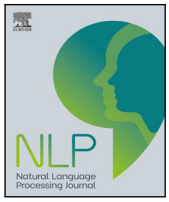
Natural Language Processing Journal 8 (2024) 100094

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



Automatic language ability assessment method based on natural language processing

Nonso Nnamoko [a,*], Themis Karaminis [b], Jack Procter [c], Joseph Barrowclough [a], Ioannis Korkontzelos [a]

[a] Department of Computer Science, Edge Hill University, St Helens Road, Ormskirk, L39 4QP, United Kingdom
[b] Department of Psychology, City University of London, Northampton Square, London, EC1V 0HB, United Kingdom
[c] Experimental Psychology Department, University College London, Gower Street, London, WC1E 6BT, United Kingdom

ARTICLE INFO

Keywords:
Cognitive assessment
Natural Language Processing
Language ability test
Cosine similarity
WASI-II
Word embedding

ABSTRACT

Background and Objectives: The Wechsler Abbreviated Scales of Intelligence second edition (WASI-II) is a standardised assessment tool that is widely used to assess cognitive ability in clinical, research, and educational settings. In one of the components of this assessment, referred to as the Vocabulary task, the assessed individuals are presented with words (called stimulus items), and asked to explain what each word mean. Their responses are hand-scored based on a list of pre-rated sample responses [0-Point (poor), 1-Point (moderate), or 2-Point (excellent)] that is provided in the accompanying manual of WASI-II. This scoring method is time-consuming, and scoring of responses that do not fully match the pre-rated ones may vary between individual scorers. In this study, we aim to use natural language processing techniques to automate the scoring procedure and make it more time-efficient and reliable (objective).

Methods: Utilising five different word embeddings (Word2vec, Global Vectors, Bidirectional Encoder Representations from Transformers, Generative Pre-trained Transformer 2, and Embeddings from Language Model), we transformed stimulus items and pre-rated responses from the WASI-II Vocabulary task into machine-readable vectors. We measured distance with cosine similarity, evaluating each model against a rational-expectations hypothesis that vector representations for stimuli should align closely with 2-Point responses and diverge from 0-Point responses. Assessment involved frequency of consistent representation and the Pearson correlation coefficient, examining overall consistency with the manual's ranking across all items and sample responses.

Results: The Word2vec model showed the highest consistency with the WASI-II manual (frequency = 20 out of 27; Pearson Correlation coefficient = 0.61) while Bidirectional Encoder Representations from Transformers was the worst performing model (frequency = 5; Pearson Correlation coefficient = 0.05). The consistency of these two models with the WASI-II manual differed significantly, $Z = 2.282$, $p = 0.022$.

Conclusions: Our results showed that the scoring of the WASI-II Vocabulary task can be automated with moderate accuracy relying upon off-the-shelf embedding models. These results are promising, and could be improved further by considering alternative vector dimensions, similarity metrics, and data preprocessing techniques to those used in this study.

1. Introduction

There is often a need to assess individuals' cognitive ability – verbal and non-verbal – in educational, research and clinical settings. The Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) (Wechsler, 2011) is a standardised assessment of cognitive ability that is widely used for this purpose and it is appropriate for individuals aged 6–90 years old. WASI-II which is a shorter version of the more comprehensive Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V) (Wechsler, 2014) assessments, consists of four sub-tests (Vocabulary, Similarities, Block Design, and Matrix Reasoning). The four sub-tests are typically administered by a trained expert and used to calculate two scores commonly known as Full-Scale intelligence quotient estimates (FSIQ-4 and FSIQ-2). The FSIQ-4 is deduced from individual scores of all four sub-tests, while the FSIQ-2 uses scores from only Vocabulary and Matrix Reasoning and is recommended when shorter administration time is warranted. In addition, the WASI-II produces two composite scores, namely: a Verbal Comprehension Index (VCI)

* Corresponding author.
E-mail address: nnamokon@edgehill.ac.uk (N. Nnamoko).
URL: https://research.edgehill.ac.uk/en/persons/nonso-nnamoko (N. Nnamoko).

https://doi.org/10.1016/j.nlp.2024.100094
Received 1 January 2024; Received in revised form 14 July 2024; Accepted 25 July 2024

consisting of scores from Vocabulary and Similarities; and a Perceptual Reasoning Index (PRI) containing scores from Block Design and Matrix Reasoning. The full WASI-II test kit includes 25 record forms, a stimulus book, nine block design cubes and an examiner's manual (McCrimmon and Smith, 2013) which are used to administer all four sub-tests.

Several studies have shown the validity of WASI II and examined its psychometric properties (Axelrod, 2002; Gontkovsky, 2017; Hasson et al., 2019; McGeehan et al., 2017; Sharratt et al., 2020; Irby and Floyd, 2013; McCrimmon and Smith, 2013). This study focuses only on the Vocabulary sub-test of WASI-II. In this sub-test, the individual who is assessed is presented with a list of words. These words (henceforth stimuli) are presented one by one, and for each stimulus, the assessed individual is asked to explain what the word means. The responses are recorded in a specialised record form by a trained expert, who administers the Vocabulary sub-test and scores these as 0-Point (poor), 1-Point (moderate), or 2-Point (excellent). The scoring process is based on the examiner's manual which accompanies WASI-II and provides a list of alternative potential responses for each stimulus, pre-rated as 0-Point, 1-Point or 2-Point responses. The quality of the responses provided by an individual also informs the flow and the length of the Vocabulary sub-test. In particular, the assessed individual might be asked to offer further detail after some 0-point responses, while the administration should stop after two consecutive erroneous responses.

Currently, the WASI-II administration and scoring procedures are undertaken manually. This implies that the trained expert who administers the Vocabulary task to the individual being assessed should (at the same time) record their responses, interpret meaning and score them in order to align with the administration guidelines of the Vocabulary sub-test. Furthermore, the administrator will typically spend time at the end of the assessment for a more fine-grained ranking of an individual's responses. Overall, the manual procedures for administering and coding (i.e., interpreting and scoring) the responses in the Vocabulary sub-test are complex, tedious, time-consuming, and expensive. Furthermore, as the responses in this sub-test are open-ended, coding can be prone to errors and subjective biases, especially when multiple administrators are involved.

In this study, we propose to use natural language processing (NLP) techniques to automate the scoring procedure of the Vocabulary sub-test and make this more time-efficient as well as reliable and objective. The NLP methods implemented in this study are largely based on word embeddings or word vectors (Mikolov et al., 2013), a widely-used resource from the fields of computational linguistics and machine learning. Broadly speaking, word embeddings enable the representation of meanings of words as vectors in a multi-dimensional Euclidean space in a way that captures semantic regularities and relationships between words (Levy and Goldberg, 2014b). For example, *cup* would fall close to *mug* and to *tea*. Furthermore, subtle semantic relationships between words can be described in linear algebra terms. A widely used example is that $woman - man = queen - king = aunt - uncle$; and thus $queen = king - man + woman$. These strengths of word embeddings have supported several recent advances in computational linguistics applications, for example, machine translation (Garcia et al., 2015), information retrieval (Roy, 2017), and question answering (Medveď and Horák, 2018).

Earlier studies have demonstrated the psycholinguistic plausibility of word-embedding models. For example, Mandera et al. (2017) showed that word embeddings account for human performance in a range of psycholinguistic tasks including vocabulary knowledge, semantic/relatedness ratings, semantic priming, and association norms. Paetzold and Specia (2016) developed a bootstrapping algorithm that employed word embeddings to infer four psycholinguistic properties of words including familiarity, age of acquisition, concreteness, and imagery. Relatedly, other recent studies used word embeddings to examine language use in autistic and typical children. For example, Prud'hommeaux et al. (2017) analysed children's responses in a semantic-fluency task (i.e., producing as many words as possible from a given category); while Goodkind et al. (2018) compared children's performance in the Autism Diagnostic Observation Schedule-2 (ADOS-2) communication assessment (Lord et al., 2012).

More recently Pérez et al. (2022) applied NLP techniques on both the 2019 and 2020 versions of the eRisk corpora (Losada et al., 2019, 2020) to predict depression severity. Vu et al. (2020) focuses on predicting responses to psychological questionnaire from social media participants. Sonabend et al. (2020) tried to derive dimensional measures of psychiatric symptoms, while Wawer and Chojnicka (2022) tried to detect ASD from picture book narratives within ADOS-2. Of the existing studies examined, only Wawer and Chojnicka (2022) and Goodkind et al. (2018) tried to encode language ability test on a standardised questionnaire (i.e., ADOS-2) which is similar to WASI II. However, ADOS-2 and other similar instruments for accessing language ability are proprietary (including WASI-II), thus limiting direct comparison. As a compromise, we applied the NLP methodologies found in these studies to the WASI II assessment tool to evaluated their usefulness in automating the scoring of the Vocabulary sub-test.

Specifically, this study assessed the extent to which individual word-embedding models support consistency with the gold-standard scoring scheme for the Vocabulary sub-test of WASI II. We considered five word embedding models, namely Word2vec (Mikolov et al., 2013), GloVe: Global Vectors (Pennington et al., 2014), BERT: Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), GPT-2: Generative Pre-trained Transformer 2 (Radford et al., 2019) and ELMo: Embeddings from Language Model (Peters et al., 2018). For all these models, we represented the alternative potential responses as vectors (points) in the word-embedding multidimensional spaces. For responses that are 'phrases' (rather than single words), we applied NLP techniques that enable representing phrases and/or sentences into a single vector (commonly known as document embedding). Subsequently, we evaluated whether the patterns for the closeness of potential response(s) to the corresponding stimulus are consistent with the gold standard scheme of the WASI-II scoring manual. Our rational-expectation hypothesis is that the stimulus vector would be closest in the multi-dimensional meaning space to 2-Point responses and farthest from 0-Point responses.

We used cosine similarity metric (Li et al., 2004; Li and Han, 2013) to evaluate the closeness in meaning between each stimulus and the corresponding alternative potential responses in the scoring manual. Several text preprocessing techniques were employed to reduce noise during experiments such as stop-word removal for filtering function words (Luhn, 1960), spelling/bias correction, tokenisation and word inflection. We also employed document centroid vector technique (Rossiello et al., 2017) for computing document embedding (Palachy, 2019) and explored the performance of Term Frequency Inverse Document Frequency (TF-IDF) for evaluating word relevance in a collection (Rajaraman and Ullman, 2011). The overarching goal of our experiments is to address the following research questions:

RQ1: How does text preprocessing influence performance of the proposed WASI-II automation approach?

RQ2: How does TF-IDF as a weighting factor influence the performance of the proposed WASI-II automation approach?

RQ3: Which vector representation model (i.e., Word2vec, GloVe, BERT, GPT2 or ELMo) produces optimal result on the experimental data (with and/or without) text preprocessing and TF-IDF?

By exploring these questions, the study makes the following contributions to the domain of psycholinguistic plausibility with NLP techniques:

- An attempt to automate the WASI-II questionnaire. To the best of our knowledge, no other study has done this.

- Performance evaluation of standard embeddings to establish the most effective in representing the intended meanings of the WASI-II stimulus items.
- Performance evaluation of TF-IDF as a weighting factor on the word or document embeddings generated from WASI-II response set and subsequent relationship mapping with the stimulus items.

The rest of the paper is organised as follows. Section 2, provides the details about related work and the necessary background for the techniques and tools commonly used in automating the analysis of qualitative data. A detailed description of the methods and materials used for experiments are presented in Section 3, including the experimental data and methodology approaches, details about the experiment setup and evaluation measures used. The findings are discussed in Section 4, with further discussions about issues likely to threaten the validity of results presented in Section 5. Section 6 summarises the study and points out future work.

2. Background and related work

Generally, researchers have explored use-cases, challenges, and approaches of NLP on linguistic tasks in many domains including industrial logistics organisation (Garg et al., 2021), analysis of social media posts (Onikoyi et al., 2023), software engineering (Nnamoko et al., 2019), and more relevant to our study, in education (Botelho et al., 2023; Hwang and Kim, 2022; Shipurkar et al., 2022; Lam and Nnamoko, 2024). For example, Shipurkar et al. (2022) used a conjunction of a page-to-word segmentation algorithm, a convolutional neural network (CNN) and bi-directional long short-term memory (BLSTM) network for recognition of handwritten text assignment documents. The method which provides similarity scores between documents produced validation accuracy of 82.10% and was subsequently implemented as plagiarism detection module.

Hwang and Kim (2022) investigated the impact of constructional diversity on second language writing proficiency using the Constructional Diversity Analyzer (CDA) on 3,284 essays. Results from regression analysis showed that higher diversity in constructions correlated with better writing proficiency. Additionally, less frequent and more complex constructions contributed significantly to proficiency levels. These studies were made possible due to advances in word embeddings to translate natural language into machine readable vectors.

From a theoretical perspective, word embeddings correspond to a seminal approach to meaning, commonly referred to as distributional semantics (Firth, 1957; Harris, 1954; Mcdonald and Ramscar, 2001; Sahlgren, 2008). The main tenet of distributional semantics is that words appearing in similar contexts have similar meanings, and thus semantic relationships can be accounted for by patterns of co-occurrence in text corpora. Earlier computational models of distributional semantics, e.g., Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) and Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996), relied on explicit counts of word co-occurrences in sizeable text corpora (Baroni et al., 2014; Mandera et al., 2017). In contrast, many word embedding models are based on prediction (Baroni et al., 2014; Mandera et al., 2017). These are typically derived from artificial neural networks, i.e., computational learning systems that are loosely based on principles of computation in the brain, that are exposed to text corpora to learn how to predict a word given its context or viceversa (Levy and Goldberg, 2014a). Arguably, in addition to being grounded on the influential distributional semantics approach, such word embedding models incorporate artificial processes which present analogies to principles of human language learning (Levy and Goldberg, 2014a), as well as the highly-influential framework of predictive learning (Clark, 2013); including its extension to language processing (Kuperberg and Jaeger, 2015) and language acquisition (Zettersten, 2019).

