Journal of Siberian Federal University. Humanities & Social Sciences 12 (2016 9) 2904-2912
~ ~ ~

УДК 81’33

On Methodology of Knowledge Presentation: Terms and Translation in Seismic Domain

Larissa Beliaeva[a]
and Valeria Chernyavskaya[b]*
[a] *Herzen University*
48 Moika river Str., St. Petersburg, 182100, Russia
[b] *Peter the Great St. Petersburg Polytechnic University*
29 Politekhnicheskaya Str., St. Petersburg, 195251, Russia

Received 14.12.2015, received in revised form 30.04.2016, accepted 28.10.2016

The underlying assumption of the paper is that the text is the result of information transfer and the source (starting point) o of information mining and extraction. Thereafter the scientific text content is universal and can be extracted in case of minimum coincidence of the author’s and recipient’s thesauri. In text translation the proper information extraction is determined with correct interpretation of terms/objects names. The paper focuses that adequacy of text perception and understanding on the lexical level is dictated by the text saturation with noun phrases, the degree of their compression and / or completeness of object nominations and their proper translations based on adequate bilingual dictionaries. Text translation in such a hazardous domain as aseismic construction is a crucial point in information exchange and mining and a safety challenge. For terminology analysis of this domain text a special research text corpus (1 mln tokens) had been created. The analysis results had been used in the specialized automatic dictionary creation for machine translation system WORD+.

Keywords: knowledge mining, data mining, text structure, text translation, terms, term extraction, noun phrases, aseismic construction domain.

DOI: 10.17516/1997-1370-2016-9-12-2904-2912.

Research area: philology.

Introduction

The process of constructing scientific knowledge includes two different stages: generating a piece of new scientific knowledge and incorporating it in the system of the existent amount of scientific knowledge. Cognition-centred and anthropo-centred aspects of text analysis are based on the fact that a scientific text is the result of cognitive and communicative activities of the certain subject in a scientific domain aimed at a specific object – facts and processes of actual reality. The structure of a scientific text reflects the main components of a human scientific cognitive activity and contains the scientific knowledge as its product.

The peculiarity of the modern situation in data and information mining is related to the fact that a huge amount of international

* Corresponding author E-mail address: laurabel@gmail.com; tcherniavskaia@rambler.ru

communication uses English as *lingua franca*, that means that lexical and syntactic structures of the sentences used in these texts are frequently under strong influence of the mother tongues of their authors. Text translation in such a hazardous domain as aseismic construction is a crucial point in information exchange and mining and a safety challenge. Text is to be considered as the result of information transfer and the source (starting point) of information mining and extraction. Thereafter the scientific text content, especially the part of its sense which is universal and can be extracted in case of minimum coincidence of the author's and recipient's thesauri, is mainly determined by information on the objects under consideration, linguistically described by their names – nouns and noun phrases.

Scientific Knowledge Presentation

The process of constructing scientific knowledge includes two different stages: generating a piece of new scientific knowledge and incorporating it in the system of the existent amount of scientific knowledge. It is necessary to differentiate, on the one hand, the type and way of thinking in the process of cognition (Crombie, 1994) and, on the other hand, the way of presenting scientific results by linguistic means in a special text. Scientific text and to be more exact scientific paper is a type of special text.

Thus, it is the text that is at the intersection of cognitive and communicative aspects in scientific communication, in fact the text is the source of information on the author, addresser and on scientific and communicative situation, in which it was created.

Two aspects play a key role when we analyze a scientific text: cognition-centred and anthropo-centred. These two aspects are based on the fact that a scientific text is the result of cognitive and communicative activities of the certain subject in a scientific domain aimed at a specific object – facts and processes of actual reality. The structure of a scientific text reflects the main components of a human scientific cognitive activity and contains the scientific knowledge as its product, such text can be used in the processes of data and information mining (Delgado et al., 2002). The textual data are not unstructured and a scientific text has a very complex implicit structure and is the source for solving the problem of the Knowledge Discovery in Texts (Feldman, Dogan 1995).

