Construction of Evaluation Dataset for Japanese Lexical Semantic Change Detection

Zhidong Ling
Tokyo Metropolitan University
ling-zhidong@ed.tmu.ac.jp

Taichi Aida
Tokyo Metropolitan University
aida-taichi@ed.tmu.ac.jp

Teruaki Oka
Hitotsubashi University
teruaki.oka@r.hit-u.ac.jp

Mamoru Komachi
Hitotsubashi University
mamoru.komachi@r.hit-u.ac.jp

Abstract

Semantic changes in German and some other European languages is an activate field of study. These languages have diachronic corpora, lists of words with known semantic changes, and evaluation datasets manually annotated with semantics or the degree of semantic change. Several researchers studying semantic change have used datasets of these languages in their evaluations. Since the Japanese language did not have a corpus-based manually annotated dataset, we developed the first dataset to study diachronic semantic changes in the Japanese language. We based our study on two diachronic corpora, which covered the near-modern and modern periods, respectively. We selected target words from a previous study, and compared sampled usages each other. In total, we have collected judgments from three experts for 540 usage pairs across the target words. Our dataset will be available on GitHub[1].

1 Introduction

The study of linguistic phenomena called diachronic semantic change has gained considerable popularity in natural language processing (NLP). For example, diachronic semantic change detection has been carried out to automatically detects words whose meanings have changed over time (Hamilton et al., 2016b; Schlechtweg et al., 2019; Giulianelli et al., 2020; Kutuzov et al., 2021). Evaluation is an essential topic in the diachronic semantic change detection task. Previous studies have identified a set of target words that have either undergone semantic changes or have retained meaning over a given timeframe within the target corpora. The outputs of their models have mainly been assessed based on the extent to which the words under evaluation were included at the top when the words in the dataset are sorted by degree of semantic change. This evaluation covers several languages and samples from different domains and periods, thus providing enriched data for the evaluation. However, owing to several previous studies using varied data, there is a lack of direct comparison between the performance of these models (Hamilton et al., 2016b; Kulkarni et al., 2015; Frermann and Lapata, 2016). Moreover, further detailed evaluation and analysis of semantic changes cannot be performed using only semantically changed or stable words.

To address this issue, Schlechtweg et al. (2018) introduced Diachronic Usage Relatedness (DURel), which is a general framework to determine the degree of semantic change in words. Their framework calculates the degree of semantic change in words over time by annotating examples obtained from the target corpora. Based on this framework, we can conduct more detailed evaluation and analysis. Evaluation datasets that follow this framework for semantic change detection have been created and published for various languages (Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021; Giulianelli et al., 2020; Chen et al., 2022). In the SemEval-2020 Task 1, Schlechtweg et al. (2020) provided evaluation datasets in four languages created by the extended DURel framework, and presented a shared task for semantic change detection. However, in the Japanese language, although there is a list of Japanese words whose meanings may have changed, no datasets created by the DURel framework for evaluating the task are not available yet.

In this study, we constructed a dataset with manually annotated degrees of semantic change for the Japanese language. We followed Mabuchi and Ogiso (2021) and chose two periods: Meiji & Taisho Eras (1860s−1920s) and Heisei Era (1990s−2010s). During both eras, the Japanese language underwent a significant change owing to social and linguistic factors. Our dataset con-

[1] https://github.com/tmu-nlp/JapaneseLSCDataset

tains nine target words, including six semantically changed words and three semantically stable words. Each target word has 60 usage pairs that contain two usages. Each usage has up to 50 context tokens on both sides of the target word, resulting in a maximum usage length of 101 tokens. For each usage pair, they are sampled from the same corpus or different corpora. In this setting, we used two Japanese corpora, Corpus of Historical Japanese (CHJ)[2] and Balanced Corpus of Contemporary Written Japanese.[3] Subsequently, three experts gave judgments based on the of DURel framework. Overall, our dataset contains nine target words with 1,620 judgments.

The contributions of this study are as follows.

- We constructed a dataset for studying diachronic semantic change in the Japanese language. Using the DURel framework, we asked three experts to perform the annotations, and successfully captured and quantified the degree of semantic changes on a diachronic corpus of the Japanese language.
- We consulted non-experts in the Japanese near-modern and modern languages to perform the same annotation and compared their results with those of experts, demonstrating no difference in agreement between experts and non-experts.

