UDK 811.111'42
Pregledni naučni rad

Dr. sc. Nejla Kalajdžisalihović

THE LEXICAL SYLLABUS – AN APPLICATION OF VOCABULARY DATA MINING TO MULTIMODAL TEXTS

Abstract

The traditional concept of 'text' is being replaced by 'hypertext', a non-linear and intrinsically intertextual artifact made up of internally linked chunks of information of various sorts (Caballero, 2005, p. 58). As a virtual structure, contemporary text allows for a hybrid mixture of language, images, and sound. For that reason, it is certain that new technologies are affecting the ways knowledge is being transmitted or constructed. In this paper, one method that supports the application of a multimodal approach to text will be presented to demonstrate how the four key skills may be improved: critical thinking, collaboration, communication, and creativity (Galante, 2011). The four skills will be discussed in the context of vocabulary acquisition from multimodal texts. The aim of the paper is to propose that these skills may be significantly improved by means of vocabulary data mining from multimodal texts, and integrating, in the process, the lexical syllabus into the contemporary English language course.

Key words: contemporary English, multimodal text, data mining, lexical syllabus

Introduction

When reflecting on reforms in education, Burke (2009) claims that, for the sake of the future, educators need to cultivate seven personae in students:

(1) storyteller, (2) philosopher, (3) historian, (4) anthropologist,
(5) reporter, (6) critic, and (7) designer.

Burke developed this approach due to the fact that he, among others, found it oppressive that standardized tests and the announcement of high pass rates was expected by the government in the No Child Left Behind reform (2001), which led many schools to manipulate results in their annual reports. That the reform to make students proficient in a certain skill by a certain date failed is one of the reasons this paper emphasizes a life-long approach to acquiring knowledge and integrating the four skills, the 4 Cs (communication, creativity, critical thinking, and collaboration) identified by the Partnership for 21st Century Learning (2002) as the most important skills required for learning in the 21st century, into curricula and language learning programmes. The NCLB reform insisted on three skills only—reading, writing, and arithmetic as these three skills are considered to be "economically relevant". Even though the NCLB was replaced by the Every Student Succeeds Act (2015), today, a question arises regarding what skills should be integrated into curricula, not only in the USA, but worldwide, to allow students to develop the seven personae referred to by Burke. Language and literature departments may find this question difficult to answer as the government has been expanding its role in public education.

In the case of foreign language learning, e.g. studying at a language department, instructors need to encourage students to join the community of those who *communicate*, *create, collaborate*, and *think critically*, textbooks aside. One way to join such a community is to interact with *multimodal texts*, i.e. texts the structure of which is composed of text itself and audio-visual effects. Multimodal texts, as cyber genres, allow for the development of the four Cs and the examples provide a focus on developing skills for lexical decisions in the translation process and general vocabulary acquisition. The aim of the study is to demonstrate how the 4 Cs may be integrated into contemporary curricula, which may result in a life-long approach to learning and working with multimodal texts.

To illustrate this approach to vocabulary acquisition and activation, the paper suggests one method of using multimodal texts as sources for data mining and describes a set of tasks incorporated into one segment of the syllabus for two courses of Contemporary English in the period from October 2017—January 2018. Taking into consideration the fact that the lexis of contemporary English is changing to adjust and conform to the needs of contemporary society, the model proposed focuses on learning from data about English language usage.

Certainly, the internet may also be used as a corpus for individual queries, and it may also be used to assess vocabulary when reading or listening. The multimodal text, being a broad concept, is in this paper used as a source for finding data about vocabulary that would enhance not only the four Cs, but also allow students to, by means of learning from their own language data, assess critically both knowledge and all aspects of living in the 21st century. Although the data-driven learning (DDL) approach has not been widely accepted in the teacher community (Boulton, 2009) due to various aspects of technophobia, it has been widely used for compiling dictionaries and grammar books. In the language classroom, the challenge of the DDL method, among other challenges, is its student-centred approach as "the task of the learner is to 'discover' the foreign language and the task of the language teacher is to provide a context" (Johns, 1991, p. 1). Therefore, the DDL method would not be easily welcomed by teachers who do not favour the student-centred approach to instruction.

