LDK 2023

The Cardamom Workbench for Historical and Under-Resourced Languages

Adrian Doyle and **Theodorus Fransen** and **Bernardo Stearns** and **John P. McCrae** and **Oksana Dereza** and **Priya Rani**
Data Science Institute
University of Galway
Galway, Ireland

Abstract

This paper describes the creation of a workbench tool designed to make technologies developed throughout the lifespan of the Cardamom project easily accessible to researchers who could most benefit from them, but who may not have the technical expertise to apply bleeding edge technologies to their own datasets. The workbench provides an intuitive graphical user interface (GUI) and workflow which abstract users away from underlying technical tasks, while providing them with a suite of powerful NLP tools developed by the Cardamom team. These include tokenisers, POS-taggers, various annotation tools, and ML models. The performance of workbench tools can be improved as text and annotations are added by users. It is envisioned that this workbench will provide a simple route to digital publication for academics in the humanities, or more specifically, for linguists working with under-resourced or historical languages, who have collected text data but are unable to make it available online as a result of financial or technical restraints. This has the added benefit of increasing the availability of high quality, annotated text data to NLP researchers, thereby providing value to both communities of researchers.

1 Introduction

Some of the most cutting edge Machine Learning (ML) and Natural Language Processing (NLP) techniques require large quantities of data for use in training and testing increasingly complex models (Brown et al., 2020; Shoeybi et al., 2019; Patil et al., 2022). A relative abundance of digital text data is readily available for some of the most widely used world languages, however, it is well established that many of the world's languages are severely under-resourced in terms of technologies to support language use (Bender, 2019; Joshi et al., 2020; Hedderich et al., 2021). As more complex resources, like machine translation tools, are built upon the foundation of rudimentary resources, like parallel corpora, a vicious cycle can emerge whereby under-resourced languages remain under-resourced, while resources for better resourced languages multiply.

Many of the most severely under-resourced languages can lack even a sufficiently large corpus of machine-readable text, never mind resources like tokenisers, part-of-speech (POS) taggers, and more advanced processing tools. NLP researchers are forced to either abandon the hope of developing ML models for such languages, or to devote time to creating basic resources like text corpora. For this reason Cieri et al. warn that, "If the language has too few resources, the project could mire in [language-resource] creation" (2016, 4548). At the same time, linguistic researchers often accumulate text which, for a variety of reasons, they may be unable to make easily accessible to other researches. Quantities of text, which may not be substantial enough to justify a print edition, are regularly produced during the course of research projects, and it can be difficult for researchers to make these texts available online if they do not have access to the required technical skills, funding or IT resources. As such, texts are often abandoned once research projects conclude. In the case of under-resourced languages, such texts could be particularly valuable in the creation of NLP tools like spell-checkers and machine translation resources. They could be harnessed to improve research prospects for humanities scholars working with languages for which little technology is readily available.

The aim of this paper is to present a workbench tool designed to provide linguistic researchers with easy access to NLP tools developed by Cardamom researchers, and to reduce the barrier to entry for digital publication of their texts. As such, these tools include preprocessing tools like tokenisers and POS-taggers, annotation tools so that a wide variety of metadata can be stored, as well more complex tools such as word-embedding models

which improve search and query options for corpora. Section 2 of this paper will discuss the value and availability of digital text resources. Section 3 will give an overview of the Cardamom project. The state of digital text availability for historical languages will be discussed as a case study in section 4. It will be demonstrated that there exist certain obstacles to the production of freely available digital text which could be harnessed to improve ML resources. Section 5 will describe the workbench itself, and how it aims to overcome these obstacles.

2 Resources and Research Communities

It is self-evident that linguistic researchers, whether their focus be on language processing or traditional linguistics, stand to benefit from freely available and easily accessible digital text corpora. Such corpora can be used as teaching aids for language students, and many traditional avenues of linguistic research can be improved or supported by the availability of a machine readable corpus of text (Lynn, 2012). For NLP researchers, ever larger quantities of digital text are becoming more important as computer processing power improves and state-of-the-art techniques become more reliant on large quantities of training data. For example, Villegas et al. report that "CLARIN NLP services prove efficient when processing large corpora but large corpora are not always available" (2012, 3287). Where text data is available to NLP researchers, they in turn can develop tools to support or enhance traditional linguistic research areas. Areas of study such as linguistic typology and syntax greatly benefit from corpus-based and data-driven research (Nivre, 2015; Alves et al., 2023). Tools for machine translation, as well as machine-readable lexicons, for example, can greatly reduce the time-investment required for otherwise laborious tasks, allowing scholars more time to focus on research questions. These same tools' performance can be improved further as larger quantities of text data become available.

