

LSC-Pattern: Interpretable Lexical Semantic Change Analysis via Diachronic Sense Distributions

Zhidong Ling

Hitotsubashi University
shito@scl.sds.hit-u.ac.jp

Mamoru Komachi

Hitotsubashi University
komachi@scl.sds.hit-u.ac.jp

Abstract

Lexical semantic change detection identifies words whose meanings change over time, but it often cannot explain which senses and semantic properties are responsible for the change. We propose LSC-Pattern, a framework for interpretable lexical semantic change analysis that combines dictionary-grounded diachronic sense distributions with interpretable semantic dimensions. The framework first assigns each target-word usage to an explicit dictionary sense and then analyzes the resulting sense distributions along dimensions such as valence, semantic scope, and metaphoricity. We instantiate it for Japanese with LLM-based word sense disambiguation, which reaches 0.815 accuracy against majority-vote human annotation and outperforms five baselines; its sense-distribution change scores correlate with JaSemChange human judgments (SHC-BCCWJ; Spearman's ρ of 0.821 on seven overlapping entries and 0.580 on all 20). Case studies illustrate how sense redistribution and operationalized semantic signals can jointly support the discovery and interpretation of candidate types of change.

1 Introduction

Word meanings change over time. Lexical semantic change (LSC) studies how such changes can be identified from diachronic corpora, and computational work has made substantial progress with diachronic embeddings, contextualized representations, and clustering-based usage modeling (Hamilton et al., 2016; Kutuzov et al., 2018; Schlechtweg et al., 2020; Kutuzov et al., 2022). These methods make it possible to estimate whether a word has changed, but they often leave the linguistic source of the change underspecified.

Detecting that a word has changed is not enough because we also need to explain which sense is responsible for the observed change. Semantic change has long been described through typologies

of the relation between an old and a new meaning and of the cognitive motivations behind the shift (Blank, 1999; Blank and Koch, 1999; Koch, 2016; Juvonen and Koptjevskaja-Tamm, 2016). A word may become more negative because a negative sense becomes more frequent, or it may broaden because a previously specific sense becomes dominant. Without linking such property changes to explicit word senses, semantic change analysis remains difficult to inspect and difficult to use as linguistic evidence.

Existing lexical semantic change detection (LSCD) methods and semantic-dimension approaches provide useful signals, but they do not always connect semantic-property changes to concrete dictionary senses. Clustering methods based on contextualized embeddings can compare usage distributions across periods, yet the resulting clusters are latent and may not correspond to lexicographic senses (Kutuzov et al., 2022). Recent dimension-based work analyzes semantic change along interpretable properties such as sentiment, intensity, or breadth (Baes et al., 2024). These directions motivate a framework that keeps the interpretability of explicit senses while retaining the flexibility of dimension-level analysis.

We propose LSC-Pattern, a two-step framework that connects diachronic sense distributions with interpretable semantic dimensions. Figure 1 summarizes the framework. Step 1 assigns each target-word usage to a dictionary sense and aggregates the assignments into a sense distribution for each time period. Step 2 analyzes the resulting sense distributions with semantic dimensions selected for the target type of semantic change. Our central claim is that diachronic sense distributions provide the missing link between corpus-level change detection and linguistically interpretable semantic change patterns. Rather than optimizing for change-detection alone, the framework focuses on making semantic change evidence inspectable through dictionary

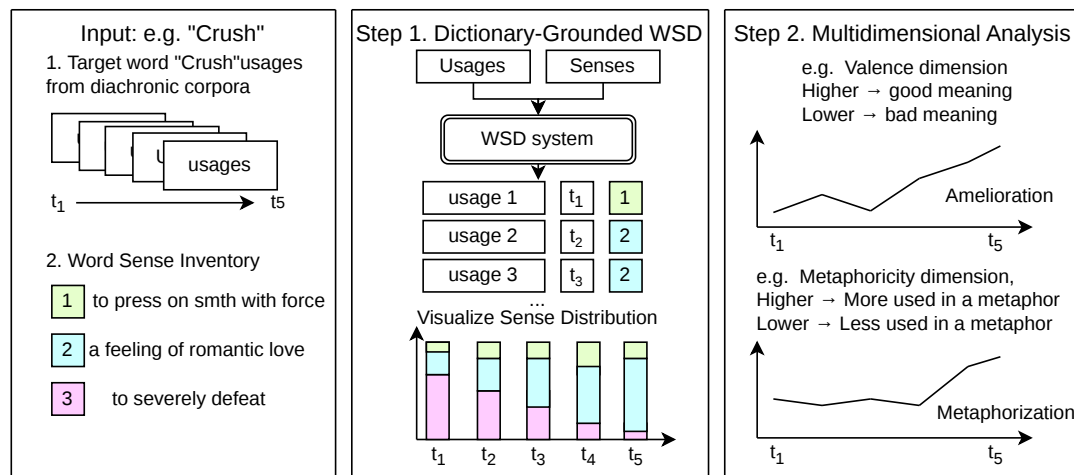


Figure 1: Overview of LSC-Pattern. The framework first maps diachronic usages to dictionary senses and then analyzes the resulting sense distributions with semantic dimensions selected for the target semantic change pattern.

senses, semantic dimensions, and attested usages.

In this study, we instantiate the framework with LLM-based dictionary-grounded WSD and three semantic dimensions corresponding to common semantic change patterns (de Sá et al., 2026). Valence corresponds to amelioration and pejoration, semantic scope corresponds to generalization and specialization, and metaphoricity corresponds to metaphorization. The framework itself is not tied to a particular WSD model or a fixed set of dimensions.

We evaluate the framework on Japanese lexical semantic change using three diachronic and contemporary corpora spanning the 1870s to the 2010s. We ask two research questions:

- **RQ1:** Can LLM-induced dictionary sense distributions capture lexical semantic change in Japanese?
- **RQ2:** How can semantic change patterns be interpreted by combining sense distributions with semantic dimensions?

This paper makes three contributions:

1. We propose LSC-Pattern, an interpretable framework that connects explicit dictionary senses with semantic-dimension analysis.
2. We show that dictionary-grounded WSD can construct interpretable diachronic sense distributions for Japanese over roughly 150 years.
3. We show how sense distributions combined with semantic dimensions can be used to analyze classical semantic change patterns such as

amelioration, pejoration, generalization, specialization, and metaphorization.

The code and annotation labels are available on GitHub¹.

2 Related Work

2.1 Lexical Semantic Change Detection and Interpretation

LSCD has been studied with both type-level diachronic embeddings and usage-level contextualized representations. Early approaches compare word embeddings trained for different time periods (Hamilton et al., 2016). More recent approaches represent individual usages with contextualized models and compare usage distributions across time (Kutuzov et al., 2022; Tang et al., 2023). Evaluation resources such as DUREl and SemEval-2020 Task 1 have further standardized graded semantic change evaluation through usage relatedness and human change judgments (Schlechtweg et al., 2020, 2024).

Recent work improves interpretability either by comparing sense distributions or by analyzing semantic change along interpretable dimensions. Sense-based approaches compare how usage clusters or induced senses are distributed across periods (Lau et al., 2014; Schlechtweg et al., 2021). Dimension-based approaches instead describe change through properties such as sentiment, breadth, intensity, or taxonomic relations (Baes et al., 2024). These two directions motivate the present paper: sense distributions identify which

¹<https://github.com/Lexuss-D/LSC-Pattern>

meaning is changing, while dimensions describe what kind of semantic property is changing. Recent work has also moved beyond fixed sense inventories and aggregate change magnitude in complementary directions. Fedorova et al. (2024) induce usage-specific senses via generated contextual definitions, while Cassotti et al. (2024) classify change types from WordNet relations, with SenseRel (Cassotti et al., 2026) further annotating denotational and connotational relations between senses.

