Digital Humanities Tools in Information 4.0 Space: Towards language worker competences

Valeria Chernyavskaya¹, Larisa Beliaeva², and Olga Kamshilova²

¹ Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation
² Herzen State Pedagogical University of Russia, St. Petersburg, Russian Federation
* tcherniavskaia@rambler.ru

Abstract. Induced by the rise of technology level due to automation processes growth in industries, the coming forth industrial revolution (Industry 4.0) will inevitably bring along dramatic changes in job market. Among new jobs a language worker is definitely a competitive one, yet little is done today both in education and industry to provide a proper training for this “job of tomorrow”. The paper focuses on supportive competences and practice conditions of a trained language worker prepared to perform information and text processing in Industry/Information 4.0 space. This paper considers IT and Web resources potential to be used by would-be language workers in their practice and studies. Methodological analysis is based on theoretical and practical conclusions about applied linguistics development which deals with knowledge engineering, i.e. methods and means of extracting, presenting, structuring and using of knowledge. In the framework discussed the focus is on relevant linguistic tools and internet resources and a modeled automated workstation to ensure the working process.

Keywords: digital humanities, Information 4.0, technical communication, language worker

1. Introduction

Since 2011 German prognosis that the combination of cyber-physical systems, the Internet of Things, and the Internet of Systems, now known as Industry 4.0, will revolutionarily change the world, there has been much discussion of the challenges and risks that this new industrial setting will inevitably bring along. Since much of the information component of Industry 4.0 will be machine-to-machine communication with no human involvement one of the most debated topics today is the challenge of technology-led destruction and creation of jobs [1]. Among new occupations technical communication specialists will definitely be in demand, yet little is done today both in education and industry to provide their proper training.

Paradoxically, Russia, where, due to globalization and international cooperation, home economy has already accepted world experience in technical communication, is seriously behind in educating and training specialists who can professionally perform it. While America and Europe have accredited university programs for teaching technical communication and functioning technical writing associations and centers, see, for instance, [2,3], this professional activity in Russia is obviously lacking academic support. Moreover, training in technical communication is becoming a commercial venture for numerous private companies and individuals that copy international standards and offer educational services on their sites.

To meet the demands of modern technology and science defined by industrial processes automatization (Industry 4.0) and appropriate presentation of information (Information 4.0) we need a new generation professional, presumably with a basic linguistic education. A linguist up-taking functions of a technical communicator is what really happens in Europe, see, for instance, the published opinion exchange in tekom Intelligent Information blog (https://intelligent-information.blog/en/iirds-how-a-novice-sees-the-information-delivery-standard/).
A technical communicator today is, in a broad sense, a language worker – terminologist, translator, lexicographer, technical writer, grammarian, etc., prepared to deal with new forms of information and the technologies and trends that will make it work. The core characteristics of content under Information 4.0 are:

- real-time
- ubiquity
- online availability
- context-awareness,

which demand not only new professional competences but an appropriate working space and skills to manage it in line with one’s professional needs, for more details see [4-8].

2. Research Question

The goal of this paper is to focus on a novice language-worker supportive competences that reflect, to a large extent, alongside with linguistic aspect of their text-processing and text-productive performance, a “technical”, e.g. profession-oriented aspect of their job, namely being able a) to manage large numbers of documents (different types, versions and formats), b) to collect, analyze and select appropriate information to design an information product, c) to manage and to overview large amounts of information, d) to know about hardware and software that is used in technical communication, e) to manage communication processes across languages and cultures; f) to know the laws, standards and regulations that are relevant for technical communication in order to provide correct information, g) to understand the basic principles and methods of terminology science. We will highlight relevant linguistic tools and internet resources and suggest a framework to furnish them (as a specialized automated workstation) to ensure the working process.

