Sanskrit Word Sense Disambiguation Based on Lexicographic Definitions

Oliver Hellwig
ISLE / University of Zurich
oliver.hellwig@uzh.ch

Sven Sellmer
Institute of Oriental Studies
Adam Mickiewicz University
Poznań
sven@amu.edu.pl

Sebastian Nehrdich
Center for Integrated
Japanese Studies
Tohoku University
nehrdich@tohoku.ac.jp

Paul Widmer
ISLE / University of Zurich
paul.widmer@uzh.ch

Rico Sennrich
Department of Computational Linguistics
University of Zurich
rico.sennrich@uzh.ch

Abstract

We present a word sense disambiguation (WSD) system for Vedic and Classical Sanskrit that extends the gloss reader architecture of Blevins and Zettlemoyer (2020). Our work addresses two key questions: how semantic inventory choice affects WSD accuracy, and which data augmentation strategies can improve performance. Using the Sanskrit-English dictionary of Monier-Williams as a test case, we explore whether traditional lexicographic resources with overlapping and periphrastic definitions can be used to train a WSD system. Our model achieves high accuracy on dominant senses but struggles with secondary senses, a pattern confirmed by statistical error analysis. While model architecture has limited impact, the granularity of the semantic inventory substantially affects performance. For augmentation, we identify alignment-based and LLM-based methods as promising approaches, particularly for increasing coverage of underrepresented senses across genres and time periods.

1 Introduction

Word sense disambiguation (WSD), i.e. assigning meanings to words in a text, is an NLP technique with diverse applications in Linguistics and Digital Humanities, including electronic lexicography (Lau et al., 2014; McCrae et al., 2022), lexical semantic change detection (Hamilton et al., 2016; Dubossarsky et al., 2016; Schlechtweg et al., 2024), semantic indexing and information retrieval (Voorhees, 1993; Zhong and Ng, 2012), and machine translation (Rahul et al., 2023). Recent years have witnessed increasing interest in semantic annotation and disambiguation for ancient languages (Perrone et al., 2019; McGillivray et al., 2022; Santoro et al., 2025), and Sanskrit is no exception. Here, work focused on building and integrating Sanskrit WordNets (Bhattacharyya, 2017; Kulkarni, 2017) and annotating word senses in context (Hellwig, 2017; Hellwig and Biagetti, 2025). Due to this previous work, Sanskrit Studies are in a comparably comfortable position when it comes to building a supervised WSD system because the Sanskrit Sembank (SSB) introduced in Hellwig and Biagetti (2025) provides a large semantic resource covering important domains of Vedic and Sanskrit literature.

In this paper, we employ the SSB to develop a large scale WSD system for Sanskrit. We aim to make two main contributions. First, we study how the choice of the semantic inventory influences WSD accuracy. Many WSD systems rely on manually crafted semantic inventories such as WordNet (Miller, 1995). Quite often, these resources are optimized for their semantic granularity, which is chosen to be fine enough to conform to lexicographic standards, but coarse enough not to overwhelm machine learning systems by too many nearly synonymous senses. However, there exist scholarly bi- and monolingual dictionaries for most languages for which WSD methods were developed. These dictionaries were often built in decades of painstaking lexicographic and philological work, and their direct use for WSD can save immense amounts of time for resource building. Yet, they were created to provide scholars and translators with clusters of near-synonymous glosses rather than discrete, disambiguated senses. Therefore, it is unclear how WSD systems perform when trained with lexicographic data which abound in periphrastic and overlapping

definitions. Using the Sanskrit-English dictionary of Monier-Williams[1] (Monier-Williams, 1899) as a test case, we explore if a Sanskrit WSD system can be trained directly on a traditional dictionary with possibly overlapping and redundant lexicographic definitions. Our second contribution concerns strategies for data augmentation. Based on a comprehensive statistical error analysis of our WSD system, we will discuss possible pathways for data augmentation, concentrating on LLM and alignment based methods which have gained in popularity over the last years.

The rest of the paper is structured as follows. After a brief overview of related work (Section 2), Section 3 describes the dataset extracted from the digitized Monier-Williams. Section 4 introduces the model architecture. In Section 5 we evaluate the model performance (5.1) and identify critical design choices in two ablation studies (5.2). In addition, we perform a Bayesian analysis of the errors made by our WSD model (5.3), enabling us to give qualified recommendations for how to augment the semantic dataset for higher labeling accuracy (Section 6).

Code for model training and inference is available at `https://github.com/OliverHellwig/sanskrit/tree/master/papers/2026iscls`.

2 Related research

Most WSD methods are based on the distributional hypothesis according to which the lexical context determines a word's meaning (Firth, 1957). This hypothesis is most directly implemented in Bayesian WSD methods that condition the sentence context on latent word meanings (Frermann and Lapata, 2016; Perrone et al., 2019) and even enable to determine the degree of polysemy (Inoue et al., 2022). While the Bayesian formulation gives a statistically complete representation of WSD, it is typically based on atomic word meanings. This introduces two related challenges. First, data sparsity prevents the reliable estimation of cooccurrence probabilities even for medium frequency words, necessitating the use of strong priors in these cases. Second, it is difficult or even impossible to share statistical power between atomic word units, both on the target (i.e. the word to be disambiguated) and the context side. Therefore, the majority of current WSD systems, including the one in this paper, focuses on the likelihood part of the complete Bayesian formulation, approximating it with a deep learning framework.

Pre-transformer deep learning approaches suffered from the problem that static distributional embeddings such as word2vec (Mikolov et al., 2013) represent each word type with a single context-independent dense vector, making it difficult to capture the appropriate sense of a polysemous word in a given context and to map vectors to senses in semantic inventories. This motivated the development of non-parametric multi-embedding models (Huang et al., 2012; Neelakantan et al., 2014).

The advent of deep contextualized models such as BERT and ELMo introduced a new paradigm (see e.g. Loureiro et al. (2021) for an overview). Contextualized models create token representations based on the surrounding words and can therefore represent different senses of the same word in different contexts, yielding substantial gains for WSD (Huang et al., 2019; Hadiwinoto et al., 2019). Our system is most closely related to Blevins and Zettlemoyer (2020), who propose a gloss-informed bi-encoder architecture where one encoder encodes the sentence context and another encoder encodes the lexicographic definition. The two resulting dense vectors are compared to select the appropriate sense. This approach improves performance particularly for rare senses by directly matching contexts to glosses.

More recently, some approaches try to bypass the development of dedicated WSD systems by prompting LLMs for semantic annotations. This paradigm is especially relevant for under-resourced languages where word semantic annotation is constrained by the lack of (human) resources. For example, Riemenschneider and Frank (2023) conduct experiments for Ancient Greek, Kaše et al. (2025) use LLM-annotated corpora for a diachronic Latin case study, Santoro et al. (2025) explore Latin WordNet annotation with LLMs, Riemenschneider (2025) employs LLM components to address challenging instances in Akkadian lemmatization tasks, and Lugli (2025) studies how to prompt LLMs for augmenting semantic annotations of Buddhist Sanskrit.

3 Data

3.1 Composition of the dataset

To develop a large-scale WSD model, we make use of the Sanskrit Sembank described in Hellwig and Biagetti (2025). Each Sanskrit lemma in this resource is associated with at least one synset from an adapted version of the English WordNet (see Figure 1), and individual corpus occurrences of a lemma can be annotated with one or more synsets, indicating their meaning in a given sentence context. Moreover,

[1] When talking about the Monier-Williams in the following, we refer to the dictionary of the Digital Corpus of Sanskrit (DCS) which was converted from an early digitization of the MW, but manually extended ever since.

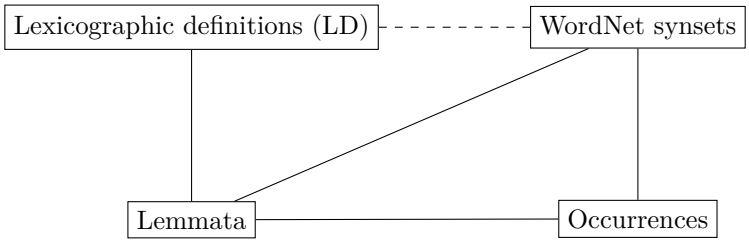

Figure 1: Simplified schema of the Sanskrit Sembank. Dashed links indicate that a type of information is not available for all entries.

each lemma is associated with one or more lexicographic definitions, most of which are taken from the Sanskrit-English dictionary of Monier-Williams (1899). Many synsets are manually linked to individual LDs of their lemma.

