A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles

Adrian Doyle
National University of Ireland Galway
a.doyle35@nuigalway.ie

John P. McCrae
National University of Ireland Galway
john.mccrae@insight-centre.org

Clodagh Downey
National University of Ireland Galway
clodagh.downey@nuigalway.ie

Abstract

This paper examines difficulties inherent in tokenization of Early Irish texts and demonstrates that a neural-network-based approach may provide a viable solution for historical texts which contain unconventional spacing and spelling anomalies. Guidelines for tokenizing Old Irish text are presented and the creation of a character-level LSTM network is detailed, its accuracy assessed, and efforts at optimising its performance are recorded. Based on the results of this research it is expected that a character-level LSTM model may provide a viable solution for tokenization of historical texts where the use of *Scriptio Continua*, or alternative spacing conventions, makes the automatic separation of tokens difficult.

1 Introduction

Dating from about the middle of the 8th century (Stifter, 2006), the Würzburg glosses on the Pauline epistles provide one of the earliest examples of Irish text contained in manuscript contemporary with the Old Irish period of "roughly the beginning of the 8th century to the middle of the 10th century A.D." (McCone, 1997, p. 163). Aside from the Würzburg collection, the later Milan and St. Gall glosses account for the only other large collections of Irish text in manuscripts from the period. As such, the contents of these glosses are of immense cultural significance, preserving some of the earliest dated writings in the language of the Irish people. All three sets of glosses have been collected in the two-volume *Thesaurus Palaeohibernicus* (Stokes and Strachan, 1901, 1903), where the relatively diplomatic editing of the text has retained orthographic features and information from the original manuscript content (Doyle, et al. 2018). Along with faithful reproduction of the text, however, come faithful reproductions of anomalies in word spacing and spelling. Section two of this paper will detail the difficulties associated with tokenizing the Würzburg glosses as they appear in *Thesaurus Palaeohibernicus* (TPH), and of tokenizing Old Irish text more generally. Section three will address the existence of comparable tokenization issues in modern languages, and research which has been carried out in order to provide solutions in these areas. Section four will provide a rationale for the creation of tokenization guidelines specifically for use with Old Irish text in a natural language processing (NLP) context, as well as discussing the results of an inter-annotator agreement experiment which has been carried out to assess these guidelines. Finally, section five will address the creation of a character-level, long short-term memory (LSTM) based recurrent neural network (RNN) model for tokenizing Old Irish, the effects of training the model on different standards of Old Irish text, and an evaluation of its performance at the task of tokenizing the Würzburg glosses.

2 Old Irish Orthography and Linguistic Considerations for Tokenization

The language encountered in Old Irish manuscripts is surprisingly uniform, with most

variation being diachronic, "the result of morphological development" (Thurneysen, 1946, p. 12). Despite this, the text is not as orthographically consistent as readers of Modern Irish will be accustomed to, and there are certain peculiarities to be observed. These peculiarities impact the potential to carry out even rudimentary pre-processing of text by conventional means for NLP purposes, and raise questions as to how different morphemes should be combined or separated to form tokens in the first place.

It is noted by Stifter that "the orthography of Irish changed over the course of time ... so that you may find in a manuscript one word written in Old Irish, the next in Modern Irish spelling and the third in a completely odd attempt at combining different standards" (2006, p.10). While this is more evident in later manuscripts, McCone has identified features more suggestive of Middle Irish than Old in manuscripts as early as that of the Würzburg glosses (1985), and there are linguistic differences evident between the three scribal hands of the Würzburg codex, with the text of the *prima manus* suggesting a more archaic form of Irish than that of the second and third hands (Stokes and Strachan, 1901; Thurneysen, 1946).

Additionally, the division of words in Old Irish manuscripts is not directly comparable to Modern Irish. Instead, word separation is based on linguistic stress patterns with spaces occurring between accentual units. In accordance with this spacing convention, "all words which are grouped round a single chief stress and have a close syntactic connexion with each other are written as one in the manuscripts" (Thurneysen, 1946, p. 24). As such it is common for conjunctions to fall together with a following verb (*articfea = ar ticfea*, "[it] will come"), for the article to fall together with a following noun (*indindocbál = ind indocbál*, "the glorification"), for the copula to fall together with a following predicate (*isdiasom = is dia-som*, "he is God"), as well as a variety of other combinations. There are also rarer instances where separate morphemes of what may be considered the same part of speech will be separated. Take, for example, the gloss, *.i. is inse ṅduit nitú nodnail acht ishé not ail* (Wb. 5b28), "i.e. it is impossible for you (sg.); it is not you (sg.) that nourishes it, but it is it that nourishes you (sg.)." In this example the verb, *ailid*, "to nourish", is used twice, in both cases combined with the empty prefix, *no*, used to infix a pronoun. While the infixed pronoun changes between the first usage, *nodnail*, "nourishes it", and the second, *not ail*, "nourishes you", the spacing introduced between the pronoun and the verbal root in the second instance is the more notable difference. What this example demonstrates is that, not only can spacing be lacking in Old Irish manuscripts where it would be desirable to inform tokenization at the boundaries of different parts of speech, but it can also be inserted within constituent parts of a verb. An automatic tokenizer capable of processing manuscript text will need, therefore, not only to introduce spacing where it does not already exist within the text, but also to remove it where it has been employed within one part of speech to separate two accentual units.

