# Language Pivoting from Parallel Corpora for Word Sense Disambiguation of Historical Languages: a Case Study on Latin

**Iacopo Ghinassi**[1], **Simone Tedeschi**[2,3], **Paola Marongiu**[4], **Roberto Navigli**[2], **Barbara McGillivray**[5]
[1]Queen Mary University of London , [2]Sapienza University of Rome,
[3]Babelscape, [4]University of Neuchâtel, [5]King's College London
i.ghinassi@qmul.ac.uk, paola.marongiu@unine.ch, {tedeschi, navigli}@diag.uniroma1.it,
barbara.mcgillivray@kcl.ac.uk

## Abstract

Word Sense Disambiguation (WSD) is an important task in NLP, which serves the purpose of automatically disambiguating a polysemous word with its most likely sense in context. Recent studies have advanced the state of the art in this task, but most of the work has been carried out on contemporary English or other modern languages, leaving challenges posed by low-resource languages and diachronic change open. Although the problem with low-resource languages has recently been mitigated by using existing multilingual resources to propagate otherwise expensive annotations from English to other languages, such techniques have hitherto not been applied to historical languages such as Latin. In this work, we make the following two major contributions. First, we test such a strategy on a historical language and propose a new approach in this framework which makes use of existing bilingual corpora instead of native English datasets. Second, we fine-tune a Latin WSD model on the data produced and achieve state-of-the-art results on a standard benchmark for the task. Finally, we release the dataset generated with our approach, which is the largest dataset for Latin WSD to date. This work opens the door to further research, as our approach can be used for different historical and, generally, under-resourced languages.

**Keywords:** Word Sense Disambiguation, Less-Resourced/Endangered Languages, Digital Humanities, Corpus (Creation, Annotation, etc.)

## 1. Introduction

Word Sense Disambiguation (WSD) is a key task whose objective is to assign the correct meaning to a target word in context (Navigli, 2009). Apart from practical uses of the task in NLP, identifying the correct senses for given polysemous words can greatly help the field of computational historical linguistics and advance our understanding of diachronic semantics (McGillivray, 2020). In this context, Latin is in a particularly favourable position among historical languages, as we have more than two thousand years of uninterrupted recorded documentation for it. Nevertheless, Latin is a case of an under-resourced language, as the availability of annotated data for NLP tasks for this language is much more limited compared to many modern languages (Passarotti et al., 2020). To address the issue of data scarcity, recent literature has shown that pivoting a low-resource language to a high-resource one such as English via parallel corpora is an effective strategy for efficiently producing WSD annotations in the under-resourced language (Pasini et al., 2021). No such attempt, however, has been made for Latin data. To fill this gap, in this work we introduce a simple yet effective, language-agnostic sense propagation framework and evaluate it on Latin data. By doing so, we make three major contributions. First, using Latin as a case study, we demonstrate how this general framework can be employed effectively for a historical language for which a great number of parallel corpora are available, but for which few annotated data exist. Second, by employing the dataset produced with our methodology as auxiliary training data for the task, we achieve state-of-the-art results on an existing benchmark for Latin WSD. Finally, we publicly release the dataset produced – which is the largest dataset for Latin WSD to date – and model checkpoints at `https://github.com/Ighina/LatinWSD`. We hope to encourage further research in this direction as our approach can be used for different historical languages.

## 2. Related Work

While NLP research has developed sophisticated WSD systems for contemporary and high-resourced languages (Bevilacqua et al., 2021), only very few attempts have been made so far to develop WSD methods on historical texts, including Bamman and Crane (2009); Bamman and Burns (2020); Lendvai and Wick (2022) on Latin, Beelen et al. (2021) and Manjavacas Arévalo and Fonteyn (2022) on 19th-century English. All these systems use the mapping between historical dictionary senses and related quotations to train automatic classification systems. The classification is based on contextualised embeddings such as BERT embeddings. All three previous Latin WSD

*LREC-COLING 2024*, pages 10073–10084
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

methods led to promising results, reaching an average macro F-score between 70% and 80%. However, these experiments only dealt with the first two macro-senses of a lemma (Bamman and Burns (2020) in the Lewis & Short dictionary and Lendvai and Wick (2022) in the Thesaurus Linguae Latinae), and their analysis only covered the eras up to Late Latin.

Most of the problems which WSD systems face in a language like Latin involve the scarcity of the resources required to train and test such systems properly. This problem is also shared by most languages other than English, as existing evaluation benchmarks are focused primarily on English (see, for example, the SemEval tasks during the years (Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2013; Moro and Navigli, 2015)). To overcome the problem of annotated datasets for WSD, previous studies proposed a number of solutions, such as automatic methods for producing sense distributions (Pasini et al., 2020), automatic creation of datasets in a target language by using high-performing WSD systems and multilingual resources on English translations of the target language (Pasini et al., 2021) or sense label propagation (Barba et al., 2020; Procopio et al., 2021). The approach of propagating annotations from high-resource languages to lower-resource ones is called language pivoting and it has been used for a variety of different tasks, starting with machine translation (Wu and Wang, 2007).

Crucially, no attempt to apply the language pivoting framework to historical languages exists and, thus, hitherto there has been no proof that such a method can work with the specific problems and resource requirements that such languages present. In this work, then, we aim to fill this gap.

## 3. Methodology

Most of the issues concerning the Latin language in the context of WSD, as highlighted in the previous section, are due to the scarcity of resources. In this section, we present our approach to addressing this shortage by automatically transferring WSD annotations from English to Latin.