Text production takes place in the framework of specific epistemic situation. It has been defined as the realization consequence of four main aspects of cognitive activity by the subject (the author of a scientific text) of textual activity, these aspects are based on the mental mechanisms of the author interaction with the world that are reflected in his/her the terminology used (Orel, 2007). They are ontological aspect, connected with the subject matter of scientific knowledge, methodological aspect, related to the procedure of knowledge acquisition (knowledge mining), axiological aspect, connected with evaluation-based nature of the subject of science and communicative-pragmatic aspect.

The ontological knowledge aspect means interaction of crucial constituents of science continuum – 'old' knowledge acquired earlier and presented in the available texts and the author's thesaurus, 'new' knowledge acquired personally by the author of a newly produced text, and predictive knowledge which acquisition is potentially possible. Interpretation of the existing ideas implies an active role of the subject of cognition and is closely related to the methodological aspect of the epistemic situation. It characterizes the ways of acquiring, presenting, interpreting scientific knowledge and presupposes conscious control over the cognitive process. Specifics of the methodological aspect is predetermined by the nature of science as

a specific form of cognition – rational and argumentative cognition. Reflection of reality in texts is not the only thing which is typical for the methodological aspect. Conscious control over the cognition forms and conditions is also important. Moreover, the resulting text understanding is based on the terminology used as on the cognitive and language braces that hold together the author's reasoning and the domain of knowledge and make possible text adequate (or nearly adequate) understanding by specialists in the same domain. If the terminology used is not harmonized or not known these braces are broken and adequate understanding is impossible.

Subjective anthropocentred component of science is further revealed in the process of the subject's reflecting upon the object of cognition, the result of this reflection depends on the author's subjective reality *(qualia)* (Edelman, 2004; Chernigovskaia, 2013).

The reflexive aspect of scientific cognition is expressed both in evaluative orientation of the subject of science in the sphere of 'old' and 'new' knowledge and his/her values, beliefs, psychological patterns, emotional reaction to the problem. Another aspect, communicative-pragmatic, is closely related to text production which is in its turn aimed at presenting the acquired knowledge to the recipient. This implies that there occur complex transformations of all extra-linguistic factors into linguistic, text-oriented ones (Bazenova, 2001; Chernyavskaya, 2011).

As a scientific text is considered to be the result of a targeted cognitive and creative process, the semantic structure of a scientific text reflects this creative process and stages of problem-solving. It means that a scientist in text-producing models a system of mental operations which leads to acquiring a certain piece of knowledge with the help of linguistic and compositional forms.

The process of constructing knowledge can be shown as a sequence of the following stages: problem situation – problem – idea – hypothesis – hypothesis confirmation – conclusion. Thus, knowledge construction is a dynamic process that presupposes moving from one cognition phase to another. Only after a step-by-step process a scientific idea is finally organized in the text form. Both semantic structure and composition of the text correlate with these cognition stages (Delgado et al., 2002).

It is the structure of the cognition stages that serves as a "skeleton" for further arranging and organizing separate text parts and providing coherence and divisibility of its composition. This structure determines the strategy of text production. The typological characteristics that have been pointed out form the basis of modern understanding about the processes of scientific text organization.

The main ideas of text organization can be described in the following way: the subject of speech deals with a certain problem by choosing a certain way of text organization (Antos, 1982). Thus, there is a mode of text organization i.e. a specific set of speech actions undertaken by the speech subject in the process of text organizing in order to solve specific cognitive-communicative tasks. Taking into account the definition of mode as a variable characteristic inherent to the subject only under some conditions in contrast to constant characteristics permanently inherent to the subject it can be claimed the following. A mode of text organization characterizes some characteristics of a text complementary to those that are compulsory and invariable in text production.

The patterns a particular text follows the nominations used for the object of the study, a representation of real world and subjective reality is related to knowledge representation (Delgado et al., 2002).

Scientific Text and Terminology Translation: Discussion

Now the need for a fast, accurate and cheap translation from and into English has become very pronounced, the problem of information flow organization and processing in different languages acquires a new practical significance. The peculiarity of the modern situation in data and information mining is related to the fact that a huge amount of international communication uses English as *lingua franca*, that means that lexical and syntactic structures of the sentences used in these texts are frequently under strong influence of the mother tongues of their authors. Text translation and data mining in such a hazardous domain as aseismic construction is a crucial point in information exchange and mining and a safety challenge and should be based on vast investigation of real texts. For terminology analysis of this domain text a special research text corpus (1 mln tokens) had been created, the analysis results had been used in the specialized automatic dictionary creation for machine translation system WORD^{+}.