Hellwig and Biagetti (2025) report results of experiments performed with shallow WSD models that predict WordNet synsets for individual occurrences of words in the DCS. While they observe an encouraging accuracy rate of 76.6% for a stratified sample from the SSB, their approach, like most WSD methods, relies on a dedicated sense inventory which is represented by the 'WordNet synsets' node in Figure 1. In this paper, we aim to explore how to circumvent the time consuming creation of such an inventory, and therefore use the Sanskrit Sembank in a slightly different way: instead of WordNet synsets, we predict lexicographic definitions. To this end, we employ the graph structure shown in Figure 1. If a corpus occurrence of a lemma is annotated with a WordNet synset, we follow the link from occurrences to synsets, and then use the link emerging from synsets to retrieve the corresponding lexicographic definitions. In this way, we are able to leverage the lexicographic information contained in Monier-Williams (1899) which is often much more exhaustive and fine-grained than that in the SSB. For example, the verb *sam sañj-* is associated (and annotated) with the single WordNet synset 'adhere' in the Sanskrit Sembank. However, its eleven lexicographic definitions listed in Monier-Williams (1899) span a much wider semantic space, including, for example, the metaphorical use 'to falter (in voice)', which is realized in phrases such as *asaṃsaktayā vācā* 'with unfaltering voice'. By finetuning pretrained transformer based modules, we are in principle able to make predictions for such lexicographic definitions that are not or only rarely annotated, which represents a significant advantage over the approach discussed in Hellwig and Biagetti (2025).

To simplify data collection and model architecture, we employ only those LDs for our WSD system which can be assigned unambiguously to annotated corpus occurrences. This choice excludes two cases:

1. multiple synsets connected to one word occurrence, used to deal with semantic underspecification; see Hellwig and Biagetti (2025, §3.2).

2. synsets not linked to any LD. This case occurs when the Sembank annotator has created a lemma-synset association on the fly without providing the additional link to an LD, or when a specific meaning found in a text is not recorded in Monier-Williams (1899). To illustrate, the verb *abhi dhā* is annotated with the synset 'enumerate' ('specify individually') in the SSB, but this meaning is not recorded in Monier-Williams (1899) although semantically related speech verbs are also listed here. Currently, we are linking such "orphaned" synsets to the lexicographic definitions, using strategies described in Patel and Kulkarni (2024) (and related to vector space based methods described in Goyal et al. (2012)). Due to the time-consuming nature of this step, its results are not yet reflected in this paper. In the training data overview in Table 1, both cases are represented by the columns labeled as 'Unconnected'.

In this paper, we perform WSD for open word classes, i.e. nouns (N), verbs (V), and adjectives (A), thereby excluding indeclinables, pronouns and numerals, all of which are semantically annotated in the SSB. The DCS, from which the data is sourced, provides a coarse chronological categorization of texts into Vedic (1), epic (2), classical (3), "medieval" (4) and late (5) works. The Vedic texts can further be divided into the early Saṃhitās and later prose treatises that explain the Vedic ritual. The prose texts feature numerous citations from the earlier metrical texts; from among the 1,189,446 words in the Vedic part of the DCS, 113,708 are parts of mantras. To obtain historically unbiased data, we exclude words forming part of mantras from the training data because these occurrences often show meanings no longer in active use in the citing texts. Filtering is performed using the digital version of M. Bloomfield's *Vedic*

	Tokens				Types					
	Connected		Unconnected		Connected		Unconnected		Lex. type	Sem. types
Class	#	%	#	%	#	%	#	%	#	#
N	248,894	78.6	67,604	21.4	22,278	65.2	11,905	34.8	20,520	12,535
V	51,071	55.3	41,223	44.7	5,929	52.5	5,361	47.5	4,334	2,314
A	43,693	73.9	15,437	26.1	5,580	69.1	2,496	30.9	4,968	2,982
Total	343,658	73.4	124,264	26.6	33,787	63.1	19,762	36.9	29,822	17,785

Table 1: Overview of the semantically labeled data. The training dataset consists of the columns labeled 'Connected'. #: counts. %: proportion of connected vs unconnected per row.

Concordance (Hellwig et al., 2023).[2] Note that mantra citations are retained in sentence contexts. For example, in the sentence *atha yajamānaḥ somakrayaṇīm īkṣate māhaṃ rāyaspoṣeṇa viyoṣam iti* 'Then the sacrificer looks at the cow used for buying soma [while speaking the mantra]: "Do not let me be separated from increase of wealth" ' (Baudhāyana Śrautasūtra 6.13), we do not use the verb *vi yav* (*viyoṣam*) contained in the mantra part as training target. However, when disambiguating the verb *īkṣ* (*īkṣate*), the full mantra is retained in the context sentence because mantras might provide content clues for WSD. Moreoever, we do not deduplicate repeated mantras in the metrical Vedic texts. Finally, we exclude the early grammatical literature, namely Pāṇini's Aṣṭādhyāyī and Yāska's Nirukta, because these texts employ a highly technical meta-language coined to describe derivational and morpho-syntactic processes.

We use 90% of the available data for training, 5% for validation, and 5% for evaluation. Statistics of the training dataset are given in Table 1, split by the three open word classes. Due to the gaps in the mapping from synsets to lexicographic meanings (see above), we cannot use the complete sense annotation of the SSB for the system developed in this paper. For example, while about 80% of all noun tokens annotated with a WordNet synset are linked to a lexicographic definition in the MW, this rate is significantly lower for verbs. The empirical distribution of lexicographic definitions per word follows an expected, strongly right-skewed pattern (median = 2, mean = 3.9, SD = 4.7, range = 1–87). Fitting a Zipf distribution yields an exponent of $s = 1.59$ (SE = 0.0035), indicating that the frequency of highly polysemous words decreases approximately as an inverse power of their number of senses.

3.2 Case study: The term *kratu-* in Vedic

Using a scholarly dictionary instead of WordNet introduces several challenges some of which are best illustrated with a case study. To this end, we choose the term *kratu-* which is known to be a semantically complex word (see e.g. Strunk (1975)). *kratu-* originally denotes a kind of action-directed inner energy with both volitional and intellectual dimensions; subsequently, also the products of that energy (plans etc.); and lastly a 'sacrifical rite' (see Rönnow (1932, p. 51–54) for possible explanations of this semantic shift).

In Monier-Williams (1899), the meanings of *kratu-* – as of every polysemous lemma – are formally differentiated on two levels: "mere amplifications of preceding meanings are separated by a comma, whereas those which do not clearly run into each other are divided by semicolons" (Monier-Williams, 1899, xv). The following list gives all meanings of *kratu-* that are marked as occurring in Vedic texts, with bullet points instead of semicolons:

- plan, design, intention, resolution, determination, purpose
- desire, will
- power, ability
- deliberation, consultation
- intelligence, understanding
- inspiration, enlightenment
- a sacrificial rite or ceremony, sacrifice (as the Aśva-medha sacrifice), offering, worship

The DCS lists the same meanings, but without the hierarchy inherent in the Monier-Williams, so that they form an unstructured set of 18 meanings. Obviously, the meanings of the original groups are almost synonymous in many cases although Monier-Williams claims to have reduced the number of synonyms given in his dictionary significantly in the second edition, which is the one used by us (Monier-Williams, 1899, xv). Indeed, a look at the first edition (Monier-Williams, 1872) shows that the number of Vedic meanings for *kratu-* amounted to 23 there.

[2]Data available at `https://github.com/OliverHellwig/sanskrit/tree/master/papers/2023wsc/data`.

