# Measuring Historical Word Sense Variation

David Bamman
The Perseus Project
Tufts University
Medford, MA, USA
david.bamman@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA, USA
gregory.crane@tufts.edu

## ABSTRACT

We describe here a method for automatically identifying word sense variation in a dated collection of historical books in a large digital library. By leveraging a small set of known translation book pairs to induce a bilingual sense inventory and labeled training data for a WSD classifier, we are able to automatically classify the Latin word senses in a 389 million word corpus and track the rise and fall of those senses over a span of two thousand years. We evaluate the performance of seven different classifiers both in a tenfold test on 83,892 words from the aligned parallel corpus and on a smaller, manually annotated sample of 525 words, measuring both the overall accuracy of each system and how well that accuracy correlates (via mean square error) to the observed historical variation.

## Categories and Subject Descriptors

H.3.7 [**Information Systems: Information Storage and Retrieval**]: digital libraries

## General Terms

Design, Documentation, Performance

## Keywords

Word sense disambiguation, linguistic variation, digital libraries

## 1. INTRODUCTION

Words in all languages naturally possess a range of possible senses, and the ambiguity of which sense is valid in any particular context is dependent not only on the genre and register of the discourse but also on its historical place. The Oxford English Dictionary [1], for example, dates the first recorded political use of *radical* (meaning "advocating thorough or far-reaching political or social reform; representing or supporting an extreme section of a party") to 1783; its first use as a slang term (meaning "excellent, fantastic") comes from 1964. We can imagine, however, that the distribution of these senses has not been uniform over this time: we might guess that the slang sense was used far more frequently in 1970s and 80s, and has decreased in frequency since then as its popularity waned. At the same time, we might guess that the use of its political sense increased in the early years of the 21st century, amid heightened public discourse of religious fundamentalism.

The ability to quantify this sense variation over time has clear value not only for linguistic research, but also arguably for information retrieval and extraction as well [23, 41, 38] – if we want to locate documents based not on a token-based search query but rather on its underlying semantics, understanding the likelihood of a word sense at a given historical time will help in the precision of retrieval (i.e., not returning documents prior to 1964 with the slang sense of *radical*).

We describe here our strategy for measuring this kind of word sense variation over time – for creating a probabilistic sense inventory in which we calculate the likelihood of any given sense for a word at any particular time. Human-created reference works such as the OED are natural sources of sense inventories – indeed, they often capture the judgments of experts over decades of labor – but our goal is to induce one automatically in a language-independent fashion, enabling its broader use in a number of languages without requiring the labor of experts.

## 2. METHOD

Our work is based on aligning a small collection of parallel texts to induce a bilingual sense inventory and using that alignment as training material for a broad-coverage classifier to automatically tag the word senses in a much larger collection [13, 21]. Based on the intuition that one language tends to use different words to translate different senses in another (e.g., "the bank" in English corresponds to "la banque" in French when referring to a financial institution but to "la rive" when referring to a side of a river) [5], the core of the approach involves word aligning a set of source documents in language $e$ with their translations in language $f$, inducing a probabilistic translation lexicon from that alignment to form the sense inventory (in which the possible translations in language $f$ for a word in $e$ approximate its different senses) and then leveraging the context around each source word in $e$ that has been aligned with a sense "label" $f$ to provide a training instance for classification.

One drawback to this approach, however, is the need for large amounts of parallel text data, especially when existing


*JCDL'11,* June 13–17, 2011, Ottawa, Ontario, Canada.

resources tend to center around a small set of genres (such as newswire and parliamentary proceedings [36, 24]), whose linguistic register may only include a small subset of any word's possible senses. A fertile source for a much wider variety of such parallel data, however, can be found in the million book libraries now emerging online.

At just over 2 million and 15 million works respectively, the online libraries of the Internet Archive and Google Books have already begun to provide the primary material for research into linguistic variation and cultural trends within a single language [10, 26]. These collections, however, are also multilingual in the extreme: the publicly available collection of the Internet Archive contains, as of January 20, 2011, not only 1,888,944 works catalogued as being written in English but a significant number in French (234,281), German (186,600), Latin (28,844), Italian (43,241) and Spanish (38,112) as well.

Latin presents a unique opportunity within this collection: with over 25,000 works containing 2.7 billion words composed over 2000 years, it arguably spans the greatest historical distance of any major textual collection today, capturing not only the language of Caesar and Cicero from the Classical era (ca. 200 BCE–200CE), but also religious texts from Late Antiquity and the Middle Ages (such as those by Augustine and Aquinas) and scientific treatises from figures such as Galileo, Newton and Kepler. This large textual collection coupled with its deep historical distance provides a natural proving ground for leveraging modern techniques of word sense induction and disambiguation to begin uncovering how word senses have evolved over two thousand years: e.g., how a Latin word like *oratio* can shift from predominantly meaning "the power of speech" in the Classical era to "prayer" in the Middle Ages and beyond.

This strategy consists of four steps: a.) mining a 389-million-word dated Latin corpus from the much larger collection of the Internet Archive; b.) aligning a small set of 129 known translation book pairs (in Latin and English) to induce a sense inventory and training instances for several WSD classifiers; c.) leveraging the trained classifiers to automatically tag the word senses for the remainder of the 389-million-word collection; and d.) measuring sense variation over 2000 years using that automatically tagged corpus. Our results suggest that the size and deep historical distance of these million book collections provide an adequate scope for measuring the development of word sense variation over time.

## 3. RELATED WORK

This work builds on two separate strands of research. One is the use of parallel corpora for unsupervised word sense disambiguation. While parallel texts have been the foundation on which modern statistical machine translation now stands [5, 6], their utility extends to semantically related tasks as well, such as inducing bilingual dictionaries [22] and translating collocations [39]. Their use in word sense disambiguation [21] rivals or surpasses other unsupervised methods, and they have been used with success with translation pairs in English/French [13], English/Chinese [8, 31], English/Portuguese [40], and English/Vietnamese [14].