Text is to be considered as the result of information transfer and the source (starting point) of information mining and extraction. Thereafter the scientific text content, especially the part of its sense which is universal and can be extracted in case of minimum coincidence of the author's and recipient's thesauri, is mainly determined by information on the objects under consideration, linguistically described by their names – nouns and noun phrases. Furthermore, the proper information extraction from a scientific text is determined by correct interpretation of terms (objects names). Thus the adequacy of text perception on the lexical level is determined by text saturation with noun phrases, the degree of their compression and/or completeness of object nominations. In synergetic aspect we can assume that when the translated text is oversaturated with long (Tab. 1 presents the list of the 8-components noun phrases) or excessively convoluted lexical complexes, text is destroyed and its perception is impeded or even impossible.

Thus the text is the source for terminology extraction and harmonization, the only source for dictionary creation, the problem solving of which in such a hazardous domain as aseismic construction. Modern approach to dictionary creation assumes preliminary formation and use of parallel corpora of modern texts, which can be considered as a database for solving not only research tasks, but practical lexicographic tasks as well. Written text corpora, as a rule, include the texts as they are, as well as text layouts: format boundaries and features, parsing results necessary for establishing morphological characteristics of lexical units. These texts can be used serve for concordance creation, word and collocation lists in case of monoligual corpora, as well as for creation multilingual lexicons and concordance if we have parallel corpora.

Noun phrases as the terminology items are the objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are rather units of syntax, not lexicon. Thus we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the appropriate sentence and reflect the peculiarities of the nomination object. The problem is to find a procedure or approach to recognize this structure in a convolution.

One of the most serious problems of English scientific text analysis and machine or human translation is determination of dependency structure in noun phrases (NP). The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the NP

Table 1. Eight-Component Noun Phrases in Aseismic Construction Text Corpus[1]

Noun Phrase Model	Corresponding Noun Phrases
A1+N1+N2+N3+A2+N4+N5+N6	electric power research institute seismic margins assessment methodology
A1+N1+A2+N2+N3+N4+N5+N6	high amplitude long period pulse type ground motion
A1+A2+N1+N2+N3+A3+N4	immediate nuclear power plant operator post-earthquake actions
N1+A1+N2+A2+A3+N3+N4+N5	MWe light water commercial nuclear power plant facility
A1+N1+Pr+N2+N3+N4+N5+N6	single degree of freedom circle curve fit method
N1+N2+A1+N3+PII+A2/N4+N5+N6	test-bed nuclear facility reinforced concrete frame structure
H+A1+A2+A3+C+A4+A5+N1	three-dimentional nonlinear static and dynamic structural analysis
N1+A/N2+N3+N4+PII+N5+N6+N8	fiber cross-section flexibility-based beam-column element
A1+A2+A3+A4+N1+N2+N3+N4	multi-cylindrical vertical flat bottom liquid storage tanks
A+N1+N2+N3+N4+N5+N6+N7	state-of-the-art frequency domain soil-structure interaction analysis software
A1+A2+N1+A3+N2+N3+N4+N5	programmable high speed multi-channel data acquisition system

components. In scientific text simple noun phrases are multicomponent units with large number of attributive elements in preposition to the NP head. Being dependent members of a sentence, these phrases form one syntactic group with its head, syntactic function of which coincides with syntactic function of the phrase as a whole.

Since a NP is a sentence convolution, a compression of this structure, such external simplification of both the structure and the form causes the NP semantic complication. The markers of relations between actual components and types of relations between elements, which sentence shows with the help of different means, are absent in the English NP. Basic noun phrases in English are two-element combinations with a head noun, frequency of which in the aseismic construction domain exceeds the frequency of three-component phrases in three times (Table 2, 3).

However external simplicity of the most frequent English NP structures is misleading. The fact is that this simplicity could be the result of initial noun phrase or sentence compression. Such compression, formal simplification of NP structure leads to its semantic complication.