To understand how these lexicographic definitions relate to the annotations in the SSB, the second author of this paper annotated all Vedic occurrences of *kratu-* in the DCS with the lexicographic definitions from Monier-Williams (1899). 36 from among these passages are also annotated with synsets in the SSB. We used the link-following method described in Section 3.1 to retrieve the corresponding lexicographic definitons. Quite surprisingly, we found that only four out of 36 passages (11%) are annotated with the same lexicographic definition. However, the situation appears less severe when the following factors are considered. First of all, many of the divergent annotations may actually be treated as matching if we consider the meaning sets defined in Monier-Williams (1899). In this way, we achieve the following additional matches:

a sacrificial rite or ceremony	sacrifice (as the Aśvamedha sacrifice)	7
intelligence	understanding	3
intention	plan	1
purpose	plan	1

This increases the number of matching annotations to 18 (50%). While still far from what is considered a sufficient inter-annotator agreement, several factors explain this figure. One of them is the double, intellectual-volitional nature of *kratu-*, which means that the annotator – in the absence of similarly ambiguous English terms – has to opt for one of the two aspects that often are present simultaneously, so that the decision typically is not clear-cut; these remarks apply in particular to the five diverging annotations of 'intelligence' vs. 'will'. Most of the remaining differences are due to the fact that *kratu-* denotes different noun types, which in many cases are not easy to distinguish and so are liable to result in diverging annotations: objective nouns (i.e., 'plan'), instrumental nouns (i.e., 'will', 'intelligence'), and action nouns (i.e., 'deliberation'). In addition, all of the difficulties just mentioned should be seen in the context of cultural and temporal distance as well as the highly poetical and deliberately cryptic character of Vedic literature, especially its older parts.

4 Model

The structure of the WSD model is inspired by the gloss encoder of Blevins and Zettlemoyer (2020); see also Manjavacas Arevalo and Fonteyn (2022) and Rachinskiy and Arefyev (2022) for related approaches. To determine the sense of a Sanskrit word in its sentence context, the WSD model is presented with the Sanskrit sentence, the Sanskrit word itself, and a list of possible English meanings extracted from Monier-Williams (1899). Its task is to select the English meaning that best fits the Sanskrit word in the given sentence. To share statistical power between similar configurations of words, sentences, and senses, these three elements are encoded as learnable dense vectors.

More formally, the model determines the compatibility of n LDs $D^{(l)}$ of a Sanskrit lemma l with a given Sanskrit sentence s by creating dense embeddings $e(\ldots)$ of all $d_i^{(l)} \in D$, of l, and of s:

- Embeddings of LDs are created with a SentenceTransformer (Reimers and Gurevych, 2019) trained to judge semantic similarity. We evaluate two settings for creating the definition embeddings $e(d_i)$:

  Trainable We integrate the SentenceTransformer as a submodule into the WSD model, enabling full end-to-end adaptation of its weights. The output of the submodule's pooling layer is used as dense representation $e(s)$. Due to memory constraints, this setting is evaluated with the `all-MiniLM-L6-v2` architecture.[3]

  Cached Instead of further finetuning an integrated SentenceTransformer on the WSD task, we use it to precalculate embeddings of all LDs, corresponding to the model's penult layer activations; see Kumar et al. (2019) for a related approach. These embeddings are cached and adapted during training. We use the more powerful `all-mpnet-base-v2` in this setting.

- We evaluate two methods to create the lemma embeddings $e(l)$:

  Atomic Each lemma obtains its own, randomly initialized embedding vector which is adapted during WSD training.

  Character-level Lemmas are encoded with a small character-level transformer.[4] Due to its tiny size, this transformer is not pretrained.

[3]Each training batch consists of B records, each of which contains up to 87 lexicographic definitions, each tokenized into 128 subword tokens, and each token is represented by an E-dimensional dense vector. Even very moderate batch sizes of $B = 8$ produce OOM errors on a V100 GPU with 32GB of VRAM if the 110M parameter `all-mpnet-base-v2` with $E = 768$ is employed.

[4]Settings: embedding dimension: 256, 2 transformer layers, 4 heads, feedforward dimension: 512, maximum sequence length: 64

- Sanskrit sentence embeddings are obtained with a character-level transformer pretrained with a masked language modeling task on 4GB of partly noisy Sanskrit text, i.e. the dataset used in Nehrdich et al. (2024).[5] Different from the LDs' encoder, this sentence transformer has not been finetuned on a semantic similarity task. Therefore, its complete transformer architecture is finetuned during WSD training. Again, we evaluate two settings for creating the sentence embedding $e(s)$:

  Full The complete masked output of the sentence transformer is averaged and passed to the classification layer.

  Range Only the substring containing the target word is averaged and passed on. For example, let the verb *śru* 'hear, listen to' be disambiguated in the phrase *janamejayas tu nṛpatiḥ śrutvākhyānam anuttamam* 'after king Janamejaya had listened to this excellent story …'. In the setting 'Range', we average the character embeddings of the substring *śrutvākhyānam*, which contains the target word. This setting corresponds more closely to the setup employed by Blevins and Zettlemoyer (2020) who average the word piece vectors describing the target word. Due to Sandhi and morphological richness, averaging the range of the actual word form *śrutvā* cannot be reproduced easily in Sanskrit and is left for future work.

For each occurrence of a sense labeled lemma, we calculate one query vector and multiple key vectors. The query q describes the context of the semantic unit. It is obtained by concatenating $e(l)$ and $e(s)$, and applying a non-linear affinity that reduces the length of $e(l) \oplus e(s)$ to that of $e(d_i)$. The keys $e(d_i)$ encode the possible meanings in this context. Dot products $q \cdot e(d_i)$ are calculated for the n LDs. The training objective is to select, from among the n LDs of the focus lemma l, the correct LD annotated in a given sentence context. To this end, the elements $q \cdot e(d_i)$ are interpreted as logits of a categorical distribution, and the model is trained to minimize the cross-entropy between the one-hot encoded ground truth and its categorical output distribution. If a given lemma has only one LD, the model is trained to maximize the cosine similarity between $e(d_1)$, i.e. the embedding of its only sense, and the query embedding.

5 Evaluation

5.1 Overview

Table 2 displays the word level accuracy[6] of our model in its best setting (on which see the discussion of Table 5 in Section 5.2.1). The accuracy rates are split by word classes (columns) and the coarse chronological layers of the DCS (rows). Overall, the model achieves 83.3% accuracy, which is a high value for an all-words task.

As in Hellwig and Biagetti (2025), one can observe very pronounced differences in WSD accuracy. Across all word classes (columns of Table 2) and historical periods (rows), nouns have the highest accuracy rates, and verbs the lowest ones. Similarly pronounced differences between nouns and verbs have been observed elsewhere (see e.g. Table 4 in Raganato et al. (2017)), and are typically explained by the higher semantic flexibility of verbs and their greater dependence from sentence context. In addition, accuracy rates are higher for epic (period 2), "medieval" and late Sanskrit (periods 4 and 5) than for Vedic (1) and classical texts (3). As detailed in Hellwig and Biagetti (2025), this result is probably due to the higher annotation density in periods 2, 4 and 5, and the fact that densely annotated texts from these periods have similar genres (epic, alchemical literature).

Large accuracy differences are also observed when the data is split by the MFS (most frequent senses) status. As becomes apparent from Table 3, the model achieves substantially higher accuracy rates for dominant senses (MFS) than for secondary ones, whose accuracy rates are more than halved for verbs and adjectives in comparison to the MFS. Similar effects are reported in Blevins and Zettlemoyer (2020, Table 2) and Kumar et al. (2019, Table 4).

Since the scores for secondary senses of adjectives and particularly verbs are remarkably low, it is worthwhile to inspect the actual errors made by the model. Table 4 lists the top ten Sanskrit verbs, their MW definition derived from the SSB ('Gold'), the meaning predicted by the WSD model ('Silver'), and

[5] At its input layer, this model concatenates the character embeddings with convolutions of width 3 and 5 to capture syllable level and subword patterns. These input vectors are passed through a standard transformer architecture with 8 transformer layers, 8 attention heads, and feedforward dimension of 512; maximum sequence length: 256. Following Devlin et al. (2019), 15% of the characters are masked, and the model is trained to reconstruct them.

[6] Most WSD papers use the scorer found at `https://github.com/ysenarath/SemEval-2015-task-13/blob/main/SemEval-2015-task-13-v1.0/scorer/scorer_original.py` or a variant thereof to report F1 scores. However, we report accuracy rates because our system outputs the sense with the highest model probability as prediction and therefore never abstains nor yields multiple proposals.

