TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish

Lauren Cassidy[1], Teresa Lynn[1], James Barry[1], Jennifer Foster[2]
[1,2] School of Computing, Dublin City University
[1] ADAPT Centre
[1] {lauren.cassidy, teresa.lynn, james.barry}@adaptcentre.ie
[2] {jennifer.foster}@dcu.ie

Abstract

Modern Irish is a minority language lacking sufficient computational resources for the task of accurate automatic syntactic parsing of user-generated content such as tweets. Although language technology for the Irish language has been developing in recent years, these tools tend to perform poorly on user-generated content. As with other languages, the linguistic style observed in Irish tweets differs, in terms of orthography, lexicon, and syntax, from that of standard texts more commonly used for the development of language models and parsers. We release the first Universal Dependencies treebank of Irish tweets, facilitating natural language processing of user-generated content in Irish. In this paper, we explore the differences between Irish tweets and standard Irish text, and the challenges associated with dependency parsing of Irish tweets. We describe our bootstrapping method of treebank development and report on preliminary parsing experiments.

1 Introduction

Irish is a minority language spoken mostly in small communities in Ireland called 'Gaeltachtaí' (CSO, 2016) but social media sites, such as Twitter, provide a platform for Irish speakers to communicate electronically from any location. Users may reach a wide audience quickly, unconstrained by the conventions of standard language upheld by editors in publications, revealing the orthographic, lexical, and syntactic variation abundant in informal Irish. Analysis of up-to-date, real-world language data can provide an insight into how Irish is used in everyday communication and how such informal texts compare to prescriptive norms of standardised language to which published texts tend to adhere.

User-generated content (UGC), such as tweets, is a valuable, highly available resource for training syntactic parsers that can accurately process social media text. UGC is a genre with features different from those of both spoken language and standardised written language more traditionally found in natural language processing (NLP) corpora. Plank (2016) notes the advantages of utilising *fortuitous data* in order to create more adaptive, robust language technology.

Given that the accuracy of syntactic parsing tools has been shown to decline when evaluated on noisy UGC data (Foster et al., 2011; Seddah et al., 2012) and that domain[1] adaptation has been shown to improve parser performance for dependency annotation of English tweets (Kong et al., 2014) and POS-tagging in Irish tweets (Lynn et al., 2015), the need for genre-specific resources is clear in order to reliably process this variety of data. The prerequisite, therefore, for research in this area is a data set of Irish UGC. This research attempts to fill this gap through the development of TwittIrish, a treebank of Irish tweets, within Universal Dependencies (UD) (Nivre et al., 2020), a cross-lingually consistent framework for dependency-based syntactic parsing. TwittIrish provides linguistic information for Irish in a digitally accessible format valuable for linguistic research and the development of NLP tools.

Open-source projects such as UD facilitate collaboration and rapid evolution of ideas among linguists internationally. In order to maintain optimum consistency with other UD treebanks, the annotation methodology employed in this research closely follows the general UD guidelines and the language-specific guidelines for Irish while aiming to incorporate the most up-to-date recommendations (Sanguinetti et al., 2022) for UGC in this evolving area of NLP. UGC, especially social media text, has recently become a popular focus within UD and NLP research more broadly (Silveira et al., 2014; Luotolahti et al., 2015; Albogamy and Ram-

[1] The terms genre and domain are used interchangeably throughout this paper to refer to the category of text such as standard published text or Twitter text.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6869 - 6884
May 22-27, 2022 ©2022 Association for Computational Linguistics

say, 2017; Wang et al., 2017; Zeldes, 2017; Bhat et al., 2018; Blodgett et al., 2018; Van Der Goot and van Noord, 2018; Cignarella et al., 2019; Seddah et al., 2020) and has encouraged active conversation around how best to represent it within this framework among the UD community.

We carry out preliminary parsing experiments with TwittIrish, investigating the following two questions: How effective is a parser trained on the Irish UD Treebank (Lynn and Foster, 2016), which contains only edited text and no UGC, when applied to tweets? And what difference do pre-trained contextualised word embeddings make? We observe a difference of approximately 23 LAS points between TwittIrish and the IUDT test set and find that the use of monolingual BERT embeddings (Barry et al., 2021) improves performance by over 10 LAS points.

The paper is structured as follows: Section 2 details the existing Irish NLP resources we use for our research, Section 3 outlines the development of the treebank, Section 4 describes the characteristics of UGC evident in Irish tweets, and Section 5 presents parsing experiments and error analysis.

2 Irish NLP Resources

We use the following resources:

Indigenous Tweets (IT)[2] This project compiles statistics on social media data of 185 minority and indigenous languages including Irish. All tweets in the TwittIrish treebank were sourced via IT.

Lynn Twitter Corpus (LTC)[3] **(Lynn et al., 2015)** A corpus of 1,493 lemmatised and POS-tagged Irish language tweets randomly sampled from 950k tweets by 8k users posted between 2006 and 2014, identified by IT. The LTC data also contains code-switching information (Lynn and Scannell, 2019).