The psycholinguistic plausibility of word-embedding models has been demonstrated in many studies (Mandera et al., 2017; Paetzold and Specia, 2016; Pérez et al., 2022; Vu et al., 2020; Wang et al., 2019; Wawer and Chojnicka, 2022; Shahamiri and Thabtah, 2020; Sonabend et al., 2020). These studies converge on a central goal to enhance and validate advanced computational techniques in psychological assessments. Each study addresses this aim by investigating specific research questions or hypotheses related to individual behaviour patterns on a set task and their implications. For instance, Pérez et al. (2022) applied NLP techniques on both the 2019 and 2020 versions of the eRisk corpora (Losada et al., 2019, 2020) to predict depression severity. Vu et al. (2020) focuses on predicting responses to psychological questionnaire from social media participants using BERT embeddings. Sonabend et al. (2020) tried to derive dimensional measures of psychiatric symptoms using NLP and word embeddings. Meanwhile, other studies concentrate on developing and validating deep learning algorithms for ASD screening using heterogeneous data sources (Wang et al., 2019) or evaluating the effectiveness of deep neural networks for ASD detection from textual narratives (Wawer and Chojnicka, 2022); and Shahamiri and Thabtah (2020) tried to create intelligent ASD screening systems using Convolutional Neural Networks (CNNs) for improved diagnostic accuracy and accessibility. Collectively, these studies emphasise the integration of machine/deep learning and NLP techniques to enhance the accuracy, efficiency, and applicability of psychological and medical assessments.

These sophisticated techniques are applied to diverse datasets, including social media posts and responses to psychological questionnaires (Vu et al., 2020), historical ASD cases (Shahamiri and Thabtah, 2020), and self-report symptom measures from a cellular biobanking study (Sonabend et al., 2020). The methodologies employed exhibits significant commonalities as most studies leverage deep learning techniques, ranging from deep embedding representations for categorical variables (Wang et al., 2019) to CNNs integrated into mobile apps (Shahamiri and Thabtah, 2020). Text encoders like BERT embeddings (Vu et al., 2020), Word2Vec (Pérez et al., 2022), ELMo and USE, combined with classification algorithms (Wawer and Chojnicka, 2022) are also used to transform and analyse textual data.

The results obtained across these studies demonstrate the efficacy of advanced computational methods particularly NLP in behavioural and psychological assessments. For example, Wang et al. (2019) achieved sensitivity and specificity rates of 99% with deep embedding representation learning for ASD. Generally, CNN-based systems showed higher accuracy, sensitivity, and specificity compared to traditional methods, emphasising their diagnostic potential (Shahamiri and Thabtah, 2020). Similarly, solutions using embeddings significantly outperformed traditional methods as evidence by Vu et al. (2020) who used BERT to predict questionnaire responses. In fact, Sonabend et al. (2020) used embedding-based measures effectively to distinguish psychiatric disorders with high accuracy and AUC.

Despite the high performances, these studies are focused on diagnosing psychological conditions with single modality like measuring psycholinguistic patterns in text. This is impractical because behavioural patterns from this single modality represents only one of the many indicators that can be used to make informed conclusions about a diagnosis for conditions like ASD. Only a few studies employed word embeddings specifically for characterising language use in typically developing and autistic children. For example, Prud'hommeaux et al. (2017) analysed responses of typical and autistic children in a semantic-fluency task (i.e., producing as many words as possible from a given category). The computational analysis suggested that although the two groups of children performed similarly in terms of the sheer number of responses, they differed in the cognitive mechanisms they used to carry out the task (the typical children employed longer 'semantic chains' in their responses). Goodkind et al. (2018) compared the responses of autistic and typical children in the Autism Diagnostic Observation Schedule-2 (ADOS-2) (Lord et al., 2012), a widely used assessment of autistic symptomatology. They found that when mapped onto a multi-dimensional word-embedding space, the responses of autistic children

were more dispersed than the responses of autistic children, which were more uniform.

The findings from Prud'hommeaux et al. (2017), Goodkind et al. (2018) and Wawer and Chojnicka (2022) have practical implications for the automatic language ability screening task presented in our study. These studies applied various embeddings to evaluate language ability, especially Wawer and Chojnicka (2022) and Goodkind et al. (2018) who used a standardised questionnaire (i.e., ADOS-2) which is similar to the WASI II. Thus, the research presented in this paper explored all the embeddings observed in similar studies albeit only on WASI II due to restricted access to other standardised questionnaires.

2.1. Open-ended responses

Both the Vocabulary sub-test of WASI-II, and the ADOS-2 assessment referred to in the previous subsection, include open-ended questions. In open-ended questions, respondents formulate a response in their own words and express this verbally (or in writing) without being steered in a particular direction by predefined response categories. Open-ended questions are thought to be closely aligned with human nature as people communicate freely in everyday life by speaking (or writing) (Roberts et al., 2014). Furthermore, open-ended questions are thought to be more suitable than closed-ended (e.g., yes/no or multiple choice) questions for measuring knowledge as they yield more rich information, whilst also minimising the chance that respondents (will try to) guess the right answer (Krosnick and Presser, 2009).

However, it can be challenging to collect and analyse data from open-ended questions. The 'difficulty in the coding analysis of the responses' is commonly cited as a reason for which open-ended questions are rarely used by researchers (Schuman and Presser, 1981; Roberts et al., 2014). Similar limitations apply to the Vocabulary sub-test (and more generally, the WASI-II), which currently supports only manual administration and manual coding based on the accompanying assessment manual. These limitations translate into costs in time and effort as well as risks for errors and subjective biases (which may require double-coding to mitigate risks to reliability).

2.2. Methods for the analysis of open-ended responses

A variety of methods exist for analysing and coding open-ended responses. Among these methods, quantitative content coding (Züll, 2016) is a commonly used approach, which is also pertinent to the approach implemented in the WASI-II. In quantitative content coding, one or more individuals code the open responses on the basis of a predefined categorisation scheme. The process typically begins with the development of themes that describes the relevant coding categories. The themes could be either flat-framed, i.e., all codes are of the same specificity and importance; or hierarchical-framed implying a taxonomy of how the codes relate to one another (Nowell et al., 2017). The taxonomy approach enables the consideration of different levels of granularity during the coding and the analysis of the results. In other words, an ontology shows the properties of the subject area and how they are related, by defining a set of concepts and categories that represent the subject. Each category in the categorisation scheme is assigned a label and a category number, followed by a category definition and examples. Fig. 1 shows a simple example of such categorisation scheme for coding an open-ended question about respondents' associations with 'the meaning of **lamp**'.

An alternative to the categorisation scheme is the computer-assisted content coding which involves scoring the responses on the basis of a dictionary that has the same function as the categorisation scheme in the manual quantitative content coding discussed above. In this case, the coding rules are based on a lists of words defined such that they unequivocally indicates a particular category, instead of a verbal/theoretical definition of the categories. Whenever these words or phrases appear in a response, the corresponding code is assigned.

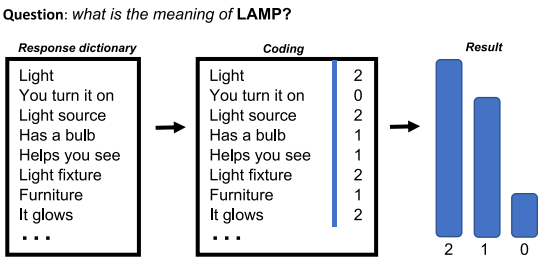


Fig. 1. A simple qualitative content coding example.

This approach allows for automation through suitable text analytics software that supports dictionary development and automatic coding such as ATLAS.ti (Paulus and Lester, 2016), Thematic,[1] WORDSTAT,[2] QDA Miner Lite,[3] RapidMiner[4] and MAXDictio, an add-on module of MAXQDA.[5] The advantage of computer-assisted content analysis is that large volumes of data can be coded quickly and reliably. However, the effort involved in the definition and validation of suitable lists of words should not be underestimated.

Another approach to coding and analysis of open-ended questions is the semi-automatic coding (Züll, 2016). This could be either through 'supervised machine learning' approach in which textual responses are automatically coded on the basis of a manually coded text sample (i.e., a training set of answers) (Giorgetti and Sebastiani, 2003); or topic modelling approaches proposed by Roberts et al. (2014). These approaches not only offer a unique blend of the strengths of both quantitative content and computer-assisted content coding, but also the coding rules are automatically formulated without explicit definition and validation. However, they present the limitation that they require large volumes of data for optimum performance.

There are a number of different procedures for direct analysis of open-ended questions without assigning one or more codes to each individual response. One example is conceptual mapping, a procedure proposed by Jackson and Trochim (2002). Another example proposed by Kronberger and Wagner (2000) is co-occurrence analysis, which focuses on words that occur together within a response. This co-occurrence forms the basis of the analysis as the generated similarity or distance matrix is further analysed, for example by subjecting it to cluster, correspondence, or multidimensional scaling analysis. Tools such as TLab[6] and Alceste[7] are suitable for this type of analysis. It is based on a similar co-occurrence analysis using similarity or distance matrix that we proposed to automate WASI-II Vocabulary sub-test.

3. Materials and methods

This section presents the experimental method and materials, including details of the experimental dataset and ethical considerations, data pre-processing techniques considered, word embeddings applied, and the detailed experimental setup to enable reproducibility of experiments and results.

3.1. Experimental data and ethics

We used the question and response data from the Vocabulary sub-test of WASI-II questionnaire (Wechsler, 2011). The Vocabulary sub-test consists of 31 items including:

[1] https://getthematic.com/
[2] http://provalisresearch.com/products/content-analysis-software/
[3] https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/
[4] https://rapidminer.com/
[5] https://maxqda.de
[6] https://www.tlab.it/
[7] https://www.image-zafar.com/Logicieluk.html

Table 1
Sentence and word count statistics of original dataset.

Stimulus	Sentences	Words	Min| Max| Avg Sentence per point	Min| Max| Avg Words per point
shirt	23	63	5 | 10 | 7.67	1 | 8 | 2.74
car	35	83	1 | 27 | 8.75	0 | 5 | 2.37
lamp	43	164	6 | 19 | 14.33	1 | 7 | 3.81
bird	28	78	5 | 12 | 9.33	1 | 6 | 2.79
tongue	29	106	6 | 14 | 9.67	1 | 10 | 3.66
pet	42	140	5 | 25 | 14.00	1 | 10 | 3.33
lunch	30	111	9 | 11 | 10.00	1 | 9 | 3.70
bell	49	191	6 | 25 | 16.33	1 | 8 | 3.90
calendar	65	263	2 | 36 | 21.67	1 | 8 | 4.05
alligator	48	217	10 | 19 | 16.00	1 | 9 | 4.52
dance	47	136	8 | 27 | 15.67	1 | 8 | 2.89
summer	31	116	6 | 14 | 10.33	1 | 9 | 3.74
reveal	23	61	2 | 12 | 7.67	1 | 8 | 2.65
decade	19	45	4 | 10 | 6.33	1 | 4 | 2.37
entertain	37	119	4 | 19 | 12.33	1 | 7 | 3.22
tradition	60	257	10 | 30 | 20.00	1 | 10 | 4.28
enthusiastic	58	153	7 | 32 | 19.33	1 | 11 | 2.64
improvise	31	115	5 | 15 | 10.33	1 | 10 | 3.21
haste	29	79	2 | 20 | 9.67	1 | 5 | 2.72
trend	33	101	9 | 14 | 11.00	1 | 6 | 3.06
impulse	39	102	6 | 26 | 13.00	1 | 7 | 2.62
ruminate	25	70	2 | 13 | 8.33	1 | 8 | 2.80
mollify	23	50	3 | 12 | 7.67	1 | 5 | 2.17
extirpate	24	45	5 | 10 | 8.00	1 | 5 | 1.83
panacea	23	63	3 | 12 | 7.67	1 | 5 | 2.74
perfunctory	25	46	6 | 10 | 8.33	1 | 5 | 1.84
insipid	17	27	3 | 7 | 5.67	1 | 5 | 1.59
pavid			Not considered	

Note The **car** stimuli does not have 1-point response, hence the '0' in **Min| Max| Avg Words per point**. The **pavid** stimuli was not considered because the word was not found in pre-trained word embedding.

Table 2
Item 8 from WASI-II Vocabulary sub-test.

Word	Points	Responses
TONGUE	2-Point	Organ; Body part; Part of body; Muscle for (tasting, eating, talking); The strip under the laces of your shoe; It is in your mouth and (you taste with it, has taste buds)
	1-Point	(Helps you, Use it) to (talk, eat, taste, swallow) **(Q)**; In your mouth **(Q)**; Has taste buds **(Q)**; Muscle **(Q)**; On (my, your) face **(Q)**; Put food on it **(Q)**
	0-Point	You (move, brush) it **(Q)**; [Points to tongue] **(Q)**; (Part of, On) your shoe **(Q)**; Red; Bumpy

Note In the table, **(Q)** means that additional query (provided) can be used for marginal, generic or functional responses including hand gesture. Text in parenthesis $(t_1, t_2, \ldots t_n)$ were extrapolated.