Period	N	V	A	Overall
1	83.4	66.7	68.9	76.9
2	90.4	69.9	78	86.1
3	81.9	61.1	71	77.7
4	88.6	77.1	80.1	85.8
5	88.1	79	84.2	86.3
All	87.5	70.5	76.3	83.3

Table 2: Accuracy of the WSD method, split by word classes (N = nouns, V = Verbs, A = adjectives) as columns and the historical layers of the DCS as rows.

Sense	N	V	A	Overall
MFS	96.2	89.2	91.9	94.7
Secondary	49.7	32.2	32.8	42.9

Table 3: Accuracy of the WSD method, split by word classes (columns) and most frequent sense (MFS) status of the record (rows)

the frequency of each misclassification ('#'). The left half of the table contains cases in which a secondary sense is predicted instead of the correct MFS. Obviously, most predicted senses are semantically very close to the MFS, or even synonyms. For example, 'besmear' is an applicative synonym of 'smear', 'select' and 'choose for one's self' are near-synonyms belonging to different registers and diatheses, and 'press …' and 'crush' primarily differ in lexical aspect. In addition, these words may even be valid alternatives when translating a given Sanskrit passage. The only real errors seem to be made for *sthā-* and *kṛ-*.

The right half of Table 4, which lists secondary senses wrongly predicted for other secondary senses, shows a similar picture. Here, *ā nī-* and, with restrictions *ā veś-*, are problematic, whereas the other cases present semantically highly related definitions. Table 4 thus suggests that the classification performance of the WSD model is substantially higher than suggested by the accuracy values in Table 2, and other metrics and strategies may be more appropriate for model evaluation.

5.2 Ablation studies

5.2.1 Model architecture

As described in Section 4, the architecture of the WSD models offers several choices that modify or extend the gloss reader of Blevins and Zettlemoyer (2020). To understand their influence on model accuracy, we perform an ablation study. The most time-critical component is the trainable sense encoder: activating it increases training time from four to more than thirty hours and additionally requires a dedicated GPU (see fn. 3). In order to save resources, we evaluate all settings not involving this component in the first step. In the second step, we only evaluate the worst and the best from among these settings, but with the trainable sense encoder activated.

The upper part of Table 5, containing configurations 1-6, shows the results of the experiments with cached sense embeddings, again split by POS classes. Judging from the overall accuracy scores (last column), there exist clear, though not dramatic differences between the configurations. Most relevantly, extending the gloss encoder of Blevins and Zettlemoyer (2020) with lemma embeddings consistently improves scores; compare the rows with 'Lemma=None' with the other configurations. Interestingly, verbs and adjectives suffer most from not having access to lemma information (configuration 3), but this can be healed to a certain degree by using sentence ranges (configuration 4). Moreover, averaging the range of the sentence containing the target word instead of averaging the whole sentence increases accuracy in two out of three cases. Encoding the lemma with a mini-transformer provides another moderate gain in accuracy, yielding the best configuration with cached word embeddings (#6).

Configuration 7 displays the results with the trainable sense encoder enabled. As could be expected, this setting yields another clear accuracy gain of 0.8% as compared to configuration 6. However, this gain comes along with significantly increased training and inference time (see above).

5.2.2 Influence of the dataset

For two reasons, the accuracy rates obtained with the model in this paper cannot directly be compared with those reported by Hellwig and Biagetti (2025). First, Hellwig and Biagetti (2025) do not report results for an all-word model, but evaluate a stratified sample from the SSB to control for typical factors influencing WSD accuracy. Second, the discussion in Section 3 has shown that the dataset composition

$M:S$				$S_1:S_2$			
Lemma	Gold	Silver	#	Lemma	Gold	Silver	#
marday	to press or squeeze hard	to crush	17	jan	to become	to come into existence	8
sthā	to stay	to stand	17	ānī	to pour	to lead towards or near	6
lip	to smear	to besmear	9	āviś	to get or fall into	to go or drive in or towards	5
bandh	to tie	to bind	5	vraj	to travel	to move	4
grah	to take (by the hand)	to grasp	5	bhū	to happen	to occur	4
vṛ	to select	to choose for one's self	4	jan	to come into existence	to become	3
rudh	to shut	to close	4	viś	to enter	to enter (a house etc.)	3
jan	to be born	to come into existence	4	vraj	to move	to go away	3
kṛ	to make like or similar	to do	4	vas	to dwell	to abide with or in	2
dviṣ	to show hatred against	to hate	4	jan	to be changed into (dat.)	to become	2

Table 4: Frequent errors made for verbs. Left half: Model predicted secondary sense instead of MFS; right half: predicted a wrong secondary sense.

Dataset	#	Sense	Lemma	Sent.	N	V	A	Global
Lex. def.	1	Cached	Atomic	Full	86	69.6	74.6	81.9
	2			Range	86.2	68.6	74.5	81.9
	3		None	Full	84	64	71.9	79.2
	4			Range	85.2	67.4	74.7	81
	5		Char.-level	Full	85.8	69.2	76.6	81.9
	6			Range	86.6	69.2	76.2	82.5
	7	Tr.	Train.	Range	87.5	70.5	76.3	83.3
SSB	6				87.3	79.3	83.9	85.5

Table 5: Accuracy scores for the ablation studies in Sections 5.2.1 and 5.2.2

Signal	Model	k@1	k@3	k@5	k@10
Random order	–	26.44	47.45	61.77	81.80
Sanskrit Original	–	27.72	57.23	72.31	88.52
MT (full sent)	Gemini 2.5 Flash-Lite	42.70	73.13	83.66	94.41
MT (full sent)	Gemini 3 Flash	44.39	74.71	**85.14**	**95.18**
MT (lemma)	Gemma 3 Flash	**52.90**	**76.10**	84.85	94.63

Table 6: Top-k accuracy (%) for semantic reranking over the meaning candidate space (up to 87 lexical definition candidates).

differs significantly between the SSB and this paper, both in terms of its absolute size and the number and granularity of senses to be distinguished. The interplay of these factors challenges any theoretical statement about their influence on WSD quality. To understand how employing lexicographic definitions instead of WordNet synsets influences WSD accuracy, we perform another ablation experiment. Using exactly the same 90/5/5% data split as for the results in Tables 2, 3 and 5, we replace each record's distribution over lexicographic definitions with the corresponding distribution over WordNet synsets. We derive the definition strings from the synsets' names. If multiple synsets share the same name, we differentiate them by concatenating the synset name n with its WordNet gloss g, employing the template 'n, i.e. g'. We train a model in configuration 6 with this dataset; the reason for this choice is again to save resources through shorter training time.

The results of this experiment are printed in the last line of Table 5. Using the SSB synsets instead of MW definitions substantially increases the overall accuracy by 3%. Gains are especially large for verbs (almost 9%) and adjectives (more than 7%), whereas nouns suffer a tiny decrease. This outcome can certainly be explained by the different degrees of granularity of the two datasets. As detailed in Hellwig and Biagetti (2025, §3.1), the SSB annotations are often underspecified, resulting in coarser and above all less senses per word than given in Monier-Williams (1899). This statement finds statistical support in a Wilcox test that compares the number of senses per word in the MW evaluation dataset with that in the SSB evaluation dataset, and yields a highly significant test statistic of 81774750 ($p < 2.2e^{-16}$). Though still not fully comparable, the overall accuracy of 85.5% in the SSB setting is by far higher than the scores reported in Hellwig and Biagetti (2025), aligning with similar performance gains reported in other WSD papers using transformer based models.

As the size of the meaning space seems to matter, we evaluate semantic reranking as a way to shrink the effective meaning candidate space (up to 87 lexical definitions), reducing memory usage and making training and large-scale labeling more practical. Since the target meanings are in English, we use machine translations of the containing sentences to construct reranking signals. We use BGE-M3 (Chen et al., 2024) as the reranking model. We evaluate in two settings: One where the entire sentence is translated into English and used as the retrieval signal, and the other is where we prompt the model to only translate the given lemma into English, having the full sentence as context included in the prompt. We test Gemini 2.5 Flash-Lite and Gemini 3 Flash (accessed in January 2026) against an unsorted random baseline, and additionally probing a no-translation condition that reranks using the original Sanskrit sentence. Table 6 shows that full-sentence MT yields a large improvement over the random baseline across all k, and model quality does make a difference in retrieval accuracy with the more costly Gemini 3 Flash showing better performance that 2.5 Flash Lite. While reranking with the original Sanskrit sentence brings a slight improvent over the randomized baseline, it performs substantially worse than English MT, suggesting that bridging into the target meaning language is critical in this setup. The MT (lemma) setting shows better performance in the @1 and @3 granularity, but falls behind sentence-based retrieval when more candidates are included. The results are encouraging because a recall of more than 95% can be obtained in the @10 setting, which means that significant pruning of the target space without severe performance degradation is possible.