A final consideration, related to the previous example, should be given to the phenomenon of infixing pronouns within compound verbs in Old Irish. A variety of Old Irish verbs are formed by prefixing one or more preverbal particles to a following verbal root (Thurneysen, 1946; Stifter, 2006). The simple verb, *beirid*, "to carry", forms the root of the compound verbs, *dobeir*, "to give", and *asbeir*, "to say", for example. Thurneysen (1946) refers to the preverbal particles as prepositions, this being their historical origin, however, the prepositional function of these particles is often obscured by combination with the verbal root. In this sense Old Irish compound verbs might be compared to Modern English counterparts such as "oversee" and "withdraw", where the combination takes on a new sense of its own as a completely separate verb in meaning, whereby that meaning would be lost if the verbal root were to be split from the preposition element. In such cases the compound verb is typically considered to be a word in its own right, rather than the combination of its constituent parts, and hence, it requires its own token. This poses a minor problem as regards automatically tokenizing Irish compound verbs in that a tokenizer must not split these apart when encountered. A more challenging problem is presented, however, in the way Old Irish deals with pronouns which form the objects of these compound verbs. These are infixed between the preverbal particle and the verbal root, effectively splitting what might ideally be considered a single token and requiring that another token be placed within it. To exemplify this issue, where the verb mentioned above, *dobeir*, "he gives", appears with the first singular infixed pronoun, *-m*, it becomes *dombeir*, "he gives me".

Webster and Kit (1992) make the point that the "simplicity of recognising words in English [results] from the existence of space marks as explicit delimiters". It is, perhaps based on this

same notion that Hông Phuong et al. (2008) claim "a tokenizer which simply replaces blanks with word boundaries ... is already quite accurate" for alphabetic scripts. Unfortunately, for the reasons outlined above, such an approach is not necessarily feasible with Old Irish texts. Before tokenization can be carried out decisions must be made regarding the treatment of issues outlined in this paper. These decisions will necessarily depend on the ultimate goal of the NLP tasks for which tokenization is to take place. It will, in any case, be necessary to decide whether to separate parts of speech which have been combined into accentual units, or to leave the manuscript spacing stand. It will also be important to consider how compound verbs, especially those bearing infixed pronouns, should be tokenized. The treatment of such issues, for the purposes of this paper, will be discussed further in section four.

3 A Review of Tokenization Solutions for Comparable Languages.

While the combination of issues outlined above, which hinder automatic tokenization prospects for Old Irish texts, are uncommon, particularly in European languages utilising the Roman alphabet, they are not all necessarily unique. Latin itself was typically written in *scriptio continua*, a writing style devoid of any spacing or marking to indicate word separation, until about the seventh century when Irish scribes introduced practice of word spacing to the European continent (Saenger, 1997). This timeframe would suggest that the Würzburg glosses, dating from about the middle of the eighth century, are quite an early example of text which demonstrates such spacing. The practice would not become the standard in European texts until about the thirteenth century. Tolmachev et al. (2018) present a toolkit for developing morphological analysers for *scriptio continua* languages, which utilises RNN and linear neural net models.

Turning towards modern natural languages further comparisons can be made. Tokenization solutions which have been developed for languages including Finnish (Haverinen et al., 2013; Lankinen et al., 2016), Arabic (Habash and Rambow, 2005) and Vietnamese (Hông Phuong et al., 2008) may provide a basis for developing an Old Irish tokenizer. In the case of Vietnamese, Hông Phuong et al. explain that the language uses an alphabetic script, but that spacing is used not only to separate words, but also the syllables which make up words. Furthermore, syllables, taken in isolation, are typically words themselves. When combined with other syllables, words of complex meaning are created. As such, the problem faced by Vietnamese in terms of word segmentation is comparable to that of Old Irish where compound verbs are formed by combining two or more commonly occurring parts of speech. The solution presented by Hông Phuong et al. combines a technique using finite-state automata, regular expression parsing, and a matching strategy which is augmented by statistical methods to resolve segmentation ambiguities. While these linguistic ambiguities are more comparable to the case of Old Irish, the solution requires the creation of rule-based finite-state automata, which is unfeasible in the case of Old Irish, where morphological complexity, spelling irregularities, and relative scarcity of text would suggest that manually morphologically analysing the text may be a more time efficient approach. By contrast, the approach adopted for Finnish by Lankinen et al. (2016) may provide a more viable solution for tokenizing Old Irish text. This approach utilises an LSTM based language model which uses characters as input and output, but which still processes word level embeddings.