A second strand of research informing our own is the increasing use of historical corpora to measure trends in language, history and culture. Researchers in linguistics and the digital humanities have long used large textual corpora to discern variation in language, both for studies in dialectology [42] and stylistics [28, 20]. The rise of large digitization projects such as Early English Books Online and of more open transcription efforts such as the EEBO-Text Creation Partnership, however, ushered in collections an order of magnitude larger (to date, EEBO-TCP contains transcriptions of 11,462 documents for a total of 521 million words from 1475-1700), enabling research into the development of English over a period of 300 years [4].

With the arrival of broad-purpose digital libraries such as the Internet Archive and Google Books, however, researchers now have an even larger collection of material to work with. While most of Google's 15 million book collection is still under copyright and unavailable for wider use, the Internet Archive, in contrast, houses a smaller collection of 2 million works all available for public download. Based on our own subset of 1.2 million books from this collection, we estimate that the total word count in the 2-million-book library of publicly available material in the Internet Archive comprises a total of 243 billion words.

The size of this collection has made it a natural target for topic modeling [27] (including evaluating topic coherence [30]), where the immensity of the data encourages automatic methods for characterizing it, as well as research into automatic methods for adding structure [3]. More recently, however, researchers have begun to exploit these massive datasets for measuring historical change. Using data from Google Books, Cohen and Gibbs [10] track the rise and fall of words in 1,681,161 book titles published in English in the United Kingdom from 1789-1914, and Michel et al. [26] chart lexical trends and linguistic phenomena such as the regularization of English verbs over the past 200 years. In general, we can expect to see more and more such projects in the future, due in part to growing awareness of large data not only in linguistics, but also in the social sciences [9] and humanities [11, 19].

## 4. MINING THE CORPUS

The first step in tracking sense variation over time is the construction a large, dated Latin corpus. We assembled our collection from two sources: first, the Internet Archive provided the source texts and metadata for 1.2 million books (a snapshot of their entire collection from early 2009). We supplemented this broad foundation by downloading all more recent works that had originally been catalogued as being written in Latin (according to the catalog records from the primary digitizing library) and scanned since that date. This resulted in a total collection of 25,886 Latin books.[1]

As others have pointed out, however, problems plague these massive collections in their use for scholarly research, not only in the quality of the image scans and the resulting OCR but also in the metadata itself that describes the text [33]. In order to compensate for this error, we automatically classified each document into its major and minor languages using an ngram language identifier [43] trained on 24 different language editions of Wikipedia along with the open source collection of Greek and Latin texts found in the Perseus Digital Library [12]. This resulted in identifying

[1]This figure represents the total number of books catalogued as Latin in the Internet Archive on November 11, 2010; since then, this number has increased as more Latin works are still being digitized.

10,263 non-Latin books that had been manually classified as Latin, along with 6,790 Latin books that had been originally classified as *not* Latin,[2] leading to a net total of 22,413 books containing 2,971,407,550 words.

The second problem that confronts a historical linguistic analysis is the lack of metadata specifying the date of composition. While all books are attended by their reported date of publication, this date only records the year of the printing of that specific edition, not the date the work was originally produced. While for modern works these dates are very closely related, this is much less true for historical works and certainly untrue for any work originally composed prior to Gutenberg.

To address this, we commissioned undergraduate student researchers in Classics to find as narrow a window as possible for the original date of composition – while more recent authors may have more-or-less established composition dates, others (such as more obscure medieval authors) do not. This work (representing approximately 1,000 person-hours of labor over the course of a summer) resulted in a collection of 7,055 works containing a total of 389 million words. Figure 1 shows the distribution of works charted by their date of composition. Major peaks immediately rise up around the Classical era of ca. 200 BCE–200 CE (including authors such as Cicero and Vergil), the works of church fathers such as Augustine (ca. 400 CE), and voluminous scholastic writers such as Thomas Aquinas (ca. 1200 CE), before yielding to an explosion in the number of printed works following the invention of the printing press (and especially following the Industrial Revolution of the 19th century).

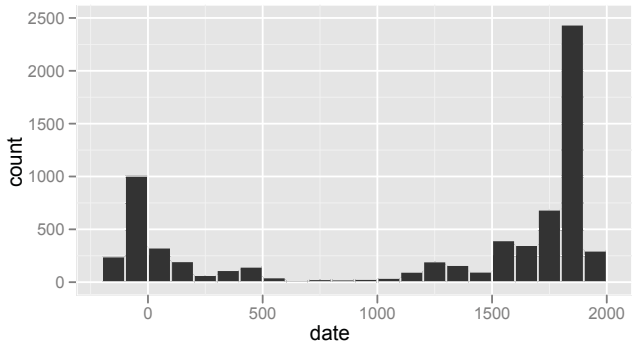


**Figure 1: 7,055 Latin works in the Internet Archive, charted by date of composition.**

[2]To test the cost of ignoring the manual classification completely, we evaluated the precision of the automatic classification on a small random sample of 50 texts where it found Latin to be the dominant language and the library metadata did not. In 49 cases, the automatic classification was correct in assigning Latin as the major language. While 16 of the manual misclassifications were outright mistakes (e.g., mistaking Dutch or Italian for Latin), the most common error seemed to be assigning the language of the full text based on the language of the first few pages (which in Classical works are often introductions in other languages). While the small sample size can only be suggestive of the real precision, we feel the gain in text data (6,790 works) helps compensate for the error introduced.

