Han2Han: A Denoising Autoencoder for Improved Korean Language Modeling and a Use Case in Tracing the Discursive Formation of Modern Korean Art

Anonymous ACL submission

Abstract

When using natural language processing for Korean historical texts, it is common for corpora to include texts in scripts belonging to several languages, leading to poor compatibility. To address this, we introduce a sequence-to-sequence denoising pre-training method with enhanced word embeddings designed to model the characteristics of Korean etymology and written scripts. Our results show a significant increase in performance on almost all Korean Language Understanding Evaluation (KLUE - *NeurIPS Datasets* (Park et al., 2021)) tasks, suggesting that the representations created by language models benefit from learning language-specific information. We then use our method in a use case to track discursive changes in the 20th century in Korean art-critical textual data, enabling the modeling of diachronic semantic change in historical Korean texts.

1 Introduction

Word embeddings represent the distributional semantic aspects of the specific textual data on which they are trained. Though transformer-based models allow for a wider set of use cases and higher performance than static embedding models, it is difficult to represent conceptual information of domain-specific corpora. This is especially true with historical Korean texts, where models are unlikely to have been pre-trained and few weights exist. From this problem as a starting point, we began to wonder how we could train a transformer-based model capable of embedding words and documents in a self-contained system, with analyzable weights that could make further meaning of the text.

This work makes two contributions: first, by proposing a novel language model architecture and pre-training approach designed specifically for the analysis of mixed-script historical Korean texts; and second, by showing that deep, corpus-based analysis can be conducted using only the hidden states of a model trained from scratch on the corpus itself. We apply this methodology to a focused study of nearly 200,000 art-related newspaper articles written between 1920 and 1999 in South Korea. This case study effectively tracks the evolution of art and culture-related terminology from colonial Korea, through modern time until the turn of the century, demonstrating the utility of our method in uncovering semantic shifts through time.

	Form
Hanja	韓國語
Hangul	한국어
Jamo	ㅎㅏㄴㄱㅜㄱ-ㅇㅓ

Table 1: The word for "Korean" (as in, the Korean language), represented in the three forms central to this study. Note that the Jamo form is simply the Hangul form decomposed into its constituent characters.

2 Background

2.1 Sino-Korean Characters (Hanja) and the Korean Characters (Hangul)

The Korean writing system, 한글, *hangul*, is an alpha-syllabic script that combines features of both alphabetic and syllabic writing systems. Each Hangul syllable block is composed of individual letters called 자모, *jamo*, which can represent either consonants or vowels. Additionally, Korean text written before the most recent decades often also includes 한자, *hanja*, which are Chinese characters used to represent Sino-Korean words. Each Hanja is represented by a single Hangul syllable, compounded to make larger units of meaning. Because of the finite number of phonemes in Hangul, a single syllable in Hangul could refer to any number of Hanja syllables with the same Korean pronunciation. Until recent decades, word meaning was conventionally disambiguated by replacing Hangul syllables with Hanja, and this is still the case with

Periods	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	Total
Number of Documents	718K	1.0M	185K	470K	758K	959K	1.1M	1.9M	7.1M
Number of Tokens	144M	245M	39M	107M	177M	248M	307M	516M	1.74B
Average # of Tokens Per Article	200	234	211	227	233	259	286	276	241
Ratio of Hangul to Hanja Terms	0.49	0.44	0.42	0.38	0.27	0.13	0.07	0.03	0.28

Table 2: Statistics of original collected newspaper corpus from *Chosun Ilbo*, *Dong-a Ilbo*, and *JoongAng*. Total number of documents and total number of tokens per decade, average number of tokens per article per decade, and average ratio of Hangul to Hanja terms per decade. Tokens are counted according to the SentencePiece tokenization scheme and are not necessarily equal to the number of words.

some academic texts, but in typical modern-day Korean, context is emphasized more to distinguish between homonyms.

2.2 Related Work

The dual-script nature of written Korean creates a need for language-specific adaptations in NLP techniques and is generally a blind spot in Korean NLP methods, although there is a handful of research that deals with textual data that uses Hanja (Yoo et al., 2019, 2022; Son et al., 2022; Yang et al., 2023). However, most of this work deals with premodern textual data, whereas the current work is intended to facilitate research of 20th-century Korean texts that use both Hanja and Hangul.

Previous research has also taken advantage of the sub-word characters in the Hangul writing system to make more performant language modeling systems (Park et al., 2018; Seonwoo et al., 2019; Cognetta et al., 2023), but there has yet to be a system that combines the etymological knowledge available from Hanja in combination with sub-word character information.

In terms of increasing the performance of language models by pre-training on historical data, Manjavacas Arevalo and Fonteyn (2021) made an interesting contribution that improved the results of a historical English-trained BERT model. We were also interested in the work by Bhattacharya and Bojar (2023) that explored the language-specific features stored in the feed-forward networks of multilingual transformer models, as its results showing the differences depending on the location of the weights aligned with our findings shown in our use case where syntactical information is stored closer to the inputs that encode sub-word information, while semantic information is stored closer to the output side of the model.

2.3 Korean Language Understanding Evaluation (KLUE) Benchmark

The Korean Language Understanding Evaluation (KLUE) benchmark is a comprehensive collection of eight natural language understanding tasks specifically designed to evaluate the performance and capabilities of Korean language models, first introduced in the NeurIPS Datasets and Benchmarks Track (Park et al., 2021). KLUE covers a diverse range of tasks including Topic Classification, Semantic Textual Similarity, Natural Language Inference, Named Entity Recognition, Relation Extraction, Dependency Parsing, Machine Reading Comprehension, and Dialogue State Tracking. The dataset is designed to be challenging and to test the generalization capabilities of Korean language models across a wide range of tasks. In our case, we used the KLUE dataset to evaluate whether etymological and character-based information could improve the performance of a Korean language model meant for use with contemporary Korean text.