5.3 Statistical error analysis

In addition to word classes, historical periods, and the MFS status of individual records (see the discussion of Table 2), WSD accuracy crucially depends from characteristics of the training data. While previous WSD studies often evaluate these factors in isolation, we hypothesize that various training set statistics interact in determining accuracy rates, and should therefore be evaluated in combination. This suggests the use of a statistical error model that describes WSD model success as a function of the aforementioned factors. Apart from insights in the behavior of our WSD model, such an error model offers practical benefits: it makes it possible to provide confidence levels for model predictions, and helps focusing on

error prone parts of the dataset for future augmentation.

We perform a Bayesian error analysis that aims to understand how the composition of training data for individual words influences the accuracy with which their senses are predicted. To this end, we introduce the binary variable y_i which is true if the model predicted the correct meaning of record i, and false otherwise. Then, we collect the following statistics for each model prediction i, following similar setups described, for instance, by Navigli (2009) and Raganato et al. (2017):

Word class Its word class (N, V, A)

MFS A binary indicating whether or not the predicted sense is the most frequent sense of w_i in the training dataset (i.e. silver MFS). Note that this MFS indicator differs from that employed for subsetting Table 3 which reflects the MFS gold status.

Training samples (lemma) The number of training samples for lemma w_i

Training samples (sense) The number of training samples for the predicted sense s_i. – Lexicographic definitions such as 'king' or 'a name of Śiva' are recorded for numerous different lemmas. Since they employ the same sense embeddings, we expect that accuracy rates increase along with this statistic. – More elaborate measures are conceivable here. For example, one may expect that the lexicographic definitions 'king', 'a king' and 'a mighty king' are placed in close neighborhood in the sense embedding space. Under this assumption, the counts used here may be replaced with a kernel function on the graph defined by pairwise cosine similarities of embeddings. We leave this type of evaluation to a future publication.

Training entropy The entropy of the training targets for lemma w_i, which measures the concentration around the dominant sense of w_i. – In practical terms, we stack the one-hot target distributions used to train the WSD model into one matrix, form its column sums, normalize the resulting count distribution $\mathbf{p}$ across the K senses of w_i, and calculate its normalized entropy by dividing the observed entropy (denominator) by the maximally possible entropy on a K-dimensional simplex:

$$e_i = \frac{-\sum_k^K p_k \log p_k}{-\sum_k^K \frac{1}{K} \log \frac{1}{K}} = \log K - \sum_k p_k \log p_k \tag{1}$$

Senses The number of lexicographic definitions of w_i. Highly polysemous words can be expected to have high WSD error rates.

Sense proportion The proportion of training samples annotated with s_i among all annotations of w_i

Probability CDF The highest probability value in the model output for record i, is often used to quantify confidence in the model prediction. However, raw probability values cannot be employed for the model discussed here because the categorical output distributions of each record have different dimensions, depending on the number of senses of w_i. Intuitively, a maximum probability of $\theta = 0.65$ predicted by the WSD model is high for a word with twenty senses, but low for one with two senses only. To account for different numbers of senses, we use the cumulative density function (CDF) of the maximum of a uniform Dirichlet distribution instead of the raw maximum probability. This distribution has a closed form; see e.g. Holst (1980, Theorem 2.1). It quantifies the probability to observe a maximum value x less or equal to θ on a K-dimensional simplex. Continuing the example from above, $p(x \leq 0.65, K = 20)$ yields 1, i.e. very high confidence, whereas $p(x \leq 0.65, K = 2)$ only yields a low confidence value of 0.3.

Period proportion Proportion of training records of w_i that are from the same period as record i. This and the two following measures are meant to capture domain (mis-)match.

Genre proportion Proportion of training records of w_i that have the same DCS genre classification as record i.

Text proportion Proportion of training records of w_i that are from the same text as record i.

Period Table 2 as well as the discussion in Hellwig and Biagetti (2025) have shown that accuracy levels differ across the coarse historical periods of the DCS/SSB. Therefore, we include the historical period of record i as a contrast encoded predictor.

Model	elpd_diff	se_diff
Joint partial pooling	0.0	0.0
Pooling by MFS	-52.7	14.4
Pooling by POS	-129.2	16.2
No pooling	-206.2	20.7

Table 7: Comparison of various error-detection models using loo values

Setting	N	V	A	MFS	Sec.	Global
PPC	0.859	0.821	0.852	0.972	0.542	0.862
CV	0.856	0.813	0.847	0.971	0.552	0.858

Table 8: PRAUC values for the posterior predictive checks (row 1) and the tenfold cross-validation, split by POS tags and MFS status (columns)

We use the 5% evaluation split of the semantic dataset for Bayesian error analysis. The predictors 'Training entropy', 'Sense proportion', 'Probability CDF', 'Period' and 'Text proportion' are naturally bounded in $(0, 1)$. The count predictors 'Training samples (word)', '…(senses)', and 'Senses' are bounded in $(0, 1)$ after min-max scaling. To stabilize parameter estimation, we logit transform the values x of these scalar predictors, i.e. work with $u = \log(x/(1-x))$, and additionally logit-transform them. No transformations are applied to the contrast encoded 'Period' predictors.

The evaluation of Table 2 has demonstrated that error rates differ significantly across word classes. Moreover, WSD systems are typically better in predicting the MFS of a word than its secondary senses. To emphasize the importance of POS tag and MFS status, we employ a Bernoulli model with partial pooling based jointly on word classes and the binary MFS indicator. Note again that this is the silver MFS status, i.e. whether the WSD model predicted the MFS. Employing the indicator functions $\mathrm{c}(i)$, which returns the word class of record i, and $\mathrm{mfs}(i)$, which represents the MFS status of s_i, this model predicts the binary outcome label y_i based on the predictors defined above:

$$\begin{aligned}
\alpha &\sim \text{Normal}(0,1),\ \sigma_\alpha \sim \text{Exponential}(s),\ \alpha_{j,k} \sim \text{Normal}(\alpha, \sigma_\alpha) \\
\beta_l &\sim \text{Normal}(0,1),\ \sigma_\beta \sim \text{Exponential}(s),\ \beta_{j,k,l} \sim \text{Normal}(\beta_l, \sigma_\beta) \\
m_i &= \alpha_{\mathrm{c}(i),\mathrm{mfs}(i)} + \sum_l \beta_{\mathrm{c}(i),\mathrm{mfs}(i),l} \cdot x_{i,l} \qquad (2) \\
y_i &\sim \text{Bernoulli}(1/(1+\mathrm{e}^{-m_i})) \qquad (3)
\end{aligned}$$

We fit this model using `rstan` (Stan Development Team, 2025), assigning 500 epochs to warmup and 500 to sampling. To ensure that the chosen pooling structure appropriately describes the data, we fit three additional models to the same data: one that performs partial pooling only for POS tags, one that does the same for MFS, and one without partial pooling. We compare the four resulting models using leave-one-out cross-validation (loo, Vehtari et al. (2017)). The results in Table 7 demonstrate that joint partial pooling shows the best predictive performance. While this model's elpd is almost four standard errors larger than that of the MFS-only model, this difference is much more pronounced when compared with the POS-only model and that without partial pooling.

To further assess model fit, we perform posterior predictive checks (PPC) by sampling y values from the model's posterior distribution (Equation 3). Averaging the predictions across all posterior draws and relating these means to the ground truth values y_i, we obtain the AUC values reported in the first row of Table 8. Similar AUC scores are obtained in a tenfold cross-validation where the model is trained with 90% of the data and evaluated on the remaining 10%; see the second row of Table 8. The overall AUC of .86 indicates good predictive capabilities and thus seems to qualify the Bayesian model as an effective error detector. However, the second segment of Table 8 shows that the overall performance is dominated by the records annotated with the MFS. These cases alone achieve an AUC of more than .97 in PPC and cross-validation. In contrast to that, the AUC for secondary senses is considerably lower (.54/.55).