### 3.1. Data Collection

In order to perform annotation propagation we used a parallel corpus of Latin texts aligned to their English translations and an inventory of 40 polysemous lemmas with their Part of Speech (PoS), each linked to their possible candidate synsets.

The source data for the parallel corpus was provided by the work carried out by Yousef and Berti (2015) for automatically building a bilingual lexicon Ancient Greek/Latin, namely the Dynamic Lexicon[1], a project supported by the Perseus Digital Library. The lexicon was built by using parallel corpora of Ancient Greek and Latin texts aligned to their English translations, and exploiting English as a bridge language to get the pairs of Ancient Greek and Latin lexical items. Starting from the two parallel corpora, Yousef and Berti (2015) automatically aligned the texts at both sentence and word level. The sentences were aligned with Moore's Bilingual Sentence Aligner (Moore, 2002). Word alignment was performed on the previously aligned sentences by using the Giza++ toolkit (Och and Ney, 2003). The parallel corpus contained 123 thousand sentence pairs and 2.33 million Latin words. The systems with which the corpus has been aligned are quite old, and therefore, since our approach relies entirely on alignment, there is considerable room for improvement. However, gold data is lacking to train sentence and word alignment systems from English to Latin, so we leave this as future work.

For the inventory of lemmas, we used the dictionary definitions of the 40 lemmas from the SemEval dataset (described in section 4.1). We focused on these lemmas as they were the only ones included in our reference gold-standard benchmark, which remains the only expertly-annotated dataset for word-in-context in Latin (see section 4.1 for further details). Each of these lemmas was assigned its PoS, and then each sense was manually mapped onto its corresponding synset in version 3.0 of Princeton WordNet (PWN) (Fellbaum, 1998).

To perform the manual linking between the word senses and WordNet senses we started with the dataset provided by the LiLa project[2], which contains a sample of 10,314 lemmas from Latin WordNet (LWN)[3]. In this dataset the synsets assigned to each lemma are manually checked and corrected, if necessary. The offset of version 3.0 of PWN is provided for each synset. For the lemmas not covered by the LiLa dataset we used LWN, which uses offsets from version 1.6, and converted the offsets to version 3.0 of the PWN. When we could not find the synset in either LWN or the LiLa dataset, we looked for the most suitable synset directly in PWN 3.0. For four lemmas it was not always possible to map all the senses of each lemma to PWN and the lemmas were excluded from the annotation propagation phase: *consul*, *sacramentum*, *templum* with the sense ‘a space marked out, an open place for observation’, and *virtus* with the senses ‘Virtue, personified as a deity’ and ’Christian virtue’. The nouns *consul* and *sacramentum* have very specific senses, which refer to concepts or institutions related to the ancient world for which we could not find any corresponding synset. *Consul* is associ-

[1] http://www.dynamiclexicon.com/

[2] https://lila-erc.eu/

[3] https://latinwordnet.exeter.ac.uk/

ated with the meanings 'consul' (a high magistrate in ancient Rome) and 'epithet of Jupiter', among others. Both meanings do not exist in PWN nor in LWN (Minozzi, 2017; Franzini et al., 2019; Biagetti et al., 2021), therefore they could not be mapped onto any synset. Same for the sense 'military oath of allegiance' of *sacramentum*.

### 3.2. Pre-processing

Finding the correct lemmas in the parallel corpus requires lemmatization and PoS tagging. To perform this step, we used the Cracovia lemmatizer and PoS tagger (Wróbel and Nowak, 2022), two separate state-of-the-art BERT-based models which won the EvaLatin-2022 challenge evaluating Latin lemmatization and PoS tagging on temporally diverse benchmarks (Sprugnoli et al., 2022). In both cases, the original implementations were used.

### 3.3. Annotation Propagation

As a result of the steps described in sections 3.1 and 3.2, we obtained an inventory mapping each of the above-mentioned 40 lemmas to their candidate synsets and a collection of parallel sentences in which such lemmas occur. In this section, we describe two variants of our annotation propagation strategy.

$Propagation_{w/inter}$. Let $I = \{(k, v) \mid k \in K, v \in V\}$ be the inventory that maps a lemma-PoS pair $k = \langle l, p\rangle$ to a list of candidate synsets $v = [s_1, s_2, \ldots, s_n]$, where $K$ is the set of all possible Latin lemma-PoS pairs, and $V$ is the set of all possible lists of candidate synsets. Additionally, let $P = [\langle la_1, en_1, k_1\rangle, \langle la_2, en_2, k_2\rangle, \ldots, \langle la_n, en_n, k_2\rangle]$ be the list of Latin-English parallel sentences with the Latin lemma-PoS pair $k_i$ pre-identified in $la_i$. Now, given a generic pair of parallel sentences $p_i = \langle la_i, en_i\rangle$ in which $k_i \in K$ occurs, we can easily retrieve the list of its candidate synsets $I(k_i) = v = [s_1, s_2, \ldots, s_n]$ associated with the Latin lemma-PoS pair. We then apply a state-of-the-art English WSD system on $en_i$ and obtain a list of synsets $w = [s_1, s_2, \ldots, s_m]$, each of them corresponding to the disambiguation of an ambiguous word in $en_i$. Finally, to select the correct meaning of $k_i$ among those in $v$, we calculate the intersection between the two sets of synsets as $v \cap w$. The resulting dataset has 6412 annotated sentences.