LSC-Pattern connects dictionary-grounded sense distributions with semantic-dimension analysis to interpret dimension-level changes in valence, semantic scope, and metaphoricality through explicit sense-level mechanisms.

2.2 Dictionary-Grounded WSD with LLMs

Dictionary-grounded WSD links a usage to an explicit sense definition, which makes it useful for interpretable semantic change analysis. WSD identifies the intended sense of a word in context (Raganato et al., 2017). Recent neural methods often formulate WSD as context–gloss matching, where the model selects the dictionary definition that best matches the target usage (Huang et al., 2019; Blevins and Zettlemoyer, 2020). This setting is useful for our framework because it directly links corpus usages to dictionary senses.

LLMs provide a practical way to perform WSD by selecting a sense from candidate dictionary definitions. Recent studies show that LLMs can perform dictionary-grounded WSD effectively with glosses and contextual information (Sumanathilaka et al., 2024, 2025; Meconi et al., 2025).

Gloss-based WSD has also been used in LSC-related tasks. Rachinskiy and Arefyev (2022) use a gloss-based WSD model to obtain sense-informed contextualized representations for zero-shot cross-lingual LSCD, while Kokosinskii et al. (2024) combine gloss-based WSD with WSI and novel-sense detection to assign diachronic usages to existing or newly induced senses.

Our goal is not to propose a new WSD model. We use LLM-based WSD as a practical module for building diachronic sense distributions, while the main contribution lies in the semantic change framework and analysis.

2.3 Japanese Lexical Semantic Change

Compared with LSCD research on English and European languages, Japanese lexical semantic change remains relatively underexplored. Recent resources

such as CHJ, SHC, BCCWJ, and JaSemChange provide a basis for diachronic corpus analysis and evaluation in Japanese (Ling et al., 2023).

Our work uses these resources to analyze Japanese lexical semantic change through explicit sense-level evidence identified from dictionary senses and semantic dimensions.

3 The LSC-Pattern Framework

LSC-Pattern is a two-step framework that first builds diachronic sense distributions and then analyzes them with interpretable semantic dimensions. The two steps are modular and can be instantiated with different WSD systems, sense inventories, and semantic dimensions.

3.1 Dictionary-Grounded Diachronic Sense Distributions

The input of Step 1 is a target entry, its usages in each time period, and its dictionary sense inventory. Let w be a target entry. Let $\mathcal{S}_w = \{s_1, s_2, \dots, s_{|\mathcal{S}_w|}\}$ be the set of dictionary senses of w , augmented with a fallback label s_0 for usages that no listed sense fits. Let $\mathcal{U}_{t,w}$ be the set of usages of w in time period t . Each usage $u \in \mathcal{U}_{t,w}$ contains the target word and its surrounding context:

$$u = (\dots, c_{-2}, c_{-1}, w, c_1, c_2, \dots).$$

Each usage is mapped to one dictionary sense by a dictionary-grounded WSD function. We write this mapping as:

$$\hat{s}(u) = f_{\text{WSD}}(u, \mathcal{S}_w), \quad \hat{s}(u) \in \mathcal{S}_w.$$

The assigned usages are aggregated by period to obtain a diachronic sense distribution. The usages assigned to sense s_i in period t are:

$$\mathcal{U}_{t,s_i} = \{u \in \mathcal{U}_{t,w} \mid \hat{s}(u) = s_i\}.$$

We then compute the sense distribution of w in period t :

$$p_t(s_i \mid w) = \frac{|\mathcal{U}_{t,s_i}|}{|\mathcal{U}_{t,w}|}.$$

The output of Step 1 is $p_t(s_i \mid w)$, which shows how the relative frequency of dictionary senses changes over time. Because the sense inventory is fixed across periods, changes in $p_t(s_i \mid w)$ can be interpreted as changes in the relative prominence of dictionary senses rather than changes in latent clusters. When we need a scalar change score between two periods t and t' , we use Jensen–Shannon divergence between their sense distributions.

3.2 Semantic-Dimension Trajectories

Step 2 analyzes the sense distributions from Step 1 with semantic dimensions to produce interpretable semantic change trajectories. The researcher first chooses one or more semantic dimensions that match specific properties of word meaning, such as valence for positive or negative meaning. These dimensions are not fixed requirements of the framework; they can be replaced or extended when the analysis targets other types of semantic change.

For usage-level dimensions, such as valence and metaphoricality, the word-level trajectory can be computed directly as the average score of all usages in a period. In these cases, WSD is not required to create the scalar trajectory itself; its key role is to decompose that trajectory by dictionary sense. The decomposition shows which senses are frequent in each period and what dimension scores their usages receive.

For sense-level dimensions, such as semantic scope, the dimension score is attached to dictionary senses, and the period-specific sense distribution determines the word-level trajectory.

The output of Step 2 is a diachronic trajectory for each target word and each selected semantic dimension. This trajectory gives a compact view of how a property of meaning changes over time. The trajectory is interpreted together with the sense distribution.

For each target word, the framework supports the interpretation of observed semantic change patterns by jointly inspecting the sense distribution and the dimension trajectory. The dimension trajectory identifies the direction of change, while the sense distribution shows which senses become more or less frequent. This combination allows the analysis to move from a scalar change score to an explanation grounded in dictionary senses and corpus examples.

4 Japanese Case Study

4.1 Data and Sense Distribution Construction

Target Entries. The main analysis uses 150 target entries processed with the same WSD and semantic-dimension pipeline. We use the term target entry to refer to a lemma–reading pair in the dictionary inventory. We first filter the Iwanami Japanese Dictionary for polysemous entries. To ensure overlap with JaSemChange, we include polysemous entries from the benchmark and randomly sample additional entries from the remaining candidates until

Statistic	WSD eval set	Main analysis set
Target entries	50	150
Usages	976	213,650
Dictionary senses	162	501

Table 1: Statistics of the WSD evaluation set and the main analysis set. The WSD evaluation subset contains 976 manually annotated usages sampled from 50 entries.

Corpus	Period	Decades	Words
CHJ	1870s–1920s	6	12.3M
SHC	1930s–2010s	9	31.1M
BCCWJ	2000s	1	4.3M
Total	1870s–2010s	15	47.7M

Table 2: Statistics of the three corpora used in diachronic analysis. The 2000s decade is covered by both SHC and BCCWJ, so the Total row counts it once.

obtaining 150 target entries in total.

Diachronic Corpora. We apply LSC-Pattern to Japanese lexical semantic change from the 1870s to the 2010s. We use three Japanese corpora: the Corpus of Historical Japanese (CHJ) (National Institute for Japanese Language and Linguistics, 2026), the Showa-Heisei Corpus (SHC) (Ogiso et al., 2024), and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014)². All usages are grouped by decade³.

Table 2 shows the statistics of three corpora. We restrict the analysis to the magazine register to reduce genre variation across corpora and periods.

Dictionary-Grounded WSD. For each usage, the WSD model receives the target entry, usage context, and candidate dictionary senses, and outputs one sense index. Table 1 shows statistics of target entries. For each target entry, candidate senses are taken from the Corpus of Iwanami Japanese Dictionary Fifth Edition⁴. Each candidate sense includes a sense index, a dictionary definition, and example expressions when available. The target occurrence is marked in the usage context, and candidate senses are presented with their fixed Iwanami sense indices. Tables 8 and 9 show prompt templates. The model outputs the most suitable sense index and a short explanation. Only the predicted sense

²We use Long Unit Word segmentation to avoid extracting long phrases that merely contain the target entry. <https://clrd.ninjal.ac.jp/bccwj/en/morphology.html>

³Due to corpus license restrictions, we cannot redistribute the original usages, although they are obtainable from the original corpus providers.

