УДК 821.161.1 + 81'322.4

Шереметьева С.О.

Прикладной анализ дискурса как основа повышения качества профессиональной документации

В докладе рассматривается вопрос использования прикладного анализа дискурса для повышения качества профессиональной документации. Выявляются проблемы, возникающие у людей при понимании или переводе текстов, в том числе и машинном переводе. Показывается как результаты анализа дискурса использованы при разработке инструментария, поддерживающего авторскую деятельность и обеспечивающего создание документов с высоким уровнем понимаемости и переводимости. Поддержка авторская деятельности понимается как интерактивная процедура, которая информирует автора о проблематичных фрагментах текста и контролирует лингвистические параметры создаваемого документа. Методика, описанная в статье универсальна и дает хорошие результаты. Подход иллюстрируется на примере компьютерного инструмента поддержки авторской деятельности в области машиностроения.

Ключевые слова: профессиональная документация, компьютерная поддержка авторской деятельности, удобочитаемость, переводимость

Шереметьева Светлана Олеговна, доктор филологических наук, профессор Южно-Уральского государственного университета, e-mail: linklana@yahoo.com

UDC 821.161.1 + 81'322.4

Sheremetyeva S.O.

Applied discourse analysis targeted at improving the quality of professional documentation

I the paper an applied aspect of discourse analysis targeted to effective support in professional writing is put into focus. Barriers to human understanding (readability), human and machine translation (translatability) of professional documentation are discussed. A methodology to develop an authoring system for improving the quality of a document based on the results of discourse analysis is presented. Authoring is viewed as an interactive procedure that makes professionals aware of the typical areas of concern and controls linguistic parameters of a document to make it more readable and translatable. The methodology is universal and provides for intelligent output. It is illustrated on the example of a computer tool for research papers on engineering in the Russian language but can be applied to other languages and domains.

Keywords: professional documentation, computer-aided authoring, readability, translatability

Sheremetyeva Svetlana Olegovna, Doctor of Philology, Professor, the Department of Linguistics and Cross-Cultural Communication, Faculty of Linguistics of the SUSU (NRU) tel.: 8-351-900-92-36, E-mail: linklana@yahoo.com

1. Introduction

Discourse analysis is nowadays a popular trend in many fields of research. There are a lot of approaches to and definitions of discourse and discourse analysis. The frame and the purpose of this paper do not allow us dwelling on the theoretical aspects of these issues, we just refer the reader to a very good review given in (Khomutova, 2010). We will note here that applied discourse analysis seeks to identify ways in which discourse analytic research can provide for recommendations for human practice (see, e.g., Willig, ed, 1999). We claim that parameters of applied discourse analysis should be defined by its specific purpose as particular type of domain. The purpose of our research is to use its results to improve professional writing by suggesting both recommendations and a computer tool.

Professional documentation is an indispensable means of scientific and technical progress in the human society. Being an important communication media in the dissemination and assimilation of domain specific knowledge professional texts should be highly comprehensible for the interested audience both in the native and foreign languages. This is directly related to such parameters as text readability and translatability. Readability is related to the level of the clarity of a text for human understanding. By the interested audience we mean the native language professionals and human translators, the latter are responsible for the comprehensibility of a document in a foreign language. Professional texts are often extremely difficult to understand (low readable) for both mentioned categories of the human audiences not only because of the abundance of specific terminology but also due to the complex syntax and syntactic ambiguity. This is especially relevant for translation, whose operativeness given the exploding volume of professional publications and ever increasing demand for international information exchange is nowadays put in focus. This, in turn, leads to the wide use of machine translation (MT), not withstanding its quality problems. The efficiency and quality of machine translation depends on the translatability indicators of a source language text. Among the translatability indicators that lower the quality of machine translation are such linguistic phenomena as lexical ambiguity, sentence length, coordination and syntactic complexity, etc. (Underwood and Jongejan, 2001). Translatability correlates (though does not coincide) with the notion of readability. Normally, if a source language document is both highly readable and translatable it guarantees the success of professional communication on both domestic and international levels.

