[Mays et al., 1992] Mays, E., F.J. Damerau, R.L. Mercer. Context-based spelling correction. Information Processing and Management. 1992, Vol. 27, No. 5, p. 517-522.

[Vossen, 2000] Vossen, P. (ed.). 2000. EuroWordNet General Document. Vers. 3 final. www.hum.uva.nl/~ewn.

[Wagner & Fisher, 1974] Wagner, R.A., M. J. Fisher. The string-to-string correction problem. J. ACM, Vol. 21, No. 1, 1974, p. 168-173.

Author information

Igor A. Bolshakov – CIC-IPN, Research Professor; Center for Computing Research (CIC), National Polytechnic Institute (IPN), Av. Juan Dios Bátiz s/n esq. Av. Miguel Othón Mendizábal, Unidad Profesional "Adolfo Lopez Mateos", Col. Zacatenco, C.P. 07738, D.F., Mexico; e-mail: igor@cic.ipn.mx

Alexander Gelbukh – CIC-IPN, Research Professor and Chung-Ang University, Visiting Professor; Center for Computing Research (CIC), National Polytechnic Institute (IPN), Av. Juan Dios Bátiz s/n esq. Av. Miguel Othon Mendizabal, Unidad Profesional "Adolfo Lopez Mateos", Col. Zacatenco, C.P. 07738, D.F., Mexico; e-mail: gelbukh@cic.ipn.mx, Internet: www.gelbukh.com

TOWARDS COMPUTER-AIDED EDITING OF SCIENTIFIC AND TECHNICAL TEXTS

E. I. Bolshakova

Abstract: The paper discusses facilities of computer systems for editing scientific and technical texts, which partially automate functions of human editor and thus help the writer to improve text quality. Two experimental systems LINAR and CONUT developed in 90s to control the quality of Russian scientific and technical texts are briefly described; and general principles for designing more powerful editing systems are pointed out. Features of an editing system being now under development are outlined, primarily the underlying linguistic knowledge base and procedures controlling the text.

Keywords: scientific and technical texts, automatic editing, linguistic knowledge base.

Introduction

Scientific and technical writing is by no means easy, even for skilled and experienced authors. Usually, the elaboration of a good scientific or technical (sci-tech) text is iterative and time-consuming process, with several persons taking part in it. Besides an author of the document, colleagues, reviewers, and an editor participate in the process, helping the author to improve the text.

Scientific papers and technical documents are essential means of communication between scientists and engineers; therefore the efficacy of the communication depends on the quality of texts. A professional editor of sci-tech texts not only looks for grammar and spelling mistakes, but also accomplishes editing specific for functional style of scientific and technical prose: controlling word usage, revealing drawbacks in logic of reasoning, judging text organization, etc. [10]. The editor explains revealed defects and drawbacks, as well as proposes possible ways of how to overcome them, thereby helping the author to improve the text and to enhance its stylistic uniformity. Almost all sci-tech writers need some aid of professional editor, and without it they lack computer systems automating certain editor functions.

Of course, well-known universal computer text editors and spellers (e.g., MS Word) are widely used for preparing texts. These systems reveal many mistakes, including spelling and simple syntactic mistakes, and their facilities are permanently extended. But the universality of these systems means that they do not account for specificity of the particular text style and genre, in particular, sci-tech prose with its intensive usage of

terms and the other highly standard units. Therefore, additional computer tools are needed for checking scientific and technical texts.

As a whole, sci-tech editing involves a wide spectrum of checks concerning different text levels, so that a deep syntactic, semantic, and logic analysis of the text are required. For this reason, it cannot be fully automated in the nearest future. Nevertheless, several computer systems were built for improving the quality of sci-tech documents, e.g., [1, 5, 7], demonstrating useful features. The systems can do many checks to provide initial editing that earlier has been done by editors or by the writer personally. Most systems are special-purpose editing systems, such as CRESS [5] designed to simplify texts in a narrow problem domain, namely, navy technologies.