3.1 Potential for Adapting Solutions to Old Irish Text

Conventional knowledge would suggest that, where limited text resources exist, a rule-based approach is likely to produce more accurate results than statistical or neural alternatives, albeit, often requiring more human effort. While this largely holds for languages with relatively simple morphology, like modern English, the comparatively complex morphology of Old Irish may make such an approach more difficult. Uí Dhonnchadha (2009) has produced a rule based morphological analyser for modern Irish using finite-state transducers, however, Fransen suggests in a forthcoming publication that a comparable approach may pose more difficulty for Old Irish where "Unpredictable inflectional patterns resulting from irregular syncope and analogy in inflectional patterns challenge a linguistically motivated, rule-based derivational approach." This extra complexity is compounded, Fransen continues, by a lack of resources necessary for the task, for example, "the absence of an exhaustive list of Old Irish verbs and information about stem type and stem formation." The human effort required to create such resources, and to encode rules to account for most textual eventualities, must be weighed against the

effort required for a human to manually carry out a given task on the reasonably sized, but limited, extant corpus of Old Irish literature. Given these particular circumstances, an argument may be made for the application of neural approaches to aid philologists in such tasks, even if it is unreasonable to expect particularly high accuracy without a large corpus on which to train.

As some repositories of machine-readable Old Irish text are available online a character-level LSTM based RNN approach may provide a more feasible solution than a purely rule-based model for Old Irish tokenization. The Milan and St. Gall glosses are available in online databases (Griffith, 2013; Bauer et al., 2017), meanwhile the 3,511 Würzburg glosses which appear in TPH are available in digital text (Doyle, 2018). The Corpus of Electronic Texts (CELT) (Färber, 1997) contains a collection of digital texts in Irish from the Old, Middle and Early Modern periods, and POMIC (Lash, 2014) contains a small collection of parsed Old and Middle Irish texts. As the large majority of word spacing used in the text of the Würzburg glosses does occur at word boundaries it may be possible to train a language model on these glosses themselves, and thereafter to use this model to recognise word boundaries in Old Irish text. Hence, it may be possible to tokenize the glosses using a model based on those same, untokenized glosses. As many word forms, particularly those which are always unstressed, almost never occur in the glosses without forming part of an accentual unit, however, the ability of such a model may be limited, and only common word boundary types may be recognisable. Another option is to train the model on texts drawn from the CELT collection. As many of these texts have been rigorously edited by scholars, both before and after being digitised, they not only provide a large source of text on which to train a language model, but a source of text in which word spacing is highly normalised and not based on accentual units. Normalisation standards vary from one editor to another, however, and the content of prose texts on CELT may not accurately reflect the religious vocabulary of the glosses. For these reasons, a set of guidelines for tokenizing Old Irish text have been created, and these will be discussed in section four. These guidelines will provide a standard against which to assess the accuracy of tokenizers built using LSTM RNN based language models which have been trained on text from the Würzburg glosses and from CELT.

4 Guidelines for Tokenizing Old Irish

Without consistent word spelling and consistent spacing at word boundaries tokenization by the conventional means of dividing a text into tokens based on spacing is not plausible. It has been shown in section two that the spacing conventions typically employed in Old Irish text do not permit such conventional tokenization into separate parts of speech. While, for some NLP tasks, it may be preferable to allow manuscript spacing conventions to stand and, thereby, compile a lexicon of accentual units which occur in a text, for many downstream NLP tasks it will be preferable to split such units into their component words. Fransen's (forthcoming) work, for example, outlines that "Morphological parsing operates on the word level, and words are defined as strings surrounded by space", hence, for this task it is a necessary prerequisite for words to be bounded by spaces. This necessity requires, if not a clear definition of what a word is considered to be in a given language, then, at least, a vague general notion of which combinations of morphemes constitute a word, and which constitute lower-level parts of speech. While this paper makes no attempt to provide such a definition, it has been necessary to develop a set of guidelines for tokenization, and these will be outlined in this section.

