Automatic Student Answer Assessment Using LSA

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

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 somewhat struggles with detecting topics in short texts. After obtaining LSA semantic spaces, answer accuracy was assessed by calculating the cosine similarity between a student's answer and the golden standard. The answers were classified by accuracy using the the K-Nearest Neighbor algorithm (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.

1 Introduction

Employing technology to improve language learning outcomes is a problem scientists have wrestled with since the 1960s. In this paper, we present a beta version of a model for an automatic assessment of student answers using Latent Semantic Analysis (LSA) implemented in a use-case scenario, i.e. for assessing vocabulary knowledge of students and associates at the Faculty of Mining and Geology, University of Belgrade. In further development, the aim is for the model will be implemented within digital lexical flashcards for learning vocabulary in English as a Second Language (ESL) classes.

Previous research (Landauer et al., 1998; Lemaire and Dessus, 2003; Lifchitz et al., 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 et al., 2003; Picca et al., 2015). Flashcards have proven to be a good tool for L2 vocabulary acquisition, combining interval (Ashcroft et al., 2018) and conscious learning (Nation, 2006; Hung, 2015) —- two approaches that enhance learning outcomes, especially at the lower levels of language knowledge (Ashcroft et al., 2018). In this phase of work, we will tackle several methodological problems, such as using LSA on short text, and finding means to contribute to the digitalisation of L2 classroom at the Faculty of Mining and Geology, University of Belgrade.

The research aims to examine the current general and geological vocabulary knowledge of the Faculty's students and associates and to improve teaching methods at the Faculty by utilising Natural Language Processing (NLP). Also, we examine LSA's application in the geological domain, and on shorter text, i.e. definitions. Conforming to the aforementioned aims, our hypotheses are: (1) the creation of the system will help digitalise learning materials; (2) LSA will be successful in assessing student answers.

The paper is organised as follows: in Section 2 we will go through previos research of vocabulary acqusition and LSA implementation in educaiton technologies, proceeding to data and model description in Section 3. After that, we will analyse the results in Section 4, starting from testng LSA model validity (Section 4.1) and going thorugh topic distribution (Section 4.2), and finishing with answer assessment (Section 4.3) and classification

Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, pages 36–44
Varna, Sep 7, 2023
https://doi.org/10.26615/978-954-452-090-8_005

(Section 4.4). Finally, we will present concluding remarks in 5 and end with the limitations of our approach.

2 Related Works

In exploring L2 acquisition, vocabulary acquisition is widely researched. It is considered that vocabulary learning has the best outcomes when combining spaced (or distributed/interval) learning with explicit learning. Spaced learning is learning in many small sessions increasing the breaks between each session (Nation, 2006), while explicit learning assumes that the student is aware of the learning process (Nation, 2006; Hung, 2015; Ma, 2009). As flashcards provide simultaneous explicit and interval learning, together with learning word form, meaning and use in context (Ma, 2009), they make a great learning tool. Several researches display that flashcards significantly enhance L2 vocabulary acquisition outcomes, especially at the lower levels of language knowledge (Spiri, 2008; Nakata, 2008; Hung, 2015; Averianova, 2015; Yüksel et al., 2022). Given that students who are non-native English speakers can enter university with different levels of language knowledge, using flashcards as a teaching tool can help students reach the necessary level of English to follow classes and learning materials. Our case is no different. One of the main problems in ESL classes at the Faculty of Mining and Geology, University of Belgrade emphasised in (Beko et al., 2015) is a low level of language knowledge at the beginning of studies. Beko et al. (2015) also points out that students have in finding a suitable learning method, and lack of translation of geological terminology to Serbian, which makes translational tasks even more difficult. Our model will be monolingual, so we will not address the last-mentioned issue.

Currently, the Faculty uses a variety of language tools, a thesaurus of geological terminology in Serbian and English, comprised of roughly 2800 words (Beko et al., 2015), and a digital mining terminology platform *RudOnto*.[1] Additionally, a system of flashcards *RGF Flashcards* was developed, using *Anki* and integrated into the Faculty's *Moodle* platform.[2]

The presented system of flashcards will be tailored to the learning materials and adapted to the CLIL methodology used in the Faculty's English 1-4 subjects. CLIL integrates learning content from a certain domain with language learning (Beko, 2013; Djerić, 2019; Baten et al., 2020), whereby C1 entry language knowledge is expected. Thus, flashcards could facilitate learning for students with lower levels of English and make following of the learning materials and classes easier.

