{antoine.blais,iana.atanassova,jean-pierre.descles,mimi.zhang,leila.zighem}@paris4.sorbonne.fr

Abstract

The exploitation of the discourse structure of a text and the identification of the discourse categories are essential elements for the automatic summarization, as well as for the textual information retrieval. In this paper we will describe an automatic summarization strategy that uses these elements as the basis for the extraction of the most relevant textual segments that will constitute the summary. Certain linguistic markers allow us to annotate automatically a text according to discourse categories, in order to make visible the discourse structure and the discourse categories in the text. Our approach is domain independent and the discourse categories that we use for summarization are general for all natural languages. This makes it possible to apply our method to articles in various domains and in different languages.

Introduction to Automatic Summarization

We present bellow the two main approaches to build automatically the summary of a text (see for more details Mani 2001).

The first approach is the automatic summary production by *comprehension*. Originating from the domain of Artificial Intelligence, this approach considers the process of automatic summarization as being similar to some extent to the human summarization activity and the automatic summarization is based on the partial or total comprehension of the text. The program must be able to build a representation of the text, which might eventually be modified, in order to generate from it a summary. However, this method is quite difficult to carry out as it requires automatic text comprehension, text representation, as well as automatic text generation. The existing methods for these tasks are still quite unsatisfactory.

The second approach is the automatic summarization by *extraction*, which is inspired by the domain of Information Retrieval. The goal of this approach is to provide quickly a simple informative summary, without making a deep analysis of the text. In this method, we search and extract the most relevant textual segments (often sentences and paragraphs) in order to constitute an extract that we consider as the summary. The central procedure consists in evaluating the relevancy of textual segments according to one or more criteria. There exist two major methods to do this. The first one is the statistical method. It uses numerical methods to measure the relevance of a given text segment according to the presence of certain terms that are representative of the text (using a frequency calculation). Different heuristic criteria, such as the position in the textual structure or the presence of title's terms, could also be used. Other methods that rely more on linguistic knowledge use the presence of surface linguistic markers to establish the relevance of a textual segment. Some particular linguistic markers allow us to attribute a semantic (discourse or rhetorical) value to a textual segment, according to a linguistic theory, and thus to find out its relevance to the summary.

Among the advantages of the extraction method is that it does not make a deep analysis of the text and does not use text generation. On the other hand, it provides a summary by using simple algorithms of extraction and does not rely on any kind of text comprehension. The disadvantages of this method are often attributed to the lack of coherence of the summary and the fact that the broken interconnections between the different textual segments that have been juxtaposed could change the text's interpretation. Nevertheless, this approach remains the only possible one for the moment from the point of view of computer implementation.

Presently, the automatic summarization is more and more oriented towards the production of flexible summaries that correspond to some specific user needs. That is why the summarization strategies should allow the production of different summaries of the same text according to the needs of the user. Furthermore, the notion of automatic summarization tends to be integrated more and more with other similar applications that rely on common text processing methods.

The LaLICC laboratory of the university of Paris-Sorbonne (ParisIV) has been working for several years in the domain of automatic summarization. The realization of different projects, such as SERAPHIN (Berri 1995), SAFIR (Berri et al. 1996) and ContextO (Crispino et al. 2003), has led to some discussion and the production of

software in this field. So far, the automatic summarization task (Blais et al. 2006) has been incorporated into the EXCOM system (Djioua et al. 2006), the main function of which is to annotate texts automatically on the semantic and discourse levels. Up to now, the method presented here has been used for the automatic processing of a number of languages, such as French, Korean, Bulgarian, English, Arabic and Chinese.

Relevant Information in Texts

First we would like to mention that a summary must not only present a shorter reproduction of a text, but it should above all give information about this text. In the case of an automatic summary, it should provide enough information about the text, so that the reader could be able to decide whether to consult the text itself or not. In our approach to automatic summarization we intend to build up a text by eliminating some of the information in it but at the same time keeping only those textual segments that convey the most essential information expressed in the original text. We qualify this information as relevant because it expresses sufficiently the content of the text. We consider a textual segment as relevant for the summary if it contributes to construct a set of segments that provides a general and coherent idea of the text content.

