Available online at: http://www.ijmtst.com/vol3issue10.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 03, Issue No: 10, October 2017

Ontology-Based Text Simplification for Dyslexics

Tatyana Ivanova Ivanova

KEE, TU, Sofia, Sofia, Bulgaria

To Cite this Article

Tatyana IvanovaIvanova, "Ontology-based text simplification for dyslexics", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 10, October,2017, pp.-34-47.

ABSTRACT

Ensuring high school students, university students and adults, having learning disabilities, as dyslexia with actual and easy to read learning and scientific content is of great importance in our days. The main problem, discussed in this paper is how to make growing scientific and learning content more accessible for peoples, having dyslexia, or other learning disabilities. We analyze many recent researches, related to text simplification and Controlled Natural Languages and its application for development of textual learning resources for dyslexic learners in several languages. We propose ontology-based methodology for development of easy to use by dyslexics Bulgarian language textual resources using English language text, text simplification and ontology management.

KEYWORDS: *Text Simplification, Ontology Learning, Dyslexia, E-Learning, Controlled Natural Languages*

I. INTRODUCTION

As dyslexia is a disability that affects language, the learning can be made more effective for dyslectics by developing more accessible learning resources. It is needed to propose learning content that will make reading and comprehension more easy and effective for dyslexics. There are three main approaches to make textual learning content more accessible for dyslectics:

- Text presentation approach;
- Ensuring presentation flexibility;
- Text simplification approach.

Text presentation approach aims to ensure dyslexia-friendly learning content by presenting the content in the most appropriate way. This includes readability by Screen Readers [1], usage of appropriate schemes, images, numbered lists, simple and clear navigation, formatting (font size of 16 or more points Sans Serif Fonts or other specialized for dyslexia as freely available fonts on http://opendyslexic.org/get-it-free/). This approach does not include any linguistic changes in the text. Plugins for browsers, formatting automatically text fonts and colours to make the text more readable for dyslexics are available (https://chrome.google.com/webstore/detail/beeline-reader/ifjafammaookpiajfbedmacfldaiamgg?hl=en).

Presentation flexibility means to make presentation customizable, i.e. to propose possibilities for easy change of fonts, colors, etc. Different people with dyslexia have different reading abilities, and different preferences, so textual content should be easily adaptable by users. Usability of mobile content readers is one approach to achieve this need. Text presentation and its flexibility has a significant effect on the reading speed of people with dyslexia. This approach is well studied and good guidelines, recommendations and tools are presented for developing Web sites and other textual resources, accessible to users with dyslexia [2]. We will not discuss this approach in our research.

Text simplification approach is about modifying textual sources to ensure dyslexia-friendly learning materials. This includes usage of short and simple sentences, understandable and non ambiguous terminology, proposing synonyms [3],

[1]E-mail: tiv72@abv.bg, Botevgrad, Preslavstr, 48, Bulgaria

understandable explanations. As all the modifications, needed to achieve dyslexia-friendly learning materials make the text more simple, every such text modification is a variant of text simplification.

Focus of this work is on text simplification and translation approaches (not on presentation) for achieving dyslexia-friendly learning materials. There are very few learning content, especially developed for dyslexics, and development of such content requires specific pedagogical and expert qualification. All the textual content, having scientific value (as research papers, new specialized book editions or news in the web) is difficult to read by dyslexics. Our main aim is to analyze existing text simplification approaches and procedures and find or develop appropriate ones for automated or semi-automatic modification of textual content, to get one, more understandable for dyslectics.

Learning in native language is the most easy and effective, but the most actual scientific content, best books and learning resources are published only in international languages, as English and Russian. So, to automate simplification and translation to learner's native language are the two important tacks to make original English language content more accessible for Bulgarian dyslexics. Text simplification research is language-specific and the most significant results and tools are also related to English language texts.

There is rapidly growing research related to usage of ontologies in automatic translation and text simplification fields (see fig. 1, fig. 2) One of the most important fields of usage of ontologies is e-learning (Google Scholar returns 34 300 results from query e-learning ontology).

In this research we will discuss how multilingualism and ontologies can benefit text simplification and translation to achieve dyslexia-friendly textual content for Bulgarian readers. As graphical and short, well-structuredis easy to read and understand for dyslectics, we will also discuss building, evolution and usage of ontologies, representing the knowledge for supporting learning. In many domains it is important for Bulgarians to use knowledge both in English and Bulgarian (resources in English are usually of higher quality, but these in Bulgarian are easy to use for most Bulgarian peoples). So, we will take attention to bilingual aspects of text simplification and bilingual ontologies.

We will make a short review and classification of textaccording to semantic principles representation of knowledge simplification approaches, and then discuss the possibilities of usage of Controlled Natural languages for text simplification and ontology development.

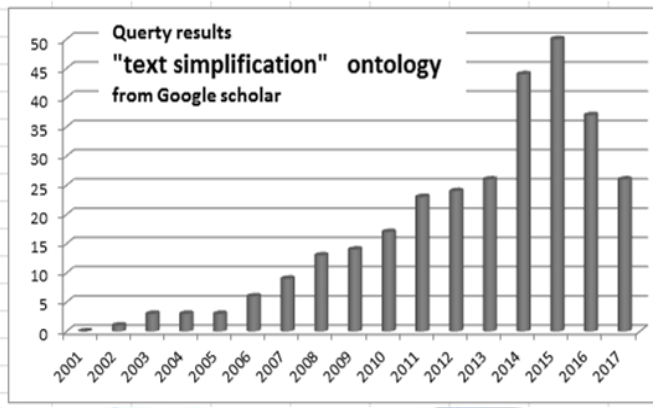

Fig. 1. Google Scolar's papers on text simplification and ontologies

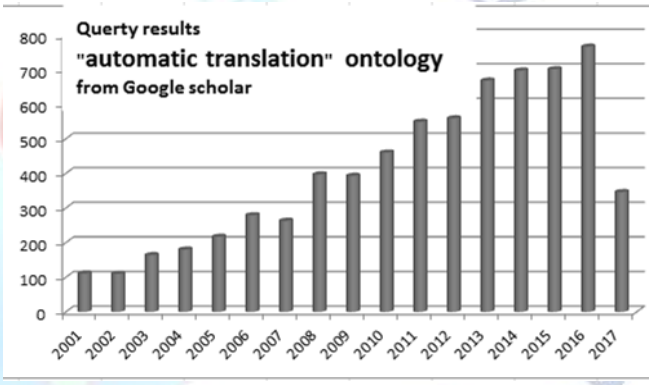

Fig. 2. Google Scolar's papers on translation and ontologies

II. TEXT SIMPLIFICATION PROBLEM

Text simplification is a NLP task that tries to reduce the linguistic complexity of the text while still retaining the original information and meaning [4]. The goal of textsimplification is to rewrite complex text into simpler language that is easier to read and understand [5].