Two experimental editing systems named correspondingly LINAR and CONUT were developed in 90s at Moscow State University (MSU) [1, 7]. While the former system was intended to control the quality of technical documents in the narrow subdomain of computer science, the latter was built to support students' practice in writing theses, primarily, to check students' texts with respect to the formal rules of text design, e.g., regularity of abbreviations, bibliography list, references, etc. Having encouraging results in the development of these systems, we aimed at development of a system with advanced facilities, regarding it as a further step towards automation of intellectual functions of sci-tech editor.

The paper shortly describes the systems LINAR and CONUT and summarizes experience of their development, in order to propose designing principles for future editing systems. Then, the research effort going on at the MSU to design more powerful system for checking the quality of sci-tech texts is discussed, including the incorporated linguistic knowledge base and procedures controlling the text. For the sake of clarity, specific features of sci-tech prose are outlined first, along with frequent defects of sci-tech texts.

Sci-Tech Prose: Norms and Defects

Functional style of scientific and technical prose comprises texts of various genres and particular types – research papers and monographs, theses and manuals, reviews and abstracts, technical reports and instructions, patents, etc. This style is admittedly the most distinctive one; its specialty stems from the necessity to express ideas in precise and simultaneously concise manner. The specialty concerns various language levels: lexis and phraseology, syntax, discourse, and composition. Norms and rules, as well as standard language devices and units optimizing sci-tech communication were formed on each level [8, 10].

Sci-tech lexis and phraseology comprises both terms of the particular terminology (e.g., *compiler*, *square root*) and common scientific words and expressions (e.g., *to test the hypothesis, for this reason, summing up*). Whereas specific terms denote concepts, objects, and processes of the particular domain, domain-independent common expressions are used to organize sci-tech text narrative, namely, to express the logic of reasoning, to connect text fragments devoted to different topics, and to structure the text. Among collocations taken from common scientific lexicon, there are clichés, which are relatively stable standard expressions exploited as ready-for-use colloquial formulas (e.g., *to outline directions of further research, the paper reports on*).

As regards the other language levels and corresponding devices and units, we should point out discourse devices. Sci-tech text narrative is organized in accordance with a number of typical discourse-composition frames, including specific ones used in texts of particular genres.

The commonly accepted norms and rules governing sci-tech texts are more of less completely explained in the books devoted to sci-tech writing, e.g. [6, 11]. The writer should follow these rules in order to obtain an accurate, informative, easy readable and understandable text. In particular, on the lexis level, terminological consistency and stability within the given document are required, which implies usage of a standard terminology, unambiguous nominating of concepts within a paper, as well as correct introducing of new terms into the text. We should note, by the way, that newly introduced terms, we call them author's terms, are inevitable in sci-tech prose. Indeed, in scientific papers they denote new concepts and ideas, while in technical documents author's terms designate certain processes and devices. Thereby, author's terms should be properly defined or explained as they are introduced.

To follow all commonly accepted norms and rules is rather difficult task for writers, so sci-tech texts often have various faults. Especially, many students' works (theses and abstracts) are full of defects, since students usually acquire writing skill mainly through their writing practice. Our experience in reading scientific papers in the domains of computer science and computational linguistics, as well as in reading and editing texts written by students shows that observed text faults vary on their nature. We have compiled and classified a list of

frequent defects, which is presented below along with some illustrative examples (given in English and Russian languages) and short explanations of violated rules or requirements.

Inaccuracy in usage of special terms, both author's and generally accepted ones (the latter are usually referred to as dictionary terms): usage of new terms without their definition or explanation; unjustified usage of multiple term variants (i.g., grammar synonyms *swing smoothing method* and *method for smoothing of swings*, Rus. *сцепление* и *сцепка*). It is worth noting that terminological synonymy is not advisable, only a few term variants is acceptable.

Stylistic and grammar mistakes in combining words of common scientific lexicon and in standard clichés: wrong collocations (e.g., *to examine a problem* instead of *to study a problem*), mistakes in syntactic agreement and government, omission of clichés elements (e.g., Rus. *обращает внимание сходство* instead of *обращает на себя внимание сходство*).

Awkward phrases and sentences with multiple complex constituents, such as subordinate clauses, homogeneous parts of the sentence, and nested constructs in parenthesis; as well as several similar syntactic dependencies with the same preposition (e.g., *process of revision of complex fragments of expository texts*).