In this stage of development of the flashcards system, we aimed to create a model for an automatic assessment of the semantic similarity of student answers and the golden standard. For that purpose, we exploited LSA — 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 broadly represent human cognitive abilities, such as vocabulary acquisition, word categorisation, discourse comprehension, and essay assessment (Landauer et al., 1998). 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 et al., 2003; Dikli, 2006; Lafourcade and Zampa, 2009; Picca et al., 2015). In the light of previously said, we believe that the method is suitable for our task as well.

The idea of context-based representation of word meaning is by no means a new one in linguistic theory. Harris (1954) fist posed that elements of a language appear relative to one another. Later an automatic clustering-based algorithm for word sense disambiguation was presented by Schütze (1998), where a cluster consists of contextually similar occurrences of a word. Distributional semantics also transcended to computational linguistics, and model such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) were developed (Landauer et al., 1997; Blei et al., 2003). More recently, with the development of neural networks, models such as Word2vec have became increasingly popular in representing word meaning (Mikolov et al., 2013). However, in a case study presented in Altszyler et al. (2016), LSA outperformed Skip-gram model when the size of the corpora was reduced from medium to small.

[1] RudOnto thesaurus, accessed 20 May 2023
[2] Moodle, accessed 20 May 2023

3 Experimental Setup

3.1 Data

Data Collection The data is a mixture of long and short texts, which enabled us to compare the LSA's performance between the two. The parts of the data used are as follows: 1) **unit texts** from the English-language textbook in preparation for subjects English 1-4 at the Faculty of Mining and Geology, University of Belgrade; 2) **vocabulary** following each textbook unit - general vocabulary (663 words), geological vocabulary (280 words), minerals (18 words); 3) **participant answers** collected via Faculty's Moodle platform enhanced with HP5 extension.[3] The test was an adapted test battery presented in (Jhean-Larose et al., 2010) and was split into three groups with different examples. Some questions (e.g. question six) were adjusted to the research aims. The test was completed by 14 participants. The participants were associates from the faculty - professors and teaching assistants, with good knowledge of geological terminology in both Serbian and English. After the completion of the testing process, 451 answers were collected. For anonymity purposes, we created a unique numerical ID for each participant. After analysing the test results, and extracting only answered open-ended questions, 72 answers remained for the analysis. Some answers were omitted from the analysis due to an unclear and inconsistent output in Moodle results. The length of unit texts, vocabulary, and participant answers in tokens is 46 888, 14 051, 5505, respectively.

Selecting the Assessment Criteria First, all answers were manually checked and assessed by the evaluator. The criterion was the answers' similarity to the golden standard - a definition from the textbook vocabulary, as well as the evaluator's English language competency. Since our model does not take into account grammar and spelling, neither did the evaluator in the assessment process. However, spelling was checked and corrected using the Grammarly[4] tool prior to feeding data to the model.

The answers were graded on a scale from 1 to 5, where 1 was completely incorrect and 5 was a correct answer. Subsequently, all answers that scored 1 were labelled as incorrect, while the rest were labelled as correct. We opted for adding the two-category assessment due to the small size of the dataset because, during the classification, our model accuracy might not be well represented when classifying 72 answers into 5 categories.

Data Preprocessing Text preparation was conducted in accordance with methods found in the literature (Deerwester et al., 1990; Dikli, 2006; Lifchitz et al., 2009), which we adapted to our goals and our data. The first step in text preparation was text lemmatisation using *SpaCy* library.[5] 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 1. The examples are extracted from different texts.

3.2 Models

For developing our LSA model, as well as for the the K-Nearest Neighbor (KNN) classification algorithm, we used the *Scikit-Learn* Python library.[6]

First, we constructed a TF-IDF matrix, 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). Trying out options between 700 and 5000 terms, we opted for a 1000-dimension TF-IDF matrix for unit texts, with a minimal term frequency of 3, and a maximal frequency of 80% of documents. In this step, we also removed stop words, which were a concatenation of the NLTK[7] stop words for the English-language, and corpus-specific stop words (*km, km/h, mm, meter, yet, well*, etc.). Initially, the same TF-IDF parameters were applied to short texts as well 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 documents, while the stop word list contained only definite and indefinite article —- *a/an, the*.