3. Methods

Methodological analysis is based on theoretical and practical conclusions made on applied linguistics developments which deal with knowledge engineering, i.e. methods and means of extracting, presenting, structuring and using of knowledge. The purpose of the paper is to discuss what has been researched so far about Industry 4.0, and its impact on the changes in the human environment. To achieve this objective, a review of books, journal papers, conference papers was performed. Thus, we recommend modern tools and data that have proved efficient in natural language and information processing, such as systems of machine translation and translation memory, terminology management, terminological databases, etc.

The structure of an automated workstation (AWS) is proposed in accord with early developments of Leningrad-St. Petersburg Engineering Linguistic School and Russian State Standard (Russian GOST 34.003-90).

4. Results and Discussion

4.1. Tools and resources for a language worker

A language worker’s tools are not only constantly developing information technologies (IT), but IT-based language resources, designed to store both relevant information (lexicographic, expert, corpora) and systems for machine translation and information processing. Language resources comprise text resources proper, arranged as powerful databases, and numerous language knowledge sources, such as thesauri, ontologies, web semantic frames and models.

Within this paper we lay special stress on modern lexicographic resources as linguistic tools for terminological task support. We presume that a terminologist shall react quickly to satisfy the requirements of content under Information 4.0 - real-time, ubiquity, online availability, context-awareness – to perform high-quality information processing and to mark out not previously registered or new terminological unit both to keep to production string and to enter the result into proper lexicographic resources.
Similarly, the work of a technical translator in a certain production string which includes properly organized language resources: a chosen MT system, a complex of automated dictionaries, lexicographic resources: under such conditions the speed of performance and high translation quality acquire special importance as the translation, which is delayed or incorrect, can bear critical consequences (in such high risk domains as seismic protection, nuclear engineering, medical science and others).

Terminology management systems were first created as early as in the 70s of the previous century: big companies and government institutions began to develop language machine funds, since new terminology permanently appeared simultaneously with economical and technical growth. Among them was the TEAM data bank, developed by Siemens company for European languages, which included about 700000 lexical units from various domains respectively grouped (natural sciences, business, engineering, etc.) [12]. Later in the 80s the idea of national language machine funds spread broadly (Russian language fund included). The purpose of such funds was developing universal language databases.

Terminological databases (TDB) are computer-aided storages for terms. Here terms are supplied with additional information of both linguistic (combinatory power, frequency, semantic field, etc.), and extralinguistic (domain, definition, regularity, etc.) type to provide knowledge on separate terms or term collocations (descriptions, examples, translations). These banks were used for compiling special text glossaries and new specialized translation dictionaries [10].

The multilingual lexicographic resources of today are web-based and can be classified as governmental, for instance, supported by European Commission, and initiative, developed by corporations or research groups. Now consider the most popular governmental bases of terminological data.

Among first developed governmental TDB is Eurodicautom databank [11], covering all the languages of European Union. European Union documents are issued in 24 languages and allow for 552 language combinations. The dictionary database was updated every week due to the results of Brussels and Luxemburg translation units, new terms and their translation proposed were systematized by Eurodicautom group, and, partly, imported from private companies and experts in separate knowledge domains.

In 2008 the European Parliament launched the IATE (InterActive Terminology for Europe) project, which objectives were to co-ordinate, support and provide tools in terminology studies. The IATE format is a terminological relational database. In addition, the European Parliament organized a TermCoord department for terminology co-ordination, which implements access to EC terminology through public site and free tools, as well as through interinstitutional terminological portal EurTerm.

EuroTermBank is now one of the most powerful free Web governmental terminological bases. This linguistic resource unites 133 local resources, developed in various EC translation units, 2 650 976 terms (the number is permanently increasing), 710 705 dictionary entries, 221 512 definition in 33 languages. EuroTermBank suggests various options for the choice of a source/target language, data domain, information format presentation, etc. The user is also provided with information on translation variants in different data domains and on terminological collocations.

EuroTermBank can be taken for a proved model for multilingual Web resource, actual both for Russian Federation languages and for Customs Union languages, since it can provide proper terminological and lexicographic support for document translation and transfer in various co-operation and knowledge spheres.

