APODICTUS: Automatic Processing Of DICTionary Update candidateS

Felix Blessing[1], Johannes S. Sax[1], Julian Kaufmann[1], Wei Zhao[3], Nikolay Arefyev[2], Dominik Schlechtweg[1]
University of Stuttgart[1], University of Oslo[2], University of Aberdeen[3]
first.last@{ims.uni-stuttgart.de/abdn.ac.uk}

Abstract

Dictionaries have to be regularly updated. Some dictionary-makers gather proposals for updates of sense entries in internal databases. We automate the process of verifying and prioritizing such sense proposals, and facilitate their addition to a dictionary, by building a sophisticated processing pipeline relying on state-of-the-art language models. Our pipeline presents the first systematic, large-scale, and comprehensive solution for processing candidates for inclusion in a dictionary, which is tested in an industry-relevant context. We conduct several experiments to evaluate the pipeline and provide an annotated dataset for future work. Model performance is acceptable for words which are not yet in the dictionary, but low for in-dictionary words. Through an error analysis and model component ablation, we gain further insight on directions of future model improvements.

Keywords: Unknown Sense Detection, Word Sense Disambiguation, Sense Definition Generation

1. Introduction

Humans have systematically recorded the **senses** of words in **dictionary** entries since ancient times. Today, millions of users worldwide access Oxford Dictionaries every month. Keeping these dictionaries up-to-date is a crucial challenge to lexicographers, as language changes quickly over time and resources are finite. Although computational processes are widely used in lexicography, there is considerable further potential to increase the efficiency and exhaustiveness by assisting the essential human component in lexicographic work (Rundell, 2023).

Oxford University Press (OUP) has an internal database called "Lexical Maintenance (and) Unifying Resource" (LEMUR) where editors can propose new sense entries for its various dictionaries including the Oxford Dictionary of English (ODE), a large dictionary of current English. Consider the following example of such a proposed sense in **LEMUR** for the headword *spam*:

- Slang. To press or strike (a computer key, button, etc.) many times in quick succession

Additionally, the headword has the following already recorded senses in **ODE**:

- irrelevant or unsolicited messages sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc.
- a tinned meat product made mainly from ham
- send the same message indiscriminately to (a large number of internet users)

The main task we solve is to find **corpus usages** of the proposed sense, such as the following one for *spam*:

- In dramatic sequences, God of War might ask the player to **spam** "X" or twirl the control sticks…

Many of the proposed senses have a low corpus frequency making their detection a highly challenging **needle-in-the-haystack** problem.

Reasons for adding a sense proposal to LEMUR are diverse. A lexicographer may have come across the new word or sense personally, or identified it as a gap when drafting or updating a related word, or a term may have come up in lists of failed website searches or trending words in a corpus. The LEMUR database currently contains around 65,000 such candidates for dictionary inclusion. Verifying them requires lexicographers to find and analyse evidence of the proposed new sense being used in a variety of different sources. Sense proposals are currently prioritized using corpus frequency and editorial judgement to identify those which are likely to become a permanent feature of the language or which are culturally significant. However, given the very large size of the LEMUR database, many sense proposals have not yet undergone prioritization.

The aim of this study is to automate the above-described process of verifying and prioritizing sense proposals.[1] For this, we build a sophisticated processing pipeline relying on state-of-the-art language models for Unknown Sense Detection

[1] OUP provided the LEMUR and ODE data used for this study.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11800–11812
11–16 May 2026.

(USD), Word Sense Disambiguation (WSD), Word Sense Induction (WSI) and Sense Definition Generation (SDG). We provide the first systematic, large-scale, and comprehensive solution for processing candidates for inclusion in a dictionary, which is tested in an industry-relevant context. Our system can retrieve usages for a given headword, remove usages of senses already recorded in a dictionary, propose unrecorded senses for inclusion into the dictionary, and suggest example usages for a given sense proposal. We further publish an annotated dataset for further development of USD systems.

2. Related Work

Erk (2006) gives the first systematic approach to automatic USD framing the problem as a binary classification task, where word usages that are not covered by any entry in a dictionary need to be assigned label 1 while covered usages need to be assigned label 0. Erk proposes two types of models, one relying on a WSD classifier exploiting word sense definitions from the dictionary, and one relying purely on word usages and an outlier detection model, which is more related to WSI because sense definitions are ignored. Both modeling directions have been pursued subsequently (e.g. Lautenschlager et al., 2024; Schlechtweg et al., 2025). Moreover, related tasks have been formulated, such as Novel Sense Detection (Lau et al., 2012; Cook et al., 2014) and Lexical Semantic Change Detection (Schlechtweg et al., 2020). These are more WSI-like as the assume no sense inventory. More recently, a shared task on USD was organized mixing aspects of WSD and WSI (AXOLOTL-24, Fedorova et al., 2024). Modern models for these tasks employ transformer-based language model architectures (Blevins and Zettlemoyer, 2020; Blevins et al., 2021; Barba et al., 2021; Eyal et al., 2022; Homskiy and Arefyev, 2022). In the field of electronic lexicography, several studies have explored related aspects of automatic dictionary updating such as neologism detection (Kerremans and Prokić, 2018; Annette Klosa, 2018; Martin, 2019). There are few studies attempting to automatically generate full dictionaries from scratch (Baisa et al., 2019; Jakubícek et al., 2021). Especially in lexicography for the low-resource languages automation and digital tools play an important role (Ogilvie, 2011; Lugli, 2019; Duijff and Kuip, 2018). More recent studies leverage large language models (LLMs) for different aspects of automatic dictionary updating (Hawkes et al., 2025; Widmann, 2025; Stöckle et al., 2025).

