Dragana Špica and Benedikt Perak

ENHANCING JAPANESE LEXICAL NETWORKS USING LARGE LANGUAGE MODELS

Extracting Synonyms and Antonyms with GPT-4o

Abstract This study presents an innovative approach to crafting and enhancing Japanese lexical networks by incorporating large language models (LLMs), especially GPT-4o, utilizing data from Matsushita's (2011) Vocabulary Database for Reading Japanese to accommodate various proficiency levels. Through this process, we extracted a total of 137,870 synonym relations and 54,324 antonym relations, forming a network comprising 104,427 nodes. A portion of the dataset underwent manual evaluation to determine the accuracy of the extracted synonym relationships, yielding an average evaluation score of 4.08 out of 5. Our findings demonstrate that almost 20% of extracted nouns are (near) synonyms, while the rest have various relation types to the source word including hyponymy, hypernymy, meronymy, class membership etc. The study emphasizes the synergy between AI-driven data generation and traditional lexicographic expertise, offering a scalable and adaptable framework for diverse linguistic applications, with implications for computational linguistics and NLP technologies.

Keywords Japanese lexical networks; large language models; GPT synonyms and antonyms; Semantic graph analysis; AI-enhanced lexicography

XXI EURALEX

1. Introduction

Lexical networks, graph representations of semantic and syntactic relationships between words, have proven invaluable in various natural language processing (NLP) tasks. These networks, when accurately constructed, offer insights into language structure, facilitate tasks like word sense disambiguation, and support language education (Navigli, 2009). Traditionally, building such networks has been a labor-intensive process, relying on expert lexicographers and time-consuming manual annotation.

Recent advancements in large language models (LLMs) have opened new avenues for automating and scaling the creation of lexical resources. These models, trained on vast amounts of text data, have demonstrated remarkable abilities in generating text, translating languages, and understanding complex instructions. However, their potential for constructing lexical networks remains largely unexplored.

This study aims to bridge this gap by proposing a novel methodology for constructing and enriching Japanese lexical networks using GPT-4o. We leverage the model's capacity to generate synonyms and antonyms, coupled with graph-based techniques, to create a comprehensive and interconnected representation of the Japanese lexicon. By incorporating additional linguistic features and evaluating the network's quality, we seek to demonstrate the effectiveness of our approach and its potential for broader applications in NLP and language education.

We introduce a methodology for extracting and examining lexical relationships, emphasizing the derivation of associative concepts through synonym and antonym connections, as well as semantic class labeling in the Japanese language. This approach utilizes prompts to query a large language model about word relationships, subsequently aggregating the responses to construct a comprehensive semantic network.

The integration of LLMs and graph procedures in the study of lexical relationships offers a modern approach to understanding and organizing vocabulary, particularly in the context of Japanese language learning and linguistic research (Aotani & Takahashi, 2021). The resulting lexical network is a kind of abstraction of the mental repository of word knowledge that facilitates efficient information retrieval and processing during language comprehension and production (Kovács et al., 2021).

State-of-the-art Large Language Models (LLMs) are pretrained on extensive text data from multiple languages, with the vast majority sourced from publicly available Internet resources (e.g., the latest Common Crawl dataset, which includes data from over 3 billion pages). Among this data, the Japanese corpus constitutes 0.11% of the total training dataset for GPT3 (Kawahara et al. 2024:266, gpt-3/dataset_statistics/languages_by_word_count.csv). Despite its seemingly modest proportion, this still represents a vast volume of data, challenging the scale of traditional NLP-labeled corpora.

The contributions of the paper include:

- Methodology for extraction of lexical relationships using large language models
- Edge Dataset of extracted and partially evaluated lexical synonymy[1] relations
- Methodology for construction of lexical networks using graph models
- The Dash Cytoscape graph application (available at the development link: http://liks.ffri.hr:8002).

2. Previous Research and Theoretical Background

There have been numerous efforts to organize Japanese vocabulary on semantic principles since the first edition of thesaurus Word List by Semantic Principles (WLSP) (NINJAL 1964) and its revised edition (NINJAL 2004). It classifies Japanese lexicon into hierarchical semantic categories, enhancing understanding and teaching. Most recently, efforts have been made to develop more sophisticated corpora and searching tools e.g., BCCWJ (Maekawa et al., 2014).

[1] Dataset with extracted synonyms is available at: https://docs.google.com/spreadsheets/d/13YdAdyecikMYV-pA6eQ8ZNibAgYRlZD5b/edit?gid=54417379#gid=54417379

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In addition to these efforts, there have been developments of the Japanese WordNet (Isahara et al., 2008; Goodman & Bond, 2021), which aligns with the structure of the English WordNet to illustrate semantic relationships such as synonymy and antonymy within the Japanese language. Many online resources, e.g., online dictionary *Goo jisho*, use Japanese WordNet. The EDR Electronic Dictionary also organizes Japanese vocabulary with semantic categorizations besides phonetic annotations (Takebayashi, 1993).

2.1 Applications of Graph Theory in Linguistics

The integration of semantic principles with computational methods naturally leads to the advancement of lexical graphs, which use graph theory to model complex relationships between words. The study of lexical graphs involves visualizing the relationships between lexical items, providing a structured and quantitative approach to understanding the lexicon (Vitevitch, 2008; Veremyev et al., 2019; Diessel, 2023). Lexical graphs are formed from nodes, representing words or lexemes, and edges, symbolizing the relationships or associations between these nodes (Perak & Kirigin, 2023). This method offers a visual and analytical framework to explore how words are interconnected within a language, helping uncover insights into the organization of the mental lexicon. For instance, graph analyses often reveal that lexical networks exhibit small-world characteristics, where high clustering coefficients and short path lengths enhance the efficiency of lexical retrieval and communication (Agustín-Llach & Rubio, 2024).

The construction of lexical graphs typically begins with the collection of lexical data. Classical natural language processing (NLP) techniques involve phases such as tokenization, lemmatization, and syntactic dependency tagging. Using these tools, lexical graphs can be constructed with edges denoting relationships determined by syntactic dependencies or collocations. For instance, the method presented by *Construction Grammar Conceptual Network* research group utilizes syntactic dependencies to create multilayer networks, which capture different lexical semantic features (Perak & Kirigin, 2023). These graphs facilitate tasks such as word sense disambiguation, sentiment analysis, and semantic search by providing a rich contextual framework for interpreting lexical data (Kirigin et al., 2021; 2022).

