

Discovering Latent Lexical and Syntactic Factors in English Reading Comprehension Responses Using Unsupervised Machine Learning

Nora Darjazini 

Department of Educational Psychology, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: n.darjazini16@gmail.com
<https://orcid.org/0009-0007-4857-810X>

Mohammad Hossein Zarghami* 

Corresponding Author, Behavioral sciences Research Center, Life style institute, Baqiyatallah University of Medical Sciences, Tehran, Iran. E-mail: zar100@gmail.com
<https://orcid.org/0000-0001-8084-8148>

Reza Ghorban Jahromi 

Assistant Professor, Department of Educational and Personality Psychology, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: rrghorban@gmail.com
<https://orcid.org/0000-0002-0295-5295>

Leila Shobeiry 

Assistant Professor, Department of French Language, Faculty of Literature, Humanities and Social Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: shobeiri@srbiau.ac.ir
<https://orcid.org/0000-0002-4933-8104>

ABSTRACT

Reading comprehension is a complex interdisciplinary skill involving cognitive, linguistic, and educational processes. With the rapid advancement of artificial intelligence and machine learning, new opportunities have emerged to model and understand the linguistic underpinnings of comprehension. Unsupervised machine learning, in particular, offers the potential to uncover hidden patterns in learner language without relying on pre labeled data. Lexical and syntactic features extracted via natural language processing (NLP) have been widely used to analyze written and spoken language, but their latent structure in the context of reading comprehension responses remains underexplored. Previous research has often focused on individual, manually selected variables or supervised models, leaving the underlying dimensionality of a broad set of NLP derived features unexamined. This study aims to fill this gap by applying unsupervised factor analysis to a rich set of lexical and syntactic features extracted from EFL learners' constructed responses to reading comprehension questions. The study addresses two research questions: (1) What latent factors underlie the lexical and syntactic features of learners' spoken and written responses? (2) How do these factors predict reading comprehension performance?

Keywords: reading comprehension, unsupervised machine learning, natural language processing, latent factor, education

Cite this Article: Darjazini, N., Zarghami, M. H., Ghorban Jahromi, R., & Shobeiri, L. (2025). Discovering the latent lexical and syntactic factor structure in responses to English reading comprehension questions using unsupervised machine learning. *Technology of Instruction and Learning*. 9(31), 49-88
doi: <https://doi.org/10.22054/jti.2026.85160.1569>



by Allameh Tabataba'i University Press 2016 ©
Publisher: Allameh Tabataba'i University Press

Extended Abstract

Introduction

Reading comprehension is a complex interdisciplinary skill involving cognitive, linguistic, and educational processes. With the rapid advancement of artificial intelligence and machine learning, new opportunities have emerged to model and understand the linguistic underpinnings of comprehension. Unsupervised machine learning, in particular, offers the potential to uncover hidden patterns in learner language without relying on pre-labeled data. Lexical and syntactic features extracted via natural language processing (NLP) have been widely used to analyze written and spoken language, but their latent structure in the context of reading comprehension responses remains underexplored. Previous research has often focused on individual, manually selected variables or supervised models, leaving the underlying dimensionality of a broad set of NLP-derived features unexamined. This study aims to fill this gap by applying unsupervised factor analysis to a rich set of lexical and syntactic features extracted from EFL learners' constructed responses to reading comprehension questions. The study addresses two research questions: (1) What latent factors underlie the lexical and syntactic features of learners' spoken and written responses? (2) How do these factors predict reading comprehension performance?

Research Question(s)

1. What latent factors can be identified in the lexical and syntactic features extracted via NLP from EFL learners' constructed responses using unsupervised machine learning?
2. How do these factors predict learners' reading comprehension scores?

Literature Review

NLP techniques have proven effective in capturing fine-grained lexical and syntactic properties of text, such as vocabulary diversity, grammatical complexity, word specificity, and semantic similarity (Zhi-wei, 2006; Nugues, 2014). Unsupervised learning methods—

including clustering, association rules, and factor analysis—have been employed in educational data mining to identify hidden student profiles, detect outliers, and predict academic achievement (Vlachos, 2011; Hodeghatta & Nayak, 2017; Saarela et al., 2020). In language assessment, exploratory factor analysis (EFA) has been used to validate the construct validity of self-efficacy scales and to uncover underlying dimensions of learner perceptions (Seyd et al., 2015; Xiangdong et al., 2022). However, few studies have applied unsupervised ML to simultaneously model the latent factor structure of a large number of NLP-derived linguistic features from both spoken and written modalities and link these factors to objective reading comprehension measures. Recent work by Yapp et al. (2023) used supervised ML to predict reading comprehension from linguistic features, but the underlying dimensionality of those features was not examined. The present study extends this line of research by employing unsupervised factor analysis to reveal the latent organization of lexical-syntactic features and assess their predictive validity.

Methodology

Design and Participants. A quantitative causal-comparative design was adopted. Participants were 360 Iranian EFL learners (180 high proficiency, 180 low proficiency) recruited from language institutes in Tehran. High-proficiency learners were enrolled in advanced levels, while low-proficiency learners were in beginner levels; this categorization was confirmed by their teachers.

Instruments and Procedure. Reading comprehension was measured using the BALA (Balanced Assessment of Literacy Ability) instrument (Jang, 2019). High-proficiency learners completed the regular version ($\alpha = .80$), and low-proficiency learners completed the revised version ($\alpha = .85$). Participants answered open-ended comprehension questions based on narrative and expository texts. Additionally, a written opinion task on the use of social media by adolescents was administered to elicit extended written responses. Spoken responses were collected via two tasks—picture description and story retelling—delivered through a custom-built application. All spoken responses were transcribed verbatim.

Feature Extraction and Data Analysis. A total of 260 lexical and syntactic features were extracted from the transcribed responses using the COVFEFE (Sinclair, Jang, & Rudzicz, 2021) and quanteda (Benoit & Nulty, 2017) R packages. Features included indices of grammatical complexity (e.g., mean length of T-unit, number of complex nominals), lexical diversity (e.g., type-token ratio, moving average type-token ratio), word properties (e.g., age of acquisition, frequency, imageability, concreteness), and semantic similarity (WordNet-based measures). Unsupervised factor analysis with principal axis factoring was conducted separately on the spoken and written datasets using the psych package (Revelle & Revelle, 2015). The number of factors was determined via scree plots and the eigenvalue-greater-than-1 criterion. Factor scores were saved and entered as predictors in multiple linear regression analyses, with reading comprehension scores as the outcome variable.

Results

The purpose of tables and figures in documents is to enhance your readers' understanding of the information presented in the document. It is much lucid and efficient if the information is communicated in tables or figures.

Factor Structure. For both spoken and written datasets, scree plots and eigenvalues indicated four factors, accounting for 54% and 55% of the total variance, respectively.

Spoken responses: Factor 1 loaded strongly on grammatical complexity variables (e.g., complex nominals, sentence length); Factor 2 reflected lexical complexity (age of acquisition, word length, low-frequency words); Factor 3 was defined by word specificity and similarity (positive loadings) with negative loadings on personal pronouns and clausal complexity; Factor 4 captured lexical diversity (moving average type-token ratios).

Written responses: Factor 1 combined grammatical complexity and standard type-token ratio (negative loadings on moving average TTRs and word similarity); Factor 2 represented lexical diversity; Factor 3 comprised clausal complexity (clauses, verbs) with negative loadings

on text length; Factor 4 was characterized by word specificity and concreteness (positive) and negative loadings on personal pronouns and word frequency.

Regression Analyses. For spoken data, only Factor 3 (word specificity/similarity) significantly and negatively predicted reading comprehension ($\beta = -4.00$, $*p* < .01$); the model explained 7% of the variance (adjusted $R^2 = .07$, $F(4,90) = 2.79$, $*p* = .03$). For written data, Factors 1 (grammatical complexity; $\beta = -2.63$, $*p* = .02$), 3 (clausal complexity; $\beta = -3.04$, $*p* < .01$), and 4 (word specificity; $\beta = 2.49$, $*p* = .04$) were significant predictors, accounting for 13% of the variance (adjusted $R^2 = .13$, $F(4,94) = 4.72$, $*p* < .01$). Factor 2 (lexical diversity) was not significant in either dataset.

Discussion

The findings reveal that lexical and syntactic features in learner responses are organized into interpretable latent dimensions, with grammatical complexity emerging as a dominant factor across both modalities. This aligns with theoretical perspectives that syntax plays a foundational role in language production (Klein, 2008; Nation & Snowling, 2000). However, the predictive patterns differed by modality. In spoken responses, only word specificity/similarity negatively predicted comprehension, suggesting that overuse of specific, similar words may reflect a limited lexical repertoire or strategic avoidance of complex structures. In written responses, grammatical and clausal complexity negatively predicted comprehension, whereas word specificity positively predicted it. This may indicate that for written tasks, higher-proficiency learners use more specific vocabulary but also produce syntactically complex sentences that do not necessarily lead to higher comprehension scores—possibly due to task difficulty or cognitive load. The negative prediction of clausal complexity is counterintuitive but consistent with some prior studies showing that syntactic complexity does not always correlate with comprehension (Gottardo et al., 1996; Cain, 2007). The positive role of word specificity in written responses echoes research highlighting the importance of precise vocabulary in academic writing (Crossley et al., 2014). The absence of emotional/affective factors in

the latent structure, except in low-proficiency spoken data, suggests that such features may be task-dependent and less structurally stable (Harris, 2018). These results underscore the value of unsupervised ML for uncovering the hidden organization of linguistic features and provide a data-driven basis for designing targeted instructional interventions.

Conclusion

This study demonstrates that unsupervised machine learning, combined with NLP, can effectively identify latent lexical and syntactic factors underlying EFL learners' reading comprehension responses. Four factors were consistently extracted from both spoken and written datasets, explaining substantial variance. The predictive models revealed modality-specific relationships between these factors and reading comprehension, highlighting the importance of word specificity as a consistent predictor and the differential role of grammatical complexity. These findings have practical implications for reading instruction: encouraging the use of specific vocabulary, avoiding over-reliance on personal pronouns and lengthy sentences, and providing learners with varied lexical exposure. Limitations include the cross-sectional design and the lack of a control group. Future research should replicate these findings with diverse samples and longitudinal designs, and further explore the interaction between affective variables and linguistic complexity. The methodology presented here offers a novel framework for leveraging AI in educational assessment and personalized learning.

Acknowledgments

This article is derived from the first author's doctoral dissertation at the Islamic Azad University, Science and Research Branch. The authors gratefully acknowledge the support of the university and the language institutes that facilitated data collection.



کشف ساختار عوامل نهفته واژگانی و نحوی در پاسخ به سؤالات درک مطلب انگلیسی با استفاده از یادگیری ماشین بدون نظارت

نورا درجزینی

دانشجوی دکتری، گروه روانشناسی تربیتی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران. رایانامه: n.darjazini16@gmail.com

محمدحسین ضرغامی*

نویسنده مسئول، استادیار، مرکز تحقیقات علوم رفتاری، دانشگاه علوم پزشکی بقیه الله، تهران، ایران. رایانامه: zar100@gmail.com

رضا قربان جهرمی

استادیار گروه روانشناسی تربیتی و شخصیت، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه: rghorban@gmail.com

لیلا سهرابی

استادیار گروه تخصصی فرانسه، دانشکده ادبیات، علوم انسانی و اجتماعی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه: shobeiri@srbiau.ac.ir

چکیده

این مطالعه به بررسی ساختار عوامل نهفته واژگانی و نحوی در پاسخ به سؤالات درک مطلب انگلیسی با استفاده از تکنیک‌های یادگیری ماشین بدون نظارت می‌پردازد. طرح تحقیق مدلسازی کمی در قالب طرح علی-مقایسه‌ای است. داده‌ها از ۳۶۰ زبان آموز که انگلیسی زبان دوم آن‌ها به شمار می‌رود و از طریق پاسخ‌های کتبی و شفاهی آن‌ها به سؤالات درک مطلب جمع‌آوری شد. در مجموع ۲۶۰ ویژگی واژگانی و نحوی از داده‌ها استخراج شد. چهار عامل برای داده‌های شفاهی و چهار عامل برای داده‌های کتبی شناسایی شد که بیش از ۵۰ درصد از واریانس نمرات درک مطلب را توضیح می‌دهند. عوامل به شرح زیر نامگذاری و تفسیر شدند. برای داده‌های مربوط به پاسخ‌های شفاهی: عامل اول پیچیدگی دستوری، عامل دوم، پیچیدگی واژگانی، عامل سوم ویژگی و تشابه کلمات و عامل چهارم تنوع واژگانی. چهار عامل داده‌های مربوط به پاسخ‌های کتبی عبارتند از: عامل ۱، پیچیدگی دستوری، عامل ۲ تنوع واژگان، عامل ۳ پیچیدگی بند، و عامل ۴ اختصاصی بودن کلمه. تحلیل رگرسیون چندگانه در پیش‌بینی درک مطلب بر اساس عوامل کشف شده نشان داد که عامل سوم (ویژگی و تشابه کلمات) تنها پیش‌بینی‌کننده معنادار درک مطلب در داده‌های شفاهی بود، در حالی که عوامل ۱ (پیچیدگی دستوری)، ۳ (پیچیدگی بند) و ۴ (اختصاصی بودن کلمه) به طور معناداری درک مطلب را در داده‌های نوشتاری پیش‌بینی کردند. این یافته‌ها نشان داد که استفاده از کلمات خاص و مشابه با درک خواندن بالاتر در پاسخ‌های شفاهی و نوشتاری همراه است، در حالی که پیچیدگی دستوری و تنوع واژگان بسته به نحوه پاسخ، روابط متفاوتی با درک مطلب دارند. این مطالعه بینش‌های ارزشمندی را در مورد عوامل زمینه‌ای که به درک مطلب در زبان‌آموزان که انگلیسی زبان دوم آن‌ها است، کمک می‌کند و دستاوردهایی در توسعه روش‌های آموزش خواندن و ارزیابی مؤثرتر دارد.

