Automatic Assessment of Short Answers Using Latent Semantic Analysis

UDC 81'322.4

DOI 10.18485/infotheca.2023.23.1.4

Teodora Mihajlov
teodoramihajlov@gmail.com
University of Belgrade
Belgrade, Serbia

ABSTRACT: Implementing technology in a modern-day classroom is an ongoing challenge. In this paper, we created a system for an automatic assessment of student answers using Latent Semantic Analysis (LSA) – a method with an underlying assumption that words with similar meanings will appear in the same contexts. The system will be used within digital lexical flashcards for L2 vocabulary acquisition in a CLIL classroom. Results presented in this paper indicate that while LSA does well in creating semantic spaces for longer texts, it fell somewhat short of detecting topics in answers and word definitions. The answers were classified using KNN, for both binary and multinomial classification. The results of KNN classification are as follows: precision $P = 0.73$, recall $R = 1.00$, $F_1 = 0.85$ for binary classification, and $P = 0.50$, $R = 0.47$, $F_1 = 0.46$ score for the multinomial classifier. The results are to be taken with a grain of salt, due to a small test and training dataset.

KEYWORDS: LSA, CLIL, L2 vocabulary acquisition, cosine similarity, KNN

PAPER SUBMITTED: 1 April 2023
PAPER ACCEPTED: 7 May 2023

1 Introduction

Employing technology to improve language learning outcomes is the problem scientists have wrestled with since the 1960s. In this paper, we will present a model for an automatic assessment of student answers using Latent Semantic Analysis (LSA). In further development, the model will be implemented within digital lexical flashcards (henceforth referred to as: *flashcards*) for learning vocabulary in English as a Second Language (ESL) classes.

Previous research (Landauer, Foltz, and Darrell 1998; Lemaire and Dessus 2003; Lifchitz, Jhean-Larose, and Denhière 2009) shows that many cognitive abilities in humans, including vocabulary acquisition, are well-represented by LSA. Furthermore, assessments provided by LSA largely correlated with those done by evaluators (Landauer et al. 1997; Graesser et al. 2000; Lemaire and Dessus 2003; Landauer, Laham, and Foltz 2003; Picca, Jaccard, and Eberlé 2015). Flashcards have so far proven to be a good tool for textscl2 vocabulary acquisition, combining interval (Ashcroft, Cvitkovic, and Praver 2018) and conscious learning (Nation 2006; Hung 2015) —- two approaches that enhance learning outcomes, especially at the lower levels of language knowledge (Ashcroft, Cvitkovic, and Praver 2018). Taking everything previously said into account, we believe that developing this model will help us tackle several methodological problems, and contribute to digitalisation of L2 classroom at the Faculty of Mining and Geology, University of Belgrade.

In accordance with the paper's interdisciplinary approach, our aims are twofold — pedagogical and methodological. The former is to examine current general and geological vocabulary knowledge of the Faculty's students and associates, as well as to improve present teaching methods by helping develop digital learning materials. The latter aim is to look into LSA's application in assessing student answers in the geological domain, and with a form of a definition, and to see to which extent the model's assessment correlates with the evaluators'. Conforming to the aforementioned aims, our hypotheses are as follows — (1) creation of the system will help digitalise learning materials; (2) LSA will be successful in assessing student answers.

2 Vocabulary acquisition using flashcards

Knowing a word includes knowing its form, meaning and use (Nation 2006). The time that a student spends learning, along with the student's involvement in the learning process, affects vocabulary learning outcomes. Spaced (or distributed) learning, i.e. learning in many small sessions increasing the breaks between each session, showed the best results when it comes to vocabulary learning (Nation 2006). In reference to the level of student involvement in the learning process, some authors presume that word form is learned implicitly as a result of frequent exposure to a word, while word meaning that is abstract and belongs to the domain of semantics, is acquired by explicit learning. Explicit learning uses an array of tools, such as dictionaries, word lists, and flashcards (Nation 2006; Hung 2015; Ma 2009).

Several researches display that flashcards give the best results when it comes to L2 vocabulary acquisition outcomes, especially at the lower levels of language knowledge (Spiri 2008; Nakata 2008; Hung 2015; Averianova 2015; Yüksel, Mercanoğlu, and Yılmaz 2022). Flashcards provide simultaneous explicit and interval learning of vocabulary, together with learning word form, meaning and use in context (Ma 2009). Nowadays, students can also learn using digital flashcards, which have several advantages compared to their paper predecessors —- user receives feedback on their spelling and grammar errors at once, the system can use pictures and sound and can be accessed from different devices. Moreover, students can now learn whenever and wherever making digital flashcards a great tool for spaced learning. Additionally, they can be used for both learning and tests, as well as gamification of the learning process. Today, the most commonly exploited systems are *Anki*[1] and *Quizlet*[2] (Ashcroft, Cvitkovic, and Praver 2018).

Learning geological terminology in ESL classes is explored in Beko, Obradović, and Stanković (2015). The paper highlights several issues, including difficulties students have in finding a suitable learning method, low level of language knowledge at the begging of studies, and lack of translation of geological terminology to Serbian, which makes translational tasks even more difficult (Beko, Obradović, and Stanković 2015). Given that our model will be monolingual, we will not address the last-mentioned issue.

Faculty of Mining and Geology currently uses language tools such as a thesaurus of geological terminology in Serbian and English, comprised of roughly 2800 words (Beko, Obradović, and Stanković 2015), and a digital mining terminology platform *RudOnto*.[3] Additionally, a system of flashcards *RGF Flashcards* was developed, using *Anki*. The flashcards were integrated into the Faculty's *Moodle* platform.[4]

Given that our flashcards will be used in a CLIL classroom, that integrates learning content from a certain domain with language learning (Beko 2013; Đerić 2019; Baten, Van Hiel, and De Cuypere 2020), whereby C1 entry lan-

1. Anki flashcards, accessed 20 May 2023
2. Quizlet flashcards, accessed 20 May 2023
3. RudOnto thesaurus, accessed 20 May 2023
4. Moodle, accessed 20 May 2023

guage knowledge is expected, the flashcards can be used to facilitate vocabulary acquisition process for students with lower levels of English-language knowledge, consequently making following learning materials and classes easier.

3 Latent Semantic Analysis lsa

Latent Semantic Analysis LSA is a theory and method for extraction and representation of word meaning in context, whereby statistical calculations are applied to a large text corpus (Landauer et al. 1997). Thus far, research has shown that LSA can largely represent human cognitive abilities, such as vocabulary acquisition, word categorisation, semantic priming, discourse comprehension, and essay assessment (Landauer, Foltz, and Darrell 1998).