Our work employs the Outlier2Cluster method which was proposed in Kokosinskii et al. (2024) and was the best-performing method in the first subtask of AXOLOTL-24. In this subtask, competing methods were given a list of target words, a list of usages of each target word from a modern corpus, and an entry for this target word from an older dictionary. The entry describes senses of the target word listed in the old dictionary, each description consisting of a sense definition and in some cases a few example usages of this sense. Models had to annotate all modern usages with one of the provided senses or a new sense id if none of these senses fit. To achieve this, Outlier2Cluster solves three related NLP tasks. Among the provided sense definitions it selects the most suitable for each given usage, this task is known as WSD. Also it clusters all given usages by sense, which is known as WSI. Finally, for each given usage it estimates the probability that this usage does not fit any of the given senses (USD) and based on this probability selects either the WSD or WSI output as the final prediction for the usage. Technically, Outlier2Cluster builds an embedding for each usage and each sense definition using the usage encoder and the gloss encoder of the GlossReader WSD model (Rachinskiy and Arefyev, 2021) correspondingly.[2] For WSD the most similar gloss embedding is selected for each usage embedding, for WSI usage embeddings of each target word separately are clustered. Finally, for USD a logistic regression classifier is employed to estimate the probability that none of the given definitions fit the given usage based on the various distances from the gloss embedding to the usage embedding and a few additional features such as the number of senses in the dictionary entry for the target word.[3]

3. System Architecture

We believe that NLP methods can significantly help lexicographers in creating and updating dictionaries, but the creation of trustworthy, reliable dictionary entries, which are clear and intelligible to human readers, still requires the input of specialist lexicographers. Thus, we propose a system architecture that does not try to fully automate the lexicographic process but instead tries to assist a lexicographer with the more laborious and time-consuming aspects of this work.

Our system takes a dictionary, a set of sense proposals with corresponding headwords (single-

[2]More precisely, there are two GlossReader models, the original one and one fine-tuned on the AXOLOTL train sets in Finnish, Russian and English. For WSD and WSI the embeddings from the second one are used, while for USD distances between the embeddings from both models are used as feature to improve performance.

[3]Taking the Manhattan distance between the L1-normalized embeddings of the usage and the gloss selected during WSD serves as a simple but strong baseline almost approaching the best USD classifier.

id	lemma	context
u1	spam	In dramatic sequences, God of War might ask the player to **spam** "X" or twirl the control sticks to mimic the action happening on screen
u2	spam	click the "X" in the upper right corner of the comment box to report **spam** or abuse. We are using Facebook commenting. Visit our FAQ page for more information.
u3	spam	**Spam**, trout, fried chicken, moon pies and anything slathered in mayonnaise – those are some of the flavors of South Korea's home cooking that might seem just a bit familiar to the U.S.
u4	spam	For big, elaborate boss battles, Barlog said, players can expect the "Track and Field" design, referring to the classic NES game in which players quickly **spammed** buttons to create a feeling of physical exertion

Table 1: Set of *spam* usages extracted from NOW corpus, with simplified identifier, shortened usages, excluded columns and modified values for illustrative purposes.

id	lemma	PoS	gloss/description	source
s1	spam	N	irrelevant or unsolicited messages sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc.	ODE
s2	spam	N	unwanted or intrusive advertising on the internet	ODE
s3	spam	N	a tinned meat product made mainly from ham	ODE
s4	spam	V	send the same message indiscriminately to (a large number of internet users)	ODE
l1	spam	V	Slang. To press or strike (a computer key, button, etc.) many times in quick succession	LEMUR

Table 2: Existing dictionary entries for the headword *spam*, taken from the ODE dictionary and LEMUR sense proposal on the bottom.

or multi-word expressions, SWEs/MWEs), and a corpus as its inputs.[4] It aims to find usages for the sense proposals, but can also find usages of completely new senses for which no proposal exists. For newly discovered senses, our system drafts sense descriptions and finds example usages.[5]

The system consists of four main modules: **S0** retrieves usages of a given headword from a corpus. **S1** selects usages of senses that are not recorded in the dictionary (we will call them "unrecorded usages" for brevity). **S2** assigns usages to sense proposals and groups remaining ones by sense. **S3** drafts a sense definition for each group. Finally, proposed sense definition drafts can be potentially refined by a lexicographer. Find an overview diagram of the whole pipeline in Figure 3 in Appendix A.

3.1. S0: Usage Retrieval

At the first step, usages of each given headword are extracted from the corpus. More specifically, we find all occurrences of the given headwords in all grammatical forms and extract a text fragment of specified length around each occurrence. Find examples for the headword *spam* in Table 1.

For our experiments we use the News On the Web (NOW) corpus (Davies, 2016-) which consists of texts from online magazines and newspapers from 2010 up to 2025, containing around 40M texts and 23B words in total. The corpus provides PoS tags and lemmas for almost all tokens, which helps us find headwords in all grammatical forms efficiently by simple string comparison of lemmas against the query headwords. When a lemma is not specified for a token, we use its lowercased form instead.