کلیدواژه‌ها: درک مطلب، یادگیری ماشینی بدون نظارت، پردازش زبان طبیعی، عامل نهفته، آموزش.

استناد به این مقاله: درجزینی، نورا، ضرغامی، محمدحسین، قربان جهرمی، رضا، و شوبیری، لیلا. (۱۴۰۴). کشف ساختار عوامل نهفته واژگانی و نحوی در پاسخ به سؤالات درک مطلب انگلیسی با استفاده از یادگیری ماشین بدون نظارت، فناوری‌های آموزشی در یادگیری، ۹ (۳۱)، ۸۸-۴۹
doi: <https://doi.org/10.22054/jti.2026.85160.1569>



مقدمه

در زمان نگارش این مقاله، هوش مصنوعی بسیار رشد کرده است و متخصصان در حال تعامل با آن هستند، اما هنوز راه زیادی برای استفاده عمومی در پیش دارد. بسیاری از سیاستمداران درباره هوش مصنوعی حاکم بر جهان صحبت کرده‌اند. برخی از آنها از این پدیده‌ی محتمل استقبال کرده‌اند، درحالی که برخی دیگر نگران عواقب آن هستند. در سال ۲۰۱۷، ولادیمیر پوتین، رئیس جمهور روسیه در دیدار با دانشجویان گفت: "هوش مصنوعی یکی از فرصت‌ها و چالش‌های بزرگ برای بشریت است. کسی که در زمینه هوش مصنوعی پیشگام شود، بر جهان حکومت خواهد کرد." روزانه شاهد تولیدات علمی و فناوری در این زمینه هستیم و شرکت‌های آینده‌نگر دنیا بورسیه‌های مختلف تحصیلی و شغلی را برای محققان مشتاق ارائه می‌کنند.

یادگیری ماشین یکی از روش‌های هوش مصنوعی است که به ماشین‌ها توانایی یادگیری از داده‌ها و الگوریتم‌ها را می‌دهد. یادگیری ماشین در حوزه روان‌زبان‌شناسی یا روان‌شناسی زبان می‌تواند برای مطالعه و تحلیل زبان و رفتارهای زبانی انسان مورد استفاده قرار گیرد. برخی از کاربردهای یادگیری ماشین در روان‌زبان‌شناسی عبارتند از: الف) شناخت و تولید زبان طبیعی: این کاربرد شامل درک و پردازش متن یا گفتار و همچنین تولید متن یا گفتار به زبان انسان است. به عنوان مثال چت‌بات‌ها، ترجمه ماشینی، تحلیل احساسات و خلاصه‌سازی متن مربوط به این کاربرد است. ب) مدلسازی زبانی: این کاربرد شامل ساخت مدل‌های احتمالی برای توصیف و پیش‌بینی ساختار و توالی زبانی است. به عنوان مثال مدل‌های N-gram، مدل‌های مبتنی بر شبکه‌های عصبی و مدل‌های مبتنی بر تحلیل عاملی در این دسته قرار می‌گیرند و ج) یادگیری زبان: این کاربرست شامل بررسی و شبیه‌سازی فرآیندهای یادگیری زبان در انسان و ماشین می‌باشد. به عنوان مثال، یادگیری زبان اول و دوم، یادگیری زبان از صفر و یادگیری زبان از روی نمونه در این گروه قرار می‌گیرند. پژوهش حاضر در راستای کاربرد دوم یعنی مدلسازی زبان بوده و در آن ساختارهای پنهان در گفتار و نوشتار زبان‌آموزان در پاسخ به سوالات درک مطلب را مدلسازی کرده است.

درک مطلب به عنوان یک موضوع بین رشته‌ای در نظر گرفته می‌شود و شامل شاخه‌هایی مانند علوم شناختی، آموزش زبان، زبان‌شناسی کاربردی، نشانه‌شناسی، معناشناسی،

روانشناسی زبان و علوم کامپیوتر است. زمینه‌ی این تحقیق درک مطلب است و به طور خاص به توسعه روش‌هایی برای توضیح فرآیند درک محرک‌های دیداری مربوط به نوشتن (که علاوه بر متن می‌تواند نمادها، تصاویر، نمودارها و... باشد) پرداخته است. توضیح و تبیین چنین فرآیندی منجر به کاربردهایی مانند پیش‌بینی، کنترل، آموزش، اصلاح و بهبود و توسعه فرآیندهای شناختی مرتبط می‌شود.

این توانایی متکی بر دو توانایی دیگر است: خواندن کلمات یعنی توانایی رمزگشایی نمادهای مشاهده شده در صفحه و درک زبان که به معنای توانایی درک معنای کلمات و جملات است. درک مطلب، یادآوری دقیق مطالب خوانده شده نیست، بلکه خواننده یک مدل ذهنی از آنچه که متن توضیح می‌دهد با یکپارچه‌سازی درک کلمات و جملات و دستیابی به یک معنای کلی دارد که به آن درک مطلب گفته می‌شود. برای درک متن، درک مطلب نه تنها بسیار حیاتی است، بلکه تنها پس از درک مطلب می‌توان ارتباطی بین متن و خواننده برقرار کرد. با گسترش هوش مصنوعی و موفقیت آن در زمینه‌های مختلف دانش و به ویژه با تاکید بر الگوریتم‌های موفق در یادگیری ماشینی و یادگیری عمیق، نیاز به آزمون کاربرد این فناوری در زمینه اندازه‌گیری و یادگیری و آموزش زبان پر اهمیت‌تر است.

دو ساختار اساسی و مقدماتی در همه زمینه‌های زبانی عبارتند از حوزه لکسیس یا واژگان (عمق و دامنه کلماتی که انسان می‌تواند بفهمد و بازتولید کند واژگان نامیده می‌شود) و دیگری حوزه سینتکتیک یا نحو که نحوه نوشتن و قرار دادن صحیح کلمات و عبارات در یک زبان را مشخص می‌کند. این ساختارهای زبانی به‌طور گسترده در فرم‌ها یا قالب‌های نوشتاری و گفتاری با هدف فهم رابطه آنها با درک مطلب مورد مطالعه قرار گرفته‌اند. به عبارت دیگر، چه مقدار از واریانس درک مطلب را می‌توان از طریق ساختارهای ذکر شده که در پاسخ خوانندگان وجود دارد، توضیح داد. یکی از دلایل بررسی نشدن رابطه درک مطلب با ویژگی‌های زبانی ذکر شده در سوالات پاسخ باز را می‌توان دشواری استخراج شاخص‌ها از متون پاسخ باز دانست. به عبارت دیگر، هدف در اینجا، مانند سایر علوم، تبیین واریانس است. به عبارت دیگر، منشأ تغییرات درک مطلب فرد در موقعیت‌های مختلف یا بین افراد مختلف چیست و تا چه اندازه می‌توان واریانس درک مطلب را از طریق این منابع

توضیح داد؟ محوریت در این پژوهش، شناسایی عوامل پنهان زیربنای ویژگی های لغوی و نحوی پاسخ های تشریحی نوشتاری و گفتاری دانش آموزان در پاسخ به سوالات درک مطلب است که می توان آن را به صورت زیر خلاصه کرد:

"چه عوامل پنهانی را می توان در ویژگی های واژگانی و نحوی استخراج شده (با استفاده از رویکرد NLP) از پاسخ های تشریحی زبان آموزان با استفاده از روش ML بدون نظارت یافت؟ و این عوامل درک مطلب را چگونه پیش بینی می کنند؟"

پیشینه پژوهش

استخراج ویژگی های واژگانی و نحوی از طریق رویکرد پردازش زبان طبیعی تا کنون به کار گرفته شده است. حوزه رو به رشد پردازش زبان طبیعی اهمیت حوزه ی واژگانی و نحوی در درک زبان را پررنگ تر کرده است (Zhi-wei, 2006). برنامه های تشخیص سرقت ادبی از تکنیک های پردازش زبان طبیعی برای مقایسه ساختار نحوی و معنایی جملات استفاده می کنند و امکان شناسایی دقیق محتوای سرقت شده را فراهم می کنند (Yuen-Chi-Hon, 2007, Yan). مبانی تشخیص سرقت علمی حتی با انجام ویرایش روی متون سرقت شده، بر قوانین حاکم بر ساختار نحوی و معنایی چه به صورت جداگانه و چه به صورت ترکیبی متمرکز اند. اصول پردازش زبان طبیعی از یک طرف شامل مفاهیم کلیدی یادگیری ماشین مانند طبقه بندی و رگرسیون است که برای تجزیه و تحلیل و پردازش متن استفاده می شود (Nugues, 2014) و از طرف دیگر بر تجزیه و تحلیل نحوی و معنایی زبان که شامل وظایفی مانند برجسب گذاری بخشی از گفتار و ساختن درخت های نحوی است (Giersch, 2023) تمرکز دارد. رویکردهای پردازش زبان طبیعی همچنین از ویژگی های نحوی، مانند n-gram ها گفتار، برای تمایز بین انواع مختلف یک زبان استفاده می کنند (کریس، لی، آنتال، بوش، 2017).

تکنیک های یادگیری ماشین بدون نظارت، مانند قوانین خوشه بندی و تداعی، در پردازش زبان طبیعی (NLP) برای شناسایی الگوهای پنهان در داده ها استفاده می شوند (Umesh, Hodeghatta & Umesh, 2017). این تکنیک ها بخشی از تجزیه و تحلیل اکتشافی هستند و برای درک ویژگی های داده و شناسایی موارد پرت استفاده می شوند. در

این بین خوشه بندی به طور خاص به شناسایی گروه های پنهان در یک مجموعه داده کمک می کند (Vlachos, 2011). روش های یادگیری بدون نظارت در NLP برای یادگیری یک مدل به داده های برجسته گذاری شده نیاز ندارند با این وجود ارزیابی آنها با استفاده از استاندارد طلایی برجسته گذاری دستی دشوار است (Chen & Liao, 2016). یادگیری ماشینی مادام العمر (LML) یک حوزه تحقیقاتی است که هدف آن حفظ و انباشت دانش آموخته شده در گذشته برای کمک به یادگیری آینده است و پتانسیل هایی دارد که می تواند به پیشرفت عمده در وظایف NLP منجر شود (Emms, 2011). تکنیک تحلیل عاملی معمولاً در روش های یادگیری بدون نظارت برای تعیین ساختار عاملی زیربنایی پاسخ های زبان آموزان استفاده می شود (Chuyuan, 2019). در مطالعه ای که روی زبان آموزان انگلیسی انجام شد از شرکت کنندگان خواسته شد تا در مورد اهمیت آیتم های کتاب های درسی زبان انگلیسی اظهار نظر کنند و سپس از تحلیل عاملی اکتشافی (EFA) برای اعتبارسنجی ساختار مقیاس خودکارآمدی کتاب های درسی زبان آموزان انگلیسی استفاده شد. SES-ELLT (سیده و همکاران، 2015). نتایج EFA پنج عامل اصلی را نشان داد که بخش قابل توجهی از واریانس کل را تشکیل می دهند و شواهد تجربی برای عوامل الفاکتنده خودکارآمدی در طراحی کتاب های درسی و توسعه مواد درسی ارائه می دهند (Yimin & Zhou, Xiaoyan, Xiangdong, 2022). این پژوهش در راستای کاربرد یادگیری بدون نظارت، به ویژه تحلیل عاملی، در درک ابعاد اساسی پاسخ های زبان آموزان است که به اهداف و زمینه تحقیق حاضر نزدیک است.