# 5. INDUCING LATIN SENSES

Latin lexicographers have long built dictionaries focussing on wide coverage for specific eras, such as the Classical period [17, 25] and the Middle Ages [32], along with those tailored for the more specialized vocabularies of individual authors (such as Thomas Aquinas [37]), and some even exist in digital form as viable sense inventories. However, a comprehensive dictionary spanning all two thousand years of usage does not exist, necessitating the automatic creation of one. Our goal is a purely data-driven approach – one that can learn the meaning of Latin words over the course of its 2000-year lifetime without the bias of dictionaries composed for any specific era.

## 5.1 Identifying translations

The sense inventory we create for Latin is based on the alignment of parallel texts consisting of a collection of source texts and their book-level translations. To compile a parallel corpus, we manually identified a set of 129 Latin-English book pairs, representing an even distribution over the two thousand years that Latin was used as a *lingua franca* throughout Europe. Even with this high-level information, however, automatically aligning sentences in the source document with sentences in the target document is not trivial due to the inherent asymmetry of the translation pairs. Unlike parallel corpora such as the Canadian Hansards [36] or Europarl [24] parliamentary proceedings, book level translations often contain information on one side that is not present in the other – a book-level translation pair of Vergil's *Aeneid*, for example, may consist of an English book containing only the *Aeneid* along with the Latin original found in the complete works of Vergil. Even presuming an equal translation pair containing the same canonical text, one work may contain an extensive introduction or notes not found in the other.[3]

This situation more closely resembles the task of extracting sentence fragments from non-parallel and other comparable corpora [29] but our experiments with these methods generally met with consistently low accuracy on this collection due, in large part, to the level of OCR errors affecting the Latin texts. Rather than rely on equally misplaced assumptions of *non*-equality, we attempted instead to leverage the document structure of the translation pair: while the source text and the translation may indeed be asymmetrical, the text in common will generally be found in the same sequence in both. To this end, we segmented both the source and target documents into sentences, indexed the target sentences and attempted to find an initial document structure by translating the source sentence word for word (using a translation dictionary induced from 2.9 million words of cleanly transcribed parallel texts in the Perseus Digital Library) and finding the best target match among those indexed. Figure 2 displays the document structure for a single Latin-English translation pair (Augustine's *Confessions*) revealed using this method.

[3]Indeed, asymmetry is the norm: of the 129 translation pairs, only 42 were more or less "equal" translations (32.6%); in 46 pairs (35.7%), the translation was a subset of the original book (i.e., the original work included significant text not found in the translation); in 20 pairs (15.5%) the original work was a subset of the translation; and 21 pairs (16.3%) held an asymmetrical relation to each other (each book containing significant text not found in the other).

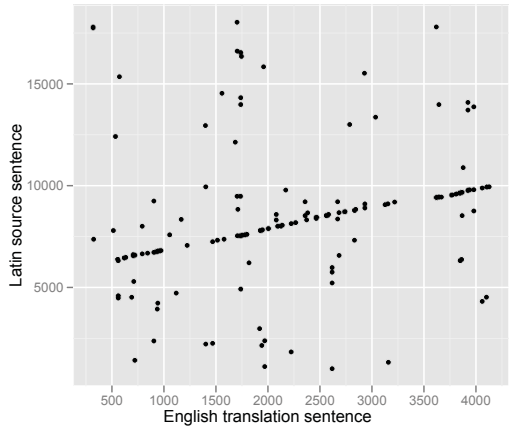


**Figure 2: Sentence alignment for Augustine's *Confessions* in Latin and English.**

In this example, the text common to both can clearly be seen on the line spanning sentences 6,000-10,000 in the Latin source document and sentences 500-4,750 in the English translation – while the translation contains only the text of the *Confessions*, it accompanies other works in the Latin original. Using linear regression, we calculate the slope of the line and exclude all sentences found outside two standard deviations of it. This resulted in 40,323 sentence pairs for subsequent word alignment.

## 5.2 Word alignment

These 40,323 sentence pairs are then aligned at the level of individual words using MGIZA++ [18], a multi-threaded version of GIZA++ [35]. Prior to alignment, all of the tokens in the source text and translation are stemmed (to account especially for Latin's rich inflection).[4] After alignment, the original Latin word forms were restored and then lemmatized, using the English sense as a feature for lemma disambiguation.[5] In order to find only high-quality translation equivalents, the word alignment is performed in both directions (i.e., from Latin to English and English to Latin) and the final word alignment is found using Och and Ney's "refined" combination (in which the intersection between the two is first determined and then extended by adding non-competing and/or neighboring links [34]. This resulted in clean alignments for 504,857 words.

## 5.3 Inducing a sense inventory

Next, a sense inventory is induced by aggregating the English translations for each Latin lemma and enforcing both a minimum threshold on the number of observations required to form a valid sense ($n = 3$) and a log likelihood thresholding to filter out common but uncharacteristic translations.[6] This filtered a set of 109,432 possible Latin-English translations pairs down to a working inventory of 3,412.

# 6. WORD SENSE DISAMBIGUATION

Armed with this induced sense inventory and set of aligned parallel texts, we treat each instance of a Latin source word aligned to a viable sense in the English translation as a a training instance for word sense disambiguation. Each training instance includes the lemma to classify, the aligned sense in the English translation, and the 20-word context on either side of each target word. We trained several standard word sense disambiguation systems [2, 7] and evaluated their performance in a tenfold test. Seven different classifiers were evaluated: language model classifiers trained on 5-grams ("5-gram LM"), 6-grams ("6-gram LM"), token unigrams ("token unigram LM"), and token bigrams ("token bigram LM"); naive Bayes classification trained on token unigrams ("Bayes"); unigram TF/IDF classification (TF/IDF); and $k$-nearest neighbor classification using cosine distance (KNN). For each classifier, the training context surrounding each instance included a window of 20 words (tokenized by space and lowercased) around each word. In addition, we also evaluated the performance of a baseline measure of simply selecting the most frequent sense (MFS) from the probabilistic translation lexicon.