While the usage retrieval task may look simple at a first glance, there are various important details to consider. News articles in the NOW corpus are scraped from the web and contain fragments of HTML. We remove HTML tags while preserving the inner text, and also replace entities like `<` with the corresponding characters. Since each document in NOW is represented as a list of tokens, one token with all its corresponding attributes per line, we do not know exactly if a space was present before each token in the original document or not when reassembling text fragments. Thus, we insert spaces before all tokens except for punctuation where none is needed, e.g. full stops, closing brackets, or apostrophes, to obtain as naturally looking text as possible. Quotation marks are grouped in pairs to determine start and end, by which unnecessary space can be omitted. One issue remaining is the censoring introduced in NOW to avoid copyright infringement, which replaces some tokens with the `@` sign.

Other challenges include retrieving usages of multi-word expressions and suffixes. During matching, MWE tokens are merged. If they contain a placeholder, any token can be matched in its place. Suffixes also need to be matched as such, that is, as suffix of the token it is compared to.

To keep the output size and processing time within reasonable limits, for each query headword we randomly sample 10K usages or return all usages if there are fewer, and only export a 150 token context around the target word. Additionally, we remove duplicated usages but store the number of duplicates of each usage.[6]

[4] We assume correspondence of headwords between proposals and dictionary, and between headwords and corpus lemma forms.

[5] Find our code at https://github.com/Garrafao/apodictus.

[6] This will allow estimating sense frequencies with and without duplicates in future versions of our system.

3.2. S1: Unrecorded Usage Detection

At this step, usages of those senses that are described in the dictionary are removed and unrecorded usages are passed to the following steps. This step minimizes the number of irrelevant usages for the following ones. This allows sense proposal generation at S3 to focus on unrecorded senses and reduces the number of false positives when mining example usages for the proposed sense definitions of unrecorded senses at S2. As an example, compare the usages for the headword *spam* in Table 1 to the dictionary entries in Table 2. Our aim is to map usage u2 to s1 and u3 to s3 in order to filter them out.

Unrecorded usage detection is implemented using Outlier2Cluster (see Section 2). For each headword, we feed all of its usages retrieved by S0 and all sense definitions from the dictionary.[7] Then we return usages with USD scores above a certain threshold *T1*, i.e., those usages that do not fit well any of the given definitions. For headwords without existing dictionary entries (out-of-dict) we skip filtering and return all usages.

3.3. S2: Proposal Usage Detection

At this step we aim to assign usages to the sense definition proposals. As an example, compare again Tables 1 and 2 and assume u2 and u3 have been filtered out in S1. At this stage, we want to map u1 and u4 to l1, the sense proposal.

Implementation-wise, similarly to S1 we employ Outlier2Cluster but this time giving to it only one definition. For each usage survived after S1 we get an USD score showing how well this usage fits the sense definition proposal. All usages with USD scores below a specific threshold *T2* are considered good example usages and returned from this step. This step is optional, i.e., if the lexicographer has no sense proposals available this step is skipped.

3.4. S3: Sense Proposal Generation

Unrecorded usages identified at the previous step are clustered by sense using the WSI model of Outlier2Cluster. Then for each cluster its usages are passed to a model that generates a sense definition proposal. For the example in Tables 1 and 2, assuming the model works perfectly, no usage is detected as unrecorded/survives S1+S2, and S3 would be skipped. If instead any cluster of unrecorded usages was detected, we would generate a new sense definition for this cluster.

We developed a model named *Sense Definition Generation from Usages* (SDG_{usages}), which employs the *Gemma3-12b-it* (Team et al., 2025) LLM through the *VLLM* library (Kwon et al., 2023) to enable fast and efficient inference. This LLM was selected due to its large context window of 128K tokens and support of over 140 languages. A large context window facilitates information extraction from many word usages in order to generate better sense definitions. We take all usages from a single cluster and insert `"NEXT SENTENCE:"` before every usage except for the first one, then concatenate and embed them into a prompt. Model hyperparameter settings are given in Appendix B.

4. Data

LEMUR is an internal database maintained by Oxford University Press, containing sense proposals that editors will then score to decide whether to add them to the dictionary or not. We were provided a total of 1,300 sense proposals to predict for. These came as CSV files with one line per proposal containing information such as the headword, the PoS, the preliminary definition and whether the word was already recorded in the Oxford Dictionary of English (ODE).

Further, we were provided with XML files containing the related entries from the ODE. These entries contain the headword, PoS, sense definitions, and usage examples.

NOW is a large English text corpus compiled from online newspapers and magazines covering the period from 2010 to the present (Davies, 2016-, 2025). The corpus is regularly updated with approximately 250 million words added each month. We acquired access to the NOW corpus for the years 2010-2024 and use it for retrieving usages of the target words associated with the LEMUR sense proposals, which we then search for relevant evidence.

4.1. Test Dataset

To evaluate the performance of our system, excluding the usage retrieval step (S0), we built a test set consisting of 2,746 annotated usages for 47 different words. Its construction involved the sampling of usages and their annotation with dictionary senses and sense proposals.[8]

[7] For our experiments the Oxford Dictionary of English (ODE) is employed. However, the system can use any dictionary that can provide a list of sense definitions for each headword, possibly an empty list for out-of-dict headwords.

[8] Find the data at https://github.com/Garrafao/apodictus.