⁴<https://www.gsk.or.jp/en/catalog/gsk2021-d>

index is used for quantitative analysis. The explanation is used only for inspection and error analysis. Appendix A.2 gives the exact prompt template and output parsing policy.

We implement the WSD module with QWEN3-8B in thinking mode (Qwen Team, 2025). Detailed inference parameters are reported in Appendix Table 7. All usages are processed independently with the same prompt format and decoding setting. The WSD outputs are then aggregated by entry, decade, and sense to obtain diachronic sense distributions.

Handling of Sense 0. Neither the model nor the annotators are forced to choose a listed sense. Both the WSD prompt and the annotation guideline provide an explicit fallback, Sense 0, for usages where none of the listed senses is suitable, where the context is insufficient, or where the usage is ambiguous. Sense 0 therefore mixes genuine gaps in the dictionary inventory with ordinary disambiguation failures, and we do not read it as a novel-sense label. Its downstream treatment depends on the dimension: valence and metaphoricity are computed from the usage itself and retain Sense-0 usages, whereas semantic scope requires an Iwanami-to-WordNet alignment and therefore excludes them. Section 5.1 reports how often Sense 0 occurs and what it is caused by.

Intrinsic and Extrinsic Evaluation. We first evaluate whether the model assigns correct dictionary senses to individual usages.

For this intrinsic evaluation, we use 50 target entries that are randomly sampled from the 150 target entries. For each target entry, we sampled at least one usage from each valid decade and up to 20 usages for annotation. Three annotators independently assign Iwanami dictionary senses to the candidate usages. The annotation guideline is shown in Appendix Table 10. All annotators are native speakers of Japanese with expertise in Japanese linguistics⁵. We report model accuracy against each annotator and against majority vote when at least two annotators agree.

We then evaluate whether the induced sense distributions align with human judgments in JaSemChange, a Japanese LSCD dataset based on diachronic usage-pair similarity judgments (Ling et al., 2023). In addition to the polysemous JaSemChange entries included in our 150-entry analysis set, we apply the same WSD and sense-distribution

pipeline to all 20 JaSemChange target words, including entries that Iwanami treats as monosemous. For each evaluation set, we compare sense distributions across corpus pairs and compute Spearman’s rank correlation with human change scores. We report results for the CHJ–BCCWJ and SHC–BCCWJ settings.

4.2 Semantic-Dimension Instantiation

We instantiate Step 2 with three semantic dimensions selected from common patterns in semantic change studies. Valence corresponds to amelioration and pejoration, semantic scope corresponds to generalization and specialization, and metaphoricity corresponds to metaphorization. Other dimensions, such as intensity, concreteness, register, or social meaning, can be added when the target analysis concerns other types of semantic change. All three dimensions are resource-based operationalizations, and we state up front what each one actually measures. Valence is computed from the words surrounding the target and therefore measures the evaluative environment in which the entry occurs, not the affective content of the entry itself. Metaphoricity is the output of a detector trained on contemporary Japanese and therefore measures what a present-day model recognizes as metaphorical, which systematically undercounts metaphors that have since been conventionalized. Semantic scope inherits whatever specificity the Japanese WordNet hierarchy encodes for the aligned synset. We consequently use the three trajectories as discovery signals: a trajectory identifies a candidate word and a candidate direction, and the candidate is accepted, qualified, or rejected only after inspecting the sense distribution and the attested usages behind it. None of the reported patterns rests on a dimension score alone.

Valence. Valence measures whether a target word is used in more positive or negative contexts. We compute this dimension with a Japanese sentiment polarity dictionary (Kobayashi et al., 2005), excluding the target word itself from the context. This score is therefore interpreted as the evaluative environment of the target usage, not as a direct annotation of the inherent polarity of the entry. Let $\text{val}(x)$ be the sentiment polarity score of a context word x . For each usage u , let $C(u)$ be the set of context words excluding the target word. Context words without a valid polarity score will not be included in later computations. The usage-level

⁵Details of annotation cost are in Appendix A.6.

valence score is:

$$\text{val}(u) = \frac{1}{|C(u)|} \sum_{x \in C(u)} \text{val}(x).$$

The word-level valence score is computed directly by averaging the usage-level valence scores of all usages of w in period t :

$$\text{Val}_t(w) = \frac{1}{|\mathcal{U}_{t,w}|} \sum_{u \in \mathcal{U}_{t,w}} \text{val}(u).$$

An increase in $\text{Val}_t(w)$ marks an amelioration candidate and a decrease marks a pejoration candidate.

Semantic Scope. Semantic scope measures whether a target word is used in more general or more specific senses. We use the term semantic scope to refer to relative semantic specificity/generality operationalized through WordNet depth. We compute this dimension with Japanese WordNet (Bond et al., 2009). Because our sense inventory is based on the Iwanami Japanese Dictionary, we first align Iwanami senses to Japanese WordNet synsets. We use SARASHINA-EMBEDDING-v2-1B⁶ to encode dictionary definitions and examples for this alignment. The semantic scope analysis is restricted to the 454 dictionary senses that can be aligned to Japanese WordNet. If none of a word’s senses can be aligned, it is assigned a score of 0. We calculate the specificity of a sense by the depth of its aligned Japanese WordNet synset, namely the distance from the synset to the WordNet root. A larger depth indicates that the sense is located deeper in the hierarchy and is therefore more specific, while a smaller depth indicates a more general sense. This operationalization is approximate because WordNet depth is a taxonomic proxy rather than a manually validated semantic scope annotation. For each dictionary sense s_i , we define $\text{scope}(s_i)$ as its depth in this lexical semantic hierarchy. The word-level semantic scope score is:

$$\text{Scope}_t(w) = \sum_i p_t(s_i | w) \cdot \text{scope}(s_i).$$

An increase in $\text{Scope}_t(w)$ suggests specialization, while a decrease suggests generalization.

Metaphoricity. Metaphoricity measures whether a target word is used metaphorically. We compute this dimension with a Japanese metaphor detection

model (Zhu et al., 2025). The model takes a usage context with the target word marked and outputs a binary metaphor label for the target usage:

$$m(u) = \begin{cases} 1, & \text{if } u \text{ is metaphorical,} \\ 0, & \text{otherwise.} \end{cases}$$

The word-level metaphoricity score is computed directly as the proportion of metaphorical usages of w in period t :

$$\text{Meta}_t(w) = \frac{1}{|\mathcal{U}_{t,w}|} \sum_{u \in \mathcal{U}_{t,w}} m(u).$$

An increase in $\text{Meta}_t(w)$ marks a metaphorization candidate. Because the detector is trained on contemporary Japanese, we describe $\text{Meta}_t(w)$ as model-detected metaphoricity rather than as metaphoricity in the linguistic sense: a meaning that arose metaphorically but has become fully conventionalized is likely to be labeled literal, so the measure may underestimate historical metaphorical usage and lexicalized extensions. A flat or falling trajectory is accordingly weak evidence against metaphorization, while a rising trajectory that coincides with a shift toward an abstract dictionary sense is the configuration we treat as informative.

5 Results and Discussion

5.1 Reliability of Dictionary-Grounded Sense Distributions

Intrinsic Evaluation. The LLM-based WSD produces reliable dictionary-grounded sense labels for aggregate diachronic analysis. In Appendix Table 11, human annotations show high agreement, with mean pairwise Cohen’s $\kappa = 0.798$ and mean word-level Krippendorff’s $\alpha = 0.689$. As shown in Table 3, the model reaches 0.793 accuracy against individual annotators and 0.815 against majority vote, outperforming the two trivial baselines and the three retrieval baselines by a clear margin. The weakest retrieval baseline is the one that predicts a Japanese WordNet synset and then maps it into the Iwanami inventory. Its low accuracy is consistent with the coverage and granularity mismatch between the two inventories that also limits our semantic scope dimension.

