Cross-lingual and Word-Independent Methods for Quantifying Degree of Grammaticalization

Ryo Nagata
Konan University / AIST / RIKEN
nagata-acl2025 @ ml.hyogo-u.ac.jp.

Daichi Mochihashi
The Institute of Statistical Mathematics
daichi@ism.ac.jp

Misato Ido and **Yusuke Kubota**
NINJAL
{ido.misato,kubota}@ninjal.ac.jp

Naoki Otani
Tokyo University of Foreign Studies
otani@tufs.ac.jp

Yoshifumi Kawasaki
University of Tokyo
ykawasaki@g.ecc.u-tokyo.ac.jp

Hiroya Takamura
AIST
takamura.hiroya@aist.go.jp

Abstract

Grammaticalization denotes a diachronic change of the grammatical category from content words to function words. One of the intensively explored directions in this area is to quantify the degree of grammaticalization. There have been a limited number of automated methods for this task and the existing, best-performing method is heavily language- and word-dependent. In this paper, we explore three methods for quantifying the degree of grammaticalization, which are applicable to a wider variety of words and languages. The difficulty here is that training data are not available in the present task. We overcome this difficulty by using Positive-Unlabeled learning (**PU-learning**) (Elkan and Noto, 2008) or Cross-Validation-like learning (hereafter, **CV-learning**). Experiments show that the CV-learning-based method achieves middle to high correlations to human judgments in English deverbal prepositions and Japanese nouns being grammaticalized. With this method, we further explore words possibly being grammaticalized and counterexamples of the unidirectionality hypothesis.

1 Introduction

Grammaticalization (Hopper and Traugott, 2003) is a diachronic change of the grammatical category from content words to function words. An example is the present English auxiliary *can*, which used to be a main verb to mean *to know*.

One of the intensively explored research directions in this area is to quantify the degree of grammaticalization; that is, linguists have worked to reveal how far target words have been grammaticalized. It plays a central role in discovering words being grammaticalized and in investigating their degrees with related linguistic features. A good example of this is linguistic studies (Kortmann and König, 1992; Fukaya, 1997; Hayashi, 2015) targeting English deverbal prepositions such as *following* and *according to*. These linguistic studies quantify the degree of grammaticalization based on either linguistic tests or human judgments on linguistic questionnaires. Both approaches examine whether target words satisfy certain properties of typical grammar items such as prepositions. While these approaches have been shown effective, it is difficult to increase the size of investigations in terms of both the number of target words and the number of instances examined, as Hayashi (2015) already points out.

Although automated methods for quantifying the degree of grammaticalization will facilitate the research in this area, their availability is limited; to our best knowledge, there have been only three studies (De Troij and Van de Velde, 2020; Saavedra, 2021; Nagata et al., 2024). Besides the existing methods have several of the following problems as discussed in detail in Section 2: (i) they are word- and/or language-dependent; for example, one of Nagata et al. (2024)'s methods is only applicable to English deverbal prepositions; (ii) they lack a theoretical or mathematical basis and are somewhat *ad-hoc*; (iii) effective linguistic features are controversial among the studies; for example, De Troij and Van de Velde (2020) argue that recurrence time distribution is effective whereas Saavedra (2021) reports on the opposite; (iv) their performance is not appropriately evaluated; for example, Saavedra (2021) targets only typical function words (e.g., *a*) and content words (e.g., *British*) and does not use grammaticalizing words in evaluation.

In this paper, we propose classification-based

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4775–4787
March 24-29, 2026 ©2026 Association for Computational Linguistics

methods for quantifying the degree of grammaticalization to attack the above four problems. The difficulty here is that training data for classification are not available in the present task; each target word must be annotated with its degree of grammaticalization. To overcome this difficulty, we solve the problem by using Positive-Unlabeled learning (**PU-learning**) (Elkan and Noto, 2008) or a Cross-Validation-like learning (hereafter, **CV-learning**).

The contributions of this paper are three-fold: (i) we present methods for quantifying the degree of grammaticalization based on static word embeddings and simple classification techniques; they are cross-lingual and word-independent methods; (ii) we create new datasets for evaluation to alleviate the current situation that evaluation data are scarce; the datasets are available with the software programs implementing the proposed methods on our web site[1]; (iii) using the datasets, we show that recurrence time distribution and contextual diversity, which have been regarded as effective in previous studies, do not satisfy the sufficient conditions but rather necessary conditions of grammaticalization; instead, contextual similarity is effective in this task. In addition, as a pilot study, we also explore the unidirectionality hypothesis that grammaticalization progresses in the single direction, content to function words, by applying the best-performing proposed method to the cleaned version (Alatrash et al., 2020) of Corpus of Historical American English (CCOHA) (Davies, 2012), revealing words possibly being grammaticalized and counterexamples of the hypothesis.

2 Related Work

2.1 Grammaticalization in Linguistics

There have been a wide variety of linguistic studies on grammaticalization (e.g., Meillet (1912); Hopper (1991); Lehmann (1995); Hopper and Traugott (2003); Bybee (2006), to name a few). They have revealed grammaticalized items with related linguistic phenomena. Typical phenomena are:

(a) Category expansion
(b) Meaning change: bleaching or generalization
(c) Decategorization
(d) Phonetic Reduction

(a): Grammaticalizing items expand their contexts and in turn their categories. For example, *can* used to be used with a small number of infinitives, and then it started to be used with any verb (Bybee, 2003), expanding its usage as the present-day auxiliary. This implies that contextual diversity can be a good source of information about grammaticalization, which has been used in the existing methods (Saavedra, 2021; Nagata et al., 2024). (b): As expanding their contexts and categories, the lexical meanings of grammaticalizing items are bleached or generalized. For example, during the Middle English, as *can* was spreading to be used with more verbs, it lost the meaning of *mental ability*, and shifted to *general ability*, and further to more general *root possibility* (Bybee, 2003). (c): As grammaticalizing constructions lose some aspects of their meanings, they can become disconnected from instances of their lexical usages. For example, *can* was disconnected and decategorized from the main verb usage. This means that contextual similarities reflecting lexical and syntactic usages can also be a source of information about grammaticalization. (d): Grammaticalizing items often involve phonetic reduction. An example of this is *going*, which can be contracted as *gonna* as in *I'm gonna*.

Previous linguistic studies often view grammaticalization as a gradual process; the change from a content word to a function word occurs by a gradual series of individual shift. The overlapping stages of grammaticalization form a chain called a *cline* (Hopper and Traugott, 2003); a line at one end, fully lexical items, and at the other, fully grammatical items. Diachronically, clines represent a path along which words change over time. Synchronically, clines can be seen as an arrangement of words. The target task in the present study, then, can be regarded as a task to align target words along a cline. For example, *following* is being decategorized from the main verb *follow* and considered to be between main verbs and prepositions.

