EtymoLink: A Structured English Etymology Dataset

Yuan Gao[1, 2], Weiwei Sun[1],
[1]Department of Science and Technology, University of Cambridge, U.K.
[2]ALTA Institue, University of Cambridge, U.K.
{yg386,ws390}@cl.cam.ac.uk

Abstract

Etymology, and the field of lexicography, is often constrained by unstructured data formats buried in scholarly articles and dictionaries. This paper presents a methodology and an empirical study for creating a structured etymological dataset suitable for computational analysis. Using data from the Online Etymology Dictionary (Etymonline), we manually annotated a subset of entries to establish a high-quality ground-truth dataset and fine-tuned the FLAN-T5-base model to extract structured etymological relationships automatically. The resulting dataset contains over 103,000 relationships covering 63,603 English lexical terms. Our findings emphasise feasibility of using large language models for structuring lexicographical data, exploring the transferability of the model to other dictionary datasets with no additional manual annotation.

1 Introduction

Etymology, is the study of the origin and historical development of words. The etymological understanding of words not only reveals their origins, but also the cultural and historical contexts that have shaped their contemporary meanings. Figure 1 shows the etymology of the English word "research", meaning diligent and systematic inquiry. It is commonly understood that the word is made up of the prefix *"re-"*, meaning 'again', and the root *"search"*. The etymological trace leads further back to the reduplicated form of the Proto-Indo-European (PIE) root **sker-*, which means to cut or divide. The duplication of **sker* suggests a repetitive action, along with the prefix *"re-"* which adds another sense of intensity and repetitiveness. This understanding, traced all the way to Proto-Indo-European roots not only uncovers the origin, but also offers a deeper understanding of how the concepts of high scrutiny and repeated examination evolved into the modern concept of "research".

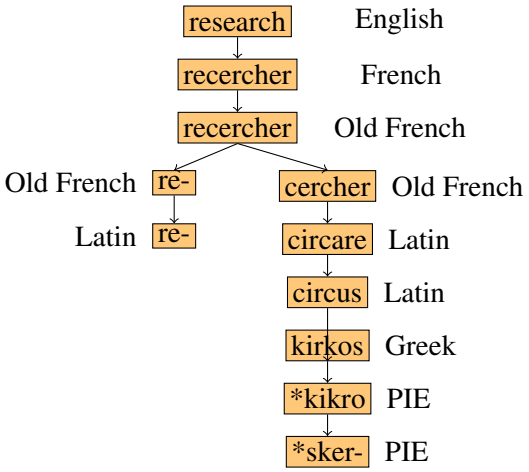

Figure 1: The etymology of the English word "research".

Traditional etymological studies have been limited to philosophical and comparative methods, relying heavily on linguistic expertise with a focus on specific languages and historical periods. This specialisation, while valuable, restricts the broader application etymology in large-scale comparative and computational linguistics. Most etymological data lives in scholarly articles, etymological dictionaries, or web resources. Such formats, although rich in detail, are inherently unstructured and not suited for computational approaches that require systematic data to process language on a large scale.

The field's reliance on verbose descriptions poses another challenge. These prose descriptions, typical of most etymological entries, make it difficult for computational tools to extract and analyze the relationships between words across languages and time periods. Consequently, the absence of structured, computation-suitable etymological databases has been a notable gap, leaving computational linguists without the resources necessary to quantitatively analyze historical linguistic

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 126–136

data effectively.