$Propagation_{w/align}$. The intersection computed as the last step of the previous strategy can contain zero, one, or more elements. If it is empty, either $\langle la_i, en_i\rangle$ were not one the translation of the other, or the English WSD system misclassified the target word, in both cases leading to an impossible propagation. If the intersection contains exactly one element, we can safely propagate the annotation from English to Latin. If, instead, the intersection contains more than one element, two or more senses in $w$ were found in $v$, hence failing to disambiguate the Latin word. To address this, we experiment with the inclusion of word alignment as provided by Yousef and Berti (2015) to uniquely identify the English target word $t$ associated with $k_i$ and then directly propagate the output of the WSD system $s_i$ specific for $t$ to $k_i$ when available in $v$. The final dataset constructed with this propagation strategy consists of 3886 annotated sentences. Figure 1 depicts an example of annotation propagation with the described methods.

## 4. Experimental Setup

### 4.1. Datasets

We used the Latin portion of the SemEval 2020 annotated dataset for training and testing, henceforth called 'SemEval 2020 Latin dataset'. This dataset was derived from the LatinISE diachronic corpus of Latin (McGillivray and Kilgarriff, 2013), a 10-million word token corpus of Latin texts from the fifth century BCE to the 2000s.[4] The in-context annotation was done as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020): of 40 Latin lemmas, 20 were selected because they changed their meaning in relation to Christianity (for example, *beatus*, which shifted its meaning from 'fortunate' to 'blessed'), and 20 as control words. 60 sentences were annotated for each of these lemmas: 30 randomly extracted from BCE texts and 30 from CE texts. The annotation was conducted according to a variation of the DuReL framework (Schlechtweg et al., 2018), as explained in Schlechtweg et al. (2020) and in McGillivray et al. (2022). The annotators assigned each usage of a target word a value in a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical) according to its closeness to each of the word's possible dictionary definitions, drawn from the Logeion online dictionary,[5] encompassing the Lewis and Short's *Latin-English Lexicon* (1879) (Lewis and Short, 1879), Lewis' *Elementary Latin Dictionary* (1890) (Lewis, 1890), and the dictionary by Du Fresne Du Cange et al. (1883-1887).

For our experiments, we reserved a portion of the Latin 2020 SemEval dataset for model evaluation by selecting one sentence in every 10 and leaving

[4] Openly available at `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2506`

[5] `https://logeion.uchicago.edu/`

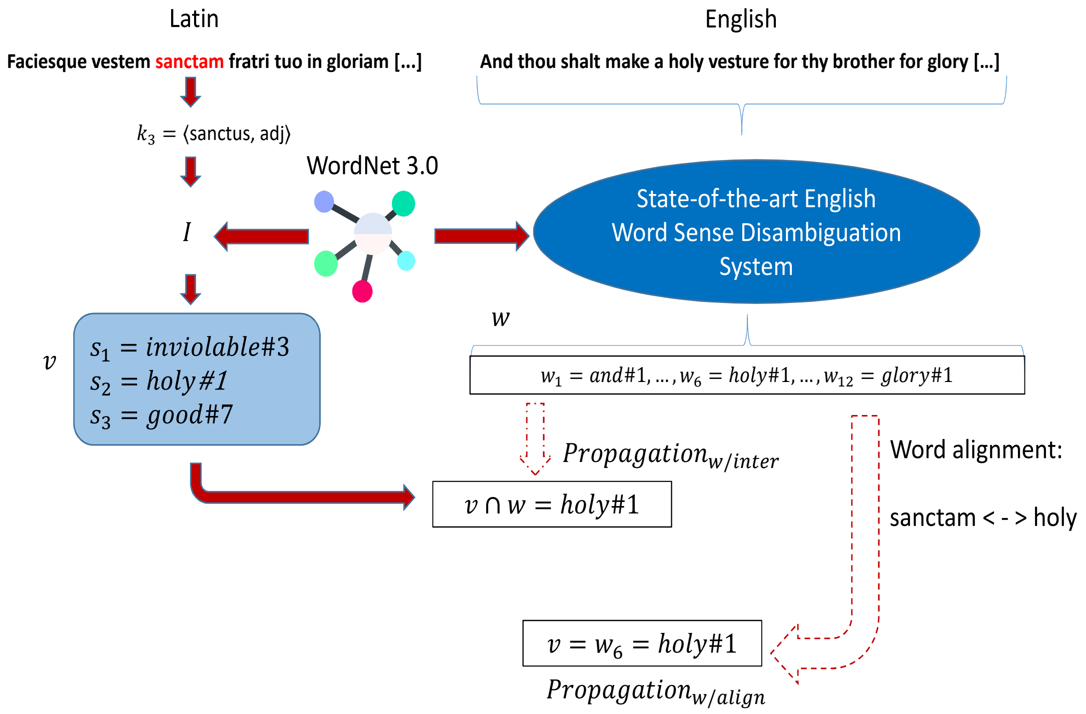
Latin
English
Faciesque vestem sanctam fratri tuo in gloriam […]
And thou shalt make a holy vesture for thy brother for glory […]
$k_3$ = ⟨sanctus, adj⟩
WordNet 3.0
$I$
State-of-the-art English Word Sense Disambiguation System
$w$
$v$
$s_1 = inviolable\#3$
$s_2 = holy\#1$
$s_3 = good\#7$
$w_1 = and\#1, \ldots, w_6 = holy\#1, \ldots, w_{12} = glory\#1$
$Propagation_{w/inter}$
$v \cap w = holy\#1$
Word alignment:
sanctam < - > holy
$v = w_6 = holy\#1$
$Propagation_{w/align}$

Figure 1: An example of the two annotation propagation methods described. Two parallel Latin-English sentences are used so that the English sentence is disambiguated by a state-of-the-art WSD system, which obtains the WordNet synset identifiers for each lemma in the English sentence. At this point the synset for the target Latin word (i.e. sanctam) is chosen in one of the two methods: a) $Propagation_{w/inter}$: we take the union of the English synsets $w$ and the possible synsets for sanctam $v$. b) $Propagation_{w/align}$: we use a word alignment module to obtain the English word associated with the target word sanctam (i.e. holy) and we assign the English synset $w_6$ as the synset of the target word.