2.2 Facilitating Lexical Extraction With Large Language Models

Traditional NLP processing steps are challenging because they necessitate high computational resources and linguistic expertise to manage the subtleties of a linguistic task. Moreover, mapping syntactic features to semantic features introduces further complexity as it involves understanding broader context-dependent meanings, often not readily available from classical corpus methods.

Top tier Large language models (LLMs), like GPT-4 or GPT-4o significantly streamline the extraction and generation of lexical relationships by leveraging their extensive pre-training on diverse linguistic data. Key advantages can be summarized as follows:

- Dynamic Generation: LLMs can quickly produce multiple lexical relationships for any given word, reducing the need for manual data extraction and ensuring comprehensive coverage across diverse contexts (Zhang et al., 2023).
- Contextual Relevance: The model's ability to reason about lexical context (Perak et al., 2024) and generate contextually appropriate linguistic patterns reinforces relevant selection of the synonyms and antonyms.
- Scalability: Utilizing LLMs scales well with large numbers of queries, enabling the efficient processing of vast data inputs and the generation of detailed lexical networks without the exhaustive preprocessing required in traditional methods.

This modern generative approach allows for the extraction of rich lexical relationships, ensuring both comprehensive and contextually relevant lexical networks construction.

2.3 Graph Methods vs. Large Language Models: Rationale for Combining Approaches

While LLMs can generate extensive linguistic data, there are compelling reasons to combine graph methods. Graph methods provide a transparent and interpretable representation of lexical relationships, unlike the often-opaque nature of LLMs, where the internal representations are not easily accessible.

Secondly, graphs enable the visualization of complex lexical networks in a manner that is easy to comprehend. Using tools like Python's NetworkX or Igraph library, lexicographers can create visual representations that highlight the density and centrality of connections among lexemes. This visual format is particularly useful for identifying key semantic clusters, understanding the spread and influence of particular lexemes, and detecting patterns that might be obscured in the raw outputs of LLMs.

By employing graph methods, researchers and lexicographers can more effectively validate and fine-tune the lexical relationships generated by LLMs. Once the LLMs generate initial lexical relationships, these can be mapped onto a graph and examined for accuracy and relevance.

Furthermore, graphs allow for the incorporation of rich attributes for both nodes and edges. For instance, each node (lexeme) can store various attributes such as part of speech (POS), frequency, pre-2010 JLPT (Japanese Language Proficiency Test) level, and semantic domains, different attributes from diverse lexical resources, as well as different centrality measures. Edges (relationships) can include attributes like relational strength, type (synonym, antonym), and domain-specific context. This rich attribute management facilitates more nuanced analyses and the creation of robust lexical resources. Furthermore, graph methods excel at enabling easy updates and modifications. As new lexical data are generated or discovered, they can be seamlessly integrated into the existing graph structure.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In practical applications such as language learning tools, search engines, and natural language processing (NLP) applications, having a well-organized and transparent lexical network is invaluable. For example, language learners may benefit from visual aids that show the relationships between words, aiding their comprehension and retention (Bollegala et al., 2015; Tokuhiro, 2016). Similarly, NLP applications can leverage the structured data in graphs to improve algorithms for tasks like semantic search, text summarization, and sentiment analysis.

3. Methodology

The procedure for extracting lexical relations is illustrated in Figure 1 is detailed in the following section.

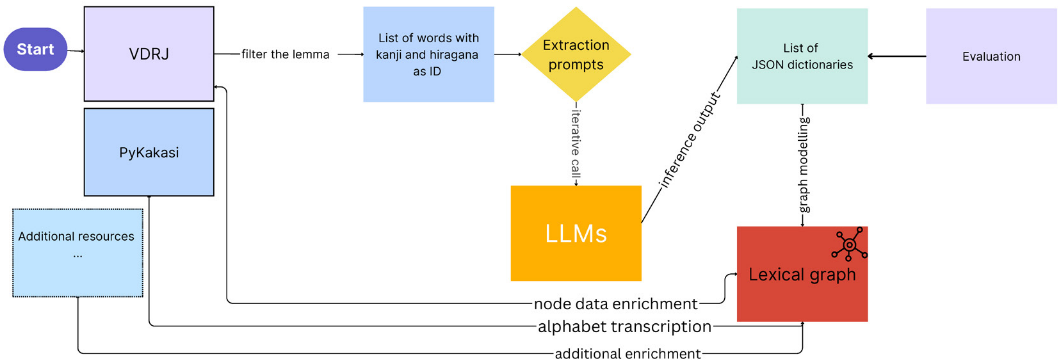

Fig. 1: Schematic representation of the lexical graph construction.

3.1 Lexicon Definition and Filtering

As a starting point for creating Japanese lexical network, we used *Vocabulary Database for Reading Japanese* (VDRJ) of over 60,000 Japanese words (Matsushita, 2011)[2], initially filtered to a subset of common nouns at the levels from N4 (the lowest) to N1 (the highest), according to pre-2010 JLPT scale[3].

[2] See also Matsushita (2016) for building corpus-based vocabulary syllabus for Japanese language learners.

[3] Pre-2010 JLPT scale corresponds to the current one in such a way that there is another level between levels 2 and 3 (see: https://jlpt.jp/e/reference/pdf/guide2011_e_02.pdf). Although the official correspondence between the JLPT scale and CEFR (Common European Framework of Reference for Languages: Learning, teaching, assessment) will not be available until 2025 (https://www.jlpt.jp/e/cefrlevel/index.html), some sources (Center for Language Education and Research Sophia University), provide the following correspondence of the levels of Japanese language courses to CEFR. (https://www.sophia-cler.jp/study/ja/pdf/LevelsofJapaneseLanguageCourses.pdf) JLPT5: A1-A2; JLPT4: A2; JLPT3: B1; JLPT2: B2; JLPT1: B2-C1.