Subsequently, we set the SVD parameters that were the same for all parts of the data. The number of topics was determined by examining the first 10 terms with the highest weights in order to determine an appropriate number of topics, we extracted 15 terms weights for each topic. Finally, we opted for 10 topics. Then, we assigned a name to each

[3] HP5 extension, accessed 22 May 2023
[4] Grammarly, last accessed 25 August 2023
[5] SpaCy library, accessed 22 May 2023
[6] Scikit-Learn library, accessed 22 May 2023
[7] NLTK 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

Table 1: Processed text

topic based on the first 100 terms with the highest weights. Separate semantic spaces were created for unit texts, i.e. long texts, and word definitions and participant answers, i.e. short texts. The names of the topics and their respective terms can be found in Appendix A.

After obtaining topic vectors, we measured cosine similarities between all texts, and between all the answers, and extracted the most similar ones, to check the LSA model validity for both long and short text. Next, we calculated a final score for each answer as a mean of cosine similarity of answer A and: a) vector of the unit text in which the defined term appears; b) vector of the correct answer (*golden standard*); c) vector of the previously obtained most similar answer B. The higher the similarity score of document A and document B, the higher the connection between the documents (Rahutomo et al., 2012). Finally, the answers were classified by accuracy using KNN, for both binomial (*Correct / Incorrect*) and multinomial classification (Li et al., 2003; Peterson, 2009).

4 Results and Analysis

4.1 Testing Model Validity

In order to check LSA validity for long text, we computed cosine similarity between all unit texts, and then detected the most similar ones. For short text, we did the same with participant answers.

Analysing the results, the supposition is that latent topics in unit texts are well-detected and that the most similar texts indeed convey similar topics. This has proven to be true, so the text about Wagner's hypothesis which 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 2).

The similarity between answers spreads from about 0.7 to 0.9. Unlike with unit text, the model was somewhat inconsistent with detecting the most similar answers, for example, answers that do not share the same terms were evaluated as most similar. However, so were the 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*: *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* and *seabed*: *land at the bottom of the ocean* both containing terms *earth, ocean, surface* (Table 3).

Based on these results, we can argue that our model did better in detecting topics in longer texts than in short ones.

4.2 Topic Distribution

The highest standard deviation of topics was observed in unit texts, while it was somewhat lower in vocabulary and answers. We believe that the reason behind the lower standard deviation 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*. 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* has high values in *EarhFormation, Weathering* and *Landslides*, the model failed to recognise latent topics in others, which is shown for exam-

Text A	Text B	Similarity
palaeozoic era	mesozoic era and cenozoic	0.9776
wegener s hypothesis	tectonic plates	0.8980
volcanoes	igneous rocks	0.8229
the causes of metamorphism	metamorphic textures	0.9606
coal as a fossil fuel	oil and natural gas mineral oil	0.7726

Table 2: Examples of the most similar texts

Text A	Text B	Similarity
hydrological cycle	seabed	0.8685
unconsolidate	backlash	0.0000
urbanisation	urbanisation	0.9840

Table 3: Examples of the most similar answers

ple in a low value of topic *Fossils* in definitions of terms *fossil, fossilised, fossilisation*. In participant answers, we find that the 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* have the highest values for the topic *EarthFormatoin*, and the lowest for *Erosion* and *Landslides*. All topic values for short, incomplete answers, consisting of just 1 or 2 words, are 0.

In all parts of the data, topic *EarthFormation* is the most frequent one, appearing in 34 out of 36 unit texts, and in most definitions and answers. The high frequency of this topic does not come as a surprise, as it contains vocabulary that is woven through most of the texts. Other frequent topics include *Volcanology*, *Weathering*, and *Landslides*, while the least frequent ones are *Minerals* and *Dating*. Relative frequencies of all topic, as well as values of the dominant topic for the first 30 data points in all parts of the data are displayed in Figure 1.