the remaining 90% for training. We experimented with 6 different training configurations:

- **SemEval**: only the above-mentioned 90% of the SemEval 2020 Latin dataset;
- **Pers**$_{inter}$: the data obtained through the automatic annotation propagation methodology described in section 3.3 applied over the Perseus parallel corpus and referred to as $Propagation_{w/inter}$;
- **Pers**$_{align}$: same setting as the previous one, but using the data produced with the $Propagation_{w/align}$ strategy described in section 3.3;
- **Semeval+Pers**$_{inter}$: both the 90% of the SemEval training set and the data produced with $Propagation_{w/inter}$ are used for training;
- **SemEval+Pers**$_{align}$: the same setting as before, but with $Propagation_{w/align}$ instead.
- **SemEval+Pers**$_{rare}$: same setting as Semeval+Pers$_{*}$, but in this case we filter out all the senses occurring more than a threshold $\sigma$ times in the SemEval dataset, with $\sigma = 18$, which is the average sense frequency in the dataset.

Table 1 compares SemEval with Pers$_{inter}$. Since Pers$_{align}$ is very similar to Pers$_{inter}$ and Pers$_{rare}$ is just a subset of the latter, in the table and in the discussion section below, we use the general term *Pers* when comparing SemEval with our propagated datasets in more general terms, where Pers=Pers$_{inter}$.

### 4.2. Word Sense Disambiguation

We used the current state-of-the-art model for WSD in Latin, which is LatinBERT fine-tuned as a classifier (Bamman and Burns, 2020). The LatinBERT model follows the original architecture of BERT (Devlin et al., 2019), specifically the base version including 12 layers and 768 hidden units, and pre-

| Dataset | SemEval | Pers |
|---|---|---|
| total data | 2400 | 6020 |
| total lemmas | 40 | 33 |
| total word senses | 125 | 84 |
| $A \setminus B$ | 52 | 11 |

Table 1: Statistics of the datasets used in our experiments. The last row shows the relative complement of the set of word senses of the dataset from the current column (A) and the set of word senses from the other dataset (B).

trained on a vast corpus of different Latin texts. Bamman and Burns (2020) showed how LatinBERT can be fine-tuned for a variety of tasks to reach state-of-the-art performance, including WSD for Latin. In fine-tuning they minimised the following loss function:

$$\mathcal{L} = CE(Y, \hat{Y}) \tag{1}$$

where CE is the cross entropy function and $\hat{Y}$:

$$\hat{Y} = Softmax(W^t E) \tag{2}$$

with $E \in \mathcal{R}^{nxd}$ being the $n$ contextualised word embeddings of dimension $d = 768$ extracted from LatinBERT and $W \in \mathcal{R}^{cxd}$ is a learnable weight matrix, with $c$ being the number of possible senses for the given lemma. This framework treats the problem of WSD as a simple supervised learning task, where the WSD system learns to classify each target lemma as one of its possible senses.

In training our Latin WSD system, we kept the learning rate fixed at 5e-5 for all the experiments. We used the Adam optimizer and trained all systems for 20 epochs. For training all the systems, we used a single Nvidia T4 GPU, resulting in an average training time of 37 minutes. For the English WSD system described in section 3.3 we used AMuSE-WSD (Orlando et al., 2021), an end-to-end multilingual system that is particularly suitable for the integration into real-world applications.

In evaluation, we used a weighted average of F1, precision and recall for each lemma, where we assigned a weight to each class (i.e. the possible sense for the given lemma) proportional to the relative occurrence of that lemma's sense in the test set. The reported results are the average of the metrics thus computed for all the 40 lemmas from the SemEval 2020 Latin test set described above.

# 5. Results

## 5.1. Quantitative Results

| Training Data | F1 | Precision | Recall |
|---|---|---|---|
| SemEval | 60.30 | 55.63 | 67.15 |
| Pers$_{inter}$ | 36.93 | 36.16 | 45.38 |
| Pers$_{align}$ | 36.93 | 36.21 | 45.78 |
| SemEval+Pers$_{inter}$ | 61.45* | **57.83** | **68.01** |
| SemEval+Pers$_{align}$ | 61.48* | 57.46 | 68.00 |
| SemEval+Pers$_{rare}$ | **61.83*** | 58.17 | 67.93 |

Table 2: F1, precision and recall scores for our Latin WSD system on the test set, using the training configurations described in Section 4.1. * indicates results are significantly better w.r.t. the SemEval baseline as indicated by a one-tail t-test performed on bootstrapped results ($p < 0.01$).

Table 2 shows the results we obtained using the different training data. Although the gold-standard data from SemEval are comparatively better than the ones obtained through our automatic annotation procedure, this outcome was expected. Indeed, the Pers dataset includes a considerable amount of noise due to the automatic sentence alignment, as well as possible artefacts from the sense propagation procedure. However, if we consider that the dataset required no manual annotation at all, the result is encouraging as it shows that a WSD system for a historical language can be successfully trained just by means of a parallel corpus and an English WSD system, with 17.64 being the score of a random baseline obtained by drawing word senses from a uniform distribution.

Our findings also demonstrate that by adding our silver data to the SemEval training set, it is possible to obtain statistically significant improvements over just using the latter on its own, across all metrics. Furthermore, we point out that our procedure can easily be extended to an arbitrary number of lemmas and that, by employing more accurate sentence (Molfese et al., 2024) and word alignment (Procopio et al., 2021) systems, our strategy could achieve considerably higher results, currently not feasible due to lack of training data for both tasks.