Our work on automatic summarization is focused on scientific articles and more generally on texts where argumentation takes an important place. The main function of this kind of texts is relatively precise and, in general, their task is to convince (Swales 1990) and also to inform the reader about a research work, a theory, etc. Thus, the most relevant elements of the text are those that convey in the clearest way the information of the text and above all the information that the author wants to express, such as the topic announcements of the author or the general conclusions of the scientific article.

Some of the current (in particular statistical) methods of automatic summarization process the text without taking into account the discourse organization of the text (Charolles 2002) and the discourse categories present in it. They consider only the physical structure of the text and do not make any discourse decomposition of the text. Therefore, they place all textual segments at the same semantic and discourse level. A significant weakness in many approaches to summary construction is the fact that there is no differentiation between, for example, the general topic announcement of the author or the general hypothesis in a scientific article and the rest of the text.

Our approach, which uses a surface linguistic analysis of texts, differentiates the textual segments (on the discourse level) by a procedure of automatic annotation of the text which will be presented below. Prototypical discourse categories in our work are TOPIC ANNOUNCEMENTS, CONCLUSIVE REMARKS, JUDGEMENTS and COMMENTS.

The discourse annotation of texts could also serve as the basis for other tasks, apart from the automatic summarization, such as information retrieval for special requests. We can retrieve, for example, conclusions, opinions or hypotheses of the author given in the text. We note the approach of Teufel and Moens (Teufel and Moens 1999, Teufel 1998), which also uses differentiation of sentences in scientific articles at the level of their rhetorical value according to the general argumentation. We note too the works of Marcu (Marcu 1998) on automatic summarization who uses (also takes into account) the rhetorical structure of the text on the basis of the Rhetorical Structure Theory (Mann and Thompson 1988). Our method is also based on the differentiation between discourse categories that enter in a strict hierarchy, but unlike the two approaches above, we introduce the categories in a different order since we use our own relevance measures of the categories for the informative summary of scientific articles. Moreover, in our algorithm the linguistic markers and rules are based on linguistic studies of corpuses and not on machine learning.

Discourse Annotation and Relevance Retrieval in a Text

Which Textual Segments Must We Process?

The textual segmentator SEGATEX (Mourad 2001), designed in the LaLICC laboratory, carries out the segmentation of a text by an analysis of the textual typography. From a text file, SEGATEX creates a new file in XML format (Fig. 1) where the physical structure of the text is presented through tags delimiting the different textual elements (titles, sections, paragraphs and sentences).

```
<section ID=1>
<titre>Title</titre>
<para ID=1>
<phrase ID=1>First sentence.</phrase>
<phrase ID=2>Second sentence. </phrase>
</para>
<para ID=2>
<phrase ID=3>Third sentence. </phrase>
<phrase ID=4>Fourth sentence. </phrase>
</para>
</section>
```

Fig. 1. Annotated file

In our work we consider the sentence as the basic extraction unit. Therefore, sentences will be evaluated according to their relevance to the summarization.

Some works prefer the paragraph as an extraction unit for the summary construction, and also for the task of textual information retrieval. In general, the reason for the extraction of paragraphs for the summary is that they provide better cohesion and readability (Halliday and Hasan 1976) of the final result. For example, non-resolved anaphoras are less frequent because the antecedent of the anaphora is usually in the same paragraph. In the case of

sentence extraction, the anaphoric links are more often broken. However, we note that a paragraph can contain sentences that do not express any relevant information. The same information could be sufficiently and clearly expressed by only one part of the sentences in the same paragraph, hence our approach limits the noise.

Discourse Annotation for Relevant Information Retrieval

The Relevant Discourse Categories for Summarization

We have chosen in our work to summarize scientific articles and documents. For each text type (argumentative, informative, narrative ...), the relevant information is not necessarily located in the same parts of the discourse structure. Actually, in each text type, some textual segments tend to hold more relevant information than others. In scientific articles some categories are more important than others for the summary: for example, the general topic announcement of the article is more informative than an example, a quotation or a part of an argument. Therefore, in our approach the retrieval of topic announcements has a priority over other discourse categories.