4.3 Answer Assessment

After analysing topic distribution, we proceed to assess the participant answers, by the criteria explained in 3.2. The lowest value in the final score is 0, which is the score of previously explained very short answers (1-2 words), while long answers show little variance between the three values used for computing the final score.

As displayed in Table 4, some correct answers have lower similarity with the corresponding unit texts than incorrect answers, and the final score of correct and incorrect answers is relatively similar. This raises the question if our method should be revised. To a human evaluator, a similarity difference of 0.1 might be significant when they look at the broader picture, but we will see if that will be the case with our classification model as well.

In a two-category distribution, the final score has lower values in correct than in incorrect answers, and values of correct answers have a greater range. In the five-grade answer assessment, 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.

Cosine similarity of the most similar answers probability greatly contributed to high values of incorrect answers, since only the highest similarity values were taken into account. Furthermore, answer accuracy was best represented by the cosine similarity between an answer and a golden standard (correct answer). Nonetheless, we opted for keeping the previously determined final score computation, because we wanted to see how the model does in comparing long and short text. Another reason behind this is that the description of a term in unit text and its respective definition in the unit vocabulary can slightly differ, f.x. the description of a geological term in the unit text could be longer, or synonyms can be used. Since synonyms will rarely appear together, we thought this might be a way to overcome this obstacle. In further research, extracting only the sentences of unit texts actually explaining a certain geological notion might be the solution. Additionally, high weights of functional

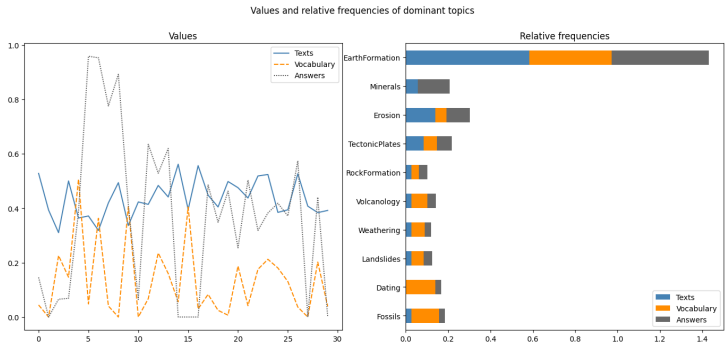

Figure 1: Dominant topics values in vocabulary

QID	PID	Answer A	Text	Def.	Answer B	Final Score	C/I	1-5
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 4: Final answer score; QID – question ID, participant ID, C/I – Correct/Incorrect, 1-5 – grade on a scale from 1 to 5

words in the semantic space of definitions and answer topics (*be, of, in, to, or, by*, etc.) might have contributed to the results. In future work, we could solve this by removing functional words with high weights from definitions and answers, and see if the results improve.

4.4 Answer Classification

The final step was a multinomial and binary answer classification. For classification purposes KNN algorithm was employed (Li et al., 2003; Peterson, 2009; Chen, 2018), labels corresponded with the evaluator assessment, while the classification criteria was the final answer score. Recall, precision and F_1 score were used for evaluation (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. Calculated model precision was 73%, recall 100%, and $F_1 = 0.85$. The size of our data might have affected the performance of the classification algorithm, given that our data set is rather small, containing only 72 observations, consequently so is the test set with mere 15 observations. 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.

In the multinomial classification, category frequency is uneven. Consequently, the model did poorly in classifying the underrepresented categories, i.e. grades 1 (incorrect) and 5 (completely correct). For the multinomial classification, model parameters are as follows: precision was 50%, recall 47% and $F_1 = 0.46$. Since the model was classifying mere 15 answers into 5 categories with uneven distribution in the data, it is expected that the results are worse than those of binary classification.

5 Conclusion

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 ad-

vantages of the model is good topic modelling 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.

To overcome model downsides, the first step in further research would be to expand the answer 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 could try and lower the number of dimensions. 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. Our claim is supported by the Faculty's students' positive attitude towards using digital flashcards in an L2 classroom expressed in previous research. Ultimately, we intend to accomplish the project's main goal —– the development of a digital flashcard system that will be implemented in the classroom.