This dataset will be available on GitHub upon acceptance.[4]

2 Related Works

2.1 Semantic change detection methods

Semantic change has been widely studied in NLP. In Kulkarni et al. (2015), the diachronic corpus was divided into several time snapshots. For each time snapshot, word vectors were learned for the target words and then the vectors were aligned into the oldest time snapshot vector space. A time series of semantic changes can be constructed using the distance between the vector spaces of the arbitrary snapshot and the oldest snapshot. To quantitatively evaluate their method, they artificially introduced semantic changes of words to target corpora. Yao et al. (2018) learned word embeddings across all time slices while incorporating regularization terms to ensure smooth transitions in the embedding changes over time. In evaluating the performance of this method, they selected technical words with known variations and verified whether the proposed method could detect them. Giulianelli et al. (2020) used BERT to obtain contextualized embeddings and group them into clusters, and their labels can be considered as a label of word senses. Subsequently, by comparing the distribution of word sense labels across two different periods, the diachronic shift in meanings was quantified.

2.2 Evaluation sets of semantic change detection

When quantitatively analyzing model performance, a typical approach is selecting certain predetermined words to evaluate binary judgments (Sagi et al., 2009; Hamilton et al., 2016a). However, these studies used different standards for word selection, which inhibited the comparison of different models and methods. Furthermore, binary evaluation leads to an evaluation that overlooks the degree of the semantic change, which limits the possibility of expanding the evaluation dataset.

Schlechtweg et al. (2018) presented a general framework DURel for language-independent annotation of diachronic usage relatedness. It asks annotators to compare and grade the semantic relatedness of target words across the usage pairs, rating from unrelated to identical (corresponding to scores from 1 to 4) as shown in Table 1. Starting with a German dataset along with DURel, datasets for many languages have been created following this framework, such as Russian (Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021), English (Giulianelli et al., 2020) and Chinese (Chen et al., 2022).

Recently, the DURel framework was improved to provide a benchmark for evaluating unsupervised semantic change detection tasks. SemEval-2020 task 1 provided evaluation datasets for English, German, Swedish, and Latin as benchmarks for shared tasks (Schlechtweg et al., 2020). These datasets were created using the Diachronic Word Usage Graph (DWUG), which is an extension of the DURel framework (Schlechtweg et al., 2021). The DWUG used usage graphs to show the changes in the senses of target words over

[2] https://clrd.ninjal.ac.jp/chj/

[3] https://clrd.ninjal.ac.jp/bccwj/en/index.html

[4] Owing to copyright reasons, we cannot release the original usage of the corpus used in this study. Instead, we will release a search ID for each usage which usage can be downloaded using the CHUNAGON search system.

time. The graph was weighted and undirected, with nodes representing the usages of the target words and weights indicating semantic relatedness scores provided by human annotators. As a result, benchmarks were provided for various evaluations based on the DURel framework, allowing them to be used to evaluate the performance of the models and approaches on different languages (Kutuzov et al., 2022; Zamora-Reina et al., 2022; Giulianelli et al., 2022).

2.3 Japanese semantic change analysis

Aida et al. (2021) analyzed pointwise mutual information (PMI)-based models to measure semantic changes and applied those models to English and Japanese corpora. When analyzing the target words, they assumed that the semantically changed word list provided by Mabuchi and Ogiso (2021) were changed words, while words not included in the list were unchanged. Additionally, the degree of semantic change was ignored and treated uniformly. They calculated the mean reciprocal rank on the word list and compared the performances of the BERT and PMI-based approaches. Kobayashi et al. (2021) applied and compared two different BERT-based methods, dictionary-based (Hu et al., 2019) and clustering-based (Giulianelli et al., 2020), to analyze the sense-level semantic change in Japanese. They did not conduct a quantitative evaluation owing to the lack of a Japanese evaluation set. Instead, they focused on the word list created by Mabuchi and Ogiso (2021) and performed an intense qualitative analysis.

3 Annotation

We applied the DURel framework to create a dataset for semantic change in the Japanese language. In this framework, the degree of semantic change of words is defined by the semantic relatedness across usage pairs from the target corpora. The semantic relatedness between the usage pairs is manually annotated. The target words in usage pairs with similar meanings received a high score, whereas those with distant meanings received a low score.

We followed the same procedure in DURel as shown in Table 1 to annotate the semantic relatedness with a 4-point scale score. Table 2 presents an example of an actual annotation. Previous studies have divided target words by their part-of-speech

score	relatedness
4	Identical
3	Closely Related
2	Distantly Related
1	Unrelated
N/A	Cannot Decide

Table 1: A 4-point scale of semantic relatedness score in the annotation.

such as nouns, verbs, and adjectives. However, for example, *sahen* nouns (e.g., '**優勝**' (win)) are allowed to occur independently as a noun and can also be used as a verb by adding '**する**' after it (e.g., '**優勝する**' (to win)) in the Japanese language. The short unit word[5] used in this study combines the above variations into one part-of-speech tag. Therefore, we do not differentiate the part-of-speech of the target words in the usage when selecting the target words and sampling the usage of the words.