Despite the clear benefits to both research communities, NLP and traditional linguistics, close cooperation between the two is not necessarily easy to coordinate. As will be demonstrated in subsection 4, it is often difficult for humanities-based researchers to ensure text data they may have accumulated can be made available and remain easily accessible. In some instances, it will be shown, it may even be beneficial to researchers to avoid creating digital text corpora. On the other side of the house, NLP researchers are often content to demonstrate improved results over state-of-the-art techniques in some task or research area, however, it is not always prioritised that these improved techniques are easily accessible to those who stand to benefit from them. McGillivray et al. "draw attention to the lack of communication between the communities of NLP and DH" and further suggest that "In spite of its damaging effect on the progress of the disciplines, we believe this lack of communication and miscommunication are underestimated" (2020). It is almost meaningless from the perspective of a language community to demonstrate even significant improvements in an NLP area, like machine translation for example, if members of that community must become proficient in one or more programming languages, as well as command line interface, before they can benefit from it. This is not to mention the types of troubleshooting and version control issues which can often cause headaches even for highly technically proficient NLP researchers. The workbench which is the focus of this paper aims to empower researchers to work more closely together and ultimately provide beneficial resources to both camps.

3 Cardamom Project

The Cardamom project (McCrae and Fransen, 2019) got underway in 2019 with the aim of developing deep-learning-based NLP techniques to close the resource gap for historical and otherwise under-resourced languages. Throughout the project's lifespan Cardamom technologies have been applied in a variety of areas ranging from text preprocessing tasks like tokenisation (Doyle et al., 2019) to sentiment analysis (Chakravarthi et al., 2020) and detection of language and dialect (Goswami et al., 2020; Rani et al., 2022). Cardamom research has focused on reducing resource requirements, both for data and for processing power, with the aim of reducing the NLP barrier to entry for under-resourced languages. This has been accomplished by developing more efficient approaches to common tasks (Goswami et al., 2021a,b) as well as by exploiting commonalities between closely related languages to improve NLP prospects for individual low-resource languages (McCrae et al., 2021).

In aiming to improve language processing prospects for both under-resourced modern lan-

guages and historical ones, Cardamom is unlike many other projects. Because historical language stages can form diachronic links between modern languages, the benefits of transfer learning can be exploited not only laterally, from one modern language to another, but temporally forward and backward also, adding new dimensionality to such NLP solutions (Dereza et al., 2023b). Inclusion of historical language stages as a means of bridging divides between modern languages which have descended from them is a somewhat novel solution, and promises to bolster further research areas such as computer-assisted diachronic terminology mapping.

As historical languages are typically very under-resourced themselves, they too stand to gain from research which aims to reduce resource requirements for NLP. Moreover, historical languages can present challenges which are not common in modern languages. One such example is that many features of manuscript orthography are unsupported by modern standards like Unicode which "gives higher priority to ensuring utility for the future than to preserving past antiquities" (Becker, 1988, 5) and therefore, "aims in the first instance at the characters published in modern text". Therefore, many such features cannot be accurately or consistently captured in digital text without employing workarounds like discreet annotations (Doyle et al., 2018, 69–70). Another example relates to orthographies which predate the standardisation typical of modern languages. These can result in a high degree of spelling variation in historical language texts, which can be particularly problematic when processing languages which are morphologically rich (Dereza et al., 2023a). Moreover, in languages which predate modern word separation using spacing, even fundamental tasks like tokenisation can pose significant difficulties (Doyle et al., 2019).

Issues such as these have been been the subjects of investigation during the course of the Cardamom project. Problem areas specific to historical languages, which have to date received little attention, have been addressed and technologies have been developed to meet the specific needs of these and other under-resourced languages (see subsection 5.2). The focus of the Cardamom project has now shifted to ensuring these technologies are easily accessible to users who may find value in them.

4 Historical Languages; a Case Study

Historical languages like Old Irish and Old English suffer from many of the same resource deficits which afflict modern under-resourced languages. As no communities of native speakers exist for these languages, no new text can be generated by native speakers. Instead, NLP researchers must rely primarily on text which has survived for centuries or even millennia, from the times when these languages were still in use. Such texts are generally preserved in manuscripts, or in some cases, engravings in stone, clay and other materials. By the very nature of their antiquity, such sources of text can be scarce. Even where a text has survived, however, a digital transcription of it may not be available to NLP researchers.

Typically, historical linguists who transcribe the contents of a manuscript will aim to release the resulting text as a print edition rather than in digital format. There are many valid reasons for this, chief amongst which may be the perception that it is more advantageous to produce texts in print. Stifter et al. stress the importance of "ensuring that scholars receive due credit for their work for the purposes of career progression" (2021, 17), and it stands to reason that scholars will aim to produce whichever form of publication is more likely to receive engagement in the form of peer reviews and citations. However, Stifter et al. also identify "a reluctance to rely on and cite digital resources" (2021, 10) among linguists working with historical Gaelic varieties, "particularly when there is a print alternative, even if more out of date". This reluctance appears to be rooted in the belief that such resources are somewhat unreliable or capricious, and Stifter et al. report that "the perceived authority and trustworthiness of digital resources" (2021, 17) was a recurring theme in their workshop. Scholars do not feel confident citing a resource which they believe could be altered at any time, with little warning or oversight. Unfortunately, for as long as there is a reluctance to interact with digital resources by humanities scholars, linguists will be actively incentivised to generate print editions at the expense of digital text resources. This, in turn, contributes to a shortage of digital text available to NLP researchers for historical languages.