4.1.1. Sampling of Headwords and Usages

We randomly sampled 24 **in-dict** LEMUR headwords which have an entry in ODE and 24 **out-of-dict** words which do not have such an entry. We calculate metrics on these two subsets separately because we observed differences in performance in previous executions of the pipeline, with precision and recall being lower for in-dict words.

For each headword we sampled a random subset of usages from those retrieved by S0. The key difference between the two types of headwords is that for the in-dict headwords one or several more frequent senses are usually already recorded in the dictionary and the unrecorded senses are the more rare ones. Thus, usages of unrecorded senses rarely occur and to catch unrecorded senses we have to sample and annotate more usages than for the out-of-dict headwords which have no recorded senses at all. To balance annotation efforts we sampled 100 and 30 usages for the in-dict and out-of-dict headwords respectively.

4.1.2. Annotation Process

The annotation was performed by two external annotators who are native English speakers. For each sampled headword separately we provided its sampled usages and a list of sense descriptions from both ODE and LEMUR. We did not show the source of each sense description to the annotators explicitly, however, in some cases they could guess about the source based on stylistic differences. For each sense its description contains its level in the hierarchy of senses (main sense / subsense), a part-of-speech tag, a gloss, and example usages when available.

Annotators had to assign each usage to one or more of the listed sense descriptions. If more than one sense matched, they should all be listed, in order of relevance. If no sense matched, it had to be added to the dictionary first, along with a description. Usages where the annotator was unsure should be marked with '0' and corrupted usages with 'x'.

After inspection of the annotations, one usage was removed due to a missing annotation, and one in-dict word was removed entirely because all its annotations were corrupted. Table 3 provides an overview of the resulting dataset.

4.1.3. Annotator Agreement

In order to calculate inter-annotator agreement and control the annotation quality, we sent one common in-dict headword with 100 usages to both annotators, and also annotated it ourselves. Based on all three annotations, we calculate four types of Krippendorff's α scores (Krippendorff, 2018).

Annotation	in-dict	out-of-dict	Total
Total	2177	569	**2746**
ODE Sense	1406	0	**1406**
LEMUR Sense	70	305	**375**
Unrecorded Sense	636	255	**891**
0	62	8	**70**
x	3	1	**4**

Table 3: Test dataset composition.

When multiple senses are annotated for a usage, we only use the first one, which is supposed to be the most relevant sense according to our annotation guidelines.

Agreement on exact sense assignments ($\alpha_s = 0.693$) and at the main-sense level ($\alpha_m = 0.888$), where subsenses of the same main sense are treated as equivalent, was generally high. These metrics were computed only for usages where all annotators assigned a recorded sense. In contrast, agreement on whether a usage is recorded or unrecorded in the first place was low ($\alpha_r = 0.228$), reducing overall quality. However, since our evaluation focuses on correctly identifying LEMUR senses, the relatively high agreement on whether a usage expresses a LEMUR sense ($\alpha_p = 0.721$) is sufficient to provide a reliable basis.

5. System Development

5.1. S0: Usage Retrieval

Evaluation of our usage retrieval is based on 60 headwords, 30 SWEs and 30 MWEs sampled randomly from LEMUR, and usages retrieved from the NOW corpus for the years 2020 to 2024.

Estimating recall of S0 is difficult due to the missing information regarding how many usages are available in NOW for each headword. Therefore, we compare the number of usages retrieved by S0 with the number of usages returned by the web interface of NOW, and estimate recall under an assumption that the web interface returns all usages matching each headword and only them. Estimated this way, the median recall across the 60 headwords is 94.2%. Copyright token replacement restricts recall of S0 on the downloaded corpus: Words that occur in the censored sections are found via the web interface, but are impossible to locate in the downloaded data.

To measure precision, for each of the 60 headwords up to 5 usages were randomly sampled and inspected. In total 228 usages were inspected as some headwords had less than 5 usages retrieved. We did not find any usages that do not match the corresponding query, thus, our estimate of precision for S0 is 100%.

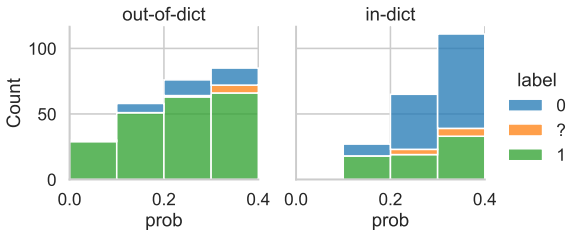

Figure 1: Quality control annotation results.

Error Type	Affected False Positives
All Errors (False Positives)	156
Loose Lexical or Semantic Overlap	61 (39.1%)
Problematic Definition	47 (30.13%)
PoS Mismatch	19 (12.2%)
Corpus Artifacts and Corruption	18 (11.5%)

Table 4: Error categories.

5.2. S1/S2: Unrecorded and Proposal Usage Detection

To provide a basis for system development and assessing our system's performance, we ran it on all 1,300 LEMUR sense proposals available to us. Usage retrieval (S0) was performed on the entire NOW corpus, limiting the number of usages per word to 10,000. For the subsequent steps S1 and S2, we initially selected the thresholds $T1 = 0.19$ and $T2 = 0.4$, based on observations from previous experiments on a subset of annotated word usages. For the USD logistic regression classifier within the Outlier2Cluster method, we employed a set of weights we call **Own USD Weights**. They were trained using a small but specialized train set of two in-dict headwords from ODE with 50 usages retrieved from NOW for each headword.