The corpora for sampling were divided into two parts, C_1 and C_2, for the early and the late time periods, respectively. Then for each target word, the DURel framework sampled three groups of usage pairs from the corpora as $Eariler$, $Later$, and $Compare$. Pairs in $Earlier$ contain two usages from C_1, pairs in $Later$ contain two usages from C_2, and pairs in $Compare$ group contain usage from C_1 and C_2 each. After that, the semantic relatedness of the target word between usages was annotated in each group.

To quantify the semantic change of each target word w, we calculated the average of all scores in each group $\mathrm{Mean}(groupname_w)$. N/A (cannot decide) will be ignored in the calculation. When the $\mathrm{Mean}(groupname_w)$ is high, it indicates that the target words in the usage pairs within the group are more semantically similar, while a low $\mathrm{Mean}(groupname_w)$ suggests that the semantics of the target words in some of the usage pairs within the group are far apart from each other. The $Earlier$ and $Later$ groups represent the distribution of semantic relatedness scores by annotation of extracted usage pairs from C_1 and C_2, respectively, and the $Compare$ group captures the semantic shifting across C_1 and C_2 by directly compare the usages from both periods.

[5] https://clrd.ninjal.ac.jp/bccwj/en/morphology.html

Target	Usage 1	Usage 2	Score
林檎	. . . **３日間**りんご**だけを食べるというアップルダイエットを覚えている？** . . .	. . . **リンゴを表現するときに、全部の**リンゴ**を「固い、固い、固い、固い」と書く子がいると思えば** . . .	4
apple	. . .Do you remember the Apple Diet, where you eat only **apples** for three days?. . .	. . .When describing apples, I thought there is a child writing 'hard, hard, hard, hard' for all **apples** . . .	
写真	. . . **趣味ある挿畫も多く、彩色畫**寫眞**版等の口畫、例によりて麗はしとも麗はし（定價十五錢博文舘發行）** . . .	. . . **ブーケ・ブートニア・ヘア＆メイク・**写真**・ビデオ・教会使用料・牧師謝礼・送迎など合計（２名）￥五十万前後** . . .	3
photo	. . . Many of the illustrations are of tasteful taste, and the coloring **pictures** and plates are as beautiful as ever (published by Hakubunkan for 15 sen). . . .	. . . Bouquet, boutonniere, hair, makeup, **photo**, video, church fee, pastor's gratuity, and transportation total (for 2 persons) around \ 500,000 . . .	
椅子	. . . **然るに當時米國政府には、フヒッシユなる人、正に大藏卿の**椅子**に在りしが、氏は今政府の甘んじて** . . .	. . . **余輩或る夕べそのプロメネードに遊び、**椅子**によりて、橙色に赤色に紅色に色と光とを變へて** . . .	2
chair	. . . However, at that time, in the U.S. government, there was a man named Fish, who was the **chair** of the Department of the Treasury, but he was not a member of the current government . . .	. . . One evening, I played at the promenade, leaned back in my **chair**, the color and light changed to orange, red, and crimson. . .	
結構	. . . **犬のふんの始末など**結構**泥臭い作業もあるので、いい加減なスタイルへ徐々にシフト。** . . .	. . . **お忙しいなか、こんな勝手なお願い、お聞き届け頂けないかもしれません。それは、もうそれで**結構**でございます。** . . .	1
structure; quite	. . . Some of the work is **quite** muddy, such as cleaning up dog feces, so we gradually shifted to a more lax style. . . .	. . . I know you are busy and may not be able to hear my selfish request. That is **fine**. . .	

Table 2: Examples in the actual annotation task. The target word in the usage is enclosed by **. Each example of the target word has an English-translated version below the Japanese version.

Based on this value in each group, the DURel framework uses two metrics to calculate the degree of semantic change of each target word w:

$\Delta Later_w = \text{Mean}(Later_w) - \text{Mean}(Earlier_w)$: It is the mean of all annotation scores given by annotators in $Later$ subtracted from the mean of all scores in $Earlier$. This metric measures the changes in meaning relatedness from the $Earlier$ to the $Later$ group. A positive value calculated by this measure indicates the meaning increases and the target word undergoes an innovative change, and a negative value indicates a reductive change.

$\text{Mean}(Compare_w)$: It is the mean of the annotated scores in $Compare$ group. In this measure, a higher/lower score indicates a weak/strong change between two periods, regardless of the type of change.