یادگیری ماشین بدون نظارت و تحلیل عاملی در خاستگاه آموزشی برای درک عملکرد دانش آموز و شناسایی الگوهای پنهان در داده ها استفاده شده است. یادگیری بدون نظارت برای شناسایی اینکه چگونه ویژگی های جمعیت شناختی و مشارکت در فعالیت های یادگیری آنلاین بر پیشرفت دانش آموزان در موقعیت های آموزش از راه دور تأثیر می گذارد، استفاده کردند. Razaque و Alajlan (2020) عملکرد مدل های مختلف یادگیری ماشین را برای تجزیه و تحلیل و ارزیابی دستاوردهای دانش آموز مقایسه کردند و دریافتند که نزول گرادیان تصادفی بالاترین دقت را ایجاد می کند. Hodeghatta و Nayak (2017) خوشه بندی را به عنوان یک تکنیک بدون نظارت برای شناسایی گروه های پنهان در داده ها، از جمله داده های آموزشی، مورد بررسی قرار دادند. وو (2020) از

رگرسیون خطی برای تجزیه و تحلیل عوامل مؤثر بر نتایج آموزشی دانش آموزان استفاده کرد و دریافت که عادات مطالعه فردی مهم تر از تأثیرات خانوادگی است. Saarela و همکاران (2020) رویکرد جدیدی را با استفاده از الگوریتم های یادگیری بدون نظارت و با نظارت برای پیش بینی عملکرد دانش آموز در نمرات ریاضی با استفاده از داده های غنی شناختی که در بستر نظام آموزشی جمع آوری می شوند، پیشنهاد داد. این رویکرد در مطالعات ارزیابی آموزشی در مقیاس بزرگ کاربرد دارد.

روش های یادگیری بدون نظارت با هدف استخراج ابعاد زیربنایی ویژگی های واژگانی و نحوی در یادگیری زبان نیز استفاده شده است. هدف این روش ها کشف الگوها و ساختارها به طور مستقیم از داده های متن خام بدون نیاز به برچسب گذاری دستی و کاربست ابزارهای تجزیه و تحلیل است. به عنوان مثال یک رویکرد شامل استفاده از آمار بیزی ناپارامتریک و نمونه گیری گیس برای یادگیری واحدهای واژگانی برای تشخیص گفتار و ترجمه ماشینی بود. رویکرد دیگر از یک تکنیک تجزیه احتمالی برای ساخت واژگان و برچسب گذاری موارد مختلف در یادگیری گرامر دسته بندی (CG) استفاده می کند. علاوه بر این، از روش های معناشناسی محاسباتی برای محاسبه شباهت های بین کلمات و عبارات استفاده شده است که می تواند به درک ساختار زبان کمک کند (Riedl, 2016). این روش های بدون نظارت دارای چنان انعطاف پذیری و پتانسیل می باشند که برای پردازش زبان های کم منبع یا مجموعه های متنوع بدون نیاز به برچسب گذاری می باشند.

رویکردهای یادگیری ماشینی بدون نظارت (ML) در وظایف درک مطلب استفاده شده است. هدف این رویکردها شناسایی بخش های دشوار یک متن برای دانش آموزان و درک ایده های انتقادی اصلی است. یک مطالعه یک روش نحوی ساده و بدون نظارت را برای مشخص کردن جملات در درس ساز در یک متن پیشنهاد کرد. مطالعه دیگری از خوشه بندی k-means چند بعدی همراه با طبقه بندی بلوم برای تعیین مجموعه های مهارت های شناختی مثبت و منفی مرتبط با وظایف درک مطلب استفاده کرد.

رویکردهای ML بدون نظارت برای کشف ساختار عوامل نهفته واژگانی و نحوی در پاسخ به سؤالات درک انگلیسی نیز استفاده شده است. مقالات منتشر شده از این تحقیقات

بینش‌هایی را در مورد کاربرد رویکردهای ML بدون نظارت برای درک ساختار عوامل واژگانی و نحوی در پاسخ‌های درک انگلیسی ارائه می‌دهند. به عنوان نمونه در یکی از این مقالات رویکردی برای انتقال دانش پیشنهاد شده است که از مدل‌های تجزیه بدون نظارت برای استخراج برجسب‌گذاری بخش‌های مختلف متن استفاده می‌کند و یک شبکه عصبی سلسله مراتبی بازگشتی (HRNN) وجود دارد که از این برجسب‌گذاری استفاده می‌کند و شکاف بین تقسیم‌بندی نظارت‌شده و بدون نظارت را پر می‌کند (Anup et al, 2021). از سوی دیگر، این مقاله به بررسی نحوه‌ی فعال‌سازی ساختار نحوی در درک اصطلاحات واژگانی با استفاده از پتانسیل‌های مرتبط با رویداد می‌پردازد. نتایج نشان می‌دهد که ساختار نحوی مستقل از آشنایی با معنا و یا حتی شفاف بودن معنا فعال می‌شود (Meichao et al, 2017). علاوه بر این، در یک مطالعه سودمندی مدل‌سازی کلاس پنهان در حل مسائل اندازه‌گیری مرتبط با ساختارهای سلسله مراتبی مهارت‌ها، به‌ویژه در ارزیابی درک مطلب نشان داده شده است. جدول مربوط به خلاصه‌ی پیشینه تحقیق حاضر ارائه شده است.

جدول ۱: خلاصه یافته‌های مربوط به پیشینه پژوهش

منبع	یافته‌ها	حوزه کاربردی	روش تحقیق	عنوان مقاله / پژوهش	نویسنده (ها)	سال
Zhi-wei, 2006	استفاده از NLP به درک عمیق‌تر ساختارهای زبانی کمک می‌کند.	زبان‌شناسی کاربردی	NLP	نقش پردازش زبان طبیعی در درک ساختارهای واژگانی و نحوی	Zhi-wei	۲۰۰۶
Chi-Hon amp; Yuen-Yan, 2007	الگوریتم‌های NLP قادر به شناسایی سرقت ادبی حتی پس از ویرایش هستند.	هوش مصنوعی	NLP	تشخیص سرقت ادبی با استفاده از ویژگی‌های نحوی و معنایی	Chi-Hon, Yuen-Yan	۲۰۰۷
Nugues, 2014	NLP ابزاری قدرتمند برای تجزیه و تحلیل متن است.	علوم کامپیوتر	NLP	مبانی پردازش زبان طبیعی	Nugues	۲۰۱۱
Vlachos, 2011	خوشه‌بندی به شناسایی گروه‌های معنادار در داده‌های بدون برچسب کمک می‌کند.	داده‌کاوی و NLP	خوشه‌بندی بدون نظارت	استفاده از خوشه‌بندی در شناسایی الگوها در داده‌های متنی	Vlachos	۲۰۱۱
Emms, 2011	این رویکرد به بهبود یادگیری در مدل‌های	هوش مصنوعی	یادگیری مادام‌العمر	یادگیری ماشینی مادام‌العمر و	Emms	۲۰۱۴

کاربرد آن در NLP		NLP کمک می‌کند.	
۲۰۱۵	Seyd et al.	اعتبارسنجی مقیاس خودکارآمدی کتاب‌های درسی انگلیسی	EFA آموزش زبان انگلیسی
Seyd et al, 2015		تحلیل عاملی اکتشافی (EFA) به شناسایی ۵ عامل اصلی کمک کرد.	
۲۰۱۶	Chen, Liu	چالش‌های ارزیابی مدل‌های بدون نظارت	یادگیری بدون نظارت
Chen & Liu, 2016		بدون استاندارد طلایی، ارزیابی مدل‌ها دشوار است.	
۲۰۱۷	Umesh .et al	کاربرد تکنیک‌های بدون نظارت در NLP	پردازش زبان طبیعی
Umesh et al, 2017		این تکنیک‌ها برای شناسایی الگوهای پنهان در متن استفاده می‌شوند.	
۲۰۱۷	Hodeghatta, Nayak	استفاده از خوشه‌بندی در داده‌های آموزشی	داده‌کاوی آموزشی
Hodeghatta & Nayak, 2017		خوشه‌بندی به شناسایی گروه‌های پنهان در داده‌های آموزشی کمک کرد.	
۲۰۱۹	Chuyuan	استفاده از تحلیل عاملی در پاسخ زبان‌آموزان	EFA آموزش زبان
Chuyuan, 2019		تحلیل عاملی به شناسایی ابعاد پنهان در پاسخ‌ها کمک کرد.	

۲۰۲۰	Razaque, Alajlan	مقایسه مدل های ML در تحلیل دستاوردهای دانش آموز	یادگیری بدون نظارت و با نظارت	داده کاوی آموزشی	نزول گرادیان تصادفی بالاترین دقت را داشت.	Razaque & Alajlan, 2020
۲۰۲۰	Saarela et al	پیش بینی عملکرد دانش آموز با استفاده از داده های غنی	یادگیری بدون نظارت و با نظارت	آموزش و پردازش داده	تلفیق روش ها به بهبود دقت پیش بینی کمک کرد.	Saarela et al, 2020
۲۰۲۱	Anup et al	استفاده از HRNN در تحلیل متن	شبکه عصبی + بدون نظارت	NLP	HRNN قابلیت استفاده بدون برچسب گذاری دستی را فراهم کرد.	Anup et al, 2021
۲۰۲۲	Xiangdong et al	استفاده از EFA در تحلیل پاسخ زبان آموزان	EFA	آموزش زبان انگلیسی	۵ عامل اصلی از مجموعه داده ها استخراج شدند.	Xiangdong et al, 2022

روش

روش تحقیق این پژوهش مدل سازی کمی است که در قالب طرح علی-مقایسه ای اجرا شده است. تمرکز این پژوهش بر روی روش شناسی در زمینه یادگیری زبان است و داده ها با هدف مدل سازی درک مطلب زبان انگلیسی دانش آموزان دوره دوم متوسطه جمع آوری شده است. اساس مقایسه، استفاده از تکنیک های مختلف یادگیری ماشین است که همگی مبتنی بر روابط هستند و از این رو، تحقیق حاضر یک تحقیق همبستگی در نظر گرفته می شود. برای درک جنبه های مختلف روش شناسی این تحقیق، لازم است طرح تحقیق مشخص شود. در اینجا متغیر اصلی تحقیق (متغیر وابسته) درک مطلب دانش آموزان است. تحقیقات نشان

می‌دهد که نحوه پاسخ زبان آموزان (شفاهی یا نوشتاری) با اتخاذ راهبردهای درک مطلب مرتبط است. بدین منظور در این پژوهش زبان آموزان شرکت کننده در تحقیق به دو صورت شفاهی و کتبی به سوالات پاسخ داده‌اند.

ویژگی‌های واژگانی و نحوی پاسخ‌های زبان آموزان از طریق پردازش زبان طبیعی استخراج شده و به عنوان داده‌های خام تحلیل عاملی در نظر گرفته می‌شود. براساس طرح جمع‌آوری داده‌ها، چهار مجموعه داده جمع‌آوری شد که شامل مجموعه پاسخ‌های کتبی زبان آموزان با مهارت بالا، مجموعه پاسخ‌های شفاهی زبان آموزان با مهارت بالا، مجموعه پاسخ‌های کتبی زبان آموزان با مهارت پایین و مجموعه‌ای از پاسخ‌های شفاهی زبان آموزان با مهارت پایین می‌باشد. ویژگی‌های واژگانی و نحوی مبتنی بر پردازش زبان طبیعی از این چهار مجموعه داده استخراج شده است. ۱۸۰ نفر زبان آموز داوطلب با سطح مهارت درک مطلب زبان بالا و ۱۸۰ دانش آموز با مهارت درک مطلب پایین به‌طور داوطلبانه از بین زبان آموزانی که به مراکز آموزش زبان در تهران مراجعه کرده بودند و سطح تحصیلات آنها متوسطه دوم بود انتخاب شدند. در تحقیقات مشابه، حداکثر حجم نمونه ۵۰ نفر در هر گروه در نظر گرفته شده است (Yapp et al, 2021, Gomez Merino, Fajardo & Ferrer, 2021, Trukenmiller, Shen and Sweet, 2021). اما با توجه به احتمال ترک دوره‌ی زبان آموزی، افزایش پایایی پارامترهای آماری برآورد شده و همچنین افزایش قدرت تعمیم نتایج در این تحقیق تعداد ۱۸۰ دانش آموز برای هر گروه انتخاب کرده‌ایم. زبان آموزانی که دوره‌ی زبان آموزی آن‌ها ترم‌های بالا و نزدیک به اتمام بود به عنوان گروه نمونه با مهارت‌های زبانی بالا در نظر گرفته شد و ترم‌های یک و دو به عنوان گروه با مهارت درک مطلب پایین در نظر گرفته شد. این معیار تقسیم‌بندی به تأیید معلمان زبان انگلیسی رسیده بود. با وجود این علت گروه‌بندی شرکت کنندگان فرض محققان است. پژوهشگران فرض کرده‌اند که یادگیری زبان در سطوح بالا به مهارت‌های زبانی بیشتری نسبت به سطوح پایین نیاز دارد.