Employing the error detection model enables the focused retrieval of new semantic annotations. For example, aiming at a precision level of at least 80% or, if this level cannot be reached, the maximal accuracy available, it is possible to calculate confidence levels of the error detection model for specific POS/MFS combinations, and use them to filter WSD results. To obtain realistic estimates, we reuse the cross-validation data from above by estimating thresholds using nine training folds, and apply these thresholds to the tenth left-out fold. Table 9 shows the outcome of this approach. Its left compartment

							MFS			Secondary		
POS	M:M	S:S	M:S	S:M	S_1:S_2	Rej.	P	R	F	P	R	F
N	2306	405	64	9	31	9931	99.6	96.0	97.8	81.0	97.8	88.6
V	1805	2	1	450	1	772	80.0	99.9	88.9	50.0	0.4	0.9
A	171	15	3	1	0	1870	99.4	98.3	98.8	83.3	93.8	88.2

Table 9: Results of applying the error detection model to the data, targeting at a minimum of 80% precision. The left table segment gives the raw counts, and the right segment p(recision), r(ecall) and f (score) for the non-rejected cases ("pseudo PRF"), split by gold MFS and secondary senses.

shows the resulting raw counts, with M:M denoting that the MFS of a record was correctly labeled as MFS, S:S the same for secondary senses, and $S_1 : S_2$ for a secondary sense mislabeled as another secondary sense. The column 'Rej.' records the number of records below the specific threshold. Cases involving secondary senses are particularly relevant for extending the semantic database. Therefore, their "pseudo-PRF" scores, i.e. scores excluding the 'Rejected' counts, are listed in the right half of Table 9. Both for nouns and adjectives, the method yields acceptable recall rates of more than 90% although the precision scores are clearly not high enough for a reliable philological application, thus necessitating manual postprocessing. The corresponding scores for verbs are by far too low to be usable.

Let us now discuss what the Bayesian model reveals about the factors influencing WSD quality. The parameter m_i of the inverse logit (Equation 2) is a weighted linear combination of the predictors. Therefore, it makes sense to compare the absolute values of the individual coefficients and their separation from zero to understand their influence on model predictions. Figure 2 displays the 90% confidence intervals (CI) of the partially pooled coefficient values, accumulated across all posterior draws. Each subplot contrasts the global mean value for one predictor (see the labels at the right side of the figure) with that for the three POS classes that depend from the mean (see the sub-labels at the left side of the plot). The left and right columns additionally indicate whether or not the predicted senses is the MFS. CI error bars are printed in red if they do not intersect the zero line (i.e. are "significant"), and blue else.

Figure 2 shows clearly that the outcomes of the Bayesian model are positively determined by the values of 'Sense proportion', 'Probability CDF' and 'Training samples (sense)', meaning that high values of these predictors lead to high proportions in the Bernoulli distribution. This result makes sense since high proportions of a given word at word level ('Sense proportion') and at the global level ('Training samples (sense)') typically entail higher WSD accuracy, as does a highly confident model output ('Probability CDF'). An opposite effect of comparable magnitude is observed for the number of senses per word ('Senses'). Again, this makes sense because highly polysemous words are typically harder to label correctly than words with few senses. A weaker negative effect is detected for 'Period 3', aligning with the discussion of Table 2 which demonstrated low accuracy for this period. The CIs of the remaining predictors mostly contain zero and therefore exert a limited influence on the value of m_i In addition to the grouping factor MFS, the POS tags, which act as the second grouping level, have a pronounced influence in some configuration; see e.g. 'Sense proportion' for secondary senses of nouns or 'Probability CDF' for secondary senses of verbs both of which are clearly below the global coefficient estimates for their groups.

While the inspection of coefficient estimates provides an initial idea of which predictors steer the behavior of the Bayesian model, it does not fully address the question of which training set characteristics exert the strongest influence on the prediction accuracy of the WSD model. This is due to the model structure: although the predictors contribute linearly to the mean parameter m_i, their influence on the resulting proportion is not linear in their size due to the non-linearity of the inverse logit and correlations between predictors. For example, if the value of 'Sense proportion' alone has moved the inverse logit of m_i close to one, the actual magnitude of 'Probability CDF' will not make a great difference in the final outcome because the inverse logit squashes its contribution.

To address this issue, we perform systematic interventions in the value of each predictor j, keeping all other predictors at their observed values. We set $x_{i,j}$ to its 5% conditional quantile,[7] evaluate Equation 3 for the whole dataset, and determine the AUC for this configuration. Then we repeat these steps for the 95% conditional quantile of all values of predictors j. Finally, we calculate the difference of the two AUC values which quantifies the change in model accuracy when changing the value of predictor j from

[7]We tested whether the scalar predictors follow a joint multivariate Normal distribution, but found that they definitely fail to do so, preventing the calculation of conditional quantiles with a closed form Gaussian linear system (on which see Bishop (2006, 2.81-82)). Instead we use quantile regression (Koenker and Bassett, 1978) to determine the conditional distribution of $x_{i,j}$, given $X_{i,-j}$, using the function `rq` from the `R` package `quantreg`. Coverage tests show that this method yields meaningful approximations.

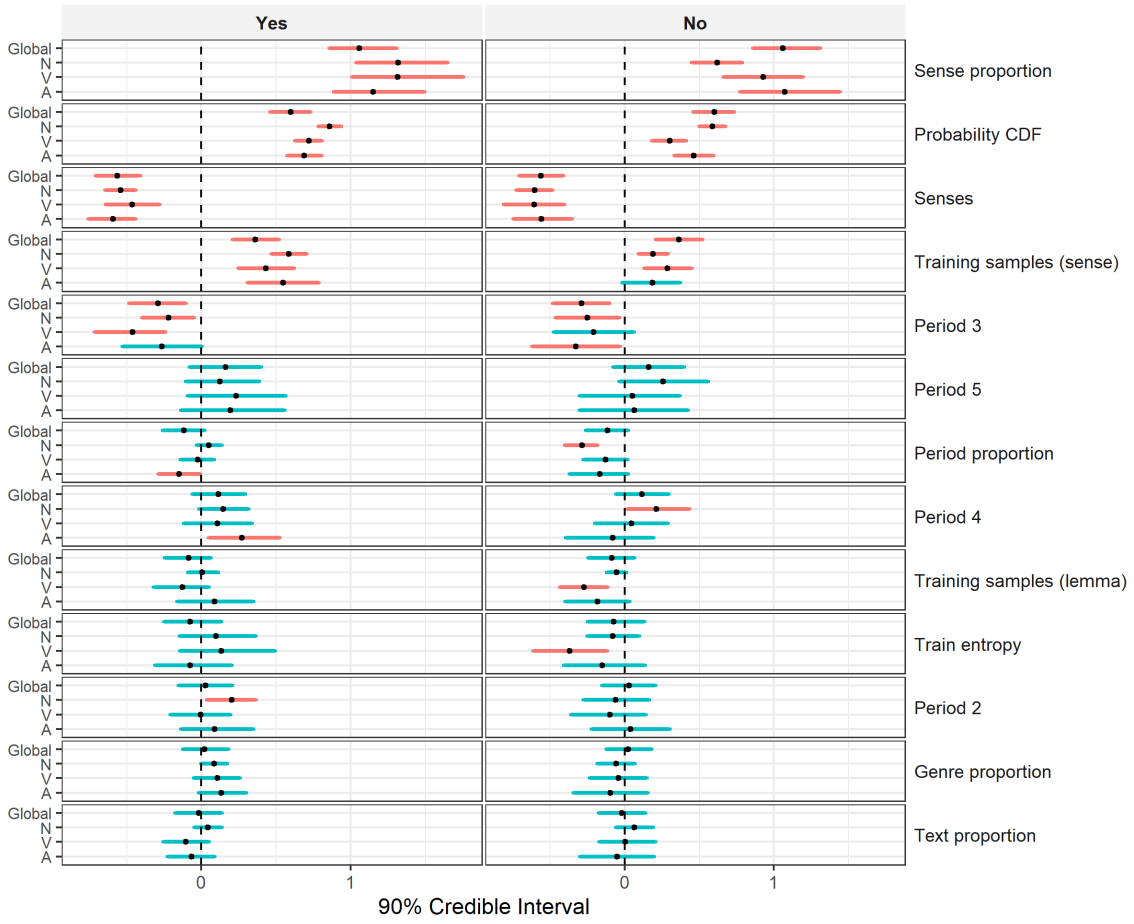

Figure 2: 90% confidence intervals (CIs) of the coefficients of the Bayesian error detection model, split by features (right labels), POS tags (sublabels) and MFS status (columns; 'Yes': the WSD model predicted the MFS). Rows ordered by the median of the 'Global' coefficient (i.e. β_i in Equation 3). Red: the CI does not include zero.