4 Setting

4.1 Annotator

Three annotators, who are highly educated native speakers of Japanese with extensive knowledge of the language, were involved in the evaluation process. All annotators are researchers at the National Institute for Japanese Language and Linguistics.[6] Similar to Schlechtweg et al. (2018), before beginning the annotation task, annotators were required to read the guideline shown in Figure 3 and instructed to complete a prepared tutorial. This tutorial followed the same format as the actual annotation, providing unambiguous answers and ensuring that the decisions of authors were correct. Consequently, all participants achieved almost perfect scores, with no unreasonable mistakes.

4.2 Corpora

Following the setting of Mabuchi and Ogiso (2021), we used two corpora, the Corpus of Historical Japanese (CHJ)[7] and the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[8] (Maekawa et al., 2014). Table 3 shows the statistics of two corpora. To avoid the influence of genre, we followed (Mabuchi and Ogiso, 2021) to use only the magazine parts of both corpora. Both

[6] https://www.ninjal.ac.jp/english/
[7] https://clrd.ninjal.ac.jp/chj/
[8] https://clrd.ninjal.ac.jp/bccwj/en/index.html

	Corpus name	Period	Domain	Tokens
C_1	CHJ	1874–1925 (Meiji & Taisho Eras)	Magazines	12.6M
C_2	BCCWJ	2001–2005 (Heisei Era)		4.4M

Table 3: Statistics of corpora.

Changed	結構 優勝 椅子 教授 免許 適当
Stable	写真 林檎 主張

corpora can be accessed using the corpus search engine CHUNAGON[9].

CHJ is a diachronic corpus covering a long range of Japanese texts from the Nara era (800C–) to the Meiji and Taisho eras (1860s–1920s). Texts are annotated with morphological information such as lemmas, readings, and part-of-speech tags for the short unit word. In our setting, we used only the Meiji and Taisho eras as our target periods.

BCCWJ is the only balanced corpus of contemporary written Japanese texts covered from the later Showa era (1970s–1988) to the Heisei era (1989–2010s). Texts in BCCWJ are annotated with similar morphological information to CHJ. It contains over 100 million words across genres, and the magazine part used in our setting has approximately 4.4 million short unit words.

4.3 Target words

We used nine words for annotation, as shown in Table **??**, with six words selected as semantically changed words and three words designated as semantically stable words. Six semantically changed words were selected by experts based on previous research by Mabuchi and Ogiso (2021). These experts compiled a list of potential words that may have changed meaning, drawing from literature and dictionaries. They subsequently investigated the frequency changes and word sense categorization to eliminate inappropriate words, as part of the preliminary process for dataset construction. Regarding the selection of the three stable words, since no existing research has provided insights into words that have maintained a consistent meaning throughout the target period, we consulted the Daijisen Japanese Dictionary.[10] From the dictionary, we randomly chose three words that possessed only one explanation, strengthening their status as semantically stable words.

4.4 Sampling

The usage of the target word was collected from the corpora using an online corpus search engine called CHUNAGON. CHUNAGON displays the usages of a target word in the corpus by entering the surface or lemma of the target word. During the Meiji and Taisho eras, the old Japanese language may have adopted different Chinese character representations compared to modern Japanese. For example, in the Heisei era, the word '**写真**' (photo) was written differently than '**寫眞**' (photo but old notation) in Meiji and Taisho eras. To collect all usages of the word, we employed a lemma search in CHUNAGON. When searching for word usage in the CHUNAGON system, the results are unique when using lemma, lemma reading, and word type. For the current study, we used only lemma when sampling, which is appropriate because the target words we selected were unique with only a lemma in the search. However, this is an ambiguous search.

Specifically, we selected 20 pairs from CHJ for the $Earlier$ group and 20 pairs from BCCWJ for the $Later$ group. For the $Compare$ group, we created 20 usage pairs by sampling one usage each from CHJ and one usage from BCCWJ. All sampled usages were different and marked with a unique ID in the CHUNAGON search system.