Linguists often deal with clines ranging over content words, function words, clitics, and inflectional affixes in this order. In this paper, however, we only focus on the first two stages, that is, the change from content words to function words as in the previous studies (Kortmann and König, 1992; Fukaya, 1997; Hayashi, 2015; De Troij and Van de Velde, 2020; Nagata et al., 2024).

The unidirectionality hypothesis is also intensively studied in linguistics. It claims that grammaticalization follows only one way from content items to grammatical items. Many linguists (e.g., Haspelmath 1999) support this hypothesis

[1] https://github.com/nagata-github/grammaticalization_tools/

while others including Campbell (2000) and Joseph (2000) provide counterexamples of this hypothesis.

2.2 Previous Methods for Quantifying Degree of Grammaticalization

As mentioned in Section 1, researchers (Kortmann and König, 1992; Fukaya, 1997; Hayashi, 2015) have intensively studied English deverbal prepositions to quantify the degree of their grammaticalization. For example, Hayashi (2015) proposes two linguistic questionnaires to judge the degree of grammaticalization of deverbal prepositions, which ranges from 0 to 10. Based on the questionnaires, he provides the degree of grammaticalization of 37 deverbal prepositions, which is shown in Appendix A. We will use the results for our evaluation in Section 5.

De Troij and Van de Velde (2020) propose using deviation of proportion as the degree of grammaticalization, which was originally proposed by Gries (2010) to measure how uniformly a word occurs throughout a corpus. The idea behind this is that function words tend to distribute regularly or uniformly throughout a corpus than content words. They report that it is effective in measuring the degree of grammaticalization of the Dutch *soort* (*sort*) while they did not provide mathematical basis for the measure. It can be formalized as recurrence time distribution through Weibull distribution as will be discussed in Subsection 3.1.

Saavedra (2021) proposes a classification-based method where the features are word frequency, deviation of proportion, and contextual diversity. He argues that word frequency and contextual diversity are effective whereas deviation of proportion is not, which contradicts De Troij and Van de Velde (2020)'s report. Contextual diversity is measured simply by counting the number of different words in the contexts of the target word (e.g., four words left and right to the target word), which is also *ad-hoc*. Besides, in his study, the evaluation was not done against grammaticalizing words, but only against typical function and content words. Also, the predicted degrees of grammaticalization were not compared to human judgments.

Nagata et al. (2024) propose methods for quantifying grammaticalization of English deverbal prepositions. One of their methods successfully predicts the degree of grammaticalization, but is only applicable to English deverbal prepositions. They also present a more general method based on contextual diversity that uses the von Mises-Fisher distribution as a mathematical basis, reporting that it is not effective. This does not agree with Saavedra (2021)'s findings.

3 Methods for Measuring Degree of Grammaticalization

3.1 Recurrence Time Distribution-based Method

As explained in Subsection 2.2, the distribution of recurrence time of words has been regarded as effective in quantifying the degree of grammaticalization as in De Troij and Van de Velde (2020). For better comparison with the proposed classifier-based methods, we reformulate it with a mathematical basis. We define the recurrence time, denoted as τ, hereafter, by the number of words between two successive uses of the target word (strictly, the number plus one). For example, the recurrence time of the word *the* in the sentence *This is the molt that lay in the house.* is $\tau = 5$. Altmann et al. (2009) show word recurrence patterns follow a stretched exponential distribution, also known as Weibull distribution. Its probability density function is defined as[2]

$$p(\tau) = \frac{\beta}{\eta} \left(\frac{\tau}{\eta} \right)^{\beta - 1} \exp \left[- \left(\frac{\tau}{\eta} \right)^{\beta} \right] \qquad (1)$$

where β and η are the shape parameter and the scale parameter of the distribution, respectively. The shape parameter β reflects how skewed the distribution is from the exponential distribution. Altmann et al. (2009) show that content words tend to follow skewed distributions with smaller values of β while function words exhibit the opposite tendency. This can be visually exemplified as in Figure 4 in Appendix B. Considering this, we use the parameter β of Weibull distribution as the degree of grammaticalization.

3.2 PU-Learning-based Method

This method directly models the probability that the target word w is grammaticalized into a function word, which will be denoted as $p(w)$. The estimated value of $p(w)$ naturally corresponds to the degree of grammaticalization of w.

One can think of various probabilistic models for classification as $p(w)$ such as logistic regression. The target word w can effectively be represented

[2] Recurrence time denoted by τ is a discrete variable while Eq.(1) is a probability density function. We treat recurrence time approximately as a continuous variable in this study.

by its static word embedding (hereafter, simply word embedding). Then, we can use any probabilistic model for classification that takes word embeddings (i.e., vectors) as input. Word embeddings represent the contexts of their corresponding words, which in turn reflect their semantic and syntactic usages. The probabilistic model, then, is expected to quantify the degree of grammaticalization, considering semantic and syntactic similarities through word embeddings.

The problem is how to obtain the training data for the probabilistic model. There exist no training data where words are annotated with the degree of grammaticalization. It would be very difficult to create such training data from scratch.

To overcome this problem, we apply a PU-learning algorithm (Elkan and Noto, 2008) to the task of quantifying the degree of grammaticalization. PU-learning is a machine learning problem where only part of positive instances are labeled in the given training data. In other words, in the PU-learning setting, only part of positive instances are available and other instances are unlabeled.

This setting fits the present task well. In most languages, a list of function words, at least part of them, is available. Then, these function words, or their word embeddings, can be regarded as positive instances. The rest are unlabeled instances. We can apply a PU-learning algorithm to these labeled and unlabeled word embeddings to estimate $p(w)$ that a given word (vector) is a function word. We interpret the estimated value of $p(w)$ as the degree of grammaticalization; the higher the value is, the more grammaticalized the word is.

The overall procedure is as follows: (i) obtain word embeddings from the input corpus; (ii) label, as positive instances, vectors whose corresponding word is in the given list of function words; (iii) apply PU-learning to the labeled and unlabeled word embeddings; (iv) predict $p(w)$ for the target word w; (v) output the predicted value as the degree of grammaticalization.

3.3 CV-Learning-based Method

The CV-learning-based method uses POS-tagging to alleviate the difficulty in the PU-learning-based method that only part of positive instances are available in training. Running a POS-tagger on the given corpus provides the distributions of POS tags of each word in it. The CV-learning-based method uses them to determine the initial sets of function and content words. Specifically, words whose POS tags fall into function words most of the time (say, 95%) are included in the set of function words. Likewise, those whose POS tags fall into content words most of the time are included in the set of content words. The resulting lists may contain erroneous classifications. Especially, words being grammaticalized such as deverbal prepositions often receive a POS tag of a content word. Besides, most words do not meet the condition (e.g., 95% of the time) and remain as unlabeled. Accordingly, the CV-learning-based method uses the results as a tentative set of content words.