The lexical syllabus as the method

One context that may be provided to students to become better thinkers, to better collaborate and communicate with the world, their peers, and their instructors is the context provided by the lexical syllabus. The lexical syllabus is the approach to teaching language advocated by Willis in the 1990s. This approach to analysing and studying language is to focus on authentic texts in learning "rather than text specially written to illustrate some aspect of language" (Willis, 1990, p. 27). Therefore, the DDL method and the lexical syllabus are interrelated, the lexical syllabus being a more tangible framework for data mining.

In this paper, we propose that for efficient language learning and developing the four Cs, a learners' corpus from written or spoken texts is a convenient corpus for concordancing, or learning from multimodal texts. In the case of written texts, a lexical syllabus approach is given by means of using concordances, online dictionaries, and forums on language usage, whereas more focus is given to learners' corpora.

In further text, we illustrate how multimodal texts were used to create learners' corpora for both spoken and written texts. Learning from data in our context, consists of using multiple multimodal sources for the purpose of lexical decisions in the translation process and vocabulary activation such as:

1. *the blog*, incorporated into the syllabus to discuss specific translation-related issues in the classroom (to enhance critical thinking, communication, collaboration, creativity);

2. *typing speed*, incorporated into the syllabus to demonstrate vocabulary activation (to enhance critical thinking, creativity);
3. *online forums* accompanying dictionaries, incorporated into the syllabus to critically assess the emergence of new vocabulary items and their application in real life (to enhance critical thinking, communication, collaboration, creativity);
4. *concordances*, incorporated into the syllabus to make vocabulary-related decisions when reading critically (to enhance critical thinking, collaboration, creativity);
5. *spoken texts,* i.e. the video, incorporated into the syllabus for vocabulary data mining and creating learners' corpora (to enhance collaboration, creativity, communication).
The multimodal sources integrated into a student-centred syllabus could be used as a one-semester project on incorporating the lexical syllabus into the language classroom.

Application of multimodal sources in vocabulary activation

The blog

In the first task, 74 undergraduate students of English, who participated in the translation practicals, used blogs and forums to activate vocabulary needed for in-class translation by means of think-aloud protocols. As Ss were translating literary and non-literary texts from the syllabus, they were asked to go online and find answers on blogs and forums when they were not sure about a translation equivalent. It has been found that the dilemma students had on a particular vocabulary item was often the dilemma of native speakers of English when writing. The most frequently discussed items were *verbs*, for instance, verbs to describe how light moves and verbs of sound emission. For instance, the following blog entry was discussed when translating *Gobleni*, a short story by Mustafa Zvizdić:

> I'm writing a sentence where pale gray morning light is being viewed through window blinds. I'm trying to think of a way to describe its entrance without sounding cliché. What are some good verbs to use for how light moves? (English Language and Usage, March 22nd 2014).

Although users of the blog provided ten answers to this question, the target group opted for: *As the light gently wove its way through the blinds*. *Morning light filtered* had 13,200 results (in Google books), whereas *morning light slipped* gained only 264 results. In both of the cases presented above, a monolingual or bilingual dictionary was compensated by multimodal texts and this particular example, along with others, has been added to the lexical archive. In another story, *Slučajan susret*, by the same author, Ss also used the blog to find appropriate lexical items when making translation decisions during the think alouds.

Blogs and forums have also been used for discussing and critically assessing word formation in English and the creativity of the contemporary digital natives by gaining access, communicating or voting on the Cambridge Dictionary blog and forum to discuss a new category in vocabulary, the so-called 'new senses' or items such as: *supertasker*, *funsultant*, *vibe manager*, *WhatsApp diplomacy*, *gameboy disease*, *deep learning*, *cybersoldier*, *sharenting* etc. Integrating the blog into the lexical syllabus has proven to offer memorable examples while simulating dialogue with peers or speakers whose L1 is English.

