Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3252–3259
Marseille, 11–16 May 2020

Computational Etymology and Word Emergence

Winston Wu, David Yarowsky
Center for Language and Speech Processing
Johns Hopkins University
(wswu, yarowsky)@jhu.edu

Abstract

We developed an extensible, comprehensive Wiktionary parser that improves over several existing parsers. We predict the etymology of a word across the full range of etymology types and languages in Wiktionary, showing improvements over a strong baseline. We also model word emergence and show the application of etymology in modeling this phenomenon. We release our parser to further research in this understudied field.

Keywords: etymology, wiktionary, word emergence

1. Introduction

Since antiquity, scholars have been fascinated by etymology, the study of words' origins. In modern days, there exist numerous etymological dictionaries for select languages (e.g. English (Skeat, 1884; Partridge, 2006), Albanian (Orel, 1998), or Old Chinese (Schuessler, 2007)) as well as language families (e.g. Italic (De Vaan, 2018), Slavic (Derksen, 2007), or Altaic (Starostin et al., 2003)). Many of these improve and expand upon existing dictionaries as new evidence comes to light about the relationships between languages and their words. However, until very recently, the discovery of these relationships has not been computational driven.

In recent years, researchers have developed computational methods for determining relationships between languages (see Nichols and Warnow (2008) and Dunn (2015) for surveys of the field of linguistic phylogenetics), but there is little work on computationally learning the etymological relationships between individual words. Researchers have constructed a Proto-Indo European lexicon (Pyysalo, 2017), and showed that knowing a word's etymology can help with text classification tasks (Fang et al., 2009; Nastase and Strapparava, 2013) and reconstructing language phylogenies (Nouri and Yangarber, 2016).

In an era of abundant linguistic data, we seek to address the dearth of computational approaches to modeling etymology. To this end, we develop a parser that extracts etymology information and translations from Wiktionary, an open multilingual dictionary. Using this data, we present several approaches to model when (*word emergence*), from where, and how a word enters a language. We employ RNN-based models to accurately predict a word's formation mechanism, parent language, and year of emergence. For emergence, we also experiment with various historical data-driven models. Our methods are language-independent and are applicable for improving existing etymology determinations that may be incorrect, as well as providing etymology for words that may not have an existing etymological entries, both in low- and high-resource languages.

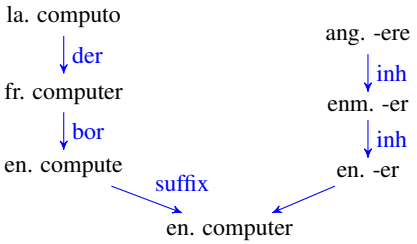

Figure 1: Wiktionary etymology graph of the English word *computer*. Etymological relationships are shown in blue.

2. Wiktionary Etymology

Wiktionary[1] is a large, free, online multilingual dictionary that is editable by anyone in the world. In addition to containing information found in traditional dictionaries (pronunciations, part of speech, definitions), it is rich source of other information that help one understand a word, including etymology, synonyms, antonyms, translations, derived terms, related terms, and even quotations. In this work, we focus on etymology, though our parser does extract these other types of information.

The etymological relationships between words[2] can be represented as a directed graph, where the nodes are words and the edges are etymological relationships. For example (Figure 1), according to Wiktionary, the etymology for the English word *computer* is *compute* + the suffix *-er*. The word *compute* is borrowed from the French *computer*, which is derived from the Latin *computo*. The *-er* suffix is inherited from the Middle English *-er*, which is inherited from the Old English (Anglo-Saxon) *-ere*.

There are a few existing efforts to parse etymological information from Wiktionary at different granularities to construct such graphs: Etymological WordNet (de Melo, 2014) contains coarse-grained relations between pairs of words. The relations include is-derived-from, has-derived-form, etymologically-related, etymological-origin-of, etymology, and variant:orthography. This data covers 2.8 million terms. EtymDB (Sagot, 2017; Fourrier and Sagot, 2020)

[1] wiktionary.org

[2] Wiktionary contains separate entries for affixes like *-er*, so we call them "words" here.

Label	Count	Label	Count
affix	28366	derived	132404
back-form	24	inherited	159239
blend	144	mention	265220
borrowed	104817	noncognate	188
calque	964	prefix	18169
clipping	44	semantic loan	15
cognate	32095	short for	3
compound	42524	suffix	49505
confix	2185		

Table 1: Etymological relationships we extracted from Wiktionary. Note that cognate and noncognate relationships are bidirectional relations, while the rest are unidirectional.

extracted more fine-grained relations including borrowing, compound, cognate, derived, derived-prefix, derived-suffix, and inherited. Both of these projects do not make use of the full range of etymological relationships present in Wiktionary. Thus, we were motivated to develop our own Wiktionary parser that is both comprehensive and extensible: it can extract the etymological information and many other types of information annotated in Wiktionary, and it is easy to use and extend for further research.

