





From Text to Intelligence: An Intro to ML and NLP

Joshua M. Paiz, Ph.D.

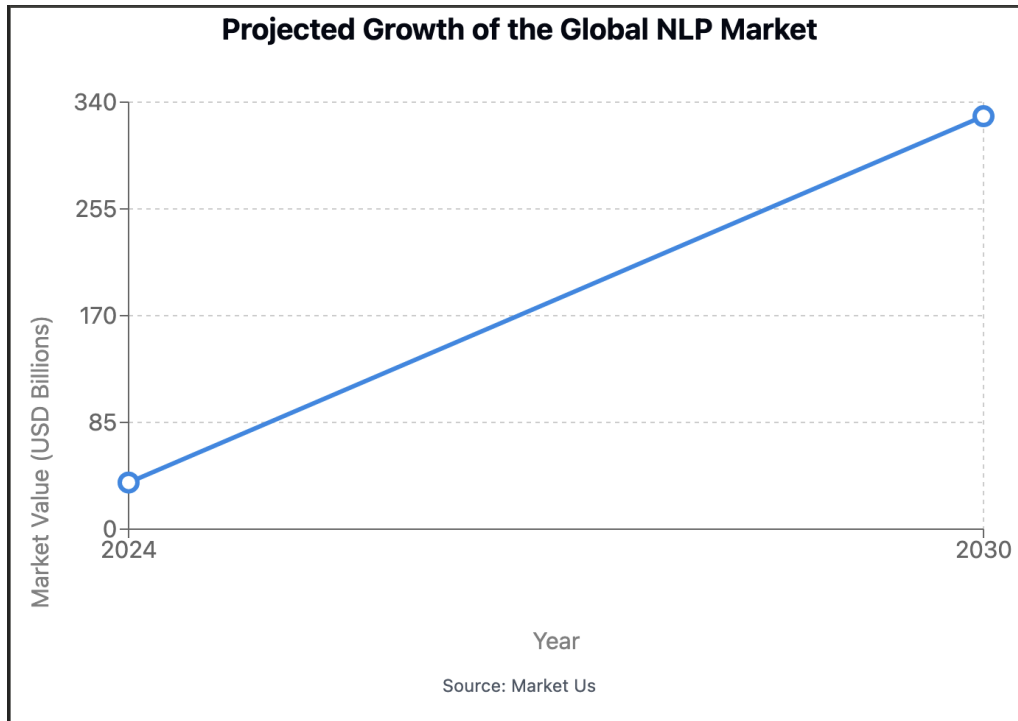
27 March 2025

CSCI 6442: Database Systems II

What to Expect

-  Core machine learning (ML) concepts
-  Key components of natural language processing (NLP)
-  Practical examples
-  Preview of AWS ML Foundations Course

Why ML & NLP Matter



- Considerable real-world applications
 - ML: face recognition, fraud/spam detection, sales forecasting
 - NLP: translation, advanced text editing, voice assistants, traditional chatbots
- Foundational to modern approach to AI (see Halper, 2017; Kamath et al., 2024)
- Opens the door to exciting inter/transdisciplinary work
- ML & NLP in Applied Linguistics
 - Automated essay scoring, tutor-bots, guided learning (see Vajjala, 2012)

So, how do we automate a complex language task that relies on both holistic and analytical measures?

The screenshot displays the Criterion Writing for Success interface. At the top, it shows the title 'Criterion', the text 'Writing for Success', and the submission date 'Submitted November 28, 2006, 02:35:30 PM EST'. Below this is a 'Trait Feedback Analysis Menu' with tabs for Grammar, Usage, Mechanics, Style, and Organization & Development. The 'Style' tab is selected. On the left, a 'Summary of Style Comments' lists various issues: Repetition of Words, Inappropriate Words or Phrases, Sentences Beginning with Coordinating Conjunctions, Too Many Short Sentences, Too Many Long Sentences, and Passive Voice. Below this list, statistics are provided: Number of Words: 399, Number of Sentences: 26, and Average number of words per sentence: 15.3. The main content area is titled 'Repetition of Words' and contains a 'View Question' section. It highlights a specific instance of word repetition in a sample text, providing feedback on why it is problematic and suggesting improvements. The feedback text states: 'The above quotation a concrete example of the mark on the path education needs to for curriculum needs to change also. You have repeated these words several times in your essay. Your essay will be stronger if you vary your word choice and substitute some other words instead. Ask your instructor for advice.' Below this, the sample text is shown with the word 'curriculum' highlighted in blue. The text discusses the need for curriculum change and the role of educators. At the bottom of the interface, there are buttons for 'View Score Analysis', 'Print Combined Feedback Report...', and 'Close Report'. A footer note says: 'Remember, for more information, click on the Writer's Handbook link for each feedback message.'

