The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy

ISSN: 2357-1330

https://dx.doi.org/10.15405/epsbs.2018.12.02.66

18th PCSF 2018
Professional Culture of the Specialist of the Future

LANGUAGE WORKER IN THE FRAMEWORK OF INFORMATION 4.0

Valeria Chernyavskaya (a)*, Larisa Beliaeva (b), Sergey Nefedov (c)
*Corresponding author

(a) Peter the Great St. Petersburg Polytechnic University (SPbPU), Polytechnicheskaya 29, Saint Petersburg, 195251 Russia, tcherniavskaia@rambler.ru
(b) Herzen State Pedagogical University of Russia, 48 Moika emb., St. Petersburg, 182100, Russia, laurabel@gmail.com
c) St. Petersburg State University, 7 - 9 Universitetskaya nab., St. Petersburg, 199034, Russia, s.nefedov@spbu.ru

Abstract

The paper focuses the profession of language workers - those prepared to solve text processing tasks in the new technology space. This is discussed in the concept of Information 4.0. This concept reveals that information can be presented as a cloud of “information molecules”. Information 4.0 examines the form, production, interaction and curation of information components. This means changing priorities in using linguistic techniques and tools upon scientific and technical texts. Information and knowledge are represented and transferred as a text. The paper reflects that Information 4.0 demands new approaches in how we work on content. New research directions and tasks to be solved are seen in advancing from the procedures to formalize text semantics to the procedures providing machine-readable automatic information structuring and text production. Thus, the methodological focus is not on the ready texts but on linguistic tools and operations explaining how to produce texts according to given models and content.

Keywords: Content analysis, information 4.0, language worker, text structure.

1. Introduction

In the current context there have been considerable changes in the circulation of scientific and technical information, the tasks and the format of expert review of the research outcomes. Previously a major focus was brought on verbal formulation and rhetoric of promoting research outcomes, text types/genre forms related to abstracting, summarizing and peer review of scientific knowledge. These included information profiling and advancing (Akopova & Chernyavskaya, 2014; Chernyavskaya, 2016; Chernyavskaya, 2017; Kabanova & Kogan, 2017; Tognini-Bonelli, 2005; Politaeva, Bazarnova, & Boguk, 2017; Porter, Bazarnova, & Boguk, 2002; van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003). Digitalization of publication culture, open access trends and most significantly, information computerized creation, search and usage have changed the traditional approaches. A present-day researcher has to be focused mainly on the technology of information presentation rather than on the text itself as a "generator of meanings" in its linguistic, pragmatic and hermeneutic sense. It is crucially important that this technology should be considered as secondary with regard to the previously adopted understanding of text as a structure.

A modern approach to information exchange in industry is characterized by newly introduced terms of Industry 4.0 and Information 4.0 (Gollner, 2016). The key concept is that the major task to be solved is to maximize the flexibility. The level of implementation of elaborated principles of this new Industry 4.0 also depends on production, exchange and application of information about the project developed, industry sector, maintenance of a specific device and/or technical equipment and financial security. In the process of scientific and technical communication, this kind of information is generated in the form of documents at all stages of the project development. The quality of the documents produced in the source language and those translated into all the languages where the product is to be distributed determines the opportunity to apply high-level automation in the process of their interpretation and publication.

2. Problem Statement

A new research objective to be considered is to develop mechanisms of extracting information from texts in the age of its automated processing and transfer. To perform these tasks, it is important to train specialists of a new kind who would be concerned with text processing. Specialists who deal with new forms of information presentation are referred to as *language workers*. This notion is used as a common nomination for terminologists, translators and all those who produce technical documents (*technical authors, technical writers*), and transfer technical information (*technical communicators*). Modern language workers are to know typology of specific and technical texts both in their native and foreign languages; they should be skilled in producing all types of specific texts both in their native and foreign languages; they should also be able to translate texts with regard to requirement differences for specific texts in certain cultures. Information 4.0 is seen as a cloud of information molecules rather than a set of structurally complete documents; it is dynamic, i.e. regularly updated; offered rather than delivered; ubiquitous, interactive, easily accessible and findable; uncensored, i.e. produced by contexts, profiled automatically, (Gallon, 2016). Thus, specialists should proceed from formal analysis of text semantics to generating and structuring the text in accordance with new contexts of the text usage and its requirements.

http://dx.doi.org/10.15405/epsbs.2018.12.02.66
Corresponding Author: Valeria Chernyavskaya
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

3. Research Questions

A major concern for the specialists within the concept of Information 4.0 is to produce a text according to set structural models and content rather than to analyze ready-to-use text structures. Therefore, information provided in a natural language -usually in a controlled language (Muegge, 2009)- in the form of scientific and/or technical documentation must be prepared to be used in various situations and be quickly adapted to different scenarios of production, maintenance and financial security. Information must be presented in the way that it can be exchanged at any stages of project implementation. This gives a different view on the text structure and the role of separate components in structuring of the text information. Based on them one can extract important information.