5 Analysis

5.1 Lexical Semantic Change

In Figure 1, we present the ranking of the target words based on their $\Delta Later$ values. Our analysis revealed the presence of words that exhibit changes in meaning at both ends of the graph, while certain words remained relatively stable in the middle portion. There are no clear boundaries between the values in the chart, which may be owing to the limited number of target words. Nonetheless, the findings align reasonably well with the positions of the two types of words, which

[9] https://clrd.ninjal.ac.jp/en/tool.html#02

[10] https://japanknowledge.com/en/contents/daijisen/

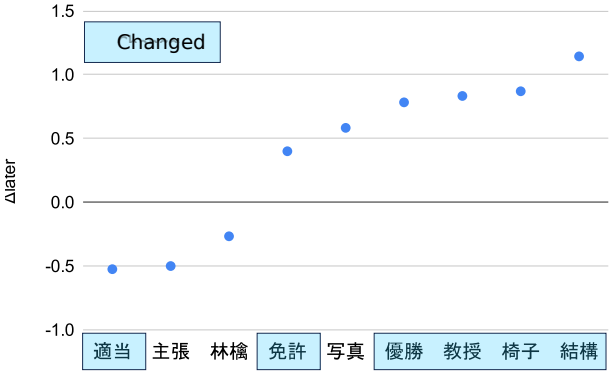

Figure 1: Rank of target words according to increase in $\Delta Later$. The absolute value of $\Delta Later$ indicates the degree of change, a positive value indicates a decrease in meanings, and a negative value indicates an increase in meanings

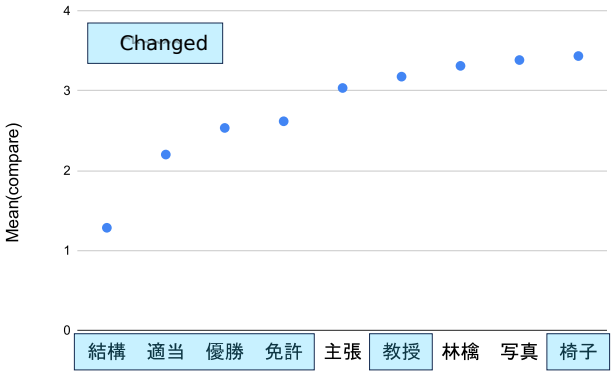

Figure 2: Rank of target words according to increase in $\mathrm{Mean}(Compare)$. The value of $\mathrm{Mean}(Compare)$ is greater than or equal to 0. The greater the value, the less the degree of change.

confirms the validity of our approach to selecting target words.

Figure 2 shows the ranks of the target words sorted by their $\mathrm{Mean}(Compare)$, which directly compares the semantics of target words in the two periods. Notably, stable words such as '**林檎**' (apple) and '**写真**' (photo) received remarkably high scores, indicating minimal semantic changes. The three semantically stable words '**主張**' (claim), '**林檎**' (apple), and '**写真**' (photo) selected in this study should not change meaning, as they are single-meaning words. However, in this setting, both $\Delta Later$ and $\mathrm{Mean}(Compare)$ showed equivalent values to the semantically changed words. This result indicates that while some previous studies assumed that most words do not change meaning (Kulkarni et al., 2015; Hamilton et al., 2016b), selecting words that do not change their meaning reliably is difficult. However, semantically changed words such as '**適当**' (appropriate → unreliable; sloppy) and '**結構**' (construction; quite; all right → quite; all right) obtained lower scores, suggesting a significant shift in their meanings.

The semantic change of word '**結構**' (structure; quite; all right → quite; all right) were successfully captured by $\Delta Later$ and $\mathrm{Mean}(Compare)$. '**結構**' has the highest $\Delta Later$ and lowest $\mathrm{Mean}(Compare)$ score, implying that it underwent a drastically reductive change during the two periods.

Table 4 shows an annotation example of the usage pair in group $Compare$. The target word '**結構**' in Usage 1 sampled from CHJ means *structure*, while Usage 2 has a word that means *quite*. A thorough examination of the usages sampled from the corpus revealed that in C_1 the word '**結構**' had several meanings: *structure* as a noun, *splendid; nice; all right* as adjective uses and *quite; fairly* as adverb uses. In C_2, the noun-meaning *construction* disappeared, and the word was used as an adjective or adverb in most usages. Another word, '**適当**', received the lowest $\Delta Later$ and the second lowest $\mathrm{Mean}(Compare)$, which indicates a drastically innovative change. As shown in Table 5, the usage pair in the $Earlier$ group has two usages such that both words '**適当**' mean *appropriate*. In the $Later$ group, we can see the meaning of *appropriate* in Usage 1, and the meaning of *sloppy* appears in Usage 2. The sampled usages reveal that in C_1 the word '**適当**' mainly means *appropriate*; however, in C_2, '**適当**' was frequently used as an adjective which means *unreliable; sloppy*, while *appropriate* retained a certain frequency of use.