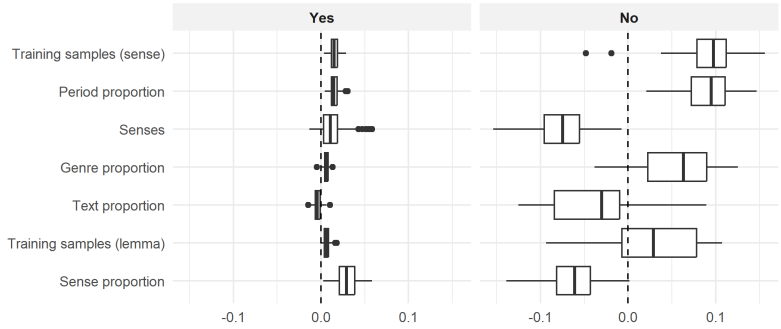

Figure 3: Results of do-interventions in the error prediction model; see Figure 2 for the interpretation of the plot. Each boxplot describes the difference in AUC when setting the respective predictor to the 95% and 5% quantiles of its conditional distribution.

a very low to a very high value. Note that we perform interventions only for those predictors that we can control by further annotation or dataset restructuring, excluding, for instance, period assignments of individual records. However, we do include the number of senses ('Senses') because its value can be modified by changing the semantic inventory.

The results of these interventions are displayed in Figure 3. As in Figure 2, results are split by predictors (rows) and MFS status (columns); note that Figure 3 is split by the gold MFS status. The most obvious result are the strong differences between sensitivity scores for MFS records (left) and those with secondary senses (right): while none of the interventions shows particularly large effects for MFS, these effects can be very pronounced for secondary senses. In particular, changing 'Training samples (senses)', 'Period' and 'Genre proportion' from their 5% to 95% conditional quantiles increases the AUC scores by 5-10% on average. The opposite effect for 'Sense proportion' is expected as higher values of this predictor for a non-MFS sense lead the error model to label this record as wrong (high values of 'Sense proportion' are typically associated with gold MFS; see the comparatively strong effect of this predictor for the MFS). However, the strongest negative effect is observed for 'Senses', meaning that reducing the number of senses per word can be expected to yield gains in accuracy. Finally, just increasing the number of training instances per lexeme has the weakest positive effect on secondary sense classification. This suggests that the WSD model in its present state is capable of learning from a limited number of attestations.

Comparing Figure 3 with Figure 2, one may ask why the two predictors 'Period' and 'Sense proportion', whose coefficient estimates oscillate around zero in Figure 2, obtain such a prominent position in Figure 3. We would like to argue that this apparent paradox arises from the non-linear nature of the inverse logit transformation: predictors with high coefficient values primarily shift probabilities in regions where the sigmoid response is already saturated, while 'Period' and 'Genre proportion' operate near decision boundaries where small changes substantially impact classification accuracy.

One predictor we did not include in the intervention analysis is 'Probability CDF', i.e. the confidence of the WSD model in its own prediction. While Figure 2 suggests that this predictor may have a strong effect on the output of the error detection model, it is obvious that its value depends from that of the other predictors. For example, lemma-noun combinations for which many training samples are available can be expected to obtain higher WSD probabilities than those with limited training data. To account for these dependencies, one would need to introduce a Bayesian submodel which conditions 'Probability CDF' on the remaining predictors. Using this submodel for marginalization would enable us to estimate the total effect sizes that predictors describing the training set composition exert on WSD model accuracy. We experimented with several variants of such a submodel. However, even the best among them (a hurdle model that models the low confidence tail of the CDF with a logit-Gaussian component) did not fit the data well enough to provide meaningful marginalization probabilities. We postpone this detailed evaluation to a future publication.

6 Discussion

We have introduced and discussed a WSD model for Vedic and Classical Sanskrit that extends the gloss reader architecture of Blevins and Zettlemoyer (2020). Our contribution aims to understand which factors drive WSD accuracy for an ancient, but well attested language like Sanskrit, and explore methods to improve it. Model evaluation in Section 5.1 shows high accuracy values. However, like related approaches,

our model is clearly better in predicting dominant than secondary word senses (see Table 2). The error detection model discussed in Section 5.3 shows a similar behavior since it is best in labeling the correctness of dominant senses. As a consequence, even when used in combination with the error detection model (Section 5.3), the overall model scores are not high enough to enable reliable high-coverage annotation of the extant Sanskrit literature that can be used, for example, to provide comprehensive corpus occurrences for a historical dictionary. Nevertheless, the WSD model can certainly serve as a good pre-labeler of word semantic annotations that are subsequently validated by a human annotator, especially when dealing with nouns (see Table 9).

What are the main sources of errors, and how can we improve WSD results? Judging from the ablation study in Section 5.2.1, the model architecture has a noticeable, though limited influence on WSD accuracy since the accuracy difference between the best and and the worst model configuration amounts to only 3.3%. This does not mean that other approaches could not improve the values reported in this paper. For example, one may consider to finetune a cross- instead of a bi-encoder, i.e. use a multilingually pretrained backbone such as XlmRoberta, concatenate a single English lexicographic definition with the Sanskrit lemma and sentence, retrieve a classification token, and use its value to calculate the logits of the multinomial sense distribution. Alternatively, one might explore more sophisticated ways to integrate lemma information in the existing model architecture. We carried out experiments with a bilinear projection of senses to sentence embeddings, conditioning the square projection matrix on the lemma embedding. In spite of its theoretical appeal, this architecture performs slightly worse than configuration 6 in Table 2, while requiring more than one day of training time.

Instead of changing the model architecture, a more viable and efficient way to increase WSD accuracy may be to augment the training data, in particular for secondary senses. Judging from the results of the statistical analysis and Figure 3 in particular, it appears crucial to increase the number of annotations of a specific sense across all lemmata annotated with it, as well as its annotations from the same time period and genre. This outcome highlights the importance of sharing statistical power between instances, and provides a practical recipe for choosing subcorpora for dataset augmentation.

From among the methods to obtain such additional data, alignment and LLM based augmentation strategies appear to be especially promising. Over the last few years, the authors of this paper have created large collections of Sanskrit texts with aligned English translations. Specifically for Vedic, we have created the Vedic Prose Corpus (Hellwig, 2026) which features manually validated alignments of 20,000 sentences from post-Saṃhitā Vedic text with their scholarly translations. To extend the Vedic dataset with translations of metrical Saṃhitā texts, the Zurich Paippalāda project (Zehnder et al., 2024) offers a modern translation of an important Saṃhitā. Moreover translations of the Rig- and Atharvaveda in its Śaunaka recension (Whitney and Lanman, 1905) are available in the open domain. The easiest way to utilize these aligned translations is applying algorithms based on string similarity, either simple string similarity and overlap, or more advanced alignment based methods that employ language models; on which see e.g. Keersmaekers et al. (2023). Prompting, as explored by Lugli (2025) for Buddhist Sanskrit, might offer another approach. For example, one may prompt a multilingual LLM with a given Sanskrit sentence, its (machine) translation, and the list of definitions from Monier-Williams (1899), and ask it to choose the best matching from among them. This approach may be improved by supplying randomly selected, but similar examples from the existing semantic database; see e.g. Li et al. (2024) on this approach in general.