Other technical factors also play a role in preventing the generation of digital text for historical languages. It is no secret that "Digital resources are expensive both to build and maintain" (Stifter

Figure 1: Cardamom Workbench: Home Page with Uploaded Texts and File Upload Options.

et al., 2021, 10). They require ongoing investment and technical support, while a print edition, once published, is relatively permanent. Publishing text online requires either developing the technical skill-set required to create a web-based text repository, or employing a web developer. Either option incurs costs, be it for hardware acquisition and maintenance, or for ongoing web-hosting services. Linguists can be easily excused for preferring to simply focus on their own specific research interests. Thus, both technical and financial restrictions contribute further to historical language varieties remaining particularly poorly resourced.

Despite the factors listed above which may obstruct linguistic communities attempting to make digital text available online, there is a clear desire to do so, and pride is rightly taken in extant digital resources. Stifter et al. note that "Medieval Irish studies have been at the vanguard of textual digitisation since the infancy of the World Wide Web" (2021, 14), and it is indeed widely reported that the first website hosted in Ireland was the *Corpus of Electronic Texts* (CELT, Ó Corráin et al., 1997; English, 2018; Burke, 2018; Ahlstrom, 2014). Other repositories like ISOS and projects like *Ogham in 3D* (White, 2012) are praised for making historical writings available to researchers and disseminating academic research to a wide public audience (2021, 7, 24–25). The value of creating digital resources is clearly not lost on humanities scholars, and it would benefit both communities of researchers, NLP and traditional linguistic, to develop a streamlined, cost-free means of publishing digital text online, whereby appropriate credit can be given to the creator of that text.

5 The Workbench

The Cardamom Workbench aims to overcome many of the problems discussed above, both those faced by NLP researchers and by those in humanities fields. It also aims to make useful NLP techniques and processes easily accessible to users. Users will be provided with an intuitive GUI through which they can interact with various Cardamom technologies, and the pipeline to digitally publishing texts online will be streamlined. If a user chooses to publish their text through the workbench, it will remain easily accessible online and will be appropriately attributed to the digital text's creator. It will also be ensured that the copyright of any earlier edition of an uploaded text is respected, and that contributed works meet quantifiable quality standards before they can be published, which should alleviate concerns about the reliability of these digital resources.

5.1 Application Design and Workflow

The application is comprised of a web-based front end and a relational database back end. The GUI has been designed to produce an intuitive workflow, intended to make the built-in Cardamom technologies easily accessible to a wide variety of users without requiring them to develop the kind of technical skill-set which would otherwise be needed. Users who make accounts can upload text files in common formats like `.pdf`, `.txt` and `.docx` at the homepage (see figure 1). The text is extracted from these files by the workbench, and stored in the database using UTF-8 encoding. Alternatively, users can create a new text from scratch using the built-in text editor. In either case, users will be asked to select the primary language of the text at the point of upload or creation. Texts can contain

multiple languages, however, some downstream tasks are language-dependent and require that a primary language is identified.

Once uploaded or created, users can select a text from the homepage. Doing so opens it in the Text Editor tab. Here changes can be made to the content of the text if necessary. Several other tabs are also available to users, each associated with a specific text processing or annotation task. These tabs, from left to right, form a workflow which is intended to guide users who may be unfamiliar with text processing though the successive steps in an intuitive manner. Certain steps are reliant on previous ones, and so some tabs will be unavailable until previous steps have been completed. For example, POS-tagging will be unavailable until a text has been tokenised. Users are not required to utilise every tab, nor to perform every type of processing which is available. For example, a user may intend only to tokenise a text, and it will be possible for them to export their token data once they have completed this step.

In each of the workflow tabs users will be able to carry out the specified task either automatically, using Cardamom technologies, or manually. This gives users manual oversight over automated tasks. For example, in the POS Tagging tab a user can manually select POS tags for individual words, or they can click the Auto-Tag button and the workbench will select the appropriate pre-trained POS-tagger model for the specified language, and use it to tag the text. The user may use the Auto-Tag function first, then manually change tags by clicking on a token, and selecting a different POS from a drop-down menu (see figure 4 below). Where a user has manually annotated text in any workflow tab, and then applies automatic annotation to the text, the automatic tool will not overwrite manual annotations. In languages which are currently unsupported by Cardamom technologies, the workbench provides generalised automation tools to support workflows where possible; for example, the workbench can attempt to tokenise text regardless of language, though results are improved where a supported language is specified. Users may have to carry out language-dependent tasks manually, however, where languages are unsupported by the workbench.

Tokenisation does not involve splitting a user's text into word-level strings and storing these. Instead, when tokenisation is carried out by a user on a text, a start index and end index are stored in the database for each token. Tokens can then be retrieved from the original text at any point using these indices. Token-dependent annotations, such as POS-tags, are applied to this index range rather than to the string itself. In a similar manner, any user-specific annotations are also applied to an index range corresponding to a string of text highlighted by the user in the GUI. This allows annotations to be provided both at token level, as well as at sub-token and super-token levels. When the user makes changes to the base text in the text editor, the indices of tokens are updated in accordance with any alterations made, ensuring that annotations remain aligned with the correct text.