However, professionals (scientists and technicians), both in Russia, as well as abroad, concentrating on the content of the document do not always express their findings in a good language. Despite many writing instructions such as, e.g., GOST (GOST, 1997), the correlation between theory and writing practice remains problematic. As this often leads to failures in professional commu-

nication in the source language and mistakes in translation a strong need for effective computer system to support professional writing is evident.

In this paper we attempt to contribute to the solution of the problem by suggesting a methodology of the computer supported improvement of the readability and translatability of professional texts. To prove the viability of the methodology it is implemented into a tool that makes professionals aware of the typical areas of concern in their texts and provides an authoring environment. Though illustrated on Russian texts on engineering, the methodology is portable between domains and languages.

The rest of the paper is organized as follows. Section 2 is devoted to related work. Section 3 defines the tasks of the research. In sections 4 and 5 we describe the methodology and the tool, correspondingly. The results are briefly discussed in Conclusions.

2. Related work

The mainstream of the research on improving text readability is carried out in connection with developing certain text simplification techniques for particular types of audience, e.g., poor literacy readers (Aluisio et al. 2010), readers with mild cognitive impairment (Dell'Orletta et al., 2011), elderly people (Bott et al., 2012), language learners of different levels (Crossley and McNamara, 2008) or just "regular" readers (Graesser et al., 2004).

These studies are mainly done within intuitive or structural approaches. An intuitive approach suggests using less lexical diversity, less sophisticated words, less syntactic complexity, and greater cohesion. It mainly relies on the developers' intuition and experience (Allen, 2009). The structural approach makes use of certain structure and word lists that are predefined by the education level of the targeted reader which is defined by the so-called readability formulas. The readability formulas, the most popular being the formulas of (Kincaid et al,1975), are, as a rule, algorithms that measure text readability based on sentence and word lengths. To improve their readability the texts in question are modified (often manually) to control the complexity of the lexicon and syntax.

Automated systems are meant to improve readability by combining linguistic and statistical techniques and penalize writers for long words and sentences. Improvement in text readability is most often carried out on the sentence level. For example, (Siddhartha, 2002) describes syntax simplification in three stages - analysis, transformation and regeneration. Among other works related to our research is (Takao and Sumita. 2003) where text simplification is treated as a translation task within a rule-based machine translation. In (Poornima et al.2011) a rule based technique is proposed to improve readability by simplifying complex sentences based on subordinating conjunctions, coordinating and relative pronouns. (Sheremetyeva, 2003) suggests a rule-based technique for decomposing complex sentences into a set of simple sentences while preserving the initial content.

There are no publications available to us that address text readability for highly educated professionals and/or translators. However, these types of audience do often experience problems in understanding poorly written professional papers. As for text translatability, one of the latest publications reports on a statistical machine translation system from English into French where the user drives the segmentation of the input text (Pouliquen et al., 2011). Another trend to cope with the source text complexity is to rewrite the source text into a controlled language to ensure that the machine translation input conforms to the desired vocabulary and grammar constraints. A controlled language software is developed with the different levels of automation and often involves interactive authoring (Nyberg et al., 2003). The users (authors, translators) have to be taught the controlled language guidelines to accurately use the appropriate lexicon and grammar during authoring. In line with these studies is the research on developing pre-editing rules, e.g., textual patterns that reformulate the source text in order to improve its translatability. Such rules implemented in software formalisms are applied for controlled language authoring (Bredenkamp et al. 2000).

Though most of the research in readability and translatability is done for English, a number of works on readability can be found for Russian as well. For example, (Oborneva, 2006) adapts the formula of Flesch and Flesch-Kincaid for the Russian language by using adjustment coefficients. (Krioni et al., 2008) define the readability of the Russian educational texts based on the complexity of linguistic structures, integrity, connectivity, functional and semantic type, information and abstractness of the text presentation, while (Karpov et al., 2014) attempt to predict a single sentence readability through the analysis of images, social networks and texts. However, we were unable to find any publications dealing with the automation of improving the translatability of Russian texts.