Here we give the main discourse categories and sub-categories that we consider as the most relevant for the summary (particularly for scientific articles):

(a) TOPIC ANNOUNCEMENT, which is divided into six sub-categories: THEMATIC PRESENTATION, THEMATIC DESCRIPTION, DOCUMENT DESCRIPTION, HYPOTHESIS, METHOD, GOAL and PROBLEM.

(b) CONCLUSIVE REMARK, which is divided into two sub-categories: CONCLUSION, RECAPITULATION.

(c) RESULT/EVALUATION, which is divided into two sub-categories : RESULT and EVALUATION.

(c) JUDGEMENT, which is divided into two sub-categories: OPINION and EMPHASIZED COMMENT.

(d) COMMENT, which is divided into three sub-categories: CONSEQUENCE, REFORMULATION and RECALL.

The different discourse categories and sub-categories above have been selected in our study as being the most relevant in the case of scientific articles. However, it could also happen that, in some cases relevant textual segments for this text type belong to other discourse categories that are not presented here. But we suppose that the categories chosen here are more frequently used by authors to express important information. We note that all these categories are domain independent; they can be found for example in biological articles, as well as in philosophical or psychological articles.

We consider that the annotated segments, such as topic announcements, are the most relevant ones for the summary because their role is to indicate what the text is about (the subject), the way it will be explained and analysed (the description and the method), and the reasons for it (the goal). These segments are the most important and the most informative about the document content; consequently, their extraction for the summary is fundamental. After them, we order and select the conclusive remarks, the judgements and the comments.

Localization of Relevant Segments

In the classical systems of information retrieval and text mining, textual segments are often extracted according to criteria based on the frequency of some terms that, by being present in the segment, can qualify it as relevant or corresponding to a specific request. These systems are nevertheless confronted to two main inconveniences.

The first inconvenience is their incapacity to recognize whether sentences or other textual segments belong to specific discourse categories localized in texts. This is an important problem because some discourse categories are more relevant than others according to the text type. Furthermore, these methods do not meet the needs for some specific user requests, such as the extraction of all the conclusions or hypotheses presented in scientific articles.

The second inconvenience is their incapacity to give the context of validity and distinguish the author's attitude towards a given sentence (Jackiewicz 1999), which is essential to establishing its relevance. In a summary, comments assumed by the author and those he takes distance from do not have the same importance for the reader.

In order to automatically distinguish which segments in texts belonging to the discourse categories presented here, we use The Contextual Exploration method (Desclés 1997). This is a method for discourse annotation of textual segments according to the presence of linguistic markers, which consists in the location in a segment of *indicators* corresponding to linguistic markers (words, expressions…), which have a particular textual function, representative of a discourse category. These linguistic markers have a fixed usage and are relatively independent from the authors' styles of writing.

However, sometimes the simple presence of an *indicator* does not permit the annotation of the textual segment, because the discourse value of the segment can change according to the context. The discourse value that we want to find by an indicator must be evaluated by removing the semantic indecision by applying contextual exploration rules. Those rules consist in localizing in the textual context of the indicator one or more linguistic clues allowing the removal of the semantic indecision and the segment annotation. More than one clue can be associated to one indicator in order to confirm or negate a specific discourse value.

The Contextual Exploration method consists of:

- Indicators that correspond to linguistic markers (words or expressions) belonging to discourse categories.
- Clues that are linguistic elements (words, expressions, typographic marks…) associated to an indicator for a

specific discourse value.

- Contextual exploration rules that apply the clues research in the textual context of the indicator to remove the semantic indecision.

Let's take an example:

(a) "***I propose*** *in this article a detailed demonstration of the disappearance of the dinosaurs*."