BALA که ابزاری جدید و جهانی برای اندازه‌گیری درک مطلب در زبان انگلیسی است و دارای دو فرمت مختلف است برای اندازه‌گیری درک مطلب زبان آموزان استفاده شد. این ابزار امکان اندازه‌گیری درک مطلب را از پاسخ زبان آموزان به صورت مداد کاغذی و تست

آنلاین کامپیوتری دارد. BALA از طریق یک متن روایی و یک متن غیرداستانی مهارت های مختلف درک مطلب را به دست می آورد. دو نسخه از BALA تاکنون توسعه یافته است، نسخه معمولی و نسخه اصلاح شده. نسخه معمولی بلا از ۱۸ سوال چند گزینه ای (با مقدار آلفای کرونباخ ۰.۸) و نسخه اصلاح شده بلا از ۱۷ سوال چند گزینه ای (با مقدار پایایی ۰.۸۵) تشکیل شده است. نسخه معمول ارزیابی سواد تعادل (BALA) (جانگ، ۲۰۱۹) برای زبان آموزانی است که مشکلات جدی در خواندن (مانند مشکلات خواندن) ندارند و مهارت های یادگیری زبان نسبتاً خوبی دارند. نسخه اصلاح شده BALA برای زبان آموزانی استفاده می شود که ممکن است مهارت های زبانی پایین یا مشکلات جدی در خواندن داشته باشند (این افراد بیشتر به ترم های پایین آموزش زبان تعلق دارند). در پژوهش ما زبان آموزان با مهارت بالا از نسخه معمولی آن استفاده کرده اند و زبان آموزان با مهارت پایین از نسخه ی اصلاح شده. از طرف دیگر، هر زبان آموز باید به سؤالات درک مطلبی که از بلا دریافت می کند به صورت باز پاسخ، پاسخ دهد. چرا که هدف این تحقیق تعیین ابعاد زیربنایی ویژگی های استخراج شده از متن پاسخ به سؤالات درک مطلب است و بنابراین لازم است که ابتدا پاسخ ها تشریحی باشد.

ابزار دوم یک تکلیف نوشتاری جداگانه است. در این تکلیف فیلمی به زبان آموز نشان داده می شود که در مورد استفاده دانش آموزان از شبکه های اجتماعی است. بعد از نمایش فیلم سپس از آنها پرسیده شد: «به نظر شما استفاده از شبکه های اجتماعی برای نوجوانان خوب است یا بد؟» پاسخ های تشریحی زبان آموزان به این سؤالات به صورت تشریحی و پاسخ باز و پاسخ آنها به قضاوت درباره استفاده نوجوانان از شبکه های اجتماعی به عنوان منبع داده های خام برای استخراج ویژگی های زبانی مبتنی بر متن (نوشتاری) در نظر گرفته می شود.

به منظور به دست آوردن داده های خام براساس گفتار (پاسخ های شفاهی) زبان آموزان، از اپلیکیشنی که برای دانش آموزان دبیرستانی طراحی شده بود، استفاده شد. سه چالش اصلی در استفاده از این اپلیکیشن وجود داشت: الف) اطمینان از بیان صحیح و دقیق دستورالعمل ها به زبان آموز، ب) ضبط صدای زبان آموز و ج) هزینه ساخت ابزار. در مواجهه با این چالش ها دستورالعمل انجام آزمون توسط یک معلم زبان خوانده شد و با کیفیت بالا ضبط شد و این صوت برای هر یک از شرکت کنندگان پخش شد. امکان پخش مجدد

دستور و متون شفاهی در صورت درخواست زبان آموز وجود داشت. زبان آموز می‌بایست دو تکلیف را انجام دهد: تکلیف اول، توضیح تصویر و دیگری تعریف مجدد داستانی که می‌شنود. بیان شفاهی زبان آموز به این دو محرک، مبنای استخراج ویژگی‌های زبانی (واژگانی و نحوی) شد. به منظور سنجش روایی صوری نیز مصاحبه‌ای با زبان آموزان جامعه‌ی هدف انجام شد که نشان می‌داد تکالیف برای موقعیتی که آن‌ها در آن قرار دارند مناسب است.

در مرحله اول قبل از تحلیل اصلی، پیش‌پردازش بر روی داده‌های جمع‌آوری شده انجام شد. همانطور که گفته شد، داده‌های جمع‌آوری شده فایل‌های متنی بود که از پاسخ‌های شفاهی و کتبی زبان‌آموزان به سؤالات درک مطلب جمع‌آوری شده بود. بسته NLP استخراج‌کننده ویژگی (فیچر) در استخراج متغیر اصلی (COVFEFE1) (Sinclair, Benoit & Rudzicz, 2021) و تحلیل کمی داده‌های متنی (Quanteda) (Nulty, 2017) که یک بسته تحت زبان برنامه نویسی R است برای پیش‌پردازش استفاده شد. از این بسته‌ها برای استخراج ویژگی‌هایی استفاده می‌شود که به آنها فیچر نیز گفته می‌شود. ویژگی‌ها در اینجا ویژگی‌های واژگانی و نحوی زبان است که الگوریتم استخراج آن قبلاً تحت برنامه‌های مختلف رایانه‌ای تنظیم و توسعه یافته است. پس از استخراج ویژگی‌ها، می‌توان بر روی آنها تجزیه و تحلیل کرد. استخراج ویژگی اصلی روشی است که برای کاهش داده‌ها بدون از دست دادن اطلاعات حیاتی استفاده می‌شود. تمام روش‌ها و تکنیک‌هایی که برای این منظور استفاده می‌شوند استخراج ویژگی نامیده می‌شوند که در بهبود کارایی و دقت مدل در مدل‌های یادگیری ماشین نقش اساسی دارند. تکنیک‌های استخراج ویژگی با مجموعه‌ای از داده‌های پایه اندازه‌گیری شده شروع می‌شوند و مقادیر یا ویژگی‌ها را به گونه‌ای تولید می‌کند که با مقدار متغیرهای کمتر همان داده‌های اصلی در مجموعه‌ای با متغیرهای بیشتر را بازتولید کند و به این صورت از متغیرهای اضافی و تکراری کم می‌کند یا آن‌ها را در متغیری جدید خلاصه می‌کند.

در مجموع ۲۶۰ ویژگی واژگانی و نحوی از پاسخ دانش‌آموزان به سؤالات درک مطلب به دست آمد. پس از به دست آوردن این ویژگی‌ها در مرحله اول تحلیل با استفاده از رویکرد ML بدون نظارت، ساختار عوامل پنهان در ویژگی‌های استخراج شده شناسایی و از آن

عوامل برای پیش‌بینی درک مطلب استفاده شد. تحلیل عاملی به طور جداگانه بر روی هر دو مجموعه داده (شفاهی و کتبی) انجام شد. نمودارهای Scree با استفاده از بسته Psych (Revelle & Revelle, 2015) در R برای تعیین تعداد مناسب عوامل ترسیم شدند. بارهای عاملی و امتیازات عاملی با استفاده از برنامه تحلیل عاملی Scikit-learn محاسبه شد (Buitinck et al, 2013). همبستگی معناداری بین ابعاد کشف شده وجود نداشت و ابعاد چرخش داده نشد. در ادامه با توجه به تعداد زیاد متغیرها، تنها بارهای عاملی بزرگ‌تر از 0.5 گزارش شده است. مرحله نهایی تجزیه و تحلیل، انجام یک رگرسیون خطی چندگانه نمرات درک مطلب بر روی نمرات عامل بود به عبارتی از طریق عوامل به مدل‌بندی پیش‌بینی درک مطلب پرداخته شد. چارت زیر خلاصه روش تحقیق این مطالعه را به تصویر کشیده است.

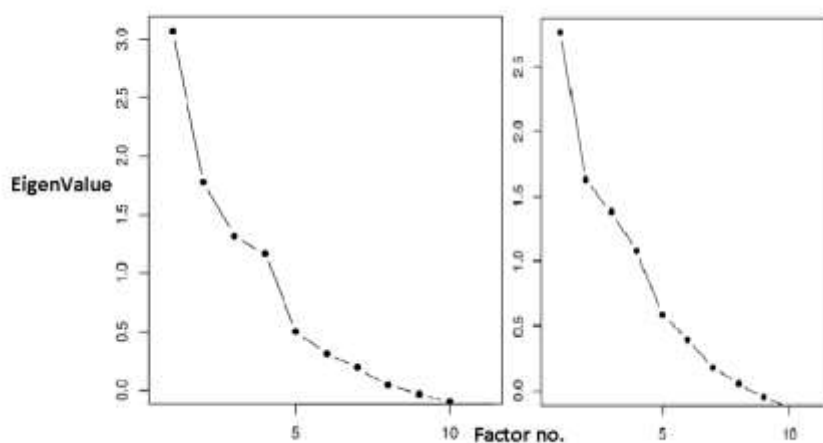
نمودار ۱: چارت روش تحقیق مطالعه حاضر



یافته‌ها

نمودارهای زیر به عنوان بازنمایی کلی بارهای عاملی به تفکیک هر مجموعه داده ارائه شده است و نحوه بارگذاری ویژگی‌های زبانی (مولفه دستوری، پیچیدگی دستوری، غنا و دامنه واژگان، ویژگی واژگانی، شباهت و ابهام و واژگان احساسات و عواطف) را نشان می‌دهد. همان‌طور که در بالا ذکر شد، به دلیل تعداد زیاد متغیرها (۲۶۰ فیچر)، تنها بارگذاری‌های بزرگتر از 0.5 در این بحث گنجانده شده است. با این حال، هنگامی که از نمرات عامل برای رگرسیون چندگانه در بخش بعدی استفاده می‌شود، همه متغیرها وجود خواهند داشت.

نمودار ۲: نمودار اسکری مربوط به ابعاد کشف شده (سمت راست داده‌های کتبی و سمت چپ داده‌های شفاهی)



شکل بالا مقادیر ویژه هر عامل را در مجموعه داده‌های شفاهی و نوشتاری نشان می‌دهد. انتخاب تعداد عوامل را می‌توان به صورت بصری از نمودار اسکری تعیین کرد. بعد از سنگریزه شدن نمودار تعداد زیادی عامل وجود دارد که واریانس معناداری را تبیین نمی‌کنند. تعداد ابعادی که به طور معنادار واریانس فیچرها را تبیین می‌کنند چهار بعد است. روش رایج دیگر برای یافتن تعداد فاکتورها استفاده از نقطه برش مقدار ویژه بزرگتر یا مساوی ۱ است. طبق این روش نیز تعداد فاکتورها ۴ عامل است. مقدار واریانس توضیح داده شده توسط هر عامل در جداول زیر بر اساس نوع داده نشان داده شده است.

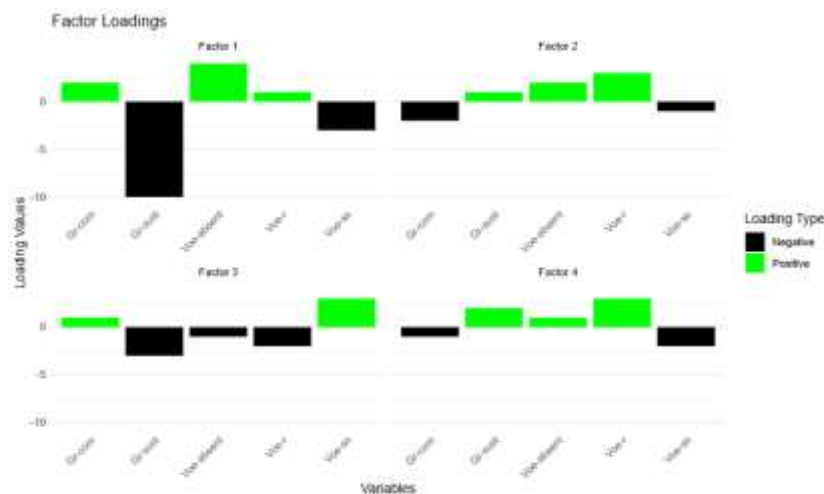
جدول ۲: میزان واریانس تبیین شده توسط هر یک از ابعاد کشف شده (۴ عامل) و میزان واریانس تجمعی آنها در داده‌های شفاهی و نوشتاری.