According to the type of changed language components text simplification can be classified into four types: lexical simplification (including changes, related to individual phrases or words), syntactic simplification (including non-semantic based changes in the sentence structure and syntax), discourse-related and semantic simplification.

Lexical simplification is a process of making a sentence more readable or understandable by replacing difficult, rarely used and\or complex words to simpler or frequently used words that

retain the same meaning. Lexical simplification is based on substitution of some words by synonyms. Synonym words can be obtained from thesauruses [6], from dictionary definition, WordNet or ontologies. Common steps of the lexical simplification process are complex word identification, substitution generation, substitution ranking. To do lexical simplification measuring of lexical complexity is needed. Most of the approaches to lexical simplification use word frequency and word length [7] as a measure of lexical complexity. The widely used metrics are variants of Lexical density (the number of content words, divided by the total number of words), lexical variation, and lexical sophistication. Decreasing of lexical variation, and lexical sophistication leads to easily readability, but can worse the scientific and learning quality of the text [8]. Dictionaries, lexicons and ontologies are main sources of synonyms, needed in the simplification process. Finding idiomatic phrases in the text is important in lexical simplification, as its components should not be changed. A simple and effective method for finding phrases in text is presented in [9].

Syntactic simplification is the process of reducing the structural and grammatical complexity of a text, while retaining its information content to make it easier to comprehend for human readers, or to process by programs. Important syntactic simplification operations are:

- Splitting long sentence into several shorter sentences (for example, by disembedding relative clauses, or separation of subordinated clauses);
- Conversion from passive to active voice;
- Decreasing the number of conjunctions and pronouns;
- Dropping (removing of unimportant parts of a sentence to make it more concise);
- Reordering of syntactic units (e. g. unclear sentences);
- Substitution of difficult phrases with their simpler synonyms;

Discourse-related text simplification is based on the fact that cohesive texts are easier to follow. It aims to make content more transparent by making discourse relations explicit. This approach makes knowledge more accessible for peoples by restructuring it in chronological order or cause/effect ordering of sentences or restructuring text to simplify argumentation. Most of these simplifications can be made only manually, by professionals or experts. This simplification is the most useful for readers with low levels of domain expertise, children,

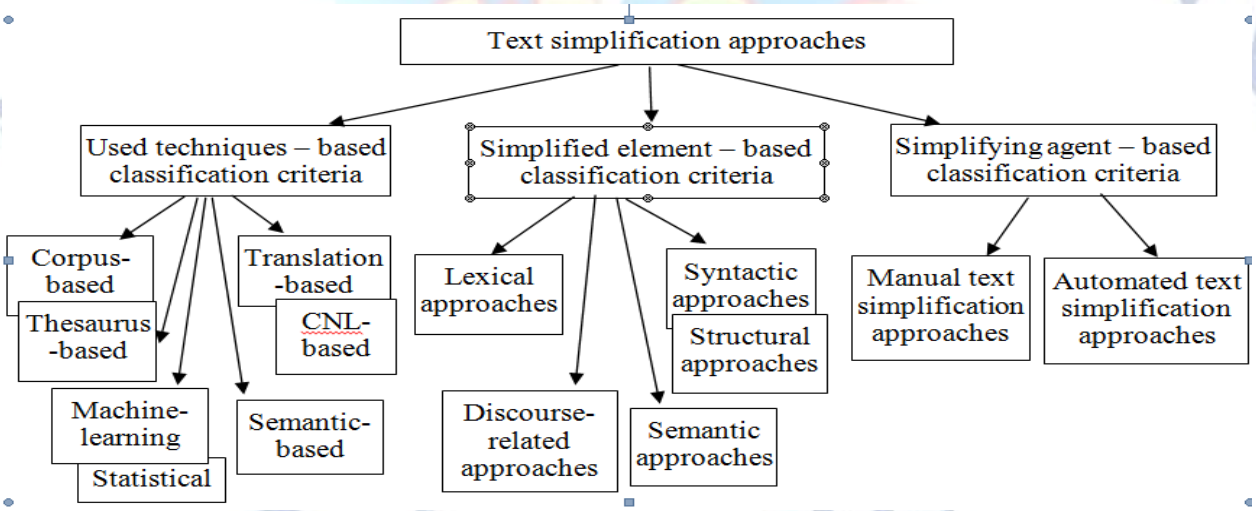

Fig. 3. Classification of Text Simplification approaches

foreign readers and readers, having learning disabilities, as dyslexia, aphasia.

Semantic simplification aims to make text more understandableby

- Background knowledge-related semantics–based conceptual simplification;
- Scientific quality-related (remove ambiguities, inaccuracies);
- Make the text more engaging; Personal narratives; humor;
- Explanations generation (in electronic texts) for difficult or new terms that have no preferred synonyms.
- Use analogy.

According to the usage of text simplification tools the simplification process can be classified as manual or automatic. Automatic text simplification has only recently become an established research

field.

According to the used simplification methods and technologies Text simplification can be classified as corpus-based, Translation-based, Semantic-based, CNL-based, Machine learning-based, Thesaurus or statistics-based (see fig. 3).

Text simplification can be seen as monolingual translation and machine translation approaches can be used in the simplification process [10]. Corpus-based translation approach (using corpora of simplified texts [11], machine learning [12] and statistical approaches) and rule-based approach, using hand-written or automatically-generated rules, and dictionaries, are the main text simplification approaches. Rules for paraphrase and lexical simplification can be acquired automatically using statistical methods, aligned corpora, or by derivation of paraphrases through existing lexical resources [13]. There is also some resent research on machine learning text simplification, based on Neural Networks. Neural network – based machine translation model is applicable for automatic extraction of rules for modification of sentence structure, substituting words, and removing words for text simplification [14].