When creating a LSA model, we start by computing a document-term matrix, thereby forming a semantic space comprised of all terms and all documents in our corpus (Deerwester et al. 1990; Landauer, Foltz, and Darrell 1998; Lemaire and Dessus 2003). The next step is applying a weighting function to each matrix cell, which assigns small weights to high-frequency terms, and high weights to terms that appear in some, but not all document (Deerwester et al. 1990; Martin and Berry 2013). After this, singular value decomposition (SVD) is performed (Equation 1). This is an instrumental part of LSA, because it makes it possible to present word meaning relative to the context it appears in (Lemaire and Dessus 2003). If m and n are integers and M is an arbitrary matrix $m \times n$, then decomposition of a matrix M_{mn} is the following:

$$M = U \times S \times V^T \tag{1}$$

where:

- M: is an orthogonal $m \times n$ matrix (document-term matrix), where m is a number of documents and n the number of terms;
- U: is an orthogonal document-topic $m \times r$ matrix, where m is a number of documents and r the number of topic;
- S: is a diagonal $r \times r$ matrix, in which all values but those on the diagonal are equal to 0. The diagonal values in S represent how much each topic explains variance in the data;
- V: is an orthogonal $n \times r$ matrix (term-topic), where n is the number of documents and r is the number of topics.

Multiplying the three SVD matrices, we can approximately reconstruct the original matrix M. Reduced matrix's M_k dimensionality ought to be optimal, and accurately represent relationships between elements in the original matrix M (Landauer, Foltz, and Darrell 1998). For checking the validity of the number of dimensions, an external validation criterion is used (Landauer, Foltz, and Darrell 1998). In this paper, we will use a cosine similarity measure between students' and correct answers (Rahutomo, Kitasuka, and Aritsugi 2012).

LSA has hitherto been used for answer assessment, providing feedback, answering student questions, as well as assessing student essay accuracy and coherency, in several smart games. In the essay assessment task, it displayed a high degree of correlation with evaluator assessments (Landauer et al. 1997; Graesser et al. 2000; Lemaire and Dessus 2003; Landauer, Laham, and Foltz 2003; Dikli 2006; Lafourcade and Zampa 2009; Picca, Jaccard, and Eberlé 2015).

4 Checking and preparing data

Our data is constituted of three parts — (1) unit texts written for Prof. Dr. Lidija Beko's English-language textbook in preparation, 12 units with 3 texts each; (2) vocabulary for each unit, split into three categories —- general vocabulary (663 words), geological vocabulary (280 words), and minerals (18 words); (3) student answers collected by tests via the Faculty's *Moodle* platform, for subjects English 1–4. There were three groups of tests, for three groups of participants. All groups had the same questions with different examples and were formed using an example from Jhean-Larose et al. (2010). Some questions (e.g. question six) were adjusted to the research aims. The test description is to be found in Table 1.

As mentioned, tests were implemented in *Moodle* platform enhanced with HP5[5] extinction and shared with the participants. The test was completed by 14 participants, and 451 answers were collected. For anonymity purposes, we created a unique numerical ID for each participant. The most answers were collected for the first group of the test, and the fewest for the third. Answer assessment by an evaluator was done in two steps. First, each answer was assessed on a scale from 1 to 5 (Table 2).

Secondly, answers marked with 1 were labeled as incorrect (I), while the rest were labeled as correct (C). The assessment criterion was answer

5. H5P extension, accessed 23 May 2023.

No.	Question	Example
1	mark definitions as TRUE or FALSE	*fossilisation is a process in which parts of a dead animal or plant being turned into a part of sediment and thereby preserved* TRUE
2	connect 11 words to their respective definitions	*fern –) a vascular plant with complex fronds and sporangia on the leaves' surface where asexual spores are found*
3	write a definition using the following words	using the following words: *collection, fragment* define *debris (a collection of fragments of rocks)*
4	connect definition parts A and B, then determine which word the definition refers to	Part A: *pertaining to complex protoplasmic life-forms*, Part B: *with a vesicular nucleus and various cytoplasmic organelles - eukaryotic*
5	mark listed characteristics as TRUE or FALSE	*embed is…) to be placed* TRUE; *) within something* TRUE; *) so that the part can be easily removed* FALSE
6	explain **why** a certain phenomenon occurs and **what** it is	*Explain what global warming is and why it happens; What is seafloor?*
7	explain **how** something happens	*How are sedimentary rocks formed?*
8	write definitions for these 10 words	*mineral (a naturally occurring inorganic substance with a characteristic chemical composition)*

Table 1: Test used for answer collection

similarity with the golden standard —- word definition from the textbook, as well as the evaluator's language competence. Since our model does not take into account grammar and spelling, neither did the evaluator when assessing student answers. Due to an inconsistent output, question 5 was not included in the analysis, while questions 3, 6, 7, and 8 were estimated fit for analysis using LSA. This selection left us with 238 answers. After removing missing values, 72 answers remained.

Mark	Meaning
1	the answer is completely incorrect
2	a part of an answer is correct
3	the answer is incomplete
4	the answer is almost completely correct
5	the answer is completely correct

Table 2: Evaluator grading scale

4.1 Text preparation

Text preparation was conducted in accordance with methods found in the literature (Deerwester et al. 1990; Dikli 2006; Lifchitz, Jhean-Larose, and Denhière 2009), which we adapted to our goals and our data. The first step in text preparation was text lemmatisation using *SpaCy* library.[6] After obtaining lemmatised text surrogates for each part of our data, we removed punctuation and special characters using regular expressions and changed text to lowercase. In addition, we removed Latin abbreviations and plurals from the vocabulary (e.g. *data sing. datum, hypothesis pl. hypotheses*). An example of text before and after preparation is displayed in Table 3. The examples are extracted from different texts.

We noticed inconsistencies when it comes to the lemmatisation of some verbs. For example, verb *split* was not lemmatised in *an act of splitting into a category*. However, going through other examples with similar structures (verb-adverb-verb or noun-verb-adverb), e.g. *an action of taking something; the process of breaking something down* we did not detect the same mistake. Furthermore, the verb *unstratified* was lemmatised to *unstratifie* when standing alone, while it remained in its original form in the answer *incoherent loose unstratified*. Latin words, such as *antennae* or *Pinaceae*, were not lemmatised at all.

