LEXICON OF COMMON SCIENTIFIC WORDS AND EXPRESSIONS FOR AUTOMATIC DISCOURSE ANALYSIS OF SCIENTIFIC AND TECHNICAL TEXTS *

Elena Bolshakova

Abstract: Various NLP applications require automatic discourse analysis of texts. For analysis of scientific and technical texts, we propose to use all typical lexical units organizing scientific discourse; we call them common scientific words and expressions, most of them are known as discourse markers. The paper discusses features of scientific discourse, as well as the variety of discourse markers specific for scientific and technical texts. Main organizing principles of a computer dictionary comprising common scientific words and expressions are described. Key ideas of a discourse recognition procedure based on the dictionary and surface syntactical analysis are pointed out.

Keywords: scientific discourse, discourse markers, common scientific words and expressions, scientific discourse operations, discourse-compositional analysis of scientific and technical texts.

ACM Classification Keywords: I.2.7 [Artificial Intelligence]: Natural language processing – Text analysis

Introduction

Functional style of scientific and technical prose is admittedly the most distinctive one, primarily due to intensive use of scientific phraseology and structuring. The phraseology includes, besides scientific and technical terms of a specific terminology, various expressions of common nature, such as English expressions *exploratory study, mentioned above, for this reason, in addition, therefore,* etc. and *Russian: вышеупомянутый, по этой причине, в дополнение к, далее мы опишем* and so on. We call such lexical items *common scientific words and expressions*.

Within scientific discourse (scientific speech), terms and common scientific expressions differ in their functions. Specific terms denote concepts, objects, and processes of the particular scientific domain, whereas common scientific expressions are domain independent: they are used to design and organize scientific text narrative by expressing the logic of reasoning, by connecting text fragments devoted to different topics and subtopics, and by structuring the text under development.

Similar to terms, common scientific expressions present a syntactically quite heterogeneous set, and to an even greater degree than terms. The set comprises, besides content (autosemantic) words, functional (auxiliary) words. Noun and verb-noun combinations, adverb and participle expressions, compound prepositions and conjunctions are included as well. Among the word combinations, one can notice stable expressions exploited as ready-for-use colloquial formulas (clichés) [1], such as *Eng. as it was stated above*, *to outline directions of further research; Rus. из вышесказанного следует, как показало проведенное исследование*. Some clichés are common for scientific and technical prose, the others are specific for particular genres. Certain common scientific words and expressions are known as discourse markers [11, 12].

The paper reports on preliminary results of an ongoing research aiming at elaboration of a procedure for discourse analysis of scientific texts, as well as development of an adequate computer dictionary of common

* Work is supported by the grant 06-01-00571-a of Russian Fond of Fundamental Researches.

scientific words and expressions. This is done within overall framework of creating computational models of scientific and technical prose. Our basic claim is that in order to attain a really effective automatic processing of scientific and technical texts (which is needed for many applications, in particular, for text summarizing [5]), we should take into account functional peculiarities of scientific prose on various levels, in the first place, levels of phraseology and discourse.

The research involved an empirical study of scientific texts in several fields of exact and natural sciences, so that scientific papers as the core of the functional style were analyzed. The study was initially performed for Russian texts, and then expanded to English. As the work progressed, the importance of the common scientific lexicon became increasingly obvious, despite of its relatively small size. In both languages the principal features of scientific lexicon and discourse proved to be the same, which emphasizes the international character of scientific and technical prose.

Advancing towards an appropriate procedure of discourse analysis, we create at the first step a computer dictionary that comprises a wide range of common scientific words and expressions and provides a classification of their syntactic and semantic features. For Russian, the dictionary is now partially implemented; for English, only the classification work was done so far.

In comparison with DiMLex lexicon of discourse markers [11], which was developed for German and English and mainly consists of conjunctions and conjunctive adverbs, our dictionary covers a wider set of lexical units, because we consider any lexical device signaling scientific discourse as its marker (e.g., English expression *by definition* or Russian *по определению*).

A discourse-analyzing procedure is also under development now. Since units of common scientific lexicon may be served as surface cues, we assume the hypothesis that shallow text analysis based on the lexicon is adequate for detecting discourse structure of scientific text, without a deeper syntactical-semantic analysis of all its sentences. Instead of concept of discourse relationship, which is proposed in well-known discourse theory RST [7] for explaining relations between adjacent phrases in text, we rely on the concept of *scientific discourse operation* for recognition of discourse-compositional text structure specific for scientific texts. Our discourse recognition procedure also differs from the procedures that were developed for Japanese texts [6, 9] and based on deep syntactical analysis of sentences, with consideration of style-independent discourse markers.