One of the main benefits of the workbench's design is that it can learn from users' content. Users, therefore, can improve the ability of the workbench to automate processing tasks for their language each time they upload or annotate text, as this provides more training data to the underlying language models. This adaptability is of great value for under-resourced languages, for which little annotated text data might yet exist. In the case of languages which are not yet supported by the workbench, users will need to manually annotate some portion of their uploaded text data themselves in the workbench. Once a sufficient quantity of text has been manually annotated, however, it will be possible to train models for the language, making automatic annotation available for that language. In order to ensure consistency of data used for model training, the streamlined annotation process requires that users tokenise and POS-tag in accordance with UD guidelines (Zeman, 2016). User-generated data will not be used as training data until it meets these criteria. While user-generated annotations may be used in resulting publications, they do not form a part of the main workflow, and will not be used in model training.

5.2 Technologies

The technologies which underlie the automatic processing and annotation options in the workbench have been developed throughout the course of the Cardamom project. As these technologies are not the focus of the current paper, technical aspects of their individual implementations cannot be discussed in detail throughout this section. Specifications of many technologies used by Cardamom have already been published (Doyle et al., 2019;

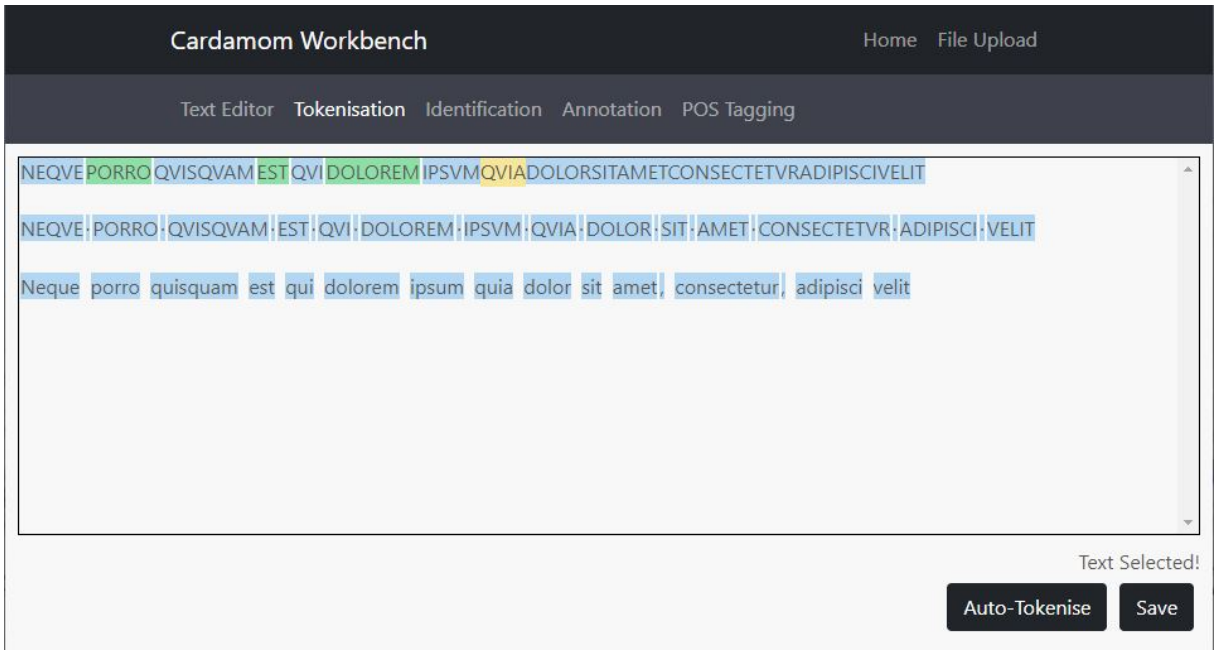

Figure 2: Cardamom Workbench, Latin Text: Tokenisation Tab with Automatically Generated Tokens (Blue), Manually Generated Tokens (Green) and Selected Text (Yellow).

Chakravarthi et al., 2020; Goswami et al., 2020; Rani et al., 2022; Goswami et al., 2021a,b; McCrae et al., 2021; Dereza et al., 2023b), and publications for other technologies are in progress. Certain tasks, such as tokenisation, which have been found to create specific difficulties for languages which have been the focus of Cardamom research will be discussed in this section, however. This section will also address tasks have been improved by Cardamom research, either by reducing the quantity of training data required to achieve sufficient results, or by reducing the processing power and time required to achieve results comparable with the state-of-the-art.

5.2.1 Tokenisation

Tokenisation has been identified as problematic for languages which predate the modern standard separation of lexical words using spaces (Doyle et al., 2019). In such cases, tokenisation requires a more targeted, language-specific approach. For example, certain Latin texts are written with words separated using an interpunct, not spacing. An example of this can be seen in figure 2. By contrast to Latin, the interpunct is often used to indicate points of stress within the verbal complex in the orthography of Old Irish editions and learning material, but not necessarily at word boundaries. Latin text requires that tokens be separated at points where an interpunct is used, however, this may be inappropriate for Old Irish where the interpunct serves a different purpose. Therefore, it was necessary to create discrete tokenisers for Latin and Old Irish, each of which treat the interpunct as appropriate for the language in question.