3. Task definition

Our ultimate goal is to create a methodology to develop a computer tool that can on-the-fly improve readability, translatability and, hence, the quality of professional texts. We are not going to calculate any readability or translatability scores as is done in many researches on these issues. The aim of our study is not to prove that professional texts are difficult to read (understand) and translate. This is a common knowledge. We target at identifying those domain specific text phenomena that make these texts difficult to understand and translate for two categories of a highly educated human audience, - i) researches and technicians in the domain in question and ii) professional translators, who do not always possess domain knowledge and, nevertheless have to understand at least the texts' syntactic dependencies clearly. Translatability indicators are to be identified related to machine translation constraints.

We conduct our research on the material of the scientific papers in the domain of engineering

in the Russian language with the perspective to extrapolate the methodology to other domains and languages. The target of our effort is thus defined by the intersection of the following criteria:

detection of readability indicators (for humans);

detection of translatability indicators (for machine translation)

automated user support in document understanding and authoring to avoid translatability indicators.

The study was conducted based on the expert judgment which, as claimed in (Pooneh and Riazi, 2012), is much more reliable than automatic text processing based on existing readability formulas. For this work we have created and analyzed a corpus of 120 scientific papers on engineering published in "Vestnk YuUrGU" (http://vestnik.susu.ru/engineering) in 2010-2014 containing in total 203,729 word forms.

To assess the difficulties in understanding (readability indicators) the texts from the corpus were given to 20 human experts including professors, instructors and students from the engineering and linguistic departments of the South Ural State University, Russia (http//: www.susu.ac.ru). Such indicators as the understanding and translation of the terminology were excluded from the examination. The targeted audience is (i) researchers and practitioners in engineering (who are supposed to know their own terminology) and (ii) linguists-translators who cannot be required to understand the professional terminology; they are only responsible to find the Russian terms foreign (English in our case) equivalents in existing professional bilingual dictionaries or other sources. The quality of such sources is not within the frame of the current research.

The experts-professionals in engineering were to mark up the fragments of the texts which were problematic to understand the technical content as such. The experts-linguists marked those fragments of the texts where they experienced problems in understanding the syntactic structure (which is necessary for human translation). To assess the problems in machine translation (translatability) all participating experts were asked to translate the texts from Russian into English with the help of any online machine translation system and mark up those source language fragments texts which caused the mistakes in MT. The results of the experiments were analyzed and systematized by the author of the current paper. It was found that the readability indicators (terminology excluded) are the syntax related ambiguities caused by

-Long sentence length

-Coordination

-Long distance dependences

-Telescopic syntactic structures

-Long participial constructions used as attributes in the preposition of a noun phrase.

-Ambiguity in the noun/verb attachment of prepositional phrases

-Grammar mistakes in agreement

-Grammar mistakes in the use of prepositions

-Style mistakes.

Analysis of the translatability indicators showed that they include practically all readability indicators. This means that what is bad for people is bad for machines as well.

But on top of the readability indicators listed above the translatability indicators also include some linguistic phenomena that do not cause problems for humans but still lead to a number of mistakes in machine translation. These are caused by the lexical ambiguity (again, we exclude terminogy which is supposed to be covered by bilingual dictionaries) or syntactic discrepancies between the source and target languages, Russian and English in our case. Thus, ***in addition to the readability indicators listed above*** the following phenomena are included in the scope of the domain and MT–related translatability indicators:

-The order of the words

>Predicate of the sentence precedes the subject

>Noun precedes the adjective used as an attribute

-Ellipsis

-Substantivated adjectives

-Verb ambiguity

-Phrasal verbs

-Prepositional ambiguity

-Nominal groups without determiners

-Grammar mistakes in assigning number (plural or singular)

-Spelling mistakes

For example, it is not uncommon for a professional paper to include fragments like the following:

В этой связи важной проблемой современной энергетики, наряду с решением задач по альтернативным источникам энергии, является проблема режима расходования топливно-энергетических ресурсов, относящихся к числу исчерпаемых. Сегодня основным средством преобразования заключенной в топливе энергии и производства механической работы, в том числе и на транспорте, является поршневая тепловая машина с кривошипно-шатунным механизмом.