3 Corpus Data

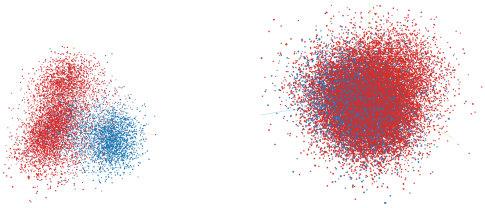
The full corpus used for this research was collected from over seven million newspaper articles written and published between 1920 and 1999 in the South Korean newspapers *Chosun Ilbo*, *Dong-a Ilbo*, and *JoongAng*. [1] All of the texts were digitized versions of the original printed material, so there were inevitably several transcription mistakes.[2] The title, author, newspaper section, publication date, and URL were all collected along with the texts. All materials are available directly on the websites

[1] All romanization of Korean terms in this paper follows the McCune-Reischauer standard, common terms with established romanization conventions omitted.

[2] For example, in the *Dong-a Ilbo* articles, the character 象 (*sang*, "form"), as in 抽象 (*ch'usang*, "abstract") was repeatedly missing, likely due to encoding inconsistencies. However, many of these errors were either consistent enough to be corrected in preprocessing or inconsistent enough to have little effect on the overall corpus.

美術展覽會 → 미술전람회 →
ㅁㅣㅅㅜㄹㅈㅓㄴㄹㅏㅁㅎㅗㅣ →
ㅁㅣ-, ㅣ-ㅅ, -ㅅㅜ, ㅅㅜㄹ, ㅜㄹㅈ,
ㄹㅈㅓ, ㅈㅓㄴ, ㅓㄴㄹ, ㄴㄹㅏ,
ㄹㅏㅁ ㅏㅁㅎ, ㅁㅎㅗ, ㅎㅗㅣ

Figure 1: Sub-word characters for the term "art exhibition," using a sliding window of three compositional characters. Note that, differently than FastText, we do not include brackets to denote prefixes and suffixes. Suffixes are already included with SentencePiece, so we omit them to reduce the size of the subword embedding matrices.



(a) Conventional token embeddings

(b) Han2Han embeddings

Figure 2: Comparison of Hangul and Hanja word embedding closeness using token-level (left) and the proposed decomposed Hangul syllable-level (right) in which red terms are Hangul, blue terms are Hanja. The embeddings in both figures were captured from the token embedding layer of the Han2Han model after training without and with subword embedding layers.

of the respective news agencies but cannot be redistributed according to South Korean law. In any case, they are fairly easily accessible and their collection can be automated. Refer to Table 2 for statistics on the original corpus.

4 Denoising Autoencoder with Embeddings Designed for Korean Writing Systems: Hangul-to-Hanja

Based on our experience trying to create better word representations of historical Korean writing using static word embeddings, we propose a new embedding method and a Hanja character prediction task to encode language-specific character-based and etymological knowledge in our model.

4.1 Subword Embeddings

Decomposed Hangul Syllable Embeddings

In the FastText algorithm, the sum of the vectors of one word's subword characters is assumed to equal the vector of the word itself (Bojanowski et al., 2017). Drawing from this, our tokenization scheme deals with Hangul syllable compositional characters. A typical convention in Korean NLP is to convert Chinese characters into Hangul before processing Korean text or to remove them altogether but our approach makes use of the extra information provided by Hanja. As demonstrated in Figure 1, the Hanja spelling (美術展覽會) for "art exhibition" is represented as the sum of its individual Hangul (미술전람회) parts.

For each entry into the vocabulary of the corpus, the text is first converted into Hangul if it is mixed script with Hanja. Each syllable character is then split into its three compositional characters, including an empty final single syllable character denoted with a hyphen as in Figure 1.

Note that the Hangul word, if also in the final dictionary, will have the same compositional character representation as the Hanja word. However, in training, the vector of the original mixed-script word is added to its Hangul compositional character vectors to ensure that the model also learns script-specific context. Refer to Figure 3 for further visualization of the token embeddings in training.

As shown in Figure 2, conventional token embeddings place Hangul and Hanja terms separately in vector space, likely because they appear in separate contexts. This is expected behavior because static embeddings represent words solely based on the words that surround them. However, because this ignores the connection between spoken and written words and between differing scripts, we decided to intervene in the embedding process to ensure that every token represented in Hanja also learns its Hangul-based representation.

When training on English data, the FastText subword embeddings are hypothesized to teach the model an understanding of prefixes and suffixes. In our case, it becomes a convenient way to ensure that even if the same word is written in different scripts, both representations will be similar.

Single Syllable Embeddings

Korean is a language that combines compounding and agglutination of free and bound morphemes to create units of word meaning. Many morphemes are composed of single-syllable units, particularly Chinese characters, so we decided also to add character-level embeddings to our word representations. A visual example of these different levels of representation can be seen in the center-left of

Figure 3, step (5), where the constituents of each of a single word's forms are combined into a single representation.

4.2 Hanja-to-Hangul Denoising Autoencoder Pre-Training

In addition to the information encoded by the additional character-based embeddings explained above, we wanted to ensure that the model could learn enhanced etymological and semantic information. Therefore, we created a simple denoising autoencoding task that requires the model to predict the original Hanja terms of a sequence based on the Hangul-transcribed terms. Following the Transformer-based Sequential Denoising Auto-Encoder scheme (Wang et al., 2021), the data is first passed through the encoder with 60% of words deleted, according to the and all of the tokens converted to Hangul. Then the original data is passed through the decoder using a causal mask, and the language modeling head predicts the original text's Chinese characters given the transcribed Hangul characters. The training process can be referenced visually in Figure 3.

(1) Shows the original text, which is first converted into Hangul in (2). The sequence is then tokenized according to the trained SentencePiece model, as shown in (3). In (4), 60% of the tokens are deleted according to the TSDAE pretraining objective. The embeddings for the original tokens, each of their syllabic characters, and their decomposed Hangul characters are summed (5) into a single vector representation before passing into the bidirectional encoder (6). A d-dimensional vector output is produced (7). In (8), the original sequence, tokenized, is shown again, but this time not transcribed into Hangul beforehand. The decoder (9) uses the encoder's output from (7) to predict the next token in the sequence (10), using causal attention. Note that step (5) occurs before both the encoder and decoder and only demonstrates the process for a single token. Through this task, the model learns to disambiguate between homonyms by predicting the correct Hanja characters based on their Hangul characters.