The overall procedure is as follows: (i) obtain word embeddings from the input corpus; (ii) apply a POS tagger to the input corpus to estimate the POS distributions of each word; (iii) create sets of function and content words with those used either as function or content words θ% of the time; (iv) split the set of content words into N sub-sets randomly; (v) train probabilistic models for classification N times, each time using word embeddings where positive and negative instances are from the set of function words and a different sub-set of content words. (vi) predict $p(w)$ for the target word w with the N trained models excluding the model where the target word is in the corresponding sub-set of content words; (vii) take the arithmetic mean of the predictions for the target word; (viii) output the mean as the degree of grammaticalization.

4 Creating Data for Evaluation

One of the challenges in the present task is that datasets for evaluation are hardly available. To our best knowledge, there exist no data for evaluation apart from those for English deverbal prepositions such as Hayashi (2015)'s study.

We alleviate this limitation by creating new datasets from Teramura (1992)'s test results of Japanese nouns possibly being grammaticalized. Unlike Hayashi (2015)'s study, Teramura (1992) does not provide the degree of grammaticalization. Instead, he only provides the test results where each question examines whether the target word has lost a property of typical nouns or has gained a property of typical conjunctional particles/modal auxiliaries. For example, one of the questions is: "Does the target word appear in the blank in the sentence? *This is a* ___ *.*," which typically takes a noun in the blank. Each of these questions has to do with grammaticalization, but is not the degree of grammaticalization itself. He provides the test results for 39 and 16 nouns possibly turned into a conjunctional particle and a modal auxiliary, respectively.

We merge the test results to obtain the rankings of the nouns with respect to their degree of grammaticalization. The idea behind this procedure is that the more grammaticalized the target word is, the more/less questions concerning function/content words it passes. Specifically, the procedure consists of the following steps: for each target noun, (i) count the number of tests concerning properties of conjunctional particles (or modal auxiliaries) that it passes; count those with partially satisfied (△ in Teramura (1992)'s term) as 0.5; (ii) count the number of tests concerning noun properties that it passes with the same 0.5 count for partially satisfied (△). (iii) rank the target nouns in descending order according to the difference between the first and second counts; (iv) ties are broken by giving a higher weight to tests that less target words pass; (v) if there are still ties even after (iv), give their average ranking to them.

These two resulting rankings are shown in Table 3 and Table 4 in Appendix A. Note that the noun *kara* (*because*, possible conjunctional particle) has two entries in Teramura (1992)'s test results since it has two noun usages and that it is excluded because the two usages cannot be distinguished from each other from its superficial information. Similarly, only a part of test results are shown for the noun *shidai* (*depending on*, possible modal auxiliary) in Teramura (1992)'s study and thus it is also excluded. As a result, the numbers of target words are 37 nouns (possible conjunctional particles) and 15 nouns (possible modal auxiliaries).

5 Experiments

5.1 Experimental Conditions

In the experiments, we target English and Japanese. The reasons for the choice are (i) test data are available for these two languages and (ii) they are typologically different. Specifically, we use Hayashi (2015)'s degree of grammaticalization of English deverbal prepositions and that of the Japanese nouns described in Section 4; we target the 22 English deverbal prepositions and the 37 and 15 Japanese conjunction particle and modal auxiliary nouns, respectively; their details are shown in Section 4 and Appendix A.

To implement the methods, we use the following packages. **The recurrence time distribution-based method**: `Fit_Weibull_2P`[3] to estimate the parameter β of Weibull distribution; **The PU-learning and CV-learning based-methods**: `word2vec` (Mikolov et al., 2013) to obtain word embeddings, `WeightedElkanotoPuClassifier` in `pulearn` `package`[4] for PU-learning, `LogisticRegression`[5] in `sklearn` as a classifier. Words whose POS tags fall into function/content words 95% or more of the time (i.e., $\theta = 95$ in step (iii) of the procedure of CV-learning-based method) are treated as initial sets of function/content words; the definition of function/content words are shown in Appendix D. The initial sets of content words are split into 10 sub-sets (i.e., $N = 10$ in (iv) in CV-learning-based method). Because we do not have other datasets, we use the one for the 15 Japanese modal nouns as a validation set to determine hyper-parameters in the PU-learning-based and CV-learning-based methods; the experimental details are shown in Appendix D. It should be emphasized that the evaluation of these methods on modal auxiliary nouns is not a blind-test in terms of hyper-parameters.

We obtain corpus statistics and word embeddings needed to implement the methods from the 2000s sub-corpus (texts published in the 2000s) in CCOHA for English and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) for Japanese. The details of these corpora are shown in Appendix C.

We use Spearman's rank correlation coefficient as a performance measure as in Nagata et al. (2024)'s study. In other words, we measure performance of the proposed methods by how well the rankings by the proposed methods are correlated with the human rankings.

For comparison, the following six methods are set as baselines: **Simple frequency**: function words tend to be more frequent and this method simply regards word frequency as the degree of grammaticalization; **POS distribution**: this method uses the ratio of the number of the target instances used as a function word to its total frequency as the degree of grammaticalization. The same definition as in the PU-learning-based method is used for function words. **Contextual diversity**: this is one of Nagata et al. (2024)'s methods based

[3] https://reliability.readthedocs.io/en/latest/API/Fitters/Fit_Weibull_2P.html

[4] https://github.com/pulearn/pulearn

[5] https://scikit-learn.org/1.5/modules/linear_model.html

on the concentration of contextualized word vectors that quantifies contextual diversity of the target word. **Deverbal preposition specific**: this is the best-performing method in Nagata et al. (2024)'s study that is only applicable to English deverbal prepositions; this method is not used in evaluation for Japanese nouns. The other two methods are those based on deviation of proportion (De Troij and Van de Velde, 2020) and classification with word frequency and contextual diversity (Saavedra, 2021).

5.2 Experimental Results

Table 1 shows the experimental results. The simple frequency-based method exhibits a mild correlation in deverbal prepositions and modal auxiliary nouns although the correlation coefficient is almost zero in conjunction particle nouns. In other words, word frequency can be an indicator of grammaticalization while it is not sufficient by itself. The POS distribution-based method sets a much stronger baseline. However, there are a number of ties in its rankings[6] according to the POS distribution. To be precise, the numbers of ties are: 16/22 in deverbal prepositions, 27/37 in conjunction particle nouns, and 11/15 in modal auxiliary nouns. The contextual diversity-based method also exhibits a mild correlation in all three test sets although it is not statistically significant.