The fragment contains such readability/translatability indicators as long sentence length, coordination, long distance dependences, telescopic syntactic structures, inverse word order (predicate before subject), coordination, verb ambiguity. This makes the fragment problematic both to understand (for example, for a human translator), and to get a correct machine translation. For example, machine translation of the fragment above with the PROMT system (http://www.translate.ru) looks as follows (mistakes are marked with "*"):

*In this regard an important problem of modern power, along with the solution of tasks of alternative energy sources, the problem of the mode of an expenditure of the fuel and energy resources *which are among the ischerpayemykh is. Today the main means of transformation of the energy *concluded in fuel *and productions of mechanical work *including on transport, *the piston thermal car with the connecting rod gear is.*

Our analysis shows that to be highly readable and translatable a document should be written within the frame of a controlled language whose rules prevent the

emergence of translatability indicators. In case of a text that has already been written the problematic passages should be rewritten (authored) in the controlled language. In the latter case before any authoring the author/translator should first clearly see the syntactic structure of the original to identify problematic passages. Professional text readability increases immediately if the reader can spot the terminology at a glance. This can be achieved by the on-the fly automatic mark up of nominal and predicate terminology, and automating the process of rewriting problematic text segments. To be suitable for a real world application the methodology should allow creating a tool with computationally attractive properties. The latter suggests the use of a combination of statistical and linguistic techniques.

4. Methodology overview

To facilitate spotting the problematic linguistic phenomena in the text the methodology suggests to first automatically markup the nominal and predicate terminology and then guide the user through the process of document authoring to avoid the readability/translatability indicators. These tasks are fulfilled by a computer environment that includes a domain tuned knowledge base and the modules of analysis, authoring and text generation. The knowledge base includes a number of lexicons, a specially developed controlled language, predicate templates and rules. The workflow consists of the following main steps:

Shallow analysis based on hybrid techniques. It serves two purposes: a) the on-the-fly visualization of the input text terminology to facilitate the identification of readability/translatability indicators, and b) the preparation of a raw document for authoring by linking it to the system knowledge base.

Authoring. The document is authored to conform the controlled language. The system controlled language specifies constraints on the lexicon, word order and syntactic complexity of sentences. It draws heavily on the readability/translatability indicators given in Section 3. The constraints of the system controlled language are mainly coded in the deep corpus-based predicate lexicon whose entries contain the explicitly listed morphological forms of the domain predicates and sets of the predicate/ argument patterns. The patterns code the domain-based information on the most frequent co-occurrences of the predicates with their case-roles (arguments), as well as the linear order of the predicate-argument text realization. For example, the pattern *(1 x 3 x 2)* corresponds to such text fragment as 1:*boards* x: *are* 3:*rotatably* x: *mounted* 2: *on the pillars.*

The controlled language restrictions are imposed on the source text semi-automatically. The system prompts the user to make correct authoring decisions by providing structural templates from the system knowledge base. In addition to the controlled language constraints built in the system, the users' awareness about

the problematic linguistic phenomena is raised by a number of instructions. For example, the users are encouraged to repeat a preposition or a noun in conjoined constructions, limit the use of pronouns and conjunctions, put participles specifying a noun in postposition, etc.

Analysis. This is the most sophisticated procedure of the document processing that includes *segmentation, lexicalization* and *content representation.* The input text is automatically chunked into noun phrases (NPs) predicate phrases (VPs) and other types of lexical units. Every VP chunk is lexicalized by associating it with a lexicon entry. The NPs are chunked based on the dynamic knowledge automatically produced by a stand-alone hybrid extractor as described in (Sheremetyeva 2012). The extractor output (lists the input text NPs in their text form) are matched against the same input text and coinciding text fragments are tagged as NPs. The remaining text fragments are then chunked into VPs and by the lexicon look-up practically without any (ambiguity) problems.

Based on the results of automatic chunking and a computer-driven interview the user can call the predicate templates from the knowledge base, to author problematic fragments by properly filling the template slots according to the control language rules. The analysis results in a set of predicate/argument structures, each representing the content of a separate sentence of the final text.