The objectives of this paper are:

- To determine the set of common scientific words and expressions we regard as discourse markers;
- To describe designing principles for the corresponding computer dictionary;
- To sketch the procedure for recognition of discourse-compositional structure of scientific texts;
- To point out potential application of the dictionary and the procedure being design.

To clarify our ideas we begin with an overview of specific features of scientific discourse, which are derived from our empirical study. The features proved to be language-independent, and we give in the paper illustrative examples from both languages: English and Russian.

Scientific Discourse and Its Devices

Discourse consists of interrelated speech acts determined by communicative goals. The global purpose of scientific communication is to convey new ideas and results of scientific research, as well as to explain and rationalize them. Therefore, scientific discourse involves reasoning that is organized as a sequence of mental operations of informing and arguing. Among typical operations we should point out assuming hypotheses,

defining new terms, determining causal relations, exemplification, resuming and so on. We will call such intellectual operations *scientific discourse operations*.

As a tendency of scientific and technical prose to be strict and plain, these discourse operations are usually introduced into texts and more or less explicitly marked by authors of texts with the aid of lexical devices – common scientific words and expressions. With this function, the words and expressions pertain to metatext component of discourse [12] and are called *discourse markers*.

We consider scientific text as composed of *discourse segments*, each segment including several adjacent sentences and corresponding to the applied discourse operation. Some sentences include discourse markers.

The most evident markers of scientific discourse are *mental performative expressions*, or *performative formulas* like *Eng. we conclude, we would assume* or *Rus. мы докажем, мы предположим*. For Russian, they are described in detail in [10]. Performative formulas are based on "mental" verbs, e.g., *Eng. to conclude, to consider, to admit, to propose; Rus. заметим, рассмотрим, выразим* and so on. As a rule, these verbs explicate particular steps of scientific reasoning and have valences (complementing arguments): *Eng. we consider N*, *we conclude that S* ; *Rus. рассмотрим N*, *подчеркивается, что S*. Besides pure "mental" verbs (*to conclude, to assume,* etc.), verbs of physical action (*to see, to show,* etc.) are used as mental. The class of such verbs is open, since various verbs may be potentially used as mental in the context of scientific discourse.

There are various forms of mental performative expressions in scientific texts:

- Canonical forms, with mental verb in the second person plural, often with the corresponding pronoun (e.g., *Eng. we resume, let us proceed to, we will proceed; Rus. мы покажем, мы рассмотрим*);
- Verbal variants (*Eng. summing up, strictly speaking; Rus. подводя итоги, строго говоря*), which are often used together will canonical forms (*Eng. refining the definition, wee see that…; Rus. суммируя вышесказанное, укажем…*);
- Impersonal forms (*Eng. it should be added, it was found, it is reasonable to assume; Rus. необходимо/нетрудно заметить, представляется, что..*), which often include words of author's estimation (*should, reasonable, necessary*);
- Descriptive variants (*Eng. N is briefly described, N are given in; Rus. N кратко описано*)

Verbal and impersonal forms are used in texts to paraphrase canonical forms (e.g., *it was found* instead of *we found*) or to give some cross-text references (e.g., *as it was stated above*). Though they are less explicit forms than canonical, they are functionally equivalent. One can also find 'hidden' performatives in scientific and technical texts, which we call descriptive variants: e.g., *These data are given in Table 3* stands for *We gave these data in Table 3*.

Mental discourse operations might be expressed by various parenthetic words and expressions: indicators of order (e.g., *Eng. first* or *lastly; Rus. во-первых, наконец*), markers of equivalency (e.g., *Eng. in other words; Rus. иными словами*), various connectives between textual parts (e.g., *Eng. nevertheless* or *so far; Rus. тем не менее, благодаря тому, что*) and so on. The metatext nature of these discourse markers is more obvious, they are typical not only for scientific and technical texts.