However, some words could not be detected successfully using these metrics. In $\Delta Later$, a semantically changed word '**免許**' (allow → license) had a lower value of 0.40 than that of a stable word '**写真**' (photo) 0.58. Hence we consider that '**免許**' is not that changed than '**写真**'. In $\mathrm{Mean}(Compare)$, the value of '**免許**' was 2.62, which is lower than '**写真**', which was 3.38. We consider the meaning of '**免許**' has changed more than '**写真**'. As a possible cause, the original sense of '**免許**' disappeared, and a new sense appeared in the modern period, and the $\Delta Later$ could not detect it effectively because of the nearly identical distribution of word senses in both periods. $\mathrm{Mean}(Compare)$, however, could success-

Data group	Usage 1	Usage 2	Score
Compare	. . . **遂に一派の法門を開き、その**結搆**布置常に人の意表に出でたり、人物も亦磊々落々、常斑の上に出でて、大にしては元祿時代の雛形とな** . . . . . .He finally opened a school of Buddhism, and his **structures** and installations were always in the forefront of people's minds, and he was a model of the Genroku era in general.. . .	. . . **萩谷今触ると、**結構**重いんです。** . . . . . .Hagiya Now that I touch it, it's **quite** heavy.. . .	1, 1, 1

Table 4: Usage pairs in group *Compare* of the target word **結構**.

Data group	Usage 1	Usage 2	Score
Earlier	. . . **亦**適當**の香味を有せざるを以て** . . . . . .As long as it exists in tea in the proper **appropriate** form. . .	. . . **茶にも正しく**適當**の形ちで存在する以上は** . . . . . .Also because it does not have the **appropriate** flavor.. . .	3, 3, 4
Later	. . . **そのために必要な技術として最も**適当**なものを、一つ選べ** . . . . . .Choose the one technology that is most **appropriate** for this purpose.. . .	. . . **車番を控えられ、照会にかけられることは間違いないので、**適当**に走らせて中野のあたりで乗り捨ててきた。** . . . . . . I had no doubt that the number of the car would be kept and I would be able to make an inquiry, so I drove it **sloppily** and abandoned it around Nakano.. . .	1, 2, 2

Table 5: Usage pairs in group *Earlier* and *Later* of the target word **適当**.

fully detect such changes because it directly compares word senses from both periods.

5.2 Inner-annotator Agreement

Non-experts are familiar with modern Japanese, but not with the old Japanese. Experts, on the other hand, are familiar with both. We were interested in whether a non-expert with little knowledge could perform this task with the same quality as an expert. To study this, we compared the inner-annotator agreement between experts and non-experts. We asked three annotators from a crowdsourcing service to investigate the influence of the background knowledge of an annotator on the annotation score. These annotators are all native Japanese and university graduates. We report four types of agreement measures: Krippendoff's alpha, Spearman ρ correlations, Cohen's Kappa, and pairwise agreement. We excluded pairs that result in NaN when calculating the mean of Spearman ρ and the mean of Cohen's Kappa. As shown in Table 6, for two groups of annotators, there was no difference across all measures.

We also present a usage pair with varied annotator scores. Notably, in the results from both experts and non-experts, we observe that when the target word appears within an idiom, it tends to receive lower agreement scores. For instance, the

	α	ρ	Kappa	pairwise
Experts	0.29	0.59	0.18	0.48
Non-experts	0.23	0.56	0.16	0.51

Table 6: Comparsion of average inner-annotator agreements between experts and non-experts.

word '**優勝**' (good; excellent) in *Earlier* had usages where '**優勝**' is a part of the idiom '**優勝劣敗**'. The idiom '**優勝劣敗**' means survival of the fittest, and when translated literally into English, '**優勝劣敗**' means the good will remain, and the bad will leave. We show the usage pair with <CHUNAGON ID> below:

1. . . . **其上二國が其地利上の**優勝**なる點を合併すれば太平洋の海權としては眞に比類なき國となるべき地勢なるは明々白々復た疑ふべきなし** . . .

   Moreover, if the two countries were to merge their superior geographical **good** points (advantages), it is no doubt that they would become a truly unparalleled nation in terms of the Pacific Ocean.

   <60M **太陽** 1901_08038>

2. . . . **世界は**優勝**劣敗の戰場なり、弱者**

の強者に制せらるるは止むを得ざるの必然なり . . .

The world is a battlefield that **the good will remain** and the bad will leave, and it is inevitable that the weak will be defeated by the strong

<60M 太陽 1901_08003>

This usage pair received scores (4, 3, 3) from three expert annotators and (1, 2, 4) from three non-expert annotators. When comparing these usages with standalone usages of '**優勝**', we observed that annotators who assigned lower scores (indicating dissimilar senses) perceived '**優勝**' in '**優勝劣敗**' as a distinct word from '**優勝**' when it appeared independently. Conversely, annotators who assigned higher scores explained that the '**優勝**' in '**優勝劣敗**' has a similar meaning to the independent usage. From the data we acquired, we could not identify marked differences in performance between experts and non-experts. However, due to the small amount of data we obtained, we did not perform a statistical analysis to investigate this phenomenon.