4. Purpose of the Study

The aim of the present paper is to justify the need for training new specialists which results from emerging high-powered technologies of extracting information from texts in scientific and technical communication. We stress that information 4.0 injects new life into the profession of technical communication and language workers - those, who are prepared to solve text processing tasks in this new technology space. Many industry sectors benefit from the idea of supporting the maintenance and operations using smart information.

5. Research Methods

Methodological analysis is based on theoretical conclusions of the sphere of applied linguistics which deals with knowledge engineering, i.e. methods and means of extracting, presenting, structuring and using of knowledge.

6. Findings

Since the pragmatic approach to the text structure in terms of text grammar and semantics claims, the text should contain structural and meaning-oriented emphasis of certain text components which bear the most significant, in the author’s viewpoint, information. This approach determines the standard pattern of production and perception of a scientific and/or technical text. It is a generally accepted IMRAD formula (introduction, method, results and discussion). Potential of automated processing sets a task of promoting a research outcome and its algorithmic marking differently.

Within the Information 4.0 framework a major concept is *structured content authoring*. It means structuring of the content into sections referred to as *topics* which are further automatically assembled in *maps*. These components allow produce a final draft of content to be used in a certain function and in a specific type of document. Topics should completely correspond with the text subjects. Then ‘information molecules’ can be marked algorithmically and form the ground for texts of a different kind. The latter can be used automatically.

Research of the text structures in order to extract information has become an important objective since the information flow expanded to the extent that its prompt and high-quality processing appeared to be really sophisticated for those who need this information. Among them are specialists and analysts who deal with data and knowledge extracting and processing. Texts began to be analyzed based on their content in the 70s of the 20th century when information retrieval (IR) became automatic. Information retrieval was used to select texts from previously produced and constantly supplemented text collections according to some topic. The topic has to be chosen by the user's query or extracted from a set number of topics. Texts could also be selected based on specific factual information, etc. IR was supposed to select automatically the texts to correspond with a certain query/topic.

A new method is based on *productivist approach.* It implies that "the granularity of the topics is determined by production issues" (Lacroix, 2016, p.103), by the objectives of scientific and technical documents. It is also potentially decorrelated from the content itself, i.e. from those topics that are actually discussed in the text. Depending on working conditions with the information flow or with the existent text corpus, preliminary text indexing can be conducted in two ways. One way is to preset a list of topics relevant for the user – analyst, researcher, librarian, etc. (manually or automatically). In this case indexing is done according to this list. In the second case the indexing is to be carried out automatically. This means that the main content of the text is described as its search profile that forms the ground for its further processing.

According to the approaches mentioned above, text indexing can be based:

- on semantic analysis of texts using special dictionaries; this method allows to establish a relationship pattern (frame) between components of a sentence and/or a text, and thus determine the topic of the whole text, i.e. its search profile;
- on lexical, statistical and/or dictionary techniques; these allow to refer a text to a certain topic relying on whether it is relevant to the preset dictionary models and texts.

Thus, when indexing, firstly, it is possible to implement a statistical approach. In this case indexing is done by selecting the key terms from a text and referring them both to diagnostic features and frequency of their occurrence in the recognized text and in the reference collection (Krippendorf, 2013; Rivers, 2011). Here a special dictionary of stop-words plays a significant role. This dictionary contains a list of words which do not bear information on a certain topic regardless of their frequency in the texts of specific subject domains (SD). Secondly, a dictionary approach is to be considered. It enables to produce a hierarchical system of models, thesauri and frames. This system represents a set of diagnostic mechanisms.

As a rule, indexing is done in terms of a certain set of topics and corresponding dictionaries. It should be noted that solving the problem of attributions of any type requires a preliminary description of the subject domain structure where indexing takes place. This description can be done in the form of the system of local dictionaries that contain such lexical units which have been selected based on preliminary research of a reference corpus and are more relevant when referring some document to a certain topic.

Therefore, to generate such a system, the first thing to be done is to investigate a set of topics and texts corresponding to these topics. A set of reference texts in each topic is selected based on consultations with experts, and the size of this set is determined by a standard approach accepted in linguistic statistics. Based on sets of reference texts frequency dictionaries are to be received. To reduce the amount of texts to be analyzed, all structural words (stop-words) and the words which do not bear information on certain topics must be rejected from these dictionaries.