Corpus-wise majority-vote accuracy is also stable: 0.803 on CHJ, 0.813 on SHC, and 0.844 on BCCWJ. The lack of a strong accuracy drop on CHJ suggests that the WSD module is usable for constructing aggregate sense distributions over historical as well as contemporary usages.

⁶<https://huggingface.co/sbintuitions/sarashina-embedding-v2-1b>

Method	Mean annot.	Majority vote
Random	0.281	0.282
MFS	0.695	0.699
WordNet BERT $\rightarrow$ Iwanami	0.476	0.435
Target token-gloss BERT	0.577	0.591
Usage-gloss embedding	0.663	0.681
Qwen3-8B (ours)	0.793	0.815

Table 3: Intrinsic evaluation of dictionary-grounded WSD against human annotations. Mean annot. is the macro average over the three individual annotators; majority vote accuracy is computed only for the 961 usages on which at least two annotators agree. Random is the expected accuracy of sampling uniformly from the candidate senses of each entry, and MFS predicts the most frequent human-labeled sense of each entry. The three retrieval baselines are described in Appendix A.3.

Coverage of the Sense Inventory. Sense 0 is rare overall but concentrated in a few entries. The model assigns Sense 0 to 3,037 of 213,650 usages (1.42%), and the three annotators use it for 3.18%, 4.82%, and 6.66% of the 976 annotated usages. Two entries account for 1,196 of the 3,037 model predictions and illustrate the two underlying causes. For *い* $\succ$ *in*, 232 of 249 usages receive Sense 0: the corpus usages mostly refer to inner clothing or to the borrowed English preposition, while the Iwanami entry lists only sports-related senses, so this is a genuine coverage gap. For *北* *kita*, 964 of 1,409 usages receive Sense 0 even though the ordinary ‘north’ sense is listed; the model repeatedly misreads a parenthetical gloss that names the opposite direction and rejects the correct sense, so this is a disambiguation error rather than a missing sense. No remaining entry exceeds 11%, and the pooled rate over the other 148 entries is 0.87%. A full breakdown is provided in Appendix A.7.

Extrinsic Evaluation with JaSemChange. We use JaSemChange as an external LSCD benchmark. For each target entry, we compute the Jensen–Shannon divergence (JSD) between its period-specific sense distributions and correlate the resulting change scores with the human semantic change scores using Spearman’s rank correlation.

As shown in Table 4, correlations are stronger on the seven-entry polysemous subset than on the full benchmark. The lower full-set correlations highlight a limitation of fixed sense inventories: when the dictionary provides insufficient sense distinctions, sense-distribution divergence has limited ability to represent graded semantic change. Further

Setting	Corpus pair	Spearman’s ρ	p
Overlap (7)	SHC–BCCWJ	0.821	0.023
	CHJ–BCCWJ	0.750	0.052
Full (20)	SHC–BCCWJ	0.580	0.007
	CHJ–BCCWJ	0.367	0.112

Table 4: Correlation between sense-distribution change scores and JaSemChange human judgments. Overlap covers the seven JaSemChange entries that are also among our 150 target entries and that Iwanami treats as polysemous; Full covers all 20 JaSemChange target words, including those Iwanami treats as monosemous.

analysis is provided in Appendix A.8.

5.2 Exploratory Trend Discovery

To reduce instability caused by sparse temporal coverage, we restrict the trend analysis to words with usages observed in at least ten decades. Among the 150 processed target entries, 130 have usages in at least ten decades and are used for slope-based trend discovery. For each word w , dimension d , and decade t , we fit a simple linear regression $y_{w,d,t} = \beta_{0,w,d} + \beta_{1,w,d}t + \epsilon_t$. The slope $\beta_{1,w,d}$ is used as an interpretable trend statistic. A positive valence slope suggests an amelioration candidate, while a negative valence slope suggests a pejoration candidate. For semantic scope, a positive WordNet-depth slope suggests specialization and a negative slope suggests generalization. For metaphoricity, a positive slope suggests metaphorization, while a negative slope suggests decreasing metaphorical usage. The slope ranking is used for candidate discovery, not as final proof of semantic change; final interpretation still requires sense-distribution inspection and representative corpus usages. Appendix A.9 reports robustness checks using Theil–Sen slopes and Mann–Kendall tests.

Overall Trend Analysis. Table 5 lists the strongest positive and negative candidates for each dimension. Appendix A.9 provides the corresponding slope and R^2 values. Across the 130 words with sufficient temporal coverage, the three semantic dimensions reveal complementary types of diachronic change. Valence slopes capture changes in evaluative environments, semantic scope slopes capture movement toward more general or more specific WordNet-aligned senses, and metaphoricity slopes capture abstract extensions that may not change hierarchical depth. These rankings are used as a discovery mechanism: they identify candidate words and candidate directions, while sense dis-

Dimension	Pattern candidate	Top 3 entries
Valence (+)	Amelioration	<i>genchi</i> (site/local place; current location); <i>youki</i> (cheerful disposition; weather); <i>zen'ei</i> (avant-garde pioneers; vanguard/front guard)
Valence (−)	Pejoration	<i>keesu</i> (case/instance; box/container); <i>karyuu</i> (downstream/lower river; lower class); <i>kaado</i> (card; credit/cash card)
Semantic scope (+)	Specialization	<i>men</i> (aspect/side; face); <i>tokui</i> (skilled at; proud/confident); <i>ashimoto</i> (immediate surroundings; local situation)
Semantic scope (−)	Generalization	<i>purasu</i> (addition/plus sign; advantage/profit); <i>hinshu</i> (variety/breed; item type); <i>en</i> (relationship/tie; edge/border)
Metaphoricity (+)	Metaphorization	<i>kaiten</i> (rotation; figurative turnover); <i>purasu</i> (addition/plus sign; advantage/profit); <i>dohyou</i> (sumo ring; arena for debate)
Metaphoricity (−)	Less metaphorical	<i>youki</i> (cheerful disposition; weather); <i>tengoku</i> (paradise; heaven); <i>tema</i> (effort/time; labor fee)

Table 5: Top slope-ranked semantic change candidates among the 130 target entries with at least ten observed decades. + and − mean the direction on the dimension. Detailed slope and R^2 values are reported in Appendix A.9.

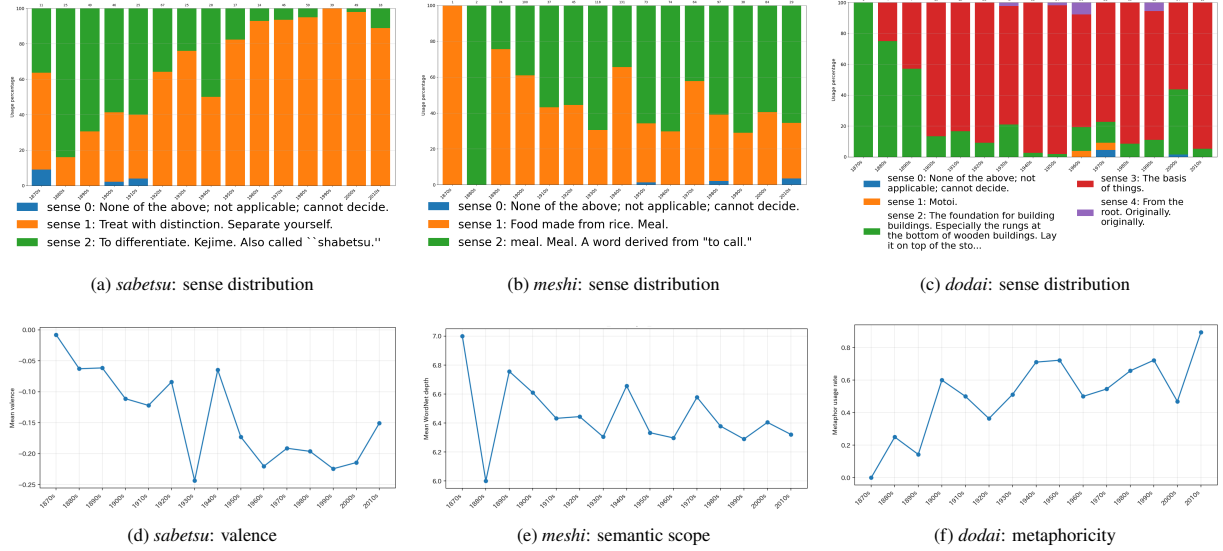


Figure 2: Case-study plots for three representative target words. The first row shows sense distributions, and the second row shows the corresponding semantic-dimension trajectories: valence for *sabetsu*, WordNet-depth semantic scope for *meshi*, and metaphoricity for *dodai*.