Pursuant to these, formation of noun phrases in a real text is based on either merging noun phrases and separate lexical units in a new, more complicated nominative construction, or on condensing multi-component NPs at the expense of deletion of the units which are implicitly obvious (Belyaeva, 209; Beliaeva, 2014).

Text-formation of a multi-component NP is realized depending on the type of nomination: either as a process of step-by-step complication and specification of the object nomination (gradual complication of a noun phrase with addition of its head element characteristics), or as a process of sequential convolution, the process is realized successively on several levels:

Level 1: transfer from a complex noun phrase to a simple one due to element inversion.

Level 2: elimination of component duplication in a new NP.

Level 3: coordination of semes and elimination of components with duplicated semes.

Referential status of noun phrases in a scientific text permits us assume that author's attitude on information transfer and its understanding requires explication of relations

Table 2. Fragment of Frequency List of English NP Structures[2]

Rang	NP type	NP frequency	Accumulated frequency	%
1	T + N	2253	2253	23,07
2	A + N	1584	3837	39,29
3	N1 + N2	1368	5205	54.32
4	A + N1 + N2	485	5690	58.26
5	H + N	241	5931	60,74
6	M + N	215	6146	62,,93
7	PII + N	208	6354	65,06
8	N1/A + N2	199	6553	67,09
9	S + N	195	6748	69,09
10	N1 + N2 + N3	173	6921	70.87

Table 3. Models of Noun Phrases in Aseismic Construction Text Corpus

Model number	Model	Length	Frequency	Number of different NP
1	A+N	2	1474	748
2	A+PII+N	3	9	6
3	N1+N2	2	1407	530
4	N1+N2+N3	3	248	128
5	A/N1+N2	2	71	47
6	N1+G/N2	2	24	18
7	A1+A2/N1+N2	3	10	10
8	A+G/N2 +N2	3	7	4
9	A1+A2+N	3	151	104
10	A+N1+N2	3	292	172
11	PII+A+N	3	25	20
12	PII+N	2	170	73
13	A1+N1+A2/N2+N3	4	3	2
14	A1+C+A2+N	4	15	9
15	A1+A2/N1+N2+N3	4	6	7

within the text. Analysis of texts in different subject domains and purposely in the aseismic construction domain had shown that occurrence a NP with length more than 2 elements is followed by occurrence a 2-component NP in the nearest context, within the limits of 2-3 sentences or combination of title, key words and abstract. Hence, at human translation we can use this situation as a key for structure diagnostics.

The peculiarities of NP formation in the text can be shown by the following example: three-component NP in case, when semantics of one of the elements is implied as a part of a new noun phrase at the expense of extralinguistic information of the domain, see the merger of NPs *seismic stability* + *direct analysis* in formation of a new NP *seismic stability direct analysis*, which in the text may be convoluted

up to a three-component NP *seismic direct analysis*.

The cases of noun phrase transformation under the condition of text coherence and cohesion give the basis for consideration of possible translation of a noun phrase with high degree of structure compression. Besides the research conducted permits to show, that exactly two-component noun phrases present special difficulties at their analysis and translation. To solve the problem of such NP translation we can propose only two approaches which can be used in translation.

The first approach includes modeling the knowledge base of the domain in question or appealing to such factual knowledge of the translator. This approach is based on vast investigations of the possible relations between both the main concepts of the domain and the items of the linguistic data base. Creation of such a thesaurus or a semantic net is extremely laborious. But the most serious disadvantage of this approach is that an unambiguous solution of the problem sometimes can't be achieved. For example, for a noun phrase *constant amplitude deformation cycle* a semantic network would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle* and it is impossible to use this information to establish the dependencies structure of the NP in translation.

The second approach could be more formal: we can use the information, which can be received on the basis of the whole text analysis. This approach seems more expedient as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different NP with the same constituents. Investigations of text structure in terms of NP composition in different subject domains (medicine, seismic isolation, space systems, power plants construction, language teaching etc) had shown that dependency structure of NP with 3 or more components can be obtained from the nearest context: a 2-component NPs would show the accurate relations relevant for this special text.

Final remarks

Now it is common place to advocate that text analysis and translation problem solving can be considered in language engineering aspect, the essence of which is development and/or adaptation of modern computer systems of natural languages text processing for specific research and technical problems, which are under investigation in various areas of science and engineering. Fast and correct translation of the proceedings of international conferences and symposia, current reports on the work of international research teams, coordination research meetings, harmonization of Norms and Standards in risky areas is a necessary tool for international cooperation and safety.