## 6.1 Evaluation

We conducted two evaluations of the different word sense disambiguation classifiers: one on the large set of automatically aligned parallel texts; and one on a smaller but manually annotated sample.

### *6.1.1 Automatic*

To evaluate the impact of the size of the training data on the overall accuracy, we conducted the tenfold test for each classifier on three different subsets of the aligned texts: words appearing in the data more than 100 times (53 distinct lemmas, 57,670 observations), more than 50 times (79 distinct lemmas, 64,163 observations) and more than 10 times (375 distinct lemmas, 83,892 observations). In each instance, the classifier was trained on 9/10 of the dataset and tested on the remaining one-tenth; this test is conducted a total of ten times, once for each held-out tenth, with the reported accuracy in Table 1 being the average of all tests.

| System | >10 | >50 | >100 |
|---|---|---|---|
| 5-gram LM | **71.6%** | **69.5%** | **69.0%** |
| 6-gram LM | 71.1% | 68.8% | 68.2% |
| Bayes | 70.2% | 68.5% | 68.0% |
| Token Unigram LM | 70.8% | 68.5% | 68.0% |
| Token Bigram LM | 70.8% | 68.5% | 68.0% |
| TF/IDF | 70.0% | 67.6% | 66.9% |
| KNN | 68.3% | 68.1% | 67.8% |
| MFS Baseline | 67.0% | 65.6% | 66.3% |

**Table 1: 10-fold test on parallel data.**

Two things immediately jump out: first, the performance degrades as more frequent lemmas are evaluated; this is due

[4]Words are mapped to a base form using the Perseus morphological analyzer [12].

[5]E.g., if a word such as *est* is aligned to the English word "eat," it is more probably derived from the lemma *edo* ("to eat") than *sum* ("to be") since many unambiguous inflections of *edo* (such as *edisti*) also align to "eat."

[6]For example, while *non* ("not") may incorrectly align to "he" many times due to the high frequency of both words in their respective languages, the pair would be filtered out due to the observed number of alignments being fewer than what would be expected if each word was indeed a translation of the other.

| System | *villa* | *pastor* | *miles* | *scientia* | *oratio* | Average |
|---|---|---|---|---|---|---|
| 5-gram LM | 54.8% | 69.2% | 90.2% | 73.7% | 61.4% | 69.9% |
| 6-gram LM | 58.3% | 61.5% | 91.2% | 65.8% | **63.8%** | 68.1% |
| Bayes | 63.5% | 62.3% | 92.6% | 70.2% | 48.0% | 67.3% |
| Token Unigram LM | 63.5% | 62.4% | 92.6% | 70.2% | 48.0% | 67.3% |
| Token Bigram LM | **64.3%** | 62.4% | 92.6% | 70.2% | 48.8% | 67.7% |
| TF/IDF | **64.3%** | 60.7% | 82.8% | 70.2% | 49.6% | 65.5% |
| KNN | **64.3%** | **73.5%** | 84.4% | 63.2% | 40.1% | 65.1% |
| MFS Baseline | 60.9% | 66.7% | **92.6%** | **79.0%** | 60.6% | **72.0%** |

**Table 2: Accuracy rates of WSD and Most Frequent Sense (MFS) classifiers on gold standard data.**

in large part to the increasing polysemy of more frequent words. Second, while all of the WSD classifiers perform better than a purely random selection, they fare only slightly better than the powerful baseline of selecting the most frequent sense in the probabilistic translation lexicon. In this test, the best performing WSD is the language model classifier trained on character 5-grams.

### 6.1.2 Manual

While the previous test gives us a measure of the consistency of the automatically aligned parallel text data, we still want to evaluate it against human-created judgments of sense. To that end, we created a test set comprised of manually annotated sense labels for 105 instances each of five Latin nouns: *villa* (villa, town), *pastor* (shepherd, pastor), *miles* (soldier, knight), *scientia* (knowledge, science) and *oratio* (speech, prayer). Each of these words was selected because it has a known shift in meaning – the original meaning of *villa* in the Classical era (ca. 200 BCE – 200 CE) is that of a "villa" or "country house" before being supplanted by "town" later in its life; the original meaning of *pastor* was "shepherd," slowly becoming a religious "pastor" over time; *miles* in the Classical era referred to a soldier in the imperial Roman army, but in the Middle Ages acquired the more specific meaning of "knight"; *scientia* in Classical Latin largely designated abstract "knowledge" though in the Early Modern era came more and more to mean the systematic study of "science"; and *oratio* in the Classical era typically signifies the power of speech, while in the Middle Ages acquired the more specific meaning of religious "prayer."

For each of these words, we classified 105 instances of its use in an even distribution across the 21 centuries from 100 BCE to 1900 CE. After training each WSD classifier on the entire parallel collection, we then evaluated its performance on this gold standard set. Table 2 presents the results of this evaluation.

While the words each show some variation as to which classifier performs best, what we find in general is that the strong baseline of selecting the most frequent sense in the probabilistic translation lexicon still largely beats out more sophisticated measures on this noisy data. In the abstract, WSD would seem to fail against this much simpler measure.

## 6.2 Evaluation Over Time

Our goal, however, is not the evaluation of word sense disambiguation in itself but rather the evaluation of whether it can be used to accurately predict sense variation over time. In this effort, the simple but powerful baseline of choosing the most frequent sense would always lead us astray, since it *a priori* prevents us from measuring any kind of variation. Figure 3 illustrates this by charting the frequency of "speech" as a sense for *oratio* according to three distributions over time: the automatic classification of the best-performing WSD classifier (6-gram LM); the gold standard test data; and the simple baseline of always selecting the most frequent sense (since "prayer" is the most frequent sense of *oratio*, this distribution leads to "speech" never being chosen).[7]

The gold standard distribution represents the actual level of sense variation in Latin for the word *oratio* over its 2000 year lifetime as measured in the 105-instance sample we manually created – in this data, we can see that "speech" is clearly the dominant sense in the Classical Latin period around the turn of the millennium, and is used with less and less frequency in the Middle Ages from ca. 700 CE – 1300 CE (where "prayer" was the dominant sense), before regaining its status as the dominant sense in the Neo-Classical period of the Renaissance and beyond. We note that the best-performing WSD classifier follows this same trajectory, while the baseline naturally cannot.