Typing speed and vocabulary activation

In another task, two pairs of students were asked to describe the photos they saw in no more than 5 or 10 minutes respectively. The aim of the task was to assess vocabulary activation and detect errors. The first pair achieved an average of 88.5 words in 5 minutes, and the second pair achieved 100 words in 10 minutes. The difference between the two pictures was that the first picture was only a picture of a bottle, whereas the other picture involved activity and agents, i.e. activation of verbs. One sample of the learners' responses with errors is given below:

> Two women depicted are professors at a university. They are two chemists working on their experiment on which they have been working for years. This picture shows them trying to create a new chemical compound. They are both excited and afraid. They are also very hopeful. If the experiment succeeds, they will become famous and acknowledged in their field of research. After conducting it, they realized that they have succeeded to make a breakthrough. Everything went according to their plan and a new compound was born.

The results were compared and discussed also in relation to the importance of the typing speed in translation when there is no extra time given for editing. The results were then analysed for 'self-references, social words, positive emotions, negative emotions, and long words' (see: Pennebaker, 2012). In this task, it was possible for learners to learn from their own errors when activating a vocabulary item under pressure.

Concordances

A beneficial tool for activating vocabulary is learning from concordances. In England, Tim Jones was one of the first language teachers at the University of Birmingham who stressed the importance of using concordances for teaching vocabulary and gaining access to one's own language data, the language data of one's peers, and language data from the language corpora (e.g. BNC, COCA). Unlike the Middle Ages, when concordances were set up manually, or the 1980s, when the concept of concordancing was being introduced into the language classroom, it is without doubt that this approach to developing the four Cs from data-mining may be integrated into a lexical syllabus.

In the third task, 120 undergraduate students watched one documentary each (or a video) on a contemporary topic (or a topic relevant for their age group). While watching the documentary, they wrote down any chunks of language from the video they found relevant for summarizing the video. The video was summarized, the sources provided and presented to the group, and a key-words page designed for each video. All the topics were orally presented and the multimodal text was, in such a way, converted to a written text, and then re-converted to an oral presentation with key words. The summaries were saved in an .rtf format and a corpus of 30,000 words was fed into concordances software for future use. This corpus corresponds to Tribble's suggestion for a corpus size of 25,000-30,000 words (Tribble and Jones, 1997, p. 11) and may be used to demonstrate several layers of vocabulary circulation and acquisition, to check students' examples against larger corpora, and to check for instances of plagiarism.

In the analysis of the students' corpus, it was concluded that the most frequently used words in the papers turned in after summarizing the previously chosen videos were: *people* (145), *social* (137), *media* (113), and *language* (99). The most frequent content words may also be compared to the most frequent function words in the corpus (*the* (1414), *and* (997), *to*

(932), *of* (899), and *a* (553)). By using their own corpus, learners can also compare any entry from the concordances compiled by the instructor with an entry from BNC, COCA, or the Ludwig Guru application. In that case, the most interesting to observe would be *hapax legomena*, or words that occur only once in the corpus, such as the occurrence of a word only once due to a spelling error (e.g. *anually*, *hieroglific*) or a typing error (e.g. *byacknowledging*). Some words occur only once because they are foreign words in the corpus (e.g. *ich*), personal names or surnames (e.g. *Alex*, *Richterova)*, abbreviations or acronyms (e.g. *EEG*), new terms (e.g. *phubbing*) or problems in choosing the dash (hyphen, em dash, or en dash) etc.

This method of learning from data and learning from multimodal texts may be applied in high school or primary school classrooms as well, especially when teaching grammar and asking students to infer grammar rules and collocations. For that reason, the benefits of joining learners' corpora to infer and assess vocabulary knowledge by mining data from multimodal texts are endless when combined with the assets of the web.