Wiktionary has a set of guidelines[3] for annotators to document etymological relations. For our parser, we developed a variety of heuristics to parse the unstructured Wikitext that makes up the the etymology section of a page (see Figure 2). Wikitext is a wiki markup language used by Wiktionary and Wikipedia. Table 1 summarizes the etymology information we extracted.

Besides the challenges of unstructured text, the human element also poses challenges: annotators are sometimes inconsistent in following the Wiktionary guidelines. According to the guidelines, `inherited` is used for words that are from an earlier stage of the same language, while `borrowed` is used for words coming from other languages. The `derived` label is intended as a catch-all label for words that are not borrowed or inherited, whereas a stricter definition of (morphological) derivation would be a word that is formed from another existing word, often with an affix. The `affix` label is another catch-all for words that do not fit into the other affixal categories prefix, suffix, or confix, or they may have multiple prefixes and/or suffixes. Table 2 samples some inconsistencies with the etymology annotations found in Wiktionary. While it is not possible to exactly determine the number of inconsistencies, the large number of etymological relationships labeled as derived and affix indicates that there are many words for which a precise relationship is not known.

3. Etymology Prediction

To improve upon and expand the etymology annotations in Wiktionary, a natural solution is to develop a computational model to solve the following task: given a (language, word) pair, we seek to predict both the *relationship* of etymology and *which language* the word came from. In the latter half of this paper, we address the question of *when* a word entered the language. Using the etymology data we parsed, we run three experimental settings spanning different granularities of etymology prediction:

1. Input: Language Code + Word
   Output: Coarse Relationship
2. Input: Language Code + Word
   Output: Fine Relationship
3. Input: Language Code + Word + Relationship
   Output: Parent Language

For the fine-grained mechanism prediction, we use the etymology labels affix, borrowing, compound, inherited, prefix, and suffix. Notably, we do not include the `derived` label due to the noise it adds to the dataset.[4] For predicting coarse-grained mechanism, we use two classes: borrowing/inheritance, and compositional, which encompasses compound, affix, prefix, and suffix. For language prediction, to make the problem computationally tractable, we predict the top five most frequent parent languages of a word, or "other" if the parent word's language is not in the top five.

We frame the task of etymology prediction as a multilabel classification task, where the input is a sequence containing the word's ISO639-3 language code and the individual characters in the word, and the output is a probability that the word belongs to one of the etymological relationship labels (note a word can have multiple labels, e.g. "apicide", which is borrowed from the Latin *apis* and contains the *-cide* suffix). For our model, we used a LSTM with an embedding dimension of 128 and hidden dimension of 128. The output of the last hidden state is passed to a fully connected layer with a sigmoid activation function. We used binary cross entropy as the loss and Adam as the optimizer with learning rate 0.001. The models were implemented using PyTorch. The data setup is shown in Figure 3.

We run these experiments on several languages around the world spanning various levels of resource-ness. In addition, we train a single multilingual system that can handle all the 3146 languages in our dataset by simply adding a language token in the input (Figure 3). We employ an 80-10-10 train-dev-test split and test with the model with the lowest loss on the dev set.

Word	Mechanism	Parent	Correct
analyst	derived	(fr) analyste	borrowed
blind	derived	(ang) blind	inherited
agricultural	affix	agriculture + -al	suffix
peatbog	affix	peat + bog	compound
acetal	compound	acetic + alcohol	blend

Table 2: Examples of noisy Wiktionary etymology labels for some English words. ang is Old English

[3] `https://en.wiktionary.org/wiki/Wiktionary:Templates#Etymology`

[4] In our initial experiments, we included words with the `der` label, but we found that the models had trouble distinguishing derivations from borrowings. Further analysis showed that words labeled as derived are noisy, as previously discussed.

Displayed Text:	From Middle English cat, catte, from Old English catt (“male cat”), catte (“female cat”), from Proto-Germanic *kattuz.
Wiki Markup:	From {{inh|en|enm|cat}}, {{m|enm|catte}}, from {{inh|en|ang|catt||male cat}}, {{m|ang|catte||female cat}}, from {{inh|en|gem-pro|*kattuz}}.

Figure 2: Etymology of the English word *cat*.