5.2.1. QC Annotation

To estimate precision, among the headwords having some example usages found we sampled 15 **in-dict** headwords and 15 **out-of-dict** headwords. From the outputs at S2, we randomly sampled up to 10 usages for each probability score bin $\in \{[0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4]\}$ for each chosen headword. Headwords without any usages in these score bins were ignored during sampling. Each example usage predicted for a sense proposal was annotated as '1' if it actually matched the sense, '0' if it did not, and '?' if the annotator was unsure. In total, 457 usages were annotated.

Figure 1 shows the results separately for in-dict and out-of-dict words. As these are S2 outputs, LEMUR usages (in green) would ideally be assigned low USD probabilities, reflecting high similarity to the corresponding LEMUR sense, while all other usages would receive high probabilities, allowing a clear separation between LEMUR and non-LEMUR usages. Although a high proportion of LEMUR usages do occur at low probabilities in our QC annotations, indicating generally reliable predictions, a number of undesirable assignments across bins remain. This problem is particularly pronounced for in-dict words, where LEMUR usages are intermixed with others across all probability bins, reducing precision. To prioritize precision over recall, we lowered the threshold T2 from 0.4 to 0.2 for the main evaluation which should lead to sufficiently high precision according to the annotation.

5.2.2. Error Analysis

We conducted an error analysis of the 156 false positive predictions from the quality control annotations, observed with the initially chosen thresholds ($T1 = 0.19$ and $T2 = 0.4$). Four categories of issues were identified as potential contributors to these errors, shown in Table 4.

Loose Lexical or Semantic Overlap Even when the target word does not match the LEMUR sense, the surrounding context may include words or phrases that overlap with the sense description, either lexically or semantically, which can lead the model to mislabel, as in the following example:

Word	Description
bodge (verb)	To turn (a chair-leg, etc.) on a lathe.
I can **bodge** some bits of **wood** together the same as anyone who has spent their life in theatre, but I'd never be able to come close to the skill required to be a master **carpenter**.	

Problematic Sense Descriptions While LEMUR sense descriptions are clear for humans, some can still challenge language models. Very short descriptions, or descriptions taking the form of context indicators rather than true glosses or definitions (as in the case of *dust ribbon* below), may lead to overgeneralization, while descriptions with unusual formatting or many special characters can make it hard for the model to interpret the content. Consider the following example:

Word	Description
dust ribbon	weather
sticker	A person who posts bills, posters, etc.; = STICKER-UP n. \\Cf. 'bill sticker' 'advertisement sticker'

PoS Mismatch Since we do not filter usages based on part of speech, instances of the target word with the wrong PoS may be matched to a LEMUR sense, even though the PoS would immediately disqualify them.[9]. This is e.g. the case for the following usage of *postgraduate*:

[9]We did not apply such a filter because we were unsure about the accuracy of automatic corpus POS tagging, especially for new creative word forms or abbreviations.

Word	PoS	Gloss
postgraduate	verb	To complete a postgraduate course; to obtain a postgraduate qualification.
During his **postgraduate** days in the United Kingdom		

Corpus Artifacts Usage examples retrieved from the NOW corpus can contain artifacts that may interfere with the model's interpretation of the text. For instance, "@" symbols are sometimes used to censor portions of text for copyright reasons. Because the NOW corpus is compiled from online articles, remnants such as placeholders for advertisements or comment sections may also appear. Consider these examples:

... What does the shortage of **@ @ @ @ @ @ @ @ @** billion promo industry? MV- I think what I am saying is ...
... to wipe out malaria in Kenya. **ADVERTISEMENT ADVERTISEMENT** Currently, the world is largely embroiled in one of the greatest health emergencies ...

6. Main Evaluation Results

We now describe the evaluation of the main pipeline components on the test set described in Section 4.1. We start with S2 as this step finds evidence for sense proposals which is our main focus task.

6.1. S2: Proposal Usage Detection

For this subtask, our main target is finding good example usages for the sense proposals from LEMUR from which a lexicographer can easily select a few to include into a dictionary. For many headwords there are no usages of the LEMUR senses in the test set and maybe even in our whole corpus. This is especially pronounced for the headwords already described in the dictionary because mostly rare senses are unrecorded for such headwords. For instance, our test set has any usages of the LEMUR senses only for 9 out of 23 in-dict headwords.

Thus, we cannot require that the system returns example usages for each sense proposal. But we want at least to reduce the proportion of irrelevant example usages returned, i.e., when some example usages are returned for a sense most of them should correspond to this sense well. Thus, our main metric is the **macro-averaged precision** of the mined example usages. Specifically, we calculate precision for each LEMUR sense proposal with at least one usage returned, then average across all such proposals. This corresponds well to a scenario when there is a huge set of sense proposals, such as the LEMUR database with tens of thousands of proposals, and not enough human resources to look through usages found for each of them. In this case, a system that returns only relevant usages even for a small subset of proposals is preferable. Of course, we also report **coverage** which is the number of sense proposals with some example usages returned as the secondary metric. On the other hand, a typical dictionary entry contains only a few example usages. Thus, we are less interested in returning all usages of a sense proposal and recall is the least important metric for us.

subset	macroP	coverage
out-of-dict (18/24)	0.84	10/24
in-dict (9/24)	0.25	4/24
TOTAL (27/48)	0.67	14/48

Table 5: Macro precision and coverage for the LEMUR sense proposals on the test set (in brackets we show the proportions of sense proposals having at least one usage in the test set).