5 Developing answer assessment model

For developing our LSA model, *Scikit-Learn* Python library was used.[7] First, we constructed a TF-IDF matrix (*Term Frequency-Inverse Document Fre-*

6. SpaCy library, accessed 22 May 2023
7. Scikit-Learn library, accessed 22 May 2023

Original text	**Processed text**
Most people today are familiar with mineral water and the perennial debate, as to whether still or sparkling is better.	most people today be familiar with mineral water and the perennial debate as to whether still or sparkle be well
Groundwater stored in subterranean aquifers has always been extracted for human use through the digging of wells.	groundwater store in subterranean aquifer have always be extract for human use through the digging of well
The conversion of sediment to rock is known as lithification transformation or diagenesis, and tends to involve two stages – compaction and cementation.	the conversion of sediment to rock be know as lithification transformation or diagenesis and tend to involve two stage compaction and cementation

Table 3: Processed text

quency), with documents in matrix rows, terms in matrix columns, and relative term frequencies in each of the documents in matrix cells (Jurafsky and Martin 2023).

When constructing a TF-IDF matrix, it is necessary we decide how many terms are to describe each document. Trying out options between 700 and 5000 terms, we decided that the number of terms in TF-IDF matrix for unit texts be 1000, that minimal term frequency would be 3, and that a term can appear in no more than 80% of the documents to be found in the matrix. In this step, we also removed stop words. Stop words list was comprised of the English language stop words from NLTK library[8] and stop words specific for our corpus (*km, kmh, mm, meter, one, two, three, etc., yet, well...*). Initially, we applied the same parameters to the rest of our data, i.e. vocabulary and participant answers, but this gave poor results. Thus, we lowered the number of dimensions to 700 and minimal frequency to 1, and increased maximum frequency to 100% of the documents, while the stop word list contained only definite and indefinite article —- *a/an, the.*

SVD parameters are the same for all parts of the data. In order to determine an appropriate number of topics, we extracted 15 terms with the highest weights for each topic and examined if there are any overlaps between topics, or if the model has missed something. Finally, we opted for 10

8. NLTK library, accessed 22 May 2023

topics. Then, we assigned a name to each topic based on the first 100 terms with the highest weights. Some topics, such as *Topic0, Topic1, Topic4*, contained more general terms that are woven through most texts. On the other hand, topics *Topic3, Topic5, and Topic7* contain terms from different geology branches, like tectonic plates, volcanology, and erosion (Table 4).

After obtaining topic vectors for all parts of the data, we measured cosine similarities between all texts so as to get the most similar ones. Next, we calculated a final score for each answer as a mean of cosine similarity of answer A and: a) vector of the text in which the word defined in the answer is used; b) vector of the correct answer (*golden standard*); c) vector of the most similar answer B to the target answer A. The higher the similarity score of document A and document B, the higher the connection between the documents (Rahutomo, Kitasuka, and Aritsugi 2012). Finally, answers were classified into two and more categories according to accuracy and based on the obtained answer score. For classification purposes, both binominal (*Correct / Incorrect*) and multinomial (criteria displayed in Table 2), KNN (*K-Nearest Neighbour*) algorithm was employed (Li, Yu, and Lu 2003; Peterson 2009).

5.1 Distribution of texts, definitions, and answers among topics

Prior to proceeding to answer classification, let us take a look at how the data is distributed among topics. This will provide us with insight into our model's validity. We observed the most uneven topic distribution in texts, while it was somewhat more uniform, although still irregular, in vocabulary and answers. We believe that the reason behind lower standard deviation (STD) (Urdan 2005) in vocabulary and answers is a more coherent text form, compared to unit texts.

In unit texts, maximal topic values vary between 0.5617 in *Volcanology*, to 0.3638 in *Dating*, while minimal values fluctuate from 0.3878 for *EarthFormation*, all the way to -0.001 for topic *Dating*, and STD in topics is high. Maximal values in definitions are, to a degree, more evenly distributed. Topic *Weathering* (0.7067) has the highest maximum value, while the lowers is that of *Landslides* (0.3547). Almost all minimal topic values are negative, apart from the topic *EarthFormation*, with a minimal value of 0.0193. While topics are assigned well to some geological terms, e.g. *debris*[9] has high values in *EarhFormation, Weathering* and *Landslides*, the model failed to recognise latent topics in others, which is shown for example in a low value of topic

9. the remainders of something destroyed

Topic	Name	Terms with the highest weights
Topic0	Earth Formation	mineral, cycle, earth, deposit, flow, sedimentary, igneous, material, soil, metamorphic, sediment, begin, grain, metamorphism, plant
Topic1	Minerals	mineral, metamorphism, grain, metamorphic, igneous, metamorphic rock, pressure, crystal, magma, ore, deposit, chemical, metallic, thermal, colour
Topic2	Erosion	flow, soil, particle, stream, slope, erosion, debris, landslide, glacial, material, groundwater, sand, glacier, velocity, move
Topic3	Tectonic Plates	plate, earthquake, wave, cycle, tectonic, magma, continental, oceanic, magnetic, magnetic field, earth, activity, stress, volcano, temperature
Topic4	Rock Formation	sedimentary, cycle, sediment, metamorphic, igneous, sedimentary rock, strata metamorphic rock, metamorphism, erosion, grain, plate, rock cycle, igneous rock, pressure
Topic5	Volcanology	magma, lava, grain, volcano, slope, eruption, volcanic, viscosity, period, landslide, hazard, volcanic eruption, debris, era, mesozoic
Topic6	Weathering	wave, earthquake, magnetic, date, particle, magnetic field, metamorphism, stress, erosion, sediment, grain, field, age, sedimentary, strata
Topic7	Landslides	slope, landslide, soil, debris, hazard, cycle, trigger, activity, fall, downslope, mitigation, metamorphism, metamorphic rock, metamorphic, angle
Topic8	Dating	earth, strata, magma, date, age, eruption, lava, idea, satellite, remote, atom, history, feature, geological, sedimentary
Topic9	Fossils	oil, wave, earthquake, coal, trap, organic, sedimentary, sedimentary rock, weathering, carbon, plant, oil gas, soil, gas, type

Table 4: Text topics and terms with highest weights for each topic

Fossils in definitions of terms *fossil*,[10] *fossilised, fossilisation.* This means that, according to our model, topic of *Fossils* does not describe the afore-

10. parts of animals or plants that have been hardened and preserved in sediment

mentioned terms in a high degree. The distribution of general vocabulary among topics was harder to evaluate since topics pertain to geology. The phrasal verb *wear away*[11] has high values for topics *EarthFormation*, Erosion and Weathering, which aligns with its meaning. However, words such as adverbs *therefore, yet, anyway* etc. were probably the most difficult ones to distribute since they appear in most topics.