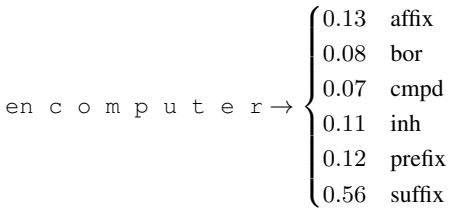

Figure 3: Setup of the fine-grained mechanism prediction task. For the language-specific setting, the leading language token (here, `en`) would not be present, and in the parent language prediction task, an additional token for the mechanism (e.g. `suffix`) would be appended.

	Coarse		Fine		Language	
Lang	Base	Ours	Base	Ours	Base	Ours
af	**0.92**	0.91	0.79	0.79	0.72	**0.81**
en	0.52	**0.76**	0.34	**0.51**	0.42	**0.80**
it	0.51	**0.84**	0.35	**0.57**	0.48	**0.68**
ja	0.89	**0.92**	0.81	**0.85**	0.58	**0.70**
sw	0.70	**0.79**	0.48	**0.59**	0.32	**0.52**
zh	0.98	0.98	0.82	**0.86**	0.36	**0.54**
all	0.66	**0.83**	0.39	**0.53**	0.67	**0.79**

Table 3: Results on the etymology prediction tasks. The metric is accuracy.

3.1. Results and Analysis

Results are in Table 3. For almost all languages and settings, our neural method beats a strong majority baseline,[5] though it falls short when the class imbalance is high. Performance on Japanese (ja) beats the high-performing baseline because of a feature of the Japanese writing system: foreign words are written in katakana, while native words are written in hiragana or kanji. Thus foreign words are easily distinguished as borrowing due to differences in the script. For Afrikaans (af) and Chinese (zh), we believe the performance is largely due to the tiny amount of training data (1.1K and 1.7K training examples, respectively), though it is remarkable that with such little data, a neural system can learn to predict etymology with such high accuracy. Equally remarkable is our finding that the spelling of a word alone is adequate to identify a word's etymology. This indicates that a language's prior on whether it prefers borrowing, inheritance, or compositional means for word formation is encoded in the spelling of the word. We will show later that a word's spelling, along with some etymology information, can predict a word's emergence year.

Due to the authors' familiarity with the language, we present analyses of some mistakes that the English models made. In the coarse mechanism prediction task (Table 4), the incorrect classification of borrowed/inherited words as compositional included borrowed words like *Prachuap Khiri Khan* that contained characters like hyphens or spaces that usually indicate compositionality, or words like *upright* that are technically inherited but could also be compositionally analyzed or were compositionally formed in an ancestor language. For words incorrectly classified as borrowing/inheritance, these are likely due to character sequences that are not common in the English language (e.g. the two components of *cumulonimbus* are borrowed from Latin).

For the English fine mechanism prediction task (confusion matrix in Table 5), the model incorrectly labels a large percentage of compounds as borrowings, and inherited words as borrowing or suffixes. Some mistakes are shown in Table 6. Many words incorrectly labeled as suffixed are due to the presence of a suffixal ending (-er or -ly); the suffixation of *drencher* and *gladfully* occurred in Middle English, so they are technically inherited, and words like *unmaidenly* and *macrobiotics* contain both a prefix and suffix. Words like *lesbro* or *Kleinberg* do not have a typical English spelling and are thus incorrectly labeled as borrowings. Other words like *appertain* and *injurious* are hard to distinguish as borrowed or inherited, due to the assimilation of Romance words due to Norman French.

Finally, for the language prediction task (confusion matrix in Table 7), the primary mistakes seem to be classifying French as other and other as Middle English. Some examples of misclassifying French borrowings include *sanitary* and *chagrin*. One explanation for these mistakes is that the

Word	Pred	Gold	Confidence
tête-à-tête	comp	borinh	0.58
Prachuap Khiri Khan	comp	borinh	0.56
upright	comp	borinh	0.54
nurturant	borinh	comp	0.70
autovacuum	borinh	comp	0.56
cumulonimbus	borinh	comp	0.64

Table 4: Mistakes in the coarse mechanism prediction task.

	affix	bor	comp	inh	prefix	suffix
affix	27	23	13	0	23	58
bor	0	1108	19	61	24	82
comp	3	132	109	9	20	53
inh	1	137	25	286	19	138
prefix	5	43	6	24	223	39
suffix	4	99	22	21	34	587

Table 5: Confusion matrix of predictions for English, where rows are the true labels and columns are predictions. For visualization purposes, this is limited to truth and predictions that only contain a single label.

[5]The majority baseline is to pick the most common etymological class within a language.