Word spacing has also been identified as problematic when tokenising historical languages. Many Latin texts were written in *scriptio continua*, without any punctuation or spacing separating words from each other (see again figure 2). Meanwhile Thurneysen notes that generally, in Old Irish manuscripts, "words which are grouped round a single chief stress and have a close syntactic connexion with each other are written as one" (1946, 24). In either case, it is difficult to create an automatic tokeniser which can accurately separate such compounded words without large quantities of training data (Doyle et al., 2019). The workbench, therefore, allows users to manually identify the exact boundaries between tokens in their texts by highlighting some quantity of text which they consider to be a single token. By this means it is even possible for users to create tokens which contain space characters, as may be required, for example, where a nasal has been separated from the following word in Old Irish (see figure 3).

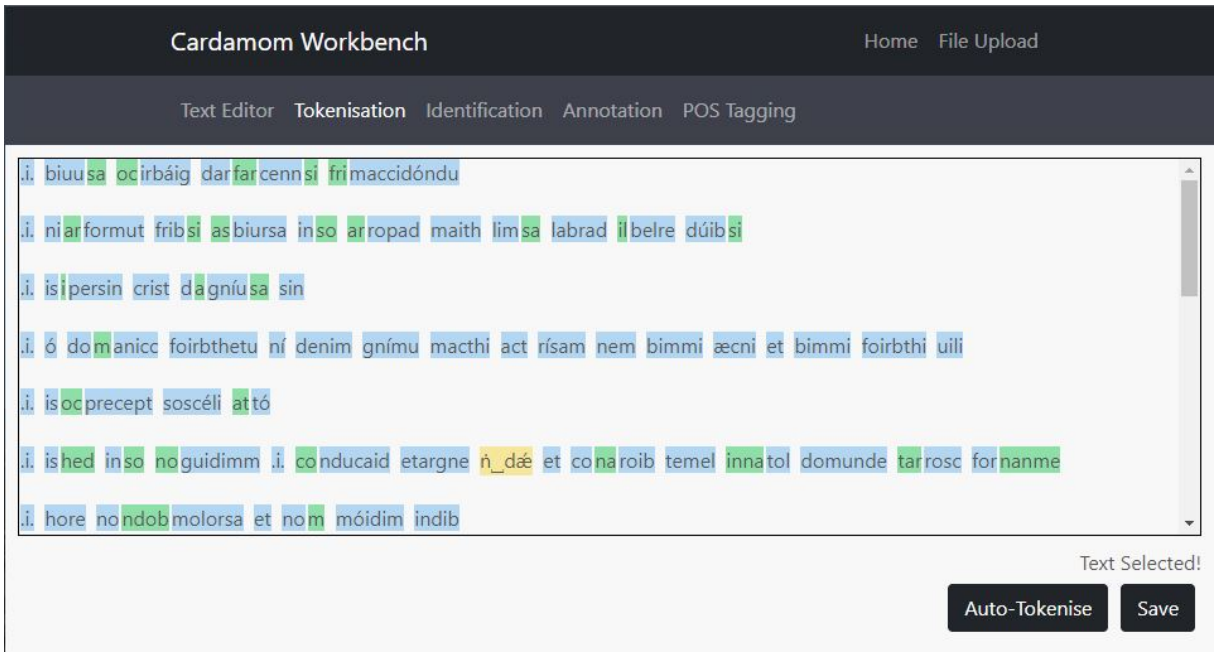

Figure 3: Cardamom Workbench, Old Irish Text: Space Character within the Selected Token (Yellow).

5.2.2 Language Identification, and Related Techniques

A considerable amount of Cardamom research has focused on the identification of various linguistic features and characteristics within a text. This includes, but is not limited to, identification of language and dialect (Goswami et al., 2020; Rani et al., 2022), authorship identification, and cognate detection. In the context of the workbench, these technologies may be of use to users working with texts which contain some degree of code switching. Identifying tokens which are not from the primary language of the text will allow for improved results in POS-tagging. These techniques may also be of interest to scholars of languages like Old Irish, for which "Contemporary divergences, such as would point to dialectal peculiarities, are very rare" (Thurneysen, 1946, 12).

5.2.3 POS-tagging

The Cardamom Workbench follows Universal Dependencies (UD) guidelines (Zeman, 2016) for tokenisation and POS-tagging. As such, the workbench utilises the same seventeen POS tags used in UD treebanks. This decision was made because UD has already established itself as a common standard, capable of facilitating the requirements of a wide range of languages. As such, it is reasonable to expect it will be suitable also for the various under-resourced and historical languages which are the target of the workbench. Moreover, adherence to such a well supported standard as UD, means that extant validation tools can be utilised to ensure the quality of data created and annotated by users.

As has been mentioned above, users can POS tag their text both automatically and manually. Automatic POS-taggers were trained for various languages using lexical data primarily drawn from UD treebanks. These models can be improved both when UD repositories are updated, and when workbench users POS tag their own text. Tagged text is colour-coded in the GUI to enable users to quickly and intuitively assess POS-tagged tokens (see figure 4). A future iteration of the workbench is expected to expand this token-level tagging to include headword identification to support digital lexicography, and lexical feature identification in accordance with UD guidelines.