The word list includes:

1. 3215 General common nouns (名詞 普通名詞一般), of pre-2010 JLPT levels N4 to N1.
2. 6200 general common nouns of N0 level (名詞-普通名詞-一般),
3. 1609 general common nouns that may function as verbal nouns (名詞-普通名詞-サ変可能),
4. 18 general common nouns that may function as nominal adjectives (-*na* adjectives) (名詞-普通名詞-形状詞可能),
5. 13 general common nouns that may function as adverbs (名詞-普通名詞-副詞可能),

Overall, we have included: 11055 unique lemmas for GPT processing.

3.2 Interfacing With GPT-4

The core of our methodology revolves around systematically interfacing the filtered lexicon with OpenAI's GPT-4o to extract lexical relationships. For this, we created a Python script iterating calls to the OpenAI API interface with a series of system prompts and precise query instructions to refine and structure the model's outputs. Each lexeme from the filtered list serves as a part of the input to GPT-4o, which is prompted to generate at least 15 synonyms and 5 antonyms for each item, as well as to infer their semantic domain.

For example, the lexeme '子供' (kodomo - child) would be introduced to GPT-4 with parts of prompts formulated as follows:

> *Here is a batch of lexemes (lemma) with their corresponding Part of Speech (POS) and katakana readings in Japanese.*
>
> *The tasks are to: 1) Translate lexemes to English and 2) to propose lexical relations in Japanese.*
>
> *For each lexeme in Japanese create a JSON dictionary with the following key: 'lexeme_data'. Inside this key determine: source_lemma, source_lemma_reading_hiragana, source_POS (from a list of {source_language_POS_list}).*
>
> *Write data about translated lexeme with key: 'lexeme_translation'. Inside this key determine: target_language, target_lemma, target_POS.*
>
> *Write the data about the synonym relation to other lexemes in Japanese with key: 'lexeme_synonyms'. Inside this key create a list of JSON dictionaries.*
>
> *For each lemma determine at least 15 synonyms in Japanese, with their corresponding POS from a {source_language_POS_list}:*

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

For each synonym determine strength of synonym relation as a float value in a range from 0.00 to 1.00.

Determine mutual sense and synonym domain in Japanese language and write this in Japanese.

Write data about the antonym relation to other lexemes in Japanese with the key: 'lexeme_antonyms'. Inside this key create list of dictionaries:

For each lempos determine at least 5 antonyms in {source_language}, with their corresponding POS from a {source_language_POS_list}:

For each antonym determine strength of antonymy relation as a float value in a range from 0.00 to 1.00. Where 0 refers to no synonymy and 1 refers to same meaning

Explain the antonymy relation and determine antonymy domain in {source_language} language and write this in {source_language}.

The prompt begins by introducing a batch of lexemes along with their corresponding Part of Speech (POS) and kana readings. This allows the GPT-4o to handle multiple lexemes in a structured manner. The tasks specified are two-fold: translating the lexemes to English as the target language and proposing lexical relations (synonyms and antonyms) in the source language.

The prompt's specificity and precision in query instructions are critical for minimizing ambiguity in the responses. By clearly defining the expected output format and the keys to be used in JSON dictionaries, the prompt guides GPT-4o to produce structured and uniform data[4]. The requirement to output data strictly in JSON format and avoid additional keys or characters ensures that the generated data is clean, easily parsable, and ready for subsequent analysis.

Furthermore, the output is rendered more controllable by setting the temperature parameter to 0. This parameter in the context of language models, such as GPT-4o, dictates the randomness of the generated text. A temperature of 0 ensures deterministic outputs, meaning the model always chooses the most probable next word, thereby reducing variability and enhancing consistency. This controllable output is particularly valuable in linguistic applications where precision and reliability are paramount, such as in the generation of synonyms and antonyms for lexical networks. By minimizing stochastic elements, the refined setting facilitates the creation of detailed and contextually accurate lexical relationships.

This structured approach supports efficient data manipulation and integration into lexical networks. Here's a simplified example demonstrating the output of the prompt:

[4] With the introduction of the model GPT4o-2024-08-06 it is now possible to define strict structured output (see: https://platform.openai.com/docs/guides/structured-outputs).

Example JSON output for the input Lexeme: 子供 (こども):

{ “source_lexeme”: { “lemma”: “子供”, “hiragana”: “こども”, “POS”: “Noun” },

“lexeme_translation”:{“target_language”:”English”,”target_lemma”:”child”,target_POS”: “Noun” },”lexeme_synonyms”: [{ “lemma”: 児童’, “hiragana”: “こども”, ‘relation’: ‘synonymy’, ‘strength’: 0.95, ‘explanation’: ‘Both refer to young individuals.’, ‘synonymy_domain’: ‘年齢’, ‘synonymy_domain_hiragana’: ‘じどう’, ‘synonymy_domain_translit’: ‘nenrei’, ‘synonymy_domain_translation’: ‘age’, ‘mutual_sense’: ‘若い人’, ‘mutual_sense_hiragana’: ‘わかいひと’, ‘mutual_sense_translit’: ‘wakaihito’, ‘mutual_sense_translation’: ‘young person’}, ...],

“lexeme_antonyms”: [{ “antonym_lemma”: “大人”,”hiragana”: “おとな”,”POS”: “Noun”, “antonym_translation”: “adult”, “antonym_strenght”: 1.0, “antonymy_domain”: “年齢”,”antonymy_domain_hiragana”:”ねんれい”,”antonym_explanation”: “child and adult are opposites in terms of age”,”antonymy_domain_translation”: “age”},...]
}

This example illustrates how the specified prompt leads GPT-4o to generate lexical data structured in JSON format.

3.3 Constructing Graph From JSON Output

To construct a graph object from the JSON output provided, we used Pandas and NetworkX Python libraries. The nodes in the graph represent lexemes (words), and the edges represent the relationships (synonyms and antonyms) between them with nodes and edges having specific attributes derived from the JSON data.