It permits language workers to search terms in various sources, to extract terms automatically, to browse term translation variants in different data domains, to search for terms in several target languages, to specify translations and to share information with other users. Access to the resource is possible directly with Microsoft Word processor.

A comparatively recent development of language resources and tools is connected with corpus technologies, namely compiling and use of real text corpora, that can be considered as databases for solving not only research, but also practical lexicographic tasks. As a rule, written text corpora used
for such purposes include texts collections, compiled according to a domain, author, function, etc. Corpus technologies, such as obtaining concordances, dictionaries of words and collocations in case of monolingual corpus, as well as creating multilingual lexicons and concordances in case of parallel or comparable text corpora are made possible due to linguistic (morphological, syntactical, semantic and other) and extralinguistic annotation (tagging) of corpus units. This tagging can be the basis for term extraction. Thus, extraction of multicomponent terms can be done on the basis of automatic syntactic analysis on a functional segments level – that of noun phrases.

Corpora known as parallel may be of two principal kinds: parallel and pseudoparallel or comparable. Parallel corpora proper compile original texts and their translation(s). Comparable corpora are built with texts in different languages semantically and functionally equal (conference proceedings, declarations, instructions and other document types for instance).

With parallel texts corpora the main task is sentence-by-sentence alignment, which relies on formal indicators of boundaries and sentence parts, conformity of volumetric and pragmatic text structures. With all these technical and linguistic complexities this process is quite realizable.

In case of comparable text corpora only terminological alignment is possible, which relies on eliciting one-word terminological units, characteristic for both corpora, and their comparison as translation equivalent candidates, as well as search of set expressions with these one-word terms. Further comparative analysis is aimed at verifying received term pairs, which requires knowledge from automated translation dictionaries and statistic instruments, which majority of Web term extraction systems apply to. They consider lexical unit frequency, likelihood ratio for two-word terms, mutual information measures. The only statistical parameter for longer collocations evaluation is term candidate frequency in the text corpora. A hybrid approach based of both linguistic and statistical methods is mostly practiced nowadays since it helps to overcome unilateral approach problems.

Another type of a language worker resource is useful for the tasks of knowledge extraction. The resource in question is what is known as Web semantic models.

Semantic models may be roughly described as conceptual, semi-formalized representations or structures that reflect data domain expert understanding and reasoning strategies in professional task solution. Knowledge extraction is a procedure of specialist interaction with a source of knowledge (another person - expert, textual documents of different structure, nature and complexity, multimedia information, etc.), that results in eliciting implicit information from the natural language texts and other sources. Conceptual analysis or knowledge representation is analysis of information, received from a knowledge source and its synthesis (or coding) into certain structures, independent of any program realizations - semantic models.

Adequate facilities for text information analysis that are absolutely necessary to obtain useful knowledge, hidden in the huge volumes of raw data, are such as FrameNet or WordNet.

A frame semantics approach is realized in FrameNet (https://framenet.icsi.berkeley.edu), that is an automatic system for laying down and documenting a whole range of semantic and syntactic combinational capabilities (valencies) of a word in its every possible meaning. In this project a semantic frame is a conceptual structure for formalized situation description, including various events, relations, objects, participants and other conceptual roles, each representing a frame component. The FrameNet lexical database contains about 1200 semantic frames, 13000 lexical units (in this system a lexical unit is a word – definition pair; polysemantic words are presented by several pairs) with more than 190000 annotated sentences used as examples. This project appeared to be effective both for linguistic tasks and natural language texts processing [9].

A constructional or differential approach underlies WordNet system developed in Princeton University (https://wordnet.princeton.edu), where lexical semantics is used for meaning representations. Under constructional approach meaning description should contain information sufficient for specifying an appropriate concept structure (by a person or a computer), that is hardly the case even with conventional explanatory dictionaries. According to differential approach to meaning a description can be done with any symbol permitting to distinguish meanings. These are simpler requirements, but they are sufficient for necessary conformity.
WordNet has four functional parts: original source files for lexicographers; software for converting these files into WordNet lexical database; WordNet lexical database proper; software for database management. It is important to note, that using the WordNet system assumes basic knowledge of English.