4.1 Extant Editorial Standards for Old Irish

In a language generally written without regard to rigid word boundaries, and instead divided at stress boundaries, the notion of a word is somewhat elusive. This factor contributes, no doubt, to the variation in standards for editing Old Irish texts, mentioned in section three. To exemplify this issue, take the commonly combined morphemes, *inso*, "this", frequently appearing in Irish manuscripts both within texts themselves and in many titles. In many editorial standards for Old Irish, these would be split apart into the article, *in*, and the demonstrative pronoun, *so*. Despite this, Stifter's practice in *Sengoidelc* (2006) is to represent the combination separated with a hyphen, *in-so*, both in a section explaining the use of demonstratives (p. 103) and in continued examples thereafter (p. 130, 26.3, eg. 6). It may not have been intended to suggest that the combination be treated as a single token, however, it nicely demonstrates the variation which can exist, even in standardised Old Irish texts.

Another area where much variation occurs in edited texts is in the treatment of enclitics, such as the emphatic suffixes and the anaphoric *suide*. In many editions the decision to present such morphemes as either enclitic, attached to a preceding part of speech by a hyphen, or as tokens separated from a preceding word, is dependent on which of the two is stressed. One edition of *Tochmarc Emire la Coinculaind* available on CELT (Färber, 1997), for example, contains the line, "*Atbert som fris-som*...", "he said to him". While significant linguistic reasons may exist for editorial decisions to treat comparable parts of speech in varying ways, this variety does not provide a good basis for tokenization. If, as suggested earlier, the goal is to split parts of speech without regard to accentual units, all occurrences of individual parts of speech which are performing an identical function should ideally be tokenized consistently. In other texts on the site preverbal particles are variously hyphenated, completely attached, or separated from the following verb by a space. In the case of particles like *ro* and *no*, the practice of separating them from the following verb may in some cases be desirable in order to identify very low level parts of speech at a later stage, however, this can create difficulty when preverbs are compounded and reduced as with *ro* in *do·á-r-bas*, "has been shown" (Thurneysen, 1946, p. 340). The problem in these cases is that the reduced particle is not typically removed or separated from the verbal root in editions. Again, this creates a situation where a part of speech with a single function is treated differently when it does not occur immediately at the beginning of a verb. Ideally a more universal editorial standard might be adhered to, however, in lieu of such a standard, the guidelines proposed below for tokenization will be based largely on extant editorial standards and will specify the reason for any variation from such standards.

4.2 Tokenization Guidelines for this Experiment

In developing guidelines for tokenization Old Irish, a balance must be struck between tailoring tokens to account for the complex morphology of the language and tailoring them to account for the relative scarcity of text resources which are digitally available. The lack of a large, universally standardised, corpus of Old Irish text limits the amount of data with which to train statistical or neural network models. As such, the guidelines for tokenization listed below have been developed so as to avoid creating a wide variety of infrequently occurring tokens. As such, frequently occurring affixes such as demonstrative and emphatic suffixes are always separated from preceding tokens and considered to be tokens themselves. An exception to this rule is made for preverbal particles, which are instead taken to be a constituent part of a following verb. While this will create a larger variety of verbal tokens, it has been shown above that the separation of these particles is not always feasible, particularly where they are compounded or reduced.

The case of verbs containing infixed pronouns requires particular attention. These guidelines recommend treating the entire verbal complex as an individual token. This will allow for verbs with infixed pronouns to be treated as morphological variants of the base verb form in part-of-speech tagging, which is necessary as the inclusion of an infixed pronoun can affect the morphology of the preverbal particle in some instances. Thurneysen points out that "the *-o* of *ro, no, do, fo* is lost before initial *a*" (1946, p. 257). For example, *dogníu*, "I do", loses the *-o* of the preverbal particle, *do*, and becomes *dagníu*, "I do it", with the third person, singular, neuter pronoun, *a*, infixed. This morphological change to the particle constitutes an alteration of the verb, and therefore would require the entry of an alternative form in a lexicon. However, as this form cannot occur without the infixed pronoun which is causing it, the entire complex should be taken as being the alternate form. Future work will look at part-of-speech tagging, and the possibility of extracting infixed pronouns and tagging them separately at that stage will be explored. In the current work, however, they will be treated, as outlined above, as internalised tokens.

Aside from internalised tokens, the guidelines account for one more form of specialised token. Where forms of a significant part of speech such as the article, the copula, or possessive pronouns occur in reduced or altered form when combined with other tokens, these forms are considered to be conjoined tokens. For example, where the article is preceded by prepositions such as *co, i* and *fri*, giving rise to combined forms such as *cosin, isnaib* and *frisna*, the separated forms of the article, *-sin, -snaib,* and *-sna* are conjoined tokens. Similarly, when possessive pronouns precede or follow vowels, they take on a conjoined form, with examples such as *id*, "in your" and *manam* (Wb. 17c4a), "my soul", containing the conjoined tokens *-d* and *m-* respectively. While conjoined tokens in the guidelines are displayed with a

hyphen to demonstrate their dependency on a preceding or following token, this is removed in implementation, hence, *manam* should be rendered *m anam*.