First, we parsed the JSON data to extract the relevant information about the source and target lexemes, i.e., synonyms, and antonyms. Importantly, the unique identifier for each node has to utilize both its kanji form and hiragana reading for a few critical reasons. (1) An orthographic form involving kanji may have multiple different words attached to it, e.g., 角 *kaku* ‘square, angle’, 角 *kado* ‘corner’, 角 *tsuno* ‘horn’ (anatomy). (2) Conversely, a word may have multiple orthographic variants - *tamago* ‘egg’ can be written using different kanji 卵 or 玉子, or can be written just in kana , i..e. たまご or タマゴ. (3) A word may have different nuances when written with different kanji, e.g., *furusato* ‘one’s hometown’ is written 故郷 or 古里 in different contexts. (4) 2 words with different phonological forms may share both orthographic form and the meaning, comprising a pair of synonyms, e.g., 故郷 *furusato* and 故郷 *kokyoo* ‘one’s hometown’. As will be shown below, intricacies of the writing system pose problems when determining synonymy.

Subsequently, we constructed edges datasets of the lexical network graph from synonym or antonym relationships along with multiple attributes derived from the JSON data, including *synonym_strenght*, *mutual_sense*, *synonymy_domain*, *synonymy_explanation*.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institute for the Croatian Language.

The *synonym_strenght* attribute quantitatively indicates the degree of semantic similarity between the source lexeme and its synonym, with a higher value indicating a closer semantic match. The *mutual_sense* with its hiragana transcript and English translation offers a shared contextual meaning, providing additional semantic context for the relation. Moreover, *synonymy_domain* and its hiragana and translation variants classify the categorical nature of the lexical relationship ensuring domain-specific clarity. The *synonymy_explanation* gives a qualitative articulation of why the terms are considered synonyms.

Edges representing antonyms are constructed with similar attributes to capture the oppositional nature of the lexical relationships. Attributes such *as antonym_strenght*, *antonymy_domain*, *antonym_explanation* are essential in this context.

The *antonym_strenght* attribute provides a quantitative measure of the strength of the antonym relationship, with a value of 1.0 indicating complete oppositeness. The *antonymy_domain* attribute categorizes the nature of this opposition by contextual domain, such as 年齢 *nenrei* 'age,' which in this case is further clarified with its hiragana representation. The *antonym_explanation* offers a qualitative description, for instance, detailing that "child and adult are opposites in terms of age".

3.4 Graph Enrichment

The initial lexical network graph, constructed using GPT-4o outputs, represents a complex relational structure of Japanese lexemes. To further enrich this graph with comprehensive lexical information, the original Vocabulary database for reading Japanese (Matsushita 2011) has been utilized to incorporate additional linguistic attributes:

- Standardized Frequency per million in 10 Written Domains (Fw) that provides a quantitative measure of how commonly each lexeme is used.
- Pre-2010 JLPT Level: The proficiency level of the lexeme.

The Jisho API (https://jisho.org/) has been employed providing a new JLPT scale where available as additional lexical attributes.

To facilitate the understanding and learning of Japanese for non-native speakers we inserted alphabet (*romaji*) transcriptions from hiragana using PyKakasi, a comprehensive Python library available on GitHub (https://github.com/miurahr/pykakasi).

3.5 Graph Analysis After Creating the Lexical Network

The network is composed of 137,870 edges representing synonymy relationships between lexemes and 54,324 edges representing antonymy relationships. In this lexical network 60,750 nodes are derived from synonym edges. When additional semantic dimensions such as mutual sense and domain are considered, the number of nodes connected by synonym edges expands to 99,068.

In terms of antonymy, the network includes 19,412 nodes connected by antonym edges. When domain nodes are included, the count of nodes linked by antonymy increases to 20,360.

The network thus contains a total of 65,807 nodes that are connected by either synonymy or antonymy edges, and inclusive with mutual sense and domain, the number of nodes rises to 104,427.

Table 1: Nodes and edges statistics for extracted Lexical Network

Type	Count
Nodes from synonym edges	60750
Nodes from synonym edges including mutual_sense and domain	99068
Nodes from antonym edges	19412
Nodes from antonym edges including mutual_sense and domain	20360
Nodes from synonym and antonym edges	65807
Nodes from synonym antonym edges including mutual_sense and domain	104427
Synonym edges	137870
Antonym edges	54324

4. Visual Representation Using Dash Cytoscape

To effectively visualize the complex relationships within Japanese lexical networks, Dash Cytoscape (https://dash.plotly.com/cytoscape) a tool for interactive graph visualizations, has been utilized. Visualization involves generating elements that represent both nodes and edges with detailed attributes (see Fig 1.). The app is currently running on a development server with address http://liks.ffri.hr:8002 .

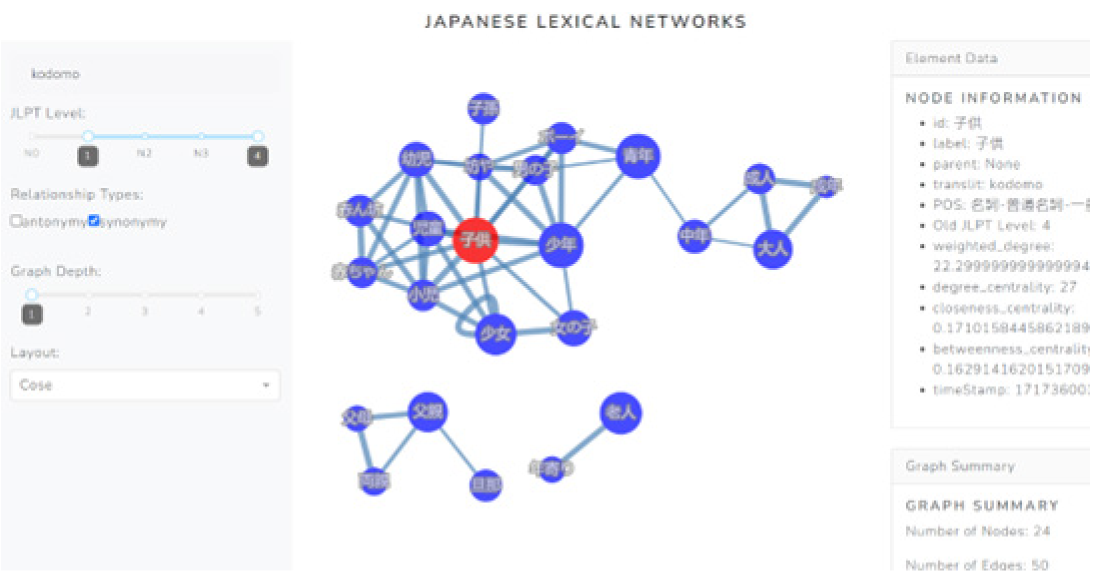

Fig. 2: Japanese Lexical networks example illustration for the lexeme '子供' *kodomo* 'child' filtered for lexemes pre-2010 JLPT level (1-4) and synonym relations in graph depth 1 using layout Cose.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The interactive features provided by Dash Cytoscape facilitate user engagement and deper exploration of the lexical network:

- Search Functionality: Users can search for specific lexemes by their transliteration or node identifiers, including kanji or hiragana.
- Filter and Range Selection:
    - JLPT levels: This slider allows users to filter lexemes based on their pre-2010 JLPT level. Additionally, there is also 'N0 level', i.e., items not included in JLPT.
    - Relationship Type Selector: Users can choose between Synonymy, Antonymy and Domain types of lexical relationships.
    - Graph Depth: This slider lets users control the depth of the network graph. Increasing depth shows more layers of connected lexemes.
- Dynamic Layouts: Multiple layout options (e.g., breadthfirst, grid, circle, cose) can highlight various properties of the graph structure, such as centrality or clustering.

Filters enable users to customize which nodes and relationships are displayed based on their proficiency level and interest.

5. Evaluation

In this chapter, output items will be evaluated and some shortcomings of the model will be pointed out. The scalar approach to defining synonymy relationships—where similarity between word senses is quantified by a *synonym_strength* attribute ranging from 0 to 1—inevitably encompasses a broad spectrum of semantically related entities. However, during the evaluation, we recognized that this broad categorization could lead to oversimplifications, failing to capture the nuanced distinctions between different types of synonymy. To address this, we introduced additional Relation Type within synonym pairs.

Let us first consider key concepts for evaluating synonymy. Synonymy is usually defined as a phenomenon whereby a single meaning is associated with more than one distinct lexical item. There have been numerous attempts to identify near synonyms, e.g., Edmonds & Hirst (2002), Inkpen & Hirst (2003), Terada & Yoshida (2007). Most recently, Ishii & Sasaki (2024) proposed a method of synonym identification by mapping word meanings using dictionary definitions and Sentence-BERT. Arguably, fully synonymous words do not exist or are very rare (Lyons, 1968, p. 448). Thus, it is a common practice to combine synonyms and near synonyms in the same category, see *Word List by Semantic Principles* (NINJAL 2004, p. 13). Partial synonymy is a type of synonymy relating to a certain meaning, e.g., Japanese words *kotoba* 'language; word' and *go* 'word' are partly synonymous. Examples are also *hot* and *spicy* in English, (Taylor, 2002, p. 266), and *joozu* 'skillful' and *umai* 'skillful' in Japanese (Ishii & Sasaki, 2004, p. 2993). Near synonyms (NS), also called plesionyms by Cruse (1986, p. 285), are words similar in meaning, which tend not to be contrastive, but are used in different contexts (Taylor, 2002, p. 263). For convenience, Partial and Near Synonyms will be included in the same category, **Near Synonyms (NS)**.

Topic-related items are those which occur within the same or closely related semantic domain. Subtypes of TR items are following:

1. **TR_hyponyms** e.g., *sandaru* 'sandals' is a hyponym for *kutsu* 'shoes',
2. **TR_hypernyms** e.g., *idoo* 'movement' is a hypernym for *ryokoo* 'travel',
3. **TR_meronyms** e.g., *shokutaku* 'dining table' is a meronym for *daidokoro* 'kitchen' (Winston 1987: 421)
4. **TR_holonyms**, e.g., *kyooiku* 'education' is holonym for *benkyoo* 'studying',
5. **TR_class** e.g., *kobachi* 'small bowl' belongs to the same class as *sara* 'plate',
6. **TR_object with an attribute** e.g., *akajiso* 'red perilla (botany)' has an attribute *aka* 'red',
7. **TR_possession** e.g., *tokeishokunin* 'clockmaker' for *tokei* 'clock',
8. **TR_subject of activity**, e.g., *fuufu* 'married couple' for *kekkon* 'marriage'
9. **TR_topology**, e.g., *obentooya* 'bento shop' for *bentoo* 'bento (lunchbox)',
10. **TR_paraphrase**, e.g., *ushi no chichi* 'cow's milk' for *gyuunyuu* '(cow's) milk'.

Other labels are following:

ERROR_kanji_based is a type of relation where a target word is apparently listed only because its first kanji is the same as the source word but they are semantically unrelated, e.g., 角, with specified hiragana reading *kaku* 'corner, angle' gave words such as 角笛 *tsunobue* 'horn (music instrument)' and 角膜 *kakumaku* 'cornea (anatomy)'. This is probably related to the fact that the same grapheme 角 has 3 words attached to it *kaku* 'angle, square', *kado* 'corner' and *tsuno* 'horn (anatomy)'.

ERROR_verb was an instance of a wrong POS, namely, verb *kaimotomeru* 'to buy' was listed as a synonym for verbal noun *kaimono* 'shopping'. ***ERROR_Chinese***: in one case, as a synonym for 言葉 *kotoba* 'word, language', "言" was listed which does not exist in Japanese. It could be the Chinese homograph meaning 'word'. ***ERROR_same*** are instances where the source word was just repeated as the target.

Cases of multiple orthographic forms are listed as ***Orth_var***. Such problems have been discussed in corpus linguistics (see Kuroda et al. 2011). If an item represents an orthographic variation, it should not be regarded as a synonym. E.g., source word 御腹 *onaka* 'stomach' has also orthographic forms お腹, which the model lists

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

as synonym of a strength 0.9, lower than 腹 *hara* ‘belly’ with 0.95. This problem should be addressed in the subsequent models.