Overall, this paper has confirmed the central role that the choice and design of the semantic inventory plays for WSD. In particular, the data ablation study in Section 5.2.2 has shown that simply employing the less fine-grained SSB synsets instead of the complete MW boosts accuracy by several percents even when cached embeddings are used (see the last line of Table 2). Furthermore, our experiments with semantic reranking (Table 6) demonstrate that the effective meaning space can be substantially reduced through cross-lingual signals, with English machine translation of Sanskrit sentences enabling BGE-M3 to consistently place correct senses in the top-k candidates, thereby mitigating the computational challenges posed by large sense inventories. Similarly, Table 4 has demonstrated that many misclassifications are semantically very close to the gold meanings, again suggesting that the MW data have a comparatively low "semantic cardinality". In addition, the do-interventions in the error detection model have demonstrated that an increasing number of senses has a detrimental effect on error detection accuracy (see the discussion of Figure 3).

An alternative, though much more time consuming solution is the use of the "synsets" found in Monier-Williams (1899) itself, i.e. the groups of senses defined by the lexicographer. As the case study in Section 3.2 has shown anecdotally, employing them instead of ungrouped MW senses increases IAA, and we hypothesize that the same holds true for WSD accuracy. This suggests that we should integrate the

MW's sense groupings into the dictionary structure of the DCS, a step that would also increase the value of the DCS for philological work. In addition, the sense groups of MW can probably be used to augment training data for our WSD model: instead of just using the lexicographic definitions linked from the SSB, we can equally use their synonyms from MW (see Janz and Maziarz (2023) for a related setup based on WordNet synonyms). Overall, our findings suggest that the path to robust WSD for ancient languages lies not in architectural complexity, but in carefully curating existing semantic inventories.

Acknowledgments

Oliver Hellwig was funded by the NCCR Evolving Language (SwCSS NSF Agreement Nr.51NF40_180888) when doing research for this paper.

References

Pushpak Bhattacharyya. 2017. IndoWordNet. In Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D. Pawar, editors, *The WordNet in Indian Languages*, pages 1–18. Springer, Singapore.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the NAACL*, pages 4171–4186.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: A bottom up computational approach to semantic change. *Lingue e Linguaggio*, 1:5–25.

J.R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Philological Society, Oxford.

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

Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, and Ralph Bunker. 2012. A distributed platform for Sanskrit processing. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 1011–1028, Mumbai, India, December. The COLING 2012 Organizing Committee.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 5297–5306, Hong Kong, China. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1489–1501.

Oliver Hellwig and Erica Biagetti. 2025. The Sanskrit Sembank. *Language Resources and Evaluation*, pages 1–24.

Oliver Hellwig, Sven Sellmer, and Kyoko Amano. 2023. The Vedic corpus as a graph. An updated version of Bloomfield's Vedic Concordance. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 188–200.

Oliver Hellwig. 2017. Coarse semantic classification of rare nouns using cross-lingual data and recurrent neural networks. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS), Long papers*.

Oliver Hellwig. 2026. The Vedic prose corpus. In *Proceedings of the Workshop on Atharvaveda, Rome January 2025*. Sapienza Università Editrice. Accepted for publication.

Lars Holst. 1980. On the lengths of the pieces of a stick broken at random. *Journal of Applied Probability*, 17(3):623–634.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2022. Infinite SCAN: An infinite model of diachronic semantic change. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1605–1616, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Arkadiusz Janz and Marek Maziarz. 2023. Data augmentation method for boosting multilingual word sense disambiguation. In German Rigau, Francis Bond, and Alexandre Rademaker, editors, *Proceedings of the 12th Global Wordnet Conference*, pages 60–66, San Sebastian, January. Global Wordnet Association.

Vojtěch Kaše, Sarah Lang, and Petr Pavlas. 2025. Embedded in the labyrinth: Investigating latin word senses through transformer-based contextual embeddings and attention. In *Anthology of Computers and the Humanities*, volume 3, pages 498–512. Anthology of Computers and the Humanities.

Alek Keersmaekers, Wouter Mercelis, and Toon Van Hal. 2023. Word sense disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 148–159, Varna, Bulgaria.

Roger Koenker and Gilbert Bassett. 1978. Regression quantiles. *Econometrica*, 46(1):33–50.

Malhar Kulkarni. 2017. Sanskrit WordNet at Indian Institute of Technology (IITB) Mumbai. In Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D. Pawar, editors, *The WordNet in Indian Languages*, pages 231–241. Springer.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Ligeia Lugli. 2025. The Mangalam dictionary of Buddhist Sanskrit: Automating lexicographic data with generative LLMs. In I. Kosem, Jakubíček, Medveď M., Zgaga M., Š. Arhar Holdt, T. Munda, and A. Salgado, editors, *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent lexicography*, pages 757–773.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter, editors, *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

John P. McCrae, Theodorus Fransen, Sina Ahmadi, Paul Buitelaar, and Koustava Goswami. 2022. Toward an integrative approach for making sense distinctions. *Frontiers in Artificial Intelligence*, 5:745626.

Barbara McGillivray, Daria Kondakova, Annie Burman, Francesca Dell'Oro, Helena Bermúdez Sabel, Paola Marongiu, and Manuel Márquez Cruz. 2022. A new corpus annotation framework for Latin diachronic lexical semantics. *Journal of Latin Linguistics*, 21(1):47–105.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Monier Monier-Williams. 1872. *Sanskṛit-English Dictionary.* The Clarendon Press, Oxford, 1 edition.

Monier Monier-Williams. 1899. *Sanskṛit-English Dictionary.* The Clarendon Press, Oxford, 2 edition.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41:10:1–10:69.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA, November. Association for Computational Linguistics.

Dhaval K Patel and Amba Kulkarni. 2024. Word sense alignment of Sanskrit lexica. In Arnab Bhattacharya, editor, *Proceedings of the 7th International Sanskrit Computational Linguistics Symposium*, pages 1–13, Auroville, Puducherry, India. Association for Computational Linguistics.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66.

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

C. Rahul, T. Arathi, Lakshmi S. Panicker, and R. Gopikakumari. 2023. Morphology & word sense disambiguation embedded multimodal neural machine translation system between Sanskrit and Malayalam. *Biomedical Signal Processing and Control*, 85:105051.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider. 2025. Beyond Base Predictors: Using LLMs to Resolve Ambiguities in Akkadian Lemmatization. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, Marco Passarotti, and Rachele Sprugnoli, editors, *Proceedings of the Second Workshop on Ancient Language Processing*, pages 226–231, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Kasten Rönnow. 1932. Ved. *kratu-*. Eine wortgeschichtliche Untersuchung. *Le Monde Oriental*, 26/27:1–90.

Daniela Santoro, Beatrice Marchesi, Silvia Zampetta, Marco Del Tredici, Erica Biagetti, Eleonora Litta, Claudia Roberta Combei, Stefano Rocchi, Tullio Facchinetti, Riccardo Ginevra, and Chiara Zanchi. 2025. Exploring Latin WordNet synset annotation with LLMs. In Chiara Zanchi, Luca Brigada Villa, Erica Biagetti, Alexandre Rademaker, Francis Bond, and German Rigau, editors, *Proceedings of the 13th Global Wordnet Conference*, pages 66–76, Pavia, Italy. Global Wordnet Association.

Dominik Schlechtweg, Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2024. Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*.

Stan Development Team. 2025. RStan: the R interface to Stan. R package version 2.32.7.

Klaus Strunk. 1975. Semantisches und Formales zum Verhältnis von indoiran. *krátu-/xratu-* und gr. *κρατύς*. In *Monumentum H.S. Nyberg*, volume II, pages 265–296. Bibliothèque Pahlavi, Téhéran/Liège.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Ellen M. Voorhees. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180, New York, NY, USA. Association for Computing Machinery.

William Dwight Whitney and Charles Rockwell Lanman. 1905. *Atharva-Veda Samhita*. Harvard University, Cambridge.

Thomas Zehnder, Oliver Hellwig, Robert Leach, Magdalena Plamada, Angelika Malinar, and Paul Widmer. 2024. Atharvaveda Paippalāda Zurich edition book 1 (version 1.0.0) [data set]. LaRS - Language Repository of Switzerland. DOI: https://doi.org/10.48656/as99-n988.

Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves Information Retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 273–282, USA. Association for Computational Linguistics.