Aside from the token types outlined in this section and those parts of speech mentioned earlier in this paper, there are few common disagreements in editorial standards. It is hoped that the guidelines outlined here will provide a reasonable baseline for measuring success in automatic tokenization, however, on the basis of varying requirements for varying tasks, a different style of tokenization may be required, and so, alteration to these guidelines.

4.3 Inter-Annotator Agreement

An inter-annotator agreement experiment has been carried out using the tokenization guidelines detailed above. Four annotators have been shown forty-one glosses selected from the Würzburg corpus (Doyle, 2018), and asked to introduce or remove spacing as necessary in accordance with the guidelines. Annotators were instructed not to introduce or remove any letters, hyphens or other non-space characters. During the timeframe of the experiment three annotators were PhD candidates in the field of Early Irish, and the fourth was a postdoctoral researcher in the same field.

Before being shown the guidelines, two of the annotators were asked to perform the task of separating words, by introducing or removing spaces only, based on their intuitive understanding of how word division should be implemented. These two annotators were shown the guidelines only after this first run had been completed, and were asked again to carry out the task, this time adhering to the guidelines. This allows a comparison to be made between annotators working both with and without the guidelines.

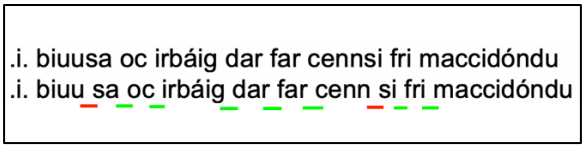

Figure 1: Agreement (green) and disagreement (red) between two annotators

Agreement between annotators was measured by determining which particular letters in a string are followed by a space in any annotator's work, then comparing two annotators work to see if they agreed on the inclusion or exclusion of a space at a given point, or if they disagreed with one including a space, and another not doing so. See an example of agreement and disagreement between two annotators in Figure 1.

	Cohen's Kappa Score
Pair 1 – (A1 + A2)	0.469
Pair 2 – (A1 + A3)	0.349
Pair 3 – (A1 + A4)	0.655
Pair 4 – (A2 + A3)	0.191
Pair 5 – (A2 + A4)	0.457
Pair 6 – (A3 + A4)	0.297
Annotator Average Score	0.403
No Guidelines	-0.058

Table 1: Inter-annotator agreement Cohen's kappa scores for each pair of annotators, and average

Cohen's kappa coefficient was used to compare the work of each pair of annotators using the guidelines. Table 1 shows that the highest agreement between two annotators using the guidelines was substantial at 0.65, while the lowest, at 0.19, was higher than would be expected by chance. The average score between pairs of annotators was 0.40 suggesting that the guidelines may require further clarification on some points. It is, however, noteworthy that the guidelines seem to ensure higher agreement than might be expected of annotators working without them, at least, when compared to the score of the two annotators work before they had been shown the guidelines, -0.058.

The results of this inter-annotator experiment will be used in section five as a means of comparing human performance at a tokenization task against that of the LSTM-based tokenizer model detailed in this paper.

5 A Character-Level LSTM Recurrent Neural Network Model for Tokenizing Old Irish

A Character-Level LSTM RNN model was created using TensorFlow and Keras. The purpose of the RNN is to model the language of the text it is trained on and develop an understanding of which sequences of characters are likely to indicate a word ending. A function has been developed so that this model can be utilised to identify points in a text where it is likely that word division should be added, and spacing is introduced at these points, thereby, allowing tokenization to be carried out by more conventional means. The development of this tokenizer and its evaluation is detailed in this section.

5.1 Pre-processing Text for Training and Evaluation

It was determined that the text of the Würzburg glosses (Doyle, 2018) contained fifty-two characters once all Latin text, and all editorial punctuation, commentary and brackets had been removed. An arbitrary, out of vocabulary character was introduced for use in padding sequences, bringing the character count to fifty-three. The only remaining punctuation in the glosses occurs in abbreviations such as *.i.* and *ł.* In these instances, the punctuation which occurs is taken to be part of the token, hence, such punctuation was not removed in pre-processing. It is also noteworthy that, with the exception of some roman numerals and Latin names, all of which had been removed by this point in the processing, very few upper-case letters are used throughout the glosses.