Generation of the authored document without readability/ translatability indicators. At this stage the final parse is submitted into the generator that automatically outputs a restructured text of a much better readability/ translatability quality, while preserving its content.

5. The tool

A screenshot of the tool authoring interface is shown in Figure 1. In the left pane it shows the original text converted into an interactive format with nominal and predicate terminology highlighted in different colours. This is the visualization of the automatic NP and VP chunking. The highlighted terminology immediately improves the text readability and helps the user quicker and better understand the input document content and structure. To author a problematic fragment of the input so as to eliminate readability/ translatability indicators the user clicks on a highlighted predicate and gets a pop-up predicate template whose slots are to be filled out with texts strings. Predicate templates are generated based on the case-role patterns in the tool lexicon. The main slot of the template is automatically filled with a predicate in a finite form, not withstanding in which form the predicate was used in the text. Other predicate slots are referenced to the particular case-roles whose semantic statuses are explained to the user by the "human" questions next to the predicate slots.

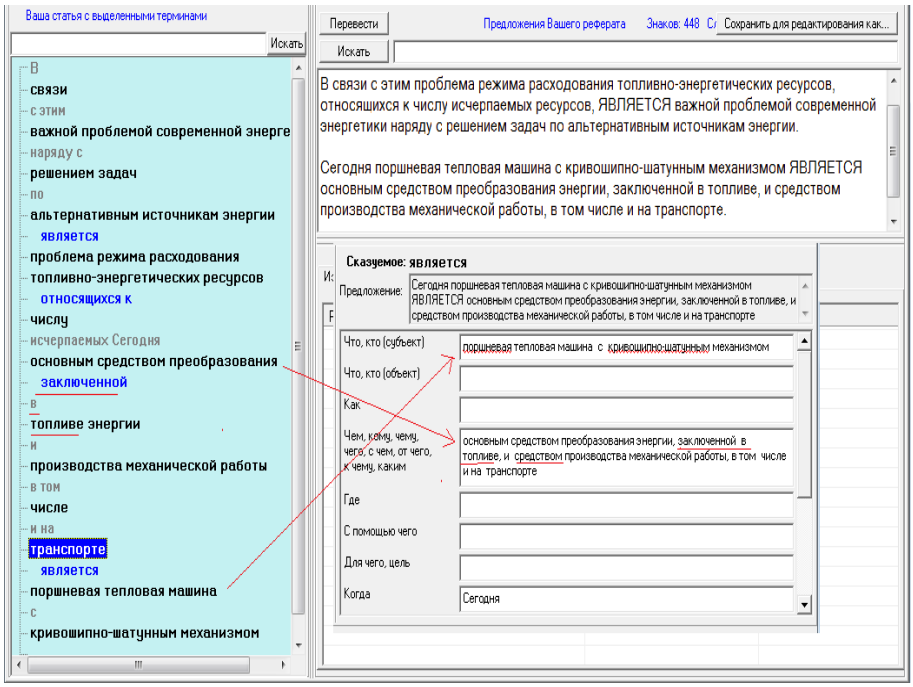

Figure 1. A screenshot of the user interface. The downloaded text with the visualized NP and VP terms is shown on the left pane. In the middle the template for the Russian predicate *является (is)* with the filled slots is shown. On the top of the right pane there are the automatically generated restructured text sentences without readability/ translatability indicators.

The user can either drag-and-drop the appropriate segments from the interactive claim text or simply type the text in the slots. Once the template is filled, the system automatically generates a grammatically correct simple sentence structured so as to avoid readability/translatability indicators. In addition to constraining the complexity of the sentence structure the predicate templates also put certain constraints on the phrase level. As the templates are meant for one-predicate sentences only, coordination of verbal phrases (predicates) that may be ambiguous is avoided. The prepositions or particles attached to the verb are put to the main (predicate) template slot that resolves a possible verb/noun attachment ambiguity. The authoring procedure completed, the content representation built by the analyzer "behind the scenes", the authored text is generated and displayed on the top of the right pane of the interface (see Fig 1). This text can be printed, saved or input in any machine translation system.