Similarly for *scientia*: Figure 4 represents the corresponding three distributions each representing the ratio of "knowledge" to all possible sense values for *scientia*. The gold standard frequency clearly shows that "knowledge" was its dominant sense until ca. 700 CE, when it starts to give way to "science." The best performing WSD classifier follows this same trend, while the baseline of always selecting the most frequent sense cannot.[8]

One method of measuring the goodness-of-fit of each of these distributions is to calculate its mean square error (MSE)[9] in relation to the true values given in the gold standard (i.e., the average of the square of the difference between the observed value $f(x_i)$ and the gold standard value $y_i$ for each value of $x$):

$$\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n}$$

In doing this, what we are measuring is not the accuracy of the word sense disambiguation in itself, but rather its appropriateness to the task of charting lexical variation over time. The gold standard represents the true variation; what we want to evaluate is how well automatically tagged data matches this historical trajectory as well. In this, we follow other evaluations that measure the accuracy of a process

[7]In this case, since *oratio* is only used with two senses, "prayer" comprises the remaining frequency for each century.
[8]Since "knowledge" is the most frequent sense of *scientia*, it is always selected; hence its observed frequency in that distribution of 1.0.
[9]I.e., the residual sum of squares divided by the number of degrees of freedom.

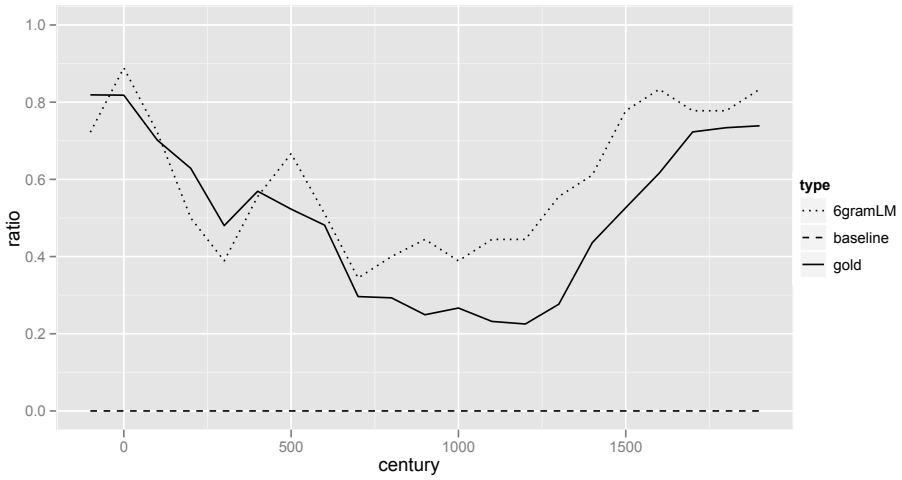


**Figure 3: Ratio of *oratio* classified as "speech" (in relation to all possible senses of the word) according to three distributions.**

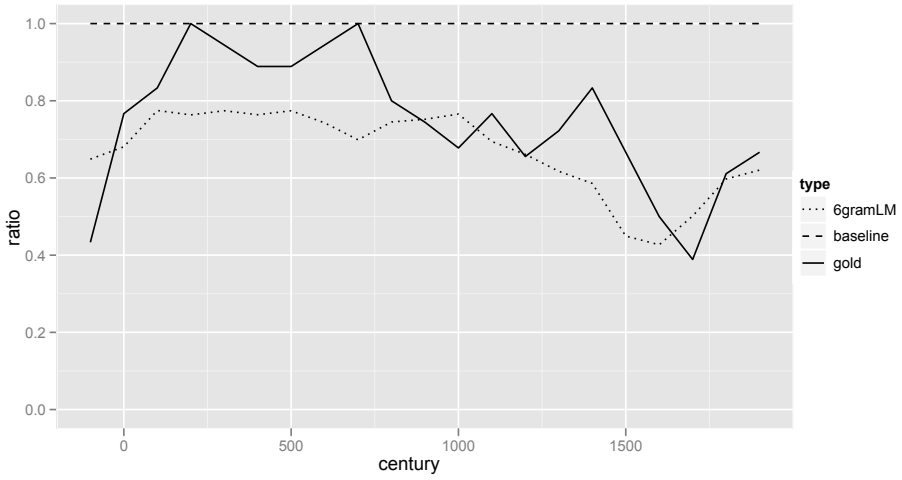


**Figure 4: Ratio of *scientia* classified as "knowledge" (in relation to all possible senses of the word) according to three distributions.**

by its downstream application, such as measuring the accuracy of word alignment through the proxy of subsequent improvements to BLEU scores in machine translation [16, 15].

Table 3 displays the MSE for each of these distributions in comparison to the gold standard for each of the words *villa*, *pastor*, *miles*, *scientia* and *oratio*.

As the table shows, the 6-gram language model classifier far outperforms the baseline and the best-performing WSD classifier (5-gram LM) when measured by overall accuracy. The baseline of selecting the most frequent sense is revealed to be the worst-performing method of classification.