This paper introduces a new structured dataset specifically designed for computational etymology. We begin by crawling data from the Online Etymology Dictionary (or Etymonline)[1], an online dictionary compiled by historian Douglas Harper from various scholarly articles and books such as *The Dictionary of Etymology* (Barnhart and Steinmetz, 1988) and *A Comprehensive Etymological Dictionary of the English Language* (Ernest Klein, 1971). We manually annotated a subset of these entries to establish a high-quality baseline and ground-truth dataset. We then use the annotated dataset to fine-tune FLAN-T5 (Chung et al., 2022), an instruction fine-tuned encoder-decoder language model to extract etymological relations from the dictionary entries. This method allows us to explore Large Language Models (LLMs) and other automation techniques to systematically extract traditional prose-based etymological entries buried in scholar articles and dictionaries into a structured format. Ultimately, the structured dataset generated through this project will provide a valuable resource for computational linguists and other researchers, facilitating more large scale analysis of language evolution and enabling new insights into the interconnectedness of languages across time and space. Furthermore, this approach explores the feasibility of leveraging LLMs to curate structured data from traditional dictionary data.

The dataset comprises 63,603 entries crawled from Etymonline, with 5,361 entries manually annotated and 58,242 entries automatically annotated using the trained system. The final dataset includes 103,322 etymological relationships and 15,931 connected components, providing a comprehensive resource for examining language evolution and etymological connections. The dataset will be publicly accessible.

2 Linguistic Background

Diachronic change is an inherent aspect of linguistic evolution, driven by the need for effective communication within and between communities. Understanding the evolution of language requires examining various factors that contribute to linguistic change. Simplification of grammatical structures is a common trend in language evolution. For example, the transition from Old English and Modern English shows a significant reduction in verb conjugation complexity (Baugh and Cable, 1993). Technological advancement also impacts linguistic development. For example, the printing press contributed to the linguistic standardization and a more uniform spelling of the English language (Okrent and O'Neill, 2021).

2.1 Current Etymology Studies

Despite extensive research, many words still have unresolved origins.The Oxford English Dictionary, a prominent resource in this field, offers etymologies for over 600,000 words but lists a significant number as "origin unknown" or "of uncertain origin".

The absence of historical documentation is a significant barrier, especially for words from prehistoric times or non-literate cultures. For example, the etymology of the English word "dog" remains surprisingly unclear, as it appears in Middle English with no clear Old English predecessors (Gąsiorowski, 2006). Language contact adds another layer of complexity, particularly for borrowed words from extinct or significantly transformed languages. The semantic shift and phonetic changes over centuries obscure the word's origins.

Etymological research also faces methodological difficulties. Deciphering ancient languages requires specialized knowledge, and distinguishing borrowed words from native ones is challenging in linguistically diverse areas. Polysemy and homophony further complicate research, as words that sound similar may have different origins or meanings. For instance, "bank" can refer to a financial institution or a riverbank, each with separate etymological paths.

2.2 Computational Historical Linguistics

Computational etymology employs innovative approaches like automated etymology extraction, using natural language processing and machine learning to identify relevant relationships in large corpora. Cognate detection is an active area of research, with traditional methods measuring lexical similarity via string similarity (Ciobanu and Dinu, 2014; Gomes and Pereira Lopes, 2011; Simard et al., 1992). Recent trends involve machine learning to identify cross-lingual orthographic transformations (Bergsma and Kondrak, 2007; Mitkov et al., 2007), and neural networks are used to trace changes in word forms over time (Kanojia et al., 2019; Goswami et al., 2023; Bollmann, 2018).

The construction and analysis of linguistic

[1] https://www.etymonline.com/

databases are essential for large-scale computational linguistics. These databases store extensive data and support analytical queries revealing patterns in language evolution. CogNet (Batsuren et al., 2019) is a large-scale cognate database extracted based on WordNet. The Database of Cross-Linguistic Colexifications (CLICS^2) is a computer-friendly framework for analyzing cross-linguistic colexification patterns with the Cross-Linguistic Data Formats initiative (CLDF) (List et al., 2018).

3 Challenges in Modern Lexicography

Lexicography, the practice of compiling, writing, and editing dictionaries, has undergone significant changes in the digital age. Historically, dictionaries have served as authoritative references for language, offering definitions, etymologies, phonetic guides, and usage examples. However, traditional lexicographical methods are increasingly struggling to keep pace with computational approaches and the rapid evolution of language in the modern era.