5 Results

We trained four different models on the collected newspaper data detailed in Section 3 for the same amount of steps before fine-tuning them on seven KLUE tasks, except for Machine Reading Comprehension, which could not converge regardless of the model used.[3]

There does not yet exist a set of evaluation data that can measure the performance of models using historical data, although this may be an appealing project for the future. Despite this, we can see how etymological knowledge encoded by our pre-training task can improve the ability of the model to create fine-grained word and token representations from its ability to disambiguate between homonyms.

5.1 Training Stability

As can be seen in Appendix A.1, the model that was trained using the proposed pre-training method and with the supplementary subword embeddings reached convergence more quickly and had overall stable training.

5.2 KLUE Evaluation Results

Using the Korean Language Understanding Evaluation (KLUE) benchmark (Park et al., 2021), we compared the performance between models trained with and without our proposed embedding techniques and with and without our proposed pre-training technique. The results can be seen in Table 3. Our proposed method outperforms the base TSDAE model on average in every task. The pre-training method also performs better overall than the base TSDAE model, but our results suggest that the sub-word character embedding supplementation is the most important factor in model performance, as the scores for Han2Han and the embeddings-only models perform higher overall than the pre-training-only model. These results suggest that language specificity is the more important factor in model performance. However, further research is necessary to explore this claim because etymological information is also encoded just by supplementing with Hangul sub-word character information.

6 Use Case: Using Han2Han to Model Diachronic Semantic Change in the 20th Century Korean Art World

A central motivation for creating a method for word embeddings designed with characteristics specific to Korean was so that we could perform a historical analysis of Korean texts that contain a signif-

[3]This may be a fault of the KLUE-baseline code, which is now several years old.

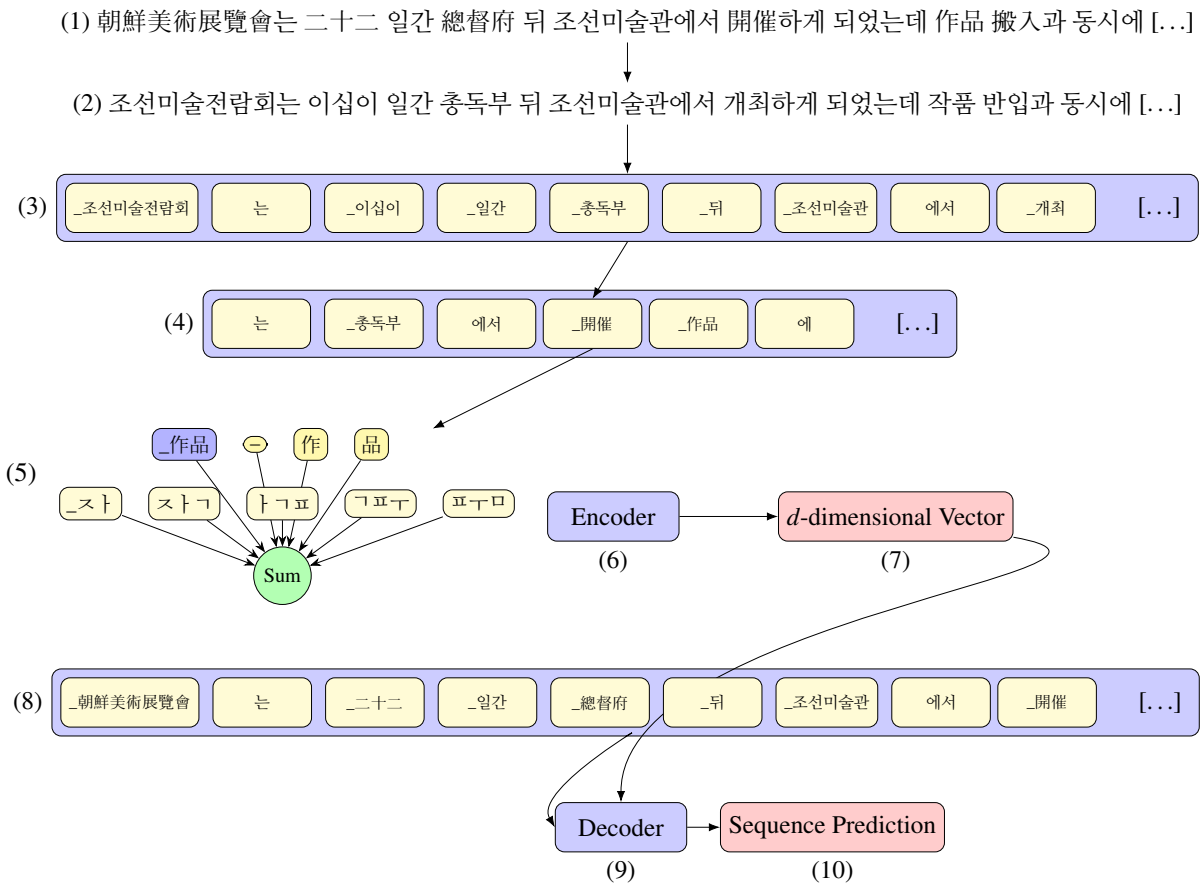


Figure 3: Visualization of the model's architectural flow, from original input sequence to next-token prediction.

icant number of Hanja characters. As illustrated in Figure 2, conventional methods using word embeddings consider Hangul and Hanja tokens to be completely different, which leads to missing information when analyzing the discourse of a Korean dataset older than a few decades. In this section, we describe our experiments involving modeling a small dataset with the proposed Han2Han methods, combined with several existing techniques, to trace discursive changes in the Korean art world in the 20th century.