نسبت واریانس تبیین شده		نسبت واریانس تراکمی تبیین شده	
داده‌های متنی	Factor1	۰/۲۳	۰/۲۳
	Factor2	۰/۱۵	۰/۳۸
	Factor3	۰/۱	۰/۴۸
	Factor4	۰/۰۷	۰/۵۵

۰/۲	۰/۲	Factor1	داده های شفاهی
۰/۳۶	۰/۱۶	Factor2	
۰/۴۷	۰/۱۱	Factor3	
۰/۵۴	۰/۰۷	Factor4	

در هر دو مجموعه داده چهار عامل بیش از ۵۰ درصد از کل واریانس داده های درک مطلب را توضیح می دهند. این مقدار توضیحی با توجه به اینکه بیش از ۲۰۰ متغیر در این تحقیق داریم کافی، مناسب و قابل ملاحظه است. از آنجایی که جداول بارهای عاملی برای تمام عامل ها طولانی می شود از ذکر آن ها در این مقاله خودداری شده است. در نمودارهای زیر ویژگی هایی که مقدار بار عاملی آنها در هر یک از ابعاد کشف شده بالاتر از 0.5 است، نشان داده شده است.

نمودار ۳: بارهای عاملی بالاتر از 0.5 توسط عوامل در داده های شفاهی (بالا) و داده های نوشتاری (پایین)



در مجموعه داده های شفاهی، عامل اول شامل تمام متغیرهای دستوری پیچیده است. به طور شهودی می توان تایید کرد که این متغیرها با یکدیگر مرتبط هستند. تعداد کلمات نیز در این بعد بار بالایی دارد که دقیقاً یک ویژگی مربوط به پیچیده نیست و از این رو جالب

است به این نکته اشاره کنیم که طول روایت (تعداد کلمات) در موقعیت های مربوط به پاسخ های شفاهی با متغیرهای پیچیدگی گرامری مرتبط است.

عامل دوم دارای بار عاملی مثبت با متغیرهای حوزه واژگانی (سن یادگیری و طول کلمه) است و کلماتی که در فرهنگ لغت وجود ندارند روی این عامل بار دارند. متغیرهای مرتبط با لغت^۱ بر روی این عامل بار منفی دارند به این معنی که وقتی سن یادگیری بالاتر است آشنایی، فراوانی و تصویرپذیری مقدار پایین تری دارد. جالب اینجاست که به نظر می رسد پیچیدگی کلمات و استفاده از کلمات کم تنوع در ارتباط با یکدیگر قرار دارند. شاخص نسبت نوع-توکن^۲ دارای بارهای عامل منفی و شاخص برونه^۳ دارای بار عاملی مثبت است. اگر بخواهیم راهی را برای بررسی مزیت کاربرد زبانی کلمات پیچیده بیابیم می توانیم از بارهای عاملی منفی اشکال پیچیده، طول کلمه و عبارات اضافه استفاده کنیم.

میانگین ابهام کلمه با تصویرپذیری و آشنایی رابطه منفی دارد که با بحث قبلی در این زمینه همخوانی ندارد. چولگی و تیزی ابهام کلمه دارای بار عاملی مثبت است که نشان می دهد استفاده زیاد از کلمات با ابهام زیاد (به جای استفاده ی متوسط از کلمات با ابهام زیاد) در راستای سن یادگیری است. کلمات تابعی^۴ و حروف (مانند حروف اضافه، حروف ربط، افعال کمکی و...) به همراه جملات ساده اعلانی^۵ روی این عامل بار منفی دارند. به طور خلاصه، عامل دوم در درجه اول به پیچیدگی کلمات^۶ مربوط است که به نظر می رسد با طول عبارات اسمی و عبارات اضافه^۷ همبستگی منفی داشته باشد (نتیجه اینکه در صورتی که نتوان کلمات را به خاطر آورد، امتیاز کاربرد عبارات طولانی از این نوع می سوزد). جملات ساده خبری با این عامل رابطه منفی دارند که بر اساس آن می توان گفت کاربرد اشکال ساده دستوری با استفاده از واژگان پیچیده رابطه منفی دارند.

¹ vocabulary

² type-token ratio

³ Bronte index

⁴ function words

⁵ Simple declarative sentences

⁶ vocabulary complexity

⁷ length of noun phrases and prepositional phrases

عامل ۳ در درجه اول با اختصاصی بودن^۱ و تشابه کلمات ارتباط مثبت با بار عاملی مثبت و با پیچیدگی دستوری و برخی از اجزای دستوری ارتباط منفی با بار منفی دارد. برای عامل سوم قوی ترین همبستگی ها منفی است که به طور خاص به استفاده از ضمائر شخصی مربوط می شود. برخی موارد مربوط به پیچیدگی گرامری نیز روی این عامل منفی بار می شوند که ممکن است به دلیل استفاده از ضمائر شخصی باشد، زیرا ضمائر شخصی می توانند به عنوان مرجع جملات عمل کنند. طول کل روایت (قصه ای که زبان آموز بیان می کند) و نسبت نوع به توکن یا علامت متحرک در پنجره ۲۰ کلمه ای نیز در این جهت بارگذاری می شود. تراکم جمله (که نسبت افعال، صفت، قید، حرف اضافه و حروف ربط به تعداد کل کلمات است) نیز دارای بار منفی با این عامل است. بارهای مثبت عامل ۳ با میانگین ویژگی و شباهت کلمه (و تنوع شباهت)، مؤلفه های اسمی و تعیین کننده ها بود. براساس نتایج بدست آمده، به طور کلی تفسیر این عامل دشوار است اما می توان آن را به عنوان عاملی تعبیر کرد که بارهای مثبت به استفاده از اسم ها، شباهت و ویژگی کلمات مربوط است و بارهای منفی به پیچیدگی دستوری، استفاده از ضمائر شخصی و تنوع واژگان. به طور شهودی نیز می توان دریافت که اگر کلماتی که فرد بکار می برد مشابه یکدیگر باشند، تنوع آنها کمتر است. یکی از نتیجه های جالب توجه این است که تعداد اسم ها با جملات کوتاه تر، بندهای کمتر در هر جمله و کلمات کلی ارتباط بیشتری دارد که به طور شهودی این موضوع قابل فهم است چون اسامی قبل از افعال آموخته می شوند یادگیری آنها آسانتر است.

عامل چهارم با نسبت میانگین نوع- نشانه متحرک نوع به توکن چند پنجره ای (۳۰، ۴۰ و ۵۰ کلمه) که نشان دهنده تنوع واژگانی است دارای همبستگی و بار عاملی مثبت بود که هر سه متغیر مربوط به عبارات اسمی پیچیده است. بارهای منفی برای عامل ۴ شامل افعال، میانگین و واریانس برانگیختگی است که از طریق تغییر وضعیت آرام تا هیجان زدگی اندازه گیری می شود.

در داده های کتبی زبان آموزان، عامل اول تنها دارای دو بار عاملی مثبت است و سایر بارهای عاملی آن با فیچرها منفی است. دو فیچری که ارتباط مثبت با عامل اول دارند عبارتند از: تعداد واحدهای T استاندارد (که با پیچیدگی دستوری رابطه معکوس دارد) و نسبت نوع به توکن. بارهای منفی این عامل با متغیرهایی مانند پیچیدگی دستوری، نسبت نوع به نشانه

¹. specificity

کشف ساختار عوامل نهفته واژگانی... | ضرغامی و همکاران | ۷۳

متحرک (پنجره‌های ده، سی، چهل و پنجاه کلمه)، میانگین و واریانس شباهت کلمه، واریانس در ابهام اسم، حروف اضافه و شاخص برون‌ته است. فهرست مطالب. عامل ۳ با تعداد بندها، افعال و عبارات جاری ارتباط و بار عاملی مثبت و با تعداد کلمات، میانگین طول بندها و نسبت نوع به نشانه متحرک پنجره چهل کلمه ارتباط و بار عاملی منفی دارد. عامل چهارم با ویژگی کلمه دارای بار عاملی مثبت است و بارهای منفی آن به کاربرد ضمائر شخصی و بسامد کلمه یا فراوانی کلمه مربوط است. تنها عاملی که با ویژگی دامن‌های کلمات در ارتباط است، عامل چهارم است که با فراوانی کلمات مربوط است. از مجموعه داده‌های کتبی زبان آموزان با سطح مهارت پایین عاملی وجود ندارد که به دسته یا طبقه‌ای احساسات و عواطف متعلق باشد.

نتایج مربوط به مدل رگرسیونی درک مطلب براساس چهار بعد کشف شده براساس نوع داده‌های جمع‌آوری شده در جدول زیر نشان داده شده است.

جدول ۳: نتایج مدل‌سازی رگرسیون چندگانه به تفکیک مجموعه داده‌های شفاهی و متنی

مجموعه داده‌های شفاهی				مجموعه داده‌های کتبی				عوامل
P	T-	SE	B	P	T-	SE	B	
val ue				val ue				
عامل ۱	۰/۴۲	۱/۰۹	۰/۳۰	۰/۷۶	-۲/۶۳	۱/۱۴	-۲/۳۴	۰/۲*
عامل ۲	-۰/۲۹	۱/۰۱	-۰/۲۰	۰/۸۵	-۱/۶۳	۱/۱۴	-۱/۴۳	۰/۱۱
عامل ۳	-۴	۱/۳۲	-۳/۰۶	<۰/۰۱**	-۳/۰۴	۱/۱۵	-۲/۶۴	<۰/۰۱**
عامل ۴	-۱/۵۸	۰/۹۹	-۱/۳۰	۰/۲۰	۲/۴۹	۱/۱۶	۲/۱۵	۰/۰۴*
ثابت	۷۴/۴۵	۱/۳۲	۶۲/۰۴	<۰/۰۱	۷۴/۱۹	۱/۱۴	۶۵/۲۹	<۰/۰۱
R ^۲)				۰/۱۳				
اصلاح شده)				۰/۰۷				
مقدار F	۲/۷۹ (۴/۹۰)	۰/۳۰*	F	۴/۷۲	(۴/۹۴)	۰/۱>		

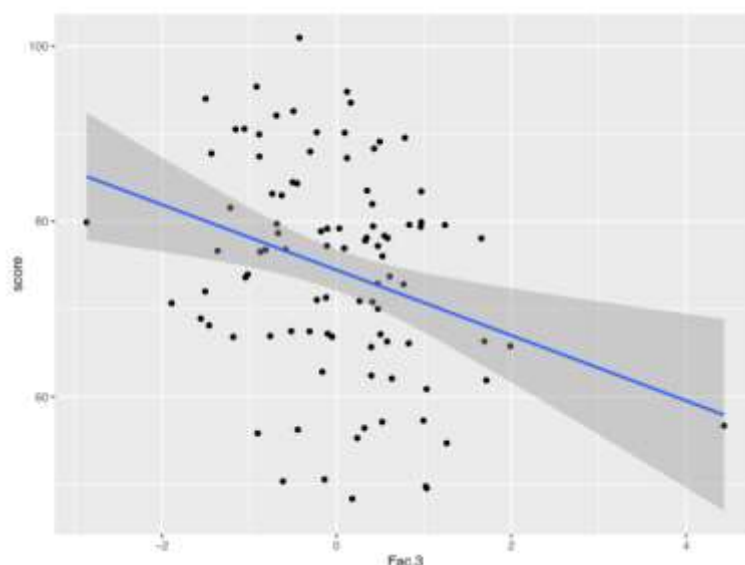
نتایج مدل رگرسیونی نمرات عاملی در مجموعه داده های شفاهی نشان می دهد که این مدل از نظر آماری معنی دار است ($p \text{ value} < 0.03$) و حدود هفت درصد از واریانس متغیر ملاک را توضیح می دهد (مقدار R^2 تصحیح شده برابر با 0.07). عامل سوم تنها پیش بینی کننده ی متغیر درک مطلب است که به لحاظ آماری معنادار است. به بیان دیگر اگر مدل پیش بین تنها دربرگیرنده ی عامل سوم به عنوان متغیر پیش بین باشد، مقدار کل واریانس توضیح داده شده ۷ درصد است. رابطه عامل سوم با درک مطلب منفی است و درک مطلب را به صورت منفی پیش بینی می کند. مقدار ضریب رگرسیونی برای این عامل ۴- است. از آنجایی که این عامل در کاربرد اسم و ویژگی و شباهت کلمه (واریانس آنها) دارای بار عاملی مثبت است می توان نتیجه گرفت که این متغیرها با درک مطلب رابطه ی معکوس یا منفی دارند اما چون این عامل در پیچیدگی گرامری، طول کلی، ضمائر شخصی، تنوع ضمائر و واژگان (در یک پنجره کوتاه) بار عاملی منفی دارد، رابطه ی آنها با درک مطلب مثبت و مستقیم است. این نتیجه ی بدست آمده مطابق با پیشینه پژوهشی در زمینه ی عوامل مرتبط با توسعه زبان و ادبیات می باشد. شاید اختصاصی بودن واژه پیش بینی کننده ی مثبت مهارت های خواندن باشد. با این حال مقدار اختصاصی بودن به دست آمده از WordNet لزوماً معیاری برای پیچیدگی واژگانی نیست.