tributions and representative usages are needed to assess whether the statistical trend corresponds to an interpretable semantic shift.

5.3 Interpretable Semantic Change Analyses

We examine three representative lexical items by directly comparing their decade-level sense distributions with the corresponding semantic-dimension trajectories. These analyses illustrate how the same sense-distribution representation supports different semantic change interpretations depending on the selected dimension. Table 16 in Appendix A.10 reports all three dimensions for these three entries; each analysis below foregrounds the one most directly related to the illustrated pattern.

Pejoration in *sabetsu* ‘distinction/discrimination’. *Sabetsu* illustrates pejoration. Figure 2a

shows a shift from the relatively neutral sense 2, ‘distinction’, toward sense 1, ‘discriminatory treatment’. In the 1880s and 1890s, sense 2 accounts for most usages, reaching 84.0% and 69.4%, respectively. The distribution begins to change in the 1920s, when sense 1 becomes dominant at 64.2%; after the 1950s, sense 1 generally remains above 80%, reaching 100% in the 1990s and 98.0% in the 2000s. The valence trajectory supports this interpretation: the slope is negative (−0.0124 per decade), and Figure 2d shows that values after the 1950s are mostly substantially more negative than those in the nineteenth-century decades. The representative usages in Table 6 clarify the contrast. The 1957 example uses *sabetsu* in the sense of a neutral distinction between types of attack, whereas the 2001 example places the word in a context of prejudice and freedom, where it denotes social dis-

Target	Time	Sense	Depth	Usage and translation
<i>sabetsu</i>	1957	sense 2: distinction	7	JA: 正攻・奇襲の区別なく、また戦略攻撃と戦術攻撃の [[差別]] なく。 EN: “With no [[distinction]] between frontal and surprise attacks, nor between strategic and tactical attacks.”
<i>sabetsu</i>	2001	sense 1: discriminatory treatment	6	JA: アメリカにおける偏見と [[差別]] から「自由」を勝ち取るための戦い。 EN: “A struggle to win freedom from prejudice and [[discrimination]] in America.”
<i>meshi</i>	1973	sense 1: cooked rice	7	JA: 米と押麦を半々に混ぜた [[飯]] と屑じゃが芋。 EN: “Rice cooked with equal parts [[rice]] and pressed barley, and scraps of potatoes.”
<i>meshi</i>	2001	sense 2: meal	6	JA: 映画観終わって [[メシ]] 食べると、さっきまで泣いてたくせに。 EN: “After watching the movie, when we ate a [[meal]], even though he had just been crying.”
<i>dodai</i>	1925	sense 2: concrete foundation	8	JA: 木造の家屋は其 [[土臺]] に用ゐた用材を取り換へねばならぬ。 EN: “For a wooden house, the timber used in its [[foundation]] must be replaced.”
<i>dodai</i>	2002	sense 3: abstract basis	8	JA: ホテルマンとしての [[土台]] を築く時期だった。 EN: “It was a period for building his [[foundation]] as a hotel worker.”

Table 6: Representative usages of the main changing senses for each target word in case studies.

crimination. Thus, the negative valence trend is consistent with the increasing dominance of the discriminatory-treatment sense.

Generalization in *meshi* ‘rice/meal’. *Meshi* illustrates generalization, but its trajectory is not monotonic. The sense-distribution in Figure 2b shows a long-term competition between sense 1, ‘cooked rice’, and sense 2, ‘meal’. Sense 2 becomes dominant in many later decades, although the trajectory is not monotonic. Since sense 1 has depth 7 and sense 2 has depth 6, the increasing prominence of the meal sense corresponds to a decrease in semantic depth. Figure 2e shows the semantic scope slope is negative (-0.0193). This case should therefore be described as a distribution-supported broadening rather than a smooth linear drift. The representative usages show the semantic contrast: the 1973 example refers to cooked rice mixed with barley, while the 2001 example refers to eating a meal after watching a movie. The latter usage is broader because it refers to a meal generally rather than specifically to cooked rice.

Metaphorization in *dodai* ‘foundation/basis’. *Dodai* illustrates metaphorization. Figure 2c shows a clear transition from sense 2, the concrete foundation of a building, to sense 3, the abstract basis of a thing or activity. Sense 2 dominates the earliest observed decades, accounting for all observed usages in the 1870s, 75.0% in the 1880s, and 57.1% in the 1890s. From the 1900s onward, however, sense 3 becomes the main sense: it accounts for 86.7% in the 1900s, 90.9% in the 1920s, 97.4% in

the 1940s, and 94.7% in the 2010s. In Figure 2f, the metaphoricity trajectory rises from near zero in the earliest decades and has a positive slope of 0.0404. This change is not captured by semantic depth, because both senses have depth 8. The later abstract-basis usage retains the source-domain schema of support and construction but maps it onto an abstract target domain.

6 Conclusion

We introduced LSC-Pattern, a framework for making lexical semantic change analysis more inspectable by combining dictionary-grounded sense distributions with semantic dimensions. It maps diachronic usages to explicit dictionary senses and analyzes their distributions with dimensions selected for the target change pattern.

Applied to Japanese diachronic corpora, LSC-Pattern shows that LLM-induced sense distributions can align with human judgments and support interpretable analysis of multiple semantic change patterns. Case studies further show how diachronic sense redistribution, semantic-dimension trajectories, and representative usages can be combined to interpret candidate semantic changes over roughly 150 years of Japanese data.

These results suggest that linking semantic trajectories to explicit dictionary senses can make diachronic semantic change analysis more inspectable and explanatory. The framework is extensible to other languages and analysis settings.

Limitations

The framework depends on the coverage and granularity of the dictionary sense inventory. The Iwanami Japanese Dictionary is a general-purpose dictionary, not an exhaustive historical inventory, so it may omit historical, obsolete, or emerging senses, and its sense boundaries need not match corpus usage. This limitation is especially important for diachronic analysis because older texts may contain senses, spellings, or constructions that are not fully represented in a contemporary dictionary inventory.

Errors in LLM-based WSD can affect the resulting sense distributions and dimension trajectories. Historical Japanese poses additional challenges because orthography, genre, and context length differ across periods. The intrinsic evaluation shows substantial agreement with human annotations, but low-agreement words show that dictionary-grounded WSD remains sensitive to sense granularity.

The evaluation remains limited in scale and in comparison coverage. We compare the disambiguation module against five baselines, but we do not compare LSC-Pattern as a whole against embedding-based or clustering-based change-detection systems, and the external benchmark is a single 20-word Japanese dataset. The results should therefore be read as evidence that a sense-grounded interpretability framework is workable for Japanese, not as evidence that it outperforms existing LSCD systems. Future work should compare against embedding-based, clustering-based and direct usage-level dimension baselines.