Word	Pred	Gold	Confidence
drencher	suffix	inh	0.55
gladfully	suffix	inh	0.72
unmaidenly	suffix	affix	0.55
aggrandize	suffix	bor	0.84
macrobiotics	prefix	affix	0.59
lesbro	bor	comp	0.75
Kleinberg	bor	comp	0.82
appertain	bor	inh	0.63
injurious	bor	inh	0.68

Table 6: Mistakes in the fine mechanism prediction task.

	en	enm	fr	la	grc	other
en	1822	0	1	11	8	34
enm	2	707	0	0	0	3
fr	34	0	110	2	13	109
grc	13	0	1	47	3	26
la	25	9	7	8	120	82
other	39	101	21	4	38	880

Table 7: Confusion matrix for predicting an English word's ancestor language.

presence of so many Romance words has diluted the Germanic spelling pool and thus confuses the model. Many of the misclassifying "other" mistakes included words that were inherited from Old English, like *font* and *cress*. Similar analysis can be performed for other languages, and future work includes collapsing languages of a single line (like Old, Middle, and Modern English) into a single label.

4. Translations

Wiktionary also contains translations. Wiktionary provides an API to access translations, but this is not convenient for bulk analysis. Within the scientific literature, there are a couple projects that have extracted data directly from the Wiktionary dumps: WIKT2DICT (Acs et al., 2013) extracts translations from the translation tables in the Wiktionary articles. This codebase supports triangulation between language to discover new translations. Kirov et al. (2016) (henceforth KIROV) also extracts translations from translation tables, in addition to morphological paradigms, which were the main focus of their work.

Our Wiktionary parser extracts translations from translation tables as well as from *definitions* of the word. Definitions are a valuable source of translations, and we are not aware of existing work that extracts translations from definitions. Extracting translations from definitions is a challenging task, since definitions are unstructured and generally freeform text, while translation tables are structured. We utilized a combination of string regex matching and some heuristics to convert the definition strings into short lexical translations.

Below, we analyze translations extracted using various systems. In these comparisons, we used the English Wiktionary dump with articles only from May 4, 2019. We ran WIKT2DICT with a small modification to the code to allow extracting translations for all languages (rather than the small subset that they previously defined). KIROV's parse is from an older (2015) edition of Wiktionary. For each parse, we removed duplicate translations and kept only foreign-English translation pairs.

Parser	Terms	# Langs
Acs et al. (2013)	1589383	2417
Kirov et al. (2016)	1577374	2165
Ours (translations)	1575392	2406
Ours (definitions)	1181666	2800
Ours (both)	2296208	3640

Table 8: Number of foreign-English translations extracted by various translation extraction systems.

Wiktionary contain 3931 languages.[6] WIKT2DICT parse contains 2367 languages, and KIROV's contains 2166. Both share 1640 languages, while separately WIKT2DICT has 727 not in KIROV, and KIROV has 526. As shown in Table 8, extracting translations from definitions covers considerably more languages and terms than just translation tables.

WIKT2DICT's and our translation extraction from translation tables are very similar, which makes sense; we are using the same data. The differences largely come from WIKT2DICT not postprocessing its output, so they include entries like Finnish *[[puhua]] [[ummet ja lammet]]* (with brackets), or words with unmatched parentheses. There is also some variation in translations, usually in proper nouns: WIKT2DICT has "Solar System", while KIROV has "the Solar System" as translations for the Zaza word *Sistemê Roci*. In terms of the number of foreign words and languages where WIKT2DICT and our method extracted more words than KIROV, this is likely due to users simply adding more words since the time KIROV's translations were extracted (we were not able to obtain the code to run their extraction). On the other hand, for some languages, KIROV was able to extract more translations due to parsing morphological information outside of the translation tables. Our innovation of extracting translation from definitions substantially increases the number of available translations.

5. Open Source

Our Wiktionary parser parses a public XML dump of Wiktionary.[7] Besides etymology and translations, our parser has basic functionality for parsing other information present in Wiktionary. Most of our efforts were focused on parsing etymology, and there are many improvements that can be made to our parser. We release our Wiktionary parser at `https://github.com/wswu/yawipa` to solicit improvements and encourage further research with our tool.

6. Word Emergence

One aspect of etymology that Wiktionary does not specifically contain is information about *when* a word entered

[6] As of April 2019. `https://en.wiktionary.org/wiki/Wiktionary:Statistics`

[7] `https://dumps.wikimedia.org/enwiktionary/latest/enwiktionary-latest-pages-articles.xml.bz2`

the language. Based on a word's parent language, one can approximate the date of entry, e.g. a word borrowed into English from Middle French would have entered sometime around 1300–1600, the lifespan of Middle French. However, this is imprecise.