از آنجایی که عامل سوم به عنوان تنها پیش بینی کننده معنادار درک مطلب است و فیچرهای بدست آمده از WordNet بیشترین بار عاملی مثبت را با این عامل دارند، نیازمند بررسی های دقیق تری هستیم. از این رو نمودار پراکندگی رابطه بین عامل سوم و درک مطلب برای داده های شفاهی / نسخه معمول ترسیم شده است. این نمودار براساس ۹۲ مشاهده ای است که در مجموعه داده های کلامی / نسخه معمول وجود داشت. همانطور که این نمودار نشان می دهد، داده ای با نفوذ یا اهرم^۱ بالا در محدوده عدد ۴ در محور X که محور مربوط به عامل سوم است مشاهده می شود. این داده ها شیب خط رگرسیون را افزایش می دهد و به عنوان یک داده ی پرت در تحلیل رگرسیونی قلمداد می شود. یک نقطه داده پرت دیگر حول

¹. leverage

مقدار ۳- در محور X وجود دارد. این دو مشاهده بیشترین و کمترین مقدار را در عامل سوم دارند.

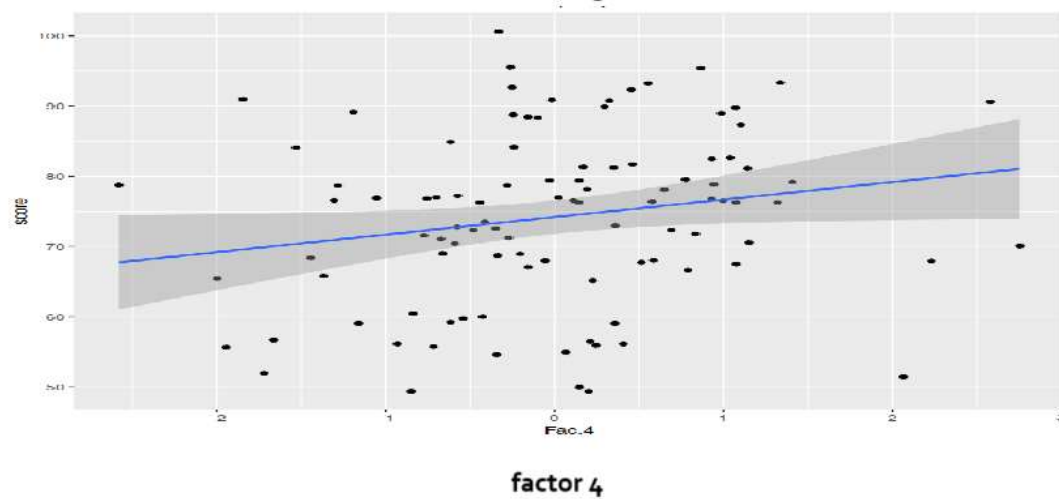
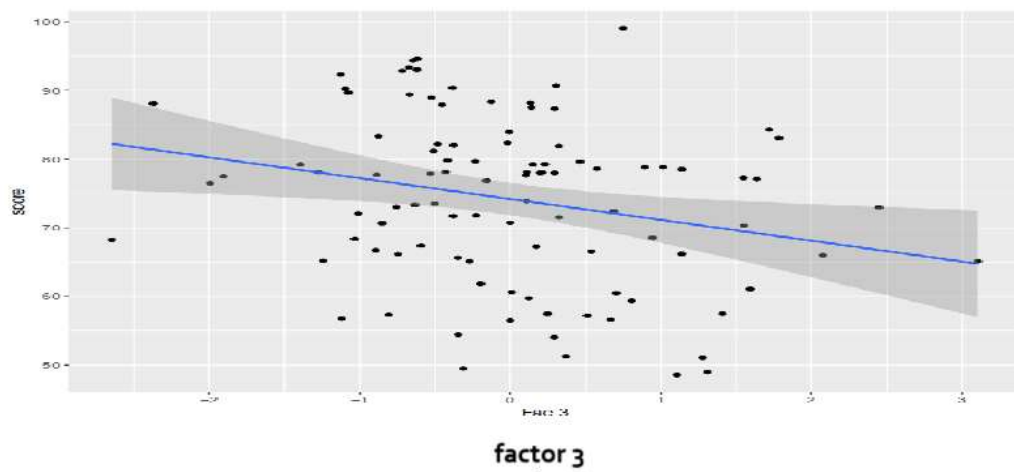
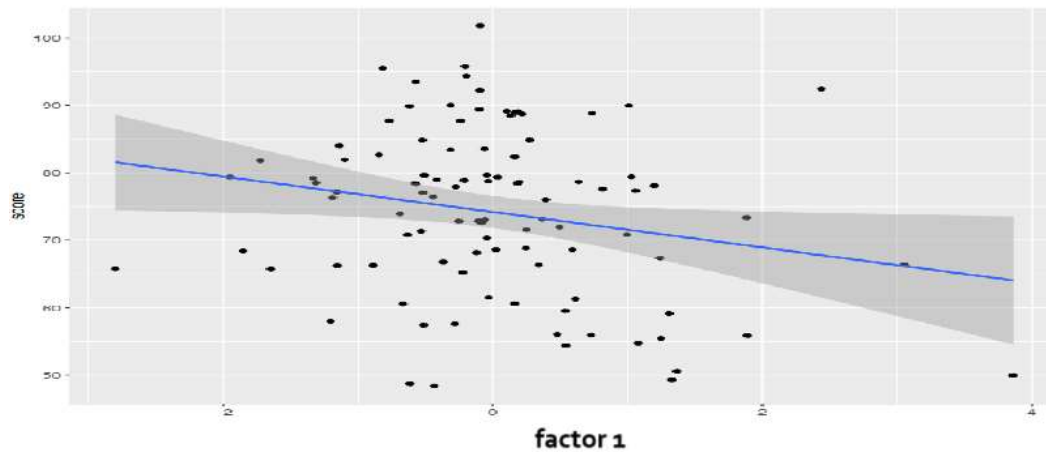
نمودار ۴: رابطه بین عامل سوم (تنها عامل معنادار در مدل پیش بینی) و نمرات درک مطلب خواندن در مجموعه داده‌های شفاهی / معمول



براساس نتایج جدول فوق، مدل رگرسیونی تشکیل شده برای داده‌های کتبی دارای قدرت پیش‌بینی آماری است ($F \text{ value}=4.72$ و $p \text{ value}<0.01$). عوامل ۱، ۳ و ۴ از نظر آماری در سطح اطمینان ۹۵ درصد به عنوان پیش‌بینی کننده درک مطلب در نظر گرفته می‌شوند. عامل دوم نمی‌تواند پیش‌بینی کننده قابل توجهی باشد. مدل رگرسیون ۵۴ درصد از واریانس درک مطلب را توضیح می‌دهد (مقدار مربع همبستگی برابر با ۰.۱۳). سه پیش‌بینی کننده معنی دار ۱۲،۳ درصد از واریانس را توضیح می‌دهند. در نمودارهای پراکندگی زیر، مدل رگرسیونی عوامل پیش‌بین معنادار با متغیر ملاک درک مطلب ترسیم شده است.

نمودار ۵: رابطه ی بین عوامل معنادار (اول، سوم و چهارم) و نمرات درک مطلب خواندن در

مجموعه داده‌های شفاهی /اصلاح شده



بحث و نتیجه‌گیری

همانطور که گفته شد دو روش گردآوری داده وجود داشت: گردآوری از طریق پاسخ‌های نوشتاری زبان آموزان (کتبی) و گردآوری از طریق پاسخ‌های گفتاری زبان آموزان (شفاهی). دو دسته تکلیف (در مجموع چهار تکلیف) به منظور جمع‌آوری داده‌های شفاهی از پاسخ‌های گفتاری شرکت‌کنندگان وجود داشت: شرح تصویر (دو تکلیف) و داستان‌سرایی (دو تکلیف). علاوه‌براین برای جمع‌آوری داده‌های نوشتاری دانش آموزان از متن پاسخ شرکت‌کنندگان به مجموعه‌ای از سؤالات نوشتاری در مورد رسانه‌های اجتماعی و پاسخ‌های کتبی آنها به سؤالات باز پاسخ درباره متون روایی و غیرداستانی گردآوری شد. برای کاهش اثرات کار و یکپارچه‌سازی داده‌ها، ویژگی‌های یا فیچرهای زبانی مربوط به پاسخ زبان آموزان از طریق پردازش زبان طبیعی NLP هم برای داده‌های شفاهی و هم داده‌های کتبی استخراج شد و میانگین آنها در هر مجموعه به‌دست آمد. بنابراین، ما دو مجموعه داده از نمرات شرکت‌کنندگان داشتیم: نمرات شفاهی و نمرات کتبی.

Hektio (2003) وظایف شنیداری و نوشتاری یکسانی را انجام داد و دریافت که مهارت‌ها در حوزه شفاهی و نوشتاری با هم ارتباط زیادی دارند. مطالعه حاضر ممکن است اولین موردی باشد که از فیچرهای بسیار جزئی استخراج شده از طریق NLP استفاده می‌کند که در داده‌های استخراج شده از گفتار و متن یکسان اند و به دلیل همین قابلیت مقایسه‌ی بین داده‌های کتبی و شفاهی وجود دارد.

رویکرد ML با هدف درک ساختار زیربنایی مجموعه داده‌های ویژگی زبانی با استفاده از تحلیل عاملی در Scikit-learn اجرا شد. نتایج تحلیل عاملی نشان داد که حدود ۵۰ درصد از واریانس در مجموعه داده‌های کتبی و شفاهی توسط این عوامل پنهان تبیین می‌شود. به طور کلی، الگوی بارهای عاملی ویژگی‌های زبانی نشان داد که ساختار پیچیدگی دستوری غالب‌تری فیچر در کل داده‌ها و به‌ویژه برای داده‌های شفاهی استخراج شده است که نشان می‌دهد زبان آموزان جوان‌تر یا دانش آموزان جدیدتر ممکن است از دستور زبان پیچیده به اندازه کافی استفاده نکنند. با این حال، جمله و طول جمله هم در داده‌های نوشتاری و هم در داده‌های گفتاری هر دو روی یک عامل بار می‌گیرند. غنای واژگان (تنوع واژگانی) و پیچیدگی دستوری در هر دو مجموعه‌ی داده همبستگی مثبت دارند با این وجود در داده‌های

به دست آمده از زبان آموزان با مهارت کم این توصیف وجود ندارد که ممکن است به دلیل تفاوت های رشدی و فقدان بارگذاری عامل پیچیدگی گرامری در داده های مربوط به یادگیرندگان با مهارت پایین باشد. این یافته ها نشان می دهند که رابطه بین غنای واژگان و پیچیدگی گرامری ممکن است در طول دوره رشد زبان تغییر کند. دامنه واژگان با هیچ بعد کشف شده ای در ارتباط نیست و بار عاملی معناداری روی آن ها ندارد. نتایج این تحقیق نشان داده است که تنها در داده های شفاهی زبان آموزان با مهارت پایین است که در آن غنای واژگان و دامنه هم جهت بارگذاری می شود. ویژگی های اختصاصی بودن واژگان، شباهت و ابهام در عوامل مختلف بار داشتند که تفسیر آن را تا حدودی دشوار می سازد. فیچرهای مربوط به احساسات و عواطف واژگانی تنها در داده های گفتاری بار عاملی معنادار داشتند.

در تحلیل رگرسیون، عاملی که متشکل بود از پیچیدگی گرامری و غنای واژگان به طور مثبت درک مطلب را در مجموعه داده شفاهی پیش بینی کرد. در داده های نوشتاری، سه عامل درک مطلب را پیش بینی می کنند: پیچیدگی دستوری و غنای واژگان، طول جمله و طول کل متن، و دامنه واژگان و ویژگی کلمه. نتایج به دست آمده پژوهش حاضر مطابق با ادبیاتی است که تنوع واژگان، استفاده از عبارات پیچیده دستوری و دامنه واژگان را به طور مثبت با درک مطلب مرتبط می داند. در مجموعه داده های نوشتاری زبان آموزان با مهارت پایین، عامل شباهت کلمه (میانگین و تنوع) پیش بینی کننده منفی و معنادار درک مطلب است که نشان می دهد نوشتن تکراری رابطه منفی با نتایج درک مطلب خواندن این زبان آموزان دارد. با این حال، این رابطه منفی در همه موارد صادق نیست، چون شباهت کلمه از انسجام نوشتاری در سایر داده ها پشتیبانی می کند. در ادامه یافته های مربوط به گرامر و واژگان توضیح داده شده است.