The semantic dimensions are resource-based operationalizations and should not be read as gold annotations of semantic properties. The valence score measures the sentiment polarity of the target’s surrounding context after excluding the target word; this can reflect discourse topic or social context rather than the lemma’s inherent evaluative meaning. The metaphor model produces a binary usage-level label and may generalize imperfectly to historical genres and orthographic variants; moreover, conventionalized metaphorical extensions may be labeled literal by a contemporary detector. These limitations affect the interpretation of dimension trajectories, so we use sense distributions and representative usages to support, qualify, or reject dimension-based candidates.

The semantic scope dimension depends on the quality of alignment between Iwanami senses and

Japanese WordNet. Only a subset of the dictionary senses can be aligned to WordNet, and WordNet depth may not always reflect semantic specificity perfectly. Consequently, semantic scope results should be interpreted as a useful but approximate operationalization of generalization and specialization.

Diachronic trends may be affected by differences among CHJ, SHC, and BCCWJ. The corpora differ in period coverage, genre composition, annotation pipeline, and orthographic conventions. The 2000s also contain usages from both SHC and BCCWJ, which improves coverage but introduces corpus overlap that should be handled carefully in future controlled comparisons.

This study uses three dimensions, but other semantic change patterns may require other dimensions. The framework is designed to be flexible, and future work can add dimensions such as intensity, concreteness, register, social meaning, or agency.

Acknowledgments

This work was supported by JST PRESTO, Japan, Grant Number JPMJPR2366, JST SPRING, Grant Number JPMJSP2174, and by JSPS KAKENHI Grant Number JP25H00470.

ChatGPT was used for language editing, table formatting, and manuscript organization, and Codex was used to assist with experimental coding. All AI-assisted outputs were reviewed, verified, and edited by the authors, who take full responsibility for the final content.

References

- Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. [A multidimensional framework for evaluating lexical semantic change with social science applications](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1390–1415, Bangkok, Thailand. Association for Computational Linguistics.
- Andreas Blank. 1999. [Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change](#). In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, pages 61–90. De Gruyter Mouton, Berlin, Boston.
- Andreas Blank and Peter Koch, editors. 1999. *Historical Semantics and Cognition*. Cognitive Linguistics Research. Mouton de Gruyter, Berlin and New York.
- Terra Blevins and Luke Zettlemoyer. 2020. [Moving down the long tail of word sense disambiguation with gloss informed bi-encoders](#). In *Proceedings of the*

- 58th Annual Meeting of the Association for Computational Linguistics, pages 1006–1017, Online. Association for Computational Linguistics.
- Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. [Enhancing the Japanese WordNet](#). In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 1–8, Suntec, Singapore. Association for Computational Linguistics.
- Pierluigi Cassotti, Naomi Baes, Stefano De Pascale, Jäder Martins Camboim de Sá, Francesco Periti, Nick Haslam, Dirk Geeraerts, and Nina Tahmasebi. 2026. [SenseRel: A sense-level benchmark for denotational and connotational meaning relations](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 499–515, San Diego, California, United States. Association for Computational Linguistics.
- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024. [Using synchronic definitions and semantic relations to classify semantic change types](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand. Association for Computational Linguistics.
- Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2026. [Survey in characterizing semantic change](#). *Natural Language Processing*, 32(3):233 – 266.
- Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. [Definition generation for lexical semantic change detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [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.
- Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. [GlossBERT: BERT for word sense disambiguation with gloss knowledge](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.
- Päivi Juvonen and Maria Koptjevskaja-Tamm, editors. 2016. *The Lexical Typology of Semantic Shifts*. De Gruyter Mouton, Berlin, Boston.
- Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2005. Collecting evaluative expressions for opinion extraction. In *Natural Language Processing – IJCNLP 2004*, pages 596–605, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Peter Koch. 2016. Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, *The Lexical Typology of Semantic Shifts*, pages 21–66. De Gruyter Mouton, Berlin and Boston.
- Denis Kokosinskii, Mikhail Kuklin, and Nikolay Arefyev. 2024. [Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling](#). In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 168–179, Bangkok, Thailand. Association for Computational Linguistics.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. [Contextualized embeddings for semantic change detection: Lessons learned](#). *Northern European Journal of Language Technology*, 8.
- Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. [Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics.
- Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. 2023. [Construction of evaluation dataset for Japanese lexical semantic change detection](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 125–136, Hong Kong, China. Association for Computational Linguistics.
- 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](#). *Lang. Resour. Eval.*, 48(2):345 – 371.
- Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavallo, and Roberto Navigli. 2025. [Do large language models understand word senses?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.
- National Institute for Japanese Language and Linguistics. 2026. [Corpus of Historical Japanese \(version 2026.3, chunagon version 2.7.2\)](#). Accessed: 2026-05-25.
- Toshinobu Ogiso, Asuko Kondo, Yuta Takahashi, and Yoko Mabuchi. 2024. [Design, Construction, and](#)

- Publication of the Showa-Heisei Corpus of Written Japanese. *IPSJ Journal*, 65(2):278–291.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Maxim Rachinskiy and Nikolay Arefyev. 2022. [Gloss-Reader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. [Word sense disambiguation: A unified evaluation framework and empirical comparison](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. 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, 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.
- Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldbberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. [The DUREl annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.
- Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. [Can LLMs assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation](#). In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.
- Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2025. [Prompt balance matters: Understanding how imbalanced few-shot learning affects multilingual sense disambiguation in LLMs](#). In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 7–15, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.
- Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. [Can word sense distribution detect semantic changes of words?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.
- Hang Zhu, Rowan Hall Maudslay, Kanako Komiya, Sachi Kato, and Masayuki Asahara. 2025. [Large-scale Japanese metaphor corpus construction: Expanding BCCWJ-Metaphor with automated annotation](#). In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 190–199, Hanoi, Vietnam. Association for Computational Linguistics.

A Additional Experimental Details

A.1 Inference Parameters

Table 7 lists the decoding and hardware settings used for all WSD inference. Every usage was processed independently under the identical prompt and decoding configuration, in a single run.

Parameter	Value
Model	Qwen3-8B
Batch size	64
Maximum tokens	16384
Temperature	0.0
Thinking mode	Enabled
GPU memory utilization	0.92
GPU number	4
GPU	NVIDIA RTX PRO 6000 Blackwell Max-Q Worksta- tion Edition
VRAM per GPU	96GB
Inference time	10h 22m 6s

Table 7: Inference parameters used for dictionary-grounded WSD generation.

A.2 WSD Prompt and Output Parsing

The WSD prompt asks the model to place the selected Iwanami sense index in square brackets at the beginning of the answer and then provide a short explanation. For quantitative analysis, we use only the parsed sense index stored in the prediction record. The explanation text is ignored when computing WSD accuracy, sense distributions, JaSemChange correlations, semantic-dimension scores, and slope rankings; it is used only for manual inspection and qualitative error analysis. Sense index 0 is reserved for cases where none of the listed dictionary senses is suitable, the context is insufficient, or the usage is ambiguous.

After inference, the generated response is first stripped of leading and trailing whitespace. If the response begins with a complete `<think>...</think>` block, this reasoning block is removed before parsing. The system then searches the remaining text for the first integer enclosed in square brackets, such as [1] or [2]. This integer is interpreted as the predicted sense index.

The predicted index is matched against the `sense_index` values in the Iwanami dictionary senses for the target word. When a matching candidate is found, its WordNet synset identifier is recorded as the predicted synset. If the response does not contain a valid bracketed integer, the prediction is marked as unresolved. For each usage

instance, the system stores the predicted sense index, the corresponding synset identifier, and the raw model response for subsequent analysis and error inspection.