6. Conclusions

We presented a methodology and an authoring environment for raising the readability and translatability of professional documentation. The efficiency of the methodology is conditioned by the controlled language framework and interactive computer-human communication. The controlled language data are created based on the domain-specific analysis of the corpus of scientific and technical papers in engineering. The constraints of the controlled language are embedded into the system knowledge base and included into a comprehensive, self-paced training material. The authoring environment is interwoven with the hybrid analysis components and completely automatic generation module. We are going to extrapolate our system to other languages and domains. Another possible way to extend our research is to raise the level of automation.

References

1. Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, *37* (4), 585–599.

2. Aluisio S., Specia L., Gasperin C. and Scarton C. (2010). Readability assessment for text simplification. In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp.1–9.

3. Applied discourse analysis : social and psychological interventions / edited by Carla Willig. Buckingham : Open University Press, viii,166 p. 1999.

4. Bott S., Saggion H. and Figueroa D.. (2012). A Hybrid System for Spanish Text Simplification. NAACL-HLT 2012 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 75–84, Montreal, Canada, June 7–8, 2012. c 2012 Association for Computational Linguistics.

5. Bredenkamp, A., Crysmann, B., and Petrea, M. (2000). Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking. *Proceedings of LREC 2000*. Athens, Greece.

6. Crossley, S. A. & McNamara, D. S. (2008). Assessing Second Language Reading Texts at the Intermediate Level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). Language Teaching, 41 (3), 409–229.

7. Dell'Orletta, F., Montemagni S., and Venturi G. (2011). READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. in Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh, Scotland, UK, 2011, pp. 73-83.

8. GOST 7.9–95. (1995) Standards on information, library indexing and publishing. Moscow.

9. Graesser, A. C., McNamara, D. D., Louwerse, M. L., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36, 193–202.

10. Karpov N., Baranova J., Vitugin F. (2014). Single-sentence Readability Prediction in Russian in Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science Volume 436, 2014, pp 91-100

11. Kincaid, J. P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Research Branch Report 8–75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis.

12. Khomutova, T.N. (2010). *Nauchnyj tekst: integralnyj podhod [Research text: an integral approach]*. Chelyabinsk: SUSU Press/

13. Krioni, N., Nikin, A., Filippova, A. (2008). Automated system for analysis of the complexity of educational texts. Manag. Soc. Econ. Syst. 11, 101–107.

14. Nyberg E., T Mitamura, D. Svoboda, J. Ko, K. Baker, J. Micher (2003). An Integrated system for Source language Checking, Analysis and Terminology management. *Proceedings of Machine Translation Summit IX,* September. New-Orleans.USA

15. Oborneva, I. (2006). Oborneva I: *Automated assessment of the textbooks' readability based on a statistical analysis (In Russian)*. Moskow: Russian Academy of Education.

16. Pouliquen Bruno, Christophe Mazenc Aldo Iorio. (2011). Tapta: A user-driven translation system for patent documents based on domain-aware Statistical Machine. *Proceedings of the EAMT Conference.* Leuven, Belgium, May.

17. Pooneh Heydari and A. Mehdi Riazi (2012) Readability of Texts: Human Evaluation Versus Computer Index. Mediterranean Journal of Social Sciences, Vol. 3 (1) January.

18. Poornima C , Dhanalakshmi V, Anand Kumar M, and Soman K. P. (2011).Rule based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications (0975 – 8887) Volume 25– No.8, July 2011.

19. Sheremetyeva S. (2012), Automatic Extraction of Linguistic Resources in Multiple Languages. In Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, Poland

20. Siddharthan, A. (2002). An Architecture for a Text Simplification System. Proceedings of the Language Engineering Conference (LEC'02), Hyderabad, India, IEEE Computer Society pp. 64.

21. Sheremetyeva S. (2003). Natural language analysis of patent claims. In Proceedings of the ACL 2003 Workshop on Patent Processing, ACL '03, Stroudsburg, PA, USA. Association for Computational Linguistics.