5.2.4 Other Annotations

Various other forms of annotation are possible aside from language and POS tagging of tokens. The Annotations tab allows users to apply annotation not only to tokens, but at sub-token and meta-token levels also. Users can highlight any quantity of text and add an annotation to it. This is useful, for example, in digital editions of historical language texts where, in the manuscript, text may have been lost due to damage, or abbreviated using a variety of symbols (Thurneysen, 1946, 25). Users may wish to indicate that they have supplied or restored text in such instances, and can do so easily by providing

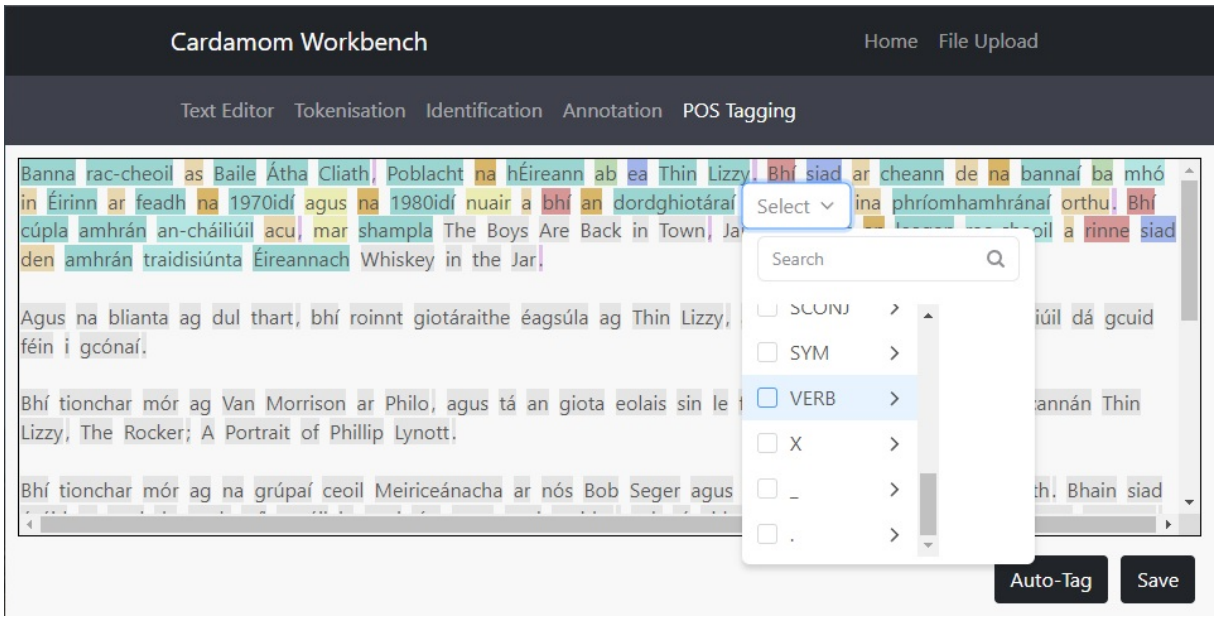

Figure 4: Cardamom Workbench, Modern Irish Text: POS-tagging with POS Tags Differentiated by Colour.

annotation in this manner. Here again, Cardamom technologies are available to help automate the process, for example, by suggesting the most likely annotation required based on the text selected by the user. In a future iteration of the workbench it is expected that users will have the option of exporting their text annotated with TEI markdown (TEI-Consortium, 1994), however, at launch the primary function of such annotations is to enhance resulting digital editions with metadata.

5.3 Value for Stakeholders and Future Work

The primary goals at launch are to ensure accessibility of current Cardamom technologies to users, and to provide a simple means of digitally publishing texts. Cardamom intends to provide free web hosting for users-submitted texts on servers owned and operated by the Insight Centre for Data Analytics, and permanent URLs will be provided for these once published. Once the period of funding has ceased for the Cardamom project itself, responsibility for continued support of the workbench, and hosting of both the application and digitally published texts, will be transferred to the Insight Centre for Data Analytics. This will ensure long-term accessibility of user-supplied content, which is beneficial both for users who will be appropriately credited with contributing the text, and for NLP researchers who will have access to more text data for under-resourced and historical languages. The quality of uploaded text and annotations can be tightly controlled using extant validation tools, and manual oversight.

As has been mentioned throughout this paper, updates to the workbench's functionality are expected as development continues after launch. Work is ongoing on a tool which utilises word embeddings to allow users to track orthographic and semantic changes in a lexeme over time, and to find words which are semantically or morphologically similar to an entered search term. It is envisioned that this functionality could be useful to historical linguists editing obscure manuscript passages, where one possible reading must be chosen over another. Generic tools such as concordancers are also intended to be implemented in future revisions, and extended functionality will be added for texts both as the workbench is developed, and in accordance with the level of annotation provided by users. For example, POS-tagging and headword annotation of tokens will enable linking to external lexical resources for a given language.

It is expected that once a sufficient interest has been demonstrated by users in the workbench, it will be possible to develop an expert peer-review and support network. This will further ensure the quality of submitted texts, allowing language experts to provide commentary and critique on a text before it is published. It will also be possible to credit reviewers when updates are made to published texts based on their recommendations. Such a network would also allow linguistic experts to advise on future development of the workbench to

support language-specific requirements, increasing its value to users going forward.