6.1 Background

In reference to existing methods for detecting semantic change, it has been shown that static embedding models like Word2Vec and FastText are more effective and that the most important factor is how the embeddings are fine-tuned in a system (Schlechtweg et al., 2020). We tried creating a system using our Han2Han embeddings using static word embeddings but found that the results were difficult to interpret. We found more appeal in the ability of language models that use attention to store concepts in their feed-forward network weights (Geva et al., 2022), so we decided to look more in the direction of creating a system that uses contextual embeddings to detect semantic change. As shown in Periti and Montanelli (2024), there are several ways to approach lexical semantic change, but we were most interested in finding an unsupervised means to explore changes occurring within a specific domain, art criticism and history in this case. Details of our approach are explained in the next section. The difficulty in historical research, in South Korea's case, especially of materials of the 20th century, is that many of them have been lost to war and the tumultuous conditions of Korea as it went through colonization, liberation, authoritarian regimes, rapid economic development, and democratization through the century. However, newspapers still operated through all of these events, giving a unique opportunity to study discursive changes with our proposed method.

	Han2Han	Embeddings Only	Pre-Training Only	TSDAE
KLUE-DP	**79.59**	75.7	78.84	78.62
KLUE-MRC	—	—	—	—
KLUE-NER (Char. Macro F1)	**80.63**	79.48	70.79	71.02
KLUE-NLI	41.4	**42.03**	39.27	40.23
KLUE-RE	**32.89**	32.09	27.59	18.54
KLUE-STS (Pearson)	**65.19**	57.6	49.65	35.85
KLUE-WOS	56.87	**59.35**	NaN	NaN
KLUE-YNAT (Macro F1)	75.73	**75.76**	76.03	76.86

Table 3: Evaluation results on all tasks in the KLUE dataset, except machine reading comprehension, with the results of the model trained with both the proposed supplemental sub-word character embeddings and pre-training method (far left), the supplemental embeddings only, the pre-training method only, and the base TSDAE model. Note that our proposed method outperforms the base model in semantic textual similarity by almost double the score.

Periods	**1920s**	**1930s**	**1940s**	**1950s**	**1960s**	**1970s**	**1980s**	**1990s**	**Total**
Number of Documents	7,342	15,263	2,680	8,932	18,485	31,217	42,774	71,795	198,488
Number of Tokens	4.5M	10.4M	1.3M	4.4M	9.3M	14.6M	20.5M	31.3M	96.3M
Average # of Tokens Per Article	609	680	496	489	503	467	480	435	485
Ratio of Hangul to Hanja Terms	0.62	0.60	0.60	0.63	0.70	0.86	0.92	0.96	0.74

Table 4: Statistics of the training data used for the diachronic semantic change experiments. Total number of documents and total number of tokens per decade, average number of tokens per article per decade, and average ratio of Hangul to Hanja terms per decade.

6.2 Experimental Details

Article filtering

Of the seven million articles mentioned in Section 3, 198,488 were collected using keywords[4] related to art and culture and used as the base dataset on which to train the model. The final dataset used for tracing diachronic change was composed only of articles that contained the Hanja or Hangul word for "art," *misul* (美術 or 미술). Corpus statistics can be referenced in Table 4.

Tokenization

We aim to preserve the convenient interpretability of the word-level tokenization used in static embedding models while employing the wide range of uses of contextual word embeddings for the historical analysis of a single corpus. Existing contextualized semantic shift detection methods require using tokens belonging to pre-trained language models, many of which are slices of words and can be tricky to interpret. Because we are training our own model, we can also define our own vocabulary, which is particularly useful in domain-specific analysis when we are looking for developments in concepts and discourse.

We first used MeCab-ko, part of the KoNLPy (Park and Cho, 2014) library, to extract all nouns from each article, and then included the most frequent nouns in the vocabulary when training the unigram language model with SentencePiece (Kudo, 2018).[5] However, because the articles contained mixed scripts, we had to make some modifications to the noun extraction method.

Although MeCab-ko is capable of parsing some Hanja words, it is not uncommon to find inconsistent results between Hanja and Hangul inputs like the following:

```
(1) >>> mecab.pos("美術展覽會")
       [('美術展', 'NNP'),
        ('覽', 'SH'),
        ('會', 'NNG')]
(2) >>> mecab.pos("미술전람회")
       [('미술', 'NNG'),
       ('전람회', 'NNG')]
(3) >>> mecab.pos(translate("美術展覽會"))
       [('미술', 'NNG'),
        ('전람회', 'NNG')]
(4) >>> hanmap("美術展覽會", "미술전람회")
       [('美術', 'NNG'),
        ('展覽會', 'NNG')]
```

[4] All of the search terms can be found in Appendix A.2.

[5] SentencePiece was used in combination with noun extraction so that the model would not run into any out-of-vocabulary terms and the document and sentence embeddings would contain more fine-grained information, but also so that we could analyze nouns from the final vocabulary without having to filter through word pieces.

Figure 4: (1,2) Different tokenization of the Hanja word and Hangul word for "art exhibition." (3,4) Part of speech parsing of the Hanja word for "art exhibition" in Korean, after transcription-based mapping.

As shown above, in (1) and (2), the Hanja word 美術展覽會 (*misulchŏllamhoe*, "art exhibition") ends up sliced incorrectly as 美術展 (*misulchŏn*, "art exhibition"[6]) 覽 (*ram*, "to see"), and 會 (*hoe*, "meeting") rather than as 美術 (*misul*, "art") and 展覽會 (*chŏllamhoe*, "exhibition") like is done with the Hangul 미술 (*misul*, "art") and 전람회 (*chŏllamhoe*, "exhibition"). This is because of the limited presence of Chinese characters in the Korean MeCab dictionary. So, we define a preprocessing function, hanmap(Hanja, Hangul) to ensure consistency of spacing whether the term is represented with Hanja or with Hangul, as in (3) and (4).