When this task is solved, specialists have quite a large amount of statistical data in order to observe general topics and the direction of information flows which reflect the spread of interest in a particular subject domain. In addition, this system allows select the most important documents for further detailed semantic analysis. Indexing subsystem produced in this way can operate efficiently if quite a full and rigidly structured description of a certain subject domain has been made. These methods make it possible to determine which key terms are characteristic of each text and what topics they reflect. Words and documents in several topics are clustered as both the key word and the text document can correspond with several topics with different probabilities.

One of the major procedures is preliminary indexing of both queries and texts. This means that topics are formulated in advance. Topics correspond with the system of arranging the topic molecules of the text when it is produced automatically. It makes the queries and retrieval corresponding the sets of key terms and relations between them that are the result of preliminary collaboration between analysts and linguists. Topics are selected based on an assumption that specific (key) units, in modern terminology – key terms which meanings are fully described, can be pointed out in the text structure. Today choosing the key words is regarded as an important objective of formulating the result obtained, (Krippendorff, 2013). Whether a term or a term combination belongs to a set of key units chosen to characterize the text content, is a critical factor in further use of the text by specialists and in spreading the results. All modern metrics used to point out key units are based on this idea. Only certain approaches and procedures vary as they employ more specialized methods of statistical analysis and big data.

Solving standard tasks both in content analysis and information retrieval traditionally relies on small sets of particular key terms (key words) used to mine relevant information on the text collection. The information can be pre-arranged as a text bank or be a permanent text flow.

Automatic analysis and further information retrieval on the uses' queries is based on the procedure of attributing formally arranged information to the text about its content (indexing). This information is organized in the way that it can be used in automatic solution of different tasks in information retrival and mining. This means that such procedure has to be multi-level and oriented to the text markup with respect to its syntactic (acoustic and graphic), semantic (combinatory and lexemic) and pragmatic (contextual) information. This procedure is based on the method of content analysis.

Both in terms of automated documentation generation and text topic determination, indexing is done with regard to a certain set of topics and corresponding dictionaries. It should be noted that solving the problem of attributions of any kinds requires preliminary description of the structure of the subject domain where indexing takes place. This description could be built in the form of a system of location dictionaries. Such dictionaries contain those lexical units which have been selected based on preliminary research of reference corpus and are more relevant when referring some document to a certain topic.

Therefore, to generate a similar system, the first thing to be done is to investigate a set of topics and texts corresponding to these topics. A set of reference texts in each topic is selected based on consultations with experts, and the size of this set is determined by a standard approach accepted in linguistic statistics. Based on sets of reference texts alphabetical frequency dictionaries are arranged. To reduce the amount of texts to be analyzed, all structural words (stop words) and words which do not bear information on certain topics must be eliminated from the dictionaries. When producing a text, preliminary analysis of a similar set of reference texts should be done. This allows distinguishing a set of topics – documentation 'molecules'. Thus, text indexing and semantic analysis of a text are not only required for solving modern tasks in the concept of Information 4.0 but are also compulsory procedures.

When solving this task, specialists have quite a large amount of statistical data in order to observe the general condition of topics and the direction of information flows which reflect the spread of interest in a particular subject domain. Moreover, this system allows selection of the most important documents for further detailed semantic analysis. Based on probability topic modeling it is possible to determine which lexical units, namely key terms, are characteristic of each text and which topics they reflect. Topic models enable to cluster words and documents in several topics-clusters as both the key word and the text document can correspond with several topics with various probabilities.

Methods applied in analyzing texts and generating documentation are implemented by preliminary manual selecting of a bank of documents. This text bank can be regarded as a reference collection of a certain system. Search in such a reference collection is conducted in three stages: 1. Extracting of key term set from a reference collection. Terms in this collection are estimated as relevant/irrelevant to the texts. It allows to reveal whether they are relevant or not in analysis and to assess whether they are necessary to be used in synthesis. 2. Mining of data on the key terms co-occurrence in the reference collection as well as in the national text corpus representative for the given language. 3. Quantitative estimation of relevance of each document based on the dictionary and co-occurrence data used to generate an estimated list of documents, for more details see: (Beliaeva, 2003; He, Chang, Lim, & Banerjee, 2009; Koval et al, 2000).