Table 5 reports the metrics for our system with the hyperparameters selected in Section 5 on the test set described in Section 4. For the out-of-dict headwords, the system returned example usages for 10 proposals (in the test set 18 out of 24 proposals do have usages) and the average precision is 84%, so false positives are rare. For the in-dict headwords the results are much worse, but it is worth noticing that the macro-precision is less reliable as we average across 4 proposals only.

6.2. S3 Evaluation: Proposal Generation Experiments

We inspected a sample of 19 of the 161 sense definitions generated for the test set from Section 4.1.[10] 6 had bad definitions while 13 were rated acceptable or good. From the 6 bad definitions, 2 were due to heterogeneous usages in the cluster mixing different senses, such as the following example for the word *Titan*:

Word	Generated gloss
Titan	A large natural satellite of a planet, especially the moon of Saturn.
While this Friday will see the second season of the **Titans** come to end...	
Manuel, **titan** of the scrum, scored in the 26th minute with Kemp doing his duty for 17-10 to the Kaupoi	

The example usages from the cluster given above have different meanings and the generated definition fits neither of them. But, also if usages are homogeneous within a cluster the generated definition may not fit as in the following example for the word *hale*:

Word	Generated gloss
hale	A person known for vigor and health.
...which might interest Justice **Hale**, a former family law specialist	
Hale threw for 2,047 yards with 17 touchdowns...	
...the early days of the pandemic, says **Hale**.	

[10] OUP lexicographers have not yet evaluated these definitions.

The definition does not fit, but it picks up the person aspect found amongst the usages and mixes it with aspects from the established adjectival sense of *hale* 'strong and healthy'. One usage also covers the health care system which may have an additional influence.

Many of the clusters contain mostly or only proper noun usages of the target word. But, the model does in some cases generate fitting descriptions for these, such as for *acker*:

Word	Generated gloss
acker	Kathy Acker was an American writer known for her experimental novels and postmodern style.
In many ways Kraus is **Acker**'s ideal biographer.	
...the last page of Kathy **Acker**'s final published book...	

Further well-fitting generated definitions include those for the words *funk*, *sideway*, *gold flakes* and *gravity bong*. Consider the following examples:

Word	Generated gloss
funk	a unique and unconventional wrestling style characterized by scrambling positions and unexpected maneuvers.
Askren reinvented folkstyle wrestling. The 174-pounder wasn't strong or fast enough to beat opponents by conventional measures. He had to create his own style. Some called it **funk**.	

Word	Generated gloss
sideway	Moving or positioned to one side; oblique.
...and threw condescending **sideway** looks at me...	
...we still expect KLCI to trend **sideway**...	

Word	Generated gloss
gold flakes	Thin flakes of gold used as a decorative or culinary garnish.
...and infused with Champagne and dusted with 24-karat edible **gold flakes**...	
These coasters are made with real flowers and sprinkled with **gold flakes**...	

Word	Generated gloss
gravity bong	A smoking device that uses water and the principles of gravity to filter and cool cannabis smoke.
...the birthday honoree is preparing to fire up a St? ndenglass **gravity bong**...	
...come sit back down at your table and hit a bong or a **gravity bong** or a dab rig...	

Of these, the generated senses of *funk*, *sideway* and *gold flakes* are currently not in ODE or LEMUR and could be considered for inclusion by OUP lexicographers.

7. S1/S2: Ablation Analysis

In this section, we study if removing usages of ODE senses at S1 helps improve precision when mining example usages of LEMUR senses at S2. Also, we compare the existing USD models with our specialized one on the test dataset. For each USD model, we perform a grid search over $(T1, T2)$ threshold pairs varying each threshold between 0 and 1 in steps of 0.01 and computing the corresponding macro precision and coverage. For the baselines that return scores outside of this range, we first normalized their USD scores using the quantile transformation. Finally, we plot a curve showing the best achievable macro precision for each coverage value observed in the grid search.

7.1. USD Models

Outlier2Cluster employs a logistic regression classifier to estimate the probability that a usage does not fit any of the given sense definitions. In Kokosinskii et al. (2024) two sets of weights were trained on the Russian and the Finnish development sets from the AXOLOTL shared task (see Section 2). We will call them **Russian USD weights** and **Finnish USD weights** respectively. In addition to these two sets of weights, we also trained our own (**Own USD Weights**) as already shortly described in Section 5.2.

Besides the three logistic regression models, we also evaluated the baseline USD models from Kokosinskii et al. (2024) which do not use logistic regression but instead calculate various distances between usage and gloss embeddings and return them as USD scores.

7.2. LEMUR Proposals

Figure 2 presents line plots illustrating the results separately for in-dict and out-of-dict words. For readability we show only two baselines on the plots, the best one **norm_l1** which is the Manhattan distance between L1-normalized embeddings, and also the popular **euclidean** distance. First we compare models for the LEMUR senses (two top subplots). It is very clear that without removing irrelevant usages at S1 we cannot achieve any reasonable precision for the in-dict headwords.