Conclusion

The aim of the present paper was to assess whether effective language learning is a form of linguistic research. As such, learning as research in the 21st century requires major reforms in curricula that would allow students to work with multimodal texts and develop the four skills of the 21st century: *critical thinking, collaboration, communication, and creativity*. Aside from the need for massive retooling and interdisciplinary approaches, the restrictions on fully applying the DDL method lie also in the authorities' requirements for measuring knowledge. Integrating multimodal texts and learning from corpora is a life-long process and may, as such, be integrated into the syllabi to promote learner autonomy and provide direct access to language thus enhancing discovery learning, motivation, noticing, sensitisation, and, finally, learning to learn. In the approach tested for the purpose of innovation in the syllabus, five multimodal sources were used to enhance vocabulary activation through the 4 Cs with the assumption that learners will, in the long run, develop Burke's seven personae having benefited from individual data mining.

Works cited

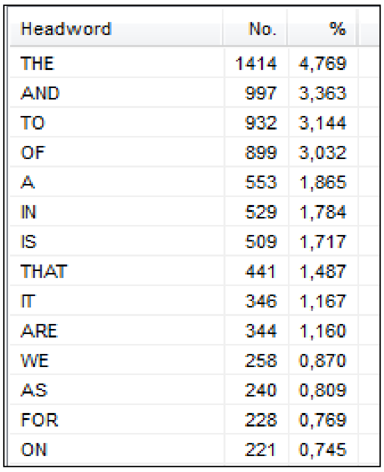
1. Boulton, A. (2009). Data-driven learning: on paper, in practice. In: T. Harris and M. Moreno Ja´en. *Corpus linguistics in language teaching*. Peter Lang, Linguistic Insights. Retrieved from https://hal.archivesouvertes.fr/file/index/docid/393809/filename/2009_boulton_LANG_paper.pdf on Jan 5 2018.
2. Burke, J. (2009). Reimagining English: The seven personae of the future. *English Journal*, V. Urbana, IL: National Council of Teachers of English, pp. 12-15.
3. Caballero, R. (2005). The influence of hypertext on genre: exploring online book reviews. In: R.C. Caldas-Coulthard and M. Toolan (Eds.). *The writer's craft, the culture's technology*. Amsterdam: Rodopi.
4. Galante, N. (2011). Teaching Generation M: Web 2.0 tools in (and out of) the classroom. In: M. Koehler and P. Mishra (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2011*. Chesapeake, VA: Association for the Advancement of Computing in Education, pp. 3211-3215.

5. Johns, T. (1991). Should you be persuaded—two examples of data-driven learning materials. In: T. John and King P. (Eds.) *Classroom concordancing*. Special Issue of *English language Research Journal 4*, Birmingham, UK: University of Birmingham: Centre for English Language Studies, pp. 1-16.
6. Pennebaker, J. (2012). *The secret life of pronouns*. Exercises. Retrieved from http://secretlifeofpronouns.com/exercises.php on Dec 12th 2017.
7. Tribble, C., Jones, G. (1997). *Concordances in the classroom*. Chris Tribble and Glyn Jones: Athelstan Publications.
8. Willis, D. (1990). *The lexical syllabus: a new approach to language teaching*. London and Glasgow: Collins E.L.T.

Secondary sources

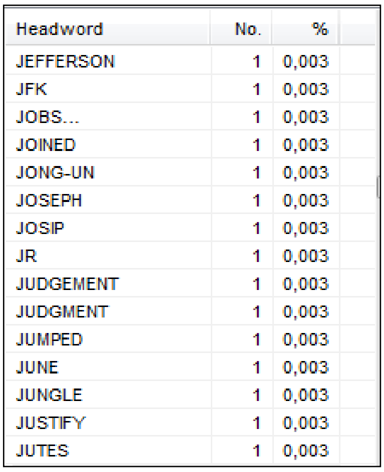
1. A blog from Cambridge Dictionary (2017). Retrieved from https://dictionaryblog.cambridge.org/ on December 9th 2017.
2. An Educator's Guide to the Four Cs—Preparing 21st Century for a Global Society. USA: National Education Association. Retrieved from http://www.nea.org/assets/docs/A-Guide-to-Four-Cs.pdf on November 22nd 2018.
3. English Language and Usage Stack Exchange. Retrieved from https://english.stackexchange.com/questions/159198/verbs-to-describe-how-light-moves on Nov 29th 2017.