Both adding word alignment (Pers$_{align}$) and filtering out common senses (Pers$_{rare}$) do seem to improve results when combined with SemEval, but not significantly more than just using Pers$_{inter}$.

## 5.2. Qualitative Results

A close analysis of a data sample shows some interesting patterns in our results. In some cases, the simple SemEval model seems to perform worse than the combined model, although the senses of the target word are clearly distinguished. For example, the verb *licet* has four senses in the SemEval dataset: (1) 'it is pemitted/allowed'; (2) 'it is possible'; (3) 'though/even if'; (4) 'yes/alright'. While senses (1) and (2) are more closely related and could be more challenging for the disambiguation task, senses (3) and (4) are clearly distinguished. But, as shown in example 1, when *licet* shows sense (2) in context, the SemEval model fails and assigns it sense (3), whereas the combined models successfully assign it the correct sense.

(1) *Corpus enim* ***licet*** *esse aliud quod fertur et una labitur omnimodis occursans officiensque [...]*
For **it is possible** that there is another body which moves and glides along at the same time (i.e., with the moon) obstructing it and impeding it in every way [...] (Lucretius, *De rerum natura* 5, 705)

At the same time, sense granularity stands out as a challenge for all models, particularly when the differences between senses are subtle. Let us take the lemma *credo*. This verb has seven abstract senses in the SemEval dataset, which are also closely related semantically: (1) 'to give as a loan, to loan, lend'; (2) 'to commit or consign something to one'; (3) 'to trust to or confide in a person or thing, to have confidence in, to trust'; (4) 'to trust one in their declarations, to believe'; (5) 'to believe a thing, hold or admit as true'; (6) 'to think, to suppose'; and (7) 'to believe in God'. Both the combined models and the simple SemEval model struggle equally to distinguish senses (5) and (6), which are very close in meaning. Consider examples 2 and 3.

(2) *Constantius tamen quam quis facturos* ***crederet*** *in tam subito periculo [...] iam multifariam scalis appositis urbem eo die defenderunt.*
However, that day they defended the city, already filled with scaling-ladders everywhere, with even more determination than one **would have thought** they would in such an unexpected emergency. (Titus Livius, *Ab urbe condita* 37, 5, 2)

(3) *Quorum militum si et in alia provincia opera uti senatus velit, utro tandem modo promptiores ad aliud periculum novumque laborem ituros* ***credat****, si persoluta eis sine detractatione prioris periculi laborisque merces sit, an si spem pro re ferentes dimittant, iam semel in prima spe deceptos?*
If the senate would like to use those soldiers also in another province, in which of these two ways would it **believe** that they will face another danger and a new toil more readily, if the reward for the previous danger and toil was paid without any deduction, or if they sent them away, bringing back hope instead of reality, when they were already deceived once in their first expectation? (Titus Livius, *Ab urbe condita* 36, 15, 7)

In example 2, *credo* has sense (6): in this case, the combined systems assign it the right meaning, but the SemEval model fails and assigns it sense (5). In example 3, *credo* has meaning (5), and the SemEval model correctly assigns it to *credo* in this context, but the combined models fail and assign it sense (6).

## 6. Discussion

### 6.1. Lemma and Sense Coverage

In the previous section we showed how using a parallel corpus for obtaining automatically annotated data (i.e. Pers) can provide a dataset on which to train a WSD system which performs considerably better than chance and we have also shown that adding the data to an existing expertly annotated dataset (i.e. SemEval) helps to improve results in a significant way. The difference in performance between the WSD system trained on SemEval and the system trained only on Pers, however, is large enough to require some additional investigation.

If the annotation propagation had had a major impact on performance, we might have expected a system trained on such data to perform as well or even better than the one trained on SemEval, as we have more training data from the annotation propagation on the Perseus parallel corpus. However, since the dataset includes a certain amount of noise, we might expect the system trained on Pers to behave in more unpredictable ways, but a closer look at the automatically obtained training data suggests that other elements determine this difference in performance, namely, the difference in lemmas and sense coverage between SemEval and Pers.

Figure 2 shows the occurrence of the 40 lemmas in SemEval and in Pers. We can immediately notice how SemEval has few, but consistent, examples for each lemma, as it was designed to be balanced among lemmas. Pers, instead, presents data which are collected "in the wild", as they come from a heterogeneous corpus which was created for very different purposes (see section 3.1). Because of this, some lemmas are extremely well covered by this dataset, while others are almost entirely, or entirely, absent.

At the same time, even if certain lemmas are well covered in the dataset, the coverage of different senses might, nevertheless, still be very different. This is confirmed by Figure 3, which shows the occurrence of different senses from the 40 lemmas in both corpora. The two heatmaps look considerably different, with Pers presenting a much more sparse distribution of senses per lemma, as most senses are either very poorly represented, or absent. Even though the scale difference is partly to blame for this result (Pers includes outliers having more than 1000 examples), overall this difference