Text simplification system has to be able to make disambiguation decisions during lexical simplification. WordNet is one of frequently used sources of synonyms for lexical simplification that also propose information for word sense disambiguation. [15] Reports for significant improvement of the quality of simplified texts, using WordNet. Semantic based approaches for text simplification are also presented. Deep Semantics and machine translation approach in syntactic simplification process are used successfully in [16].

It is clear that text simplification tack is very complex, language and target users dependent and to achieve good results, combination of manual and automated approaches, as well as different simplification level approaches should be combined.

According to [17] evaluations of the quality and correctness of results from automated text simplification have been performed on a small scale (as few as 20 sentences), and by small number of target users (in many cases 3-10 users).

As simplification of textual content is closely related to controlled natural languages, it is important to know whether controlled natural languages can be used for development of learning resources, suitable for dyslexics. So, we will continue with short analysis and classification of Controlled natural languages.

III. CONTROLLED NATURAL LANGUAGES

Controlled natural languages (CNL) are subsets of natural languages that are obtained from natural languages by restricting the grammar, vocabulary and/or style in order to reduce ambiguity and complexity. There is no generally agreed-upon definition for controlled natural language [18]. The main aim of CNLs is to make texts easily readable and clearer by reducing natural language ambiguity and complexity.

There are two main types CNLs according to its target users: languages mainly designed for good processability by computers and languages mainly designed for good understandability by human readers.

Controlled natural languages mainly designed for good processability by computers are useful mainly in ontology building and machine translation. They are used to improve text translatability to another natural language, or support automated ontological representation of knowledge.

In ontology development process controlled natural languages can support:

- Ontology learning;
- Domain experts to develop or modify ontologies;
- Documentation of developed ontologies.

Some CNLs designed for machine translation or ontology development are completely unambiguous and have a direct mapping to formal logic. So, most of CNLs, designed for good processability by computers, are defned by formal grammars. On the other hand, formal semantics of the controlled languages for humans are rarely discussed and evaluated.

The main purpose of CNLs for humans is to make the information in the text more accessible and understandable for humans, having specific learning needs. CNLs for humans are developed for supporting learning, automating documentation creation, speaking by groups of peoples, having language difficulties. Examples are the language for speaking with children, Basic English, Special English, Simplified Technical English etc.[18].

Basic English for example is an English-based controlled language (a simplified subset of regular English) created as an international auxiliary language for teaching English. *Special English* is a controlled version of the English language used for

presenting daily news by the United States broadcasting service Voice of America (VOA). Both Basic English and Special English are used in development of Simple English Wikipedia (https://en.wikipedia.org/wiki/Simple_English_ Wikipedia). Use of controlled English significantly improves text comprehension, with a particularly large effect for complex texts and non-native speakers [18]. Another language *CLCM* has been developed to have a positive effect on reading comprehension for most groups of readers under certain circumstances such as stress situations [19].

CNLs, designed for simplifying automated text processing include two main groups of languages: for automated translation and for ontology development and management.

CNLs, designed for automated translation usually are domain-specific. Most of these languages are also important for improving readability for human readers. Examples are Standard Language (SLANG), developed at Ford Motor Company, ASD Simplified Technical English, Air Traffic Control Phraseology [18]. EasyEnglish is a part of IBMs internal editing environment, used as a pre-processing step for machine-translating IBM manuals. EasyEnglish aimed for making the text easy to understand.

CNLs, designed for ontology management have a formal logical basis (formal syntax and semantics), and can be mapped to some logic-based formal language (description logic, first-order logic) [20]. At the same time, they look as natural languages and are easy for humans to write and understand. CNL is a way to bridge the gap between a natural language and a formal language. These CNLs can serve as high-level interface languages to knowledge systems. There are two types formal CNLs: general-purpose languages (they have not been developed for a specific scenario or application domain), and CNLs for the Semantic Web [21], as ACE View subset, Rabbit [22], OWL Simplified English [23] and Sydney OWL, business oriented as SBVR Structured English. Examples of general-purpose CNLs are Attempto Controlled English, Processable English (PENG), Computer Processable Language (CPL). These languages differ from human-oriented CNLs in the need to have the sufficient expressive power to represent real domain logics. Logical expressiveness sometimes leads to usage of constructions, difficult for understanding by humans.

ACE for example is a sequence of declarative sentences that can be interrelated recursively in composite sentences using coordination, subordination, quantification, and negation [24]. To constrain the ambiguity of full natural language ACE uses unambiguous alternatives. Interpretation rules are used to reformulate ambiguous constructs. Users can either accept the proposed interpretation, or they must rephrase the input to obtain another one. The vocabulary of ACE includes predefined function words (e.g. conjunctions, determiners), content words, predefined phrases (e.g. "it is true that ...", "it is visible that ..."). In ACE quantifications are used to speak about all objects, or to denote explicitly the existence of at least one object, as in description logics.

Some of CNLs can be used both by achieving humans readability and computer processability. Every CNL designed to improve human-computer communication can also be used for communication between humans; and controlling a language for enabling a better communication between humans also improves the computer processability [18]. Some results of our analysis of CNLs can be seen in table 1, table 2, table 3, table 4.

IV. Text simplification rules and controlled languages

Controlled languages can be used in generation of documents, written on simplified languages. As it is discussed above, controlled languages can be classified as Human-Oriented Controlled Languages, and Machine Oriented Controlled Languages. According to the way of generating simplified documents, controlled languages can be classified as manual and automatic. Practically only a few simplification tacks can be fully automated. Achieving maximal automation level is important for efficiency of the simplification process.

Rules are widely used to transform natural language documents into documents, written in controlled language or to generate controlled language documents. O'Brien classified rules for controlled languages into four types: lexical, syntactic, textual structure-related and pragmatic [25].

Lexical rules are about the use of particular acronyms, synonyms, conjunctions, double negations, ambiguous anaphoric reference etc. Rules for standardization of number formats, date formats and for dealing with ambiguous words also are important. These rules can be automated by

using dictionaries, thesauruses, or ontologies.