Architecture Selections

For contextual word embeddings, we modify the TSDAE pretraining objective (Wang et al., 2021) and use Attention-Free Transformer (AFT) attention (Zhai et al., 2021) and relative positional embeddings (Shaw et al., 2018; Huang et al., 2020) for quick training of a transformer-based model for historical Korean corpus analysis.[7]

Model Hyperparameters and Training

Because the training is meant not for general language understanding, and instead to model the data as-is, we did not use a test split and instead ran the training process until the model reached convergence. To cover all documents, training was staggeringly fast—it only took two hours of training, which equaled approximately two passes through the entire dataset on a single GPU, until the loss curve flattened, which we owe to the speed of the employed AFT attention mechanism. We also found during our experiments that on smaller datasets, models with only two layers would still converge, and in as little as ten minutes of total training.[8]

[6]This is an abbreviated form of the term. Although not completely incorrect, it is a misrepresentation of the original word; the latter two characters would be deleted because we only include words with two or more characters.

[7]Note that, for analysis based solely on word embeddings, Han2Han embeddings can also be used by making a few simple modifications to the preprocessing steps usually used in libraries like Gensim (Rehurek and Sojka, 2011).

[8]However, the results were more difficult to analyze through the methods described in the following section, so we chose a moderately-sized model for the final use case.

Korean Term	English Term	Spearman
印象 *insang*	Impression	0.9357
美術界 *misulgye*	Art World	0.9152
鮮展 *sŏnjŏn*	Joseon Salon	0.8239
科目 *kwamok*	Subject	0.7747
社長 *sajang*	Chief	0.7599

Table 5: Top five terms shown to have the most steady semantic change between the 1920s and 1990s.

Korean Term	English Term	Cosine
印象 *insang*	Impression	-0.0541
美術界 *misulgye*	Art World	0.0847
鍾路 *chongno*[9]	*Jongno*	0.1196
公共 *konggong*	The Public	0.1361
鮮展 *sŏnjŏn*	Joseon Salon	0.1761

Table 6: Top five terms shown to have the most significant semantic change among each decade between the 1920s and 1990s. The cosine similarity shown is between the two decades in which the distance was the largest.

Methods of Analysis

We used a method inspired by Horn (2021), collecting embeddings of each term in the vocab by tracking a weighted running average of the terms over each decade, slightly modified to consider the term frequency of each term per decade. Additionally, following recent research demonstrating that latent conceptual representations are stored in the final feed-forward networks of transformer blocks (Geva et al., 2022), we projected the token embeddings onto the weights of the final feed-forward networks in each model layer, choosing the vectors most likely to predict terms found to have semantic change, and clustered the resulting logits with UMAP (McInnes et al., 2018). Then, inspired by Grootendorst (2022), we used a class-based Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme and a combination of word and sequence embeddings to choose the most salient terms for each concept cluster. Each resulting concept can be referenced in Figure 8, arranged in order of cluster size.

6.3 Results and Discussion

Significant Terms in 20th Century Korean Art

The top five terms with the most significant shift and those with the most consistent shift between 1920 and 1999 are shown in Tables 5 and 6.

[9]The first character 鍾 (*jong*, "winecup") here is said to have replaced the original character 鐘 (*jong*, "bell") during the Japanese Occupation, when several districts in Seoul were renamed to have a more Japanese style or feeling.

Semantic Changes of *Modern Art*

The Japanese colonial era led to unprecedented change on the Korean Peninsula. The term for fine art most commonly used today, 美術 (미술—*misul*), was originally conceived through Japan in 1873 as a mistranslation from the German *kunstgewerbe*, a term referring to arts and crafts, rather than fine art (Hyeshin, 1999). The following decades would see the complete institutionalization of the arts, formed through endless similar debates over the proper Chinese characters to use to adopt Western concepts to apply to culture, society, and politics (Marra, 2010). As a country whose modernist language passed through several rounds of translation,[10] it is more appropriate to interpret the meaning of words as intrinsic to the historical and cultural contexts in which they were first used (Park, 2022). This approach lends itself well to the space between linguistic and art historical research, especially when applied to semantic change.

Evolution of Associations of *the Art World*

It was found that 美術界 (*misulgye*, "the art world"), was one of the highest-changing terms over the 20th century. Using UMAP for dimensionality reduction, we visualized the terms closest to *misulgye* in each decade in Figure 6. Globally, we see a horseshoe shape as the term develops through time, with each decade gathered in its own cluster. We also note three sub-clusters, where the 1920s-1940s are grouped, the 1950s-1970s are grouped, and the 1980s-1990s are grouped.

We then use the model to create sequence embeddings for the string of terms representing each cluster. For each decade, the cosine similarities between the vector for *misulgye* and the vector for each sense of the term are plotted to investigate the representative sense of the term in each decade, as shown in Figure 8, with the concept numbers corresponding to those from Figure 8. Although a fully-fledged analysis is out of the scope of this paper, we present a simple discussion based on the concepts unique to each of the clustered periods found in Figure 6.

From the 1920s to the 1940s, the most prominent sense of the term *misulgye* is from cluster number 2, which contains terms such as "film," "school," "Italy," "international," "economy," "manufacturing," and "competitions." From these terms, we can assume that artists were thinking about the practical applications of art and the mobility they could gain from domestic and international exposure. We also see terms that suggest a positioning towards art as an industry rather than as a pastime.

In the next period from the 1950s to the 1970s, the most prominent cluster is number 6, which contains terms such as "education and culture," "East Asia," "school," "setting a visible example," and "principle." We can assume that the art world in this period would have dealt with topics related to national identity, society, and tradition, marked by a more reflective function of the arts that prompted artists to consider how to represent themselves on a global stage.

In the 1980s and 1990s, the most similar sense of "the art world" is to concept number 1, which does not contain specific terms but is still the largest cluster. The second closest sense of the term is the same as it was during the Japanese Occupation, suggesting an emphasis on the economic realities of the industry, and giving some meaning to the horseshoe shape evident in Figure 6. However, the closeness overall of the concept clusters is considerably lower in this period, suggesting that the idea of the art world was beginning to take on a new meaning of its own that could not yet be well-defined.