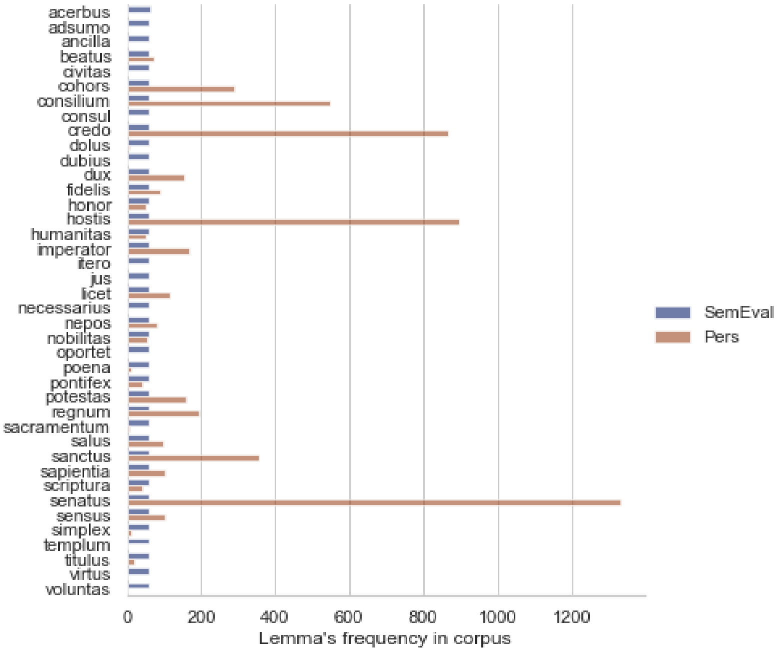

Figure 2: Occurrence of the 40 lemmas in SemEval and in Pers.

in distribution will also have an effect on the final performance, as the percentage of examples for a given word sense in Pers could be different from that represented in SemEval, resulting in a WSD system being more or less confident than it should be (according to SemEval) in outputting that sense. Moreover, when Pers does not include certain word senses at all, a system trained only on Pers will have no way of individuating those cases. By looking at the first column of the last row in Table 1, we can see that in fact there are 52 word senses that occur in SemEval but not in Pers. Both the difference in lemmas and senses' coverage, then, has a clear effect on results, as the systems are tested on SemEval and the performance reflects the lemma and sense distribution of that dataset.

The discrepancy between expertly curated datasets and data automatically collected "in the wild" is a factor that needs to be taken into account when applying these methodologies and even more so in the context of historical languages, where different datasets might cover extremely different time periods, which, in turn, has an effect on the distribution of words' senses. The SemEval dataset was created to represent different time periods in the Latin language equally (McGillivray, 2020), while more traditional corpora for Latin tend to include more texts from Latin's classical period. The mere difference in word senses distribution might depict a system trained on automatically collected data as strongly underperforming in comparison to a system trained on a balanced dataset like SemEval, but according to the context of use this might not be the case (e.g. if we want to perform WSD for data that are more similar to the ones in the Pers dataset).

The same difference in coverage from the two corpora also allows us to discover senses that were otherwise not represented in the original expertly annotated dataset, but that are instead present in the Pers one. The second column of the last row of Table 1 shows that eleven such occurrences were discovered in our experiments, which, however, could not be included in the test results, as the test set does not include these word senses. In future work, a portion of Pers could be validated so as to test the systems trained on different datasets on a different test set and give fairer evaluation of cases such as the ones just described. Notwithstanding the discussed limitations, the relative success of the technique we proposed makes it evident that the Pers dataset brings about improvements, at least on certain occasions. We hypothesise that these improvements are stronger in cases in which a sense was originally under-represented in the SemEval dataset, a hypothesis that we discuss in further detail in the next section.

### 6.2. Performance on Rare Senses

In this section, we want to see how the WSD system performs on relatively rare senses, as indicated by their lower occurrence in the SemEval dataset. We hypothesise that the improvements brought about by adding our propagated training data, compared to just using SemEval to train a WSD system, are related to these rare senses. To observe the performance for different word senses according to their