- 3 *picture stimuli* for which the examinee is required to name the object presented visually, and
- 28 *verbal stimuli* for which the examinee is asked to define words that are presented visually and orally.

The picture stimuli are closed-ended and thus, easy to analyse therefore we focused on the verbal stimuli. A set of possible responses (gold-standard) are provided for each stimulus in form of words, phrases and/or sentences that are scored as follows: 0-Point (poor match), 1-Point (good match) or 2-Point (excellent match). These are used to score examinees' responses to each stimulus depending on similarity to the gold-standard.

In this study, we considered 27 out of the 28 verbal stimulus listed in Table 1. This is because the stimuli ('pavid') does not have a vector representation in the embedding models used in this study (see Section 3.3.1 for details of embedding models). It is also important to note that one of the 27 stimulus ('car') does not have response associated to 1-Point, so only 0-Point and 2-Point scores were considered for this stimulus. The full characteristics of the dataset including stimulus and associated word and sentence counts per point scale is presented in Table 1.

The dataset includes 936 alternative potential responses (i.e., sentences) with a total of 3,000 constituent words excluding the 27 stimulus words. The average number of words within a set of response to stimulus is 34.63.

Table 2 shows a sample stimuli item (*tongue*) from the dataset with 'some of' the possible responses and associated point score. As can be observed in the Table, some of the responses are single words ($n = 254$ in the dataset) while the rest are either phrases or sentences ($n = 681$ in the dataset). The point score to response ratio is as follows — 0-Point : 158 responses; 1-Point : 396 responses; and 2-Point : 385 responses.

3.2. Data pre-processing tools

The methods implemented in this paper are largely focused on text vectorisation which involves the use of NLP tools to transform textual data (words, phrases and/or sentences) into machine readable format (vectors). However, a range of preprocessing steps were applied to the experimental dataset (described in Section 3.1) before vectorisation. Specifically, we used the Natural Language Toolkit (NLTK) (Loper and Bird, 2002) to perform spelling/bias correction, stop-word removal, tokenization, and word inflection. These are common preprocessing steps typically applied to clean-up textual data before vectorisation and an example is presented in Table 3 to illustrate their effects on a data sample. A brief explanation is also provided for each of them with reference to the sample data.

3.2.1. Spelling/bias correction

The WASI-II response set contains some words that are unlikely to feature in the word embedding models applied in our experiments

Table 3
Sample of a preprocessed response from a stimulus item.

Original response	The strip under the laces of your shoe
After Tokenisation	'The','strip','under','the','laces', 'of','your','shoe'
After Stop-word	'strip','under','laces','shoe'
After Word Inflection	'strip','under','lace','shoe'

(see Section 3.3). These include hyphenated words like 't-shirt'; words contracted with apostrophes like 'cannot' and numerical values such as '10'. Such representations were corrected in this preprocessing step. We also removed some irrelevant characters including responses that contain the stimulus word to avoid experimental bias and skewness. For example, the 0-Point response '[Points to tongue] **(Q)**' in the sample data shown in Table 2 contains the word 'tongue' which is the stimulus word. The example also contains the character '**(Q)**', used to indicate that the administrator should inquire further detail. These were removed from the dataset. We note that spelling/bias correction was performed to obtain the original dataset presented in Table 1 which includes a total of 936 sentences (made up of 3,000 words) associated to the 27 stimulus items.

3.2.2. Tokenisation

Tokenisation is designed to filter out meaningless symbols and split the remaining text (sentence or string) into tokens, i.e., set of characters that have a meaning by themselves (Jackson and Moulinier, 2002). A simple tokeniser splits a string by white space, but a more efficient tokeniser can use other techniques to separate elements eg punctuation and abbreviations (Loper and Bird, 2002). As punctuation marks do not contribute to the similarity evaluation conducted in this study, we removed them from the experimental data. This was achieved using the `word_tokenize()` function provided by the NLTK library, which effectively breaks down the text into a sequence of words based on whitespace and punctuation. We note that tokenisation did not alter the sentence and word count in the dataset which remained at 936 and 3,000 respectively. The sample presented in Table 3 illustrates why there was no change in the counts because each word in the original dataset is essentially a token.

3.2.3. Stop-word removal

Natural language often contain constructive terms (e.g., prepositions) and other language structures used to make sentences. These terms are commonly known as stop-words and their presence in the response set may increase the dimensionality of data if they make up a large portion of the textual dataset (Makrehchi and Kamel, 2008). Specifically for the experiments presented in this paper, the presence of stop-words may lead to poor efficiency of the similarity task between stimulus and response due to information loss in the centroid vector generation approach applied (i.e., averaging word vectors in a sentence).

As such, we used the `stopwords.words('english')` function of NLTK to filter out common English stopwords from the response set e.g., 'of' in Table 3. While the operation did not affect the sentence count of the dataset which remained at 936, the word count reduced to 2,131. The illustrative example in Table 3 provides insight into the rationale behind this decrease, attributed to the elimination of stopwords such as 'the' (occurring twice) and 'of' and 'your' (each occurring once), resulting in a reduction of the sample size from 8 to 4 words.

3.2.4. Word inflection

In the context of NLP, word inflection refers to the various forms a word can take based on factors like tense, number, gender, and case. This is handled by stemming and lemmatisation techniques which simply reduces words to their root or base forms. For the word 'laces', used in its verb form within Table 3; stemming might produce the stem 'lace' (noun), as it removes affixes 's' to return a base form. Meanwhile, lemmatisation considers the word's context and grammar, potentially yielding the lemma 'lace' as well (present tense).

For stemming, we used NLTK implementation of the Porter stemming algorithm (Porter, 1980) which can be accessed through the `PorterStemmer` class. For lemmatisation, NLTK offers the WordNet lemmatiser, which utilises WordNet, a lexical database of English. The lemmatisation process was performed using the `WordNetLemmatizer` class. It is important to note that stemming can sometimes produce non-standard or even non-existent words, while lemmatisation ensures valid words are returned. This ensures that we always retrieved word replacements whilst avoiding the repetition of words that share the same basic term and meaning but have different vector representation. Also, some word inflections (e.g., plural form) may not even exist in word embedding models. We note that word inflection did not alter the sentence and word count observed after stop word removal. Thus the dataset remained at 936 sentences and 2,131 words after word inflection. The overall statistics of original and pre-processed versions of the dataset is shown in Table 4.

3.3. Methods

This section presents our method to compute the similarity between the stimulus and the response set for the verbal items of the Vocabulary sub-test dataset (described in Section 3.1). The method employs multiple NLP tools and techniques and performs the following experimental steps on the dataset:

Step 1: **Data Retrieval** Retrieve the original and preprocessed WASI-II data items in textual form (see Table 3), each including the stimulus and the associated set of responses with score.

Step 2: **Vectorisation** Apply 5 embedding models (i.e., Word2vec, GloVe, BERT, GPT2 and ELMo) to obtain vector representation for each preprocessed stimulus and response set. Further details of the 5 embedding models and vectorisation is presented in Section 3.3.1.

Step 3: **Vector Weighting** Repeat **Step 2** and apply TF-IDF to weight each word in the response set based on relevance. Further details of the TF-IDF weighting approach is presented in Section 3.3.2.

Step 4: **Similarity Computation** Compute cosine similarity to evaluate the distance between stimulus and response set. This process is done independently for the resulting embeddings generated in **Step 2** and **Step 3**. Further details of the cosine similarity evaluation approach is presented in Section 3.3.3.

A high level representation of the method is shown in Fig. 2. The theoretical underpinning for this method is the rational-expectations hypothesis that the stimulus vector would be closest to that of 2-Point, then 1-Point and 0-Point responses because word embedding models map vectors in space such that words that are similar in meaning appear close to each other.

3.3.1. Vectorisation

Various embedding models exist for transforming textual data into vectors including Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019) and ELMo (Peters et al., 2018). These models have received much empirical evaluation and have been shown to be efficient for learning high quality distributed vector representations. They capture syntactic and semantic relationships between a large number of words and returns k-dimensional vector for each word.

To process the textual content of the WASI-II stimulus and response set, we transform words into fixed-length numerical vectors using the 5 embedding models. It is important to note that Word2Vec and

Table 4
Overall sentence and word count statistics of original vs. pre-processed dataset.

Dataset	Sentences	Words	Min| Max| Avg Sentence per point	Min| Max| Avg Words per point
original	936	3000	1 | 36 | 11.45	0 | 11 | 3.03
pre-processed	936	2131	1 | 36 | 11.45	0 | 7 | 2.20

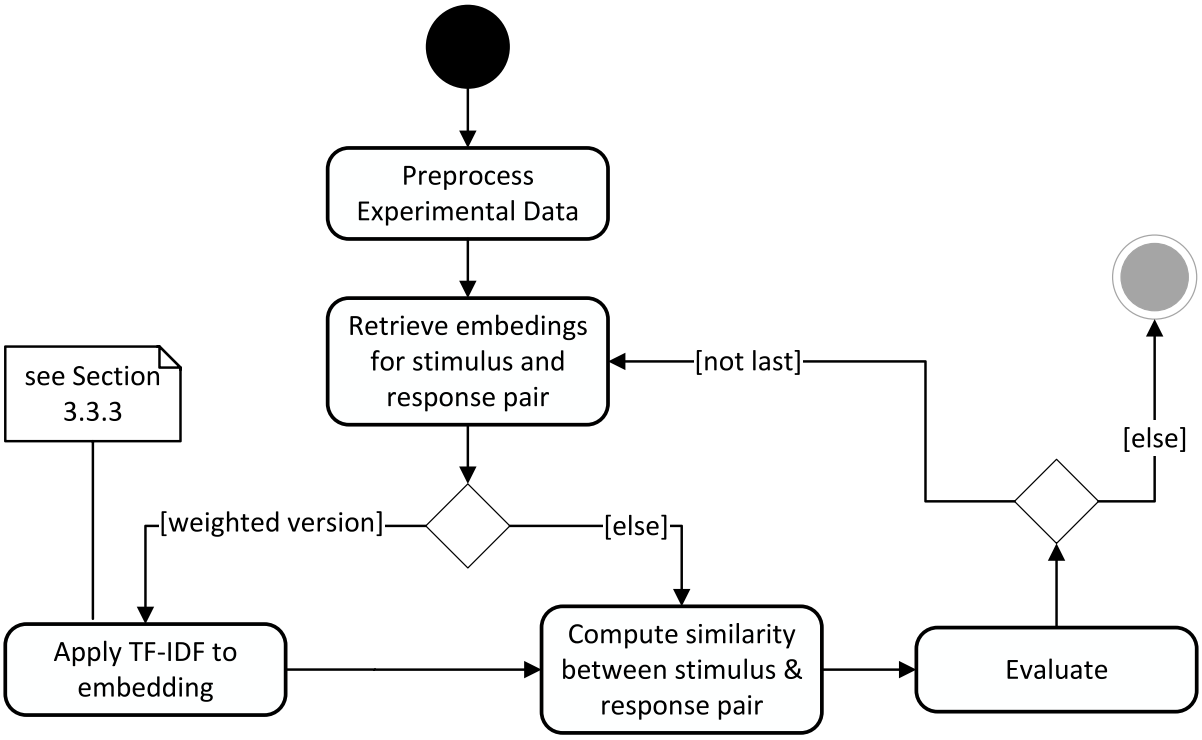


Fig. 2. High level activity diagram of the proposed method.

GloVe only accounts for single word while BERT, GPT2 and ELMo are capable of generating vector representation for words, phrases and/or sentences. For the purposes of this experiment, we used only the word vectorisation capability of the five models. In particular, for responses that consist of multiple words, we computed sentence vectors from the constituent words using the 'document centroid vector' (Rossiello et al., 2017) approach computes the average of all the word vectors in the document. Thus, for each WASI-II response r_i, we pass the sequence of words $w_1, w_2, \ldots, w_n$ through the embedding models to transform each word into a fixed-length numerical vector represented mathematically as Eq. (1):

$$\begin{aligned} \mathrm{r}_i &= \langle w_1, w_2, \ldots, w_n \rangle \\ &= \langle v_1, v_2, \ldots, v_n \rangle \end{aligned} \tag{1}$$

where $v_1, v_2, \ldots, v_n$ represents the transformation of response word sequence w_i into fixed-length numerical vectors.

To obtain sentence vector from a multi-word response within a set, we apply the document centroid vector method (Rossiello et al., 2017) by computing the sample mean of r_i which can be formalised as Eq. (2):

$$\bar{\mathrm{r}}_i = \frac{v_1 + v_2 + \cdots v_n}{n} \tag{2}$$

where $\bar{\mathrm{r}}_i$ is the mean of word vectors $v_1, v_2, \ldots, v_n$ within a sentence, and n represents the total number of words in the response.

Thus, a document vector including all responses in a single point scale (e.g., 0-Point) of a stimulus item is the mean of all individual responses within the set which can be represented as Eq. (3).

$$\bar{\mathrm{d}}_i = \frac{1}{n} \sum_{i=1}^{n} \bar{r}_i \tag{3}$$

where $\bar{\mathrm{d}}_i$ is the mean of sentence vectors $\bar{r}_i$ of a response subset in a finite response set and n is the total number of responses in the set.

This approach is known to be effective for identifying synonyms in short documents, but may be sub-optimal in long documents. Thus, we integrated a weighting factor in the computation of the centroid vector as explained in Section 3.3.2.