Interestingly, the two methods based on the recurrence time distribution (i.e., fifth and seventh rows in Table 1) exhibit small or even negative values of the correlation coefficient. To look into this point, let us observe Figure 1 showing the relationship between the human judgments and the prediction by the recurrence time distribution-based method for the 22 deverbal prepositions. It shows that a few deverbal prepositions such as *past* and *regarding* appear along the diagonal line, meaning that the predictions agree with the human judgments well. The other majority, however, exhibit a higher degree of grammaticalization than the human judgments predict. These deverbal prepositions such as *confronting* are at an early stage of grammaticalization according to the human judgments and thus they are used as both a verb and a preposition. These two types of usage are not distinguished by their recurrence time distributions, and their values of β tend to be higher. Figure 5 in Appendix E shows that a similar tendency holds in other words than deverbal prepositions. Figure 2 also shows that the same argument applies to the degree of grammaticization estimated by contextual diversity. These results suggest that recurrence time distribution and contextual diversity do not satisfy sufficient conditions but rather necessary conditions of grammaticalization and that other sources of information are necessary to quantify the degree of grammaticalization accurately.

In contrast, the two proposed methods perform much better. The CV-learning-based method achieves the highest correlation coefficients in all target datasets among all the methods except the deverbal preposition specific method; its correlation coefficients are all statistically significant even after Bonferroni correction ($p < 0.05$).

The PU-learning-based method does not perform as well as the CV-learning-based method likely because of the fundamental difficulty in PU-learning that the classifier is learned only from partial positive instances. Nevertheless, it rivals the CV-learning-based method in the two types of grammaticalization of Japanese nouns. Note that the two classification-based methods (and also deverbal preposition specific method) are based on semantic and syntactic similarities measured through word embeddings. The previous classification-based method (Saavedra, 2021), which is based on word frequency and contextual diversity, does not perform well at all, which agrees with the argument that they do not satisfy the sufficient conditions but rather necessary conditions of grammaticalization.

Even in the CV-learning-based method, there are still words where the differences between the human and predicted rankings are large: *wanting* (difference: 15.5) in deverbal prepositions; *tokoro* (*place*, *while*) (difference: 30) in conjunction particle nouns; *kai* (*worth*) (difference: 7) in modal auxiliary nouns.

A possible reason for this is word frequency, which influences the quality of word embeddings. Actually, large frequency differences exist between good and poor predictions in deverbal prepositions and modal auxiliaries; the average word frequencies for words where the differences between the human and predicted rankings are within 2 and more than 5 are 8,393 and 2,189 for deverbal prepositions, respectively; similarly 65,381 and 999 for modal auxiliaries. This suggests that the predictions based on word embeddings learned from more

[6] For examples, some words are always tagged as a function word or a content word. Accordingly, the POS distribution-based method cannot predict which is more grammaticalized for these words.

Method	Deverbal preposition	Conjunction particle	Modal auxiliary
Simple Frequency	0.450 (0.03)	−0.082 (0.63)	0.321 (0.24)
POS distribution	0.375 (0.09)	0.451 (0.01)	0.595 (0.02)
Contextual diversity	0.355 (0.11)	0.297 (0.10)	0.333 (0.23)
Recurrence time distribution	−0.221 (0.32)	0.282 (0.09)	0.313 (0.26)
Deverbal preposition specific	**0.611** (0.01)	—	—
De Troij and Van de Velde (2020)	−0.383 (0.08)	−0.032 (0.85)	−0.04 (0.88)
Saavedra (2021)	−0.014 (0.95)	0.252 (0.13)	−0.06 (0.82)
PU-learning	0.331 (0.13)	0.508 (0.001)	**0.831** (1.2 $\times 10^{-3}$)
CV-learning	0.544 (0.01)	**0.604** (7.6 $\times$ 10^{-5})	**0.831** (1.2 $\times 10^{-3}$)

Table 1: Spearman's Rank Correlation Coefficient between Human Judgments and Degrees of Grammaticalization Predicted by Automated Methods. Values in brackets are p-values. *Deverbal preposition*, *Conjunction particle* and *Modal auxiliary* in the first row refer to English deverbal prepositions, Japanese nouns possibly grammaticalized into conjunction particles and modal auxiliaries, respectively. Note that the hyper-parameters in *PU-learning*, *CV-learning*, and Saavedra (2021) are determined by using the modal auxiliary noun data.

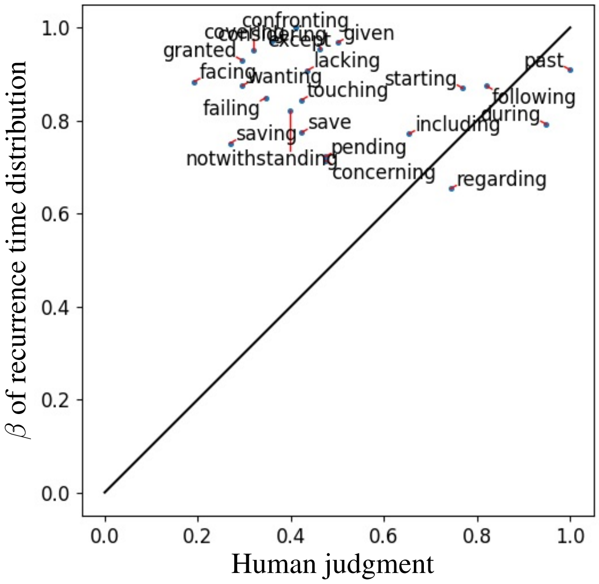

Figure 1: Relationship between Degree of Grammaticalization of Deverbal Prepositions by Hayashi (2015) and β of Weibull Distribution. The former is normalized to range from 0 to 1.

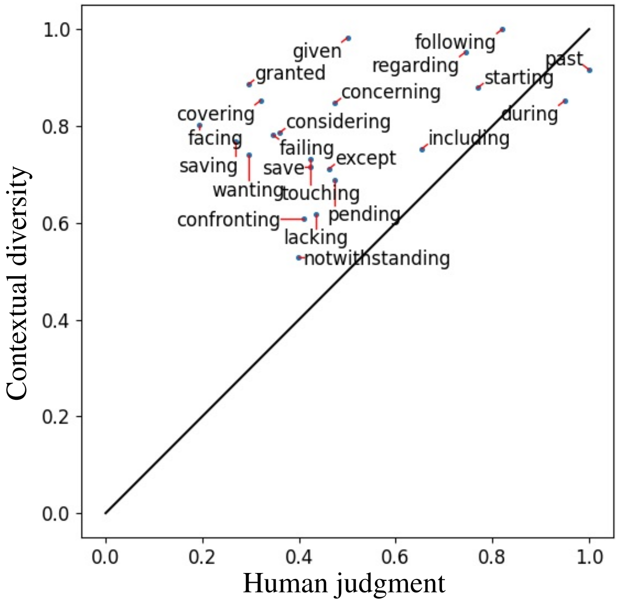

Figure 2: Relationship between Degree of Grammaticalization by Human Judgment and Contextual Diversity. Both are normalized to range from 0 to 1.

data tend to exhibit a better agreement to human rankings. However , this does not apply to conjunction particle nouns where there are considerable training instances for both cases (27,472 and 54,080 for the ranking differences within 2 and more than 5, respectively).