In participant answers, we find topic *TectonicPlates* has the highest maximal value (0.7042), while the lowest one is that of *Landslides*, with just 0.3612. Minimal values are for the most part negative, and have values between -0.5557 for Minerals and 0.0000 for *EarthFormation*. Answers to the same question mainly have similar topic distribution. Most answers to the question *global warming*[12] have the highest values for the topic *EarthFormatoin*, and the lowest for *Erosion* and *Landslides*. All topic values for short, incomplete answers are equal to 0.

6 Results

We will begin the result presentation by displaying dominant topics in texts, definitions, and participants' answers. In the following section, we will show the way in which we calculated answer accuracy, and analyse the results. Finally, we will evaluate the results of bi- and multinomial KNN classification algorithms.

6.1 Dominant topics in texts, definitions and participant answers

Three dominant topics were extracted for each sample in each part of the data. In unit texts, *EarthFormation* is the most frequent as the first dominant topic, appearing in 21 documents, then in 11 as the second dominant topic, and in 2 documents as the third dominant topic (Figure 1). Dating is the only topic not found as the first dominant, but it comes second by frequency as the second dominant one, along with topic *Minerals*. All topics appear as the second dominant topic, *Landslides* coming up just once in unit text *Mass Wasting and Types of Landslides*, while *Erosion* is the only topic not appearing in the place of a third dominant topic (Figure 1).

11. an action of gradually eroding or grinding something down

12. global warming an increase in global temperature due to various factors such as an increase in carbon dioxide emission and pollution with a potentially catastrophic outcome

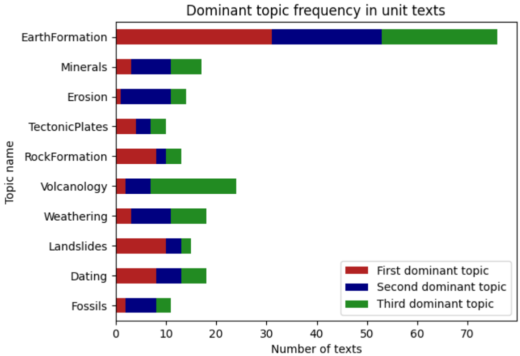

Figure 1: Dominant topic frequency in unit texts.

Analysing values of dominant topics in unit texts, we can see that the first dominant topic values are the least scattered, ranging from approx. 0.3 to 0.6. The second and third dominant topic values are more scattered and somewhat lower, stretching from a bit below 0.1 to about 0.4 for the second, and from negative values up to 0.4 for the third dominant topic (Figure 2). For the most part, the model did well in detecting dominant topics in unit texts. Nonetheless, in the text *Mineral Evolution*, *Minerals* is not to be found among the dominant topics, but it appears in the following text of the same unit, *Physical Properties*, that discusses the origin and physical properties of minerals.

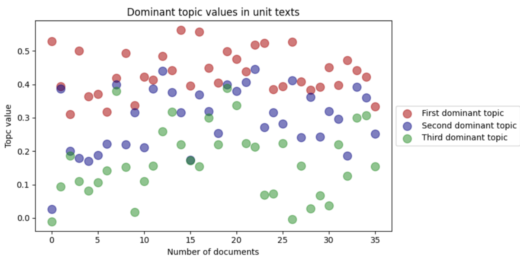

Figure 2: Dominant topic values in unit texts.

Based on the results, we can conclude that there are large differences in values of the first and second dominant topics of a document. In spite of that, the second and third dominant topics often more closely specify what a text is about. Therefore, all three topics should be taken into consideration when analysing the results.

EarthFormation is also the most frequent first dominant topic in vocabulary, and most frequent overall, the same as in texts. After that, comes *TectonicPlace*, in close frequency with *Fossils, Volcanology* and *Landslides*, while *Minerals* is the least frequent first dominant topic (Figure 3).

In a definition of the word *glaciation*,[13] topics *Weathering, EarthFormation* and *Volcanology* are found as dominant. On the other hand, *earthflow*[14] is primarily placed in topics *Fossils, EarthFormation*, and *TectonicPlates*, even though, intuitively, we would perhaps list *Erosion, Landslides* and *Weathering*.

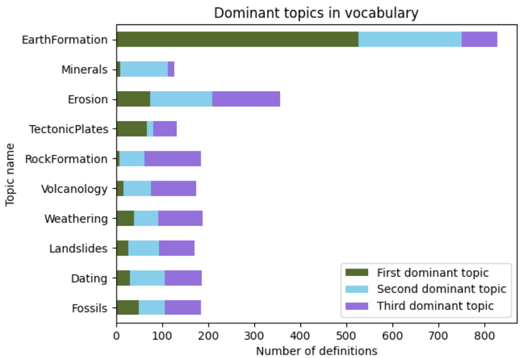

Figure 3: Dominant topics in vocabulary.

In Figure 4 can see that values pertaining to the first dominant topic are more scattered than in unit texts, fluctuating between over 0.7 and below 0.1. Second dominant topic values are slightly lower than those of the first dominant topic, but equally irregularly distributed, while values of the third dominant topic in vocabulary are predominantly low – below 0.4, and relatively homogeneous.

Lastly, *EarthFlow* scores first in frequency in all three places in the participants' answers, emerging 31 times as the first, 25 times as second, and 18

13. used for referring to geological processes of a glacier - its formation, movement, and recession

14. a downslope movement of unconsolidated material, usually caused by percolation of water between the loose particles

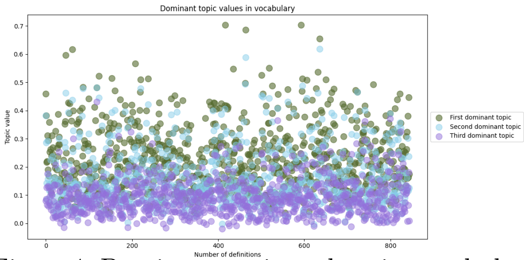

Figure 4: Dominant topics values in vocabulary.

times as the third dominant topic. *Landslides* is second most frequent as the first dominant topic, having the largest values in 10 answers, while *Erosion* is second as the second dominant topic with 12 answers. All topics are of similar frequency as the third dominant topic, apart from *Landslides*, which appears only 2 times (Figure 5).