A.3 WSD Baselines

All baselines are evaluated against the same human annotations, including the individual annotator labels and majority-vote labels. The **usage-gloss embedding** baseline encodes each usage and each Iwanami definition together with its examples using `Sarashina-Embedding-v2` and selects the candidate with the highest cosine similarity. The **target-token-gloss** baseline uses `tohoku-nlp/bert-base-japanese-v3` and compares the contextualized representation of the target token with the mean token representation of each gloss. The **WordNet BERT** baseline predicts a Japanese WordNet synset with a fine-tuned `tohoku-nlp/bert-base-japanese-v3` and maps the prediction into the Iwanami inventory using the alignment described in Section 4.2; its low accuracy reflects the coverage and granularity mismatch between the two inventories. **MFS** predicts, for each entry, the sense most frequently assigned by the human annotators, which makes it an optimistic baseline because it uses the evaluation labels themselves. **Random** is the expected accuracy of sampling uniformly from the candidate senses of each entry.

A.4 Annotation Guideline

Table 10 shows the guideline given to the three annotators, in the original Japanese and in English translation.

A.5 Inter-Annotator Agreement of the WSD Evaluation Dataset

The inter-annotator agreement statistics are computed from the 976 manually annotated WSD usages used in the intrinsic evaluation. Each usage has three independent human labels, each corresponding to an Iwanami dictionary sense index for the target word. Because sense inventories are entry-specific, we compute Krippendorff’s α separately for each target entry and report the mean across the 50 entries. Pairwise Cohen’s κ is computed for each annotator pair using entry-sense pairs as nominal labels over the full set of 976 usages, and we report the mean across the three annotator pairs.

#タスク説明# 1つの用例とその中にある対象単語を与えます。次に、その対象単語の辞書から語義項目を与えます。用例の中で角括弧に囲まれた対象単語の意味が次のどれの語義に当てはまるかを答えてください。 ##入力例## 【用例】 東京大学理学部 [[教授]] 有機物ノ合成 松井直吉 化学者物体ヲ分析シテ僅ニ六十有餘ノ単体ヲ發見セリ 【語義】 [0] どちらでもない。判断できない。曖昧性がある。 [1] 品詞：名-スル 語義：学問や技芸を教え授けること。例文：「商学を教授する」 [2] 品詞：名-スル 語義：児童・生徒・学生に知識・技能を授け、その心意作用の発達を助けること。例文：「学校で数学を教授する」；「英語教授法」 [3] 品詞：名-スル 語義：大学や高等専門学校・旧制高等学校などで、研究・教育職階の最高位。また、その人。例文：「東京大学の教授に就任する」；「教授会」；「名誉教授」；「助教授」 【答え】 ##出力例## [3] «意味の説明»... 出力例では、まず [] 内に語義項目番号をまず答えてください。そして、用例の文脈を考慮して、この語義項目を選んだ理由を教えてください。 それでは、次の用例は与えられた語義項目のどれにあたるか、答えてください。 ##入力## 【用例】 {usage} 【語義】 [0] どちらでもない。以下の項目にはふさわしいものがない。決定できない。 {senses} 【答え】	
---	--

Table 8: WSD Prompt Template in Japanese

A.6 Annotation Cost

Annotators were paid 4,000 JPY per hour. The reported working times were 8.5, 10, and 9 hours, respectively. The total annotation cost was 110,000 JPY, approximately 690.8 USD at the exchange rate on May 25, 2026.

A.7 Sense 0 Statistics

Table 12 breaks down the Sense-0 predictions summarized in Section 5.1 across the full set of 213,650 usages.

A.8 Additional JaSemChange Analysis

Effect of Sense-Inventory Granularity. We report two JaSemChange evaluation settings in the main text. The first contains the seven JaSem-

Change entries that are also included in our 150-entry analysis set and that Iwanami represents as polysemous. The second contains all 20 JaSemChange target words, to which we apply the same dictionary-grounded WSD and sense-distribution pipeline.

The full-set correlations are lower than those of the seven-entry subset. This difference highlights a structural limitation of dictionary-grounded change detection. Several JaSemChange targets are represented by only one sense in the Iwanami inventory. For such entries, the representation cannot distinguish changes that occur within that single listed sense, and multiple targets may consequently receive tied or near-identical sense-distribution change scores. This reduces the resolution of the

Task Description

You are given a sentence (usage example) containing a target word, along with a list of dictionary sense definitions for that word.

Your task is to determine which sense best matches the meaning of the target word in the given context.

Input Example

[Sentence]

After work, she went to the [[bank]] to deposit her paycheck.

[Senses]

[0] None of the above / cannot determine / ambiguous.

[1] POS: n ; Meaning: the land alongside or sloping down to a river or lake. ; Example: "They sat on the river bank."

[2] POS: n ; Meaning: a financial institution where people deposit and withdraw money. ; Example: "He opened an account at the bank."

[3] POS: n ; Meaning: a set or supply of something stored for future use. ; Example: "a blood bank"

[Answer]

Output Example

[2]

«Explanation»

The sentence mentions depositing a paycheck, which is an activity associated with a financial institution. Therefore, sense [2] is the correct choice.

In your output:

First, provide the sense number inside brackets [].

Then, explain your reasoning based on the context.

Input

[Sentence]

{usage}

[Senses]

[0] None of the above / cannot determine / ambiguous.

{senses}

[Answer]

Table 9: WSD Prompt Template in English

ranking even when the human benchmark assigns different degrees of semantic change. The seven-entry results therefore characterize performance when the inventory contains multiple distinctions that can participate in sense redistribution, whereas the full-set results more realistically reflect both the strengths and the coverage limitations of the fixed inventory.

Agreement with Usage-Pair Relatedness Judgments. JaSemChange provides graded relatedness judgments over usage pairs rather than gold usage-sense labels, so these annotations cannot be

used to compute WSD accuracy directly. As an additional descriptive analysis, we compare whether our WSD system assigns the same dictionary sense to the two usages in each annotated pair.

We consider two binarizations of the original four-point relatedness scale. Under the *strict* setting, only the highest relatedness score is treated as indicating the same sense. Under the *relaxed* setting, the two highest scores are treated as indicating the same sense. Across the 798 annotated pairs, agreement between the binarized human judgments and the WSD sense assignments is 30.2% under the

Japanese	English Translation
各スプレッドシートには、対象単語ごとに、岩波国語辞典に基づく語義項目を提示しております。対象は合計 50 単語で、単語ごとにシートが分かれております。また、用例数は合計 976 文。	Each spreadsheet presents the sense entries for each target word based on the <i>Iwanami Japanese Dictionary</i> . The dataset contains a total of 50 target words, with one sheet for each word. The total number of usage examples is 976 sentences.
各列には、以下の情報が含まれています。 • left_context: 対象語の前文脈 • target: 対象単語 (キー、判別対象) • right_context: 対象語の後文脈 上記 3 つを合わせたものが、 target 列に示された単語の具体的な用例となります。	Each column contains the following information: • left_context: the context preceding the target word • target: the target word, i.e., the key item to be disambiguated • right_context: the context following the target word Together, these three fields constitute the concrete usage example of the word shown in the target column.
作業内容としては、各用例における target がどの語義に該当するかをご判断いただき、対応する語義番号を「 label 」列にご記入ください。	Your task is to determine which dictionary sense the target corresponds to in each usage example, and to enter the corresponding sense number in the “ label ” column.
なお、以下の場合には「0」をご記入ください。 • どの語義にも該当しない場合 • 文脈から判断が難しい場合	Please enter “0” in the following cases: • The example does not correspond to any listed sense. • It is difficult to determine the sense from the context.

Table 10: Annotation guideline for word sense disambiguation in Japanese and English.