7. Conclusion

A promising investigative path is to solve the problem of training new specialists in natural language processing. Linguists working with Information 4.0 are required to have new competencies that can be described as follows:

- ability to collect, analyze and select relevant information to develop an information product,
- ability to choose the most appropriate product manufacturing strategies in order to design corresponding information products for various purposes and consumers,
- ability to ensure that information is extractable and available, that it is a coherent model, and comport with products and contexts,
- ability to select appropriate hardware and software to be used in scientific and technical communication,
- ability to design and assess e-learning courses,
- knowledge of the process of information products publishing and stages,
- sufficient level of understanding of subject domains relevant for specialists in technical information distribution (information science, mechanical engineering, physics, etc.) to be able to cooperate with experts in these subject areas,
- knowledge of basic principles and methods of the science of terminology,
- ability to build up linguistic resources, lexicographical databases and text corpora to solve professional tasks.

The two latter competencies concern working with terminology since in a new information environment exactly a technical writer, a product manager and a terminologist reveal new terms. They appear as a result of developing, certifying and documenting new products, thus all types of documentation are to be considered: catalogues, user manuals and reports, user interfaces, error reports and system messages, etc.

References

Akopova, M., & Chernyavskaya, V. (2014). Evaluation of Academic Science: Perspectives and Challenges. *Zeitschrift fur Evaluation*, *2,* 348-357

Beliaeva, L. (2003). Machine Translation Versus Dictionary and Text Structure. *Journal of Quantitative Linguistics*, *10 (2)*, 193-211

Chernyavskaya, V. (2017). Towards methodological application of Discourse Analysis in Corpus-driven Linguistics. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. Tomsk State University Journal of Philology*, *50*, 135–148.doi: 10.17223/19986645/50/9

Chernyavskaya, V. (2016). Cultural Diversity in Knowledge Dissemination: Linguo-Cultural Approach. *SGEM 3rd. International Multidisciplinary Scientific Conference on Social Sciences and Arts*, *2*, 443-450

Gallon, R. (2016). Information 4.0, the Next Steps. *Towards a European Competence Framework. Tekom-Jahrestagungundtcworldconference in Stuttgart. Zusammenfassungen der Referate*, Stuttgart: tcworld GmbH Verantwortlich

Gollner, J. (2016). Information 4.0 for Industry 4.0. *Towards a European Competence Framework. Tekom-Jahrestagungundtcworldconference in Stuttgart. Zusammenfassungen der Referate*. Stuttgart: tcworld GmbH Verantwortlich

He, Q., Chang, K., Lim, E., & Banerjee, A. (2009). Keep It Simple with Time: A Re-examination of Probabilistic Topic Detection Models. *IEEE Transactions on pattern analysis and machine intelligence,* *32*(10), 1795-1808 Retrieved from http://wwwusers.cs.umn.edu/~banerjee/papers/09/pami-tdt.pdf.

Kabanova N., & Kogan M. (2017). Needs Analysis as a Cornerstone in Formation of ICT Competence in Language Teachers Through Specially Tailored In-service Training Course. *Learning and Collaboration Technologies. Novel Learning Ecosystems. LCT 2017. Lecture Notes in Computer Science*, *10295*, 110-123. doi:10.1007/978-3-319-58509-3_11

Koval, S., Beliaeva, L., Kogan, L., Mikhailov, A., Nikolaev, V., Piotrowski, R., & Tovmach, Y. (2000). Morphological Representation in Pc-Based Text Processing Systems. *Literary and Linguistic Computing*, *15, 2,* 131-155

Krippendorff, K. (2013). *Content Analysis: An Introduction to its Methodology*. Los Angeles; London: Sage

Lacroix, F. (2016). *Writing for the 21st Century. Towards a European Competence Framework. Tekom-Jahrestagungundtcworldconference in Stuttgart. Zusammenfassungen der Referate*, Stuttgart: tcworld GmbH Verantwortlich.

Muegge, U. (2009). *Controlled language - does my company need it?* Retrieved from www.tekom.de/artikel/artikel_2756 html. 2009

Politaeva, N., Bazarnova, J., & Boguk, S. (2017). Innovative technologies secondary use of processed active source. *Proceedings of the 2017 International Conference "Quality Management, Transport and Information Security, Information Technologies", IT and QM and IS 2017, 8085865*, 471- 476

Porter, A.L., Kongthon, A., & Lu, J. C. (2002). Research profiling: improving the literature review. *Scientometrics, 53, 3,* 351-370

Rivers, N.A. (2011). Future Convergences: Technical Communication Research as Cognitive Science. *Technical Communication Quarterly*, *20 (4),* 412- 442. doi:10.1080/10572252.2011.591650

Tognini-Bonelli, E., & del Lungo Caniciotti*, G*. (2005). *Strategies in Academic Discourse*. Amsterdam: John Benjamins Publishing. doi: 10.1075/scl.19

van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & van Raan, A. F. J. (2003). The Holy Grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, *57(32),* 257-280