3.3.2. Vector weighting

In the data sample shown in Table 2, the relevance of each word within a response set (e.g., 2-Point Responses) in identifying the corresponding stimulus (i.e., Tongue) may differ. Thus, we assigned weighting to each word using TF-IDF (Rajaraman and Ullman, 2011), a term statistic commonly used in NLP to measure word relevance in a collection. TF-IDF is the multiplicative value of term frequency (TF) and inverse document frequency (IDF). To illustrate, TF — $\mathrm{tf}(t, d)$ is the frequency counter for a term t in document d while DF — $\mathrm{df}(t, D)$ is the count of occurrences of term t in N number of document set D. IDF — $\mathrm{idf}(t, D)$ is the inverse of $\mathrm{df}(t, D)$ commonly used to measure the

informativeness of term t. Thus, TF-IDF is represented mathematically as Eq. (4).

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D) \tag{4}$$

There are many different variations of TF-IDF but we used the version proposed by Ganesan (2020). To illustrate our implementation of TF-IDF with the data sample in Table 2, D represents all the textual responses within any of the point scales (0 - 2); d is a single response within a given point scale; and t is a single word within d. The weighting score wt_i for each word t is calculated as

$$\text{wt}_i = t \cdot \text{tf-idf}(t, d, D) \tag{5}$$

Thus, for each word w_i in a sequence of words within a single response r_i shown in Eq. (1), we multiply each vectorised word $v_1, v_2, \ldots, v_n$ with its TF-IDF weight. The weighted response r'_i is represented as Eq. (6)

$$\text{r}'_i = \langle v_1 \cdot wt_1, v_1 \cdot wt_2, \ldots, v_n \cdot wt_n \rangle \tag{6}$$

where $v_1 \cdot wt_1, v_1 \cdot wt_2, \ldots, v_n \cdot wt_n$ represents the weighted vector sequence of a given response.

The document vector described in Eq. (3) can then be re-represented by calculating the sample mean of the weighted response r'_i as shown in Eq. (7)

$$\bar{\text{d}}_i = \frac{1}{n} \sum_{i=1}^{n} \bar{r'}_i \tag{7}$$

where $\bar{\text{d}}_i$ is the mean of weighted sentence vectors $\bar{r'}_i$ of a response subset in a finite response set and n is the total number of responses in the set.

3.3.3. Similarity computation

Distance calculation is a common technique used in many text mining applications to measure the similarity between features of two data objects, in a dataset. Short distance between objects indicates high degree of similarity, while large distance indicates low degree of similarity. We applied this principle to assess how the WASI-II stimulus items are similar to their corresponding point scale responses. Some of the similarity metrics employed for NLP tasks include Jaccard similarity, Manhattan distance, Euclidean distance, Minkowski distance, Jensen–Shannon Divergence, Levenshtein distance and Cosine Similarity (Ladd, 2020; Wu, 2021). However, Cosine Similarity (Li et al., 2004; Li and Han, 2013), which measures the angle between two vectors, is the most popular for text mining. It is effectively calculated as dot-product of two normalised vectors as shown in Eq. (8)

$$\begin{aligned} \text{Cosine}(\vec{s}, \vec{r}) &= \frac{\vec{s} \cdot \vec{r}}{\|\vec{s}\| \cdot \|\vec{r}\|} \\ &= \frac{\sum_{i=1}^{N} s_i \times r_i}{\sqrt{\sum_{i=1}^{N} s_i^2} \times \sqrt{\sum_{i=1}^{N} r_i^2}} \end{aligned} \tag{8}$$

where $\vec{s}$ and $\vec{r}$ are stimulus and response vectors of dimension N.

To illustrate this process, we use a subset of the data sample presented in Table 2 as follows:

- **Stimulus**: "Tongue"
- **Response**: "The strip under the laces of your shoe"
- **Point**: "2-Point" is the score allocated to this response

The illustrative example shows a response r belonging to the 2-point subset of the response set $\mathbb{R}$ for a stimulus item s. Assuming that $\text{s} \notin \mathbb{R}$, the distance between the stimulus vector s and the response set $\mathbb{R}$ can be formalised as Eq. (9).

$$\text{dist}(\text{s}, \mathbb{R}) = \inf\{d(s, r) \ : \ \text{r} \in \mathbb{R}\} \tag{9}$$

where *tongue* is the stimulus s, 'strip, under, ⋯ , shoe' are the words that make up the response r which belongs to the 2-Point response set $\mathbb{R}$.

This distance between each stimulus item and the corresponding vector representation for each response subset (i.e., 0-Point, 1-Point and 2-Point) is calculated to measure similarity. It is important to note that cosine similarity value ranges between −1 (no similarity exists between compared vectors) and 1 (the compared vectors are absolutely similar). Thus, for the illustrative example data, the ideal outcome for our method is for the cosine similarity value between the stimulus item and '2-Point responses' to be closest to 1 (i.e., high degree of similarity), and then decrease sequentially from '1-Point to 0-Point responses'.

3.4. Experiment setup

The experimental data was stored as a single text file containing a series of stimulus and response set pairs in plain text. Each line is a single document $\mathbb{D}_i$ that contains the stimulus s and response subsets d_i ($i = 2$, 1 or 0) representing 2-Point, 1-Point or 0-Point. The structure of each document D_i can be formulated as Eq. (10).

$$\text{D}_i = \langle s_i; d_2; d_1; d_0 \rangle \tag{10}$$

where s_i is the stimulus, and d_2, d_1, d_0 are the corresponding response subsets for 2-Point, 1-Point and 0-Point respectively.

Each response subset D_i is a collection of textual responses t_i of a given point scale p_i. The formalisation of a given response subset D_i is presented in Eq. (11).

$$\text{d}_i = < t_1, t_2, \ldots t_n, p_i > \tag{11}$$

where $t_1, t_2, \ldots t_n$ is a series of textual responses that belongs to the p_i point scale (e.g., 2-Point).

To vectorise the textual responses, we explored different vector lengths (i.e., dimensions) of the 5 embedding models described in Section 3.3.1. For Word2vec and GloVe, we used Gensim (Ř.ehůřek and Sojka, 2010) python implementation and experimented with 50, 100, 200 and 300 vector dimensions. We observe that our method is stable when the dimension is set to a value between 200 and 300 but the best performance was obtained when the dimension is set to 300. This is consistent with recommendations from the founding study of both word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014). Thus, the dimensionality of vectors used for both models is 300. Similar experiments were conducted for BERT, GPT2 and ELMo using the TensorFlow (Abadi et al., 2015) python implementation to arrive at the optimal vector dimensions of 768, 300 and 512 respectively.

Many implementations exist in the literature for calculating TF-IDF. To perform the vector weighting task described in Section 3.3.2, we implemented a custom class based on the standard definition of TF-IDF proposed by Ganesan (2020). The custom class was implemented with `TfidfTransformer` and `CountVectorizer` classes from Scikit-Learn (Pedregosa et al., 2011).

For the similarity computation described in Section 3.3.3, we used Manning, Christopher D. Raghavan, Prabhakar Schütze (2008) cosine similarity implementation on Scikit-Learn (Pedregosa et al., 2011). This was used to obtain the cosine similarity between a given stimulus item and the corresponding response document illustrated in Eq. (9).

3.5. Evaluation & metrics

In this section, we present the evaluation metrics and processes employed to address the three research questions (i.e., **RQ1**, **RQ2** and **RQ3**) outlined in Section 1.

RQ1 compares the performance of the proposed method with and without the preprocessing steps discussed in Section 3.4. We explored this question mainly because of the verbatim transcription requirements in qualitative research (Davidson, 2009); but also to check if the full text will better capture the meaning, perception and context in which the WASI-II responses were created.

RQ2 examines the influence of TF-IDF weighting on the WASI-II responses. By making a direct comparison between the standard and

TF-IDF weighted vectors, we can determine to what extent a term weighting factor affects the performance of the proposed approach.

RQ3 investigates the performance differences observed in the 5 embedding models. We choose these models because they provide dense vector representation and to the best of our knowledge, they are the most popular embedding models used in contemporary literature (Lastra-Díaz et al., 2019).

3.5.1. Frequency of consistent representation

This metric counts the frequency of the study hypothesis in a given experiment i.e., how often the cosine similarity value between stimulus and response set decreases sequentially from 2-Point to 0-Point. Samples exhibiting this trend were considered a positive match and the rest were considered a negative match. For example, a positive matching is achieved for any given stimulus item if the computed cosine similarity coefficient is highest for its 2-Point response set and smallest for its 0-Point response set. Otherwise, the outcome of the computation is considered negative matching.

3.5.2. Pearson correlation coefficient

The Pearson correlation coefficient commonly known as r is a statistic that measures linear correlation between two variables x and y (Upton and Cook, 2008). It expresses the degree (on average) to which the x and y variables change correspondingly. Its value ranges between $+1$ and -1, where $+1$ indicates a perfect positive linear correlation, 0 indicates no linear correlation, and -1 indicates a perfect inverse (negative) linear correlation. Given a paired data $(x_1, y_1) \cdots (x_n, y_n)$ consisting of n pairs, the Pearson correlation coefficient r_{xy} is defined as Eq. (12).

$$Z_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (n\sum y_i)^2}} \tag{12}$$

where n is sample size and x_i, y_i are the individual sample points indexed with i.

Thus, if x_i increases when y_i increases, then there is a positive correlation. In this case the correlation coefficient will be closer to 1. However, if x_i decreases when y_i increases, then there is a negative correlation and the correlation coefficient will be closer to -1. In this study, the x_i represents the actual WASI-II response categorisations (2-Point, 1-Point and 0-Point) while y_i represents the corresponding cosine similarity coefficients computed for each response category. The ideal scenario (i.e., high positive linear correlation) is obtained when the calculated cosine similarity values for a given response set decreases sequentially from 1 to -1 across the WASI-II response categorisation from 2-Point to 0-Point.

3.5.3. Significance test

We conducted significance test by calculating the p-value of r. The p-value is the probability of obtaining test results at least as extreme as the results actually observed, if the correlation coefficient r was in fact zero (null hypothesis) (Wasserstein and Lazar, 2016). If this probability is lower than the conventional 5% ($p < 0.05$), then the correlation r is said to be statistically significant. In other words, p-value lower than 0.05 indicates that there is a positive relation between the WASI-II response categorisation and the calculated cosine similarity values.

3.5.4. Z-Test statistics

Correlations retrieved from different samples can also be tested against each other. For example, to test the significance of the difference between the correlation coefficient values obtained with original data and the preprocessed data. This is recommended when the correlations are conducted on the same variables (i.e., cosine similarity) but two different groups, and if both correlations are found to be statistically significant. To achieve this, we first compute the Fisher Z-Transformation (Fisher, 1915, 1921) which transforms Pearson's correlation coefficient r into a value z_r that can be used to calculate other metrics for r such as confidence interval or even comparison of correlations from independent samples. This transformation is necessary because the transformed variable z_r follows a normal distribution. The Fisher z_r of any correlation coefficient r is defined as Eq. (13):

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{13}$$

where ln is the natural logarithm function.

In this study, the transformation variable z_r was used to perform a two-tailed, two proportion z-test to compare correlations from independent samples for significance e.g., to ascertain if there is a significant difference in the correlation r of our experiments with original and pre-processed dataset. Given two transformation variables z_1 and z_2 obtained from two independent samples z-test statistic can be formalised as defined as Eq. (14):

$$z_{test} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{14}$$

where n_1 and n_2 represents the sample size used to obtain z_1 and z_2 transformations respectively.

The two tailed z-test has a single critical value (i.e., ± 1.96) for the conventional 5% ($p < 0.05$) significance, so its value ranges between -1 and 1 due to the two tailed approach. z_{test} will be a positive value if z_1 is bigger than z_2; and a negative value otherwise. For simplicity, we report only the absolute value that gives distance and discard information about direction. Statistical significance can be assessed by checking if the z_{test} value is greater than the critical value. For example, a significance level set at 0.05 indicates that the critical value is ± 1.96, so a z_{test} greater than ± 1.96 falls into the rejection region; thus statistically significance. This means that the null hypothesis can be rejected that the two correlations z_1 and z_2 are not significantly different.

4. Results

This section presents the results of experiments to address the research questions **RQ1**, **RQ2** and **RQ3**. For brevity, the actual cosine similarity values between stimulus and response set is reported in Appendix A.

Table A.1 shows the results of experiments with original dataset while Table A.2 presents results of experiments with preprocessed dataset. We also present in Tables A.3 and A.4, the experimental results obtained when TF-IDF is applied to the original and preprocessed datasets. The results follows similar presentation format in all 4 Tables (i.e.,Tables A.1–A.4) where the stimulus item is shown in the first column, followed by 15 columns that presents the cosine similarity of the different response scales obtained for each embedding model. Aggregate metrics calculated from the cosine similarity results are presented in Tables 5 and 6 to evaluate experiments with the unweighted and weighted versions of the experimental dataset respectively. For clarity, the result highlights for each dataset version is interpreted in separate sub-sections.

4.1. Unweighted dataset

As shown in Table 5, Word2vec performed better than the other embedding models with 18 out of 27 stimulus:response pair successfully matching the evaluation criteria (i.e., 2-Point > 1-Point > 0-Point). Similar performance was observed between experiments with original and preprocessed dataset and the results show significant correlation ($r = 0.61$, p-value = 0.00) between the WASI-II response categories and the computed similarity. Consequently, no significant difference was observed between the experiments with original and preprocessed dataset as evidenced by the z_{test} of 0.00 which is lower than the critical value (i.e., ± 1.96). In fact, all the embedding models yielded insignificant correlation difference between the experiments with original and preprocessed datasets.