Identifying the date of first use of a word has historically involved lexicographers scouring old literature and manuscripts. For high-resource languages like English, existing work (e.g. Fischer (1998)) details different processes of neologisms, like clipping and borrowing (which are annotated in Wiktionary; we leave the modeling of this to future work). Dictionaries of neologisms (e.g. Algeo and Algeo (1993)) list years or even specific dates of the first use of a word. In recent years, people have started investigating neologisms computationally (e.g. Ahmad (2000; Kerremans et al. (2011)), and online dictionaries and datasets provide convenient electronic versions of a word's year of first use. However, these resources vary in the amount of information they provide and are often limited to a handful of languages. Most similar to this work is Petersen et al. (2012), who analyze word birth and death, and Ryskina et al. (2020), who examine factors that affect the creation of neologisms through the lens of word embeddings.

In the remainder of this paper, we present our work on modeling word emergence, an integral part of a word's etymology. We distinguish between, word birth, the year a word was first recorded as being used, and word *emergence*, the year in which the word starts gaining popularity in usage, and we argue that the latter is more informative than the former. We examine two datasets of historical word usage, the Google N-Grams corpus (Michel et al., 2011) and Merriam-Webster's Dictionary (Merriam-Webster, 2006), and propose several methods for predicting the year of emergence in any language.

7. Historical Word Data

There are few existing sources of historical word usage, especially for languages other than English. Our work utilizes data from two sources:

Google N-Grams (GNG). The Google N-Grams project (Michel et al., 2011) collects statistics of how many times a particular n-gram appears in how many books published in a given year. Data are available for 1- to 5-grams, and the languages covered are English, Chinese, French, German, Italian, Russian, and Spanish. The oldest books date from the 1500s, while the most recent are from 2008. GNG was constructed by using OCR to extract text. This process is not perfect, and we show that our methods can potentially detect these errors. The total number of words in GNG per year is shown in Figure 4.

Merriam-Webster Dictionary (MW). This dictionary contains the year of first use for words in the English language. Before 1500, the data is more coarse-grained, and years are grouped by century; the oldest designation is *before 12th century*. The most recent words are from 2016. The data contained in MW is the first recorded year the word was used in print or writing.[8]

[8]Which is not necessarily when it was added to the dictionary. And the first attestation in print is also not necessarily the first strict usage of the word. Generally, words are introduced in speech before they are written down.

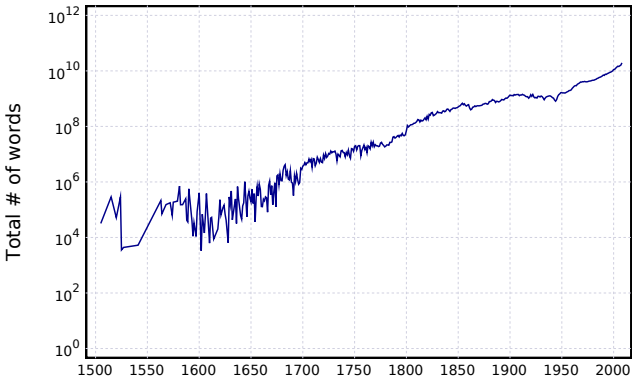

Figure 4: Total number of words in GNG per year. Note the log scale on the y-axis.

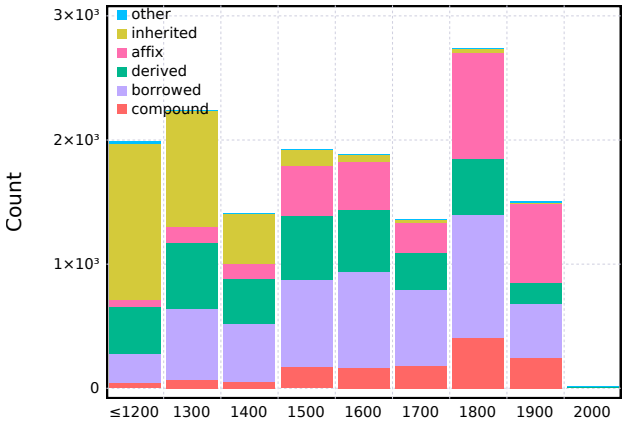

Figure 5: Sources of word formation for English words by century of word birth.

8. Models and Experiments

8.1. RNN-based

We first try the RNN-based approach as we have done for modeling etymology, as a sanity-check to see if modeling word birth is indeed possible. In this experiment, we use MW as the training data, restricting the words to those for which we have extracted etymology information (19,081 words). We know that different time periods in a language's history are characterized by different distributions of word formation (Figure 5), so we are interested in assessing the contribution of etymology to the task of predicting word birth. We train a character-based neural model in a 70-15-15 train-dev-test split using the same setup and hyperparameters as in Section 3.. We run an ablation study with four settings: only characters, characters + the parent language, characters + the word formation mechanism (bor, inh, etc.), and characters + mechanism + parent language. We experiment on these words and a reduced set whose birth year is $\geq$ 1500 (a total of 11,494 words), because in the MW dataset, years before 1500 are grouped by century. Results are presented in Table 9 (the metric is mean average error between the true year and the predicted year) and example predictions in Table 10.