Another possible reason is due to the Japanese writing system. The word *tokoro* (ranking difference: 30) can be written at least in two ways, Chinese and Hiragana characters. The former and latter seem to appear more as nouns and as conjunction particles, respectively. At the same time, these two writings are learned separately in different word embeddings. The predictions were made via the word embedding from Hiragana following the writing in Teramura's questionnaire although the answers to the questions seem to cover both writings. This may influence prediction performance.

To summarize, the CV-learning-based and PU-learning-based methods achieve both high applicability and high performance. If a POS tagger is available for the target language, one can apply the CV-learning-based method to a corpus of that language to find candidate grammaticalizing words; otherwise, they can use the PU-learning-based method. Then, they can examine from which content word (e.g., a noun) to which function word (e.g., a conjunctional particle) the obtained candidates are changing by consulting actual usages. Once possible changes in categories are found, they can apply a more language-specific and word-specific method; for example, for a change from a noun to a conjunctional particle, they can compare the candidate with typical nouns and conjunctional particles through their word vectors just as in the deverbal preposition specific method. They can also use the recurrence time distribution-based and contextual diversity-based methods to see if the candidate satisfies the necessary conditions about recurrence time and contextual diversity.

6 Exploring Possibly Grammaticalized and Degrammaticalized Words

The experimental results in Subsection 5.2 reveal that the proposed methods, especially the one based on CV-learning, are effective in quantifying the degree of grammaticalization. They will likely be useful for research in grammaticalization as will be explored below; the CV-learning-based method will be used in the analysis below, which will be referred to as the proposed method in this section.

Here, we explore words possibly grammaticalized or degrammaticalized. To achieve this, we apply the proposed method to the 1800s and 2000s sub-corpora in CCOHA to quantify the degree of grammaticalization of words in the sub-corpora. We use the same hyper-parameter setting as in Section 5 except that we create the initial sets of function and content words with those appearing as either function or content words 95% or more of the time in **both** sub-corpora.

Figure 3 shows the results where the horizontal and vertical axes correspond to the degree of grammaticalization estimated from the 1800s and 2000s sub-corpora, respectively. Considering that grammaticalization occurs in high-frequency words (and also for readability), we only show words whose frequency is equal to or more than 5,000 in both sub-corpora, which results in 835 words. Note that words that are closer to the upper left/lower right corners are estimated to be more grammaticalized/degrammaticalized in the 2000s.

In the upper left area of Figure 3, the word *around* is prominent. According to `Online Etymology Dictionary`[7], the use of *around* as a preposition is relatively newer than that as an adverb. It has recently developed its use as a preposition or a particle as in *get around to it*, which was attested in 1864. The phrase *get/gets/got around to* is 6.6 times more frequent in the 2000s. Figure 3 may reflect this development of *around*.

Similarly, the word *want/wants*, more strictly *want to*, is known as *quasi-modal* (Krug, 2000) as a result of grammaticalization. This agrees with the comparison between the 1800s and 2000s in Figure 3. Related to this is the word *wanting* that is a deverbal preposition derived from the verb *want* (Hayashi, 2015).

The words *mean* and *know* appearing around *want* often occur in fixed contexts to indicate pragmatic functions such as *I mean* and *you know*. Krug (2001) points out that whether they are content (or *lexical* in his word) items or not is debatable. The coordinates of *mean/means* and *know* in Figure 3 may correspond to this fact. Words like *please* also fall into this case. These are not examples of grammaticalization, but the bleaching in their meanings may cause it eventually.

Now let us have a look at the other area in Figure 3. Words such as *must*, *such*, and *upon* appear relatively near to the lower right corner. These may be counterexamples of the unidirectionality hypothesis.

The word *must* is actually known to be a counterexample. It mostly appears in a modal auxiliary in both sub-corpora while it is also used as a noun. In the 1800s, the fraction of noun usage is only 3% and most of them are erroneously recognized as a noun by the POS tagger, `spaCy`; the rest are homographs related to the wine sense (i.e., *fresh wine*). In the 2000s, it is used as a noun derived from the auxiliary meaning *that which has to be done, seen, or experienced*, which is estimated to be 5% of all by `spaCy`. According to `Online Etymology Dictionary`, its use as a noun is relatively new, attested in 1892. Figure 3 agrees with this well.

The word *such* often (28% of the time) appears as the phrase *such a* in the 1800s sub-corpus, which is interpreted as a predeterminer. This usage decreases to 9% in the 2000s sub-corpus. Instead, it appears without the indefinite articles (e.g., *such things*) and in the phrase *such as*. These usages superficially look as if an adjective. This might be a reason why *such* appears at a lower right corner in Figure 3. Strictly, this is not a counterexample of the unidirectionality hypothesis, but suggests a possible usage shift to a content word.

Looking into its uses in the 2000s sub-corpus reveals that the word *upon* is erroneously estimated to be degrammaticalized. In the 2000s, it appears often in hyphenated words as in *our agreed-upon signal*. These hyphenated words are indeed content