Let us now turn to more specific evaluation details. The evaluation dataset comprises a subset of synonym edges extracted from a larger lexical network focused on Japanese vocabulary related to the pre-2010 JLPT Level 4. This sample includes 1085 edges with 84 source lemmas that were randomly selected. These lemmas represent a diverse range of basic Japanese vocabulary, from everyday nouns like 人 *hito* ‘person’ to more specific terms like 結婚 *kekkon* ‘marriage’. Each source lemma has an average of 12.5 synonyms that altogether create a set with 1050 unique synonym lemmas.

The evaluation process, conducted manually by one of the authors, involved a systematic review of each word pair and generated metadata, using resources such as online dictionary *Goo jisho* https://dictionary.goo.ne.jp/ which includes Shogakukan Thesaurus (2003) and NINJAL (2004). The evaluator initially assessed each semantic relation with respect to the enlisted domain using numeric scores on a scale from 1 to 5, with 1 indicating poor and 5 excellent quality. The scores were then analyzed to produce statistical measures for each type of relationship within the dataset.

- EVALUATION_synonym: represents the assessment of the quality of the synonym relationship. The results show a mean score of 4.08 with a standard deviation of 0.70. The majority of evaluations are high, with 75% of scores being 4 or higher.

- EVALUATION_sense evaluates the alignment of mutual senses between the synonym pairs. It shows a high level of agreement, with a mean score of 4.97 and 75% of scores being 5.

- EVALUATION_domain assesses the domain consistency of the synonym pairs, with a mean score of 4.91 and a standard deviation of 0.49. Similar to the sense evaluation, the scores indicate strong domain alignment.

The synonym relationships in the evaluation set are categorized according to their types described above. The distribution of these relation types is represented in the Table 2:

Table 2: Relation types per category. Relation Type is a proposed class of semantic relationship between synonym pairs. Count is the number of synonym pairs analyzed for each relation type. Mean Strength represents the average strength of synonym relationships within each relation type. Std Dev is the standard deviation of synonym strength, indicating the variability or spread of the synonym strength values within each relation type. Min Strength is the minimum synonym strength observed within each relation type. Max Strength is the maximum synonym strength observed within each relation type. Correlation (Synonym Strength, Evaluation Synonym) represents the correlation between synonym strength and the evaluation score.

Relation Type	Count	%	Mean Strength	Std Dev	Min Strength	Max Strength	Correlation
TR_MERONYM	500	46.1	0.542080	0.18603	0.150000	0.920000	0.090661
NS	215	19.8	0.836419	0.10769	0.350000	0.980000	0.222864
TR_HYPONYM	133	12.3	0.587218	0.17512	0.100000	0.950000	0.112714
TR_CLASS	132	12.2	0.564773	0.18520	0.150000	0.880000	0.106177
TR_HYPERNYM	32	2.9	0.657812	0.230045	0.200000	0.950000	0.177710
ERROR, KANJI BASED	16	1.4	0.490000	0.172378	0.200000	0.800000	NaN
TR_OBJECT_WITH_ATTRIBUTE	13	1.2	0.419231	0.137747	0.200000	0.650000	-0.532802
TR_HOLONYM	11	1.0	0.613636	0.130558	0.400000	0.800000	0.034641
ERROR, SAME	8	0.7	0.912500	0.210017	0.400000	1.000000	NaN
TR_SUBJECT_OF_ACTIVITY	5	0.5	0.680000	0.115109	0.500000	0.800000	NaN
ORTH_VAR	4	0.4	0.925000	0.064550	0.850000	1.000000	NaN
ERROR, SEMANTIC	4	0.4	0.600000	0.070711	0.550000	0.700000	-0.471405
TR_POSSESSION	4	0.4	0.537500	0.262599	0.150000	0.700000	-0.714545
TR_TOPOLOGY	3	0.3	0.516667	0.275379	0.200000	0.700000	-0.419314
TR_PARAPHRASE	2	0.2	0.650000	0.070711	0.600000	0.700000	NaN
ERROR, VERB	1	0.1	0.650000	NaN	0.650000	0.650000	NaN
ERROR, CHINESE	1	0.1	0.750000	NaN	0.750000	0.750000	NaN
ERROR_STRUCTURAL	1	0.1	0.600000	NaN	0.600000	0.600000	NaN

Near Synonyms comprise 19.8% and Topic related words, including all 10 subtypes, account for 77.1% in the dataset. NS has a highest mean strength (0.836), indicating that GPT self evaluation coincides with classification of the synonym pairs as generally closely related, with a moderate correlation (0.222864) to the evaluation score. TR (Topic-Related) items show lower mean strengths (from 0.542 to 0.419) and lower correlations to the evaluation score, suggesting these relationships are more diverse in nature.

The model correctly lists words with strong synonymy relation. For instance, for source word *ryoori* 'cooking, cuisine, dish' GPT lists target words *choori* 'cooking', *suiji* 'cooking', and *kukkingu* 'cooking'; and also *shokuji* 'meal' and *tabemono* 'food, dish', partial synonyms on the meaning dish' which we all classified as NS.

Let us consider examples of mutual sense and synonymy domain. Item *otokonoko* 'boy', has synonymy_strength 0.95 with *shoonen* 'boy' and is labeled near synonym;

their mutual_sense is *wakai dansei* 'young male' and synonymy_domain *nenrei* 'age'; next item is *kodomo* 'child', strength 0.85, whose mutual_sense is *wakai hito* 'young person' and synonymy_domain also *nenrei* 'age'. Next 2 items, *danshi* 'male child' (0.88) and *danji* 'male child' (0.75) have both mutual_sense *dansei no kodomo* 'male child' with synonymy_domain *seibetsu* 'gender'; *jidoo* 'pupil' (0.8) has mutual_sense *gakkoo ni kayou kodomo* 'school-going child' with synonymy_domain *kyooiku* 'education' and synonymy_explanation "Both refer to children, typically of school age."

The model correctly differentiates cases when a lexeme has 2 POS's, e.g., noun and nominal adjective: *genki* 'vigorously, healthy' and 'vigor, health' yield 2 types of synonyms, *kappatsu* 'lively' *kappatsusa* 'liveliness'.