A Quick Introduction to ML

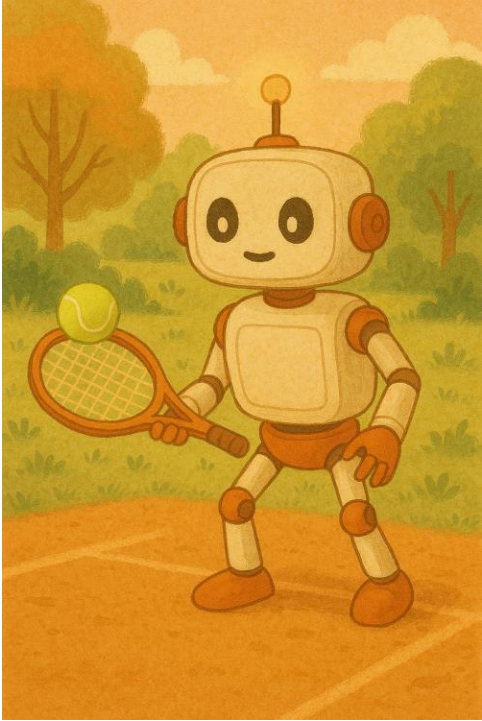
From Explicit Instruction to Reward-based Learning

A Quick, Human Pit Stop

- Take a second and think of the last time you learned to do something new...
- How did you go about learning that new skill...
- What *worked* for you...
- What *didn't*...



Core Concepts in ML

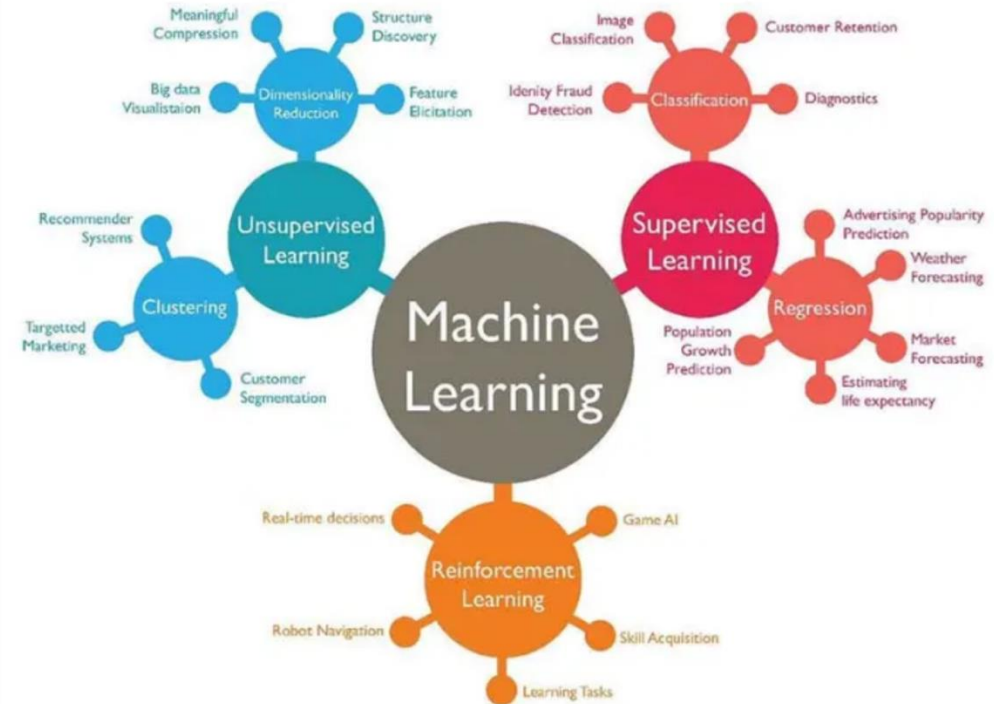


- Machine Learning (ML) – Programming computers to *learn* from data and improve their understanding a specific (type) of problem over time.
 - It's teaching by *example* instead of *explicit instructions*.
 - Enabled the move from symbolic to sub-symbolic AI systems (Mitchell, 2019).