7 Conclusion

In this paper, we have detailed our experiments in creating a novel way to represent the Korean language in contextual embedding language models. Han2Han enriches Korean word embeddings by encoding language-specific etymological and written-letter information without considerable changes to existing architectures. Our results suggest that the stronger contribution made by our work is owed to the embedding methods, leading us to suggest that language models trained in the future include language-specific information, if not also including text belonging to historical data. Future research may detail the effects on the hidden weights of models trained with extra information, giving the opportunity to see how language models not only learn general language but how they can teach us about the linguistic qualities of the languages on which they are trained.

[10] Like with *kunstgewerbe*, or terms like *informel* several decades later, it was common for terms to come first from French or German materials, sometimes abridged, translated to Japanese, and finally once more into Korean.

8 Limitations

Primarily, we must acknowledge that the evaluation benchmarks used in this paper are not designed to measure the performance of models trained on historical data. Although benchmarks evaluating the performance of models to trace semantic change do exist for a handful of languages, there is no such benchmark for Korean. The results presented in this paper are only a demonstration of the potential of our proposed method.

The models used in this paper dealt with newspaper text only from the 20th century and only published by three news agencies. For a more general-purpose pre-trained model, a significant amount of data would need to be collected including contemporary and historical Korean text.

Additionally, our main motivation was to create a working system that could explore changes within a small and domain-specific historical corpus. Although we evaluated our model's performance on present-day textual data, the results shown would not be able to compare to a large language model unless our proposed method was employed in the pre-training process of these models.

Ethics Statement

We adhere to the ethical guidelines outlined in the ACL Code of Ethics. The datasets for this study were gathered from publicly accessible web URLs. The accompanying code and experiments will be made available on GitHub.

References

Sunit Bhattacharya and Ondřej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Marco Cognetta, Sangwhan Moon, Lawrence Wolfsonkin, and Naoaki Okazaki. 2023. Parameter-efficient Korean character-level language modeling. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2350–2356, Dubrovnik, Croatia. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Franziska Horn. 2021. Exploring word usage change with continuously evolving embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 290–297, Online. Association for Computational Linguistics.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3327–3335, Online. Association for Computational Linguistics.

Kim Hyeshin. 1999. Modern japan and japanese art. *Journal of the Association of Western Art History*, 11:187–207.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Michael Marra. 2010. Chapter two. the creation of the vocabulary of aesthetics in meiji japan. In *Essays on Japan*, Brill's Japanese Studies Library, pages 23–47. Brill Academic.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

MKDAT. Multilingual Dictionary of Korean Art Terms (MDKAT) [online]. 2018.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea.

Ji Na Park. 2022. The history of acceptance of concept on johyung : from johyung to design. *Bulletin of Korean Society of Basic Design & Art*, 23:595–612.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2429–2438, Melbourne, Australia. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), Kyunghyun Cho, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

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.

Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. 2019. Additive compositionality of word vectors. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 387–396, Hong Kong, China. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. 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 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary Korean and English. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1260–1272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *Preprint*, arXiv:2104.06979.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics.

Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. HUE: Pretrained model and dataset for understanding hanja documents of Ancient Korea. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1832–1844, Seattle, United States. Association for Computational Linguistics.

Kang Min Yoo, Taeuk Kim, and Sang-goo Lee. 2019. Don't just scratch the surface: Enhancing word representations for Korean with hanja. 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 3528–3533, Hong Kong, China. Association for Computational Linguistics.

Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. 2021. An attention free transformer. *Preprint*, arXiv:2105.14103.

A Appendix

A.1 Training Stability

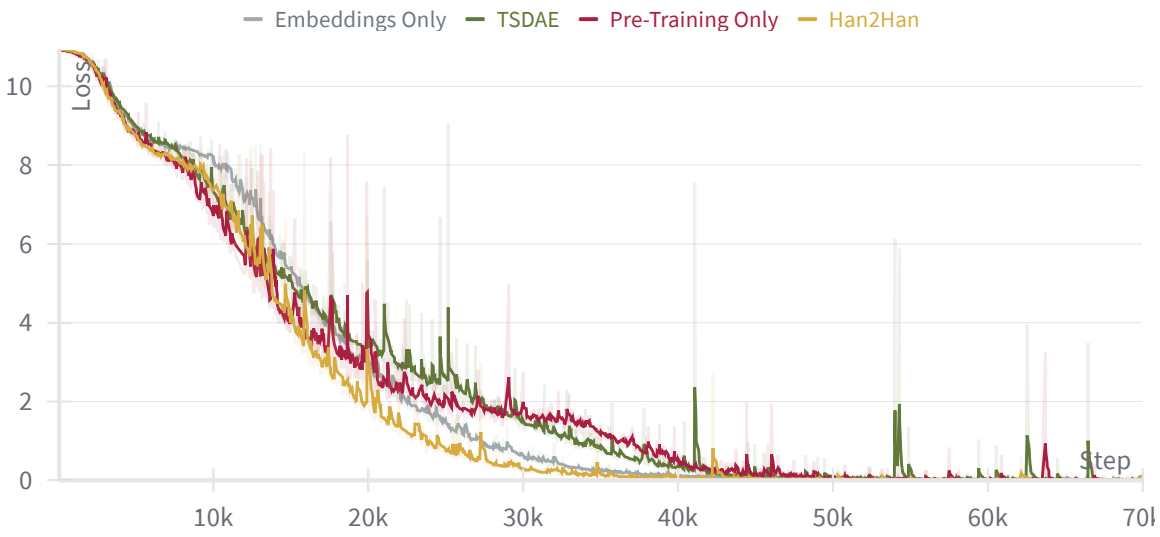


Figure 5: Training stability between the four tested models for the first 70,000 steps.