There are cases when GPT does not list multiple meanings of a word. E.g., *midori* has 2 main meanings – 'green color', 'greenery'. However, the model captures only the second one giving topic related words e.g., *kusa* 'grass', *ha* 'leaves', mori 'wood'. This is not the case with e.g., *aka* 'red color' where various nuances of red are enumerated (however, it has no multiple meanings). This could be mitigated by using a more sophisticated prompt.

As for the number of NS for each entry, there is room for improvement. For example, 菓子 *kashi* 'confectionary' includes a synonym *okashi* which differs from source only by polite prefix *o-*, 2 near synonyms *suiitsu* 'sweets' and dezaato 'dessert' and the next 11 words are hyponyms e.g., *kyandii* 'candy', *chokoreeto* 'chocolate' etc. Words such as *kanshoku* 'food eaten between meals' and *oyatsu* 'snack' are conspicuously absent although we would expect to find them here. Also, it is noticed that all hyponyms are items with sweet taste despite the fact that *okashi* includes savory snacks, too. In the future research, a more systematic comparison with available resources e.g., Japanese WordNet will be considered.

It should be added that lists of near synonyms and related words may be relevant for Japanese language education, e.g., developing vocabulary syllabus, since a teaching unit usually involves specific topic(s). Yamauchi (2013) gives a comprehensive overview on the relation between a topic and the choice of vocabulary, and research on the recently developed *Japanese Topic-Oriented Conversation Corpus (J-TOCC)* (Nakamata et al. 2023) shows such relation even more specifically.

6. Discussion

6.1 Synergy of AI Tools and Traditional Lexicography

The integration of AI tools, specifically large language models such as GPT-4o, with traditional lexicographic methods marks a significant advancement in the field of lexicography (De Schryver, 2023; Lew, 2024).

AI tools, such as GPT-4o, revolutionize this process by automating the extraction and generation of lexical relationships. The collaborative approach leverages the computational strengths of AI with the nuanced understanding of human expertise.

It extends the applicability of lexicographic resources across languages and dialects, making it possible to construct comprehensive lexical databases or for less-studied languages and dialects more quickly than traditional methods alone would allow. Additionally, this synergy supports the development of advanced language learning tools and natural language processing applications, which require robust and detailed lexical databases to function effectively.

In addition, LLMs like GPT-4o can be viewed as an extension of linguistic corpora. These models are pre-trained on vast and diverse datasets that encapsulate a wide array of linguistic patterns, usages, and contexts. However, it is crucial to acknowledge that the output generated by these models, especially in lexical items, does not always equate to words (i.e. long word units) as defined in standardized linguistic resources such as UniDic used in e.g., the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). In contrast, LLMs might generate lexical items based on statistical probabilities and patterns observed during training, which may include neologisms, colloquialisms, or less standard forms.

6.2 Refining LLM Prompts With Relationship Types

A potential avenue for further research involves refining LLM prompts to explicitly incorporate the identified relationship types. By providing specific prompts that include terms such as "near synonyms," "hyponyms," or "meronyms," we can potentially elicit more precise and nuanced lexical relationships from the model. For instance, instead of a general prompt requesting synonyms, a more specific prompt could ask for "near synonyms with a similar level of formality" or "hyponyms of the target word." This level of granularity could significantly enhance the quality and specificity of the extracted lexical data, leading to more accurate and informative lexical networks.

Additionally, exploring the potential of combining multiple LLMs or fine-tuning existing models on specialized lexical datasets could further improve the accuracy and depth of the generated lexical relationships. This could involve training LLMs on curated corpora of lexical data, including manually annotated examples of various relationship types.

6.3 Adaptability to Other Languages or Different LLMs

The methodology presented is adaptable to other languages and LLMs. By applying the same core principles and leveraging multilingual LLMs, detailed lexical networks can be constructed for various languages. This involves tailoring the lexicon to the target language and utilizing a suitable LLM to generate synonyms, antonyms, and other lexical relationships. The flexibility to employ different LLMs enhances the potential for creating nuanced and culturally specific lexical networks (Huang et al., 2023).

7. Conclusion

This study proposes a novel method for constructing Japanese lexical networks by LLMs such as GPT-4o. The approach involves:

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

1. **Extracting Lexical Relationships:** GPT-4o is used to generate synonyms, antonyms, and domain information for Japanese words.
2. **Constructing Lexical Networks:** A graph is built with nodes representing words and edges representing relationships (synonyms/antonyms) with attributes like strength and domain.
3. **Enriching the Network:** Additional information like frequency and JLPT level is added from external resources.
4. **Interactive Visualization:** Dash Cytoscape allows users to explore the network by searching, filtering, and customizing layouts.

Evaluation of a sample set showed high accuracy for synonym relationships (mean score 4.08) and strong alignment of sense and domain. The study acknowledges limitations of a single "synonym strength" measure and proposes further classification of synonym types (e.g., near synonyms, topic-related). It was established that in the sample dataset Near Synonyms comprise 19.8% and Topic related items, including all 10 subtypes, account for 77.1%. Future work will focus on using these identified relation types to refine LLM prompts, aiming to improve the precision and accuracy of generated lexical relationships.

This approach offers a scalable and adaptable framework for building comprehensive lexical networks, with significant implications for computational linguistics and NLP applications.

References

Agustín-Llach, M. P., & Rubio, J. (2024). Navigating the Mental Lexicon: Network Structures, Lexical Search and Lexical Retrieval. *Journal of Psycholinguistic Research*, *53*(2), 21.

Aotani, N., & Takahashi, S. Y. (2021). An analysis of Japanese EFL learners' lexical network changes. In *Frontiers in Education* (Vol. 5, p. 621437). Frontiers Media SA.

Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2021). Lexical sense labeling and sentiment potential analysis using corpus-based dependency graph. *Mathematics*, *9*(12), 1449.

Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2022). Graph-Based Taxonomic Semantic Class Labeling. *Future Internet*, *14*(12), 383.

Bollegala, D., Maehara, T., Yoshida, Y., & Kawarabayashi, K. I. (2015). Learning word representations from relational graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

Cruse, D.A. (1986). *Lexical Semantics.* Cambridge University Press

De Schryver, G. M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387.

De Schryver, G. M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387.