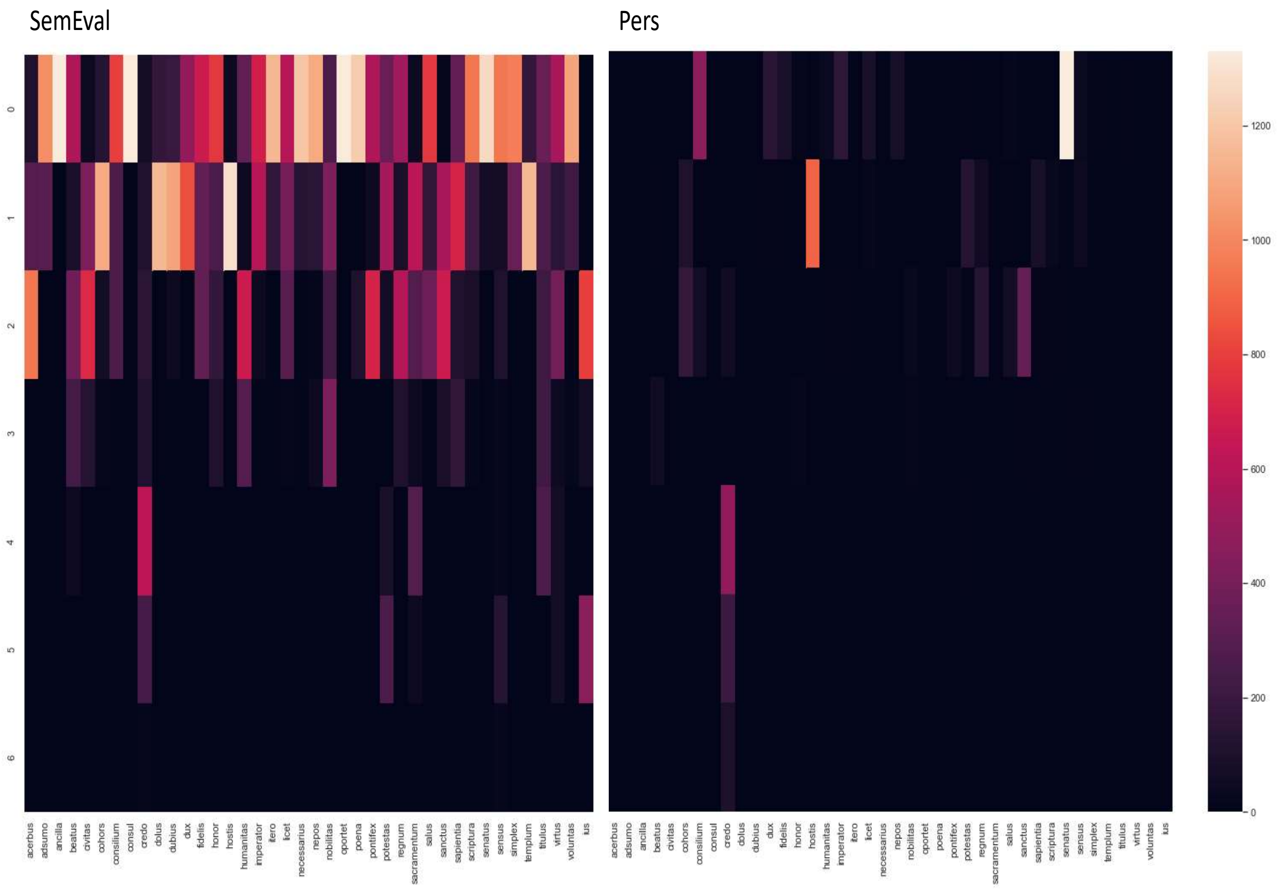

Figure 3: Occurrence of different senses per lemma in SemEval (left) and in Pers (right).

frequency, we grouped together all word senses into six frequency bins, according to the number of times they appear in SemEval. Figure 4 shows how the mean accuracy of the system increases as we consider word senses which occur more frequently in the target dataset, where the number of occurrences is indicated on the x-axis. The increase in mean accuracy is expected and we can eventually reach perfect performance after a certain frequency threshold. The system, however, has trouble disambiguating senses for which less than 10 examples exist in the SemEval training data.

If we then add the training data from Pers, we can see clearly from Figure 5 how performance for rare senses increases. This increase is even stronger if we filter the additional data to include just senses that were rare in SemEval; in this case, we can see that the rarest senses from SemEval get an almost double increase in performance and we can attribute to this the small but consistent gain of SemEval+Pers$_{rare}$ over SemEval+Pers$_{inter}$ and SemEval+Pers$_{align}$.

The SemEval+Pers$_{rare}$ configuration also shows a drop in performance for senses that were well represented in SemEval, demonstrating the trade-off between improving results for rare senses and having the same results for better represented senses. This trade-off probably occurs because more false positives are introduced for the rare senses, as the system trained on SemEval alone would often resort to never predicting such senses. This, in turn, causes more false negatives for senses that are well represented in SemEval.

Ultimately, the choice as to whether to use an approach such as SemEval+Pers$_{rare}$ depends on the use case, according to whether we are more interested in preserving the target distribution of word senses, or in capturing rare senses too. In the latter case, we have shown how this approach is effective and future research might further investigate this in other similar contexts, while exploring how to minimise the discussed trade-off.

In general, these experiments have proven that we can use annotation propagation to train a system from scratch and that if we add these annotated data to existing datasets we can obtain a significant improvement. The improvement is particularly notable for senses that were under-represented in the original dataset. In this regard, we have shown how using the propagation method to specifically augment under-represented senses can further improve results.

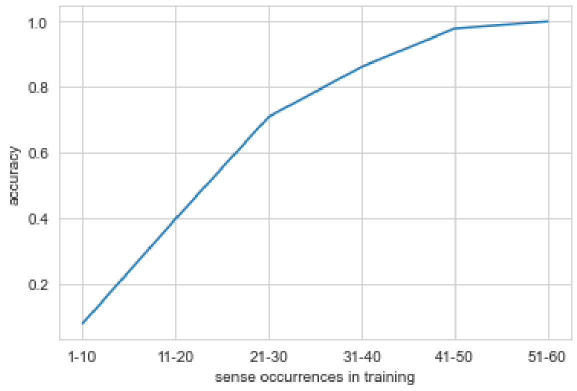

Figure 4: Accuracy (y-axis) against frequency of senses in the test set (x-axis) for the SemEval experiment.

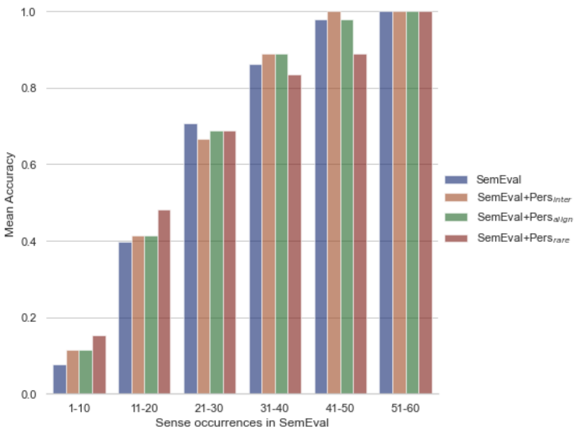

Figure 5: Mean accuracy performance for the WSD system trained with SemEval and SemEval+Pers$_{inter}$ sorted by sense occurrences in SemEval.