Table 5
Evaluation results of experiments with original and preprocessed dataset on the 5 embedding models.

Model & Data		+ve count	r	p-value	z_{test}
W2v	Original	18	0.61	**0.00**	0.000
	Preprocessed	18	0.61	**0.00**	
GloVe	Original	13	0.25	0.22	0.187
	Preprocessed	16	0.30	0.12	
BERT	Original	6	0.04	0.84	0.035
	Preprocessed	5	0.05	0.79	
GPT2	Original	5	−0.12	0.56	0.765
	Preprocessed	7	0.10	0.60	
ELMo	Original	17	0.44	**0.02**	0.267
	Preprocessed	18	0.50	**0.01**	

Note In the table, '**+ve count**' column indicates the count of samples that meets the evaluation criteria (2-Point > 1-Point > 0-Point) and '**bold**' typeface indicates that result is significant.

Table 6
Evaluation results of experiments with TF-IDF weighted original and preprocessed dataset on the 5 embedding models.

Model & Data (TF-IDF)		+ve	r	p-value	z_{test}
W2v	Original	20	0.61	**0.00**	0.000
	Preprocessed	18	0.61	**0.00**	
GloVe	Original	14	0.26	0.19	0.189
	Preprocessed	16	0.31	0.12	
BERT	Original	5	0.03	0.89	0.069
	Preprocessed	5	0.05	0.79	
GPT2	Original	7	0.11	0.58	0.000
	Preprocessed	9	0.11	0.58	
ELMo	Original	18	0.48	**0.01**	0.233
	Preprocessed	19	0.53	**0.00**	

Note In the table, '**+ve**' column indicates the count of samples that meets the evaluation criteria (2-Point > 1-Point > 0-Point) and '**bold**' typeface indicates that result is significant.

That said, all the embedding models (excluding Word2vec) performed better with the preprocessed dataset. However, only the ELMo model which produced the second best results shows significant correlation ($r = 0.51$, p-value = 0.01) between the WASI-II response categories and the computed similarity. Specifically, one more stimulus:response pair was identified successfully with preprocessed data. Further discussion is provided in Section 5 to contextualise the results and show the extent to which our experiments addressed the research questions.

4.2. Weighted dataset

As shown in Table 6, Word2vec and ELMo performed better than the other embedding models. The Word2vec model was more successful when the original dataset was weighted with TF-IDF; producing 20 out of 27 matches of the correlation frequency evaluation criteria (i.e., 2-Point > 1-Point > 0-Point). This is 2 matches more than the 18 produced with the preprocessed dataset. However, the difference between the two models is insignificant as shown by the z_{test} of 0.00 which is lower than the critical value (i.e., ±1.96). This is expected, given that both models produced similar correlation coefficient, $r = 0.61$.

The ELMo model, performed better with the preprocessed dataset (19 out of 27 matches) in comparison to the original dataset (18 out of 27 matches). The correlation difference between the two models is insignificant as shown by the z_{test} of 0.233. In fact, all the embedding models yielded insignificant correlation difference between the experiments with original and preprocessed datasets when TF-IDF was applied. Generally, the results suggests that the TF-IDF weighting steps had positive effect on some embedding models' performance, such as ELMo, GPT2, and GloVe. However, the improvements are insignificant. Contextual explanation of this result is provided in Section 5 to show how they address the research questions.

5. Discussion

It is important to put the results into context especially to show how and to what degree we addressed the research questions: **RQ1**- Usefulness of text preprocessing; **RQ2**- TF-IDF influences as a weighting factor; and **RQ3**- Best performing embedding model.

Although embedding models made it possible to analyse WASI-II response texts automatically, many issues relating to natural language structure and formality still remained. For example, informal writing such as 'cannot' is ubiquitous in English language but computers prefer its formal representation 'cannot'. Likewise, common words such as 'is' might not add much value to the meaning read by a machine, hence **RQ1**. The results presented in Table 5 shows our attempt to assess this empirically, by comparing the performance of our method with the original and preprocessed versions of the dataset. We observed that text preprocessing had a positive but insignificant effect on the models' performance.

It is important to take a task specific viewpoint when interpreting this result because several factors may have contributed to the outcomes. The original data source determines the level of noise it contains before processing so data obtained from social media is likely to contain more noise (e.g., slang, abbreviations, emoticons etc.) than the WASI-II dataset. Thus, the preprocessed version of our experimental data may not be too dissimilar from the original version. In fact, stop-word-removal is the preprocessing step that is likely to make a difference. Unfortunately, there is no universal stop words list because a word can be empty of meaning depending on the corpus in use, or the task being undertaken. Some people may consider a stop words to be any word that has high frequency on a corpus while others may consider every word that is devoid of true meaning given a context. This means that any word can be a stop word depending on task being undertaken which explains why there is a lot of debate about the relevance of stop-word-removal (Munková et al., 2014; Silva and Ribeiro, 2003). That said, the presence of stop words may increase the dimensionality of data if they make up a large portion of the textual dataset (Makrehchi and Kamel, 2008). Reducing the dataset size helps to avoid known constraints such as increased model complexity, slow training/analysis speed, and increased inferential latency (Wu et al., 2016). These can potentially limit model performance, applicability and deployment so having less tokens is often desirable.

In addition, TF-IDF which gives more value to rare words than repetitive tokens is a technique that is commonly used to boost performance. Consider the case where the WASI-II 2-Point response set for a given stimulus contains a very rare word. Since the frequency of this word is very low, TF-IDF will consider it a rare token and assign a high weight. The results presented in Table 6 shows our attempt to assess this empirically, by comparing the performance of our method when TF-IDF is used to add weights to the tokens within the original and preprocessed versions of the dataset. Again, we observed that TF-IDF had a positive but insignificant effect on the models' performance and thus addresses **RQ2**.

It is important to note that the WASI-II automation achieved in this study was made possible by the embedding models which facilitated vector representation of textual data. However, varying performance was observed with the 5 embedding models. Specifically, Word2vec and ELMo seem to be better than the rest in providing representative vectors for the cosine similarity task undertaken. The best Word2vec model was obtained when TF-IDF was applied to the original dataset with positive correlation frequency count of 20 and correlation coefficient of 0.61 as shown in Table 5. The best ELMo model however,

was achieved when TF-IDF was applied to the preprocessed dataset with one less positive correlation frequency count (i.e., 19) and a correlation coefficient of 0.53 which is 8% lower than that obtained with the Word2vec model. As such, we conclude that Word2vec produced the most optimal performance for the task undertaken in this study which addresses **RQ3**. We note that the z_{test} between Word2vec and ELMo models is 0.411 which means that the observed correlation difference is insignificant. However, significant correlation difference was observed between the best Word2vec model and the least performing model (i.e., BERT with TF-IDF applied to the preprocessed dataset). Specifically, the correlation frequency for the BERT model is 5, with r value of 0.05 which is insignificant ($p = 0.79$). Therefore the correlation difference between the best Word2vec and BERT models is 56% (i.e., 0.61 - 0.05). This equates to z_{test} of 2.282 between the two models which is higher than the recommended critical value of ± 1.96) for the conventional 5% ($p < 0.05$) significance, thus significant.

The varying performance of the embedding models could be due to various reasons that may relate to training corpus and size, vector dimension and/or operational mechanism. For example, Word2vec and GloVe are context independent while BERT, GPT2 and ELMo generates different vector representation for a given word in a way that captures the context of the word (i.e., its position in a sentence). An in-depth exploration of these factors is presented to better understand why certain models excelled in some instances while struggling in others.

Firstly, Word2vec and GloVe are context-independent models. This means that these models generate a static vector representation for each word, irrespective of the context in which the word appears. For example, the word "bank" would have the same vector representation in the phrases "river bank" and "financial bank". This can be a limitation when dealing with language assessments like the WASI-II, where context plays a critical role in understanding and evaluating language ability. Despite this limitation, Word2vec achieved 74.07% accuracy in our study, suggesting that its simplicity and efficiency in capturing general word associations still provide a reasonable performance baseline. We suspect that the word disambiguation issue did not occur because the responses to stimuli items in WASI II are already specific to task. However, the lack of contextual sensitivity in Word2vec might cause it to struggle in practical settings when capturing and comparing actual user responses with the indicative responses provided by WASI II, especially with words that have multiple meanings or are used in complex sentence structures.

On the other hand, BERT, GPT-2, and ELMo are context-dependent models. These models generate different vector representations for the same word based on its context within a sentence. This contextual awareness is particularly beneficial for tasks that require a deep understanding of language and its subtleties, such as the WASI-II questionnaire. BERT, for example, uses a bidirectional transformer architecture that allows it to consider both the left and right context of a word simultaneously. This capability enables BERT to capture more intricate relationships between words, leading to a more accurate representation of their meanings in context. Consequently, BERT and similar models may perform better in tasks requiring high semantic understanding. We suspect that BERT did not perform as expected in this study because the indicative responses from WASI II is already well structured to address the stimuli so the models' complexity and computational requirements became a downside, potentially making it less efficient than simpler models like Word2vec in this scenario.

Another aspect to consider is the training corpus and size. Models like GloVe are pre-trained on large corpora such as the Common Crawl, Twitter or Wikipedia, which contain a vast amount of general language data. While this extensive training can help the model capture broad linguistic patterns, it might not be as effective in specialised contexts like the WASI-II, where specific linguistic features are crucial. That said, GloVe still performed better than BERT and GPT-2, which are also trained on extensive datasets. The latter uses advanced architectures that enable fine-tuning on specific tasks and this adaptability allows them to excel in more specialised language assessments. Unfortunately, BERT and GPT-2 were not fine-tuned with sufficient domain-specific data in our study, hence the performance. Conversely, ELMo performed well on WASI II structured data due to its unique architecture, which includes bidirectional LSTM layers that capture both forward and backward context, and its ability to generate contextualised word embeddings based on entire sentence structures. This makes ELMo particularly adept at understanding the syntax and semantics of well-structured texts, such as the standardised content of the WASI-II questionnaire. In comparison, GloVe's pre-trained vectors, derived from sources like Common Crawl, Twitter, and Wikipedia, lack the refined context specificity that ELMo offers. Additionally, while BERT and GPT-2 also generate context-aware embeddings, their transformer architectures may not align as closely with the structured nature of the WASI-II content as ELMo's LSTM-based approach, which is inherently better at modelling sequential data.

5.1. Theoretical contributions

The theoretical contributions of this work are multifaceted, addressing significant gaps in the existing literature on automated cognitive assessment tools. Firstly, our attempt to automate the WASI-II language ability questionnaire represents a novel endeavour in the field of psychometrics. To the best of our knowledge, no prior studies have successfully automated this specific questionnaire, making our research new and relevant in this area. This effort not only broadens the scope of automated assessments but also sets a precedent for future research in automating other established psychological and cognitive evaluation instruments.

In evaluating the performance of various standard word embedding models, including Word2vec, GloVe, BERT, GPT2 and ELMo; we contribute valuable insights into their effectiveness in representing the intended meanings of the WASI-II stimulus items. Our findings highlight the potential of these models in capturing the different semantic relationships inherent in WASI II language ability assessment tool. In particular, the 74.07% accuracy demonstrated by Word2vec model emphasises the practical applicability of these embeddings in real-world cognitive assessments. This contributes to the theoretical understanding of how advanced NLP techniques can be leveraged to enhance traditional psychometric methods.

Furthermore, our exploration of TF-IDF as a weighting factor on the word or document embeddings generated from the WASI-II response set adds another layer of theoretical contribution. The use of TF-IDF in conjunction with word embeddings to improve response relevance matching and information gain is an interesting approach that has not been extensively studied in this context. Our results indicate that applying TF-IDF weighting can significantly enhance the performance of the automated scoring system. This finding enriches the theoretical discourse on the integration of classic information retrieval techniques with modern machine learning models, indicating their combined efficacy in the domain of automated cognitive assessments.

5.2. Practical implications

While integrating the automated scoring system for the WASI-II language ability questionnaire into existing assessment environments falls outside the scope of this study; it presents a unique opportunity to enhance the efficiency and accuracy of cognitive evaluations. The system can be seamlessly incorporated into digital assessment platforms commonly used by clinicians, researchers, and educators. For example, existing electronic health record (EHR) systems or educational assessment tools can be augmented with our automated scoring feature, allowing for unbiased analysis and feedback. This integration would streamline the assessment process, reducing the manual scoring burden on professionals and enabling them to focus more on interpretation and intervention.

The development of a user-friendly interface is crucial for the successful adoption of this automated system. The interface should be intuitive, providing clear instructions and visual aids to guide users through the process. For clinicians, the interface could include options for uploading patient responses (in audio format) directly into the EHR systems, along with NLP modules for translation into text, reviewing and automatic scoring. Researchers could also benefit from features that allow for easy data export and detailed statistical analysis, facilitating the integration of scoring results into broader research studies. Educators might require simplified tools for quick assessments and instant feedback, making the tool accessible even to those with minimal technical expertise.

To ensure seamless interaction, the interface should incorporate real-time feedback mechanisms, allowing users to see the automated scores and corresponding analyses immediately after submission. This could be achieved through an interactive dashboard displaying key metrics, such as the accuracy of the responses, correlation scores, and areas requiring further attention. Additionally, customisation options should be available to cater to the specific needs of different user groups, whether it be adjusting the scoring parameters or selecting different preprocessing techniques and word embedding models based on the task at hand.