Traditional lexicographic methods often lead to inconsistencies in dictionary data due to the subjective nature of language documentation and the variability in editorial practices. For instance, the noun "research" has four different definition in the Oxford English Dictionary (OED) but only three in Merriam-Webster. Another example is the etymology of "pumpkin". The word is traced to its French origin (*pompon* or *pompion*) in the OED, Merriam-Webster, and Etymonline, but each provides varying historical context. The OED links it to Classical Latin *pepōn-*, Merriam-Webster further identifies Greek *pépon*, and Etymonline traces it to the Proto-Indo-European root **pekw-*. These inconsistencies in the depth and nature of information reflect differing editorial standards and the lack of standardised lexicographical practices.

Digitizing traditional dictionaries presents another challenge due to their inherently non-structured, descriptive format, which is often incompatible with computational processing. Dictionary entries typically contain long, verbose paragraphs that are sometimes hard for humans to comprehend and even more challenging for computers to parse. This is further complicated by the lack of consistencies among dictionaries for data integration, such as differing abbreviations for parts of speech.

These challenges highlight the need for sophisticated computational approaches and collaboration between lexicographers and computational linguists. This project aims to explore using Information Extraction (IE) and NLP systems to automate structured data extraction from dictionaries, enabling systematic linguistic pattern analysis. Ultimately, the goal is to bridge the gap between traditional lexicography and modern computational linguistics, providing scalable and transferable solutions for mining valuable information.

4 Building an IE system for Lexicographical Mining

This paper presents the collection and annotation of etymological entries from Etymonline, using large language models to extend annotations. The aim is to convert unstructured dictionary data into a structured format suitable for linguistic analysis and computational processing.

4.1 Data Collection

For this study, Etymonline was selected as the primary source of data. The website aggregates etymological information from various scholarly sources, providing a detailed description of a word's origins, historical developments, and transformations within the English language. Below is an example entry of the word "research" whose structured etymology is given in Figure 1.

> **research** *(v.)*
> 1590s, "investigate or study (a matter) closely, search or examine with continued care," from French recercher, from Old French recercher "seek out, search closely," from re-, here perhaps an intensive prefix (see re-), + cercher "to seek for," from Latin circare "go about, wander, traverse," in Late Latin "to wander hither and thither," from circus "circle" (see circus).
>
> The intransitive meaning "make researches" is by 1781. Sometimes 17c. also "to seek (a woman) in love or marriage." Related: Researched; researching.

Etymonline is grounded in scholarly rigor, extensive coverage, and open accessibility. The dictionary is curated by Douglas Harper, who compiles information solely from scholarly sources, ensuring

high accuracy and reliability. Harper's consistent approach to entry composition allows for systematic extraction and analysis of etymological data. This uniformity is crucial for applying computational techniques that require standardization data inputs.

A total of 63,603 entries were extracted from Etymonline.

4.2 Data Preprocessing

Simple data preprocess was conducted. For words with the same spelling but have different parts of speech, the POS tag remains part of the lexical term to differentiate between them, as illustrated in the "research (v.)" example. Homographs, such as "bank", are differentiated by an index, such as *bank (n.1)* and *bank (n.2)*. The typsetting information given by the original Etymonline entry was ignored, while hyperlinks were preserved to build a complete etymological network.

Since all entries used similar expository language to describe the etymological relations, regex, a pattern matching tool, was used to extract a collection of candidate lexical terms. The candidate terms extracted for the verb "research" are shown below.

> recercher, recercher, re-, re-, cercher, circare, circus, circus, Researched, researching

4.3 Manual Annotation

A main challenge of this paper is transforming the prose paragraph style of the Etymonline dictionary entries into structured formats. 5,361 entries were selected for manual annotation. A key criterion for selection was diversity in word initials to prevent overrepresentation of any particular prefix. In Etymonline, suffixes are represented by unique word initials, so varied initials also account for word endings.