Three Major Types of ML

- Supervised Learning – Learning from labeled training data, where the algorithm is given input-output pairs and learns to predict outputs for new inputs
- Unsupervised Learning – Learning from unlabeled data, identifying patterns and structures without predefined labels
- Reinforcement Learning – Learning through interaction with an environment, where an agent receives rewards or penalties for actions, aiming to maximize cumulative reward

See Murphy (2022)



Supervised Learning in Action

- Labeled Data
- Prediction & Correction Loop
- Examples:
 - Spam Detection – a classification problem using Naïve Bayes or Support Vector Machines
 - House Price Prediction – a regression problem using Linear Regression or Gradient Boosting Regression (e.g., XGBoost)

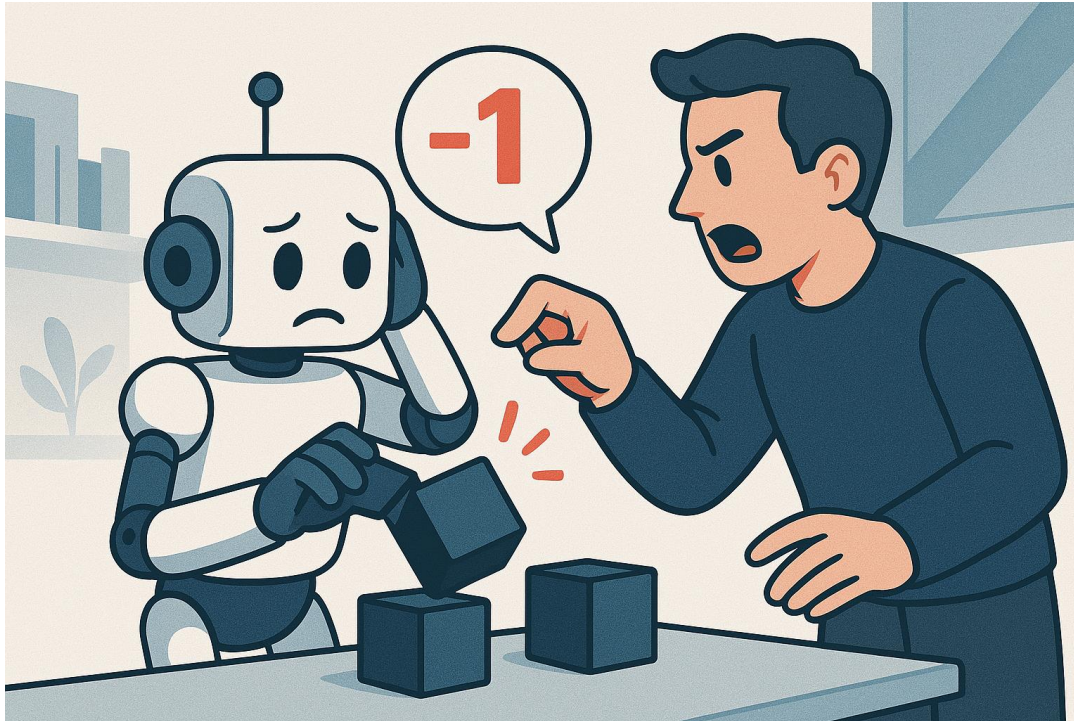
Chopra & Roopal, 2023; Géron, 2019

Unsupervised Learning

- Unlabeled Data Pattern
- Discovery & Grouping
- Examples:
 - Customer Segmentation – a clustering problem using K-Means or Hierarchical Clustering
 - Dimensionality Reduction – using Principal Component Analysis (PCA) or t-SNE for data visualization

Patel, 2019

Reinforcement Learning



- **Reinforcement Learning**
 - Learning through Interaction
 - Action & Reward Mechanism
 - Examples:
 - Game Playing – an RL problem using Q-Learning or Deep Q-Networks (DQN)
 - Robotic Control – using Policy Gradient Methods or Actor-Critic Algorithms
- Morales, 2020

How Does it all Work Under the Hood?