Syntactic ambiguity entailing semantic ambiguity: ambiguous anaphoric elements (mainly pronouns), ambiguous grammar structure of a word combination or of a sentences in the whole (e.g., Rus. *недостаток машин* – a lack of machines or a defect of machines?)

Syntactic heterogeneity of list items (e.g. *1) to consider material 2) to process data 3) analysis of results*); semantic incompatibility of homogeneous parts of the sentence (e.g., *compilation, interpretation, and translator*)

Drawbacks of discourse-composition structure: weak coherence of the text, lack of logic relations between text fragments (sentences and paragraphs). It is worth noting that such relations are normally expressed by connectors, such as *nevertheless, since, to this end*.

Violations of commonly accepted rules of sci-tech texts design, such as rules of citation, referring, numeration of text units, abbreviation of words and word combinations, etc.

Clearly, many of these defects and drawbacks are specific to sci-tech prose and are not controlled by universal commercial text editors. Meanwhile, special-purpose editing systems, such as LINAR and CONUT described below check for some presented defects.

Two Systems for Sci-Tech Text Editing

LINAR [7] seems the first system developed for editing Russian sci-tech text. It was intended to control the quality of Russian sci-tech documentation, primarily technical reports and texts of technical tasks on the theme "Architecture of multiprocessor systems".

Besides spelling control, LINAR provides a sufficiently wide set of specific checks. It reveals some style defects (such as presence in the phrase of several words with the same root: e.g., *functional, function*), defects of sentence structure (e.g., violations of neutral order of words), particular semantic defects (e.g., inconsistencies like *compiler, interpreter, and processor*), defects in text composition (such as absence of obligatory parts of document or improper order of parts). These facilities are based on relatively deep morphologic and syntactic parsing of words and phrases; elements of semantic analysis are used as well. Controlling procedures exploit several computer dictionaries, among them a large dictionary of word stems and a semantic dictionary (thesaurus) representing relations between terms of the given problem domain.

As compared with universal text editors, LINAR presents, besides diagnostics indicating revealed text defects and variants to correct them (if any), short explanations of the violated rules. To initiate checking process, the user specifies what checks are to be applied and to what text fragments. This feature of LINAR is connected with its module organization: its program kernel consists of procedures, and each of them is intended to control the particular rule or property (aspect) of text.

The system CONUT [1] was implemented in the late 90s for controlling and editing students' texts (theses and abstracts). It checks texts with respect to formal rules of text design, which comprise:

- ✓ consistency of abbreviations (their introduction and usage);
- ✓ correspondence of bibliography references in the text to the bibliography list;
- ✓ presence of obligatory parts of document (e.g., introductory part and table of contents);
- ✓ correctness of table of contents (its correspondence to headings and pages in the text);

- ✓ regularity of numeration of text units and pictures.

Such rules are considered formal because they do not concern the meaning of the text and its units.

CONUT can also estimate some aspects of style of the text, in particular, its simplicity (readability). A heuristic formula was proposed for this purpose, which accounts for various elements indicating sentence complexity (the number of words, punctuation marks, conjunctions, and pronouns in the sentence).

It is considered important that CONUT not only reveals defects in texts and estimates text style, but also explains the essence of formal rules and style estimation methods being applied. CONUT provides a reference guide accumulating information about formal rules of sci-tech text design. The guide is usable in checking process: when any defect is identified, a corresponding page of the guide (wherein the violated rule is explained) is given to the student. The reference guide is flexibly organized: its text material is represented as a hypertext, enabling both free navigation and learning the material in the recommended (predefined) order.

Comparing this system with LINAR, it should be noted that CONUT does not analyze text properties for which natural language parser is needed, most formal rules can be applied for the check without large dictionaries and syntactical-semantic analysis of phrases. Meanwhile, CONUT provides wider range of formal checks oriented just to revising students' texts and has advanced tutoring function.

Two described systems demonstrate some significant features, which ought to be considered while developing more powerful and helpful editing systems.

Designing Principles for Sci-Tech Text Editing Systems

Starting at MSU the development of new experimental editing system and having in mind the list of frequent defects described earlier, we intended to concentrate upon checks that are not fully implemented in LINAR and CONUT or are absent in them. At the same time, we take into account following crucial principles derived from our previous experience.