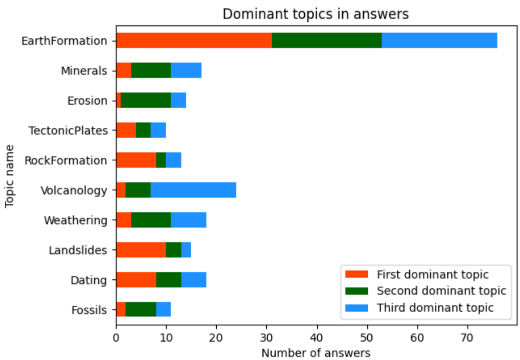

Figure 5: Dominant topics in participant answers.

Generally, the same topics are assigned to long answers to the same question. For example, all answers defining *global warming* have *EarthFormation*, *Volcanology* and either *Minerals* or *Dating* as dominant topics. In the participants' definitions of a notion of *sedimentary rock*,[15] *EarthFormation* and *Landslides* take turns as the first two dominant topics, while *Fossils*, *Dat-*

15. a type of rock formed by accumulation and cementation of transported sediments by means of water, wind, glacier, or gravity

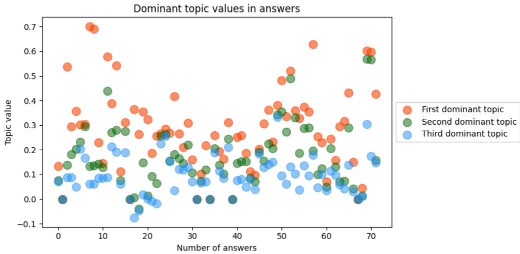

Figure 6: Dominant topic values in participant answers.

ing and *Erosion* rotate at the third place. Zero values are explained by the unsorted answers (Figure 6).

After going through topic distribution, values, and dominant topics in all parts of the data, we compared extracted dominant topics of the same notions in answers, vocabulary, and relevant unit texts (Table 5).

For example, extracted dominant topics for answer, definition, and unit text for the term *transpire* do not align, and their values differ in all parts of the data (Table 5). On the contrary, we detected a high degree of similarity between topics and their values for the term *global warming* (Table 6). We believe that the reason for this is that we have several long answers for the term *global warming*, all containing terms with big weights in topics, which made answer distribution easier for the model.

6.2 Measuring unit text similarity

To measure the similarity of unit texts in the textbook, we applied cosine similarity measure to topic vectors of all texts, retrieved after applying SVD to a previously constructed TF-IDF matrix. Based on the cosine similarity results, we can see how well our LSA model recognised latent topics in unit texts.

Analysing the results, the supposition is that latent topics in texts are well-detected and that the most similar texts indeed convey similar topics, so a text about Wagner's hypothesis that explains an assumption of the existence of Pangaea, has the highest similarity with a text about tectonic plates. Furthermore, a text about volcanology is closely matched to a text about igneous rocks (Table 7).

Term	Definition	Source	Topic value			Topic name		
			I	II	III	I	II	III
transpire	transpire an action of discharge liquid through a plant discharge perspire excrete flow		0.2037	0.1603	0,0302	Erosion	Earth Formation	Fossils
transpire	release liquid through opening in leave of a plant		01478	0.0435	00418	Earth Formation	Weathering	Volcanology
Unit	**Text**							
metamorphic rocks and gemstones	the causes and definition of metamorphism		0.4378	0.4111	0.2329	Earth Formation	Minerals	Rock Formation
	metamorphic textures		0.5223	0.4447	0.1867	Minerals	Earth Formation	Volcanology
	gemstones associated with metamorphism		0.5269	0.2629	0.0683	Earth Formation	Minerals	Rock Formation

Table 5: Dominant topics for the term *transpire*

Term	**Definition**	**Source**	**Topic value**			**Topic name**		
			I	**II**	**III**	**I**	**II**	**III**
global warming	global warming an increase in global temperature due to various factor such as increase carbon dioxide emission and pollution with a potentially catastrophic outcome		0.1256	0.1159	-0.0016	Volcanology	Earth Formation	Dating
global warming	rise in global temperature of the earth due to carbon dioxide emission		0.3257	0.2093	0.0006	Earth Formation	Volcanology	Minerals
Unit	**Text**							
fossils through times	fossil formation and palynology		0.4997	0.1832	0.0836	Earth Formation	Fossils	Volcanology
	palaeozoic era		0.3641	0.2228	0.0228	Earth Formation	Minerals	Rock Formation
	mesozoic era and cenozoic		0.3702	0.2567	0.0263	Earth Formation	Volcanology	Weathering

Table 6: Dominant topics for the term *global warming*

Text a Headline	Text b Headline	Similarity
palaeozoic era	mesozoic era and cenozoic	0.9776
wegener s hypothesis, seafloor spreading, convection cells	tectonic plates	0.8980
volcanoes	igneous rocks	0.8229
the causes and definition of metamorphism	metamorphic textures	0.9606
coal as a fossil fuel	oil and natural gas mineral oil	0.7726

Table 7: Examples of the most similar texts

6.3 Computing answer scores

In this section of the paper, we will present results of cosine similarity measured between a vector of answer A and: (1) vector of a unit text containing the target word; (2) vector of the correct answer; (3) vector of the most similar answer B. In order to see if there are any differences between scores of correct and incorrect answers, we will take a look at the distribution of answer score values among two and more grading categories obtained by the evaluator's answer assessment.

QID	PID	Question	Cosine similarity				C/I	1–5
			Text	Def.	Answer	Score		
6	3	global warming	-0.2713	0.7325	0.8629	0.8004	C	3
6	6	global warming	0.7591	0.8259	0.9680	0.8998	C	4
7	6	sedimentary rock	0.5107	0.6346	0.9422	0.8033	C	5
7	1	sedimentary rock	-0.2183	0.5181	0.9317	0.7538	C	5
3	4	hypothesis	0.0000	-0.0094	0.7816	0.5527	C	5
3	3	hypothesis	0.7104	0.7017	0.8666	0.7885	C	4

Table 8: Cosine similarity results — Examples for answer-text similarity; QID – question ID, PID – participant ID.

When comparing answers and texts, we measured the cosine similarity of an answer and each of the three texts of a unit in which the defined notion is explained and then computed a mean of the three values so as to obtain the similarity between an answer and the entire unit. The results of this computation should be taken with a grain of salt because text vectors assuredly differ from unit vectors. We noticed greater similarities between units and long answers, especially the ones referring to geological terms, such as *global warming*, and *sedimentary rock*. Contrastingly, we did not obtain steady results for an answer to the definition of *hypothesis*,[16] where we have a similarity of 0 for one answer, and a rather high similarity for another answer (Table 8).