One question that remains, however, is whether the baseline of selecting the most frequent sense *overall* in the data is the best point of comparison – a better baseline may in fact be selecting the most frequent sense *in a given century* (MFS/C), since this could still reveal variation across time. Since our sense inventory is induced from our aligned parallel text data, and the texts in that data are attended by date information, we can easily calculate the most dominant sense for any word in any century. Comparison with this new baseline raises an even more fundamental question: whether it may be more accurate to measure historical sense variation directly from the parallel text data itself (rather than introducing noise from WSD classification errors). The fundamental question here is which is better for this task: a small set of relatively accurate data (504,857 words of aligned parallel texts, each with a sense "label" from the translation) or a much larger set of noisy data (389 million words of automatically tagged senses). To evaluate this new baseline, we calculated the sense distribution over time for only the aligned data and calculated the mean square error between that distribution and the gold standard. Since the aligned data is very sparse (*oratio*, for example, occurs a total of 89 times with an aligned label of "speech" or "prayer" and does not occur at all in 9 of the 21 centuries), we smoothed

| System | *villa* | *pastor* | *miles* | *scientia* | *oratio* | Average |
|---|---|---|---|---|---|---|
| 5-gram LM | .056 | .034 | .052 | .044 | .137 | .065 |
| 6-gram LM | .053 | .053 | .052 | **.022** | **.022** | **.040** |
| Bayes | .047 | .060 | .055 | .040 | .228 | .086 |
| Token Unigram | .047 | .060 | .055 | .040 | .228 | .086 |
| Token Bigram | .047 | .060 | .055 | .044 | .230 | .087 |
| TF/IDF | **.037** | .050 | .049 | .040 | .189 | .073 |
| KNN | .101 | **.028** | .054 | .039 | .248 | .094 |
| MFS Baseline | .228 | .170 | **.014** | .091 | .388 | .178 |

**Table 3: Mean square error of WSD systems and Most Frequent Sense (MFS) baseline in relation to the gold standard.**

the data for centuries with fewer than two observations by assigning it a value on the slope between the closest dates before and after it with two or more observations.

Table 4 displays the performance of this new baseline – while the average mean square error of the baseline of picking the most frequent sense given the century (.146) is slightly better than the simple baseline of selecting the most frequent sense overall (.178), it still performs worse than any WSD classifier above, and far worse than the best-performing classifier, the 6-gram LM (.040).[10] In this particular task, a larger volume of noisy data trumps a smaller set of more accurate data.

| Word | MFS/C MSE | 6-gram LM MSE |
|---|---|---|
| *villa* | .178 (59) | .053 (14,499) |
| *pastor* | .191 (43) | .053 (23,386) |
| *miles* | .037 (233) | .052 (45,818) |
| *scientia* | .064 (61) | .022 (29,320) |
| *oratio* | .260 (89) | .022 (96,313) |
| Average | .146 (97) | .040 (52,742.4) |

**Table 4: Mean square error of Most Frequent Sense by Century (MFS/C) baseline in relation to the gold standard. The figures in parentheses represent the total number of observations comprising the distribution.**

# 7. TRACKING SENSE VARIATION OVER 2000 YEARS

After identifying the best-performing WSD classifier for this specific task of tracking historical variation (6-gram LM), we trained it on the full parallel text data and used it to classify the senses for the entire 389-million-word dated Latin corpus. While existing resources to chart the historical rise and fall of lexical trends have so far focused on word forms, we can now include sense information with this broader trend data.

Figures 5, 6 and 7 show three applications of this information: tracking variation in Latin senses; tracking variation in Latin word choice for a fixed sense; and comparing multiple lexical trends refined by specific sense.

[10]The single word for which the most frequent sense by century performs better than the 6-gram language model (*miles*) illustrates an important point: that MFS is a powerful predictor for words with low entropy in their sense distribution – e.g., those, like *miles*, in which a single sense is used $> 90\%$ of the time (thanks to an anonymous reviewer for pointing this out).

## 7.1 Variation in sense

Figure 5 displays a stacked chart representing the frequency with which "speech" and "prayer" were used as senses of the Latin word *oratio* over 2,100 years. Around year 0, for example, *oratio* was used with an overall frequency of .0003 (compared to all words), with its sense of "prayer" accounting for .00005 of that mass (16.7%) and "speech" accounting for the remainder. Year 1000 records a dip in the overall frequency of the Latin word form, but we can see that at this point the dominant sense has flipped: of the total frequency of .00015 with which *oratio* is used, approximately .00012 of that mass (80%) is comprised of "prayer." While we may have intuited this increasing use of *oratio* to signify "prayer" rather than "speech" over the course of the increasingly religious Middle Ages, we may not have suspected its return to "speech" with the rise of scientific discourse after the Renaissance (ca. 1500 CE). This data gives us a starting point to begin investigating this observation further.

## 7.2 Variation in lexical choice

While we have focused so far on leveraging parallel text data to induce sense information for Latin words, we can also use the same data to induce Latin senses for English words. One value of working in this direction is to identify the changing Latin lexical items that correspond to a fixed sense. One example of this is the changing distribution of Latin words corresponding to the English word "knight." In the Classical period, the Roman *equites* were a ruling class that held a social rank between the senate and the common people; this term is often translated into English as "knight." In the Middle Ages, the warrior class who followed a code of chivalry (whom we more commonly today refer to as "knights") were known as *miles* ("soldier" in Classical Latin). By tracing the changing Latin equivalents for a single English sense, we can see (in figure 6) how this variation unfolded over time – with *eques* the dominant sense of "knight" in the Classical era, suddenly giving way to the *miles* of the Middle Ages.