First, future sci-tech text editing systems will inevitably be based on semi (not fully) automatic editing procedures. One reason is obvious: the reliability of all automatic text processing programs is ranging approximately from 70 to 97%. Thus, the result of automatic checks of the text might be wrong, and proposed variants of how to revise text defects (if any) might be improper. So the author of the text should control all results and should ultimately choose ways of text revising. Hence, revising process implies a dialog between the user and the editing system, the latter indicates problems areas to be corrected and proposes variants of revision (as a human editor usually does) while the user makes appropriate decisions.

Second, it is not convenient to straightway apply all possible checks at user's text, since the spectrum of checks and estimations is sufficiently wide even in LINAR and CONUT, and their accomplishment might lead to a time-consuming process with vast diagnostics. It seems more reasonable when the user sequentially chooses desirable checks, initiates checking, and then analyzes obtained results, making necessary decisions. This feature of user interface determines the principle of system organization – the program kernel performing various text checks and estimations should be implemented as a set of procedures, each one controlling the particular rule or norm (for example, correctness of abbreviations) or estimating the particular aspect of the text. This principle was successfully tested within LINAR and CONUT; it enables to easily increment the editing power.

Third, although tutoring function of editing systems is clearly an auxiliary one, it is of no little significance. Comprehensive explanations of particular text defects, and also the explanatory information about formal and informal requirements to sci-tech documents can facilitate writing. Thus, it makes sense to reinforce editing systems with special dictionaries (such as systematic dictionary of common scientific words and expressions) and a reference guide explicating various norms and rules of sci-tech writing – even if some rules can not be checked so far in the particular system. For example, the guide can present explanations and typical examples of how to properly introduce new terms into scientific and technical texts. For students, such a guide can serve as a tool for systematic learning of sci-tech writing.

Forth and finally, in order to implement checks and estimations specific to sci-tech prose, an editing system should include vast linguistic knowledge base comprising both domain specific and domain independent components reflecting features of the prose.

According to discussed principles, main components of our novel system are the following: module implementing user interface, procedures controlling and estimating text properties, linguistic knowledge base, and reference guide that provides (besides explanation material) browsing data from the knowledge base.

Linguistic Knowledge Base

In order to ensure special-purpose control and estimation of Russian sci-tech texts, several linguistic components are being built into the system:

- Terminological dictionary [3] accumulating units, i.e., single and multi-word terms of commonly accepted terminology of computer science, gathered from several available text dictionaries. The dictionary includes known synonymous variants of terms, in particular, acronyms and other abbreviated forms (e.g. *central processing unit, CPU*). Most terms are nouns and noun combinations, but verbs are included as well. The dictionary represents relations between terminological units, primarily, class-subclass and part-whole relations, thus the dictionary can be regarded as thesaurus.
- List of definition templates [2] describing typical single-sentence definitions of new terms. Such definitions were compiled through manual scanning of sci-tech texts, they contain standard lexical units, such as nouns *term*, *name*, etc., verbs *call, refer, define*, etc. Definition template specifies both immanent lexical components and empty slots (places), their syntactic and semantic properties. An example of definition template is <*Ph*> *we will be call* <*N*>, where *N* denotes an author's term, and *Ph* is a noun phrase explaining its meaning.
- Dictionary of words and word expressions of common scientific lexicon [4]. It comprises both autosemantic and auxiliary words, noun and verb-noun combinations, adverb and participle expressions, compound prepositions and conjunctions. The dictionary unit represents adequate information: syntactic properties of word combinations (interrupted / uninterrupted, stable / free, semantic and syntactic valences, etc.), and semantic class and group of the unit within the proposed semantic classification. The classification comprises 5 main classes: text structuring and composing (e.g., *in addition, next*); expressing logical relations (e.g., *provided that, hence*); indicating sources of information (e.g., *in their opinion*), author's estimates (e.g., *essentially, it is quite likely*); structuring scientific knowledge via common scientific variables (generic nouns), such as *analysis*, *result.* For the latter class, dictionary units describe also syntactic combinability of the noun (e.g., *significant result, to derive result, to question result*).
- Dictionary of standard clichés (stable colloquial expressions) comprising both phrasal formulas (e.g., *the paper describes main features, argument can be made against*) and predicative constructs (e.g., *to outline directions of further research, to take as starting point for*). Some clichés are common for sci-tech prose, the others are specific for particular genres. Clichés are described by templates similar to definition templates: empty slots are indicated, and their syntactic and semantic properties are specified.
- Morphological dictionary of word stems, which covers all words encountered in the other dictionaries (separately or within any multi-word combination). The dictionary unit represents adequate morpho-syntactic information, e.g., part of speech and flexional class (if any), as well as pointers to units of the other dictionaries describing available combinations with the given word (stem). Thus, morphological dictionary connects units of different dictionaries, facilitating their recognition in texts.
- Inventory of prototype discourse frames specifying discourse-composition structures of sci-tech texts. Some frames are domain-specific, e.g., a frame specifying composition of texts that describe particular technical devices. Each slot of prototype frame corresponds to typical subtopic (e.g., functionality of device) and contains pointers to dictionary clichés and common scientific expressions that signal the subtopic in the text.