(b) "*To give an idea of it,* ***I propose*** *to you to look at the image bellow."*

We consider the segment *I propose* as an indicator for a topic presentation of the author. Nevertheless, its presence is not enough to consider the sentence as a topic presentation. As shown in example (b), this segment can occur also in a sentence that does not refer at all to what is presented in the document. In example (a), there are two clues that confirm the discourse role of the sentence. The first one is *a detailed demonstration* which is the result of an act of thought or speech. The second one is *in this article*, which links the topic presentation with the current document. Therefore, we could say with certainty that this sentence is a topic presentation of the author referring to the current document. On the other hand, in example (b), we do not have enough clues to confirm that the sentence is a topic presentation.

The EXCOM system uses contextual exploration rules (Djioua et al. 2006), that aim to annotate the text on the discourse level by searching for indicators and clues. Applying those rules the system finds automatically textual segments (here sentences), that belong to discourse categories considered as relevant for the summary. The rules are written in XML (fig. 2) format and processed with XSL rules, the latter being included in the general application, the EXCOM system for automatic annotation.

```
<!-- RPresentationThematiqueAuteur4 : je presenterai de
facon detaillee ... -->
- <regle nom_regle="RPresentationThematiqueAuteur4"
  tache="resume"
  point_de_vue="annonce_thematique|annonce_thematique_adv"
  type="EC">
  - <conditions>
    <indicateur espace_de_recherche="phrase"
      type="annotation" valeur="forme-presentation-
      auteur" />
    <indice contexte="droit" espace_de_recherche="."
      type="annotation" valeur="indice-thematique" />
  </conditions>
  - <actions>
    <annotation type="ajout_attribut" espace="identique"
      annotation="presentation-thematique-auteur1" />
  </actions>
</regle>
```

Fig. 2. Contextual exploration rules

The different linguistic markers (indicators or clues) are stored in files either as lists of words or regular expressions if they are complex (fig. 3).

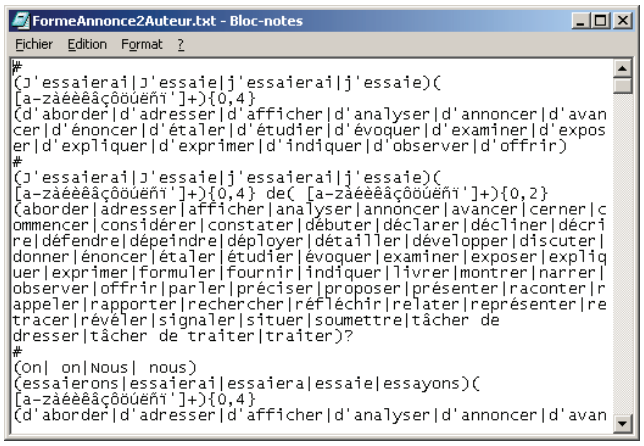

Fig. 3. Complex linguistic markers

Using the linguistic resources, EXCOM carries out the discourse annotation of the text. The EXCOM system adds discourse information to every sentence of the input segmented file, to which the rules of contextual exploration are applied. The discourse annotation attributed to sentences corresponds to new meta-data associated to the text: they are either in the same file that contains the text structured in XML format, or in a separate file containing all the discourse meta-data (XLink structures, see fig. 4 for an annotation example).

```
- <annotation ID="annotation_semantique1">
    <segmentRef xlink:type="simple" xlink:href="#12"
      type_segment="phrase" />
  - <annonce_thematique
      libelle_annotation="lalic.excom.resume.annonce_thematique.annonce
      -thematique-auteur">
    - <regle num_regle="RAnnonceThematiqueAuteur">
        <indicateur>Nous présenterons ensuite,</indicateur>
      </regle>
    </annonce_thematique>
  </annotation>
</Excom>
```

Fig. 4. Discourse annotation in XML format

The Process of Summary Construction

We present below the different steps in the process of summary construction. It uses software modules that are completely automatic.