A.2 Original Search Terms

Hangul (Hanja)	English
조선미술전람회 (朝鮮美術展覽會)	Joseon Art Exhibition
대한민국미술전람회 (大韓民國美術展覽會)	National Art Exhibition
— (鮮展)	Joseon Art Exhibition (abbr.)
— (國展)	National Art Exhibition (abbr.)
— (寫生)	Sketching
— (繪畵)	Painting
— (美展)	Art Exhibition (abbr.)
미술평(美術評)	Art Review
미술 (美術)	Fine Art
예술 (藝術)	Art
추상 (抽象)	Abstraction
앵포르멜	*Informel*
단색화 (單色畵)	Dansaekhwa
비정형 (非定型)	Unformed
표현주의 (表現主義)	Expressionism
박서보(朴栖甫)	Park Seobo
권영우 (權寧禹)	Kwon Youngwoo
정창섭 (丁昌燮)	Chung Changsup
하종현 (河鍾賢)	Ha Chong Hyun
김환기 (金煥基)	Kim Whanki
김기린 (金麒麟)	Kim Gui-line
이우환 (李禹煥)	Lee Ufan
윤형근 (尹亨根)	Yun Hyongkeun
이경성 (李慶成)	Lee Kyung-sung
윤명로 (尹明老)	Youn Myeong-Ro
김창열 (金昌烈)	Kim Tschang-yeul
이구열 (李龜烈)	Lee Ku-yeul
하인두 (河麟斗)	Ha In-du
문우식 (文友植)	Moon Woosik
오광수 (金秉騋)	Oh Kwang-su
방근택 (方根澤)	Pang Kŭnt'aek

Table 7: Search terms for original corpus. All English translations and Chinese characters sourced from MKDAT (2018). Terms were only searched in Korean. Certain Hangul terms marked with "—" were also omitted to prevent redundant or irrelevant search results.[11]

[11]For example, in the case of the word 선전 (*sŏnjŏn*), the same Hangul term could refer to the Joseon Art Exhibition, or to the term for propaganda, 국전 (*kukchŏn*) could refer to the National Art Exhibition, or instead to the word for national war, 회화 (*hoehwa*) to painting or to the word for conversation, etc.

A.3 Diachronic Visualization

The following figures show the visualized word embeddings for each term closest to *misulgye*, "the art world," over the 20th century.

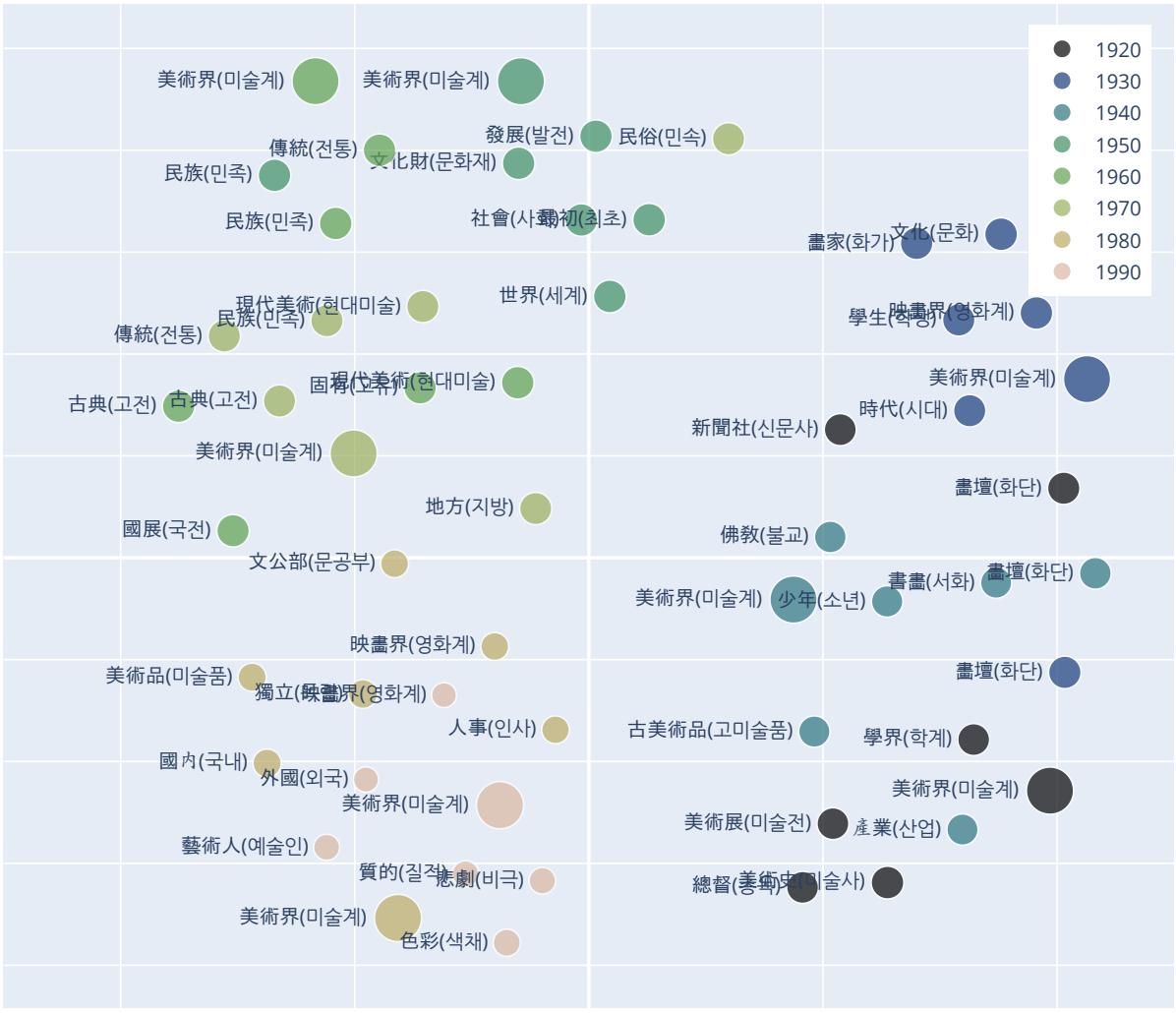


Figure 6: Closest word vectors to the word *misulgye* over the 20th century.