Diessel, H. (2023). *The construction: Taxonomies and networks.* Cambridge University Press.
Edmonds, P., & Hirst, G. (2002). Near-Synonymy and Lexical Choice, *Computational Linguistics*, *28*(2), 105–144.

Goodman, M. W., & Bond, F. (2021). Intrinsically interlingual: The Wn Python library for wordnets. In: *Proceedings of the 11th Global Wordnet Conference* (pp. 100–107).

Huang, S., Mei, Y., Wang, Y., Yang, Z., & Gong, S. (2023). *Ranking of Large Language Model (LLM) Cultural Bias.* World Association of Artificial Consciousness.

Inkpen, D., & Hirst, G. (2003). Near-Synonym Choice in Natural Language Generation. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III: Selected papers from RANLP 2003.*

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation.* (pp. 2420–2423).

Ishii, Y., & Sasaki, M. 2024. Sentence-BERT to gogi teigibun o riyoo shita gogikan no ruigi hantei shuhoo [Synonym Detection between Senses of Japanese Words Using Sentence-BERT and Sense DefinitionsSentence-BERT]. *Shizen gengo shori gakkai happyoo ronbunshuu* (The Association for Natural Language Processing).

Kovács, L., Bóta, A., Hajdu, L., & Krész, M. (2021). Networks in the mind–what communities reveal about the structure of the lexicon. *Open linguistics*, *7*(1), 181–199.

Kuroda, K, Kuribayashi, T., Bond, F., Kanzaki, K., & Isahara, H. (2011). Nihongo waadonetto no ihyooki taioo to heikoo koopasu e no gogi tagudzuke. [Orthographic variants and multilingual sense tagging with the Japanese WordNet]. In *17th Annual Meeting of the Association for Natural Language Processing*, p. A4–1, Toyohashi.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1–8.

Lyons, J. (1968). *Introduction to Theoretical Linguistics.* London: Cambridge University Press.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, *48*, 345–371.

Matsushita, T. (2011). *Nihongo o yomu tame no goi deetabeesu, Ver. 1.11* [The Vocabulary Database for Learners of Japanese, Version 1.11] Retrieved December 7, 2023, from http://www17408ui.sakura.ne.jp/tatsum/database.html#vdrj.

Matsushita, T. (2016). Koopasu shutsugendo kara mita goi shirabasu [Vocabulary syllabus based on corpus frequency]. In A. Mori, & K. Shuppan (Eds.), *Niizu o fumaeta goi shirabasu* (pp. 53–77).

Nakamata, N., Ota, Y., Kato, E., Sawada, H., Shimizu, T., & Mori, A. (2023). "Nihongo wadaibetsu kaiwa koopasu: J-TOCC" no kaisetsu [Introduction to Japanese Topic-Oriented Conversation Corpus: J-TOCC]. In Nakamata, N. (Ed.), *Wadaibetsu koopasu ga hiraku nihongo kyooiku to nihongogaku* (pp. 15–51). Hituzi syobo.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys* (CSUR), *41*(2), 1–69.

NINJAL (National Institute for Japanese Language and Linguistics). (2004). *Bunruigoihyoo – zooho kaiteiban* [Word List by Semantic Principles – Revised and Enlarged Edition], Dainippon tosho

Oliveira, H. G., & Rodrigues, R. (2023). GPT3 as a Portuguese Lexical Knowledge Base?. In *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 358–363).

OpenAI. (2023). GPT-4 Technical Report. *ArXiv:2303.08774*

Perak, B., & Kirigin, T. B. (2023). Construction Grammar Conceptual Network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, *29*(3), 584–614.

Perak, B., Beliga, S., & Meštrović, A. (2024). Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)* (pp. 220–229).

Takebayashi, Y. (1993). EDR electronic dictionary. In *Proceedings of Machine Translation Summit IV* (pp. 117–126).

Taylor, J. R. (2002). Near synonyms as co-extensive categories: 'high' and 'tall' revisited. *Language Sciences, 25*, 263–284.

Terada A., & Yoshida M. (2007). A System for Constructing a Synonym Dictionary, *Journal of Natural Language Processing Journal of Natural Language Processing, 15*(2),39–58.

Tokuhiro, Y. (2016). Goimappu o mochiita shokyuu kanji goi kyoozai no kaihatsu. [Development of beginner's kanji vocabulary teaching materials using vocabulary maps]. *JSL Kanji gakushuu kenkyuukai-shi*, *8*, 15–19.

Veremyev, A., Semenov, A., Pasiliao, E. L., & Boginski, V. (2019). Graph-based exploration and clustering analysis of semantic spaces. *Applied Network Science*, *4*, 1–26.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval?. *Journal of Speech, Language, and Hearing Research*, *51*(2). https://doi.org/10.1044/1092-4388(2008/030)

Winston, M. E., Chaffin, R., & Herrmann, D. (1987). Taxonomy of Part-whole relations, *Cognitive Science, 11*, 417–444.

Yamauchi, H. (Ed.) (2013). *Jissen nihongo kyooiku standaado* [*Practical Japanese Language Education Standards*], Tokyo: Hituzi shobo

Zhang, Z., Li, B., & Zeng, Z. (2023). Review of research on synonym equivalence relation mining. In: *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)* (pp. 1–6). IEEE.

Zhang, Z., Pu, P., Han, D., & Tang, M. (2018). Self-adaptive Louvain algorithm: Fast and stable community detection algorithm based on the principle of small probability event. *Physica A: Statistical Mechanics and Its Applications*, *506*, 975–986.

Acknowledgements

This work was supported by projects: a) UNIRI CLASS1 A1 Mikrokvalifikacije: Jezične tehnologije i digitalna obrada teksta (JTDOT), b) UNIRI PROJEKTI ISKUSNIH ZNANSTVENIKA 2023: Matematičko modeliranje u obradi prirodnog jezika (uniri-iskusni-prirod-23-150). This paper was prepared as part of the institutional project *Japanese language and society: basic concepts and empirical analysis* (10/1/2023-9/30/2027), Juraj Dobrila University of Pula, Croatia.

Contact information

Dragana Špica
Sveučilište u Puli
dragana.spica@unipu.hr

Benedikt Perak
Sveučilište u Rijeci
bperak@uniri.hr