As displayed in Figure 7, measured cosine similarity values between answers and units are somewhat higher in correct than in incorrect answers. In the five-category distribution, we have the greatest variance in categories

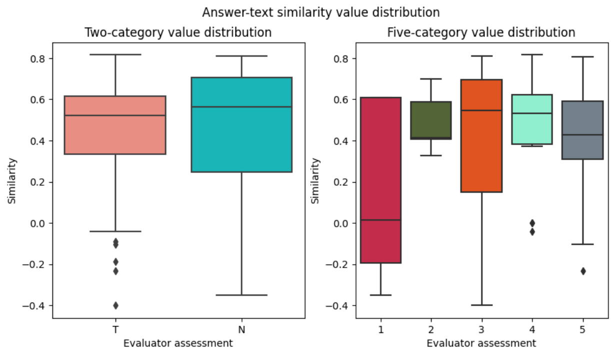

Figure 7: Answer-text similarity value distribution.

3 and 5. The lowest values are detected in category 1. Outliers are answers to questions *chemical weathering*[17] and *convergent*[18] (Figure 7).

The completeness and correctness of an answer are best represented by the cosine similarity of an answer and the correct answer. Each of the an-

16. an assumption resting on previous knowledge, but has not yet been proven true or false

17. transformation of rocks through chemical reactions when exposed to air or water containing dissolved elements

18. used for describing two or more objects or ideas moving towards the same point or developing in the same direction

swers was compared to only one correct answer, so not only are we comparing similarities of two short texts, but also gain a direct insight into the LSA's comprehension of short text and definitions. For example, a participant with ID 4 defined the word *convergent* as a mathematical term, which is not incorrect, but it is not the meaning we were looking for, so the answer's similarity to the correct answer is rather low (Table 9). Furthermore, one of the definitions of *straightforward*[19] is an incomplete answer that has low similarity to the correct answer as well. On the other hand, some correct answers have negative cosine similarity values to the correct answer, such as *petrologist*[20] (Table 9). Answer length might explain this kind of result.

QID	**PID**	**Question**	**Cosine similarity**				**C/I**	**1–5**
			Text	**Def.**	**Answer**	**Score**		
3	3	convergent	0.0000	0.2885	0.9992	0.7354	C	4
3	4	convergent	-0.3299	0.2999	0.9992	0.7377	I	1
8	3	straightforward	-0.0291	0.0011	0.9627	0.6807	I	1

Table 9: Cosine similarity results — Examples for similarity of an answer and the correct answer; QID – question ID, PID – participant ID.

Additionally, high weights of functional words (*be, of, in, to, or, by*, etc.) might have contributed to the results. In further research, we could solve this by removing functional words with high weights from answers and definitions, and see if the results improve.

As we can see, the range of correct answers' values is greater than that of incorrect. The reason behind low values of correct answers is probably short answers evaluated as correct (Figure 8). In the five-category evaluation, completely correct answers (grade 5) have the greatest value range. Low values are mostly answers where a participant used synonymous words in their definition with respect to the correct answer (Figure 8).

The third measured value is a cosine similarity of all answers, obtained in the same manner as the similarity of all texts in the textbook. Unlike texts, the model fell short in detecting most similar answers. Similarity values

19. ccurring without obstacles or irregularities, in a simple manner
20. a scientist in the field of petrology, studying rocks and how they are formed

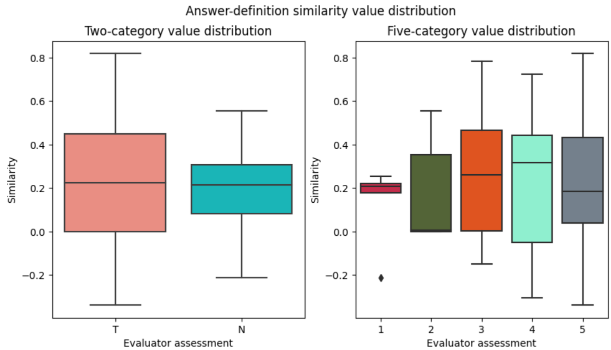

Figure 8: Answer-definition similarity value distribution.

spread from about 0.7 to 0.9. If the answer vector is 0, the first answer in the data was computed as the most similar one, with a similarity of 0 (Table 10). Answers that do not share the same terms are evaluated as most similar but also answers to the same question that do share many terms, as well as answers to different questions that share the same terms, such as answers to questions *hydrological cycle*[21] and *seabed*,[22] both containing terms *earth, ocean, surface* (Table 10).

Question a	**Answer a**	**Question b**	**Answer b**	
hydrological cycle	the hydrological cycle of the earth be the sum total of all process in which water move from the land and ocean surface to the atmosphere and back in form of precipitation	seabed	the seabed be the bottom of the ocean or the top surface of the earth in sea and ocean	0.8685
unconsolidate	unstratifie	backlash	adverse reaction to a recent development	0.0000
urbanisation	make an area more urban	urbanisation	be the process of make an area more urban	0.9840

Table 10: Most similar answers.

21. the representation of a continuous, circular movement of water through the atmosphere, where the physical state of water alters as it flows through the cycle
22. land at the bottom of the ocean

Since we only took the highest similarity score for each of the answers, the value distribution of correct and incorrect answers is nearly equal, while in the five categories, answers with grade 2 have the most scattered values (Figure 9).

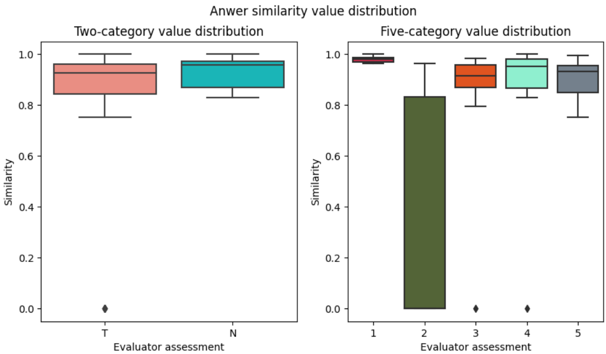

Figure 9: Answer similarity value distribution.

Finally, we calculated a final score for each answer, as a mean of the three aforementioned values. The lowest value in the final score is 0, which is the score of previously explained short answers, while long answers show little variance between the three values. In two-category distribution, the final score has lower values in correct than in incorrect values, and values of correct answers have a greater range. Answer similarity in all probability contributed to high values of incorrect answers (Figure 10). In five categories, we can see that values of answers graded 2 are most scattered, while the densest ones are those of answers with grade 1, and higher grades have relatively similar final scores (Figure 10).