Syntactic rules are about the use of specific prepositions; specify location of prepositions to reduce ambiguity, about usage of present participle, forms of conjunction, punctuation. Some of them can be also automated.

Textual structure-related rules are about usage of lists, tables, constrain maximum sentence and paragraph lengths, specify keywords to use for coherence. Pragmatic rules are about the use of metaphor, slang or idiom. Most of these rules are not standardized and can be applied only manually.

Specific set of rules are used to translate text from English to controlled languages. Every CNL has its own set of rules.

Basic Simplified Technical English rules for example are:

- Restrict the length of noun clusters to no more than 3 words;
- Restrict sentence length to no more than 20 words (procedural sentences) or 25 words (descriptive sentences);
- Restrict paragraphs to no more than 6 sentences (in descriptive text);
- Avoid slang and jargon while allowing for specific terminology;
- Make instructions as specific as possible;
- Use articles such as "a/an" and "the" wherever possible;
- Use simple verb tenses (past, present, and future);
- Use active voice;
- Do not use present participles or gerunds (unless part of a Technical Name);
- Write sequential steps as separate sentences;
- Put commands first in warnings and cautions, with the exception of conditions.

This language can be used to improve comprehension for people whose first language is not English, and also for making human translation easier, faster and more cost effective, and facilitate computer-assisted translation and machine translation.

Drawbacks of CNLs for humans are related to the fact, that some of rules can be performed only manually, and there are not effective rules for removing natural language ambiguities. Usage of some CNLs for some type users can lead to loose of some of scientific quality of the text (for example, text for children will not be understandable, if it contains original scientific terminology, but substitution of difficult terms decrease scientific value of the text.

Main aim of controlled languages for machine use is not to make texts simpler, but to reduce the potential for misunderstanding by controlling ambiguity. So, these languages allow long sentences and complex logical dependencies for example that makes text difficult for understanding by poor readers. They also can use infrequent terms, long terms and unpopular abbreviations, as all these terms are important in some cases and for some domains and they can be added to dictionary, thesaurus or ontology.

For our research is very important to find text simplification approaches and rules, useful in text simplification for people with disabilities, as well as rules, making text more suitable for automated ontology development. As both textual resources and ontologies are needed, both type CNLs may be useful. To find the best strategy of some CNLs rules or usage of CNLs in the development of learning resources for dyslexics, we made generalized comparative analysis of the three type CNLs.

What CNLs will be useful in the development of learning resources for dyslexics? There are only a few researches in this area. The query ["controlled natural language" dyslexia], sent to Google Scholar returns only 5 results since 2013 and no one result since 2016! On the other hand research paper number, related to text simplification for dyslectic readers has significant grown for the last few years (see fig3). So, text simplification for dyslexic readers is very yang research area having increasing importance, and specific CLN for this simplification is not proposed yet.

V. Text simplification for dyslexic readers

Different target groups of users have different simplification needs, so text specific simplification should be made for every target users group. Not all text simplification techniques lead to the more accessible text for dyslectics. Frequent usage of abbreviations for example makes text shorter, but difficult for dyslectics, if they are not familiar with abbreviations.

Dyslexics typically encounter problems when reading infrequent words, long words [26], and difficulties in comprehension of homophonic words, orthographically similar words long compound sentences [27]. Usage of more frequent words helps the learners with dyslexia to read faster [2], and proposed terminology synonyms are useful in deep comprehension of the reading text. So, the rules for textual content simplification for

dyslexics should ensure these requirements.

Lexical simplification operations for dyslectic readers aim to decrease usage of Irregular words, homophonic words or pseudo homophonic words (weather and whether for example), foreign words. Orthographically similar words, including letter reversals (trail for trial), words, having differences only in one or two letters (as addition and audition, similar than the words toffee and coffee,) in many cases is better to be substituted by some synonyms. Dyslectic readers should take special attention to new words, so only most important such words should be used in the learning material. Dyslectic readings also should use rarely pseudo–words, and fantabulous words. Words, started with "non"(so-called non–words), less frequent words, long words, are also difficult for dyslectics. Confusions of small words (for example in by is) also are source of reading difficulties. Decreasing of the usage of pronouns and its replacement with names can make text more understandable for dyslectic readers.

Numbers represented as digits instead of words, as well as percentages instead of fractions, improve readability of people with dyslexia [28]

Syntactic simplification. Long and compound sentences should be reformulated or braked into short, simple and clear sentences in texts for dyslexics. Text length should be reduced by reformulation and by deleting peripheral information. Reduction of the usage of passive voice make text more understandable for dyslectic readers.

Semantic rules for dyslexia-friendly texts should reduce ambiguity and difficulty in reading, make all important knowledge explicit and clear. Conceptual simplification, including short and clear definitions of important concepts, is highly recommended. Numerical simplification, including writing numbers by digits (not to use words) is also useful.

Lists, tables and schemas in textual structure and short sentences should be used in texts for dyslexics. Semantic simplification is the most-difficult. Full automation of them is impossible. Systems for semiautomatic text simplification should detect ambiguity, ungrammaticality and complicated constructs and help an author to revise the text manually.

Graphical schemes improve the subjective readability and comprehensibility of people with dyslexia[28].

Lexical simplification via automatic substitution of complex words readability and comprehensibility of people with dyslexia [28] by simpler synonyms is less helpful than

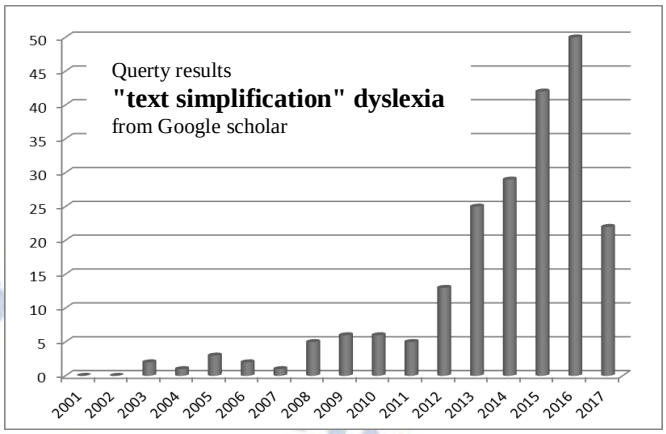

Fig. 4. Google Scolar's papers on text simplification and dyslexia

showing synonyms on demand for improving the subjective

Reading comprehension can be improved for dyslexic or poor reader by usage of rules for:

- Splitting long complex sentences by transforming them into ones that are more easily;
- Substituting difficult words or rarely –used words;
- Showing synonyms on demand;
- Avoiding pre-posed adverbial clauses;
- Presenting information in cause-effect order;
- Making discourse relations explicit;
- Use analogy.