Comparing USD models on the out-of-dict headwords, our specialized model clearly outperforms other models for the high-precision scenario. E.g. we cover up to 10 out of 18 unrecorded senses with precision above 80%. The second-best model is the norm_l1 baseline. For the in-dict words it is hard to draw reliable conclusions from the plots. This is because the in-dict subset includes only nine words having some LEMUR usages and for our high-precision scenario we have to limit coverage resulting in only 3-4 proposals to estimate macro precision from.

7.3. Proposals from the Test Set

To address the small number of LEMUR usages in the test set, especially for the in-dict headwords, we additionally compared models on the sense proposals from the annotators of the test set. Only headwords with at least one sense proposal usage were selected giving 21 in-dict and 15 out-of-dict

Figure 2: Macro precision (y-axis) vs. coverage (x-axis).

headwords, with 2374 usages in total. To estimate the effect of sense frequency, we separately mined example usages for the least frequent (**LFS**) and the most frequent (**MFS**) sense proposals of these headwords.

For the in-dict headwords and both their least and most frequent unrecorded sense proposals norm_l1 is evidently the best model, while our specialized USD model is among the worst ones. For the out-of-dict headwords the results are less clear, norm_l1 is still among the best performers but the highest precision is achieved by classifiers that are different for the high-precision and the high-coverage scenarios, and also for MFS and LFS.

Overall, no single model clearly outperforms the others across all scenarios. Further improvements of our system performance could likely be achieved by tuning hyperparameters and selecting models individually for different types of headwords and sense proposals.

8. Conclusion

We described the first systematic, large-scale and comprehensive solution for processing candidates for inclusion in a dictionary. Our pipeline first extracts corpus usages, then filters out recorded usages and identifies evidence of sense proposals, and finally generates sense proposals for completely unrecorded senses. We conducted several experiments and annotations to tune and adjust our models. An error analysis showed that loose lexical overlap between usage and proposal as well as short or imprecise sense definition proposals are major factors in prediction errors. Precision and coverage are acceptable for words which are not in the dictionary, but low for in-dictionary words. An ablation and comparison of model components showed that a filtering step for recorded senses is essential for performance, especially with in-dictionary words. Further experiments on additional data suggest that model modifications could improve performance, but show no clear pattern. A manual analysis of the definition generation component showed that the majority of definitions have acceptable quality.[11] Overall, we conclude that finding unrecorded senses is a very challenging problem requiring the processing and filtering of large amounts of data to find the needle in the haystack.

We applied the final pipeline with optimized parameters to predict evidence usages for the full 1,300 LEMUR entries and provided these to OUP lexicographers for evaluation. In the future, we would like to (i) improve the pipeline and (ii) focus on sense frequency estimation for better prioritization. For (i), we see potential in model selection, fine-tuning neural network models on the target data, POS filtering and LEMUR entry modification or finetuning. For (ii), we see potential in weighting frequencies according to correct prediction proportions from our human evaluation.

Limitations

The usage retrieval part of our pipeline is constrained by the corpus we use, which contains only online news content. Further text sources, such as social media posts should be included. Further, we perform no evaluation of our WSI component although heterogeneous clusters are one reason for generated definitions with low quality.

In the present study, we focused on the English language. Finding usages of rare new senses requires the availability of large representative and recent corpora, which may not be available for less-resourced languages. Also, model performance may further decrease for such languages.

Acknowledgements

We thank Oxford University Press and the lexicographers from Oxford Languages for providing the necessary data, explanation and interpretation to conduct this study. We are particularly grateful to Kate Wild and Iona Ogilvie for their continuous support and helpful feedback to the paper draft. This paper is based on the bachelor theses of Felix

[11] OUP lexicographers were not involved in the analysis of the output.

Blessing, Johannes S. Sax and Julian Kaufmann (Blessing, 2025; Sax, 2025; Kaufmann, 2025).

9. Bibliographical References

Harald Lüngen Annette Klosa. 2018. New german words: Detection and description. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 559–569, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Vít Baisa, Marek Blahus, Michal Cukr, Ondrej Herman, Miloš Jakubíček, Vojtěch Kovár, Marek Medved, Michal Boleslav Mechura, P. Rychlý, and Vít Suchomel. 2019. Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. Brno, Czech Republic: Lexical Computing CZ s.r.o.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Felix Blessing. 2025. Automatic prioritization of dictionary update candidates. Bachelor thesis, University of Stuttgart.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 1624–1635, Dublin, Ireland.

Mark Davies. 2016-. Corpus of news on the web (now). Available online at https://www.english-corpora.org/now/.

Mark Davies. 2025. Full-text corpus data. https://www.corpusdata.org/now_corpus.asp. [Accessed 2025-08-23].

Pieter Duijff and Frits van der Kuip. 2018. Lexicography in a minority language: A multifunctional online Dutch-Frisian dictionary. *International Journal of Lexicography*, 31(2):196–213.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 128–135, New York City, USA. Association for Computational Linguistics.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.

Elinor Hawkes, Phoebe Nicholson, and Will Rogers. 2025. Making sense of the past: AI-assisted historical word sense disambiguation and the OED. In *Electronic lexicography in the 21st century*.

Daniil Homskiy and Nikolay Arefyev. 2022. DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland. Association for Computational Linguistics.