[7] `https://www.etymonline.com/search?q=around`

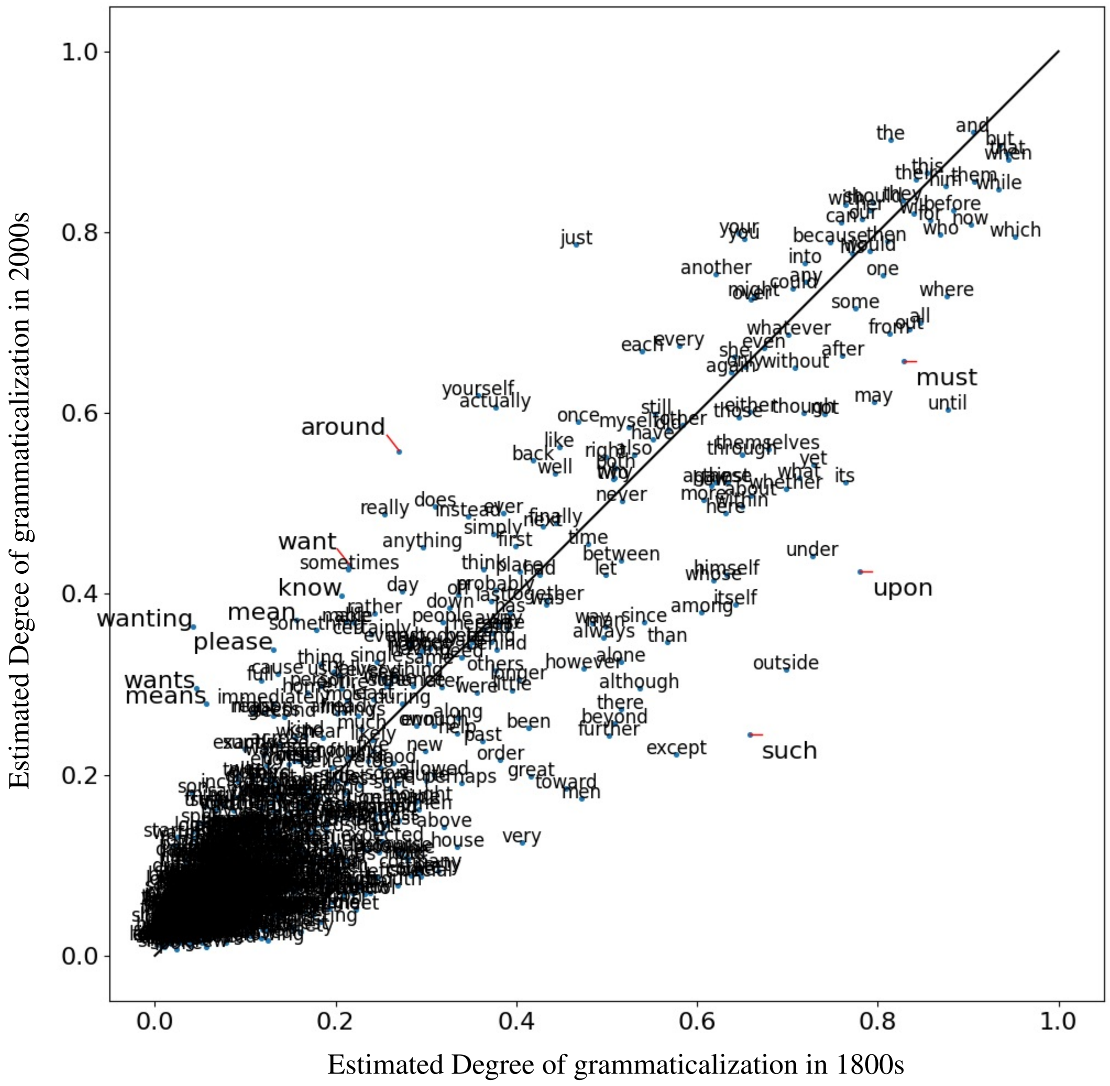

Figure 3: Scatter Plot of Estimated Degree of Grammaticalization. Horizontal and vertical axes are CCOHA 1800s and 2000s, respectively. Target words are those appearing more than 5,000 times in both sub-corpora.

words as a whole (adjectives in the above case). They were, however, split into several tokens as in *agreed*, -, and *upon* by the used tokenizer, and thus the *upon*s were merged into those of the isolated *upon* during word embedding learning. Then, the resulting word embedding of *upon* may reflect adjective properties at least to some extent.

As having been discussed, the analysis based on the proposed method enables us to explore words possibly being grammaticalized and counterexamples of the unidirectionality hypothesis. The results themselves are not a proof, but facilitate constructing new hypothesis and supporting existing hypotheses about grammaticalization just as we have seen. Ultimately, it will lead to examining the unidirectionality hypothesis although the time period examined in this study is too short to be able to tell if the hypothesis is correct or not. It would be interesting to explore the hypothesis with corpora of a wider period of time with the proposed methods.

7 Conclusions

In this paper, we have presented cross-lingual and word-independent methods for quantifying the degree of grammaticalization. We have also created new datasets for evaluation. Experiments have shown that the CV-learning-based method exhibits middle to high correlations to human judgments of grammaticalization. They have also revealed that recurrence time distribution and contextual diversity, whose effectiveness had been controversial in previous studies, are rather necessary conditions of grammaticalization and that they are not sufficient by themselves. With the CV-learning-based method, we have further explored words possibly being grammaticalized and counterexamples of the unidirectionality hypothesis.

8 Limitations

It should be emphasized that the degree of grammaticalization quantified by the proposed methods is not fully legitimate as reflected in the experimental results in Section 5. The proposed methods are only suitable for finding words possibly being grammaticalized and for supporting hypotheses about grammaticalization that one already has.

Related to this are the sizes of the evaluation data. Although we have used three types of grammaticalizing items over two languages, each data set consists only of dozens of target words. This is an inevitable limitation in grammaticalization research because the number of function and grammaticalizing words is generally limited in any language. One should consider the sizes of evaluation data when they interpret the evaluation results.

Another limitation is that the proposed methods and also the other existing methods all assume that the form of a word is constant as it undergoes grammaticalization. In reality, word forms change as grammaticalization progresses, ultimately to inflectional affixes. Similarly, grammaticalization can occur in phrases such as *be going to* and *according to*. The proposed classifier-based methods consider it indirectly as can be seen in the word *want (to)* and *(I) mean* discussed in Section 6 because word embeddings are obtained from actual usages of the target word (i.e., with surrounding words). When one wants to know how a phrase such as *be going to* is being grammaticalized, more directly, they can apply the proposed methods to them by hyphenating each element, for example. However, if one wants to find out phrases possibly grammaticalized in a corpus, the existing methods are not capable of doing so. A similar challenge exists in languages where functional elements are not explicitly separated. It is a challenging problem to quantify the degree of grammaticalization and to model the forms of the target items simultaneously.

It would be interesting to use contextualized vectors instead of non-contextualized ones (i.e., static word embeddings). They represent each usage of the target word and one can take them into consideration to quantify the degree of grammaticalization. It is not straightforward, however, how to calculate the degree of grammaticalization from multiple contextualized vectors. Besides, a language model might not work well on older corpora whereas static word embeddings can easily be obtained from any corpora. The use of contextualized vectors to quantify the degree of grammaticalization will be one of the future directions of this work.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number JP23K12152, the NINJAL collaborative research project “Toward a Computationally-Informed Theoretical Linguistics,” and AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.” We used ABCI 2.0 provided by AIST and AIST Solutions.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS ONE*, 4:1–7.