## 7. Conclusions

In this work, we have used a large Latin-English corpus to propagate annotations for WSD from English to Latin. By doing so, we aimed to address the challenges of performing WSD on Latin, which are shared by similar historical languages, namely, the lack of large sense annotated corpora. In addressing this problem, we have used an existing framework which we adapted to the specific use case, by exploiting a large bilingual corpus instead of native English resources.

We were able to train a WSD system for 40 test lemmas, using just the generated dataset. We have also shown that using such data to augment a small expertly-annotated dataset also significantly improves results, reaching the best performance in all settings. Especially when looking at senses that were under-represented in the expertly-annotated dataset, the gain in performance is evident and our method can improve results in these specific situations by augmenting an existing dataset with data related to just the rare senses.

Limitations still exist, however. There are still problems involving the coverage of lemmas and senses, as the distribution of a parallel corpus collected “in the wild” can be quite different from that of a curated dataset for WSD, where examples are selected based on a specific coverage of word senses. On the one hand, some lemmas or word senses can be completely missing from the data thus collected and there is no way of training a system effectively for such occurrences. On the other hand, the expertly curated dataset can misrepresent the actual frequency of certain lemmas and word senses in the target language. The results we obtain with a system trained on different data from a general-purpose parallel corpus might underperform as a consequence of a mere difference in the datasets’ distributions. Also, this work assumes parallel word sense distribution between English and Latin, meaning that there is always a correspondence between Latin and English word senses. This is indeed a strong assumption and it can prove false in a number of cases. Further, our system demonstrates a good level of accuracy for NLP applications, but its efficacy may still fall short for meaningful integration into historical linguistic research. Also, valuating system quality presents challenges due to the low inter-annotator agreement, as well as the system’s granular sense distinctions, which may lead to divergent sense selections by different annotators for the same word occurrences.

Notwithstanding these limitations, this work is the first to our knowledge to have applied annotation propagation for WSD in the context of a historical language. Many research directions can be pursued starting from this work. For example, different historical languages could be explored, or automatic machine translation systems could be employed so as to overcome specific weaknesses of existing bilingual corpora. We leave such experiments for future research.

## 8. Ethical Concerns

Given that the domain of application is historical languages, the authors do not express any particular ethical concerns about this work. As the systems explored are automatic, however, it is important to remember that word sense disambiguation is a nuanced task for which a system like ours can be of help, but not a substitute for the opinion of an expert in the field. We therefore encourage using our findings and methodologies as an aid for researchers, rather than as a substitute.

## 9. Data and Code Availability Statement

The code and data associated with this paper are available at https://github.com/Ighina/LatinWSD.

## 10. Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR. This article is also based upon work from COST Action NexusLinguarum – “European network for Web-centered linguistic data science” (CA18209), supported by COST (European Cooperation in Science and Technology) www.cost.eu.

This work has been carried out while Simone Tedeschi were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

## 11. Authors’ contributions

IG: conceptualization, formal analysis, investigation, methodology, software, writing - original draft; writing - review and editing. ST: conceptualization, formal analysis, investigation, methodology, software, writing - original draft; writing - review and editing. PM: data curation, investigation, writing - original draft; writing - review and editing. RN: Conceptualization, methodology, supervision, funding acquisition, writing - original draft; writing - review and editing. BMcG: conceptualization, methodology, supervision, writing - original draft; writing - review and editing.

## 12. Bibliographic References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. In *ArXiv*, volume abs/2009.10053.

David Bamman and Gregory Crane. 2009. Computational linguistics and classical lexicography. *Digital Humanities Quarterly*, 3.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, Roberto Navigli, et al. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3837–3844.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751—2761.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Erica Biagetti, Chiara Zanchi, and William Michael Short. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266, University of South Africa (UNISA). Global Wordnet Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, and Federica Zampedri. 2019. *Nunc Est Aestimandum*: Towards an evaluation of the latin wordnet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Accademia University Press.

Piroska Lendvai and Claudia Wick. 2022. Finetuning Latin BERT for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41, Taipei, Taiwan. Association for Computational Linguistics.

Enrique Manjavacas Arévalo and Lauren Fonteyn. 2022. Adapting vs pre-training language models for historical languages. *Journal of Data Mining and Digital Humanities, NLP4DH*.

Barbara McGillivray. 2020. *Routledge International Handbook of Research Methods in Digital Hu-*

*manities*, chapter Computational Methods for Semantic Analysis of Historical Texts. Routledge.

Stefano Minozzi. 2017. Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In *Strumenti digitali e collaborativi per le Scienze dell'Antichita*, pages 123–134, Venezia. Università Ca' Foscari.

Francesco Molfese, Andrei Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. 2024. CroCoAlign: A cross-lingual, context-aware and fully-neural sentence alignment system for long texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2209–2220, St. Julian's, Malta. Association for Computational Linguistics.

Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proc. of AAAI*.

Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4008–4018, Online. Association for Computational Linguistics.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, pages 177–212.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Luigi Procopio, Edoardo Barba, Federico Martelli, Roberto Navigli, et al. 2021. Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In *IJCAI*, pages 3915–3921.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1–23. International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Tariq Yousef and Monica Berti. 2015. *The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon*, pages 117–123. Institute of Computer Science, Warsaw, Poland.

## 13. Language Resource References

Charles Du Fresne Du Cange, G. A. Louis Henschel, P. Carpentier, Johann Christoph Adelung, and Léopold Favre. 1883-1887. *Glossarium mediæet infimælatinitatis*. L. Favre, Niort.

Charlton T. Lewis. 1890. *An Elementary Latin Dictionary*. American Book Company, New York, Cincinnati, and Chicago.

Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short*. Clarendon Press, Oxford.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpus Linguistics*, pages 247–257, Tübingen. Narr.

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.