Milos Jakubícek, Vojtech Kovár, and Pavel Rychlý. 2021. Million-click dictionary: Tools and methods for automatic dictionary drafting and post-editing. *Book of Abstracts of the 19th EURALEX International Congress*, pages 65–67.

Julian Kaufmann. 2025. Usage retrieval for dictionary headwords with applications in unknown sense detection. Bachelor thesis, University of Stuttgart.

Daphné Kerremans and Jelena Prokić. 2018. Mining the web for new words: Semi-automatic neologism identification with the neocrawler. *Anglia*, 136(2):239–268.

Denis Kokosinskii, Mikhail Kuklin, and Nikolay Arefyev. 2024. Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 168–179, Bangkok, Thailand. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Stroudsburg, PA, USA.

Jonathan Lautenschlager, Simon Hengchen, Emma Sköldberg, and Dominik Schlechtweg. 2024. Detection of non-recorded word senses in English and Swedish.

Ligeia Lugli. 2019. Smart lexicography for low-resource languages: Lessons learned from buddhist sanskrit and classical tibetan. In *Proceedings of ELex*.

Katherine Connor Martin. 2019. New words prioritization engine: A system for evaluating multiple data inputs to prioritize neologisms for inclusion in dictionary projects. Oxford University Press.

Sarah Ogilvie. 2011. Linguistics, Lexicography, and the Revitalization of Endangered Languages. *International Journal of Lexicography*, 24(4):389–404.

Maxim Rachinskiy and Nikolay Arefyev. 2021. GlossReader at SemEval-2021 task 2: Reading definitions improves contextualized word embeddings. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 756–762, Online. Association for Computational Linguistics.

Michael Rundell. 2023. Automating the creation of dictionaries: are we nearly there? In *Proceedings of AsiaLex*, Seoul, Korea.

Johannes Sax. 2025. Sense definition generation and how it can improve wsd. Bachelor thesis, University of Stuttgart.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg, Emma Sköldberg, Shafqat Mumtaz Virk, James White, and Simon Hengchen. 2025. Automatic non-recorded sense detection for Swedish through word sense induction with fine-tuned Word-in-Context models. In *Electronic lexicography in the 21st century*.

Philipp Stöckle, Wolfgang Koppensteiner Daniel Elsner, and Katharina Korecky-Kröll. 2025. LLM-assisted dialect lexicography: Challenges and opportunities in processing historical Bavarian dialects. In *Electronic lexicography in the 21st century*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron,

Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Thomas Widmann. 2025. A pipeline for automated dictionary creation with optional human intervention. In *Electronic lexicography in the 21st century*.

A. System diagram

Find a schematic overview of our system in Figure 3.

B. Definition generation hyperparameters

Table 6 gives the hyperparameters we used for sense definition generation.

Parameter	Value
temperature	0.6
top_p	0.1
top_k	100
presence_penalty	0.7
frequency_penalty	0.5
max_tokens	60
seed	192

Table 6: LLM hyperparameters.

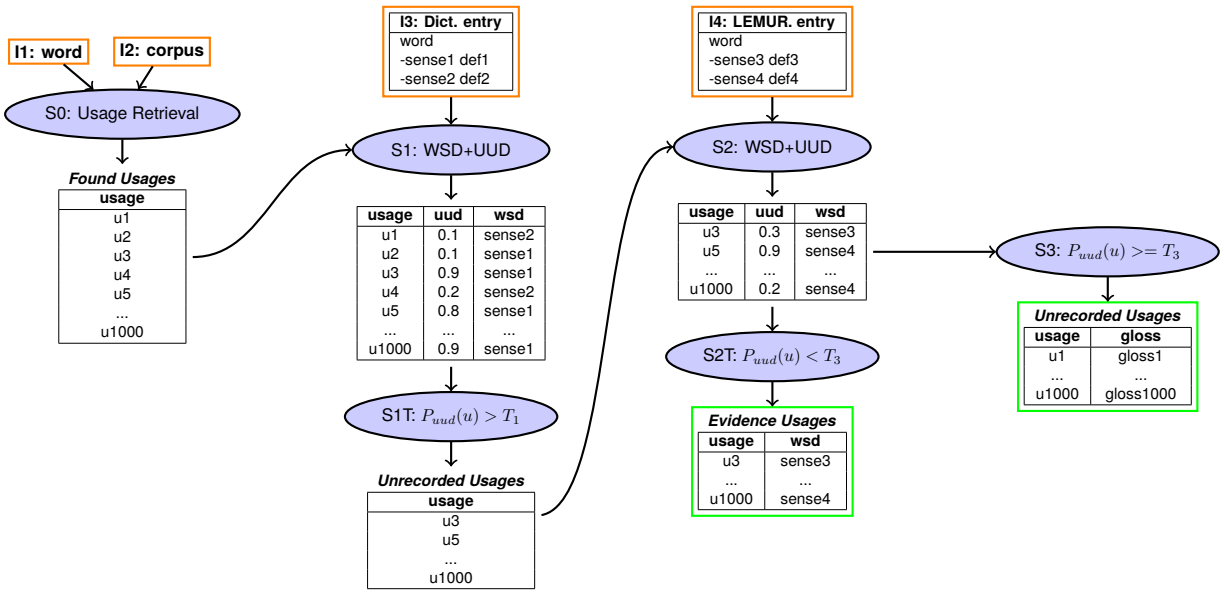

Figure 3: System pipeline diagram.