Joan Bybee. 2003. *Cognitive Processes in Grammaticalization*, pages 145–167. Lawrence ErLbaum Associates.

Joan Bybee. 2006. From usage to grammar: The mind’s response to repetition. *Language*, 82:711–733.

Lyle Campbell. 2000. What’s wrong with grammaticalization? *Language Sciences*, 23:113–161.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora*, 7(2):121–157.

Robbert De Troij and Freek Van de Velde. 2020. Beyond mere text frequency: Assessing subtle grammaticalization by different quantitative measures. A case study on the dutch *Soort* construction. *Languages*, 5(4).

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220.

Teruhiko Fukaya. 1997. *The Emergence of -ing Prepositions in English: A Corpus-Based Study*, pages 285–300. Taishukan, Tokyo.

Stefan T. Gries. 2010. *Dispersions and adjusted frequencies in corpora: further explorations*, pages 197–212. Brill, Leiden.

Martin Haspelmath. 1999. Why is grammaticalization irreversible? *Linguistics*, 37:1043–1068.

Tomoaki Hayashi. 2015. Prepositionalities of deverbal prepositions: Differences in degree of grammaticalization. *Papers in Linguistic Science*, 21:129–151.

Paul Hopper. 1991. *On some Principles of Grammaticalization*, pages 17–35. John Benjamins Publishing Company.

Paul J. Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*, second edition. Cambridge University Press, New York.

Brian Joseph. 2000. Is there such a thing as "grammaticalization?". *Language Sciences - LANG SCI*, 23:163–186.

Slava M. Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engeneering*, 2(1):15–59.

Bernd Kortmann and Ekkehard König. 1992. Categorial reanalysis: The case of deverbal prepositions. *Linguistics*, 30:671–698.

Manfred G. Krug. 2000. *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. John Benjamins Publishing Company.

Manfred G. Krug. 2001. *Frequency, iconicity, categorization: Evidence from emerging modals*, pages 309–335. De Gruyter Mouton.

Christian Lehmann. 1995. *Thoughts on grammaticalization*. Lincom Europa, Munich.

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

Antoinne Meillet. 1912. L'évolution des formes grammaticales. *Scientia*, 6(12):130–148.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Ryo Nagata, Yoshifumi Kawasaki, Naoki Otani, and Hiroya Takamura. 2024. A computational approach to quantifying grammaticization of English deverbal prepositions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 211–220, Torino, Italia. ELRA and ICCL.

David C. Saavedra. 2021. *Measurements of Grammaticalization: Developing a Quantitative Index for the Study of Grammatical Change*. Boston: De Gruyter Mouton.

Hideo Teramura. 1992. *Syntax and semantics of noun modification – part 4*, pages 1–34. Kuroshio, Tokyo.

A Details of Evaluation Data

For English deverbal prepositions, we used Hayashi (2015)'s dataset as in the previous study (Nagata et al., 2024). Out of the 37 deverbal prepositions shown, we target 22 that satisfied the following two conditions: deverbal prepositions (i) consisting of only one word and (ii) whose frequency is more than 200 in the 2000s sub-corpus in CCOHA. Table 2 shows the target deverbal prepositions with their corresponding degree of grammaticalization.

Table 3 and Table 4 respectively show the degrees (rankings) of grammaticalization for Japanese nouns possibly grammaticalized into conjunction particles and modal auxiliaries based on Teramura (1992)'s tests. The details of the data creation are described in Section 4.

B Recurrence Time Distribution Measured by Weibull Distribution

Figure 4 shows where the content word *effects* and the function word *in* appear in CCOHA[8]; the horizontal axis corresponds to the positions in the corpus, and vertical, blue lines are drawn at the positions where the target words appear. The left panel shows that the content word *effects* scarcely appears in the corpus and that its occurrences are clustered close to each other, which is known as burstiness (Katz, 1996). In contrast, the right panel shows that the function word *in* appears much more regularly, showing a rather uniform distribution in the entire corpus. If content and function words

[8] The details of this corpus are shown in Section 5.

Target word (degree of grammaticalization)
facing (1.5), saving (2.1), granted (2.3), wanting (2.3), covering (2.5), failing (2.7), considering (2.8), notwithstanding (3.1), confronting (3.2), granting (3.3), save (3.3), touching (3.3), lacking (3.4), except (3.6), concerning (3.7), pending (3.7), excluding (3.8), given (3.9), including (5.1), preceding (5.3), regarding (5.8), starting (6.0), following (6.4), during (7.4), past (7.8)

Table 2: Degree of Grammaticalization of English Deverbal Prepositions by Hayashi (2015); the higher value, the more grammaticalized.

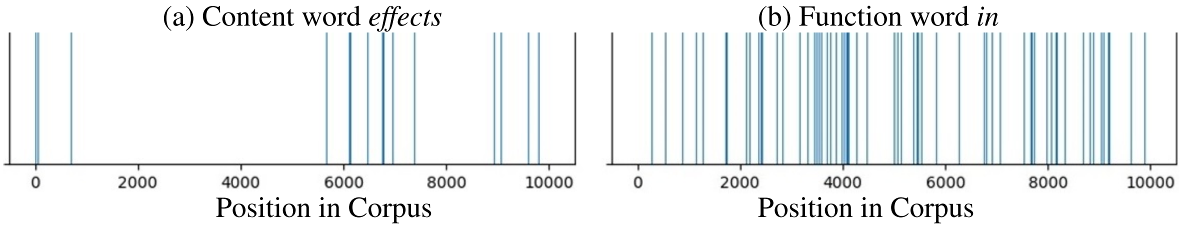

Figure 4: Occurrence Positions of *effects* and *in* in 2000s Sub-corpus of CCOHA. Vertical, blue lines are shown where the target words appear in the corpus.

Target Noun (Ranking)
kiri (1.5), nari (1.5), irai (3.0), kagiri (4.0), made (5.0), yue (6.0), dake (7.0), hodo (8.0), tabi (9.0), igo (10.0), sue (11.0), sei (12.0), amari (13.0), kurai (14.0), ue (15.0), ageku (16.0), kuse (17.5), wari (17.5), maido (19.0), izen (20.0), you (21.0), wake (22.0), tame (23.0), tori (24.0), mama (25.0), baai (26.0), gendo (27.0), teido (28.0), kekka (29.0), toki (30.5), aida (30.5), koro (32.0), gen-in (33.0), tokoro (34.0), yousu (35.0), mokuteki (36.5), riyuu (36.5)

Table 3: Degree (Rankings) of Grammaticalization of Japanese Nouns (Possible Conjunctional Particles) based on Teramura (1992)'s Tests.