Text Analysis on Different Levels

We outline a few methods and procedures proposed to implement specific checks of sci-tech texts, which concern different text levels and exploit surface syntactical analysis.

Analyzing Terms and Common Scientific Expressions

For checking regularity of term usage and usage of common scientific expressions, terms and expressions should be recognized in texts. Identification of author's terms presents the major difficulty [2]. Indeed, they are free and often unstable multi-word nominal combination (e.g., *coefficient adjustment learning*) matching several possible syntactic patterns. Moreover, author's terms might be used in texts without any definition or explication, and in order to properly recognize them, local syntactic analysis should be complemented with elements of lexical semantics and term occurrence statistics.

An automatic recognition procedure is proposed [2] based on surface syntactic analysis and dictionary information. The procedure makes use of particular syntactic patterns of terms (e.g., coordinated combination of adjective and noun) and takes into account possible syntactical-semantic variants of new terms (such as

candidate elimination algorithm and *algorithm for elimination of candidates*). The procedure exploits morphological analyzer converting words to their normalized forms and computes frequencies of term occurrences.

According to this procedure, occurrences of dictionary terms and collocations of common sci-tech lexicon are extracted first. Then, certain author's terms are identified by looking for sentences that match dictionary definition templates and by extracting lexical units from the proper places of the encountered sentences. And finally, the procedure attempts to recognize undefined author's terms: it detects word combinations of the given syntactic patterns, identifies among detected combinations different variants of the same term, and gathers them into groups of related term variants along with computed frequencies.

Units recognized at the last step of the procedure are regarded as term candidates, i.e., as potential new terms. Compiled list of term candidates is to be presented to the author of the text for validating and further revising, for example, selecting the most appropriate term in each group of related variants.

Identifying Syntax Ambiguity and Heterogeneity

We consider simple kinds of syntactic ambiguity, mainly ambiguous pronouns and syntactically ambiguous noun combinations. To check pronouns, local analysis of previous context is applied. For noun combinations we propose rather simple procedure of local syntactical analysis that checks dependencies of words within the noun combination.

To check syntactic homogeneity of items in lists, as well as homogeneous parts of the sentence, an automatic procedure is used, which differentiates noun and verb phrases from phrasal constructs and additionally distinguishes several types of noun and verb phrases.

Analyzing Discourse Structures

For recognition discourse structures of texts we propose a heuristic multi-step procedure, which exploits data from the dictionary of standard clichés and the dictionary of common scientific lexicon.

The recognition procedure first searches in the text for all sentences and composing clauses that match dictionary clichés. Occurrences of common scientific expressions signaling discourse relations are looking for as well. For this purpose, local context techniques similar to those described in [9] are used. When an instance of dictionary cliché or common scientific expression is recognized in the text, the procedure extracts lexical units from proper places of the instance and makes an attempt to fill slots of discourse frames related with this cliché or this expression. In general case, after recognition in the text of all cliché instances and common scientific expressions, slots of several discourse frames might be filled, therefore the frame with the maximum number of filled slots is selected as appropriate one. Information associated with this frame is used to estimate drawbacks of the recognized text structure, in particular, improper order of text parts devoted to particular subtopics.