Agreement statistic	Value
Mean word-level Krippendorff’s α	0.689
Mean pairwise Cohen’s κ	0.798

Table 11: Inter-annotator agreement statistics for the manual WSD evaluation dataset. Krippendorff’s α is nominal.

Entry	Usages	Sense 0	Rate
イン <i>in</i>	249	232	93.17%
北 <i>kita</i>	1,409	964	68.42%
Other 148 entries	211,992	1,841	0.87%
Total	213,650	3,037	1.42%

Table 12: Sense-0 predictions over the full prediction set. The two entries with the highest rates illustrate the two causes: a genuine gap in the dictionary inventory (*in*) and a disambiguation error (*kita*). Human annotators assigned Sense 0 to 3.18%, 4.82%, and 6.66% of the 976 annotated usages.

strict criterion and 68.6% under the relaxed criterion. The substantial difference between the two settings illustrates that categorical dictionary-sense assignments are considerably coarser than graded usage-relatedness judgments. We therefore report these results only as descriptive agreement rather than as WSD accuracy. Per-pair results are reported in Table 13.

A.9 Linear Regression Slope Calculation

Linear regression slopes are computed from the decade-level metric table. For each target entry and each metric, the script first groups rows by lemma

Corpus pair	Strict	Relaxed
CHJ-BCCWJ	127/399 (31.8%)	265/399 (66.4%)
SHC-BCCWJ	114/399 (28.6%)	282/399 (70.7%)
Combined	241/798 (30.2%)	547/798 (68.6%)

Table 13: Descriptive agreement between predicted sense identity and binarized JaSemChange relatedness judgments. Strict treats only the highest relatedness score as the same sense; relaxed treats the top two scores as the same sense. These numbers are not WSD accuracy: JaSemChange annotates graded relatedness over usage pairs, not gold usage-sense labels.

and lemma reading. It parses decade labels such as 1870s into numeric decade values and divides the decade start year by 10, so one unit on the regression input corresponds to one decade. For each of the three metrics, `val_t`, `depth_t`, and `meta_t`, blank or invalid values are skipped.

For every target entry with at least two valid decade-level points, the script fits a one-variable least-squares linear regression from decade to metric value. It stores the number of valid points, the slope per decade, the absolute slope, the intercept, and R^2 . For constant trajectories, the slope is zero and R^2 is not informative.

Only entries observed in at least ten decades enter the rankings, which leaves 130 of the 150 processed entries. The excluded entries are those confined to a short span, typically because they are absent from CHJ and therefore cannot appear in more than nine decades; ranking a slope estimated from three decades next to one estimated from fifteen would

not be meaningful. Within the retained entries, the script ranks separately for each metric by absolute slope magnitude. Tables 14 and 15 report the ten largest positive and negative slopes, respectively.

Robustness of the trend statistic. To assess whether the slope-based ranking depended on the linear least-squares estimator, we recomputed each eligible trajectory using the Theil–Sen estimator and tested monotonicity with the tie-corrected Mann–Kendall test. We retained 365 word–dimension series with at least ten valid decadal observations. The rankings induced by OLS and Theil–Sen slopes were strongly correlated for valence ($\rho = 0.787$), semantic scope ($\rho = 0.839$), and metaphoricity ($\rho = 0.908$; all $p < 10^{-27}$). After Benjamini–Hochberg correction across all 365 tests, however, only 20 trajectories showed a significant monotonic trend ($q < 0.05$): two for valence, six for semantic scope, and twelve for metaphoricity. All 20 retained the same slope direction under OLS and Theil–Sen and had $R^2 \geq 0.30$. These results indicate that slope magnitude provides a reasonably robust candidate ranking, while strong evidence for smooth monotonic drift is restricted to a small subset; the remaining entries are therefore treated as candidates for qualitative inspection rather than confirmed linear changes.

A.10 All Dimensions for the Case-Study Entries

We compute every applicable dimension for every target entry and do not collapse them into a composite score, because they answer different questions and are not commensurable. Each case study in Section 5.3 foregrounds the dimension most directly related to the illustrated pattern; Table 16 reports all three.

Valence	Semantic scope	Metaphoricity
現地 (0.0538, 0.531)	面 (0.3104, 0.677)	回転 (0.0763, 0.694)
陽気 (0.0402, 0.320)	得意 (0.1024, 0.501)	プラス (0.0732, 0.555)
前衛 (0.0370, 0.653)	足元 (0.0806, 0.129)	土俵 (0.0505, 0.481)
クリーム (0.0311, 0.255)	モデル (0.0683, 0.230)	重心 (0.0471, 0.271)
天国 (0.0295, 0.329)	株 (0.0621, 0.637)	土台 (0.0404, 0.576)
裾 (0.0272, 0.544)	峠 (0.0509, 0.431)	色合い (0.0374, 0.302)
迎も (0.0197, 0.400)	年 (0.0426, 0.632)	カード (0.0354, 0.232)
棘 (0.0194, 0.066)	闇 (0.0426, 0.580)	方向 (0.0331, 0.841)
斜め (0.0190, 0.257)	教授 (0.0364, 0.400)	波 (0.0319, 0.673)
春秋 (0.0188, 0.162)	カット (0.0330, 0.156)	経済 (0.0317, 0.665)

Table 14: Ten largest positive slopes for each semantic dimension, over the 130 target entries with at least ten observed decades. Values in parentheses give the slope per decade and R^2 .

Valence	Semantic scope	Metaphoricity
ケース (-0.0328, 0.233)	プラス (-0.2347, 0.468)	陽気 (-0.0674, 0.801)
下流 (-0.0218, 0.409)	品種 (-0.1403, 0.234)	天国 (-0.0439, 0.592)
カード (-0.0182, 0.306)	縁 (-0.1062, 0.168)	手間 (-0.0383, 0.393)
品種 (-0.0140, 0.145)	ドライブ (-0.0934, 0.274)	盲目 (-0.0333, 0.199)
空気 (-0.0138, 0.174)	背景 (-0.0860, 0.303)	コース (-0.0317, 0.206)
病気 (-0.0138, 0.149)	差別 (-0.0533, 0.734)	品種 (-0.0298, 0.155)
少年 (-0.0129, 0.360)	ケース (-0.0448, 0.113)	下流 (-0.0246, 0.142)
差別 (-0.0124, 0.564)	木 (-0.0401, 0.415)	速度 (-0.0210, 0.145)
イン [†] (-0.0122, 0.063)	天国 (-0.0387, 0.095)	万能 (-0.0192, 0.108)
背景 (-0.0112, 0.454)	陽気 (-0.0371, 0.245)	奥様 (-0.0168, 0.257)

Table 15: Ten largest negative slopes for each semantic dimension, over the 130 target entries with at least ten observed decades. Values in parentheses give the slope per decade and R^2 . [†] *In* is the entry for which 93.2% of usages receive Sense 0 (Section 5.1); its valence slope is computable because valence is a usage-level measure, but the candidate cannot be interpreted because the sense distribution carries no information, which is exactly why the framework requires the verification step.

Entry	Valence	Semantic scope	Metaphoricity
<i>sabetsu</i>	-0.0124 (0.564)	-0.0533 (0.734)	0.0121 (0.478)
<i>meshi</i>	0.0123 (0.215)	-0.0193 (0.132)	0.0038 (0.018)
<i>dodai</i>	0.0079 (0.021)	constant	0.0404 (0.576)

Table 16: All three dimension slopes per decade, with R^2 , for the three case-study entries; the dimension discussed in Section 5.3 is in bold. *Sabetsu* is a candidate on two axes and both are driven by the same sense shift. For *dodai*, every aligned sense sits at WordNet depth 8, so the semantic scope score is constant across all periods and the dimension is silent about a change that metaphoricity captures.