In etymological studies, prefixes and suffixes are critical for tracing word origins as they often contain significant linguistic markers of historical and morphological transformations. Ensuring varied word initials prevents bias toward specific affix patterns, which is crucial to prevent machine learning models from skewing their learning toward particular initials. This approach enhances the generalizability and accuracy of the models.

The manual annotation of the dataset was executed by a single linguistics student, tasked with converting the text into edge list format. The target format of the entry "research" is an edge list shown below. In this study, the annotation task was designed to be straightforward and did not necessitate specialized expertise. Therefore, we opted to employ a single annotator for the task. While inter-annotator agreement is important for tasks requiring trained experts to ensure reliability and consistency, we deemed it unnecessary for this simple annotation task. The clarity and simplicity of the task ensured that the single annotator could perform it with sufficient accuracy and consistency.

> **research** *(v.)*
> research_E, recercher_F
> recercher_F, recercher_OF
> recercher_OF, re-_OF, cercher_OF
> cercher_OF, circare_L
> circare_L, circus_L

The edge list format used in this project is designed to represent the descendency relationships between words. Each line in the list represents a direct etymological link from one form to another. For example, the line "recercher_OF, re-_OF, cercher_OF" indicates that the Old French word *recercher* derives directly from the Old French prefix *re-* and the Old French word *cercher*. Further more, the suffixes attached to each word after the underscore, such as "_E", specify the language of the word form in question. Table 1 reports some language and their abbreviations. The complete list of languages and their abbreviation, refereed to as language labels from here on, used in this dataset can be found in appendix A.

Language Label	Language
PIE	Proto-Indo-European
F	French
ONF	Old North French
AF	Anglo-French
MF	Middle French
OF	Old French
...	...

Table 1: Language Labels and Corresponding Languages

One observation of the edge list is that it is not yet complete compared to the graph given in Figure 1. More specifically, it is still missing the etymological relationships from the Latin word *circus* to the Proto-Indo-European root **sker-*. These missing

links are documented under the entry for the word "circus". Once the annotation process is completed for the entire dataset, these connections will be fully integrated, resulting in a complete representation of the word's etymological history.

4.4 Automatic Annotation

To fully annotate the entirety of the crawled entries from Etymonline, we employed the FLAN-T5-base model (Chung et al., 2022), a variant of the Transformer-based T5 model with 248 million parameters, which has been pre-trained on a diverse range of language understanding tasks. This section details the selection rationale, fine-tuning process, and the specific configurations used to adapt the model to the task of etymological annotation.

4.4.1 Model Selection

The open-source FLAN-T5-base model was chosen for its flexibility, strong performance in text generation tasks, and relatively small size compared to some state-of-the-art LLMs.The core architecture of FLAN-T5 is based on the Transformer model (Vaswani et al., 2017), utilizing self-attention mechanisms to process data sequences. These mechanisms, which compute the relevance of all other words in the sequence for each word in the input, are particularly beneficial as etymological relationships are buried within and across non-adjacent sentences. Unlike most current LLMs with a decoder-only structure, FLAN-T5 employs a dual structure with an encoder that processes input text and a decoder that generates output text. This setup is ideal for transforming verbose etymology descriptions into structured formats like edge lists. The encoder captures contextual relationships within the input, while the decoder uses this context to generate accurate, formatted output. Furthermore, FLAN-T5 has been fine-tuned to adapt to specific tasks with minimal task-specific data, crucial for high-quality annotation where such data is scarce. The model's robust pre-training enables it to generalize well across previously unseen tasks.

4.4.2 Fine-Tuning

The manually annotated subset of 5,361 entries from the initial data collection phase was split into a training dataset of 4,556 entries and a test dataset of 805. Each entry in the dataset was further processed and presented as a prompt to the model. As example of the input prompt to the model is given below.