Irish Universal Dependencies Treebank (IUDT)[4] **(Lynn and Foster, 2016)** A UD treebank consisting of 4,910 sentences sampled from a balanced mixed-domain corpus for Irish.

gaBERT (Barry et al., 2021) A monolingual Irish BERT model, trained on approximately 7.9 million sentences, which outperforms Multilingual BERT (mBERT) (Devlin et al., 2019) and WikiBERT (Pyysalo et al., 2021) at the task of dependency parsing for Irish.

3 TwittIrish Development

We combined 700 POS-tagged tweets from the LTC with 166 tweets more recently crawled by IT in order to leverage previous linguistic annotations while also including newer tweets. This involved converting the LTC annotation scheme to that of the UD framework and then POS-tagging the new raw tweets. We provide further detail in Appendix A.

LTC conversion With regard to tokenisation, multiword expressions were automatically split into separate tokens following UD conventions. Only minor manual adjustments were required for lemmatisation to ensure alignment with the IUDT (to enable bootstrapping – see Section 3). Finally, the POS tagset used in the LTC was automatically converted to the UD tagset. Appendix A.2 describes this process.

Preprocessing of newly-crawled tweets Due to the lack of a tokeniser designed to deal specifically with UGC in Irish, we compared two tools for this task: UDPipe (Straka et al., 2016),[5] a language-agnostic trainable pipeline for tokenisation, tagging, lemmatisation and dependency parsing, and Tweettokenizer[6] from NLTK (Bird et al., 2009), a rule-based tokeniser designed for noisy UGC. The latter proved to be more effective for tokenising UGC phenomena such as emoticons, URLs, and meta language tags. Manual corrections were then applied in order to adhere to the Irish-specific tokenisation scheme within current UD guidelines. In order to establish the best system to use for automatic lemmatising and POS-tagging, two tools, Morfette (Chrupala et al., 2008) and UDPipe (Straka et al., 2016), were analysed with Morfette achieving higher scores on both tasks.

Syntactic annotation As a method shown to reduce manual annotation efforts in syntactic annotation (Judge et al., 2006; Seraji et al., 2012), we carry out a bootstrapping approach to dependency parsing as recommended by UD. [7]

The bootstrapping process is illustrated in Figure 1. After converting the LTC and new tweets

[2] http://indigenoustweets.com/

[3] https://github.com/tlynn747/IrishTwitterPOS

[4] https://github.com/UniversalDependencies/UD_Irish-IDT

[5] Trained on IUDT v2.8 with no pre-trained embeddings.

[6] https://www.nltk.org/api/nltk.tokenize.html

[7] https://universaldependencies.org/how_to_start.html

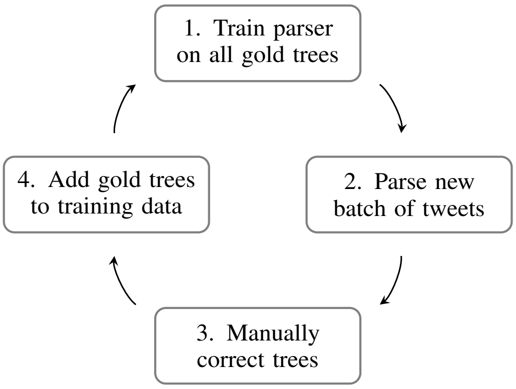
1. Train parser on all gold trees
2. Parse new batch of tweets
3. Manually correct trees
4. Add gold trees to training data

Figure 1: Bootstrapping approach to semi-automated syntax annotation.

to the CoNLL-U format, we manually annotated a small set of 166 tweets and began the bootstrapping cycle.[8] (Step 1) A parsing model[9] was trained on the IUDT in combination with the newly annotated tweets. (Step 2) The parsing model was used to automatically annotate the next batch of 100 tweets. (Step 3) These tweets were manually corrected. (Step 4) The corrected tweets were added to the training data. Steps 1 to 4 were repeated until all 866 tweets were fully parsed. This dataset represents the TwittIrish test set in the UD version 2.8 release.[10]

4 Annotating Irish UGC

This section describes the linguistic features that can create challenges when parsing Irish social media text. We provide Irish examples and discussion around the factors that influence these phenomena.

4.1 Orthographic Variation

Orthographic variation refers to deviation from the conventional spelling system of the language and is observed at the token level. Therefore, it can affect the lemmatisation of a token in an NLP pipeline, potentially affecting other downstream areas of annotation. In the TwittIrish dataset, 2.5% of tokens contained some orthographic variation. Table 1 exemplifies some frequently-occurring phenomena in Irish tweets that deviate from standard orthography.

Diacritic variation Diacritic marks are often omitted or incorrectly added to tweets. The acute accent or *síneadh fada* is used in Irish to indicate a long vowel and is necessary to disambiguate between certain words. Example 1 shows the most probable intended word *léacht* 'lecture' rendered as *leacht* 'liquid'.