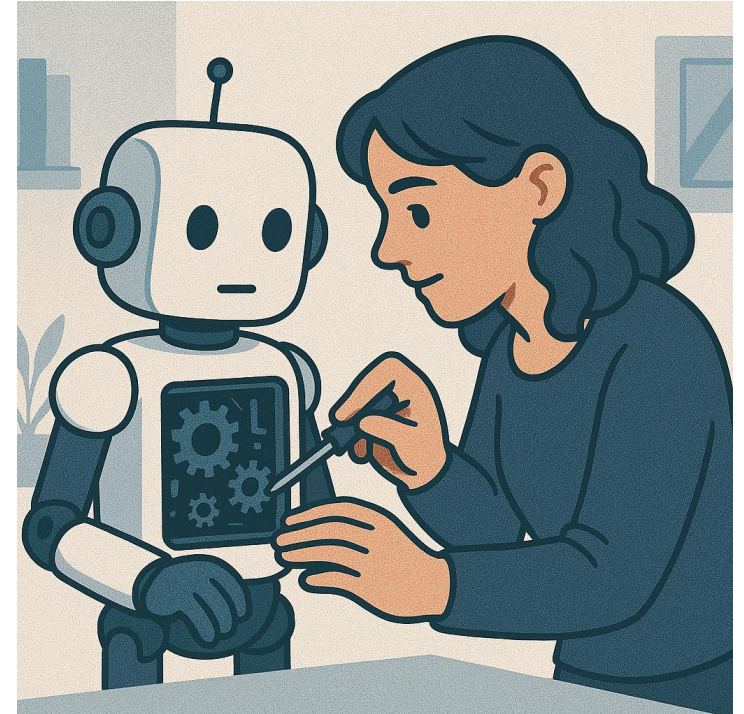
Model = “learned representation”



Algorithm = “training procedure” (e.g., Linear Regression, Decision Trees, Neural Networks)



Data Quality Matters



Natural Language Processing

How do you get a machine to understand the nuance and variability inherent in human language?

What does the following utterance mean?



- Would you like to come up for some coffee?

Is the following utterance grammatical?

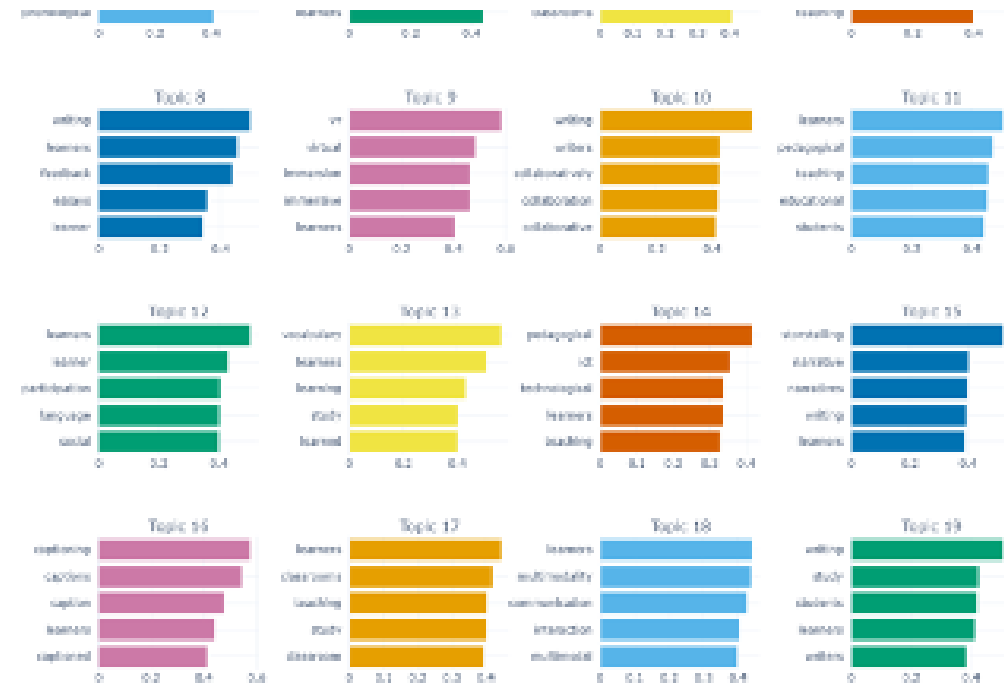
Kindly do the needful and send the report at your earliest convenience.