> ###INSTRUCTION:extracting etymological relations from text and structuring this information into an edge adjacency list.
> ###WORD: research (v.)
> ###TEXT: 1590s, from Middle French recercher, from Old French recercher ""seek out, search closely,"" from re-, intensive prefix (see re-), + cercher ""to seek for,"" from Latin circare ""go about, wander, traverse,"" in Late Latin ""to wander hither and thither,"" from circus ""circle"" (see circus). Related: Researched; researching.
> ###CAND: recercher, recercher, re-, re-, cercher, circare, circus, circus, Researched, researching"

The target output is the edge list shown in section 4.3. The list is further processed into a string format as the model can only output sequence data. Each node in the edge is separated by a comma; each edge is encapsulated in parenthesis, and edges are separated by a semicolon. An example target output for the word "research" is shown below.

> (research_E, recercher_F);(recercher_F, recercher_OF);(recercher_OF, re-_OF, cercher_OF);(cercher_OF, circare_L);(circare_L, circus_L)

The model was trained for 2 epochs with a learning rate of 5.6×10^{-4} and a weight decay of 0.01. The fine-tuning of FLAN-T5-base was performed on three NVIDIA A100 Tensor Core GPU to facilitate computation.

4.5 Evaluation

Two types of evaluations were performed, string-based and edge-based metrics. String-based evaluation metrics are the current standard in Natural Language Generation (NLG) tasks, and what LLMs used in this project was originally evaluated on. Though they are useful for accessing the presence of important etymological elements by comparing the target and generated texts, is not sufficient on its own for tasks like etymological relationship annotation where structural accuracy is important. For this task, where the correct representation of relationships between words is essential, string-based metrics may overlook errors in the logical or hierarchical arrangement of data. Therefore, an edge-based assessment focusing on the structural

and relational accuracy of the outputs is needed for a comprehensive evaluation approach.

4.5.1 String-based Evaluation

The string-based evaluation focuses on measuring the textual similarity between the model-generated output and the target (manually annotated) output. This involves the use of several well-established metrics in NLG tasks. Bilingual Evaluation Understudy (BLEU; Papineni et al., 2002) calculates the precision at word or phrase level between the model's output and the reference text. Recall-Oriented Understudy for Gisting Evaluation (ROUGE; Lin, 2004) emphasizes recall, ensuring all necessary etymological components are included. ChrF (Popović, 2015) evaluates the similarity at the character level, making it useful in this scenario where morphological differences between languages are significant. The results are reported in table 2. We observe a consistent and high accuracy among all three metrics.

String-based Evaluation	
BLEU	0.902
Rouge	0.920
ChrF	0.929

Table 2: String-based evaluation results

4.5.2 Edge-based Evaluation

The edge-based evaluation assesses the structural and relational accuracy of the outputs.

Edge-based Evaluation	
Edge Recall	0.905
Language Label Detection	0.990
Language Label Accuracy	0.909
Word Root Accuracy	0.905
Word Root Levenshtein Distance	0.321

Table 3: Edge-based evaluation metrics. Edge recall is the proportion of the etymological relationships (edges) in the data that the model identified, accurately or not. Language label detection reports the proportion of word roots that received a language label, accurately or not, while language label accuracy reports the proportion of word roots with the correct language label. Word root accuracy reports the proportion of the word roots correctly extracted and word root Levenshtein distance reports the average edit distance of predicted word roots from the actual roots.

Figure 2: Performance on BLEU, ROUGE, ChrF, Root Accuracy, and Language Label Accuracy over different training data size.

Table 3 reports several relevant metrics. A relatively high edge recall indicates that the model is proficient at identifying the presence of etymological relationships. A high language label detection rate at 0.990 but a comparably lower language label accuracy at 0.909 means that the model is generally reliable in applying language labels to word roots, it struggles to extract and interpret the correct source of the words.

A relatively high word root accuracy shows the model's effectiveness in identifying and extracting the foundational elements of the words, though further improvement is needed. Lastly, an average Levenshtein distance of 0.370 indicates the wrongly identified words still remain similar to the actual words.