Among words typical for scientific discourse we should mention abstract nouns, such as *problem*, *analysis*, *model, concept, conclusion* and so on. They are aimed to name mental constructs by which scientific information is semantically structured. We call such nouns *common scientific variables*, since they have the obligatory semantic valence (*problem of N, model of N*). Common scientific variables are mainly used with mental performative verbs, thereby forming stable noun-verb combinations, such as *Eng. to test hypothesis* or *to draw conclusions; Rus. подвергнуть анализу, проводить аналогию, опровергнуть гипотезу* [4]. Meanings of such verbs are close to Mel'čuk's Lexical Functions [8] with corresponding nouns as arguments.

Below we present several English text fragments from the book on artificial intelligence and from the papers [5,11], which illustrate the usage of common scientific words and expressions (they are underlined):

In fact, notice that the value of the slot Players is a set. Suppose, instead, that we want to represent the Dodgers as a class instead of an instance. … For example, we could make it a subclass of major league baseball players. (1)

According to our corpus study, we have identified three basic rhetorical configurations for summaries, that we call Meta-Schemas. (2)

For dealing with discourse markers, we do not regard this distinction as particular helpful, though. As we have illustrated above and will elaborate below, these words can carry a wide variety of semantic and pragmatic overtones, which render the choice of a marker meaning-driven, as opposite to a mere consequence of structural decisions. (3)

Besides common scientific lexicon, non-lexical devices are used to organize scientific discourse. In particular, such devices as sections, paragraphs, items, rubrics, and numeration are intended to structure scientific texts and to form their composition. All structuring and discourse-organizing devices present an interconnected system: devices can complement or substitute one another. For example, section headings are really substitutes for performative expression *we proceed to*, whereas numeration often complements performative formulas: e.g. *Let us enumerate main statements: 1)…2)…* . This interconnected system is rather excessive, since for most discourse operations there exist collections of similar lexical markers, but at the same time it allows for flexible paraphrasing.

In general, some discourse operation with its lexical and non-lexical devices can be used to implement another operation. For example, for categorization, a definition of new term is often required. Therefore certain discourse segments are embedded into some others, and in this way hierarchical structure of scientific text is formed.

Computer Dictionary of Common Scientific Lexicon

To develop a computer dictionary, collections of Russian and English common scientific words and word combinations were gathered from few available text dictionaries of scientific phraseology [3, 4] and from scientific texts in several fields of science (mainly in computer science and artificial intelligence), through their manual scanning. While selecting a word or expression for our collection, we used the following non-formal criteria. First, discourse-organizing function of the word or expression should be evident, second, it should be rather frequently used in texts in several fields. Inter-language correspondences were used: for Russian expressions English equivalents were looked for, and vise versa.

Heterogeneous collections of words and word combinations were classified according to their discourse-organizing functions in scientific texts, irrespectively of their grammatical form and syntactic features. Based on our study, we propose for classification the list of scientific discourse operations, the most significant are given in Table 1:

Table 1. Scientific discourse operations

Operation	Russian Examples	English Examples
Description or statement	*укажем, что; характеризуя*	*let us to describe; we point out that*
Elaboration or adding information	*в частности; в дополнение к*	*to be more precise; in addition*
Expressing relations of causal, conditional, and concession type	*по этой причине; следовательно*	*hence; provided that; however*
Actualization of the topic	*перейдем к; рассмотрим*	*as for; let us consider; regarding*

Emphasizing	*особо подчеркнем;* *необходимо отметить*	*first of all;* *it is necessary to emphasize*
Presupposition	*предположим/допустим, что*	*we would assume; it may be admitted*
Definition	*будем называть; по определению*	*by definition; we call it/them,*
Comparison	*по сравнению с*	*as compared with*
Contraposition	*с одной стороны*	*on the one hand; as opposite to*
Illustration or exemplification	*к примеру; например*	*as illustrated below; for example*
Generation or resuming	*суммируя вышесказанное; в общем*	*in general; summing up*
Enumeration or ordering	*во-первых; наконец*	*next; finally*
Labeling with a scientific variable	*идея; модель; результат*	*result; idea; model*
Expressing of author's attitude	*целесообразно считать;* *по всей видимости*	*in our opinion*; *it seems reasonable*

Our collection of common scientific words and expressions was divided into functional classes in accordance with the proposed list of discourse operations. Within each class, all words and word combinations that are semantically close and interchangeable in the texts as discourse markers were gathered into a group, thereby giving a subclass of functionally equivalent markers. Each group of functional equivalence often includes words of different parts of speech and contains from 2 to 9 units, the number depending on the language. For example, the resulted group of the consequence relationship includes for English: *hence, therefore, as a result, consequently, it follows that, we conclude that* etc., and for Russian: *значит*, *итак*, *таким образом*, *тем самым, как видим* etc. For both English and Russian, we obtain 53 groups corresponding to particular discourse organizing functions.