## 7.3 Comparing multiple lexical trends refined by sense

The third application of a large sense-tagged historical corpus is the ability to track the rise and fall not only of words, but also of the specific senses that they have historically been used with. Figure 7 illustrates this by charting the rise of specialized religious vocabulary in Latin denoting officers of the church – while *diaconus* ("deacon") and *papa* ("bishop, pope") are relatively unambiguous as Ecclesiastical Latin neologisms for these offices, *pastor* has several

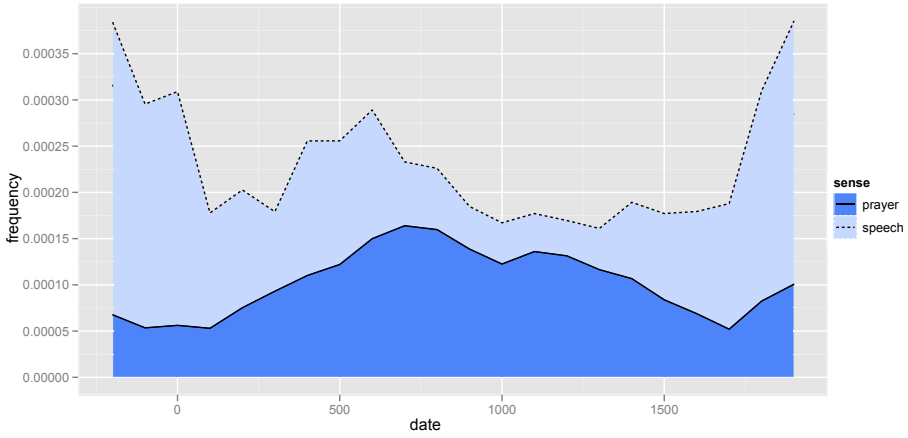


**Figure 5: Stacked charted showing English sense distribution for the Latin word *oratio* (96,313 instances).**

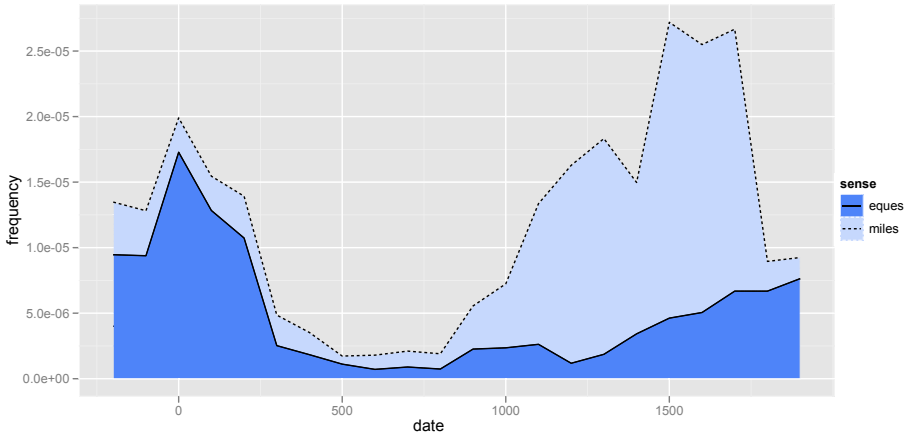


**Figure 6: Stacked chart showing Latin sense distribution for the English word *knight* (4,440 instances).**

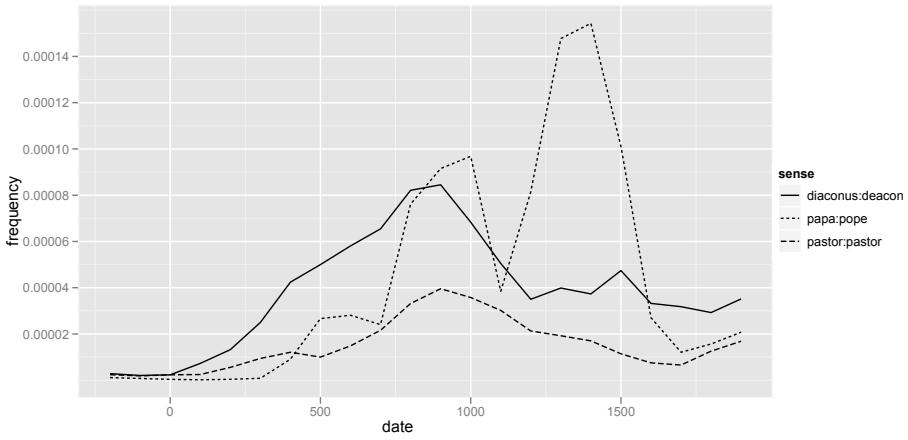


**Figure 7: Comparing multiple lexical trends refined by sense.**

possible senses (meaning both "pastor" and "shepherd"). By restricting our search specifically to the "pastor" sense of *pastor*, we are able to see its rise only after ca. 100 CE; if we considered only the trend information for the word form itself, the presence of these multiple senses would confound the distribution (preventing us from seeing quite so clearly

that *pastor* as "pastor" could only emerge in the Christian era). This may end up being the most useful application of this data, in that it allows us to move away from visualizing trend information for word forms and rather into the visualization of the concepts they embody. In our earlier example, for instance, if we were to apply this same strategy to tag the senses in a large dated corpus of contemporary English, we would be able to plot the rise of *radical* (meaning "excellent") against other slang terms with similar meaning without confounding other senses in the distribution – allowing us to see much more precisely the specific trend we're looking for.

# 8. CONCLUSION

The ability to supplement a dated historical corpus with tagged sense information for each of its words begins to open up a new dimension of both historical and linguistic inquiry. While researchers in both the humanities and sciences are now beginning to mine the huge cultural collections contained in million book libraries for such purposes, a sense-tagged collection gives us the ability to refine our searches beyond simple word tokens to the underlying concepts they embody.