Information in texts for dyslectics is more explicit and clearly represented, texts follow more strict and clear grammatical and syntactic structure. So, we think such texts will be useful in automated ontology development. The developed ontologies may be useful in learning.

We will also discuss details of how such simplified texts are useful for automatized terminology extraction and ontology learning. Important questions are what of the text changes for making texts more accessible for dyslexics make the text more appropriate for ontology learning or for automated translation.

By performing text simplification, using rule types, listed above, we usually obtain textual resources, written in some variant controlled natural language. We will briefly survey and compare well-studied controlled natural languages (for humans, ontology building and for machine translation). Our aim is to find the language, or group of rules in controlled languages, useful in the process of text simplification and ontology building

in the learning field for dyslectics.

Another important aspect of text simplification is dependence with the specific natural language. For example, translation

toffee in Bulgarian is конфекция and translation of coffee is кафе, that words are not similar. Frequency of usage of translated words also is different. This leads to two main conclusions:

- Many of discussed rules seems language – independent, its applications are language-specific;
- Parallel usage of terms of two or more languages (for example in brackets) can improve understandability and remove some ambiguities.

So, usage of English-language terms as additional words (for example in brackets) can make text less ambiguous and clearer for Bulgarian readers that can use English.

Our results from comprehensive analysis and comparison of the main simplification rules, used in CNLs for easy reading CNLs for machine translation, and CNLs for ontology development are summarized in the following tables (table 1, table 2, table 3, and table 4). We also compare these rules to the text simplification needs for dyslexics.

Table 1. Lexical text simplification rules and CNLs

rule	CL for easy reading	CL for machine translation	CL for ontology development	Text simplification needs for dyslexics
For verbs	Use the most frequentor shortest verbs and adjectives.	Use only the forms of verbs and adjectives shown in the dictionary.	Use only the forms of verbs and adjectives shown in the Dictionary.	Decrease usage of irregular words, homophonic words or pseudo homophonic words.
	Prefer simple tenses and active voice.	Prefer simple tenses and active voice.	Prefer simple tenses and active voice	Prefer simple tenses and active voice.
For nouns	Use the shortest and simplest or frequently used names.	Use an approved Dictionary words only.	Use an approved dictionary words and synonyms.	Use shorter and most popular terms.Replace technical jargon with short and clear definitions
	Use a technical name only as a noun or as an adjective.	Use a technical name only as a noun or adjective.	Clear relationship between technical namesand all other, related to themwords.	Use a Technical Name only as a noun or as an adjective. Provide explanation if needed.
	Use the official technical names as much as possible.	Use the official technical names as much as possible.	Use both the official techni-cal names and synonyms.	Use the mostpopular and short terms as much as possible.
	Rule out the use of non so popular acronyms, synonyms, use relative pronouns.	Restricted vocabulary. Rule out synonyms, that are not in thesauruses	Use both the names and acronyms and show clearly relation between the two.	Rule out the use of acronyms, and synonyms, use easiest and shortest words only.
For ad-jectives	use only important, popular and shortest ones	Use an approved Dictionary words only	Use synonyms and show clearly relation between the two	Decrease number of adjectives, use only important, popular and shortest ones.
For clear struc-ture	Do not use different Names for the same thing.	Different Names should clearly specified.	Different Names should clearly specified.	Do not use different names for the same thing.
	Rule out double negations.	Rule out double negations.	Double negations can be used.	Rule out use of double negations.
	Rule out ambiguous	Rule out ambiguous	Rule out ambiguous	Rule out ambiguous

	words, conjunctions, disjunctions andambiguous anaphoric reference.	words, conjunctions, disjunctions and ambiguous anaphoric reference.	words, and ambiguous anaphoric reference.	words, conjunctions, disjunctions and ambiguous anaphoric reference.

Table 2. Grammatik text simplification rules

rule	CL for easy reading	CL for machine translation	CL for ontology development	Text requirements for Dyslexics
senten ce-rela ted	Restrict use of parenthe-ticals, Use uniform sentence structures.	Use correct punctuation and clear sentence structures.	Use correct punctuation and clear sentence structures.	Restrict use of parenthe-ticals, Use simple and uniform sentence structures..
	Avoid too many subjects in one sentence			Avoid too many subjects in one sentence
Verb-r elated	Do not use the past participle with the verb have	Avoid use of present participle, past participle	Avoid use of present participle, past participle	Avoid use of present participle, past participle
	Avoid complicated past, future, conditional tenses.	Avoid complicated past, future, conditional tenses.	Avoid complicated past, future, conditional tenses.	Avoid complicated past, future, conditional tenses.
	Use an approved verb to describe an action, (not a noun or other part of speech).	Use an approved verb to describe an action, (not a noun or other part of speech).	Use an approved verb to describe an action, (not a noun or other part of speech).	Use an approved verb to describe an action, (not a noun or other part of speech).
	Use the active voice in descriptive writing	Rule out passive voice, use indicative mood	Rule out passive voice, use indicative mood	Use the active voice in descriptive writing
Com -mon	Restricts the grammar	Use clear and correct grammatical elements	simplicity - easily learned and applied by peoples;	
	Restrict functional words	Restrict functional words	Restrict functional words	Restrict functional words
	Avoiding gerund,	avoiding gerund,	avoiding gerund,	avoiding gerund