Step 1: Thematic terms extraction. We find and extract the thematic terms. We consider as such words (here common and proper nouns) that are representative of the subject of the document. We extract them from the titles and sub-titles of the document, assuming that these often contain thematic terms because of their function of reference introducers (Jacques 2005).

Step 2: Discourse annotation of texts according to relevant discourse categories for the summarization. The EXCOM system detects different linguistics markers in the text and applies contextual exploration rules in order to annotate textual segments. We note that all rules can use (notably as clues) thematic terms extracted during the first step. So,

after this procedure, we obtain a discourse annotated text.

Step 3: Summarization strategy. We evaluate the relevance of each of the annotated sentences in the text according to their discourse categories, their position in the textual structure, and the presence of thematic terms. As mentioned above, some discourse categories are more relevant than others; we order them according to a predefined hierarchy. The position of the sentence in the textual structure is an additional criterion of evaluation. For example, a conclusion at the end of the text is more important than a conclusion in the middle of the text, as its position at the end of the text confirms that this is the general conclusion of the document. Finally, we verify the presence of thematic terms in the sentence. If it contains thematic terms, we can link it to the subject of the document or to one of its sub-parts. Note that it is the discourse annotation attributed to the sentence that constitutes the main criterion of relevance evaluation. The position in the textual structure and the presence of thematic terms are only additional clues to the relevance evaluation of the sentence. Thus we obtain a set of relevant sentences (the number varying according to the size of the summary), in the same order as they occur in the original text; this set of sentences is used for the summary by extraction.

Step 4: Cleaning-up step. Finally, we proceed to clean-up the summary in order to improve its cohesion and readability. We remove and add certain elements in order to improve the readability of the summary. For example, we remove the enumerations (firstly, secondly…), as enumerative series are likely to appear incomplete in the summary due to their partial extraction.

Step 5: Visualization. The user can visualize the summary in a browser (fig. 5) where the various discourse categories of the summary are clearly shown. It is also possible to display some additional information related to the sentences: description of the discourse category, position in the textual structure, thematic terms, etc.

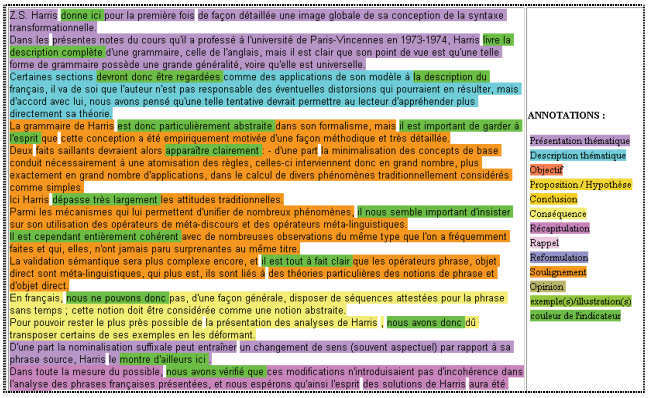

Fig. 5. Colored text presentation

Flexible Extracts and Information Retrieval

Once the text is annotated according to discourse categories, different visualizations could be created of this annotation so as to facilitate the reading and navigation in the text. What is more, the discourse annotation can also serve as the basis for information retrieval in correspondence with some specific user demands.

As another possible application of the discourse annotations we will consider here flexible extracts according to a predefined model. We create automatically structured flexible extracts that correspond to a given document type and to a class of user demands. We define a flexible extract model as a set of discourse categories to be extracted that would represent best the contents of the original document and make the information more accessible. For example, in the case of scientific articles, a flexible extract model would contain the categories that are the most important for this genre, namely the topic announcements, the hypotheses, the conclusions, etc.

Flexible extracts are very useful because they can help the user for the task of information retrieval, especially for some specific requests. For example, if the user has to go through a large number of scientific articles, they could use extracts in order to get a general idea of the articles' contents and to decide which ones are worth reading. Moreover, in such a situation flexible extracts could serve as a navigation tool (Teufel and Moens 1999). By using flexible extracts the system could answer to some more specific user requests concerning the discourse structure of the texts. For example, the user could find answers to questions like: "What are the hypotheses used in this article?", "What are the conclusions?", etc.