Appendix



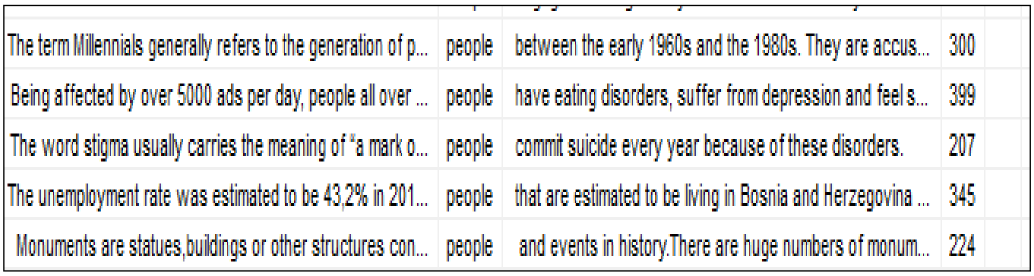
Headword	No.	%
THE	1414	4,769
AND	997	3,363
TO	932	3,144
OF	899	3,032
A	553	1,865
IN	529	1,784
IS	509	1,717
THAT	441	1,487
IT	346	1,167
ARE	344	1,160
WE	258	0,870
AS	240	0,809
FOR	228	0,769
ON	221	0,745

Fig. 1. Screenshot of the most frequent words in the Ss corpus



Headword	No.	%
JEFFERSON	1	0,003
JFK	1	0,003
JOBS...	1	0,003
JOINED	1	0,003
JONG-UN	1	0,003
JOSEPH	1	0,003
JOSIP	1	0,003
JR	1	0,003
JUDGEMENT	1	0,003
JUDGMENT	1	0,003
JUMPED	1	0,003
JUNE	1	0,003
JUNGLE	1	0,003
JUSTIFY	1	0,003
JUTES	1	0,003

Fig. 2. Screenshot of the words used only once in the Ss corpus



The term Millennials generally refers to the generation of p...	people	between the early 1960s and the 1980s. They are accus...	300
Being affected by over 5000 ads per day, people all over ...	people	have eating disorders, suffer from depression and feel s...	399
The word stigma usually carries the meaning of "a mark o...	people	commit suicide every year because of these disorders.	207
The unemployment rate was estimated to be 43,2% in 201...	people	that are estimated to be living in Bosnia and Herzegovina ...	345
Monuments are statues,buildings or other structures con...	people	and events in history.There are huge numbers of monum...	224

Fig. 3. Screenshot of the concordances for the noun 'people' the Ss corpus

LEKSIČKI SILABUS – PRIMJENA RUDARENJA PODATAKA O VOKABULARU IZ MULTIMODALNOG TEKSTA

Sažetak

Tradicionalno shvaćanje koncepta „teksta“ postepeno biva potisnuto „hipertekstom“, nelinearnim i inherentno intertekstualnim artefaktom koji se sastoji od međusobno povezanih skupova informacija različite vrste (Caballero, 2005, p. 58). Kao virtuelna konstrukcija, savremeni tekst podržava hibridni spoj jezika, slike i zvuka. Iz tog razloga, neminovno je da nove tehnologije imaju ulogu u tome kako će se znanje prenositi ili konstruirati. Budući da je usvajanje vokabulara savremenog engleskog jezika poseban izazov, u ovom radu bit će predstavljen metod kojim je moguće, korištenjem multimodalnog pristupa tekstu, pomoći studentima da razvijaju danas četiri najtraženije vještine: kritičko mišljenje, zajedničko djelovanje, komunikativnost i kreativnost (Galante, 2011). Vještine se razmatraju u kontekstu usvajanja vokabulara iz multimodalnog teksta, a cilj rada je predstaviti model razvoja navedenih vještina rudarenjem podataka o vokabularu iz multimodalnog teksta, a potom integriranjem leksičkog silabusa u nastavu savremenog engleskog jezika.

Ključne riječi: savremeni engleski jezik, multimodalni tekst, rudarenje podataka, leksički silabus