گرامر: به طور کلی، پیچیدگی گرامری در رویکرد تحلیل عاملی بدون نظارت نسبت به رویکرد ML نظارت شده حضور قوی تری دارد (Yapp و همکاران، 2023). در رویکرد بدون نظارت که مبنای تحقیق حاضر است، دو عامل از چهار عامل در هر دو مجموعه داده به پیچیدگی دستوری ربط داشتند. از این میان، عامل پیچیدگی گرامری درک مطلب را در داده های شفاهی و دو عامل در مجموعه داده های نوشتاری درک مطلب را پیش بینی

می کردند. نکته ای که از این مقایسه استخراج می شود این است که پیچیدگی دستوری یک ساختار سازماندهنده ی مهم زبان سازنده است، همانطور که نتایج به خوبی نشان داده است و عواملی که حول پیچیدگی دستوری سازماندهی شده اند درک خواندن را پیش بینی می کنند. با این حال، به عنوان یک ویژگی تنها، پیچیدگی دستوری ممکن است زمانی که بسیاری از متغیرهای نحوی واژگانی ظریف دیگر وجود دارند، به طور مستقیم درک مطلب را پیش بینی نکند. اگرچه این یافته متناقض به نظر می رسد اما با پیشینه و ادبیات تحقیق هم راستا می باشند. ادبیات تحقیق زبان شناسی در مورد نقش خاص توانایی نحوی هنوز دارای بحث ها و نظریات مختلف و گاهی متناقض یکدیگر است. درحالی که برخی از محققان توانایی نحوی را پیش بینی کننده قوی درک مطلب می دانند (برای مثال Demont، Deacon Kiefer، 2018، 1996 Gombert)، دیگران نقش منحصر به فرد پیچیدگی نحوی را در درک مطلب زیر سوال می برند. به عنوان مثال، Umesh و همکاران (2017)، با ثابت نگه داشتن حافظه کاری کلامی و حساسیت واج شناختی، دریافتند که پردازش نحوی تنها ۱،۳-۱،۵ درصد از واریانس درک خواندن را به خود اختصاص می دهد. (Jang 2019) نیز نتایج مشابهی را هنگام اضافه کردن دانش نحوی به عنوان یک متغیر پیش بین در مدل رگرسیونی به دست آورد. با تدوین مدل رگرسیون سلسله مراتبی به این نتیجه رسیدند که اگرچه گرامر پیش بینی کننده قابل توجهی برای درک مطلب در مدل های آنها بود، اما لزوماً به این معنی نیست که پردازش گرامری منجر به مشکلات خواندن می شود یا مهارت های دستوری توسعه یافتگی کمتری دارند. این موضوع ممکن است نشانگر چالش های زبانی گسترده تر باشد. در تحقیق (Nation and Snowling 2000) و (Bowie 1994) این موضوع مورد بررسی و مطالعه قرار گرفته است. از آنجایی که سایر متغیرها (به عنوان مثال، درک شنیداری) در مدل مطالعه حاضر حضور نداشتند، ملاحظات باید در اینجا در نظر گرفته شوند. برای درک اینکه چرا عوامل سازماندهی شده حول پیچیدگی گرامری بیشتر از ویژگی های پیچیدگی گرامری فردی پیش بینی کننده قوی درک مطلب هستند، به تحقیقات بیشتری نیاز است. از آنجا که اجزای دستوری، به جای پیچیدگی دستوری، پیش بینی کننده های رایج در مدل های نظارت شده بودند، یک مدل دو عاملی یا سلسله مراتبی ممکن است بهتر با این داده ها تناسب داشته باشد (به عنوان مثال، پیچیدگی دستوری به عنوان سطح بالاتر از اجزای دستوری).

با این حال، ممکن است همه مولفه های گرامری به پیچیدگی گرامری مرتبط نباشند. به عنوان مثال، به نظر می رسد که مدل های نظارت شده استفاده از اسم را با نمرات درک مطلب پایین تر و استفاده از فعل را با نمرات درک مطلب بالاتر مرتبط کنند. به نظر می رسد استفاده از فعل مخصوصاً برای تکالیف شفاهی که در آن روایات ارائه شده توسط زبان آموزان دربرگیرنده ای افعال می باشند و در آن زبان آموز نمی تواند از سبک توصیف فهرست وار استفاده کند صادق است. با وجود این که نام گذاری در زبان آکادمیک کلیدی تلقی می شود (Feng و همکاران، 2006)، این یافته ها با یافته های افراد دیگر مانند Anne, Giersch (2023) هم راستا است که استفاده از اسم ها را به عنوان عنصری کلیدی برای ارزیابی کنندگان انسانی از نوشتار در نظر نمی گیرد. در واقع، در اینجا، استفاده از اسم با نمرات خواندن پایین تر همراه است. همان طور که Graham & Neubig 2012 اشاره کرده است، یادگیری افعال دشوارتر از اسامی است و وابسته به گروه سنی شرکت کنندگان است که کاملاً منطبق بر نتایج تحقیق حاضر است.

کلمه یا واژه: اندازه گیری گستره واژگان دقیقاً همان طور که انتظار می رفت انجام شد: سن اکتساب و طول کلمه به طور مثبت نتایج خواندن را پیش بینی می کرد، در حالی که فراوانی، تصویرپذیری و آشنایی پیش بینی کننده های منفی درک مطلب بودند. با توجه به یافته های Pierre & Nuges 2014 مبنی بر اینکه آشنایی، فراوانی و تصویرپذیری با درک مطلب همبستگی مثبت دارند که همه آن طریق مجموعه WordNet تعریف عملیاتی شده اند، تعجب آور نیست که منجر به تفسیر واضح و منظم نشده است. انتظار می رود میانگین ابهام نتیجه خواندن را به طور مثبت پیش بینی کند (Kassas و همکاران، 2018)، با این وجود فقط در داده های کتبی/نسخه اصلاح شده پیش بینی کننده مثبت درک مطلب بود. تحقیقات در مورد ابهام و رشد زبان نشان می دهد که استفاده کودکان از کلمات مبهم با استفاده ای افراد بعد از ۵ سالگی بسیار تفاوت دارد. این یافته تا حدی منطقی است زیرا شرکت کنندگانی که به نسخه ای اصلاح شده بلا پاسخ دادند کمتر به زبان انگلیسی مهارت داشتند و در عین حال به طور میانگین جوانتر بودند و اختلاف میانگین سنی آن ها معنادار بود ($p < 0.001$). برخلاف انتظار، توابع ابهام پیش بینی کننده های مثبت قوی درک مطلب زبان آموزان بودند اما می توان به صورت شهودی درک کرد که اگر فقط از کلمات مبهم استفاده شود، می توان از واژگان سطح بالاتر کمتر استفاده کرد، زیرا کلمات مبهم معمولاً

کلماتی با بسامد بالاتر هستند. بنابراین، تغییر در تابع ابهام یک پیش‌بینی‌کننده مثبت برای مجموعه داده‌های شفاهی یادگیرندگان با مهارت پایین و مجموعه داده‌های نوشتاری برای یادگیرندگان با مهارت بالا بود که به تحقیقات بیشتری نیاز دارد.

شباهت کلمه: که شباهت کلمات را در یک قطعه معین اندازه‌گیری می‌کند، یکی از پیش‌بینی‌کننده‌های داده‌های نوشتاری است که به صورت مثبت درک مطلب زبان آموزان با مهارت بالا را پیش‌بینی می‌کند، این درحالی است که همین متغیر پیش‌بینی منفی در پاسخ‌های نوشتاری آزمون‌شوندگان با مهارت پایین است. مرحله توسعه یا اکتساب زبان شرکت‌کنندگان می‌تواند بر این نتیجه تأثیر بگذارد. در تجزیه و تحلیل داده‌های نوشتاری بدون نظارت کودکان کم‌مهارت عاملی وجود دارد که کاملاً از میانگین و انحراف معیار شباهت واژه‌ها شکل گرفته است و یک پیش‌بینی‌کننده منفی معنادار است که با نتایج مدل اصلاح شده در این مجموعه داده هماهنگ است. آنها دریافتند که همپوشانی کلمه-محتوا در نوشتار شرکت‌کنندگان جوان نسبت به شرکت‌کنندگان بزرگ‌تر کاهش یافته است (تحقیق آنها روی نوجوانان و بزرگسالان جوان بود). با این وجود حرکت از همپوشانی زیاد به همپوشانی کم محتوای متن ممکن است خطی نباشد و نیاز به کاوش بیشتر دارد. همان‌طور که انتظار می‌رود، ویژگی افعال حتی اگر این ویژگی به طور مستقیم با گستره واژگان مرتبط نباشد، به طور مثبت نتایج خواندن را برای داده‌های نوشتاری کودکان کم‌مهارت پیش‌بینی می‌کند. در مجموعه داده کلامی، تنوع صفات یک پیش‌بینی‌کننده قابل توجه با جهت ارتباط متفاوت است. این رابطه در داده‌های شفاهی مثبت و در داده‌های شفاهی برای مبتدیان منفی است. موضوع تنوع در میان این فیچرها نیاز به تحقیقات بیشتری دارد. تحقیقات بیشتر می‌تواند برای درک اینکه آیا نتایج به‌دست آمده به دلیل عدم توسعه واژگان مناسب در خط لوله COVFEFE است یا اینکه این ویژگی‌ها در واقع به همان اندازه مهم هستند که نتایج فعلی نشان می‌دهند، متمرکز باشد؟

تحقیق نشان داد که عواطف و متغیرهای مربوط به وجه عاطفی واژگان به ویژه برای افراد کم‌مهارت ممکن است به مراحل رشد و خودتنظیمی آنها برگردد. گنجاندن سایر منابع داده‌ای مانند گزارش‌های خودتنظیمی دانشجویان می‌تواند به این پرسش پاسخ دهد. به نظر می‌رسد اهمیت عواطف و متغیرهای عاطفی در پیش‌بینی درک مطلب را نمی‌توان به همه

تکالیف تعمیم داد. به عنوان مثال، تکالیف مربوط به استنتاج کلامی که شامل برخی رویدادهای شبه دراماتیک بود برای بازگویی دقیق نیاز به استفاده از کلماتی با احساسات منفی داشت. از این رو، ارزش پیش‌بینی‌کننده مثبت احساسات منفی در آن تکالیف معنادار است. همانطور که Heris (2018) اشاره می‌کند، ابزارهای تجزیه و تحلیل احساسات که روی زبان طبیعی آموزش داده‌شده‌اند، فاقد قابلیت تعمیم‌اند، زیرا به شدت وابسته به مجموعه‌ای هستند که در آن آموزش دیده‌اند. مقایسه نتایج احساسات و عواطف در متون مختلف می‌تواند یک سیر مطالعه سودمند قلمداد شود.

سایر ویژگی‌های عواطف و احساسات واژگان به سختی قابل تفسیراند: آیا ارزش پیش‌بینی‌کننده منفی برانگیختگی در مجموعه داده‌های شفاهی برای زبان‌آموزان کم مهارت به این دلیل است که شرکت‌کنندگان احساسات کنترل‌ناپذیری داشتند؟ واریئر و همکاران در توسعه اولیه مجموعه نوشتاری خود (2013) دریافتند که افراد جوان نمرات بالاتری را در ویژگی‌های برانگیختگی و ظرفیت واژگان دارند اما در مجموعه داده‌های کتبی برای همین جوانان یا شرکت‌کنندگان کم مهارت، برانگیختگی متوسط پیش‌بینی‌کننده مثبت درک مطلب است - آیا این نتایج به تفاوت در متن و گفتار مربوط است؟ برانگیختگی در استانداردهای ارزیابی نوشتار نشان‌دهنده‌ی مشارکت است. برای پاسخ به این پرسش‌ها به تحقیقات بیشتری نیاز است.

نکته قابل توجه عدم وجود عواطف و ویژگی‌های عاطفی در بین نتایج بدست آمده از تحلیل‌های عاملی است این ویژگی‌ها فقط در داده‌های کلامی برای افراد کم مهارت وجود داشت. تعداد کمی از ویژگی‌های مربوط به پیچیدگی دستوری پیش‌بینی‌کننده‌های برتر در مدل‌های نظارت‌شده بودند (Yapp و همکاران، 2023)، اما در مدل‌های یادگیری ماشین بدون نظارت ساختارهای سازمان‌دهنده مهم به شمار می‌روند. این موضوع نشان می‌دهد که متغیرهای هیجانی می‌توانند درک مطلب را پیش‌بینی کنند (اگرچه ممکن است مختص یک تکلیف خاص باشد و نمی‌توان آن را به تکالیف دیگر تعمیم داد)، اما ویژگی‌های هیجانی ممکن است فاقد ساختارهای کافی در زایش زبانی باشند. با وجود این زبان مشترک باید در کنار ویژگی‌های هیجانی باشد. در مدل‌سازی بدون نظارت داده‌های کلامی برای زبان‌آموزان کم مهارت از نظر آماری مهم‌ترین عامل در پیش‌بینی درک خواندن بود. بارگذاری‌های

قوی روی متغیرهای هیجانی و هیجانات واژگانی با نتایج بدست آمده از مدل‌های نظارت شده‌ی Yapp و همکاران (2023) در مدل‌های نظارت شده منطبق بود. این الگو در سایر مجموعه داده‌ها مشاهده نمی‌شود. بررسی و مطالعه‌ی نقش عواطف در توانایی خواندن در سنین و سطوح مختلف توانایی یک خط مطالعاتی دیگری است که می‌تواند مورد نظر سایر پژوهشگران باشد.