Target Noun (Ranking)
rashii (1.5), darou (1.5), sou (3.0), kai (4.0), yue (5.0), you (6.0), hou (7.0), hazu (8.0), tsumori (9.5), ki (9.5), wake (11.0), yousu (12.0), ito (13.0), yotei (14.5), zizyou (14.5)

Table 4: Degree (Rankings) of Grammaticalization of Japanese Nouns (Possible Modal Auxiliaries) based on Teramura (1992)'s Tests.

follow different distributions similar to those in Figure 4, they might be useful as a source of information to quantify the degree of grammaticalization.

In Figure 4, the values of β are respectively estimated to be 0.43 and 0.97 for *effects* and *in*. We expect words being grammaticalized to have values of β in between.

C Details of Used Corpora

Table 5 shows the sizes of the corpora used in the experiments. We used the corpora following the specified licenses. We conducted the following pre-processing for the corpora:

Corpus	Size (words)
CCOHA 1800s	111,048,657
CCOHA 2000s	68,678,659
BCCWJ	124,100,964

Table 5: Sizes of Corpora Used in Experiments.

CCOHA[9]: Documents containing the string "@@YEAR.txt" (e.g., @1525.txt), which seems to be an erroneously included filename, were removed as noise. The document tags (e.g., <P></P>) were also removed. In CCOHA, 5% of ten consecutive tokens every 200 are replaced by '@' due to copyright regulations. Sentences containing these special tokens were also excluded from the analyses. The remaining sentences were tokenized by `spaCy`[10].

BCCWJ[11]: Although manual tokenization was available in this corpus, we did not use the information assuming corpora with no manual tokenization. Instead, we applied `MeCab`[12] to the corpus for tokenization with POS labels.

D Details of Experiments

We used the following implementations complying with the licenses.

Static word embeddings were obtained by using `word2vec` in `Gensim`[13]. Following the setting shown in the previous study (Nagata et al., 2024), we set the values of the hyper-parameters as follows: dimension: 200; window size: 10, the number of epochs: 100; the others: default values. The obtained word embeddings were standardized so that each dimension had zero mean with one stan-

[9] https://licenses.library.ubc.ca/EnglishCorporaCOHA

[10] https://spacy.io/, the *en_core_web_sm* model, MIT License.

[11] https://clrd.ninjal.ac.jp/bccwj/en/doc/contract/BCCWJ_Commercial_2.pdf

[12] https://taku910.github.io/mecab/, GPL, unidic.

[13] Gensim 4.3.1: https://radimrehurek.com/gensim/models/word2vec.html, GNU LGPLv2.1 license.

dard deviation and were then normalized so that the norm equaled one.

For PU-learning, `WeightedElkanotoPuClassifier` in `pulearn` package[14] was used. As its classifier, `LogisticRegression`[15] in `sklearn` was used. The following grid-search was conducted to determine the values of the hyper-parameters using the modal auxiliary nous dataset: word frequency threshold: 100, 200, 300 (we stopped the search at 300 because we were not able to target all the modal auxiliary nouns in the dataset for more than 300); inverse of regularization strength: 0.5, 1.0, 2.0, 10.0, 100.0; the other hyper-parameters: default values. As a result, we obtained the optimal setting of 100 and 100.0 for PU-learning, and 200 and 2.0 for CV-learning.

We defined function and content words as shown in Table 6. We used the definition to create the initial sets of function and content words.

We used `Fit_Weibull_2P`[16] to estimate β of Weibull distribution. We used the default hyper-parameter setting.

To implement the previous methods, we followed the implementation details shown in the corresponding papers (De Troij and Van de Velde, 2020; Saavedra, 2021; Nagata et al., 2024). In the contextual diversity-based method, some target words were split into several sub-words because it used BERT. We excluded such words from evaluation; as a result, the number of target words were 32 (instead of full 37) in the Japanese conjunction particle noun dataset. We applied the same grid-search to Saavedra (2021)'s method.

E Relationship between Degree of Grammaticalization Estimated by CV-learning-based and Recurrence time distribution-based Methods

Figure 5 shows the relationship between the degree of grammaticalization predicted by the CV-learning-based method (instead of human judgments) and the recurrence time distribution-based method where the target words are those whose frequency is more than 10,000 in the CCOHA 2000s sub-corpus. As has been seen in Figure 1 in Subsection 5.2, many content words such as and *making* show a relatively high value of β while typical function words such as *and* and *with* show even higher values. Figure 5 suggests that content words can have as high a value of β as that of a function word and that it is not enough by itself as the degree of grammaticalization.

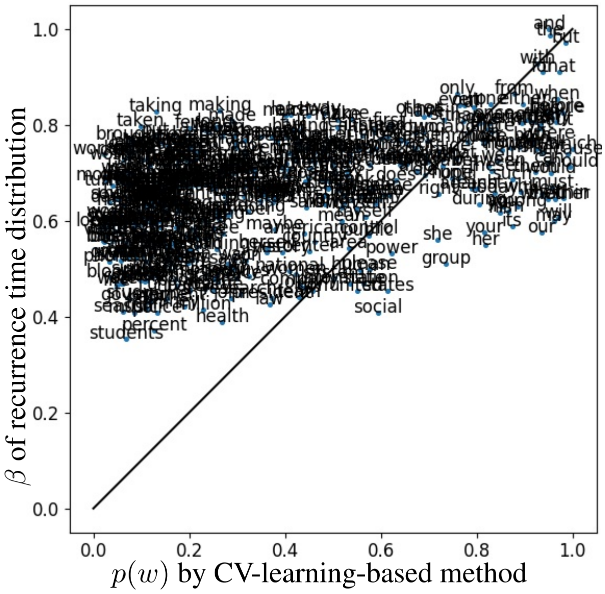

Figure 5: Relationship between Quantified Degree of Grammaticalization by CV-learning-based Methods and Recurrence Time Distribution-based. For readability, words whose frequency is equal to or more than 10,000 are shown.

English POS tag (Penn Treebank POS tag)
Function word: CC, DT, IN, MD, PDT, PRP, PRP$, RP, TO, WDT, WP, WP, WRB
Content word: JJ, JJR, JJS, NN, NNS, NP, NPS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ
Japanese POS tag
Function word: pronoun, auxiliary, particle, interjection, conjunction
Content word: adverb, verb, noun, adjective, adjectival verb

Table 6: POS Tags Used to Create Initial Sets of Function and Content Words.

[14] https://github.com/pulearn/pulearn, BSD 3-Clause license.

[15] https://scikit-learn.org/stable/modules/linear_model.html, BSD 3-Clause license.

[16] https://reliability.readthedocs.io/en/latest/API/Fitters/Fit_Weibull_2P.html, LGPL-3.0 license.