Multilingual Approach

We claim that our approach is based on discourse categories that extend across different languages and therefore it is language independent. The discourse categories that we consider are in their substance the means used by the author to enunciate the discourse structure itself. In particular, for scientific articles it is clear that certain discourse categories, such as topic announcements, conclusions, hypotheses, etc., are present in this type of text and this applies to articles written in any language. These categories are proper to the discourse structure for the scientific genre and are language independent.

Having said that, we consider that our method is valid and can be applied to any natural language. In the LaLICC laboratory, discourse automatic annotation is already carried out for French, Korean, Bulgarian, English, Arabic, and Chinese. Texts in all these languages are processed using the Contextual Exploration method.

This method is not only language independent, but also once the necessary linguistic resources are created for a given language, they can be relatively easily transmitted into other languages by a linguistic analysis of the indicators and clues corresponding to each discourse

category. We note that the indicators and clues constitute words and expressions that are language specific and they cannot be obtained by a simple translation of the linguistic markers from another language. However, the discourse categories and the annotation methodology remain the same for all languages.

Conclusions and Future Work

We have presented in this article a method of summary production by the extraction of sentences form a text.

In our work we insist on three important points:

- The goal of the summarization strategy is to extract the relevant information using the presence of linguistic markers (to determine the discourse role of sentences), the physical structure (the place of the sentence in the textual structure) and some of its elements (titles and sub-titles for the extraction of thematic terms). Several criteria of relevance evaluation and sentence selection ensure a better and reliable extraction. It is important to say also that our approach is in principle domain independent. The linguistic markers that we use can be found in any type of scientific articles in any domains.
- The author's attitude toward a sentence is also an essential element in our approach, since in the case of scientific articles, the comments that are assumed by the author do not have the same relevance as other comments. We distinguish in our analysis the comments that the author assumes (by the use of linguistic markers) and the rest. We insist on this, because in other information retrieval or automatic summarization systems this distinction is often not taken into account.
- The automatic annotation of a text with several discourse categories can be used not only for summary production, but also to make a synthesis of the text according to one or more discourse categories. Having obtained the discourse annotations of the textual segments, the system can eventually respond to more specific requests and give as a result only the annotated sentences corresponding to the request. We offer thus the capacity to find information in the text through different discourse categories: for example, the user can consult all conclusions or hypotheses contained in scientific articles.

Our future work will be directed towards processing a huge number of texts using the strategy described above and making an evaluation. We would also like to improve the post-processing algorithms, which would result in a better readability of the automatically obtained summary.

References

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., Minel J-L., 1996, SAFIR, système automatique de filtrages de textes. *TALN'96*, Marseille.

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

Charolles M., 2002, Organisation des discours et segmentation de écrits. Inscription Spatiale du Langage : structures et processus, IRIT, Toulouse.

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.

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, Floride.

Halliday M.A.K., Hasan R., 1976, *Cohesion in English*. Longman, England.

Jackiewicz, A., 1999, Causalité et prise en charge énonciative. Etudes Cognitives, n°3, *Académie Polonaise des Sciences*, Varsovie 249-269.

Jacques M-P., 2005, Structure matérielle et contenu sémantique du texte écrit. *CORELA*, Volume 3, numéro 2.

Mani I., (2001), Automatic Summarization. John Benjamins Pub Co, ISBN 1-58811-060-5, Amsterdam.

Mann W. C., Thompson S. A., (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text: An Interdisciplinary Journal for the Study of Text, 8(2)

Marcu D., 1998, Improving Summarization through Rhetorical Parsing Tuning. *Proceedings of the COLINGACL Workshop on Very Large Corpora*. Montreal, Canada.

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, Univ, Paris-Sorbonne.

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

Teufel S., 1998, Meta-discourse markers and problem-structuring in scientific articles. Workshop on Discourse Structure and Discourse Markers, *ACL 1998*, Montreal.

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