غنا‌ی واژگان که به عنوان تنوع واژگانی نیز شناخته می‌شود، در مدل‌های بدون نظارت سناریویی شبیه به پیچیدگی دستوری داشت (Yapp و همکاران، 2023). تنها دو ویژگی غنا‌ی واژگان در میان پیش‌بینی‌کننده‌های برتر مدل‌های نظارت شده (هم در داده‌های گفتاری و هم در داده‌های نوشتاری) بودند (Yapp و همکاران، 2023). در داده‌های گفتاری، نمره Honore (که تابعی از تعداد کلماتی است که فقط یک بار در یک نمونه زبانی استفاده می‌شود) پیش‌بینی‌کننده مثبت درک مطلب بود درحالی‌که در متون نوشتاری، نسبت میانگین متحرک نوع-نشان نقش پیش‌بینی‌کننده مثبتی داشت. در مدل‌های بدون نظارت، نسبت‌های میانگین نوع-توکن متحرک برای داده‌های مربوط به نسخه معمول ۲ عامل از ۴ عامل اول، در مجموعه داده‌های شفاهی / نسخه تصحیح شده ۲ عامل از ۳ عامل نخست و برای داده‌های کتبی / نسخه اصلاح شده ۱ از ۵ عامل اول وجود داشت. در هر دو مجموعه داده (شفاهی و کتبی)، عوامل با نسبت‌های میانگین متحرک نوع-نشان به طور معنادار و مثبت نتایج خواندن را پیش‌بینی کردند. در زبان گفتاری، میانگین متحرک نسبت‌های نوع-نشان در همان جهتی بارگذاری می‌شود که طول جمله و طول کل پاسخ بارگذاری می‌شود. در مجموعه داده‌های نوشتاری / نسخه معمول، نسبت‌های میانگینی نوع-توکن متحرک با پیچیدگی گرامری در یک عامل و با طول جمله و طول کل متن در عامل دیگر بارگذاری شدند. نسبت میانگین متحرک نوع-نشان یک ساختار مهم در هر چهار مجموعه داده به نظر می‌رسد، زیرا این ویژگی‌ها در عوامل هر چهار مجموعه داده وجود داشت با این وجود فقط درک را در مجموعه داده‌های مربوط به نسخه‌ی معمولی پیش‌بینی می‌کنند. این یافته ممکن است مربوط به تفاوت در رشد یا بلوغ زبان در مجموعه داده‌های زبان آموزان با مهارت بالا و کم مهارت باشد، اما این فرضیه باید توسط تحقیقات آینده تأیید شود.

عواملی که بر ویژگی های نسبت نوع-نشانه میانگین متحرک بار شده اند، بر ویژگی هایی دیگری بار می شوند که آن ها نتایج خواندن را به طور مثبت پیش بینی می کنند. نسبت نوع-توکن استاندارد که غنای واژگان را در کل نمونه زبان می سنجد، دارای ویژگی های واژگانی و نحوی است که نشان دهنده توانایی های زبانشناختی و ادبیاتی پایین تر است، درحالی که میانگین های متحرک به طور مداوم با ویژگی هایی بارگیری می شوند که نشان دهنده توانایی های زبان شناختی بالاتر است. همچنین قابل ذکر است که در عواملی که هم نسبت نوع-توکن استاندارد و هم میانگین متحرک بارگذاری می شوند دارای بارهای عاملی هستند که جهت آن مخالف یکدیگر است. این نتیجه که مستخرج از تحقیق حاضر است هم راستا با انتقاداتی است که از سوی پژوهشگران به نسبت استاندارد نوع-توکن می شود.

محدودیت های مطالعه ما شامل استفاده از یک مجموعه داده واحد و فقدان گروه کنترل است. تحقیقات آینده باید این یافته ها را با استفاده از مجموعه داده های مختلف و طرح های تجربی تکرار کند. علاوه بر این، تحقیقات بیشتری برای کشف مکانیسم های اساسی که رابطه بین ویژگی های زبانی و درک مطلب را توضیح می دهند، مورد نیاز است. در نتیجه، این مطالعه بینش های ارزشمندی را در مورد عواملی که به درک مطلب در مجموعه داده های شفاهی و نوشتاری کمک می کنند، ارائه می کند. یافته ها نشان می دهد که ساده سازی ساختارهای دستوری، قرار دادن دانش آموزان در معرض طیف وسیعی از واژگان، و اجتناب از استفاده بیش از حد از ضمایر شخصی و جملات طولانی ممکن است برای بهبود درک مطلب مفید باشد. این یافته ها پیامدهای مهمی برای آموزش و یادگیری دارند و نیاز به تحقیقات بیشتر در این زمینه را برجسته می کنند.

با این وجود، این مطالعه با روش های گذشته تفاوت های عمده ای داشته است که در پایان به ذکر آنها پرداخته می شود. روش های نوین یادگیری ماشین بدون نظارت و پردازش زبان طبیعی (NLP) به منظور شناسایی عوامل پنهان در ویژگی های واژگانی و نحوی پاسخ های شفاهی و کتبی زبان آموزان مورد استفاده قرار گرفته است و برخلاف رویکردهای قبلی که بیشتر متکی بر متغیرهای دستی و روش های نظارت شده بودند، این تحقیق با استخراج ۲۶۰ ویژگی زبانی از داده های گفتاری و نوشتاری و استفاده از تحلیل عاملی به صورت خودکار چهار عامل مهم را در هر دو مجموعه داده شناسایی کرد. این عوامل شامل "پیچیدگی

دستوری"، "تنوع واژگانی"، "تشابه و ویژگی کلمه" و "پیچیدگی بند" هستند که بیش از ۵۰ درصد از واریانس داده‌ها را توضیح می‌دهند. سپس این عوامل به عنوان متغیرهای پیش‌بینی‌کننده در مدل رگرسیونی قرار گرفتند و نتایج نشان داد که در داده‌های شفاهی تنها عامل سوم (تشابه و ویژگی کلمه) و در داده‌های کتبی عوامل اول (پیچیدگی دستوری)، سوم (پیچیدگی بند) و چهارم (اختصاصی بودن کلمه) به طور معناداری در پیش‌بینی درک مطلب نقش دارند. این روش، با تلفیق هوش مصنوعی و تحلیل‌های زبانی عمیق، امکان شناسایی الگوهای پنهان در داده‌های زبانی را فراهم کرده و زمینه را برای توسعه روش‌های دقیق‌تر آموزش و ارزیابی زبان فراهم می‌کند.

تعارض منافع

نویسندگان هیچ گونه تعارض منفعی ندارند کفایت می‌کند.

سپاسگزاری

مقاله حاضر برگرفته از رساله دکتری دانشگاه آزاد واحد علوم تحقیقات است. بدینوسیله از حمایت دانشگاه مربوطه و موسسات زبانی که زمینه اجرای این تحقیق را فراهم کردند، کمال تشکر را داریم.

References

- Abdul, Razaque., Abrar, M., Alajlan. (2020). Supervised Machine Learning Model-Based Approach for Performance Prediction of Students. *Journal of Computer Science*. doi: 10.3844/JCSP.2020.1150.1162
- Amir, Karami., Aryya, Gangopadhyay., Bin, Zhou., Hadi, Kharrazi. (2015). A Fuzzy Approach Model for Uncovering Hidden Latent Semantic Structure in Medical Text Collections.
- Andreas, V. (2011). Evaluating unsupervised learning for natural language processing tasks.
- Anne, G. (2023). Natural Language Processing. doi: 10.1007/978-1-4842-8931-0_4
- Anup, A, Deshmukh., Q, Zhang., M, Li., J, Lin., Lili, M. (2021). Unsupervised Chunking as Syntactic Structure Induction with a Knowledge-Transfer Approach.
- Benoit, K., & Nulty, P. (2017). quanteda: Quantitative analysis of textual data. R package version 0.99, 9.
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238.
- Henry, Lin. (2018). Syntactic and Lexical Approaches to Reading Comprehension. doi: 10.18653/V1/W18-3702
- Chi-Hong, L., Yuen-Yan, Ch. (2007). A natural language processing approach to automatic plagiarism detection. doi: 10.1145/1324302.1324348
- Chris, v, der, Lee., A, van, d, Bosch. (2017). Exploring Lexical and Syntactic Features for Language Variety Identification. doi: 10.18653/V1/W17-1224
- Chuyuan, Lin. (2019). Unsupervised learning on functional data with an application to the analysis of U.S. temperature prediction accuracy.
- Dan, K. (2008). Unsupervised Learning for Natural Language Processing.

- Engr., S, Bhutto., I, Farah, S., Qasim, A, Arain., M, A. (2020). Predicting Students' Academic Performance Through Supervised Machine Learning.
doi: 10.1109/ICISCT49550.2020.9080033
- Feng, Zhi-wei. (2006). Several Features of Current Development of Natural Language Processing. Journal of College of Chinese Language and Culture of Jinan University.
- Jang, E.E. (2019). Potential of machine learning approaches for advancing diagnostic language assessment. Keynote presentation at the 9th Canadian Association of Language Assessment (CALA) Symposium. Toronto, ON, Canada.
- Gomez-Merino, N., Fajardo, I., & Ferrer, A. (2021). Did the three little pigs frighten the wolf? How deaf readers use lexical and syntactic cues to comprehend sentences. *Research in Developmental Disabilities*, 112, 103908.
- Graham, N. (2012). Unsupervised Learning of Lexical Information for Language Processing Systems.
- Janel, Wu. (2020). Machine Learning in Education.
doi: 10.1109/ICMEIM51375.2020.00020
- Martin, E. (2011). Machine Learning for Natural Language Processing.
- Martin, R. (2016). Unsupervised Methods for Learning and Using Semantics of Natural Language.
- Meichao, Zh., Aitao, Lu., Pingfang, S. (2017). ERP Evidence for the Activation of Syntactic Structure During Comprehension of Lexical Idiom. *Journal of Psycholinguistic Research*.
doi: 10.1007/S10936-017-9485-Z
- Md., R, Islam., Shuji, S., Yoshihiro, Y., Andrew, W., Vargo., Motoi, Iwata., M, Iwamura., Koichi, K. (2021). Self-supervised Learning for Reading Activity Classification.
doi: 10.1145/3478088
- Minh, N. (2022). ViMRC - VLSP 2021: An empirical study of Vietnamese Machine Reading Comprehension with Unsupervised Context Selector and Adversarial Learning. *VNU Journal of Science: Computer Science and Communication Engineering*.
doi: 10.25073/2588-1086/vnucsce.344

- Pierre, N. (2014). Language Processing with Perl and Prolog: Theories, Implementation, and Application.
- Revelle, W., & Revelle, M. W. (2015). Package 'psych'. The comprehensive R archive network, 337(338).
- Seyyedeh, M, Hamed, R, Pishghadam, M, Ghazanfari. (2015). Detecting the Underlying Constructs of the Self-efficacy Scale for English Language Learners' Textbooks (SES-ELLT) through Exploratory Factor Analysis. Journal of Language Teaching and Research. doi: 10.17507/JLTR.0701.25
- Sinclair, J., Jang, E. E., & Rudzicz, F. (2021). Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing. Journal of Educational Psychology, 113(6), 1088.
- Shuangyan, Liu., Mathieu, d'Aquin. (2017). Unsupervised learning for understanding student achievement in a distance learning setting. doi: 10.1109/EDUCON.2017.7943026
- Stephen, W., Suresh, M. (2000). Unsupervised lexical learning with Categorical grammars using the LLL corpus. Lecture Notes in Computer Science,
- Truckenmiller, A., Shen, M., & Sweet, L. E. (2021). The Role of Vocabulary and Syntax in Informational Written Composition in Middle School. Reading and Writing. 34(4), 911-943. <https://doi.org/10.1007/s11145-020-10099-1>
- Umesh, R., Hodeghatta., Umesh, Nayak. (2017). Unsupervised Machine Learning. doi: 10.1007/978-1-4842-2514-1_7
- Xiangdong, Wu., Xiaoyan, Liu., Yimin, Zhou. (2022). Review of Unsupervised Learning Techniques. doi: 10.1007/978-981-16-6324-6_59
- Yapp, D., de Graaff, R., & van den Bergh, H. (2023). Effects of reading strategy instruction in English as a second language on students' academic reading comprehension. Language Teaching Research, 27(6), 1456-1479.