The forty-one glosses utilised in the inter-annotator agreement experiment were removed from the corpus to be used as a test set in a later evaluation stage. At this point the remaining glosses were concatenated to form a single string. This string was the first of two training sets used in this experiment. The second training set was drawn from texts available on CELT (Färber, 1997). Ten texts were selected which were deemed both to be reasonably long and also to be edited to a standard comparable to one another:

- Táin Bó Regamna
- Táin Bó Fraích
- Táin Bó Cúailnge Recension I
- Táin Bó Cúalnge from the Book of Leinster
- Compert Con Chulainn
- Serglige Con Culainn
- Tochmarc Emire la Coinculaind (Harl. 5280)
- Tochmarc Emire la Coinculaind (Rawlinson B 512)
- Fled Bricrend (Codex Vossianus)
- The Training of Cúchulainn

The texts were concatenated together to form one string, and all characters were changed to lower-case. A number of characters which did not transfer cleanly into UTF-8 format had to be manually corrected. Other alterations included the automatic removal of editorial notes and folio information, editions which use the letter *v*, which does not occur in the Würzburg character-set, were altered and the letter *u* was substituted in its place. Finally, in an attempt to align the various editorial standards with the tokenization guidelines, a script was written using regular expressions to identify common preverbal particles which had been separated from a following verb and attach them to it, and similarly, to find common suffixes attached to preceding words by hyphenation and detach them. This approach runs the risk of accidentally splitting genuine tokens where part of the token matches the regular expression used. It would be preferable to train on a corpus where the editor had deliberately edited using this standard, however, this was the most feasible solution with the available editions. With the two separate training corpora having been created, the following steps were applied to each before training on them.

The training corpora were sequenced into strings of ten. For every string of eleven characters in the training corpus, the first ten characters were added to a list of training strings, and the eleventh was added to a set of associated labels. Each label, therefore, is the character which directly follows the preceding string of ten characters. Finally, each sequence of characters, and label, were converted into one-hot vectors with a length of fifty-three to account for each character. Before training, ten percent of sequences and labels were set aside. During the training process these were used to validate the accuracy of the model by testing it on unseen sequences. This step helped to prevent overfitting of the model.

5.2 Developing the Model

It was decided to build a character level model so that the network could learn which sequences of characters are most likely to signify a word ending. LSTM cells were utilised in the RNN to enable dependencies to be learned by the model over long distances, as some rare morphological features may occur infrequently in a text, and hence, may be spread far apart in a string of characters. Backpropagation is used by RNNs in order to improve at a given task over time. Error signals flow backwards through the network and weights between cells are recalibrated to improve accuracy. Over time conventional networks' evaluation of backpropagated error signals tend to either increase or decrease exponentially (Hochreiter and Schmidhuber, 1997). This results in a network which may be accurate in the short term, but which becomes increasingly incapable of pattern recognition in the long term, for example, over long strings of text. LSTM RNNs attempt to overcome this issue, whereby error evaluation either explodes or vanishes over time, by intelligently "forgetting" error information as it becomes irrelevant to the system. This is an important improvement as, generally, the more data which a network can train on, the more

accurately the network can identify patterns. Sundermeyer et al. write of language modelling, "the probability distribution over word sequences can be directly learned from large amounts of text data..." (2015, p. 517). A similar approach will be used here, instead attempting to learn probability distributions over character sequences in order to identify word endings.

No. of Hidden Layers	2
Hidden Layer Size	53
Input Format	53x10 Vector
Output Format	53 (Model 1) **OR** 2 (Model 2)
Optimiser	Adam
Loss Function	Categorical Cross-entropy

Table 2: Hyperparameters for the RNN

Through experimentation it was determined that the most accurate model was achieved utilising two hidden layers of LSTM cells. The number of cells in each hidden layer was equal to the length of the one-hot vectors, as this was found to be the most accurate without causing overfitting. No attempt was made to train using batches. See table 2 for more information on the model's hyperparameters.

Two variants of the model were created. The first was designed to guess the following character based on the sequence of ten characters it was shown, and the second was designed only to guess only whether the following character would be a space or not. These will be referred to as Model 1, and Model 2, respectively.

5.3 Designing the Tokenizer

At first, a function was created in order to tokenize strings of text using the model. The function takes each character in the string and uses the model to determine if the next character should be a space or not. The next subsection will detail how the models and tokenizers were evaluated, however, this tokenizer's performance was deemed to be unsatisfactory.

To improve performance a second, reverse model was trained. This model works backwards through the training text and attempts to predict a character preceding a given input sequence. Once this model had been trained the tokenization function was adapted to include it. For each character in a string which is fed into the function, the forward model predicts whether a space should be introduced after it. If the forward model predicts a space, the reverse model is shown the following ten characters to make a prediction whether a space should precede them. A space is introduced only if the two models agree that a space should be introduced at a given point. Similarly, the function looks at spaces already in the string which is fed into it and seeks agreement from the models as to whether to remove the space or leave it in the string. Finally, the function outputs the string with new spaces included, and potentially with some spaces removed. This combined forward-reverse tokenizer was found to be more accurate than one based on either the forward or reverse models alone.