Key Concepts in NLP

- Definition: Intersection of Linguistics & Computer Science
- Goal: Teach machines to understand/generate language

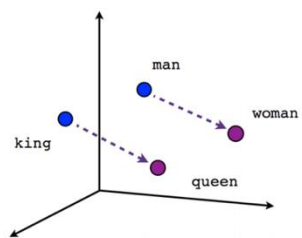
Vasiliev, 2020



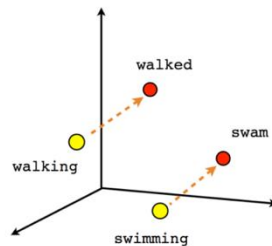
How Text Data is Handled

- **Text Preprocessing:** tokenization, dealing with punctuation, lowercasing
 - Running may just become “run” or “run” + “in progress”
 - Love and Like may get combined into “fond”
- **Ultimately → numbers (vectors or IDs)**
 - **One-hot encoding**
 - **Word2Vec**
 - **TF-IDF**

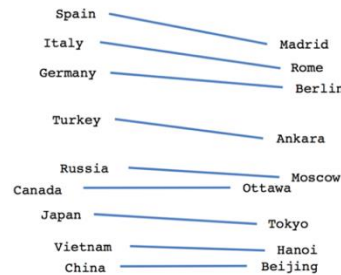
Word Embeddings: Capturing Meaning in Vectors



Male-Female



Verb tense

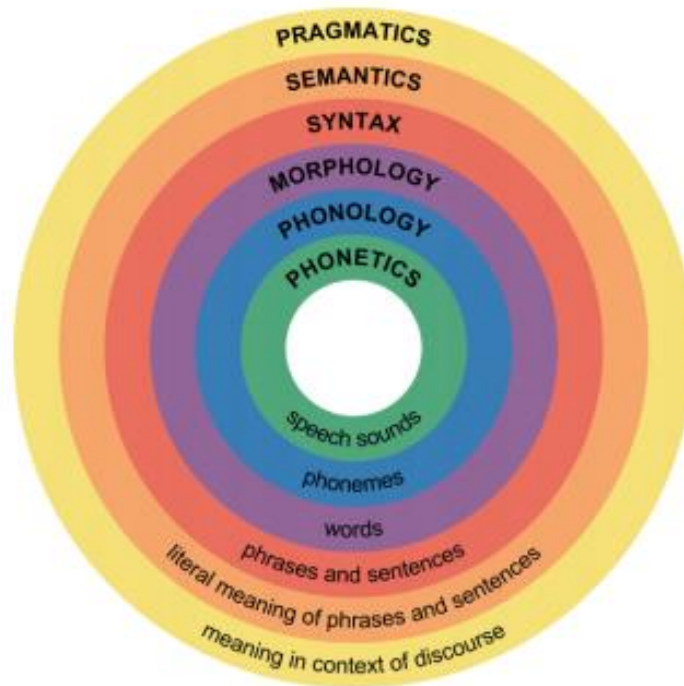


Country-Capital

- Numeric Vectors for Words
 - Words with similar meanings or relationships are positioned close together
- Example: “Paris” – “France” + “Germany” \approx “Berlin”
 - The model infers geography patterns from text data, without explicit coding
- Learned Patterns from Large Corpora
 - No direct rules; the algorithm discovers connections like “capital–country”

See Lynn, 2018

Layers of Language Understanding



- **Lexical/Syntax**: structure, parts of speech
- **Semantics**: meanings of words/phrases
- **Context/Pragmatics**: the broader intent (sarcasm, topic, etc.)
- See Yule, 2014

Core NLP Tasks

- **Language Modeling** (predict next word)
- **Machine Translation** (Google Translate)
- **Syntax & Parsing** (diagramming sentences)
- **Sentiment Analysis** (positive/negative/neutral)
- **Named Entity Recognition** (names, places, dates)