4.5.3 Effects of Training Data Size on Model Performance

One of the motivation of this project is to investigate the feasibility of leveraging LLMs to extract structured data from dictionaries. In this section, we wish to explore the effects of training data size on model performance, given most dictionary data has little to no structured annotation. The FLAN-T5-base model was trained with different subsets of the training corpus, including sizes of 1000, 2000, 3000, 4000, and the entire corpus of 4556. The hyperparameters were kept exactly the same as described in section 4.4.2. The results are reported in Figure 2.

As expected, performance generally improves with increasing training data size for all metrics, although the magnitude of improvement varies.

ROUGE, and ChrF scores both show a plateau effect, where performance gains diminish after reaching approximately 3000 training examples. Root accuracy also shows a similar trend, suggesting that the models no longer learns to extract the root words with more training samples. This might be attributed to the fact that these metrics are string-based, similar to what LLMs were pre-trained on, where they excel even with relatively small samples for fine-tuning. Hence, even limited data is sufficient to achieve strong results in these metrics.

Label Accuracy, measuring correctly predicted language labels, and BLEU show a significant surge from 4000 to 4556 samples, indicating that a larger dataset benefits these metrics. The sudden performance jump may reflect a threshold effect, where the additional 556 samples provide sufficient data to predict specific label patterns accurately. It remains unclear why this threshold occurs between 4000 and 4556 training samples. The BLEU score jump is likely due to improved Label Accuracy.

Edit Distance shows substantial improvement from 1000 to 2000 training samples, even with high root accuracy rates. This suggests that while root accuracy was high with 1000 samples, the model made significant mistakes on incorrect predictions. The extra 1000 samples helped the model better predict more challenging words.

Overall, these results emphasize the importance of training data size, particularly for non-string-based metrics like Label Accuracy. The plateau effect in ROUGE, ChrF, root accuracy, and Edit Distance suggests that LLMs can effectively structure lexicographical data with limited manual annotation. However, additional data significantly benefits more complex tasks like Language Label predictions.

5 Resulting Resource and Analysis

Out of the 63,603 entries crawled from Etymonline, 5,361 were manually annotated to fine-tune the system and the remaining 58,242 were automatically annotated with the trained system.

Using a regex-based method to pattern match the result, we found that the exact match rate for our model's output was 0.913, indicating that 53,148 out of the 58,242 output adhered to the expected format. Though the formatting of the dataset is relatively easy, LLMs, such as the one used in this study, often struggle with generating outputs that adhere to specialized formatting requirements, as they are predominantly trained to produce fluent, natural language text rather than structured or formatted data.

Upon further analysis, we discovered that a significant proportion of the mismatches were due to the absence of a language label for each node. This suggests that while the model was often successful in identifying etymological relationships (e.g., detecting the correct word roots and their connections), it frequently failed to append the appropriate language labels to these roots. To address this issue and better understand the model's capabilities in detecting relationships without the confounding factor of label generation, we modified our evaluation approach. We expanded the regular expression used in our assessment to no longer require a language label for each node, focusing instead solely on the detection of correct relationships between the word roots. This adjustment aimed to isolate the model's performance in understanding and reconstructing the etymological connections from the additional task of accurate language classification. After removing the language label constraint, 57,214 entries had the correct format, about 98.2% of the total entries, a significant improvement.

In total, 103,322 relationships were found, with 15931 connected components. The top 5 most connected lexical terms are given below in Table 4, all of which are affixes. This result is not surprising as affixes are one of the most productive morphological units in English.

Lexical Term	#Connections
'un-'	656
'-y'	552
'-ly'	388
'-al'	360
'-ism'	346

Table 4: Top 5 most connected lexical terms. The terms are all English, eliminated language labels for brevity.

More interestingly, Table 5 reports the top 3 most connected Proto-Indo-European roots. It is important to point out that the concept of most connected does not necessarily mean there are the most English words derived from it. It simply means the PIE root had evolved into the most distinct terms which then evolved into English terms.