5.4 Evaluation

During the training of models, a wide variety of parameters were experimented with in order to produce the best possible model. At this stage training accuracy was measured using TensorFlow's built-in TensorBoard. This also enabled loss to be measured over the time taken to train a given model.

As mentioned above, ten percent of all training sequences were split off and used to validate accuracy and loss scores by periodically testing the model-in-training on unseen sequences and labels. At the point in training when validation loss began to increase, training was stopped in order to prevent overfitting. This generally occurred at about 24 epochs when training on sequences from the first training set drawn from the glosses, and at about 8 epochs when training on the larger collection of texts of the second training set. It is also notable that the accuracy for Model 1 was consistently lower than that of Model 2. The highest validation accuracy score for Model 1 peaked at about 36%, while that of Model 2 reached a peak of 92% accuracy. These scores were not apparently affected by the training set used, and both training sets used with Model 1 reached the 92% accuracy score on the validation set. This suggests that the task of predicting word endings only was easier for models than the task of predicting any of the potential fifty-two characters.

While an accuracy of 92% is reasonably high for an RNN trained on a limited amount of text, it should be remembered that a tokenizer built on a model with this accuracy score would insert or remove a space incorrectly about once for every ten characters in a given string. This may explain why the performance of the forward only tokenizer design was unsatisfactory. In any case,

the accuracy score of a model is not necessarily an accurate indicator of how well a tokenizer built on that model will work. This is especially true in the case of tokenizers built on Model 1, where the tokenizer function ignores all character predictions other than ones which would introduce or remove a space.

Tokenization accuracy, therefore, needs to be measured by separate means to those described above for evaluating LSTM models. For this purpose, four tokenizers were used to tokenize the forty-one glosses used in the inter-annotator agreement assessment. Information regarding the model and training set used to create each tokenizer can be seen in table 3.

Tokenizer	Model	Training Set
T1	Model 1	Wb. text
T2	Model 1	CELT texts
T3	Model 2	Wb. text
T4	Model 2	CELT texts

Table 3: Tokenizers, models and training texts

In order to quantify the success of each tokenizer the output of each model was compared against the work of each annotator, again using Cohen's kappa coefficient (see table 4).

	A1	A2	A3	A4
T1	0.2703	0.2225	0.2842	0.2693
T2	0.0297	0.0172	0.0563	0.0355
T3	0.2494	0.1974	0.2613	0.2431
T4	0.1836	0.1408	0.1805	0.1701

Table 4: Measurement of annotators' work (A1-4) compared against output of tokenizer models (T1-4) using Cohen's kappa

These results show that no tokenizer performed worse than the two human annotators working without guidelines, while the better performing tokenizers show a higher score than at least one pairing of human annotators working with guidelines. This seems to suggest that a neural approach may provide a feasible solution for automatic word segmentation in unedited Old Irish texts. It is interesting that the best performing tokenizer (T1) was trained on the glosses themselves, rather than on a larger amount of text which has been edited to a desirable standard. It may be the case that out-of-vocabulary terminology in the glosses reduces the effectiveness of models trained on prose text. Future work, therefore, will focus on applying a bootstrapping approach to tokenization of the glosses. Models will be periodically trained on manually tokenized glosses and tested against this same test set until an improvement is noted over the current models. It is expected also that training on a corpus of edited gloss material will increase performance, therefore, going forward, attempts will be made to improve the techniques detailed here by training similar tokenizers on the text of the St. Gall glosses (Bauer, et al., 2017). Further improvements may be gleaned by the addition of a simple rule-based output layer which would make sure that easily identifiable features, such as common particles, abbreviations, and initial mutations, are appropriately bounded by spacing.

6 Conclusion

This paper has examined difficulties inherent in tokenization of Early Irish texts and presented guidelines for tokenization developed with these particular difficulties in mind. These guidelines have been shown to improve inter-annotator agreement on a word segmentation task. A character-level LSTM based RNN was developed to automatically tokenize Old Irish text and demonstrated potential. It may be possible to improve upon performance by training on a corpus of pre-processed glosses, as prose material appears to be less suitable, and by the addition of a rule-based output layer.

Acknowledgements

Particular thanks are owed to the following annotators who have, thus far, contributed their time and expertise to the inter-annotator agreement experiment detailed in this paper: Maria Hallinan (NUIG), Dr. Daniel Watson (DIAS), Theodorus Fransen (TCD), and Jody Buckley-Coogan (QUB).