Considering the displayed results, we can conclude that LSA did well in detecting topics in unit texts. Yet, when it comes to vocabulary and answers, results are not as great, especially concerning detecting latent topics in very short answers, and in answers pertaining to words that do not fall under the domain of geology. Also, there is not much difference between cosine similarities of correct and incorrect answers, which leads up to questioning our answer assessment methodology.

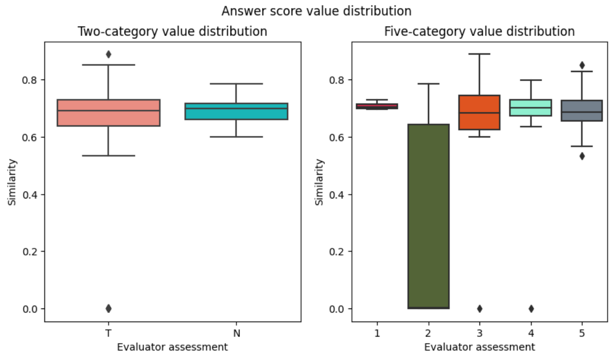

Figure 10: Final score value distribution.

6.4 Answer classification

The final step was answer classification — binary and multinomial. For classification purposes KNN algorithm was employed (Li, Yu, and Lu 2003; Peterson 2009; Chen 2018), labels corresponded with the evaluator assessment, while the classification criteria were the final answer score. Recall, precision and F_1 score were used to evaluate the KNN models (Géron 2022).

In binary classification, answers were classified as correct or incorrect. The data is comprised of 60 correct and 12 incorrect answers. Due to this discrepancy, the model classified all answers as correct (Table 11a)). Calculated model precision was 73%, recall 100%, and $F_1 = 0.85$. Given that our data set is rather small, containing only 72 observations, consequently so is the test set with mere 15 observations, the presented results do not reflect actual model validity. Since the data is randomly split into a training and test set, it can just so happen that all the observations in the test set have the same label.

The same goes for multinomial classification — category frequency is uneven, where incorrect and completely correct answers (grades 1 and 5) are least frequent, thus, we expect that this will affect the results of the classification model. Grades 3, 4 and 5 are equally represented in the data, with 17, 21 and 22 answers, respectively. As displayed in the confusion matrix, answers graded with 1 were the hardest for the model to classify. Seven out of six answers with grade 5 ended up in the test set, and our model classified 3 of them correctly (Table 11b).

Estimated model precision was 50%, recall 47%, and $F_1 = 0.46$, which is significantly worse compared to the results of binary classification. We

b) binary classification

	I	C
I	0	4
C	0	11

b) multinomial classification

	1	**2**	**3**	**4**	**5**
1	0	0	1	0	1
2	0	1	0	0	0
3	0	0	2	1	0
4	0	0	2	1	0
5	1	0	1	1	3

Table 11: confusion matrices

believe that this is a result of a small data set, uneven label representation in the training and test set, along with similar value distribution of final answer scores in different grade categories.

7 Concluding Remarks

In this paper, we discussed the application of latent semantic analysis for the assessment of short answers. In accordance with the set pedagogical goals of this paper, we extrapolated that the utilisation of flashcards for L2 vocabulary acquisition gives favourable results, particularly at the lower levels of language knowledge. As students of the Faculty of Mining and Geology come from different educational backgrounds and usually enter their studies with a low level of English, we strongly believe that using a system of flashcards that accompany the subject textbook would greatly help students to make progress faster and get to a level of vocabulary knowledge suitable for following CLIL lectures.

Reflecting on the methodological aims of the paper, we determined that developing this model helped us recognise the advantages and disadvantages of our approach. One of the greatest advantages of the model is good topic modeling of longer texts and vocabulary and answers pertaining to geology. We deem that the biggest downside is its inability to detect topics of very short answers. Supposedly, we could overcome this by expanding the data set.

In further research, we will aim to expand our data set. Our second goal is to add a system for spelling and grammar assessment. In order to improve the results obtained using LSA, we will lower the number of dimensions and try and see if the model improves. Additionally, creating separate semantic spaces for words that are not geological notions, i.e. general vocabulary, might be a good idea. When comparing answers and unit texts, we believe that we would get more meaningful results if we extract just a fragment of the text where a certain geological notion is explained or a word belonging to the general vocabulary used. Lastly, instead of computing the similarity of all answers, we would proceed to calculate the similarity of answers to the same question.

The presented model development laid a foundation for the development of a system for automatic answer assessment in digital flashcards. Comparing the goals and aims of CLIL methodology and the outcomes of using flashcards in teaching, we concluded that this technology would greatly complement the textbook in preparation, whose author is Prof. Dr Lidija Beko. Our claim is supported by the Faculty's students' positive attitude towards using digital flashcards in an L2 classroom expressed in previous research. In further research, we will aim to accomplish the project's main goal —- the development of a digital flashcard system that will be implemented in the classroom.

Acknowledgment

As this paper is a result of my Master's thesis written for an interdisciplinary master's program Social Sciences and Computing at the University of Belgrade, I would hereby like to thank my mentor, Prof. Dr. Ranka Stanović, for her patience, guidance, knowledge, and infinite support. Also, I would like to express my gratitude to Prof. Dr. Lidija Beko, for inviting me to use her textbook in preparation as the material for my thesis. The textbook will be used for courses English 1–4 at the Faculty of Mining and Geology, University of Belgrade. Finally, I want to thank Prof. Dr. Jelena Jovanović, who was a member of the committee for my thesis, because her questions contributed to improvements applied in this version of the paper.

References

Ashcroft, Robert John, Robert Cvitkovic, and Max Praver. 2018. "Digital flashcard L2 Vocabulary learning out-performs traditional flashcards at lower proficiency levels: A mixed-methods study of 139 Japanese university students." *The EuroCALL Review* 26 (1): 14–28. https://doi.org/10.4995/eurocall.2018.7881.

Averianova, Irina. 2015. "Vocabulary acquisition in L2: does CALL really help." In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy,* edited by F. Helm, L. Bradley, M. Guarda, and S. Thouësny, 30–35. https://doi.org/10.14705/rpnet.2015.000306.

Baten, Kristof, Silke Van Hiel, and Ludovic De Cuypere. 2020. "Vocabulary Development in a CLIL Context: A Comparison between French and English L2." *Studies in second language learning and teaching* 10 (2): 307–336.

Beko, Lidija. 2013. "Integrisano učenje sadržaja i jezika (CLIL) na geološkim studijama." PhD thesis, Univerzitet u Beogradu, Filološki fakultet.