To determine lexical entries of our computer dictionary, we considered requirements for its use by automatic text processing system, first of all, by discourse-analyzing procedure. The dictionary contains:

- Units corresponding to words of common scientific lexicon. They comprise both functional and content words, including those encountered only within scientific expressions.
- Units corresponding to common scientific word combinations.

For a particular word, each unit stores adequate morphosyntactic information, including the part of speech and the flexional class (if any), as well as pointers to dictionary units describing available combinations with this word.

In turn, each unit for a particular word combination accumulates necessary syntactical properties of the combination: stable *vs*. free, continuous *vs*. discontinuous. Since most word combinations have syntactic valences, we propose to represent information about valences with the aid of special ***lexicosyntactic patterns***.

Each lexicosyntactic pattern fixes lexemes (constituent words of the particular combination) and their grammatical forms, as well as specifies syntactic conditions necessary for filing its empty slots (valences of the fixed lexemes). An example of such a pattern is *"let us consider" NP* with *NP* denoting a noun phrase. Another example is *NP "we will call" T*, where *T* denotes an author's term and *NP* is a noun phrase explaining its meaning; it describes the typical English expression for definition of new terms.

A formal language for specifying lexicosyntactic patterns was elaborated, as well as a methodology for acquiring new patterns for the particular discourse operation from scientific and technical texts. Lexicosyntactic patterns proved to be a convenient device for describing stable colloquial expressions comprising both phrasal formulas (like *the paper describes main features of*, *argument can be made against*) and predicative constructs (such as *to take as starting point for*). Based on the acquiring methodology, a collection of patterns was created, which

describes typical Russian single-sentence definitions of new terms. For example, one of lexicosyntactic patterns for discourse operation of defining a new term is

«под» NP1 <case=ins> V<пониматься; tense=pres, person=3> NP2 <case=nom> < NP1.numb=V.numb>

Particular lexemes of the pattern are quoted, letter *V* denotes the verb, NP1 and NP2 denote noun phrases, and grammatical conditions are written within angle brackets – they specify values of grammatical parameters (tense, person, case, number) or establish their equality. The pattern describes both Russian sentences *«Под графемной конструкцией понимается графическая форма, построенная из базисных, проблемно-ориентированных и/или графических конструкций»* and *«Под данными при такой формализации понимаются последовательности символов в некоторых алфавитах»* (in these sentences fixed lexemes of the pattern are underlined) .

In addition, for each dictionary unit considered as discourse marker, the developped computer dictionary provides semantic information that facilitates recognition of underlying discourse operation, namely:

- Functional class and group of the unit within the proposed semantic classification;
- Contextual conditions necessary for being discourse marker within texts;
- Information about size and boundaries of implied discourse segment (the segment consists of one sentence or of several sentences; the dictionary unit marks the beginning or the end of discourse segment).

Discourse-Analyzing Procedure

Our study of scientific discourse showed that common scientific lexicon has its own functional semantics, which makes it possible to superficially read scientific texts, i.e. to derive underlying discourse operations and to comprehend logic of scientific reasoning, without deep understanding of these texts. So we are developing our procedure for recognition of discourse-compositional structure of scientific texts on the basis of shallow text analysis and the described computer dictionary.

We consider discourse-compositional structure of scientific text as hierarchical structure of sequenced and embedded discourse segments, which corresponds to applied discourse operations and applied structuring devices. The structure may be represented as a tree, with tree nodes corresponding to discourse segments, and tree links fixing semantic (in particular, causal) and structural (in particular, embedding) relations between segments. In order to construct such a tree for a given text, the proposed recognition procedure takes the following steps:

1. Grapheme analysis of words, delimiting of sentences, and detecting of text composition elements: section headings, paragraphs, items, rubrics, and numeration.
2. Morphologic analysis of words and identification of occurrences of common scientific words and word combinations.
3. Recognition of dictionary discourse markers in the given text through matching text fragments with dictionary lexicosyntactic patterns that contains identified common scientific words.
4. Delimiting of discourse segments, based on recognized discourse markers and semantic information presented in the dictionary for functional groups and classes. In general case, the result of the segmentation is ambiguous, since several plausible discourse trees fit the sequence of identified markers.
5. Selection of the most plausible discourse-compositional tree within the set resulted at the previous step. A number of heuristic rules are used for this purpose, for example, an exemplifying segment is rather embedded into another segment than embeds it.