6. Conclusion

We have investigated the feasibility of automating the WASI-II language ability questionnaire by evaluating the performance of five word embedding models. We considered various text preprocessing techniques and TF-IDF as a weighting function to improve the performance of our method. Our approach is based on established NLP tools and techniques that have been applied extensively in other research works. However, there is novelty in the way we have combined them to achieve our research goals and to the best of our knowledge, no other research has applied them to automate the WASI-II questionnaire. As the task is a correlation-based one involving a modest dataset sample of 27 stimulus items instances, we explored various preprocessing techniques to reduce noise before applying TF-IDF weighting to increase information gain. Our method shows that the WASI-II questionnaire can be automated with 74.07% accuracy (i.e., 20 out of 27 correlation frequency) between actual (textual) and calculated (vector) representation in terms of response relevance matching. This result was obtained with Word2vec model on the original dataset and there is potential to improve the performance in further experiments by exploring different vector dimensions, similarity metrics and data preprocessing techniques not considered in this study.

CRediT authorship contribution statement

Nonso Nnamoko: Supervision, Data curation, Software, Methodology, Formal analysis, Validation, Writing – original draft, review & editing . **Themis Karaminis:** Supervision, Project administration, Funding acquisition, Conceptualization. **Jack Procter:** Visualization, Software. **Joseph Barrowclough:** Writing – review & editing. **Ioannis Korkontzelos:** Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been carried out as part of the Project ‘Aut2vec: Exploring language abilities of autistic children with computational linguistics and machine learning methods’, which received funding from Edge Hill University Research Investment Fund (Reference No. 20224966), awarded to TK and IK. Special thanks to the computer science department for providing time and resources to complete the research.

Appendix A. Complete experiment result tables

This appendix contains the actual cosine similarity values between stimulus and response set obtained from our experiments with the:

1. original dataset (Table A.1),
2. preprocesed dataset (Table A.2),
3. original dataset weighted with TF-IDF (Table A.3) and
4. preprocessed dataset weighted with TF-IDF (Table A.4).

Table A.1
Cosine similarity relationship between the stimulus and responses on original data.

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
shirt	0.36	0.19	0.22	0.40	0.30	0.34	0.73	0.77	0.72	−0.01	−0.02	0.03	0.46	0.31	0.37
car	0.35	NaN	0.20	0.43	NaN	0.42	0.78	NaN	0.90	0.01	NaN	0.01	0.47	NaN	0.30
lamp	0.31	0.21	0.18	0.30	0.21	0.18	0.87	0.86	0.94	−0.04	−0.01	0.01	0.37	0.31	0.25
bird	0.42	0.34	0.27	0.44	0.33	0.34	0.70	0.66	0.71	0.03	0.01	−0.03	0.48	0.42	0.34
tongue	0.23	0.20	0.16	0.25	0.27	0.23	0.73	0.75	0.69	0.01	−0.02	0.01	0.38	0.26	0.24
pet	0.40	0.36	0.36	0.36	0.35	0.38	0.77	0.70	0.75	−0.01	0.03	0.07	0.38	0.38	0.35
lunch	0.61	0.44	0.38	0.55	0.47	0.35	0.76	0.72	0.71	−0.02	0.04	0.02	0.58	0.45	0.49
bell	0.30	0.27	0.16	0.30	0.29	0.21	0.88	0.89	0.84	−0.02	0.00	−0.03	0.37	0.33	0.25
calendar	0.29	0.24	0.18	0.32	0.30	0.22	0.89	0.87	0.88	−0.05	−0.03	0.04	0.42	0.38	0.37
alligator	0.42	0.34	0.17	0.28	0.23	0.13	0.62	0.56	0.53	−0.06	0.00	−0.01	0.50	0.46	0.30
dance	0.34	0.32	0.21	0.39	0.36	0.27	0.78	0.72	0.73	−0.02	0.00	−0.01	0.40	0.44	0.36
summer	0.47	0.43	0.25	0.57	0.53	0.41	0.73	0.75	0.73	0.03	0.01	−0.02	0.51	0.48	0.30
reveal	0.45	0.32	0.19	0.41	0.42	0.39	0.67	0.71	0.71	0.02	−0.03	−0.02	0.57	0.46	0.43
decade	0.39	0.37	0.47	0.50	0.49	0.56	0.68	0.68	0.62	0.01	0.04	0.03	0.43	0.47	0.57
entertain	0.35	0.30	0.14	0.28	0.25	0.18	0.82	0.83	0.91	0.01	0.00	0.02	0.49	0.43	0.38
tradition	0.27	0.27	0.21	0.41	0.40	0.31	0.79	0.73	0.72	−0.01	0.00	−0.01	0.41	0.41	0.40

(*continued on next page*)

Table A.1 (*continued*).

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
enthusiastic	0.43	0.35	0.27	0.40	0.37	0.25	0.77	0.76	0.74	0.02	0.02	−0.01	0.58	0.52	0.49
improvise	0.24	0.28	0.21	0.06	0.10	0.09	0.62	0.59	0.66	−0.02	−0.02	0.00	0.44	0.47	0.45
haste	0.33	0.26	0.21	0.19	0.14	0.17	0.76	0.73	0.83	0.04	−0.03	0.04	0.47	0.45	0.38
trend	0.34	0.28	0.21	0.38	0.36	0.37	0.71	0.71	0.70	0.06	0.03	0.02	0.49	0.46	0.36
impulse	0.27	0.28	0.20	0.27	0.27	0.23	0.82	0.78	0.78	−0.01	−0.02	0.07	0.41	0.44	0.42
ruminate	0.37	0.27	0.29	0.06	−0.07	−0.07	0.55	0.55	0.48	−0.02	−0.04	0.03	0.48	0.41	0.41
mollify	0.47	0.24	0.29	0.34	0.07	0.19	0.67	0.62	0.66	0.07	0.03	−0.01	0.56	0.45	0.52
extirpate	0.35	0.24	0.23	0.05	−0.03	0.06	0.58	0.55	0.56	0.02	0.06	0.02	0.42	0.38	0.38
panacea	0.42	0.29	0.15	0.19	0.17	−0.02	0.63	0.65	0.61	−0.02	−0.04	0.00	0.48	0.44	0.29
perfunctory	0.34	0.26	0.22	0.23	0.13	0.07	0.66	0.63	0.66	−0.06	−0.01	−0.02	0.55	0.49	0.49
insipid	0.47	0.30	0.24	0.41	0.11	0.13	0.58	0.51	0.51	0.04	−0.02	−0.02	0.64	0.48	0.49
pavid	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Note In the table, '**NaN**' means that WASI-II does not contain a response while '–' means that stimulus was not found in pre-trained word embedding.

Table A.2
Cosine similarity relationship between the stimulus and responses on preprocessed data.

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
shirt	0.37	0.18	0.22	0.41	0.30	0.35	0.72	0.76	0.72	0.00	−0.01	−0.01	0.48	0.34	0.37
car	0.35	NaN	0.20	0.42	NaN	0.39	0.73	NaN	0.82	−0.01	NaN	−0.04	0.49	NaN	0.31
lamp	0.32	0.21	0.19	0.36	0.23	0.19	0.83	0.85	0.92	−0.03	0.00	−0.04	0.37	0.32	0.27
bird	0.43	0.34	0.28	0.43	0.33	0.33	0.67	0.66	0.71	−0.04	0.00	−0.01	0.49	0.43	0.36
tongue	0.25	0.18	0.15	0.27	0.27	0.19	0.74	0.76	0.69	0.01	0.01	0.01	0.43	0.32	0.27
pet	0.42	0.35	0.36	0.37	0.35	0.38	0.77	0.70	0.75	0.00	0.00	−0.05	0.41	0.39	0.35
lunch	0.63	0.48	0.40	0.60	0.53	0.36	0.78	0.73	0.71	−0.02	0.05	0.00	0.61	0.49	0.48
bell	0.31	0.27	0.16	0.29	0.28	0.20	0.85	0.87	0.84	0.00	0.00	0.00	0.41	0.36	0.27
calendar	0.29	0.22	0.17	0.35	0.29	0.21	0.88	0.84	0.84	0.01	−0.02	−0.03	0.43	0.40	0.40
alligator	0.44	0.36	0.17	0.36	0.26	0.14	0.65	0.58	0.54	0.02	0.00	0.00	0.53	0.48	0.31
dance	0.34	0.33	0.21	0.41	0.36	0.27	0.78	0.71	0.72	0.01	−0.02	0.01	0.43	0.45	0.37
summer	0.47	0.43	0.25	0.57	0.52	0.42	0.73	0.74	0.74	−0.02	0.03	0.04	0.53	0.51	0.32
reveal	0.45	0.32	0.18	0.42	0.42	0.40	0.68	0.71	0.67	−0.01	−0.06	−0.06	0.57	0.45	0.42
decade	0.39	0.38	0.47	0.45	0.46	0.56	0.69	0.68	0.62	0.00	0.01	0.06	0.45	0.49	0.57
entertain	0.35	0.30	0.14	0.30	0.26	0.17	0.80	0.82	0.90	0.02	0.00	−0.01	0.48	0.43	0.37
tradition	0.28	0.29	0.24	0.41	0.39	0.32	0.79	0.73	0.72	0.02	0.03	0.01	0.41	0.42	0.41
enthusiastic	0.44	0.35	0.27	0.43	0.38	0.25	0.69	0.75	0.73	−0.04	−0.04	−0.03	0.61	0.52	0.48
improvise	0.25	0.28	0.21	0.11	0.13	0.09	0.65	0.60	0.66	0.01	−0.01	0.02	0.46	0.47	0.45
haste	0.33	0.26	0.21	0.19	0.14	0.17	0.76	0.73	0.83	0.03	−0.05	0.01	0.47	0.45	0.38
trend	0.34	0.27	0.20	0.39	0.36	0.37	0.72	0.70	0.70	0.03	0.00	0.02	0.51	0.45	0.36
impulse	0.26	0.28	0.20	0.30	0.28	0.23	0.80	0.77	0.78	−0.02	−0.01	0.09	0.43	0.44	0.42
ruminate	0.38	0.27	0.29	0.08	−0.04	−0.08	0.55	0.56	0.47	−0.01	−0.02	−0.02	0.48	0.43	0.42
mollify	0.47	0.24	0.29	0.35	0.08	0.19	0.68	0.63	0.66	0.04	−0.04	0.01	0.56	0.45	0.52
extirpate	0.35	0.24	0.23	0.07	−0.03	0.06	0.58	0.55	0.56	0.03	0.06	0.04	0.42	0.38	0.38
panacea	0.41	0.28	0.15	0.25	0.18	−0.01	0.66	0.65	0.62	0.05	0.01	−0.04	0.49	0.44	0.28
perfunctory	0.34	0.26	0.22	0.26	0.14	0.07	0.67	0.64	0.66	0.00	−0.01	−0.03	0.55	0.49	0.49
insipid	0.47	0.30	0.24	0.41	0.12	0.13	0.58	0.52	0.52	0.04	−0.03	−0.03	0.64	0.49	0.49
pavid	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Note In the table, '**NaN**' means that WASI-II does not contain a response while '–' means that stimulus was not found in pre-trained word embedding.

Table A.3
Cosine similarity relationship between the stimulus and responses on original data weighted with TF-IDF.

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
shirt	0.36	0.19	0.23	0.41	0.30	0.34	0.73	0.77	0.72	0.03	−0.02	−0.04	0.49	0.39	0.44
car	0.35	NaN	0.21	0.43	NaN	0.40	0.77	NaN	0.89	0.02	NaN	−0.01	0.52	NaN	0.38
lamp	0.32	0.21	0.18	0.32	0.22	0.19	0.82	0.85	0.93	−0.01	−0.03	0.02	0.39	0.34	0.27
bird	0.42	0.35	0.27	0.43	0.33	0.32	0.70	0.66	0.71	0.00	−0.02	−0.11	0.50	0.46	0.38
tongue	0.23	0.20	0.16	0.24	0.27	0.22	0.73	0.75	0.69	0.04	−0.01	0.01	0.43	0.34	0.32
pet	0.39	0.35	0.34	0.35	0.35	0.37	0.77	0.70	0.75	0.04	0.00	0.04	0.42	0.42	0.39
lunch	0.60	0.44	0.38	0.55	0.47	0.35	0.77	0.72	0.71	0.01	0.01	0.00	0.62	0.49	0.50
bell	0.30	0.26	0.16	0.29	0.28	0.20	0.87	0.89	0.84	−0.02	0.04	−0.02	0.46	0.42	0.31
calendar	0.30	0.24	0.18	0.32	0.30	0.23	0.89	0.86	0.87	0.00	0.02	−0.02	0.51	0.46	0.48
alligator	0.41	0.34	0.17	0.27	0.24	0.13	0.62	0.59	0.52	0.03	0.01	0.00	0.54	0.51	0.32
dance	0.34	0.32	0.21	0.40	0.36	0.28	0.78	0.72	0.72	0.01	0.03	0.06	0.51	0.48	0.39
summer	0.47	0.43	0.25	0.58	0.53	0.41	0.73	0.75	0.73	0.02	0.02	−0.08	0.59	0.57	0.38
reveal	0.45	0.33	0.20	0.42	0.43	0.37	0.67	0.72	0.72	−0.01	0.04	0.05	0.60	0.51	0.54
decade	0.37	0.34	0.47	0.49	0.47	0.54	0.68	0.70	0.62	−0.06	−0.02	0.00	0.47	0.49	0.54
entertain	0.35	0.30	0.14	0.29	0.26	0.20	0.78	0.82	0.90	0.01	0.01	0.00	0.54	0.49	0.47
tradition	0.27	0.26	0.21	0.41	0.40	0.31	0.78	0.73	0.71	−0.01	−0.01	0.02	0.50	0.49	0.44

(*continued on next page*)

Table A.3 (*continued*).