6 Conclusion

This study presented the first human-annotated dataset used to investigate semantic changes in the Japanese language. We applied the DURel framework to the Japanese language to construct the dataset. Using two metrics, $\Delta Later$ and $\mathrm{Mean}(Compare)$, we achieved reasonable annotation results and successfully captured semantic changes in the target words, proving the validity of the DURel framework for the Japanese language. Furthermore, we assigned annotation tasks to two groups of annotators with different knowledge backgrounds. The assessment of the inter-annotator agreement indicated that within our established framework, there was no disparity in efficiency between experts and non-experts.

Our future goals include expanding the dataset and exploring evaluations. First, in addition to the corpus adopted in this study, another corpus that covers the Showa era, which is between the Taisho era and the Heisei era from 1926 to 1989, is being released. We aim to sample and annotate them to increase the coverage of the current data further. Also, as long as the amount of data can be expanded, we plan to perform a statistical analysis of the difference between experts and non-experts. Second, as a possible direction, we aim to define types of semantic change such as broadening and narrowing using the usages in the corpus for further detailed analysis.

Acknowledgements

The work reported in this article was partly supported by the NINJAL collaborative research project‘ Extending the Diachronic Corpus through an Open Co-construction Environment ’ .

References

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. A comprehensive analysis of PMI-based models for measuring semantic differences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 21–31, Shanghai, China. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2022. Do not fire the linguist: Grammatical profiles help language models detect semantic change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Kazuma Kobayashi, Taichi Aida, and Mamoru Komachi. 2021. Analyzing semantic changes in Japanese words using BERT. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 270–280, Shanghai, China. Association for Computational Lingustics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625 – 635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical profiling for semantic change detection. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Yoko Mabuchi and Toshinobu Ogiso. 2021. An attempt to construct a dataset of words for semantic change analysis of modern Japanese. In *Proceedings of the annual meeting of the Association for Natural Language Processing 2021*, pages 1166–1170.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 673 – 681, New York, NY, USA. Association for Computing Machinery.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Annotation guideline

Figure 3 illustrates the guideline presented to the annotators during annotation, and Figure 4 shows the English-translated version of the annotation guideline. This guideline is based on the English version of DURel[11].

[11] https://durel.ims.uni-stuttgart.de/guidelines?lang=en

【どんなタスク？】

表記は同じでも、意味が変わる単語があります。例えば、「中国」という単語は国名でもありますが、「中国四国地方」というと日本の中国地方のことで意味が違います。
こうした「複数の意味を持つ単語」を含んだ文書（の一部分）を2つ提示します。作業者はその2つを読み比べて、そこで使われている対象語の意味が同じか違うか（意味的な関連性）を関連度という4段階のスコアで評価します。

例えば、次のような2つの文書が与えられます。評価してほしい対象単語は
対象単語
のように、アスタリスクで囲っています。この例だと、**端末** です。

1) 基本的にはネットワークというインフラがあって、そのネットワークの外側にいろいろな**端末**をつけることでネットワークの高度利用を図っていこうというのがトレンドだ。

2) かた鹿角嘴と龍廟嘴砲臺との間に亘りて設けられ、他の一部は日島より北の方劉公島に及ぶ、日島と劉公島間の防材は劉公島の陸地に近き處の**端末**に少しく間隙あれども、此處は岩石多く舟を行る能はず

1）と2）を読んだうえで、対象単語 **端末** の意味がどれくらい関連しているか、次の評価スコアを使って主観評価していただきます。

==========
4 対象単語の意味が一致した (Identical)
3 対象単語の意味が近い (Closely Related)
2 対象単語の意味が少し関連している (Distantly Related)
1 対象単語の意味が完全に異なる (Unrelated)

0 判断できない (Cannot decide)
==========
1 -> 2 -> 3 ->4 の段階的スコアで、
4に近いほど同じ意味で使われており（意味的に関連している）、
1に近いほど違う意味で使われています（意味的に関連していない）
です。

上の **端末** の例では、1）ではスマホのようなコンピュータ端末の意味ですが、2）では『末端、端（はし）』の意味なので、1 に近い評価をします。

対象単語の意味をどうしても決められない場合、
そもそも提示された2つの対象単語が異なる単語である場合（「端末」と「和歌」が同時に提示された）など、
評価スコアを1～4に決められない場合は「その他」で理由を教えてください。