Setting	MAE (all)	MAE (year $\geq$ 1500)
Chars	253.0	118.9
Chars + Mechanism	180.9	112.8
Chars + Parent Language	157.9	103.2
Chars + Mech + Lang	157.3	101.9

Table 9: Ablation study of predicting word birth.

Word	True	C	CM	CL	CML
hippopotamus (bor, la)	1563	1682	1673	1662	1650
macrobiotic (affix, en)	1965	1804	1886	1819	1852
manucure (bor, fr)	1877	1723	1718	1739	1771
tae kwon do (bor, ko)	1967	1791	1937	1878	1955
eureka (der, grc)	1603	1750	1711	1783	1731

Table 10: A sample of predictions of birth year. C, CM, CL, and CML correspond to the settings in Table 9.

Restricting the data to words born after 1500 results in a noticeable improvement, though even with the added noise of old words, the LSTM model can predict a word's birth year within two centuries. We see improvements in performance by adding etymological information, which demonstrates that while a word's spelling encodes at least some information about a word's birth year, and knowing how and what language a word came from can help narrow the predicted time range of a word, allowing an average prediction within a century. Specific examples in Table 10 reveal that adding more etymology information tends to, but does not always improve predictions. These results indicate that word birth is modelable, but there are potentially better methods for doing so.

8.2. Examining Historical Data

The year of first use is somewhat problematic. We already noted that older words have a less precise birth year. OCR errors are also common; the classic example is the long s (ſ), which was used up until around 1800. OCR software have difficulty distinguishing between this letter and the letter 'f', so words like "funk" would appear to have a much earlier year of first use than in reality. And a word's birth year is not necessarily informative: the word genomics

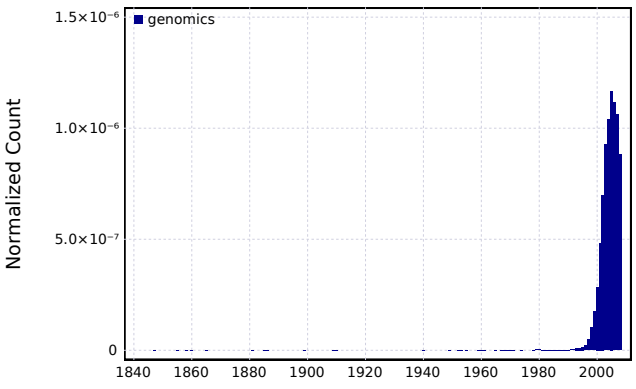

Figure 6: Normalized counts of the word "genomics" in GNG. Note the tiny bar at year 1847.

(Figure 6) was first used in 1847, but did not gain popularity until the late 1900s.[9] Thus, we are interested in when a word gains traction, or emerges into the language, rather than the absolute first use. We devise several models of word emergence, following some preprocessing:

First, we smooth the GNG count data by averaging the counts of the current year with those of the immediately preceding and following year. Then these counts are normalized by dividing by the total number of words in that year. This represents the percentage of the total number of words that a given word contributed in any given year.[10] We propose several data-driven formulas for extracting a word's emergence year from GNG data:

GNG First Attestation. Perhaps the simplest model: use the first year a word was attested in GNG. This may be problematic for younger (more recent) words, e.g. *genomics*.

% of median threshold. Petersen et al. (2012) used a threshold of 0.05 $\times$ the median normalized count. They consider the first year a word's count crosses this threshold as its emergence year.

% of max threshold. A similar threshold heuristic: the first year in which the normalized count crosses 1% of a word's maximum normalized count is considered the emergence year.

Curve Fitting. The above heuristics are simple but they do not utilize all the data. To take into account trends in the data, we employ locally estimated scatterplot smoothing (LOESS) to fit a curve to the data. LOESS is a nonparametric regression method that fits a low-degree polynomial (in our case, degree 2) to a sliding window of the data. We chose this model because in many cases, humans can look at a graph of word usage and easily identify a word's emergence year just by seeing where there is a sudden change in the shape of the curve. This curve-fitting model predicts the emergence year of a word as the most recent year[11] where the LOESS curve crosses from negative to positive. If the curve never dips below the x-axis, then it designates the emergence year as the year at the curve's minimum value. We experimented with different settings for the span parameter, which controls the size of the sliding window.

Derivative. The final model we evaluate also exploits trends in the data: we take the derivative of the LOESS regression curve and identify the first year where it becomes positive. This indicates the beginning of an upward trend in the number of occurrences.