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
enthusiastic	0.42	0.35	0.27	0.39	0.37	0.25	0.77	0.76	0.74	−0.02	−0.02	−0.01	0.64	0.58	0.51
improvise	0.23	0.28	0.20	0.06	0.11	0.09	0.63	0.59	0.67	0.04	0.00	0.00	0.51	0.52	0.45
haste	0.33	0.26	0.21	0.19	0.14	0.17	0.76	0.73	0.83	−0.02	−0.03	0.17	0.53	0.47	0.45
trend	0.33	0.27	0.21	0.38	0.36	0.37	0.69	0.71	0.69	0.09	−0.04	−0.02	0.56	0.51	0.45
impulse	0.28	0.28	0.20	0.28	0.27	0.23	0.77	0.76	0.78	−0.01	0.03	0.07	0.49	0.49	0.44
ruminate	0.37	0.27	0.30	0.07	−0.07	−0.06	0.56	0.55	0.49	0.02	0.05	−0.08	0.50	0.46	0.47
mollify	0.46	0.24	0.28	0.35	0.08	0.19	0.67	0.63	0.65	0.07	−0.01	−0.04	0.60	0.51	0.54
extirpate	0.35	0.24	0.23	0.06	−0.03	0.07	0.58	0.56	0.56	0.05	0.02	0.02	0.44	0.42	0.40
panacea	0.40	0.29	0.15	0.19	0.17	−0.02	0.63	0.65	0.61	−0.03	0.00	0.02	0.48	0.46	0.33
perfunctory	0.34	0.26	0.22	0.24	0.13	0.07	0.67	0.64	0.66	0.01	0.00	0.02	0.59	0.55	0.51
insipid	0.47	0.30	0.24	0.41	0.11	0.13	0.58	0.51	0.51	0.05	0.00	0.01	0.65	0.53	0.53
pavid	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Note In the table, NaN means that WASI-II does not contain a response while '–' means that stimulus was not found in pre-trained word embedding.

Table A.4
Cosine similarity relationship between the stimulus and responses on preprocessed data weighted with TF-IDF.

	Word2vec			GloVe			BERT			GPT2			ELMo		
stimulus	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt	2-pt	1-pt	0-pt
shirt	0.38	0.18	0.22	0.41	0.30	0.35	0.72	0.76	0.72	0.01	−0.02	0.02	0.50	0.41	0.43
car	0.35	NaN	0.20	0.42	NaN	0.39	0.73	NaN	0.81	0.01	NaN	0.04	0.52	NaN	0.39
lamp	0.32	0.21	0.19	0.37	0.23	0.20	0.82	0.84	0.92	−0.02	−0.02	0.02	0.41	0.35	0.29
bird	0.44	0.34	0.27	0.43	0.33	0.32	0.67	0.65	0.71	−0.02	−0.01	0.02	0.53	0.46	0.40
tongue	0.26	0.18	0.15	0.28	0.26	0.19	0.74	0.76	0.69	0.11	0.04	−0.04	0.48	0.34	0.32
pet	0.41	0.34	0.34	0.36	0.34	0.37	0.77	0.70	0.74	−0.02	−0.01	0.01	0.45	0.42	0.39
lunch	0.62	0.46	0.40	0.59	0.51	0.36	0.78	0.73	0.71	0.00	0.01	0.01	0.65	0.50	0.49
bell	0.31	0.27	0.16	0.29	0.28	0.20	0.85	0.86	0.84	0.01	0.00	0.00	0.46	0.42	0.31
calendar	0.30	0.22	0.17	0.36	0.29	0.21	0.88	0.84	0.84	0.04	0.03	0.02	0.49	0.44	0.48
alligator	0.43	0.35	0.17	0.35	0.26	0.14	0.65	0.58	0.53	0.00	0.04	0.02	0.59	0.52	0.33
dance	0.34	0.33	0.21	0.42	0.36	0.28	0.78	0.71	0.72	0.03	0.06	0.01	0.52	0.48	0.39
summer	0.47	0.42	0.25	0.56	0.52	0.42	0.72	0.74	0.73	0.06	−0.02	−0.04	0.60	0.57	0.37
reveal	0.45	0.32	0.20	0.42	0.43	0.39	0.68	0.72	0.68	0.03	0.04	0.13	0.60	0.51	0.52
decade	0.37	0.35	0.47	0.44	0.46	0.54	0.69	0.70	0.62	0.03	0.00	0.02	0.48	0.47	0.54
entertain	0.35	0.30	0.14	0.29	0.26	0.19	0.79	0.82	0.88	0.01	0.00	−0.07	0.52	0.48	0.45
tradition	0.28	0.28	0.24	0.40	0.39	0.32	0.79	0.74	0.72	−0.01	0.00	−0.01	0.50	0.48	0.44
enthusiastic	0.43	0.35	0.27	0.42	0.38	0.25	0.70	0.75	0.73	−0.02	−0.01	0.00	0.64	0.57	0.52
improvise	0.24	0.28	0.20	0.10	0.13	0.09	0.65	0.60	0.67	0.00	−0.01	0.00	0.49	0.52	0.45
haste	0.33	0.26	0.21	0.19	0.14	0.17	0.76	0.73	0.83	0.04	0.02	0.17	0.53	0.47	0.45
trend	0.33	0.27	0.20	0.38	0.35	0.36	0.71	0.71	0.70	0.11	0.01	−0.01	0.57	0.50	0.45
impulse	0.26	0.28	0.20	0.30	0.28	0.23	0.79	0.76	0.77	−0.03	0.02	0.03	0.49	0.49	0.44
ruminate	0.38	0.27	0.30	0.09	−0.04	−0.07	0.55	0.56	0.48	−0.02	−0.01	−0.03	0.49	0.45	0.46
mollify	0.46	0.24	0.28	0.35	0.09	0.19	0.68	0.63	0.65	0.09	0.01	0.05	0.58	0.50	0.54
extirpate	0.35	0.24	0.23	0.08	−0.03	0.07	0.58	0.56	0.56	0.07	0.03	−0.04	0.44	0.42	0.40
panacea	0.40	0.28	0.15	0.24	0.18	−0.01	0.65	0.65	0.62	0.06	0.01	−0.01	0.50	0.45	0.32
perfunctory	0.34	0.26	0.22	0.26	0.13	0.07	0.68	0.64	0.66	0.00	0.04	0.00	0.60	0.54	0.51
insipid	0.47	0.30	0.24	0.41	0.12	0.13	0.58	0.52	0.52	0.04	0.02	−0.02	0.65	0.53	0.52
pavid	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Note In the table, NaN means that WASI-II does not contain a response while '—' means that stimulus was not found in pre-trained word embedding.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Axelrod, B.N., 2002. Validity of the wechsler abbreviated scale of intelligence and other very short forms of estimating intellectual functioning. Assessment 9 (1), 17–23. http://dx.doi.org/10.1177/1073191102009001003, URL http://journals.sagepub.com/doi/10.1177/1073191102009001003.

Baroni, M., Dinu, G., Kruszewski, G., 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 238–247. http://dx.doi.org/10.3115/v1/p14-1023.

Botelho, A., Baral, S., Erickson, J.A., Benachamardi, P., Heffernan, N.T., 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. J. Comput. Assist. Learn. 39 (3), 823–840. http://dx.doi.org/10.1111/jcal.12793, URL https://onlinelibrary.wiley.com/doi/10.1111/jcal.12793.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36 (3), 181–204. http://dx.doi.org/10.1017/s0140525x12000477.

Davidson, C., 2009. Transcription: Imperatives for qualitative research. Int. J. Qual. Methods 8 (2), 35–52. http://dx.doi.org/10.1177/160940690900800206, URL http://journals.sagepub.com/doi/10.1177/160940690900800206.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL http://aclweb.org/anthology/N19-1423.

Firth, J.R., 1957. Papers in Linguistics: Language and Languages. Oxford University Piress, London.

Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10 (4), 507. http://dx.doi.org/10.2307/2331838, URL https://www.jstor.org/stable/2331838?origin=crossref.

Fisher, R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron (1), 3–32.

Ganesan, K., 2020. How to use tfidftransformer & tfidfvectorizer? In: AI Implementation. URL https://kavita-ganesan.com/tfidftransformer-tfidfvectorizer-usage-differences/#.Ysm6LnbMKUk.

Garcia, E.M., España-Bonet, C., Màrquez, L., 2015. Document-level machine translation with word vector models. In: Proceedings of the 18th Annual Conference of the

European Association for Machine Translation. Antalya, Turkey, pp. 59–66, URL https://aclanthology.org/W15-4908.

Garg, R., Kiwelekar, A.W., Netak, L.D., Bhate, S.S., 2021. Potential use-cases of natural language processing for a logistics organization. In: Gunjan, V.K., Zurada, J.M. (Eds.), Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2. Springer International Publishing, Cham, pp. 157–191. http://dx.doi.org/10.1007/978-3-030-68291-0_13.

Giorgetti, D., Sebastiani, F., 2003. Automating survey coding by multiclass text categorization techniques. J. Am. Soc. Inf. Sci. Technol. 54 (14), 1269–1277. http://dx.doi.org/10.1002/asi.10335, URL http://doi.wiley.com/10.1002/asi.10335.

Gontkovsky, S.T., 2017. Sensitivity of the wechsler abbreviated scale of intelligence-second edition (WASI-II) to the neurocognitive deficits associated with the semantic dementia variant of frontotemporal lobar degeneration: A case study. Appl. Neuropsychol.: Adult 24 (3), 288–293. http://dx.doi.org/10.1080/23279095.2016.1154857, URL https://www.tandfonline.com/doi/full/10.1080/23279095.2016.1154857.

Goodkind, A., Lee, M., Martin, G.E., Losh, M., Bicknell, K., 2018. Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In: Proceedings of the Society for Computation in Linguistics. (SCIL) 2018, pp. 12–22. http://dx.doi.org/10.7275/R56W988P, URL https://aclanthology.org/W18-0302.

Harris, Z.S., 1954. Distributional structure. WORD 10 (2–3), 146–162. http://dx.doi.org/10.1080/00437956.1954.11659520.

Hasson, R., Wu, I., Fine, J., 2019. Clinical utility of the WASI-II and its association with acculturation levels among Arab American adolescent males. Appl. Neuropsychol.: Child 8 (4), 295–306. http://dx.doi.org/10.1080/21622965.2018.1442219, URL https://www.tandfonline.com/doi/full/10.1080/21622965.2018.1442219.

Hwang, H., Kim, H., 2022. Automatic analysis of constructional diversity as a predictor of EFL students' writing proficiency. Appl. Linguist. 44 (1), 127–147. http://dx.doi.org/10.1093/applin/amac046.

Irby, S.M., Floyd, R.G., 2013. Test review: Wechsler abbreviated scale of intelligence, second edition. Canad. J. School Psychol. 28 (3), 295–299. http://dx.doi.org/10.1177/0829573513493982, URL http://journals.sagepub.com/doi/10.1177/0829573513493982.

Jackson, P., Moulinier, I., 2002. Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization, second ed. John Benjamins Publishing Company.

Jackson, K.M., Trochim, W.M.K., 2002. Concept mapping as an alternative approach for the analysis of open-ended survey responses. Organ. Res. Methods 5 (4), 307–336. http://dx.doi.org/10.1177/109442802237114, URL http://journals.sagepub.com/doi/10.1177/109442802237114.

Kronberger, N., Wagner, W., 2000. Keywords in context: Statistical analysis of text features. In: Bauer, M., Gaskell, G. (Eds.), Qualitative Researching with Text, Image and Sound. SAGE Publications Ltd, London, pp. 299–317. http://dx.doi.org/10.4135/9781849209731.n17, URL http://methods.sagepub.com/book/qualitative-researching-with-text-image-and-sound/n17.xml.

Krosnick, J.A., Presser, S., 2009. Question and questionnaire design. In: Wright, J.D., Marsden, P.V. (Eds.), Handbook of Survey Research, second ed. Elsevier, San Diego, CA.

Kuperberg, G.R., Jaeger, T.F., 2015. What do we mean by prediction in language comprehension? Lang., Cogn. Neurosci. 31 (1), 32–59. http://dx.doi.org/10.1080/23273798.2015.1102299.

Ladd, J.R., 2020. Understanding and using common similarity measures for text analysis. In: Walsh, B. (Ed.), Program. Hist. (9), http://dx.doi.org/10.46430/phen0089, URL https://programminghistorian.org/en/lessons/common-similarity-measures.

Lam, C., Nnamoko, N., 2024. Quantitative metrics to the CARS model in academic discourse in biology introductions. In: Braud, C., Hardmeier, C., Li, J.J., Loaiciga, S., Strube, M., Zeldes, A. (Eds.), Proceedings of the 5th Workshop on Computational Approaches to Discourse. (CODI 2024), Association for Computational Linguistics, St Julian's, Malta, URL https://aclanthology.org/2024.codi-1.7.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211–240. http://dx.doi.org/10.1037/0033-295x.104.2.211.