Estimating Composition

Among parameters indicating quality of text composition, we consider proportionality of units of the same level, for example, proportionality of chapters or proportionality of items in itemized list. The proportionality means that units have comparable sizes, and the size can be computed as the number of words in the unit (another option is the number of units of the level underneath). It seems reasonable to estimate the proportionality of sentences within each paragraph, items within each itemized list, paragraphs within each chapter or text section, and chapters within the whole text.

To evaluate the proportionality, dispersion D and mean square deviation σ are calculated:

$$\sigma = \sqrt{D}, \qquad D = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n^2}$$

where x_i is the number of words in i-th component of the text, and n is the number of components.

Function $\exp(-0.1*\sigma)$ is proposed to obtain a measure within the interval $(0,1)$ and with maximum values corresponding to texts with better composition.

So the formula for a proportionality value is $E_C = \exp(-0.1*\sigma)$. The formula is used systematically to estimate the proportionality of units on different levels: sentences, itemized lists, paragraphs, sections, chapters, and also headings of sections, with x_i denoting the number of words in units being estimated, and n

denoting the number of them. The estimation value for the whole text can be computed as the arithmetic mean of proportional values already computed for all text units.

Conclusion

While discussing the problem of computer-aided editing of scientific and technical texts we outlined peculiarities of special-purpose systems built for editing Russian text and pointed out the principles applicable for designing editing system with advanced facilities. We also outlined some features of a novel editing system being under development, primarily its linguistic knowledge base and procedures controlling the text. We hope that the system will be useful for a wide community of sci-tech writers, in particular for inexperienced authors, such as postgraduate students.

By now, the linguistic knowledge base of the system is partially implemented, with the terminology dictionary covering terms in certain narrow subfield of computer science and the dictionaries of common scientific lexicon and standard clichés containing more than 700 units. The first versions of the controlling procedures are tested.

Bibliography

1. Bolshakova, E. *Computer Assistance in Writing Technical and Scientific Texts*. Proceedings of 2nd International Symposium "Las Humanidades en la Educación Técnica ante el Siglo XXI", México, 27-29 September, 2000, p. 59-63.
2. Bolshakova, E. *Recognition of Author's Scientific and Technical Terms*. In: Computational Linguistics and Intelligent Text Processing. Second International Conference CICLing 2001. A. Gelbukh (Ed.). Lecture Notes in Computer Science, N 2004, Springer-Verlag, 2001, p. 281-290.
3. Bolshakova, E., Vasilieva N., Yudin D. *Extraction of Dictionary Terminological Word Combinations in Scientific and Technical Texts*. Proceedings of International Workshop on Computational Linguistics and its Applications Dialogue'2001. Russia, 2001, V. 2, p. 48-51 (in Russian).
4. Bolshakova, E. *Desingning Principles for a Computer Dictionary of Commom Scientific Lexicon.* Proceedings of International Workshop on Computational Linguistics and Intellectual Technologies Dialogue'2002. Russia, 2002, V. 1, p.19-23 (in Russian).
5. Glenda, M. *Readability Formulas: Useful or Useless?* IEEE Transactions on Professional Communications. Vol. PC-30, No 1, March, 1987, p.12-15.
6. Emerson, F.B. Technical Writing. Houghton Mufflin, 1987.
7. Malkovsky, M.G., Bolshakova E.I. Intellectual System for Control of Text Quality. In: Intellectual Systems, V. 2, No. 1-4, Moscow, 1997, p. 149 –155 (in Russian).
8. Mitrofanova, O. *Language of Scientific and Technical Literature*. Moscow University Press, 1973 (in Russian).
9. Paice, C., Jones P. *The Identification of Important Concepts in Highly Structured Technical Papers*. Proc. of 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, 1993, p.69-78.
10. Senkevich, M. *The Style of Scientific Speech and Literary Editing of Scientific Works*. Moscow, Vysshaia Shkola, 1976 (in Russian).
11. Zobel, J. Writing for Computer Science. Springer, 1997.

Author information

Elena I. Bolshakova is Docent of Moscow State Lomonossov' University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Adress: Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; E-mail: eibolsh@aha.ru ; bolsh@cs.msu.su