Table 3. Syntactic text simplification rules

CL for easy reading	CL for machine translation	CL for ontology development	Requirements for Dyslexics
Use clearnarrative sentence structure	Syntax should be well-defned, and efficiently parsed for programs.	Syntax should be well-defned, and efficiently parsed for programs.	Use short sentences
Reduction in the number of embedded clauses, conjunctions, restrict size of noun cluster	Insist on the use of article, clear structure of embedded clauses. Do not use semicolons.	Use of logically-clear embedded clauses, negation conjunctions disjunctions, quantifications.	Rare use of embedded clauses and disjunctions, conjunctions, restrict size of noun cluster.
Standardize format for numbers and dates	Standardize format for numbers and	Standardize format for numbers and dates	Standardize format for numbers and dates

	dates		
After you choose the words to describe something, continue to use the same words.	Restrict apposition,	Sentences in the CNL be understandable by people.	After you choose the words to describe something, continue to use the same words.
relative pronouns such as “who”, “which” or “ that” should not be omitted	relative pronouns such as “who”, “which” or “ that” should not be omitted	relative pronouns such as “who”, “which” or “ that” should not be omitted	relative pronouns such as “who”, “which” or “ that” should not be omitted

Table 4. Pragmatic and semantics-based rules

CL for easy reading	**CL for machine translation**	**CL for ontology development**	**Text requirements for Dyslexics**
Rule out the use of metaphor, slang or idiom.	Rule out the use of metaphor, slang or use only standard idiom.	Rule out the use of metaphor, slang or use only standard idiom.	Rule out the use of metaphor, slang or use only popular idiom
Clear describtion of the problem domain	Standardized describtion of the problem domain	Expressivity-covering the desired problem domain	Short and clear describtion of the problem domain
Urge author to be as specific as possible.	Urge author to be as specific as possible.	Urge author to be as specific as possible.	Balanced specificity.
Well-structured presentation.	Clear and standardized presentation structure.	Well-defned semantics.	Cause-effect or chronologic short presentation.

Apart from textual resources, ontologies also are useful in the learning, as they can improve reading comprehension through making discourse relations explicit. This can be done by graphical representation (as tree-like organization of terminology dependences) or as automated generation of short explanations in the text. Ontologies also can support text simplification.

As a conclusion from the comparative analysis, we can say that there are many common rules for all CNLs and dyslectic needs, but there are also significant differences. For example, comprehensive semantic elements (as negation, conjunction, disjunction) are important for ontology – related CNLs, but should be avoided in the texts for dyslectics. Rich dictionary is also needed for development of ontologies, but not for easily reading by humans. So, if we perform terminology simplification for dyslectics, resulting text will be clear for automated analysis, but will not be sufficiently reach for ontology development. So we should perform different text simplification procedures for easy reading by different groups of learners and for ontology development

VI. Text simplification strategies and a methodology for development of textual content, accessible for Bulgarian dyslexics

Most text simplification procedures can be seen as some kind of (monolingual) translation. Simplification for dyslexics is very specific, natural language-dependent, complex and not well studied yang research area. Bulgarian language corpora and linguistic sources are also very scarce. Good learning materials and newest scientific sources are usually written in English and only a few can be found in Bulgarian. Most research, related to text simplification is about simplifying English language text. Our idea is to combine monolingual with bilingual translation (by using dictionaries thesauruses or ontologies and available textual corpuses) both in the text simplification and ontology development process to produce translated and\or simplified textual resources and bilingual (English-Bulgarian) ontologies, containing English and Bulgarian language technical or scientific terminology. Machine bilingual translation also can benefit from bilingual ontologies in the translation of (simplified) textual learning resources from English to Bulgarian.

As there are very scarce research in simplification of Bulgarian language text (we have

found only one publication [30]), we have analyzed results, related to text simplification in several other languages to find general principles and well working approaches, usable in simplification and applicable for Bulgarian.

Simplification systems and simplification studies have performed for Brazilian [31], Portuguese [32], Japanese [33], Japanese [13], French [34], Italian [35], Spanish [36] and Korean [37], Bulgarian for deaf peoples [30]. Text modification is a highly language dependent task, but there are some general language-independent principles [30].

After analysis of the results of above mentioned and some other research we came to the following conclusions about general text simplification rules for easily readable by dyslexics texts (That are also applicable and useful for simplifying texts in Bulgarian):

- Removing redundant words or phrases makes text more shorter and more clear
- It is important to use WSD in the process of lexical simplification
- Choosing the most frequent synonym should not be the only criteria. (It may be the most polisemous, or can contain more confusing letters for example)
- Simple and short sentences, having clear grammatical structure should be used in simplified texts
- Active voice should be preferred
- Substitution of the important scientific terms with its popular synonyms should not be made. Instead, short explanation or synonym should be proposed on demand.
- Not so important adjectives should be omitted
- Discourse relations should be clear and explicit

Implementation of these general rules will lead to grand number of concrete for Bulgarian language simplification rules. Most of them at this time can be written only manually because of the absence of simplified textual corpora for dyslexics in Bulgarian, as well as very restricted number of domain-specific Bulgarian Thesauruses and ontologies. Every proposed set of such rules should be implemented in some text simplification tool and in such a way – evaluated by dyslexic users.Some of the approaches for text simplification in discussed above languages are implemented in tools, having some automation level. Example of such tool that use language other than English is presented in [38]. This tool is a web browser plug-in CASSA that helps in reading Spanish text on the Web by proposing useful suggestions for people with dyslexia. It processes a selected web text and shows synonyms and definitions on-demand. This tool also uses a list of difficult words to determine if the word is difficult for dyslexics.

A set of nine clause splitting rules and a set of five finite verb phrase extraction rules (formed by an auxiliary and a full verb) for Bulgarian are presented in [30]. Empty categories in the text (as null pronouns, traces of extracted syntactic constituents, empty relative pronouns, empty subject) sometimes make comprehension of text in Bulgarian difficult for dyslexics. Rules for subject recovery have been defined in [30].

As a whole, results from testing only of small number of simplification rules for Bulgarian text for dyslexics can be found in the literature. Freely available Bulgarian texts, adapted for dyslexics that can be used as a corpus for automated extraction of simplification rules are mainly for children, or only in few domains [39].