【注意】

時代による意味の変化も調べたいので、1)2)のように、現代の文書だけでなく、
明治・大正時代 の文書(漢字表記が現代と違う場合がある)も出てきます。
また、複合名詞などが本番作業で出てくる場合があります。例えば「椅子」の用例に「車椅子」の用例が出てきます。その場合はアノテータの直感でスコアをつけていただきます。

対象単語数:9単語
読み比べる文書:各単語に60ペア（20+20+20）
全540ペア

【アノテーションする手順】

1.チュートリアルを必ず受けてください。解答完了後フォームを提出して、フィードバックを提示します。フィードバックにある責任者の解答と理由を読んで、タスク作業中にご参照ください。

チュートリアルのリンク:
https://docs.google.com/forms/d/e/1FAIpQLSczQHya-Xz6Z3GVWi4v_2Ir-rzeJxffPKQ8clVYU7AFvh1M1A/viewform?usp=share_link

2.本番作業は各単語に3つずつのフォーム［Compare・Earlier・Later］に回答する必要があります。1フォームに20問あります。
本番作業する際に、アノテータを区別するためにメールアドレスを収集します。悪用はしません。

Figure 3: Annotation guideline for annotators to explain the task.

[English Version]

[What is this task?]
Some words have the same notation but have different meanings. For example, the word "Chugoku" is also the name of a country China, but "Chugoku-Shikoku region" has a different meaning as it refers to the Chugoku region of Japan.
Two documents (parts of documents) containing such "words with multiple meanings" are presented. The operator compares the two documents and evaluates whether the meanings of the target words used in the two documents are the same or different (semantic relevance) using a 4-point score called relevance.

For example, given two documents as follows The target words to be evaluated are
target word.
and are enclosed with asterisks. In this example, it is **端末**.

1) 基本的にはネットワークというインフラがあって、そのネットワークの外側にいろいろな**端末**をつけることでネットワークの高度利用を図っていこうというのがトレンドだ。
Basically, the trend is to have an infrastructure called a network, and then to make advanced use of the network by adding various **terminals** to the outside of that network.

2) かた鹿角嘴と龍廟嘴砲臺との間に亙りて設けられ、他の一部は日島より北の方劉公島に及ぶ、日島と劉公島間の防材は劉公島の陸地に近き處の**端末**に少しく間隙あれども、此處は岩石多く舟を行る能はず
(Something) was set from Rokkakushi to Ryumyoshi Artillery battery, and the other parts extend from Hinoshima to LiuGong Islet in the north, The barrier between the two islands has a small gap at the **end** near the land of LiuGong Islet, but there are too many rocks in this area for a boat to pass through.

After reading 1) and 2), you will be asked to subjectively rate how relevant the meaning of the target word **端末** is, using the following rating score.

==========
4 The meanings of the target words matched (Identical)
3 The meanings of the target words are closely Related (Closely Related)
2 The meanings of the target words are slightly related (Distantly Related)
1 The meanings of the target words are completely different (Unrelated)

0 Cannot decide
==========
Score on a scale of 1 -> 2 -> 3 ->4,
The closer to 4, the words are used in the same sense (semantically related),
The closer to 1, the differently it is used (not semantically related).

In the **端末** example above, 1) means a computer terminal like a smartphone, but 2) means 'end, edge', so it is rated closer to 1.

When the meaning of the target word cannot be determined by any means,
When the two target words presented in the first place are different words ("terminal" and "waka" are presented at the same time), etc,
If you cannot decide on a score between 1 and 4, please give a reason in the "Other" column.

[Notification]
Since we would like to study the changes in meaning over time, we will not only look at modern documents as in 1) and 2), but also documents from the Meiji and Taisho eras (where the Chinese characters may differ from those of the present day).
In addition, compound nouns, etc. may appear in the production work. For example, in the example of "chair", the example of "wheelchair" may appear.
Number of target words: 9 words
Documents to be compared: 60 pairs for each word (20+20+20)
540 pairs in total

[Annotation Procedure]
1. Be sure to take the tutorial. Submit the form after completing the solution and present the feedback. Please read the responsible person's answers and reasons in the feedback and refer to them during the task work.

Link to the tutorial:https://docs.google.com/forms/d/e/1FAIpQLSczQHya-Xz6Z3GVWi4v_2Ir-rzeJxffPKQ8clVYU7AFvh1M1A/viewform?usp=share_link

2. The production task requires you to answer 3 forms {Compare, Earlier, Later} for each word. 20 questions per form.
During the production work, we will collect email addresses to distinguish annotators. We will not abuse it.

Figure 4: English version of the annotation guideline.