Both systems described shall be used as important linguistic resources to optimize a language worker’s performance, as part of professional AWS and as a model for similar systems development.

4.2. A language worker’s workstation

A language worker’s activity includes information extraction, communication across languages and cultures, translation and dictionary creation, etc. which requires use of various linguistic and software tools. An effective way to arrange a complex of tools is an automated workstation (AWS). Competent professional use of an AWS, in 2003 successfully started in Leningrad-St. Petersburg Engineering Linguistic School, has proved at best with MT-specialists and now becomes customary in various knowledge domains besides translation. A translator’s AWS usually includes domain-specific information retrieval systems, a set of resident dictionaries, thesauri and spell-checking systems.

Russian GOST 34.003-90 describes an automated workstation as a program-technical complex of computer-aided systems for automation of special activity types. A structured complex of standard and special software with appropriate data and knowledge bases installed on a personal computer, a standard AWS was initially meant for non-humanitarian knowledge domain specialists.

Actually, the idea of AWS use in humanitarian sciences began to be broadly realized in the 90s of the last century. Introduction of rapidly developed specialized complexes for professional translator became a challenge for both users and designers, for more details [14].

That was due to the conscious need in such tool for better professional activity organization, on the one hand, and to the growth of new accessible computer facilities, enabling creating and functioning of such AWS, on the other. Realization of AWS in its present form has required considerable development in computer systems, special technologies and devices, auxiliary aids and systems.

This breakthrough was significant, however introduction of a translator’s AWS was not an absolutely new thing, since successful proposals to create its components (including translational memory) had been made more than 20 years before.

A today AWS for humanitarian use constitutes an information system, elements of which are not material objects but data (information) types. This information system, determined by the set of its functions, is a set of hardware-in-the-loop tools and algorithmic procedures, aimed at collection, input, storage, simulation and image presentation of data.

Today changes in computer and internet use caused by enormous growth of personal computers and different user categories in the Internet generate new conditions for specialists in humanitarian knowledge domain. They create new high-tech environments, which gradually expand: industrial, research, educational, for review see [15-22]. They require new forms of information transfer, knowledge extraction, including a new context of training. Hence, the growing creation of national knowledge funds in various fields of human activity, funds which give opportunity to receive information on the stored items, fill it up and modify as necessary.

The changes mentioned above inevitably cause updating computer facilities and transferring to the next generation computers, which in its turn has a negative impact on automatic information processing, since each generation transfer is necessarily followed by a system creation all over again, involving the change of principles and methods, revision of theoretical bases and realization principles. This spiral development is characteristic for all kinds of information processing systems: MT-systems, information retrieval, annotating and abstracting, computer training etc.

At different stages of computer design, it was noticed that any progress or problem solution in computer technologies much depends on success in natural language processing (NLP) development. As NLP is the key tool to generate, store and transfer information, information technologies in NLP field (linguistic technologies) are an important factor relevant to information technologies as a whole. NLP tools shape the core of a language worker’s AWS structure.
Hardware peripherals (scanners, modern information carriers, digitizers etc.) support the idea of AWS since they save time for text punching and automatically arrange natural texts for further processing.

An AWS for a language worker (a freelance translator or a translator team member, lexicographer, terminologist, technical writer, etc.) can perform a complex of functions in order to organize their on-site activity, including facilities for

- multilanguage text preprocessing,
- optical character recognition,
- information compression and extraction,
- document transfer and search over information networks
- spelling and grammar checking,
- terminology management (concordances and other dictionary types creation, access to local/remote data banks or other linguistic resources),
- translation memory management,
- machine translation.