5.4 Related Tools

A number of extant tools may be compared to the workbench presented here, both as regards providing users with similar technologies, and simplifying interaction with annotated corpora. It is important to acknowledge these tools in order to appreciate the features and use cases which distinguish the Cardamom Workbench from them. The value proposition of the workbench, as well as its intended user base, are the primary distinguishing factors. As has been mentioned above, the intent of the workbench is to create value for two groups of researchers with distinct sets of requirements in order to improve their particular research prospects.

The historical focus of Cardamom research creates value in an area for which discrete solutions are required, and certain tools have already been made available in this area in an attempt to provide such solutions. *TEITOK* is an open source, web-based tools which enables users to create and distribute corpora (Janssen, 2016, 16). Users can align manuscript pages with transcribed text, and transcribe directly from manuscript images. Annotation is enabled using TEI, and users are given tools for visualising annotations such as dependency grammars and parse trees. As such, this tool is possibly the closest extant resource to the workbench in terms of its historical focus, and its corpus creation and annotation support. A few things set the two apart, however, the foremost of which is the technology stack provided by Cardamom. Workbench users benefit from these tools not as merely as static resources, but as dynamic ones. They can play a role in improving their performance by contributing more text and annotations. Thus, while the focus of *TEITOK* appears to be to facilitate corpus creation and annotation, the focus of the workbench is to provide users with tools which will empower them to process and annotate texts more efficiently, and to constantly improve the tools available to users.

Some extant resources provide users with technologies comparable to those of the Cardamom Workbench. The *IMS Open Corpus Workbench* (Evert, 2008) provides users with open source corpus query tools and is intended for use with large text corpora. On the one hand this is very useful for users who have access to large text corpora, though it is an unrealistic scenario for under-resourced or historical languages. The aim of Cardamom research has been to close the resource gap by creating tools which can be both trained and used on relatively small text corpora. On the other hand, according to the *IMS Open Corpus Workbench*'s website, "It is intentionally not very user friendly", requiring that users interact with it using secondary software which abstracts away from the technology stack. By contrast, the Cardamom workbench was designed from the beginning with user friendliness in mind, as its intended user base is specifically those who do not have the technical skill-set to use Cardamom technologies if it means downloading scripts from repositories like GitHub and running them using command line interface. *Persides* is an editing platform for Classics texts which allows large groups of users to partake in "allows for the participation of a large group of users in the process of editing, publishing, and analyzing ancient documents" (Almas and Beaulieu, 2013, 502). It is based on the principle that "a well-organized crowdsourcing effort can accomplish far more work than any lone scholar and the work ultimately produced benefits from the variety of perspectives included" (Almas and Beaulieu, 2016, 172). This contrasts with the work presented here in that the Cardamom workbench aims to empower individual scholars to annotate and publish their work with minimal effort or collaboration. Another web-based application, the *INCEpTION* annotation environment (Klie et al., 2018), provides users near free rein over how they annotate their corpora. While it provides predefined elements, like knowledge bases, layers and tag-sets, it also allows users to modify these, or to create their own annotations. While the Cardamom workbench allows users to provide their own annotations where desired, the streamlined annotation process is designed to ensure users' output meets a single common NLP standard as closely as possible for tasks like tokenisation and POS-tagging (Zeman, 2016). Moreover, the workbench provides users with a suite of NLP tools specifically designed to aid in such annotation for historical and under-resourced languages.

Possibly the most well known extant tool in this area is *Sketch Engine*, a web-based corpus management system which also provides users with text analysis functionalities. Some of the analysis tools provided by *Sketch Engine* overlap with those of the Cardamom Workbench, for example,

it allows POS-tagging for a wide range of supported languages. It also provides a "summary of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2014, 9), however, to support such features *Sketch Engine* requires that tools like a tokeniser, lemmatiser, POS-tagger, and morphological parser must already exist for a given language (2014, 18). Being a commercial tool, it is not free to use, however, a feature-limited free counterpart, *NoSketch Engine*, does exist. While *Sketch Engine* provides very valuable technologies to lexicographers, translators, language learners, and institutes like universities, its primary focus seems to be on making extant tools more accessible rather than developing or improving language tools. Here again the Cardamom Workbench provides value to users. Both *Sketch Engine* and the Cardamom Workbench cater more to some languages, for which more language resources are readily available, than to other less resourced languages. Cardamom, however, provides users with the possibility of creating such resources, and harnessing them to improve built-in language tools as they use the workbench. The suite of technologies built into the Cardamom Workbench is also more extensive than that of *Sketch Engine*, and these are targeted towards the kinds of processing and annotation tasks which will allow users to create the most useful language resources using their supplied text.

6 Conclusion

This paper has presented the Cardamom Workbench, a tool which provides language experts with modern NLP tools which can be easily applied to their own texts. It also aims to provide users with a streamlined means of digitally publishing text content which may be of value to both traditional linguists and to NLP researchers, meanwhile allowing appropriate credit to be given to users who produce and annotate the digital text.