The application of grand number of varioussimplification rules can't guarantee the quality of the simplification process. [40] did not found clearand categorical correlationbetween word frequency and preference. We think, this means that to be useful, text simplification should be intended for specific category of users (as children, poor readers, dyslexics). A word is familiar or not for some peoples, and it is frequent in some textual corpuses. More common familiar word has little or no benefit to a reader over a slightly less common, but still familiar one if for example the more familiar word is related to slightly different context, or is more polisemous. This makes automation of lexical simplification more difficult and shows the importance of involvement of target users in the simplification process. So, to produce good and actual simplified textual learning sources, both automated and manual (expert driven) approaches should be used. And the simplification and translation processes should be continuous, to ensure good and actual from scientific point of view content.

So, we will propose a general methodology for simplification of English language text, bilingual ontology learning or evolution and its usage for generation of textual representation in Bulgarian.

- The main steps are (Fig. 5 on page 9):Simplify English language textual reading resource content, using existing corpora, lexical and syntactic rules (including ontology-based ones). Save removed domain terms (or its

synonyms);

- Evaluate how simplified text is good for (semi) automated translation and intelligent processing (for ontology development). Detect and remove ambiguities manually, if needed;
- (Semi) automatic ontology or concept map extraction from simplified text, original text and saved terms. Synonyms and other type related terms should be included in developed ontologies;
- Translate English language terminology, presented in ontology to Bulgarian, making bilingual ontology;
- Use this ontology in the learning process (for example to visualize interconcept relations);
- Use this ontology in resource-development process (for example to generate short textual learning content in Bulgarian, or to generate tests);
- Perform some text simplification on the generated textual resources in Bulgarian.

We will show, that the text, simplified for learning is appropriate for ontology learning, if all the terminology, removed during simplification is also used. For recognizing some of these terms, original text can be used.

How to perform text simplification? First standard tacks as lemmatization, parsing, part of speech tagging, named entity recognition, and disambiguation are performed.

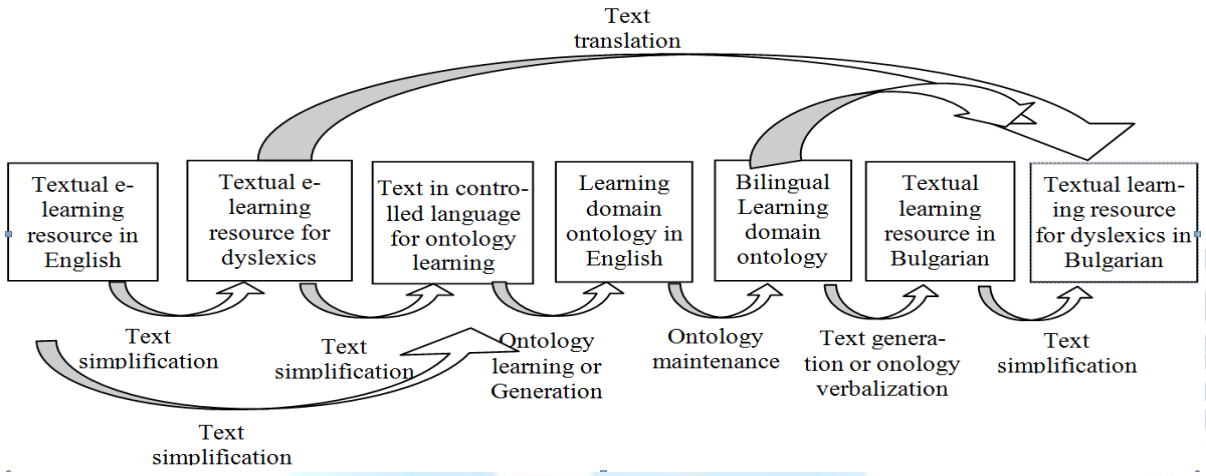

Fig. 5. Our methodology for development of textual content, accessible for Bulgarian dyslexics

Many researches rely on Wikipedia and Simple English Wikipedia, a subset of Wikipedia using simplified grammar and terminology, to learn simplifications. Simplification techniques are not yet of high enough quality for fully automated scenarios. In [3] two different strategies are evaluated: one that automatically substitutes each complex word by a simpler one and another one that allows the user to see and choose between several synonyms. Results show that semi-automatic strategy leads to better understanding of the learning content, and produces higher quality learning content for dyslexics. Involving learners in the simplification process will be the most useful in simplification of Bulgarian language resources, or in the automated translation from English to Bulgarian. In such way we can compensate insufficiencies of corpora or linguistic Bulgarian language resources. Ontology maintenance includes addition of all extracted English or Bulgarian language domain terms during translation or simplification process in the bilingual future translation or simplification tacks

Lexical simplification is more important for dyslexics than syntactic, but as more frequent and shorter words are usually more polysemous, this simplification can add ambiguities, that can worse the text quality in the context of ontology learning. Parallel usage of original, simplified text and stored simplifications will solve this problem. Only the most popular abbreviations (as HTML) should be used in the text, and others should be substituted with synonyms. and English. These terms will be used in future translation or simplification tacks.

Lexical simplification is more important for dyslexics than syntactic, but as more frequent and shorter words are usually more polysemous, this simplification can add ambiguities, that can worse the text quality in the context of ontology learning. Parallel usage of original, simplified text and stored simplifications will solve this problem. Only the most popular abbreviations (as HTML) should be used in the text, and others should be substituted with synonyms. In the ontology both abbreviations

Fig. 5. Our methodology for development of textual content, accessible for Bulgarian dyslexics

stored, as the ontology

can be used also during the analysis of original (non-simplified) text written both in Bulgarian and english .

Technical terms, close to English language terminology should be used if they are short or popular (as algorithm, program, online) and long or non popular English language technical terms (as implementation, translation) should be substituted by synonyms in the simplified text in Bulgarian. In the ontology every used term and its synonyms both in Bulgarian and English languages should be stored

The ontology is useful in the process of lexical and semantic simplification. A set of Bulgarian language structural and syntactic rules should be used for structural and grammatical simplification. Learners also can be involved in the generation and approval of some of these rules.

VII. Conclusion

Our analysis of the research on text simplification and proposed classification of the used text simplification approaches will clarify achieved results and problems in this area and its possible usage for making textual content more accessible for dyslectics. Research results in this area are language-specific and have achieved mainly for English text and partially for several other languages.