A prototype language worker’s AWS structure then should include a number of programmes and linguistic modules listed below:

1. A linguistic automaton as a complex of automatic text processing tools, used for translating, editing and annotating texts, information retrieval and analysis and other kinds of text processing operations. It presents a hierarchical system of program modules, each responsible for a certain operation and can function both independently and in complex with other tools.

2. A full-text database for text storage, modification and browsing. It is now evident that such a base should certainly include references to national text corpora and net resources. This part of an AWS structure can be immediately used for specific linguistic facts analysis, and textual and comparative analysis as well. One more important application of such a base is that of being an important source for building dictionaries of different composition and needs;

3. A terminological database for storing, structuring, thesauri and ontologies building, as well as for term extraction and translation. It comprises electronic dictionaries developed and routinely used by translators, such as Lingvo (www.abbyy.ru) or Multitran (https://www.multitran.com), translation servers (https://www.promt.ru) and portals (www.slovari.ru). Lexicographic resources to be included in the AWS structure can be oriented to extract and process terms and/or concepts, to work with bi-, multi- or monolingual resources;

4. A reference base for specialized glossaries, automatic and educational dictionaries, united in an integrated complex, that makes it possible to use any collected dictionary information for educational and scientific goals;

5. A base of specialized linguistic software, among the other things the machine translation systems, translational memory, systems of terms extraction from parallel and comparable texts corpora;

6. A reference base for actual informational standards requesting and delivering intelligent information between individual and independent companies, such as iiRDS (intelligent information Request and Delivery Standard);

7. A set of programmed data models for processing and generating texts of fixed formats (DocBook, DocsVision, DITA).

A set of similar AWS (for example, a Canadian Alis Translation Solutions (ATS) system) can be also found in the Internet today. They are provided with a series of natural language processing tools and a set of auxiliary services: systems for supporting human and machine translation, on-line dictionaries, browsers, tools for operation control.

A modern AWS for a language worker should provide an opportunity for individual language worker’s resources to be used with on-line ones. A distributed network which unites individual AWSs
should assign terms for personal access to dictionary databases, since terminological units are extracted and translated by different users and at different stages of dictionary base formation. Creating common terminological resources implies a working structure for different users, which will regulate changing the term bases and translation dictionaries, following the potential situations [6]:

- A terminological database user finds a new text term, absent in the dictionary base.
- A user-proposed pair “term – translation” can be entered into the resource of an individual AWS, then transferred to the terminological data manager, who delivers it for the expert evaluation, and hereinafter a decision on term introduction into common dictionary resource is to be made.
- A designer of actual terminological resource enters a new term and its translation.
- In this case a decision on introducing the term in question into the integrated base is required.
- A product designer enters a new term, which should be approved by experts.

Thus, a language worker’s individual AWS resource base, remaining an individual user tool, should be permanently evaluated as meeting requirements of a potential common resource.

5. Conclusions and Implications
Following the research goal, we have pinpointed the forthcoming changes in job market induced by automation processes growth in industries, the so-called forth industrial revolution (Industry 4.0). We also argue that among jobs competitive under new circumstances there is one that actually meets the demands of Information 4.0 forms and technologies, namely a language worker. Focusing on a language worker’s supportive competences we describe facilities for their successful performance. Analyzing current applied linguistics researches and developments which deal with language and knowledge engineering we claim that:

1. Language resources for a language worker comprise text resources proper, arranged as powerful databases, and numerous language knowledge sources, such as thesauri, ontologies, web semantic frames and models;
2. A language worker shall be taught to use multilingual lexicographic resources of today that are web-based, both governmental and initiative, developed by corporations or research groups (Eurodicautom, EuroTermBank and the like);
3. The use of corpus technologies, such as obtaining concordances, dictionaries of words and collocations in case of monolingual corpus, as well as creating multilingual lexicons and concordances in case of parallel or comparable text corpora is another way to solve terminology problems in a language worker performance;
4. To cope with knowledge extraction issues a language worker can apply to resources known as Web semantic models (FrameNet, WordNet);
5. An effective way to optimize a language worker’s performance and to arrange a complex of tools and resources is an automated workstation (AWS);
6. The proposed prototype language worker’s AWS can perform a complex of functions in order to organize their on-site activity. We suggest a module structure for the AWS and describe terms for individual and distributed work.