Limitations

The main limitation of work presented in this paper is a small data set. Not only did scarce data made it more difficult to find the right parameters for creating semantic spaces, but it also hindered the classification task. Additionally, feature extraction in short texts, i.e. definitions and answers, should be revised. By removing just articles, we left too much noise in these parts of the data, which resulted in topics having similar terms with the highest weights. Methodologically, the biggest downside, in our opinion, is a lack of demographic questionnaire, where the participants would fill out their English language levels, by either self-evaluation, or state if they possess an English language certificate, as well as their age, gender and professional qualification. This should be included in further research. Having the level of participants knowledge would have provided us with an additional criteria for LSA assessment, but also help us make conclusions on the needs of our user target group.

Acknowledgements

This paper is a result of Master thesis at the Social Sciences and Computing MSc program at the University of Belgrade. The author would hereby like to thank their mentor prof. Dr Ranka Stanković for her help and advice, as well as prof. Dr Lidija Beko, for landing her textbook in preparation for the purposes of this work.

References

Edgar Altszyler, Mariano Sigman, Sidarta Ribeiro, and Diego Fernández Slezak. 2016. Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Robert John Ashcroft, 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.

Irina Averianova. 2015. Vocabulary acquisition in l2: does call really help. In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, pages 30–35.

Kristof Baten, 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.

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

Lidija Beko, 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.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Shufeng Chen. 2018. K-nearest neighbor algorithm optimization in text categorization. In *IOP conference series: earth and environmental science*, volume 108, page 052074. IOP Publishing.

Scott C. Deerwester, 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.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).

Miloš Djerić. 2019. Doprinos clil-a savremenim tokovima nastave stranog jezika. *Philologia*, 17(17):23–38. 3.

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

Arthur Graesser, 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.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hsiu-Ting Hung. 2015. Intentional vocabulary learning using digital flashcards. *English Language Teaching*, 8:107–112.

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

Daniel Jurafsky and James Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3 edition, volume 2. Draft of January 7, 2023.

Mathieu Lafourcade and Virginie Zampa. 2009. Pticlic: a game for vocabulary assessment combining jeuxdemots and lsa. In Alexander Gelbuch, editor, *Advances in Computational Linguistics*, volume 41, pages 289–298. Center for Computing Research of IPN.

Thomas Landauer, Darreil Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. In Mark D. Shermis and Jill C. Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112. Routledge.

Thomas K. Landauer, Peter W. Foltz, and Laham Darrell. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.

Thomas K Landauer, 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*, pages 412–417.

Benoit Lemaire and Philippe Dessus. 2003. A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24.

Baoli Li, Shiwen Yu, and Qin Lu. 2003.

Alain Lifchitz, 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.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Tatsuya Nakata. 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.

Paul Nation. 2006. *Vocabulary: Second Language*, pages 448–454.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Davide Picca, 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.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

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

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

Appendices

A Appendix

Topic	Name	Terms with the highest weights
Topic0	Earth Formation	mineral, cycle, earth, deposit, flow, sedimentary, igneous, material, soil, metamorphic
Topic1	Minerals	mineral, metamorphism, grain, metamorphic, igneous, metamorphic rock, pressure, crystal, magma, ore
Topic2	Erosion	flow, soil, particle, stream, slope, erosion, debris, landslide, glacial, material
Topic3	Tectonic Plates	plate, earthquake, wave, cycle, tectonic, magma, continental, oceanic, magnetic, magnetic field
Topic4	Rock Formation	sedimentary, cycle, sediment, metamorphic, igneous, sedimentary rock, strata metamorphic rock, metamorphism, erosion
Topic5	Volcanology	magma, lava, grain, volcano, slope, eruption, volcanic, viscosity, period, landslide
Topic6	Weathering	wave, earthquake, magnetic, date, particle, magnetic field, metamorphism, stress, erosion, sediment
Topic7	Landslides	slope, landslide, soil, debris, hazard, cycle, trigger, activity, fall, downslope
Topic8	Dating	earth, strata, magma, date, age, eruption, lava, idea, satellite, remote
Topic9	Fossils	oil, wave, earthquake, coal, trap, organic, sedimentary, sedimentary rock, weathering, carbon