We discuss the possibilities of application of the main groups of rules for text simplification, translation, and ontology building, tested for English language in the simplification of text, written in Bulgarian. We also propose general methodology for ontology-based development of learning recourses and other actual textual content for adult Bulgarian dyslexics and students with dyslexia, using English language textual sources, automated translation and simplification.

References

[1] L. Rello, R. Baeza-Yates, "How to present more readable text for people with dyslexia", Universal Access in the Information Society, 16(1), 2017, pp. 29-49.

[2] Rello, L., R. Baeza-Yates, L. Dempere, H. Saggion, "Frequent words improve readability and short words improve understandability for people with dyslexia", Proceedings of INTERACT, Vol. 13, 2013.

[3] L. Rello, et al., "Simplify or help?: text simplification strategies for people with dyslexia" In Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (p. 15). ACM, 2013.

[4] E. Barbu et al, "Language technologies applied to document simplification for helping autistic people". Expert Systems with Applications, 42(12), 2015, pp. 5076-5086.

[5] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in current text simplification research: New data can help," Transactions of the Association for Computational Linguistics, 2015, pp. 283-297.

[6] J. De Belder and M. Francine Moens, "Text simplification for children," In Proceedings of the SIGIR Workshop on Accessible Search Systems, 2010, pages 19

[7] S. Devlin and G. Unthank, "Helping aphasic people process online information," In Proc. ASSETS, 2006, pp. 225–226.

[8] S. Bautista, et al., "Empirical identification of text simplification strategies for reading-impaired people." In European Conference for the Advancement of Assistive Technology, the Netherlands, 2011.

[9] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," In Advances in neural information processing systems, 2013 , pp. 3111-3119.

[10] S. Wubben, et al., "Sentence simplification by monolin-gual machine translation," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012

[11] W. Coster and D. Kauchak, " Simple English Wikipedia: a new text simplification task," In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011, pp. 665-669.

[12] C. Horn, C. Manduca, and D. Kauchak, " Learning a Lexical Simplifier Using Wikipedia" In ACL (2) , 2014 , pp. 458-463

[13] K. Inui, et al., " Text simplification for reading assistance: a project note." In Proceedings of the second international workshop on Paraphrasing-Volume 16, 2003, pp. 9-16

[14] T. Wang, et al.,"An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification.", arXivpreprint , 2016.

[15] N. Yakovets, and A. Agrawal, "SimpLe: Lexical Simplification using Word Sense Disambiguation.", 2013

[16] S. Narayan, and C. Gardent, "Hybrid simplification using deep semantics and machine translation.", In the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 435-445.

[17] A. Siddharthan, A. A. Mandya, "Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules.", In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014).

[18] T. Kuhn, “A Survey and Classification of Controlled Natural Languages “, Computational Linguistics, Volume 40 , Issue 1, 2014, pp.121-170.
[19] I. Temnikova, “Text Complexity and Text Simplification in the Crisis Management Domain.” Ph.D. thesis, University of Wolverhampton, 2012.
[20] T. Kuhn, "A survey and classification of controlled natural languages." Computational Linguistics , 2014, pp. 121-170.
[21] R. Schwitter, "Controlled natural languages for knowledge representation." Proceedings of the 23rd International Conference on Computational Linguistics, 2010.
[22] G. Hart , M. Johnson, C. Dolbear, ”Rabbit: Developing a control natural language for authoring ontologies.” In: ESWC 2008.
[23] R. Power, “OWL Simplified English: a finite-state language for ontology editing.” In International Workshop on Controlled Natural Language, 2012, pp. 44-60, Springer Berlin Heidelberg.
[24] K. Kaljurand, “ACE View---an Ontology and Rule Editor based on Attempto Controlled English.” In OWLED, 2008.
[25] S. O’Brien, “Controlling controlled English. An analysis of several controlled language rule sets.” Proceedings of EAMT-CLAW, 2003, pp. 105–114.
[26] Rello, L., et. al. “DysWebxia 2.0!: More accessible text for people with dyslexia” Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, 2013
[27] A. Siddharthan, “A survey of research on text simplification.”,ITL-International Journal of Applied Linguistics, 2014, pp. 259-298.
[28] L. Rello, “DysWebxia: a text accessibility model for people with dyslexia.”, 2014
[29] M. Shardlow, ”A survey of automated text simplification.” International Journal of Advanced Computer Science and Applications, 2014, pp. 58-70.
[30] S. Lozanova, et al., “Text Modification for Bulgarian Sign Language Users.”, Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, 2013.
[31] S. M. Alu´ısio, “Towards brazilianportuguese automatic text simplification systems.”, Proceedings of the eighth ACM symposium on Document engineering, 2008.
[32] C. Scarton, et al., “ SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments.”, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
[33] M. Hading, Y. Matsumoto, and M. Sakamoto, ” Japanese Lexical Simplification for Non-Native Speakers.”, NLPTEA 2016.
[34] L. Brouwers, et. al., “Syntactic Sentence Simplification for French.” Proceedings of the ThirdWorkshop on Predicting and Improving Text Readability for target reader populations, 2014.
[35] G. Barlacchi, S. Tonelli, “ ERNESTA: A Sentence Simplification Tool 35 for Childrens Stories in Italian.”, Computational Linguistics and Intelligent Text Processing, 2013, pp. 476–487. Springer.
[36] S. Bott, , H. Saggion, , and S. Mille, “Text simplification tools for Spanish.”, LREC, 2012
[37] J.-W. Chung, e al., ”Enhancing readability of web documents by text augmentation for deaf people.” Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, 2013.
[38] L. Rello, et. al., “A plug-in to aid online reading in Spanish.” In Proceedings of the 12th Web for All Conference, 2015
[39] В. Терзиева, П. Кадемова-Кацарова, “ Уеб-ресурси и услуги за допълващо обучение на деца със СОП,” 2012
[40] A. Walker, A. Siddharthan, A. Starkey, “Investigation into human preference between common and unambiguous lexical substitutions.” In Proceedings of the 13th European Workshop on Natural Language Generation, 2011, pp. 176-180