We also analyzed the immediate word origins of the English words. Immediate word origins refer to the most recent source language from which

Term	Meaning	#Connections
'*kwo-'	stem of relative and interrogative pronouns	12
'*gno-'	to know	12
'*gene-'	give birth, beget	11

Table 5: Top 3 most connected PIE roots.

modern English words were borrowed or derived. We can better understand the linguistic influences that have shaped contemporary English Vocabulary. The top five immediate word origins, other than English itself, are

1. Latin
2. Old French
3. French
4. Old English
5. German

6 Related Work

6.1 Computational Resources in Lexicography

The Oxford English Dictionary was one of the earliest digital lexicographical projects, bringing traditional practices into the digital era by offering searchable and downloadable lexical data (Simpson and Weiner, 1989). Merriam-Webster's online dictionary similarly provides API access for integrating curated lexical information with computational systems. WordNet (Miller, 1995), a landmark in lexical resource development, is organized as a network of synonym sets (synsets) and provides rich semantic relationships between words. It has inspired projects like BabelNet (Navigli and Ponzetto, 2012), which integrates WordNet with multilingual resources. Wiktionary, a collaborative and open-source dictionary project, has grown into a significant resource for structural lexical information. However, due to its collaborative nature, quality and consistency issues arise, necessitating data refinement and filtering for computational applications (Meyer and Gurevych, 2012).

6.2 Computational Resources in Etymology

Etymological WordNet (de Melo, 2014) was one of the first significant attempts to create a structured multilingual etymological database. It aggregates etymology sections from Wiktionary and organizes them into a machine-readable network. Etymological WordNet contains over 500,000 lexical items from various languages and more than 2 million links, offering the first structured multilingual view of word origins and relationships across languages. Despite the significant contributions of Etymological WordNet, it relies entirely on data extracted from Wikitionary, which suffers from inconsistencies due to its collaborative natures and lacks granularity in tracking etymological relationships. Some entries on Wikitionary presents folk etymologies, such as the word "pumpkin". Out of the 2 million links within the network, a major portion of those emphasize cross-lingual cognates and derivational links, rather than genuine etymological relationships. Futhermore, de Melo (2014) uses custom pattern matching techniques to mine data, making it only applicable for Wikitionary, and thus not transferable to other dictionaries one wishes to structure.

7 Conclusion

In this paper, we presented a comprehensive methodology for building an information extraction system that transforms the unstructured textual data of the Online Etymology Dictionary (Etymonline) into a structured, computation-friendly format. Our system achieved 94.4% accuracy in correctly identifying relationships between word roots, demonstrating the feasibility and potential of leveraging large language models for structured data extraction from unstructured lexicographical sources.

Future work will focus on exploring is the transferability of the current model on different dictionary data, potentially eliminating the need for time-intensive manual annotation of other similar datasets.

Acknowledgments

We want to thank Li Liang for the data collection and annotation, and Junjie Cao for the initial implementation of a ranking system. We would like to also thank the anonymous reviewers for the insightful and valuable suggestions. This work is partly supported by Cambridge University Press & Assessment.

References

Robert K. Barnhart and Sol Steinmetz. 1988. *The Barnhart Dictionary of Etymology*. H.W. Wilson Company.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A Large-Scale Cognate Database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Albert C. Baugh and Thomas Cable. 1993. *A History of the English Language*. Taylor & Francis.

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-Based Discriminative String Similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic. Association for Computational Linguistics.

Marcel Bollmann. 2018. Normalization of historical texts with neural network models.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *Preprint*, arXiv:2210.11416.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105, Baltimore, Maryland. Association for Computational Linguistics.

Gerard de Melo. 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. European Language Resources Association (ELRA).

Ernest Klein. 1971. *A Comprehensive Etymological Dictionary Of The English Language By Ernest Klein*.