Figure 7: Closest word vectors to the word *misulgye*, in English, over the 20th century.

A.4 Feed-Forward Network Concept Clusters

Concept #	Most Representative Terms
1	미술계, 藝術(예술), 경북, 미술, 개최, 美術(미술), 전체, 美術界(미술계), 文化(문화)
2	東亞日報(동아일보), 映畵(영화), 學校(학교), 이탈리아, 國際(국제), 經濟(경제), 製作(제작), 美術界(미술계), 평북, 콩쿠르
3	문화, 작품, 미술, 교류, 文化(문화), 不能(불능), 生活(생활), 考慮(고려), 말썽
4	한국, 外交(외교), 사장, 연구회, 교양, 주최, 희곡, 작업, 硏究(연구)
5	主義(주의), 國有(국유), 滿洲(만주), 朝鮮(조선), 時事(시사), 김주, 경희, 단독, 中心(중심), 입시
6	文教(문교), 童謠(동요), 東亞(동아), 學校(학교), 文化(문화), 시범, 主義(주의), 배제, 組合(조합), 日曜日(일요일)

Table 8: Top ten most representative terms for each sense of the word *misulgye*, as learned by the model.

Concept #	Most Representative Terms
1	the art world, art, Kyŏngbuk, fine art, holding (an exhibition), fine art, totality, the art world, culture
2	*Dong-a Ilbo*, film, school, Italy, international, economy, production, the art world, Pyŏngbuk, competition
3	culture, artwork, fine art, exchange, culture, incompetence, livelihood, consideration, trouble
4	South Korea, diplomacy, chief, research association, refinement, host (a gathering), theater, work, research
5	principle, national property, Manchuria, Joseon, current events, Kimju, Kyŏnghŭi, being independent, the center, entrance examinations
6	education; culture, nursery rhyme, East Asia, school, culture, setting an example, principle, exclusion, an association, Sunday

Table 9: Top ten most representative terms for each sense of the word *misulgye*, in English.

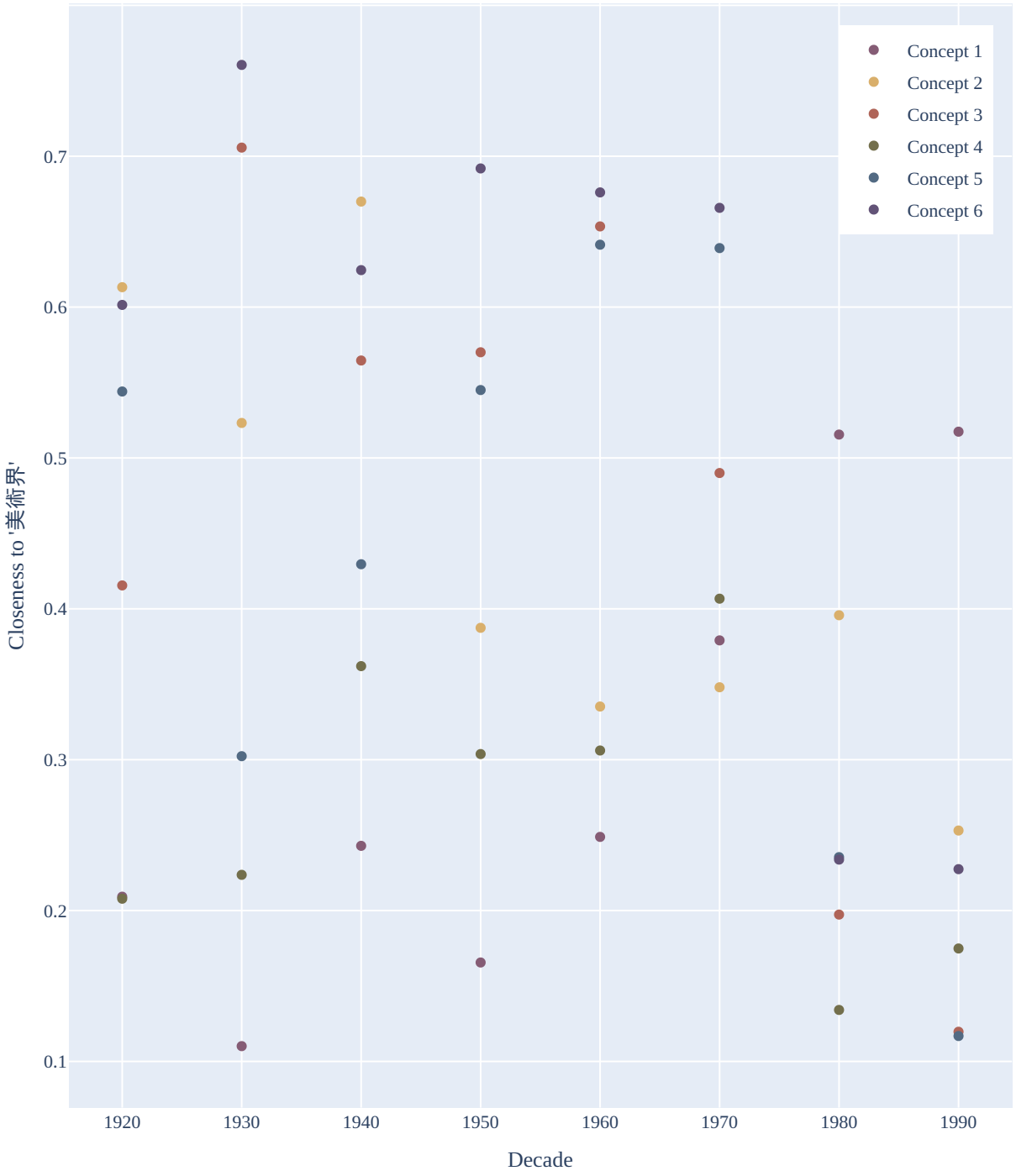


Figure 8: Varying senses of the term *misulgye* over the 20th century.