Beko, Lidija, Ivan Obradović, and Ranka Stanković. 2015. "Developing Students' Mining and Geology Vocabulary Through Flashcards and L1 in the CLIL Classroom." In *The Second International Conference on Teaching English for Specific Purposes Developing students' mining and geology vocabulary through flashcards and L1 in the CLIL classroom.* Faculty of Electronic Engineering, University of Niš.

Bettig, Ronald. 2018. *Copyrighting Culture: The Political Economy of Intellectual Property.* Routledge.

Chen, Shufeng. 2018. "K-nearest neighbor algorithm optimization in text categorization." In *IOP conference series: earth and environmental science,* 108:052074. 5. IOP Publishing. https://doi.org/10.1088/1755-1315/108/5/052074.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41:391–407.

Đerić, Miloš. 2019. "Doprinos CLIL-a savremenim tokovima nastave stranog jezika." *Philologia* 17 (17): 23–38. ISSN: 1451-5342. https://doi.org/10.18485/philologia.2019.17.17.3.

Dikli, Semire. 2006. "An Overview of Automated Scoring of Essays." *Journal of Technology, Learning, and Assessment* 5 (1). https://ejournals.bc.edu/index.php/jtla/article/view/1640.

Géron, Aurélien. 2022. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media, Inc.

Godwin-Jones, Robert. 2011. "Mobile Apps for Language Learning." *Language Learning & Technology* 15 (2): 2–11. https://doi.org/10.18485/infotheca.2018.18.1.1.

Graesser, Arthur, Peter Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, Natalie Person, and Tutoring Research Group. 2000. "Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor." *Interactive Learning Environments* 8:129–148. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129.

Hung, Hsiu-Ting. 2015. "Intentional Vocabulary Learning Using Digital Flashcards." *English Language Teaching* 8:107–112. https://doi.org/10.5539/elt.v8n10p107.

Jhean-Larose, Sandra, Vincent Leclercq, Javier Diaz, Guy Denhiere, and Bernadette Bouchon-Meunier. 2010. "Knowledge evaluation based on LSA : MCQs and free answers." *Stud. Inform. Univ.* 8 (January): 57–84.

Jurafsky, Daniel, and James Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3rd ed. Draft of January 7, 2023. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Lafourcade, Mathieu, and Virginie Zampa. 2009. "PtiClic: a game for vocabulary assessment combining JeuxDeMots and LSA." In *Advances in Computational Linguistics,* vol. 41 of *Research in Computer Science,* edited by Alexander Gelbuch, 289–298. Center for Computing Research of IPN.

Landauer, Thomas, Darreil Laham, and Peter Foltz. 2003. "Automated scoring and annotation of essays with the Intelligent Essay Assessor." In *Automated essay scoring: A cross-disciplinary perspective,* edited by Mark D. Shermis and Jill C. Burstein, 87–112. Routledge.

Landauer, Thomas K, Darrell Laham, Bob Rehder, and Missy E Schreiner. 1997. "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans." In *Proceedings of the 19th annual meeting of the Cognitive Science Society,* 412–417.

Landauer, Thomas K., and Susan T. Dumais. 1997. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review* 104 (2): 211.

Landauer, Thomas K., Peter W. Foltz, and Laham Darrell. 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25 (2–3): 259–284.

Lemaire, Benoit, and Philippe Dessus. 2003. "A System To Assess The Semantic Content Of Student Essays." *Journal of Educational Computing Research* 24 (October). https://doi.org/10.2190/G649-0R9C-C021-P6X3.

Li, Baoli, Shiwen Yu, and Qin Lu. 2003. "An Improved k-Nearest Neighbor Algorithm for Text Categorization." In *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, August 2003.* https://arxiv.org/abs/cs/0306099.

Lifchitz, Alain, Sandra Jhean-Larose, and Guy Denhière. 2009. "Effect of tuned parameters on an LSA multiple choice questions answering model." *Behavior research methods* 41:1201–1209. https://doi.org/10.3758/BRM.41.4.1201.

Ma, Qing. 2009. *Second language vocabulary acquisition.* Vol. 79. Peter Lang.

Martin, Dian I., and Michael W. Berry. 2013. "Mathematical foundations behind latent semantic analysis." In *Handbook of latent semantic analysis,* edited by Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, 35–55. Psychology Press.

Nakata, Tatsuya. 2008. "English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning." *ReCALL* 20 (1): 3–20.

Nation, Paul. 2006. "Vocabulary: Second Language." In *Encyclopedia of Language Linguistics,* 448–454. ISBN: 9780080448541. https://doi.org/10.1016/B0-08-044854-2/00635-0.

Peterson, Leif E. 2009. "K-nearest neighbor." *Scholarpedia* 4 (2): 1883. https://doi.org/doi:10.4249/scholarpedia.1883.

Picca, Davide, Dominique Jaccard, and Gérald Eberlé. 2015. "Natural language processing in serious games: a state of the art." *International Journal of Serious Games* 2 (3): 77–97.

Rahutomo, Faisal, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. "Semantic cosine similarity." In *The 7th international student conference on advanced science and technology ICAST.* https://www.researchgate.net/profile/Faisal-Rahutomo/publication/262525676_Semantic_Cosine_Similarity/links/0a85e537ee3b675c1e000000/Semantic-Cosine-Similarity.pdf.

Spiri, John. 2008. "Online study of frequency list vocabulary with the WordChamp website." *Reflections on English Language Teaching* 7 (1): 21–36.

Urdan, Timothy C. 2005. *Statistics in plain English.* 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Von Ahn, Luis. 2013. "Duolingo: Learn a Language for Free While Helping to Translate the Web." In *Proceedings of the 2013 international conference on Intelligent user interfaces,* 1–2. http://dummyadress.bg.ac.uk.

Wang, Aobo, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. "Perspectives on Crowdsourcing Annotations for Natural Language Processing." *Language resources and evaluation* 47 (1): 9–31.

White, Cynthia, and Hayo Reinders. 2010. "The theory and practice of technology in materials development and task design," 58–80. Cambridge University Press.

Xodabande, Ismail, Yasaman Iravi, Behzad Mansouri, and Hoda Matinparsa. 2022. "Teaching academic words with digital flashcards: Investigating the effectiveness of mobile-assisted vocabulary learning for university students." *Frontiers in Psychology,* 2903. https://doi.org/10.3389/fpsyg.2022.893821.

Yüksel, H. Gülru, H. Güldem Mercanoğlu, and M. Betül Yılmaz. 2022. "Digital flashcards vs. wordlists for learning technical vocabulary." *Computer Assisted Language Learning* 35 (8): 2001–2017.