[9] The term was coined in 1986 (Yadav, 2007).

[10] One observation with normalizing by the total number of words is that the usage of an old word may be diluted over time. For example, the normalized count of the Spanish word "agua" was 0.00298 in 1522 and 0.00023 in 2009. While in 1522, there was a smaller total number of words, the occurrences of "agua" made up a larger percentage of the total than in 2009, when the Spanish language had a much larger vocabulary size. Petersen et al. (2012) describes this phenomenon as "competing actors in a system of finite resources."

[11] There are cases where the curve may cross multiple times, especially if the word is older.

Year	# Words	First	Median	Max	C 0.3	C 0.4	C 0.5	C 0.6	C 0.7	Der	# Words	C+M+L
1500-1549	2360	**96.7**	96.7	96.8	299.5	311.4	319.3	326.4	337.6	145.3	39	199.2
1550-1599	4491	**89.9**	90.2	90.1	255.8	268.3	275.4	281.3	289.3	126.6	181	149.3
1600-1649	4230	**88.2**	88.6	88.6	214.3	225.7	232.5	236.6	240.8	111.2	288	129.4
1650-1699	3003	**81.9**	82.6	82.7	164.7	173.0	178.3	181.5	184.9	89.6	160	95.1
1700-1749	2108	80.8	81.9	81.8	117.8	127.3	132.6	135.5	138.6	**70.3**	104	65.2
1750-1799	3030	80.8	81.8	81.7	79.3	85.9	89.4	91.5	94.8	**53.1**	121	64.4
1800-1849	6053	77.8	78.9	78.7	47.4	52.8	55.3	57.2	58.6	**46.3**	195	56.2
1850-1899	8001	75.3	73.5	73.7	**34.5**	34.3	35.3	36.3	38.1	45.2	228	74.0
1900-1949	6801	83.6	75.5	75.6	30.2	**26.6**	26.7	27.0	28.0	51.6	229	95.4
1950-1999	3420	101.0	89.2	87.3	32.6	27.9	26.2	25.2	**23.4**	66.5	156	130.5
2000-2049	47	133.5	131.4	123.9	41.4	40.9	42.4	41.5	**38.7**	104.4	24	166.4

Table 11: Mean absolute error in years for different models. C 0.3 denotes the curve fitting model with span of 0.3.

8.3. Results and Analysis

As far as we are aware, there are no existing datasets for word emergence, so we evaluate each of the above models in predicting a word's birth year as a proxy for emergence year. We utilize the intersection of MW words with unigrams from GNG, for a total of 57,015 words. Each model was evaluated on mean absolute error (in years) with respect to the gold birth years of MW.

We examine the performance of each model in 50-year increments (Table 11), revealing noticeable differences in model performance. On average, the simple heuristic models (First, Median, and Max) predict birth year within a century, though accuracy decreases for more recent words. On the contrary, the curve fitting models perform poorly on older words but greatly outperform the heuristic models on recent words. The derivative model, which uses the fitted curve, performs best around 1700-1800, but accuracy falls off for older and younger words. The RNN model exhibits a similar U-shaped performance curve.[12] For the non-neural models, First, Median, and Max are consistently within 100 years of the gold, the curve fitting and derivative models can greatly improve upon these simpler models. While Median and Max do not perform as well, they more accurately model the phenomenon of word emergence than First.

Figures 7 and 8 show each model's predictions on an older word *machine* and a younger word *scam*, respectively. MW lists the first use of *machine* as 1545, though it was not found in GNG until after 1700. For *scam*, MW lists the first use year as 1963, though the word seems to have been in use at a low frequency since 1700.[13] Because of this, the simpler models give an incorrect birth year, while the curve fitting model correctly identifies the start of a period of exponential grow around 1960. Thus the curve-fitting model works well as a model for word emergence. We saw similar results for GNG Spanish and French data, though we did not have gold data to formally compare against.

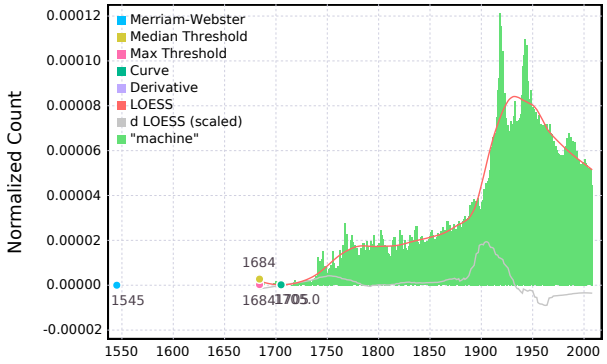

Figure 7: Plots of each model's birth year predictions on the word "machine".