[1] Conventional signs: A – adjective, N – noun, PII – participle II, C – conjunction, H – numeral, Pr – preposition, / - homonym.
[2] Additional conventional signs: T – article, S – determinative, M – pronoun, G – Participle I.

References

Antos, G. (1982). *Grundlagen einer Theorie des Formulierens. Textherstellung in geschriebener und gesprochener Sprache*. Tübingen. 362 S.

Belyaeva, L. (2009). Scientific Text Corpora as a Lexicographic Source SLOVKO. *NLP, Corpus Linguistics, Corpus Based Grammar Research*, Proc. from the Intern. Conference, November 25 – 27. Smolenice, Slovakia. pp. 19-25.

Beliaeva L. (2014). Applied Lexicography and Scientific Text Corpora Transcations on Business and Engineering Intelligent Applications, Galina Setlak, Kassimir Markov (ed.). Rzeshov, Poland: ITHEA. pp.55-63.

Baženova, E. (2001). Nautschnyj text v aspekte politekstual'nosti [Scientific Text in polytextual dimension], Perm. 270 p.

Chernigovskaya, T.V. (2013). Cheshirskaya ulybka kota Shredingera: jasyk i soznaniye. [Cheshire smile of Schrödinger's cat: language and consciousness] Moscow: Slavonic Culture Languages. (Language and Reasoning). 448 p.

Chernyavskaya,V. (2011) Interkulturelle Differenzen von wissenschaftlichen Texten. *Fach -Translat – Kultur. Interdisziplinare Aspekte der vernetzten Vielfalt,* Band 2. K.-D. Baumann, ed., Berlin: Frank& Timme. pp. 1241-1270.

Crombie, A. (1994). Styles of Scientific Thinking in the European Tradition, London. 3 vol.

Delgado, M., Martin-Bautista, M.J., Sanchez, D., Vila, M.A. (2002). Mining Text Data: Special Features and Patterns Lecture Notes. *Computer Science*. Springer-Verlag GmbH. pp. 140-151.

Edelman, G.M. (2004). Wider than the sky: a revolutionary view of consciousness. Penguin. 360 p.

Feldman, R., Dagan, I. (1995). Knowledge discovery in textual databases (KDT). *Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining* (KDD-95), AAAI Press, pp. 112–117.

Orel, T. (2007). Terminology analysis by means of frame construction (on the basis of the English telecommunications terminology). *XVI European Symposium on Language for Special Purposes (LSP) "Specialized Language in Global Communication*". Hamburg: University of Hamburg. pp. 119-121.

Методология представления знания: текст, термин и перевод (на материале предметной области «сейсмостойкое строительство»)

Л. Беляева[а], В. Чернявская[б]
Российский государственный педагогический университет им. Герцена
Россия, 182100, Санкт-Петербург, наб. р. Мойки, 48
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, ул. Политехническая, 19

Статья анализирует текст как результат передачи информации и источник (отправную точку) для ее поиска и извлечения. Соответственно, содержание научного текста универсально и может быть извлечено в случае минимального совпадения тезаурусов автора и реципиента. При переводе корректное извлечение информации определяется верной интерпретацией терминов/имен объектов. Основное внимание в статье уделяется восприятию и пониманию текста на лексическом уровне, что определяется насыщением текста именными группами, степенью их компрессии и/или полнотой номинации объектов, а также их корректным переводом на основе соответствующих двуязычных словарей. Перевод текста

в такой области повышенного риска, как сейсмостойкое строительство, имеет определяющее значение для обмена и поиска информации, а также важен для решения проблем безопасности. Для терминологического анализа этой предметной области составлен специальный исследовательский корпус текстов объемом 1 млн словоупотреблений. Результаты анализа использованы при создании предметно-ориентированного автоматического словаря для системы машинного перевода WORD⁺.

Ключевые слова: инженерия знаний, технологии извлечения знаний, структура текста, перевод текста, термин, именная группа, сейсмостойкое строительство.

Научная специальность: 10.00.00 – филологические науки.