Acknowledgements

This publication has emanated from research supported by the Irish Research Council under grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) and co-funded by Science Foundation Ireland (SFI) under grant SFI/12/RC/2289_P2 (Insight_2) and grant SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intelligence). We would like to acknowledge the hard work carried out by all Cardamom members, past and present, on the workbench presented here. We would also like to acknowledge the use of language data from Universal Dependencies treebanks in training and testing certain Cardamom models.

References

Dick Ahlstrom. 2014. How the Irish Helped Weave the Web. *The Irish Times*.

Bridget Almas and Marie-Claire Beaulieu. 2013. Developing a New Integrated Editing Platform for Source Documents in Classics. *Literary and Linguistic Computing*, 28(4):493–503.

Bridget Almas and Marie-Claire Beaulieu. 2016. *The Perseids Platform: Scholarship for All!*, pages 171–186. Ubiquity Press.

Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. Analysis of Corpus-based Word-Order Typological Methods. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 36–46, Washington, D.C. Association for Computational Linguistics.

Joseph D. Becker. 1988. Unicode 88. Standard, Unicode Consortium, Palo Alto.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Roisin Burke. 2018. Creator of Ireland's First Website Logs off after 34 Years. *EchoLIVE*.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection Criteria for Low Resource Language Programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia.

Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023a. Do not Trust the Experts: How the Lack of Standard Complicates NLP for Historical Irish. In *Proceedings of the 3d Workshop on Insights from Negative Results in NLP, EACL 2023*. In print.

Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023b. Temporal Domain Adaptation for Historical Irish. In *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), EACL 2023*. In print.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2018. Preservation of Original Orthography in the Construction of an Old Irish Corpus. In *Proceedings of the LREC 2018 Workshop: "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age"*, pages 67–70, Miyazaki, Japan.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2019. A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.

Eoin English. 2018. UCC-based Developer of Ireland's First Webpage Logs off. *The Irish Examiner*.

Stephanie Evert. 2008. The IMS Open Corpus Workbench (CWB). Retrieved: February 02, 2023.

Koustava Goswami, Sourav Dutta, and Haytham Assem. 2021a. Mufin: Enriching Semantic Understanding of Sentence Embedding using Dual Tune Framework. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2034–2039.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021b. Cross-lingual Sentence Embedding using Multi-Task Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised Deep Language and Dialect Identification for Short Texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Maarten Janssen. 2016. TEITOK: Text-Faithful Annotated Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1:7–36.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, and Iryn de Castilho, Richard Eckart anda Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Teresa Lynn. 2012. Medieval Irish and Computational Linguistics. *Australian Celtic Journal*, 10:13–27.

John P. McCrae and Theodorus Fransen. 2019. Cardamom: Comparative Deep Models for Minority and Historical Languages. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 276–279, Paris, France. European Language Resources Association (ELRA).

John P. McCrae, Atul Kumar Ojha, Bharathi Raja Chakravarthi, Ian Kelly, Patricia Buffini, Grace Tang, Eric Paquin, and Manuel Locria. 2021. Enriching a terminology for under-resourced languages using knowledge graphs. In *Proceedings of The Seventh Biennial Conference on Electronic Lexicography, eLex 2021*, pages 560–571.

Barbara McGillivray, Thierry Poibeau, and Ruiz F. Pablo. 2020. Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly*, 14(2).

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16, Cham. Springer International Publishing.

Spandan Patil, Lokshana Chavan, Janhvi Mukane, Deepali Vora, and Vidya Chitre. 2022. State-of-the-Art Approach to e-Learning with Cutting Edge NLP Transformers: Implementing Text Summarization, Question and Distractor Generation, Question Answering. *International Journal of Advanced Computer Science and Applications*, 13(1).

Priya Rani, John P. McCrae, and Theodorus Fransen. 2022. MHE: Code-Mixed Corpora for Similar Language Identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433, Marseille, France. European Language Resources Association.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *CoRR*, abs/1909.08053.

David Stifter, Nina Cnockaert-Guillou, Beatrix Färber, Deborah Hayden, Máire Ní Mhaonaigh, Joanna Tucker, and Christopher Guy Yocum. 2021. Developing a Digital Framework for the Medieval Gaelic World; Project Report. Technical report, Developing a Digital Framework for the Medieval Gaelic World.

TEI-Consortium. 1994. Text Encoding Initiative. Retrieved: February 24, 2023.

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*, 2 edition. The Dublin Institute for Advanced Studies, Dublin.

Marta Villegas, Nuria Bel, Carlos Gonzalo, Amparo Moreno, and Nuria Simelio. 2012. Using Language Resources in Humanities research. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3284–3288, Istanbul, Turkey. European Language Resources Association (ELRA).

Nora White. 2012. Ogham in 3D. Retrieved: February 24, 2023.

Dan Zeman. 2016. UD Guidelines V2. Retrieved: February 24, 2023.

Donnchadh Ó Corráin, Hiram Morgan, Beatrix Färber, Gregory Toner, Benjamin Hazard, Emer Purcell, Caoimhín Ó Dónaill, Hilary Lavelle, Seán Ua Súilleabháin, Julianne Nyhan, and Emma McCarthy. 1997. CELT: Corpus of Electronic Texts. Retrieved: February 24, 2023.