Piotr Gąsiorowski. 2006. The etymology of Old English *docga. 111:275–284.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In *Progress in Artificial Intelligence*, pages 624–633, Berlin, Heidelberg. Springer.

Koustava Goswami, Priya Rani, Theodorus Fransen, and John McCrae. 2023. Weakly-supervised Deep Cognate Detection Framework for Low-Resourced Languages Using Morphological Knowledge of Closely-Related Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 531–541, Singapore. Association for Computational Linguistics.

Diptesh Kanojia, Kevin Patel, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholemreza Haffari. 2019. Utilizing Wordnets for Cognate Detection among Indian Languages. In *Proceedings of the 10th Global Wordnet Conference*, pages 404–412, Wroclaw, Poland. Global Wordnet Association.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Johann-Mattis List, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. CLICS2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In *Electronic Lexicography*, pages 259–292. Oxford University Press.

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

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Arika Okrent and Sean O'Neill. 2021. Blame the Printing Press. In Arika Okrent and Sean O'Neill, editors, *Highly Irregular: Why Tough, Through, and Dough Don't Rhyme—And Other Oddities of the English Language*, page 0. Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montréal, Canada.

J. A. Simpson and E. S. C. Weiner. 1989. *The Oxford English Dictionary*. Clarendon Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Language Labels

Below is the complete list of languages and their labels (abbreviations) used to annotated the dataset.

Language Label	Language
PIE	Proto-Indo-European
F	French
ONF	Old North French
AF	Angolo-French
MF	Middle French
OF	Old French
L	Latin
MediL	Medieval Latin
ModL	Modern Latin
LateL	Late Latin
VL	Vulgar Latin
OE	Old English
PGer	Proto-Germanic
H	Hebrew
Avest	Avestan
IndoIr	Indo-Iranian
San	Sanskrit
G	Greek
GE	Greenland Eskimo
I	Italian
A	Arabic
Sy	Syriac
Per	Persian
Ira	Iranian
Por	portuguese
OHGer	Old High German
Adut	Afrikaans Dutch
Ger	German
AL	Anglo-Latin
Cel	Celtic
Tur	Turkish
ModG	Modern Greek
EG	Ecclesiastical Greek
OL	Old Latin
PI	Proto-Italic
Nor	Norse
ONor	Old Norse
Dan	Danish
FCan	French-Canadian
Fran	Frankish
Gae	Gaelic
Scot	Scottish
Hin	Hindi
Yid	Yiddish
Rus	Russian
Algo	Algonquian
preL	Pre-Latin
Serb	Serbian
Aben	Abenaki

ORus	Old Russian
OPro	Old Provencal
LGer	Low German
WGer	West Germanic
Ir	Irish
Nah	Nahuatl (Aztecan)
Mal	Malay
Ch	Chinese
Scan	Scandinavian
Wel	Welsh
Sem	Semitic
Norw	Norwegian
Swe	Swedish
Sla	Slavonic
Jap	Japanese
Ber	Berrichon
Afr	Africa
SerCro	Serbo-Croatian
Aram	Aramaic
Gas	Gascon
Egy	Egyptian
Tup	Tupi
Jav	Javanese
Ben	Bengali
Fin	Finnish
Kut	Kutchin
Guugu	Yimidhirr
Sio	Siouan
Nepa	Nepalese
Dra	Dravidian language
Pol	Polish
OFri	Old Frisian
Canto	Cantonese
Esto	Estonian
Lith	Lithuanian
GaRo	Gallo-Roman
CuSpan	Cuban Spanish
Araw	Arawakan
NEAL	Southern New England Algonquian
Nar	Narragansett
Flem	Flemish
Aztec	Aztec
ByG	Byzantine Greek
Que	Quechua
Afrika	Afrikaans
Ojib	Ojibwa
Hun	Hungarian
Lush	Lushootseed
Dako	Dakota
Cro	Croatian
EL	Extinct language