This work has been funded by the National University of Ireland, Galway, through the Digital Arts and Humanities Programme, and is also supported by the Irish Research Council through the Government of Ireland Postgraduate Scholarship Programme.

Dr. McCrae's research is supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

Bauer, Bernhard, Rijcklof Hofman and Pádraic Moran. 2017. *St. Gall Priscian Glosses, version 2.0* <www.stgallpriscian.ie/> (Accessed: 08/05/2019).

Doyle, Adrian. 2018. Würzburg Irish Glosses, <www.wuerzburg.ie> (Accessed: 08/05/2019)

Doyle, Adrian, John P. McCrae and Clodagh Downey. 2018. *Preservation of Original Orthography in the Construction of an Old Irish Corpus*. Proceedings of the 3rd Workshop on Collaboration and Computing for Under-Resourced Languages (CCURL 2018). Miyazaki, Japan.

Färber, Beatrix (ed.). 1997. *CELT Corpus of Electronic Texts*. <https://celt.ucc.ie/irlpage.html> (Accessed: 08/05/2019)

Fransen, Theodorus. Forthcoming. *Automatic Morphological Analysis and Interlinking of Historical Irish Cognate Verb Forms*. Elliott Lash, Fangzhe Qiu and David Stifter (eds.). Corpus-based Approaches to Morphosyntactic Variation and Change in Medieval Celtic Languages. De Gruyter, Berlin.

Griffith, Aaron. 2013. *A Dictionary of the Old-Irish Glosses*. <https://www.univie.ac.at/indogermanistik/milan_glgloss/> (Accessed: 08/05/2019).

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski and Filip Ginter. 2013. *Building the essential resources for Finnish: the Turku Dependency Treebank*. Language Resources and Evaluation.

Habash, Nizar and Owen Rambow. 2005. *Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. Proceedings of the 43rd Annual Meeting of the ACL, pp. 573-580.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. *Long Short-Term Memory*. Neural Computation, Vol. 9, No. 8, pp. 1735-1780.

Hông Phuong, Lê, Nguyên Thi Minh Huyên, Azim Roussanaly and Hô Tuòng Vinh. 2008. *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain.

Kavanagh, Séamus and Dagmar S. Wodtko. 2001. *A Lexicon of the Old Irish Glosses in the Würzburg Manuscript of the Epistles of St. Paul*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Lankinen, Matti, Hannes Heikinheimo, Pyry Takala, Tapani Raiko and Juha Karhunen. 2016. *A Character-Word Compositional Neural Language Model for Finnish*.

Lash, Elliott. 2014. *The Parsed Old and Middle Irish Corpus (POMIC)*. Version 0.1. <https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/> (Accessed: 08/05/2019).

Lynn, Teresa. 2012. *Medieval Irish and Computational Linguistics*. Australian Celtic Journal, 10:13-28.

McCone, Kim. 1985. *The Würzburg and Milan Glosses: Our Earliest Sources of 'Middle Irish'*. Ériu, 36:85-106.

McCone, Kim. 1997. *The Early Irish Verb*. An Sagart, Maynooth, 2nd edition.

Mullen, Lincoln. A., Kenneth Benoit, Os Keyes, Dmitry Selivanov and Jeffrey Arnold. 2018. *Fast, Consistent Tokenization of Natural Language Text*. Journal of Open Source Software.

Saenger, Paul. 1997. *Space Between Words: the Origins of Silent Reading*, Stanford University Press, Stanford, California

Stifter, David. 2006. *Sengoidelc*. Syracuse University Press, New York.

Stokes, Whitley and John Strachan. (Eds.). 1901. *Thesaurus Palaeohibernicus Volume I*. The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Stokes, Whitley and John Strachan. (Eds.). 1903. *Thesaurus Palaeohibernicus Volume II.* The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Sundermeyer, Martin, Hermann Ney and Ralf Schlüter. 2015. *From Feedforward to Recurrent LSTM Neural Networks for Language Modeling*. Audio Speech and Language Processing IEEE/ACM Transactions on, Vol. 23, No. 3, pp. 517-529.

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

Tolmachev, Arseny, Daisuke Kawahara and Sadao Kurohashi. 2018. Juman++: *A Morphological Analysis Toolkit for Scriptio Continua*. ACL.

Uí Dhonnchadha, Elaine. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish Using Finite-State Transducers and Constraint Grammar*. PhD Thesis, Dublin City University.

Webster, Jonathan J. and Chunyu Kit. 1992. *Tokenization as the Initial Phase in NLP*. Proceedings of the 14th Conference on Computational Linguistics, Nantes, France.