References
[1] Mashelkar R A 2018 Exponential Technology, Industry 4.0 and Future of Jobs in India Review of Market Integration 10(2) 1–20
[2] Reave L 2004 Technical Communication Instruction in Engineering Schools A Survey of Top–Ranked U.S. and Canadian Programs Journal of Business and Technical Communication 18(4) 452-490.
[3] Meex B, Karreman J 2016 TecCOMFrame. Towards a European Competence Framework tekom-Jahrestagungundtcworld conference in Stuttgart. Zusammenfassungen der Referate.
Stuttgart: tcworldGmbHVerantwortlich, 486-489

[4] 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 95-97

[5] 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 93-94

[6] Beliaeva L, Chernyavskaya V 2019 Technical Writer in the Framework of Modern Natural Languages Processing Tasks Journal of Siberian Federal University. Humanities & Social Sciences 1 (12) 20-31. DOI: 10.17516/1997-1370-0377

[7] Chernyavskaya V 2019 Corpus-Assisted Discourse Analysis of Russian University 3.0 Identity Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya - Tomsk State University Journal of Philology 2019 58 97-114. DOI: 10.17223/19986645/58/7

[8] Almazova N I, Beliaeva L N, Kamshilova O N 2018 Towards Textproductive Competences of a Language Worker and Novice Researcher 18th PCSF 2018 - Professional Culture of the Specialist of the Future: The European Proceedings of Social & Behavioural Sciences EpSBS L1 103-109

[9] Hutchins J 2001 Machine translation and human translation: in competition or in complementation? International Journal of Translation 13(4) 287-307.

[10] Hutchins J 1988 The Origins of the Translator’s Workstation Machine Translation 13 5-20

[11] Johnson I, MacPhail A 2000 IATE – Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. In LREC Workshop on Terminology Resources and Computation, Athens URL http://mt-archive.info/LREC-2000-Johnson.pdf

[12] Ruppenhofer J, Ellsworth M, Pettruck M R L, Johnson C R , Scheffczyk J. 2016 FrameNet II: Extended Theory and Practice Berkeley, California International Computer Science Institute 129 URL: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

[13] Großjean A 2009 Corporate Terminology Management: An approach in theory and practice VDM Publishing 100

[14] Beliaeva L, Chernyavskaya V 2016 Evidence-based linguistics: methods in cognitive paradigm Voprosy Kognitivnoy Lingvistiki 3 77-84. DOI: 10.20916/1812-3228-2016-3-7.

[15] Almada-Lobo F 2016 The Industry 4.0 revolution and the future of manufacturing execution systems (MES) Journal of Innovation Management 3 16-21.

[16] Whitmore A, Agarwal A, Da Xu L 2015 The Internet of things - A survey of topics and trends Information Systems Frontiers 17 261-274 DOI:10.1007/s10796-014-9489-2

[17] Kimball M 2016 The Golden age of Technical Communication Journal of Technical Writing and Communication 1 1-29

[18] Cleary Y, Flammia M 2012 Preparing technical communication students to function as user advocates in a self-service society Journal of Technical Writing & Communication 42(3) 305–322.

[19] Hamer S, Rich A 2005 Trends in undergraduate curriculum in scientific and technical communication programs Technical Communication 52(2) 209–220.

[20] Meloncon L 2009 Masters programs in technical communication Technical Communication 56(2) 137–148

[21] Meloncon, L 2012 Current overview of academic certificates in technical and professional communication in the United States Technical Communication 59(3) 207–222.

[22] Beliaeva L, Chernyavskaya V 2019 Scientific and Technical Texts in the framework of Information 4.0: content analysis and text synthesis St.Petersburg State Polytechnical