To implement steps 3 and 4 surface syntactical analysis of sentences is needed, which takes into account:

1. agreement and coordination of words;
2. overall grammatical structure of sentences.

It should be noted that reliability of discourse recognition depends on several factors, among them are the number and types of discourse markers encountered in the text. In order to increase the reliability, the other linguistic devices, in particular, anaphoric links and repetitions of lexical units in adjacent sentences are to be considered.

Conclusion

We have overviewed the features of scientific discourse and the spectrum of common scientific words and expressions, with their role in scientific discourse. We described main organizing principles of the computer dictionary of common scientific lexicon, which provides various information valuable for automatic analysis of scientific and technical texts. We have also outlined heuristic multi-step procedure for recognition of scientific discourse-compositional structure, with the aid of the dictionary and surface syntactical analysis of sentences.

The recognized discourse markers and discourse-compositional structure is apparently useful in computer systems intended for

- Text abstracting, which may be based on processing of detected markers, e.g. *we illustrate our approach with N* transforms into *the approach is illustrated with N*;
- Document browsing and intra-document information retrieval, which are especially topical for large-size technical documents;
- Computer-aided writing and editing of scientific and technical texts;
- Eliciting of knowledge represented in scientific and technical texts, in particular, extraction of new terms and their definitions introduced into a text by authors.

These applications will supposedly be investigated after implementation, testing, and refinement of the dictionary and the recognition procedure. But more actual task now is creating of computer-aided procedures for enlarging the dictionary, in order to accumulate a comprehensive set of common scientific words and expressions.

Bibliography

1. Bolshakova, E.I. Phraseological Database Extended by Educational Material for Learning Scientific Style. In: ACH/ALLC 2001: The 2001 Joint Int. Conference. Conf. Abstracts, Posters and Demonstrations, New York, 2001, p. 147-149.
2. Bolshakova, E.I., Vasilieva N.E., Morozov S.S. Lexicosyntactic Patterns for Automatic Processing of Scientific and Technical Texts. In: Proc. of 10th National Conference on Artificial Intelligence with International Participation 2006. Moscow, Fizmatlit, Vol 2, 2006, p. 506-514 (in Russian).
3. Dictionary of Word Combinations Frequently Used in English Scientific Literature. Nauka Publ., Moscow, 1968.
4. Dictionary of Verb-Noun Combinations of the Common Scientific Speech. Nauka Publ., Moscow, 1973 (in Russian).
5. Ellouze, M., Hamadou A.B. Relevant Information Extraction Driven with Rhetorical Schemas to Summarize Scientific Papers. In: Advances in Natural Language Processing. Third International Conference, PorTAL 2002. E. Ranchhod and N.J. Mamede (Eds.). Lecture Notes in Computer Science, N 2389, Springer-Verlag, 2002, p. 111-114.
6. Kurohashi, S., Nagao M. Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. In: COLING 94 Proceedings of the 15th Int. Conf. On Computational Linguistics. Vol. II, Kyoto, Japan, 1994, p. 1123-1127.
7. Mann, W.C., Thompson S.A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8 (3), 1988, p. 243-281.

8. Mel'čuk, I. Dependency Syntax: Theory and Practice. SONY Press, NY, 1988.
9. Ono, K., Sumita K, Miike S. Abstract Generation Based on Rhetorical Structure Extraction. In: COLING 94 Proceedings of the 15th Int. Conf. On Computational Linguistics. Vol. II, Kyoto, Japan, 1994, p. 344-348.
10. Ryabtseva, N.K. Mental Performatives in Scientific Discourse. Voprosy yazykoznaniya, V 4, 1992, p. 12-28 (in Russian).
11. Stede, M., Umbach C. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. Proceedings of Int. Conf. On Computational Linguistics COLING-ACM'98, Vol. 2, 1998, p. 1238-1242.
12. Wierzbicka A. Metatext w tekscie. In: O spojnosci tekstu. Wroclaw-Warszawa-Krakow-Gdansk, 1971, p.105-121.

Authors' Information

Elena I. Bolshakova – Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: bolsh@cs.msu.su