Lastra-Díaz, J.J., Goikoetxea, J., Hadj Taieb, M.A., García-Serrano, A., Ben Aouicha, M., Agirre, E., 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Eng. Appl. Artif. Intell. 85, 645–665. http://dx.doi.org/10.1016/j.engappai.2019.07.010, URL https://linkinghub.elsevier.com/retrieve/pii/S0952197619301745.

Levy, O., Goldberg, Y., 2014a. Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 302–308. http://dx.doi.org/10.3115/v1/p14-2050.

Levy, O., Goldberg, Y., 2014b. Neural word embedding as implicit matrix factorization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (Eds.), In: Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., pp. 2177–2185, URL https://proceedings.neurips.cc/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf.

Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P., 2004. The similarity metric. IEEE Trans. Inform. Theory 50 (12), 3250–3264. http://dx.doi.org/10.1109/TIT.2004.838101, URL http://ieeexplore.ieee.org/document/1362909/.

Li, B., Han, L., 2013. Distance weighted cosine similarity measure for text classification. In: Yin, H. (Ed.), Intelligent Data Engineering and Automated Learning – IDEAL 2013. In: Lecture Notes in Computer Science, vol. 8206, Springer, Berlin, Heidelberg, pp. 611–618. http://dx.doi.org/10.1007/978-3-642-41278-3_74, URL http://link.springer.com/10.1007/978-3-642-41278-3{_}74.

Loper, E., Bird, S., 2002. NLTK. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics -, vol. 1, Association for Computational Linguistics, Morristown, NJ, USA, pp. 63–70. http://dx.doi.org/10.3115/1118108.1118117, URL http://portal.acm.org/citation.cfm?doid=1118108.1118117.

Lord, C., Rutter, M., DiLavore, P.C., Risi, S., Gotham, K., Bishop, S.L., Luyster, R.J., Guthrie, W., 2012. Autism Diagnostic Observation Schedule, Second Edition (ADOS-2), Part 1: Modules 1–4, second ed. Western Psychological Services, Los Angeles, CA.

Losada, D.E., Crestani, F., Parapar, J., 2019. Overview of erisk 2019 early risk prediction on the internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10. Springer, pp. 340–357, URL https://ceur-ws.org/Vol-2380/paper_248.pdf.

Losada, D.E., Crestani, F., Parapar, J., 2020. Erisk 2020: Self-harm and depression challenges. In: Jose, J.M., Yilmaz, E., Magalhães, J.a., Castells, P., Ferro, N., Silva, M.J., Martins, F. (Eds.), Advances in Information Retrieval. Springer International Publishing, Cham, pp. 557–563.

Luhn, H.P., 1960. Key word-in-context index for technical literature (kwic index). Am. Doc. 11 (4), 288–295. http://dx.doi.org/10.1002/asi.5090110403, URL http://doi.wiley.com/10.1002/asi.5090110403.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28 (2), 203–208. http://dx.doi.org/10.3758/bf03204766.

Makrehchi, M., Kamel, M.S., 2008. Automatic extraction of domain-specific stopwords from labeled documents. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (Eds.), Advances in Information Retrieval. Springer, Berlin, Heidelberg, pp. 222–233. http://dx.doi.org/10.1007/978-3-540-78646-7_22, URL http://link.springer.com/10.1007/978-3-540-78646-7_22.

Mandera, P., Keuleers, E., Brysbaert, M., 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. J. Mem. Lang. 92, 57–78. http://dx.doi.org/10.1016/j.jml.2016.04.001.

Manning, Christopher D. Raghavan, Prabhakar Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, London.

McCrimmon, A.W., Smith, A.D., 2013. Review of the wechsler abbreviated scale of intelligence, second edition (WASI-II). J. Psychoeduc. Assess. 31 (3), 337–341. http://dx.doi.org/10.1177/0734282912467756, URL http://journals.sagepub.com/doi/10.1177/0734282912467756.

Mcdonald, S., Ramscar, M., 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In: Proceedings of the 23rd Annual Conference of the Cognitive Science Society. 611–6, URL https://api.semanticscholar.org/CorpusID:13161362.

McGeehan, B., Ndip, N., McGill, R.J., 2017. Exploring the multidimensional structure of the WASI-II: Further insights from schmid-leiman higher-order and exploratory bifactor solutions. Arch. Assess. Psychol. 7 (1), 7–27, URL https://api.semanticscholar.org/CorpusID:51948927.

Medveď, M., Horák, A., 2018. Sentence and word embedding employed in open question-answering. In: Rocha, A.P., van den Herik, H.J. (Eds.), Proceedings of the 10th International Conference on Agents and Artificial Intelligence. ICAART 2018, SciTe Press - Science and Technology Publications, Madeira, Portugal, pp. 486–492. http://dx.doi.org/10.5220/0006595904860492.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (Eds.), In: Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Munková, D., Munk, M., Vozár, M., 2014. Influence of stop-words removal on sequence patterns identification within comparable corpora. In: Trajkovik, V., Anastas, M. (Eds.), ICT Innovations 2013. Springer International Publishing, Heidelberg, pp. 67–76. http://dx.doi.org/10.1007/978-3-319-01466-1_6.

Nnamoko, N., Cabrera-Diego, L.A., Campbell, D., Korkontzelos, Y., 2019. Bug severity prediction using a hierarchical one-vs.-remainder approach. In: Métais, E., Meziane, F., Sunil, V., Sugumaran, V., Saraee, M. (Eds.), International Conference on Applications of Natural Language to Information Systems. NLDB, Springer, Manchester, pp. 247–260. http://dx.doi.org/10.1007/978-3-030-23281-8_20, URL http://link.springer.com/10.1007/978-3-030-23281-8_20.

Nowell, L.S., Norris, J.M., White, D.E., Moules, N.J., 2017. Thematic analysis: Striving to meet the trustworthiness criteria. Int. J. Qual. Methods 16 (1), 160940691773384. http://dx.doi.org/10.1177/1609406917733847, URL http://journals.sagepub.com/doi/10.1177/1609406917733847.

Onikoyi, B., Nnamoko, N., Korkontzelos, I., 2023. Gender prediction with descriptive textual data using a machine learning approach. Natural Lang. Process. J. 4, 100018. http://dx.doi.org/10.1016/j.nlp.2023.100018, URL https://www.sciencedirect.com/science/article/pii/S2949719123000158.

Paetzold, G., Specia, L., 2016. Semeval 2016 task 11: Complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation. (SemEVal-2016), Association for Computational Linguistics, San Diego, California, pp. 560–569. http://dx.doi.org/10.18653/v1/S16-1085, URL https://aclanthology.org/S16-1085.

Palachy, S., 2019. Document embedding techniques. Towards Data Sci. URL https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d.

Paulus, T.M., Lester, J.N., 2016. ATLAS.ti for conversation and discourse analysis studies. Int. J. Soc. Res. Methodol. 19 (4), 405–428. http://dx.doi.org/10.1080/13645579.2015.1021949, URL https://www.tandfonline.com/doi/full/10.1080/13645579.2015.1021949.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12 (85), 2825–2830, URL http://jmlr.org/papers/v12/pedregosa11a.html.

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1532–1543. http://dx.doi.org/10.3115/v1/D14-1162, URL http://aclweb.org/anthology/D14-1162.

Pérez, A., Parapar, J., Barreiro, Á., 2022. Automatic depression score estimation with word embedding models. Artif. Intell. Med. 132, 102380. http://dx.doi.org/10.1016/j.artmed.2022.102380, URL https://linkinghub.elsevier.com/retrieve/pii/S093336572200135X.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2227–2237. http://dx.doi.org/10.18653/v1/N18-1202, URL http://aclweb.org/anthology/N18-1202.

Porter, M., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137. http://dx.doi.org/10.1108/eb046814, URL https://www.emerald.com/insight/content/doi/10.1108/eb046814/full/html.

Prud'hommeaux, E., Van Santen, J., Gliner, D., 2017. Vector space models for evaluating semantic fluency in autism. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Vancouver, Canada, pp. 32–37. http://dx.doi.org/10.18653/v1/P17-2006, URL https://aclanthology.org/P17-2006.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners. Open AI, Massachusetts, URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Rajaraman, A., Ullman, J.D., 2011. Data mining. In: Mining of Massive Datasets. Cambridge University Press, pp. 1–17. http://dx.doi.org/10.1017/CBO9781139058452.002.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural topic models for open-ended survey responses. Am. J. Political Sci. 58 (4), 1064–1082. http://dx.doi.org/10.1111/ajps.12103, URL http://doi.wiley.com/10.1111/ajps.12103.

Rossiello, G., Basile, P., Semeraro, G., 2017. Centroid-based text summarization through compositionality of word embeddings. In: Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 12–21. http://dx.doi.org/10.18653/v1/W17-1003, URL http://aclweb.org/anthology/W17-1003.

Roy, D., 2017. Word embedding based approaches for information retrieval. In: Freire, A., Baeza-Yates, R. (Eds.), Proceedings of the Seventh BCS-IRSG Symposium on Future Directions in Information Access. FDIA 2017, BCS-IRSG, Barcelona, Spain, http://dx.doi.org/10.14236/ewic/FDIA2017.9.

Sahlgren, M., 2008. The distributional hypothesis. Italian J. Linguist. 20, 33–53, URL https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf.

Schuman, H., Presser, S., 1981. Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. Academic Press Inc, New York.

Shahamiri, S.R., Thabtah, F., 2020. Autism AI: A new autism screening system based on artificial intelligence. Cogn. Comput. 12 (4), 766–777. http://dx.doi.org/10.1007/s12559-020-09743-3, URL https://link.springer.com/10.1007/s12559-020-09743-3.

Sharratt, K., Boduszek, D., Retzler, C., 2020. Clarifying the relationship between psychopathy and intelligence using four dimensions of the WASI-II. Deviant Behav. 41 (5), 619–627. http://dx.doi.org/10.1080/01639625.2019.1582968, URL https://www.tandfonline.com/doi/full/10.1080/01639625.2019.1582968.

Shipurkar, G.M., Sheth, R.R., Surana, T.A., Shah, K.N., Garg, R., Natu, P., 2022. End to end system for handwritten text recognition and plagiarism detection using CNN & BLSTM. In: 2022 4th International Conference on Artificial Intelligence and Speech Technology. AIST, pp. 1–6. http://dx.doi.org/10.1109/AIST55798.2022.10064985.

Silva, C., Ribeiro, B., 2003. The importance of stop word removal on recall values in text categorization. In: Proceedings of the International Joint Conference on Neural Networks, 2003, vol. 3, IEEE, Portland, pp. 1661–1666. http://dx.doi.org/10.1109/IJCNN.2003.1223656, URL http://ieeexplore.ieee.org/document/1223656/.

Sonabend, A.W., Pellegrini, A.M., Chan, S., Brown, H.E., Rosenquist, J.N., Vuijk, P.J., Doyle, A.E., Perlis, R.H., Cai, T., 2020. Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec. In: Guloksuz, S. (Ed.), PLoS One 15 (4), e0230663. http://dx.doi.org/10.1371/journal.pone.0230663, URL https://dx.plos.org/10.1371/journal.pone.0230663.

Upton, G., Cook, I., 2008. A Dictionary of Statistics. Oxford University Press, London, http://dx.doi.org/10.1093/acref/9780199541454.001.0001, URL http://www.oxfordreference.com/view/10.1093/acref/9780199541454.001.0001/acref-9780199541454.

Ř.ehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50, URL http://is.muni.cz/publication/884893/en.

Vu, H., Abdurahman, S., Bhatia, S., Ungar, L., 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1512–1524. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.137, URL https://www.aclweb.org/anthology/2020.findings-emnlp.137.

Wang, H., Li, L., Chi, L., Zhao, Z., 2019. Autism screening using deep embedding representation. In: Rodrigues, J.a.M.F., Cardoso, P.J.S., Monteiro, J., Lam, R., Krzhizhanovskaya, V.V., Lees, M.H., Dongarra, J.J., Sloot, P.M. (Eds.), Computational Science – ICCS 2019. Springer International Publishing, Switzerland AG, pp. 160–173. http://dx.doi.org/10.1007/978-3-030-22741-8_12, URL https://link.springer.com/10.1007/978-3-030-22741-8_12.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA statement on p -values: Context, process, and purpose. Amer. Statist. 70 (2), 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108, URL https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

Wawer, A., Chojnicka, I., 2022. Detecting autism from picture book narratives using deep neural utterance embeddings. Int. J. Lang. Commun. Disord. 57 (5), 948–962. http://dx.doi.org/10.1111/1460-6984.12731, URL https://onlinelibrary.wiley.com/doi/10.1111/1460-6984.12731.

Wechsler, D., 2011. Wechsler Abbreviated Scale of Intelligence, second ed. Pearson, San Antonio, TX.

Wechsler, D., 2014. Wechsler Intelligence Scale for Children, fifth ed. Pearson, San Antonio, TX.

Wu, G., 2021. String similarity metrics. In: Baeldung. URL https://www.baeldung.com/cs/string-similarity-edit-distance.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. http://dx.doi.org/10.48550/arXiv.1609.08144, arXiv:1609.08144, URL http://arxiv.org/abs/1609.08144.

Zettersten, M., 2019. Learning by predicting: How predictive processing informs language development. In: Patterns in Language and Linguistics. De Gruyter, pp. 255–288. http://dx.doi.org/10.1515/9783110596656-010.

Züll, C., 2016. Open-ended questions. In: GESIS Survey Guidelines. GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany, http://dx.doi.org/10.15465/gesis-sg_en_002, URL https://www.gesis.org/fileadmin/upload/SDMwiki/Zuell_Open-Ended_Questionsa.pdf.