See Gudivada, 2018

NLP for Language Learning

- Examples: grammar correction, essay scoring, chatbots
- NLP = understanding user's input, providing feedback
- E.g., Meurers, 2012; Zilio et al., 2017



ML & NLP in Action

Examples from Applied Linguistics & Language Education

Automated Essay Scoring (AES)

- **Supervised Learning** on teacher-scored essays
- Finds **patterns** in vocabulary, grammar, structure
- **Pros:** Instant feedback, scalable grading
- **Cons:** Misses creativity, can be tricked

E.g., Ke & Ng, 2019; Ramesh & Sanampudi, 2022

Grammar & Writing Feedback

- **NLP-Powered Tools** (e.g., Grammarly)
- **“Translation”**: incorrect → more corrected sentence
 - Correctness is measured against a preferred target
 - Complicated by regional varieties of English and context-based linguistic registers.
- **Helps Learners**: Real-time error feedback
- **Caution**: Over-reliance = fewer self-corrections

See Koltovskaia, 2020

Chatbots for Conversation Practice

- **24/7 Language Partner** (lowers anxiety)
- **Powered by Language Models** (dialogue systems)
- **Example:** Duolingo Roleplay (GPT-4)
- **Benefit:** Grammar hints, real-time practice

See Jinming & Daniel, 2024; Paiz et al., 2025

Personalized Learning & Recommendations

- **Adaptive Pathways:** tracks performance, adjusts lessons
- **Spaced Repetition** for optimal review
- **Similar to Netflix:** recommends next “content”
- **Crucial** for diverse learner needs

See Istani et al., 2024; Yu & Chauhan, 2024



A C A D E M Y

Machine Learning Foundations

References

- Chopra, D., & Roopal, K. (2023). *Introduction to Machine Learning with Python*. Bentham Science Publishers
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly.
- Gudivada, V. N. (2018). Natural language core tasks and applications. In *Handbook of Statistics* (Vol. 38, pp. 403-428). Elsevier.
- Halper, F. (2017). Advanced analytics: Moving toward AI, machine learning, and natural language processing. *TDWI Best Practices Report*.
- Istanti, W., Pratiwi, S., & Saddhono, K. (2024, November). AI-Driven Personalized Learning: Revolutionizing Language Education. In *2024 International Conference on IoT, Communication and Automation Technology (ICICAT)* (pp. 329-334). IEEE.
- Jinming, D. U., & Daniel, B. K. (2024). A systematic review of AI-powered chatbots in EFL speaking practice: Transforming language education. *Computers and Education: Artificial Intelligence*, 100230.
- Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). *Large Language Models: a deep dive: Bridging Theory and Practice*. Springer.
- Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI* (Vol. 19, pp. 6300-6308).
- Khan, D. (2021). What are the types of machine learning? *Python in Plain English*. Retrieved from: <https://python.plainenglish.io/what-are-the-types-of-machine-learning-540b15dc467f>
- Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44, 100450.
- Lynn, S. (2018). An introduction to word embeddings for text analysis. *Data Science, Startups, Analytics, and Data Visualization*. Retrieved from: <https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/>
- Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Picador Press.
- Morales, M. (2020). *Grokking deep reinforcement learning*. Manning.
- Meurers, D. (2012). Natural language processing and language learning. *Encyclopedia of Applied Linguistics*, 4193-4205. Routledge.
- Murphey, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.
- Paiz, J. M., Tonecelli, R., & Kostka, I. (2025). *Artificial intelligence, real teaching: A guide to AI in ELT*. University of Michigan Press.
- Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data*. O'Reilly.
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3), 2495-2527.
- Vajjala, S. (2012). Machine learning and applied linguistics. In *The Encyclopedia of Applied Linguistics*. Wiley.
- Vasiliev, Y. (2020). *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press.
- Yu, J. H., & Chauhan, D. (2024). Trends in NLP for personalized learning: LDA and sentiment analysis insights. *Education and Information Technologies*, 1-42.
- Yule, G. (2014). *The study of language*. Cambridge University Press.
- Zilio, L., Wilkens, R., & Fairon, C. (2017, September). Using NLP for Enhancing Second Language Acquisition. In *RANLP* (pp. 839-846).