There are a number of directions in which we can see this work continuing. First, the WSD classifiers we tested were all rooted in simple character or token ngrams; we would expect improvements with classifiers making use of more elaborate features (such as parts of speech). Second, the induction of the sense inventory and creation of training instances is highly dependent on the quality and volume of the underlying parallel texts. In our experiments so far, we made use of a small parallel corpus of 1.2M words (from which we found clean alignments for just over 500,000 words). More parallel data in this respect (either manually or automatically identified) will naturally lead to higher quality WSD classification.

While our experiments have focused so far on Latin due to its unique historical position of having been a *lingua franca* for over two thousand years, our methods are designed to be language independent, and lend themselves to reproducibility with any language for which there exists a large, dated historical collection and a smaller set of translations. As more and more books make their way into publicly available digital libraries, we hope to be able to apply these methods to a much broader range of languages in the future.

# 9. ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation (IIS-910884, "Mining a Million Scanned Books: Linguistic and Structure Analysis, Fast Expanded Search, and Improved OCR"), the National Endowment for the Humanities (PR-50013-08, "The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages") and the Digging into Data Challenge ("Towards Dynamic Variorum Editions"). Thanks are also due to Alison Darling, Elise Goodman-Tuchmayer, Daniel Libatique, Lee Marmor, John Owen and Erin Shanahan for their invaluable research assistance, and to four anonymous reviewers for their constructive comments. 

# 10. REFERENCES

[1] *The Oxford English Dictionary, Third Edition*. Oxford University Press, 2010.

[2] Alias-i. Lingpipe 4.0.1. http://alias-i.com/lingpipe, 2008.

[3] David Bamman, Alison Babeu, and Gregory Crane. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 11–20, New York, NY, USA, 2010. ACM.

[4] A. Baron, P. Rayson, and D. Archer. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, 20(1):41–67, 2009.

[5] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85, June 1990.

[6] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, 1993.

[7] Bob Carpenter. Character language models for chinese word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 169–172, Sydney, Australia, July 2006. Association for Computational Linguistics.

[8] Yee Seng Chan and Hwee Tou Ng. Scaling up word sense disambiguation via parallel texts. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, pages 1037–1042. AAAI Press, 2005.

[9] Dan Cohen. From Babel to Knowledge: Data Mining large Digital Collections. *D-Lib Magazine*, 12(3), 2006.

[10] Dan Cohen and Fred Gibbs. Victorian books: A distant reading of Victorian publications. http://victorianbooks.org/.

[11] G. Crane. What do you do with a million books. *D-Lib Magazine*, 12(3), 2006.

[12] Gregory Crane (ed.). The Perseus Digital Library. http://www.perseus.tufts.edu/hopper/opensource, 2011.

[13] Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[14] Dien Dinh. Building a training corpus for word sense disambiguation in English-to-Vietnamese machine translation. In *Proceedings of the 2002 COLING workshop on Machine translation in Asia - Volume 16*, COLING-MTIA '02, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[15] Victoria Fossum, Kevin Knight, and Steven Abney. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine*

*Translation*, StatMT '08, pages 44–52, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[16] Alexander Fraser and Daniel Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007.

[17] Wilhelm Freund, editor. *Wörterbuch der lateinischen Sprache: nach historisch-genetischen Principien, mit steter Berücksichtigung der Grammatik, Synonymik und Alterthumskunde*. Teubner, Leipzig, 1834-1840.

[18] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[19] Charles Henry and Kathlin Smith. Ghostlier demarcations: Large-scale text digitization projects and their utility for contemporary humanities scholarship. In *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*, pages 106–115, Washington, 2010. Council on Library and Information Resources.

[20] David Holmes. Authorship attribution. *Computers and the Humanities*, 28:87–106, 1994.

[21] Nancy Ide, Tomaz Erjavec, and Dan Tufis. Automatic sense tagging using parallel corpora. In *NLPRS*, pages 83–90, 2001.

[22] Nancy Ide, Tomaz Erjavec, and Dan Tufis. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8*, WSD '02, pages 61–66, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[23] Sang-Bum Kim, Hee-Cheol Seo, and Hae-Chang Rim. Information retrieval using word senses: root sense tagging approach. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 258–265, New York, NY, USA, 2004. ACM.

[24] Phillip Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*, Phuket, Thailand, 2005.

[25] Charles T. Lewis and Charles Short, editors. *A Latin Dictionary*. Clarendon Press, Oxford, 1879.

[26] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 2010.

[27] David Mimno and Andrew McCallum. Organizing the OCA: learning faceted subjects from a library of digital books. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 376–385, New York, NY, USA, 2007. ACM.

[28] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

[29] Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[30] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[31] Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: an empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 455–462, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[32] Jan Frederick Niermeyer. *Mediae Latinitatis Lexicon Minus*. Brill, Leiden, 1976.

[33] Geoffrey Nunberg. Google's book search: A disaster for scholars. *The Chronicle of Higher Education*, August 31, 2009.

[34] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

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

[36] Salim Roukos, David Graff, and Dan Melamed. Hansard French/English.

[37] Ludwig Schütz. *Thomas-Lexikon*. F. Schoningh, 1895.

[38] H. Schütze and J. O. Pedersen. Information retrieval based on word senses. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas NV, 1995.

[39] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: a statistical approach. *Comput. Linguist.*, 22:1–38, March 1996.

[40] Lucia Specia, Maria Das Graças, Volpe Nunes, and Mark Stevenson. Exploiting parallel texts to produce a multilingual sense tagged corpus for word sense disambiguation. In *In Proceedings of RANLP-05, Borovets*, pages 525–531, 2005.

[41] Christopher Stokoe, Michael P. Oakes, and John Tait. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 159–166, New York, NY, USA, 2003. ACM.

[42] Benedikt Szmrecsanyi. Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):279–296, 2008.

[43] William John Teahan. Text classification and segmentation using minimum cross-entropy. In Joseph-Jean Mariani and Donna Harman, editors, *RIAO*, pages 943–961. CID, 2000.