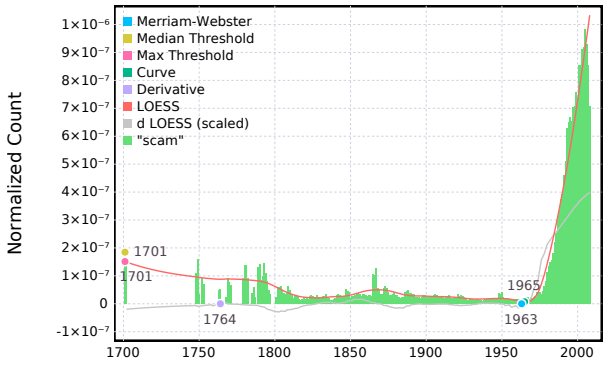

Figure 8: Plots of each model's birth year predictions on the word "scam".

9. Conclusion

We presented a Wiktionary parser with comprehensive support for parsing etymology and translations. We introduced the task of etymology prediction, where given a word, one should predict its parent word and language. We performed preliminary experiments, showing the effectiveness of multilingual models on this task. Regarding word emergence, an aspect not found in Wiktionary etymology, we experimented with numerous models in modeling word emergence using historical word data. All of our methods are language independent, and we see future application in correcting misannotations and increasing coverage of etymological dictionaries for low-resource languages.

[12] Results for the best RNN-based model (chars + mechanism + language) were included in this table for comparison, but the results are not directly comparable because unlike the other models, the neural model uses a training and development set, so the test set is substantially smaller.

[13] The etymology of *scam* is uncertain. The earlier usages in Google N-grams are likely OCR errors of the word *seam*.

10. Bibliographical References

Acs, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ahmad, K. (2000). Neologisms, nonces and word formation. In *Proceedings of the Ninth EURALEX International Congress*, page 71.

Algeo, J. and Algeo, A. S. (1993). *Fifty years among the new words: A dictionary of neologisms 1941-1991*. Cambridge University Press.

de Melo, G. (2014). Etymological Wordnet: Tracing the history of words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1148–1154, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

De Vaan, M. (2018). *Etymological dictionary of Latin and the other Italic languages*, volume 7. LEIDEN· BOSTON, 2008.

Derksen, R. (2007). *Etymological dictionary of the Slavic inherited lexicon*. Brill.

Dunn, M. (2015). Language phylogenies. In *The Routledge handbook of historical linguistics*, pages 208–229. Routledge.

Fang, A. C., Li, W., and Ide, N. (2009). Latin etymologies as features on BNC text categorization. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 662–669, Hong Kong, December. City University of Hong Kong.

Fischer, R. (1998). *Lexical change in present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms*, volume 17. Gunter Narr Verlag.

Fourrier, C. and Sagot, B. (2020). Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB-2.0. In *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France. (to appear).

Kerremans, D., Stegmayr, S., and Schmid, H.-J. (2011). The neocrawler: identifying and retrieving neologisms from the internet and monitoring on-going change. *Current methods in historical semantics*, pages 59–96.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In *LREC*.

Merriam-Webster. (2006). *The Merriam-Webster Dictionary*. Merriam-Webster, Incorporated.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Nastase, V. and Strapparava, C. (2013). Bridging languages through etymology: The case of cross language text categorization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 651–659, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nichols, J. and Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

Nouri, J. and Yangarber, R. (2016). From alignment of etymological data to phylogenetic inference via population genetics. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 27–37, Berlin, August. Association for Computational Linguistics.

Orel, V. (1998). *Albanian etymological dictionary*. Brill.

Partridge, E. (2006). *Origins: A short etymological dictionary of modern English*. Routledge.

Petersen, A. M., Tenenbaum, J., Havlin, S., and Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific reports*, 2:313.

Pyysalo, J. (2017). Proto-Indo-European lexicon: The generative etymological dictionary of Indo-European languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 259–262, Gothenburg, Sweden, May. Association for Computational Linguistics.

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D. R., and Tsvetkov, Y. (2020). Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *Proceedings of the Society for Computation in Linguistics*, 3(1):43–52.

Sagot, B. (2017). Extracting an etymological database from wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, pages 716–728.

Schuessler, A. (2007). *ABC etymological dictionary of Old Chinese*. University of Hawaii Press.

Skeat, W. W. (1884). *An etymological dictionary of the English language*. Clarendon Press.

Starostin, S. A., Dybo, A., Mudrak, O., and Gruntov, I. (2003). *Etymological dictionary of the Altaic languages*, volume 3. Brill Leiden.

Yadav, S. P. (2007). The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT*, 18(5):277.