(1) ***Leacht*** *faoi stair Príosún Dún Dealgain*
'Lecture about the history of Dundalk Prison'

Abbreviation Predictable shorthand forms can occur in standard Irish texts e.g. *lch* as an abbreviated form of *leathanach* 'page'. While more unconventional, and thus less predictable, abbreviations are observed in Irish tweets, as per Example 2 in which the word *seachtain* 'week' is shortened to *seacht* 'seven'. Abbreviations are more common in tweets than standard text as the character limit and real-time, up-to-date nature of the platform encourages the user to be efficient with time and space.

(2) *Bím de ghnáth ach sa bhaile an* ***tseacht*** *seo*
'I usually am but home this week'

Lengthening This refers to the elongation of a token by repeating one or more characters. This can be thought of as an encoding of sociophonetic information (Tatman, 2015) and is strongly linked to sentiment. Despite incentives to save time and space while tweeting, users often elongate certain words for expressive purposes (Brody and Diakopoulos, 2011). Example 3 demonstrates the lengthening of the word *buí* 'yellow'.

(3) *tá siad go léir* ***buuuuuuí***
'They are all yelloooooow'

Case variation Nonstandard use of upper- and lowercase text is another method of encoding sociophonetic information by focusing attention or emotion on a particular word or phrase. Heath (2021) discusses the association between the use of all-caps and perceived shouting as in Example 4.

(4) *Níl todhchaí na Gaeilge sa Ghaeltacht, ach in aon áit* ***AR DOMHAIN***
'The future of Irish is not in the Gaeltacht but anywhere ON EARTH'

[8] Due to the limited funding available, all manual annotation and correction was performed by one linguist annotator.

[9] Biaffine Parser (Dozat and Manning, 2017) with mBERT (Devlin et al., 2019) embeddings.

[10] The TwittIrish Treebank is available here: https://github.com/UniversalDependencies/UD_Irish-TwittIrish/tree/master

Phenomenon	**Example**	**Standard form**	**Gloss**
Diacritic variation	***nior*** *fhoghlaim* ***tu***	*níor fhoghlaim tú*	‘you did not learn’
Abbreviation	***fhoir*** *rugbaí na* ***hÉir***	*fhoireann rugbaí na hÉireann*	‘Irish rugby team’
Lengthening	***obairrrr***	*obair*	‘work’
Case variation	*ceolchoirm* ***DEN SCOTH***	*ceolchoirm den scoth*	‘excellent concert’
Punctuation Variation	**folúntas**	*folúntas*	‘vacancy’
Transliteration	*go* ***wil***	*go bhfuil*	‘that is’
Other spelling variation	*O ’Bama*	*Obama*	‘Obama’

Table 1: Examples of orthographic variation in Irish tweets.

Transliteration The practice of transliteration, in which a word in one language is written using the writing system of another, is common within the language pair of Irish and English. In the TwittIrish treebank, the English language phrase ‘fair play’ occurs twice while variations ‘fair plé’, as shown in Example 5 and ‘féar plé’ occur once each.

(5) ***Fair plé*** *daoibh* 💗’
‘Fair play to you 💗’

Punctuation variation Punctuation is used creatively in UGC to format or emphasise strings of text. However, due to the lack of standardisation, occurrences of unconventional punctuation can make text difficult to parse for both human and machine, as in Example 6 which shows a phrase from an Irish tweet appended by two punctuation characters ‘-)’. It is unclear whether this should be interpreted as some form of punctuation, creative formatting, or a smiley e.g. ‘:-)’.

(6) *sin a dhóthain**-)***
‘That‘s enough-)’

Other spelling variation These are mostly slight variations very close to the intended word and may occur due to typographical error. Typos are very common in UGC due to lack of editing or proofreading and may occur via insertion, deletion, substitution, or transposition of characters. Example 7 shows *sraith* (season) rendered as **staith*. Due to their phonetic dissimilarity and the fact that ‘t’ and ‘r’ are adjacent on the QWERTY keyboard layout, it is reasonable to infer that the substitution was unintentional. Less commonly, disguise or censorship of words or phrases may occur to encrypt profanity or taboo language.

(7) tus **staith** 6 de Imeall
‘start of season 6 of Imeall’

4.2 Lexical Variation

Just 38.32% of the set of unique lemmata that make up the vocabulary of the TwittIrish treebank occur in the IUDT training data. Table 2 shows examples of lexical variation in Irish tweets.

Dialectal vocabulary Irish has three major dialects; Connaught, Munster, and Ulster. Distinctive features of these dialects in the form of lexical variation are evident in spoken language and informal text such as tweets. Example 8 shows the use of *domh*, the Ulster variant of *dom* ‘to me’.

(8) *Ba chóir* ***domh*** *rá!*
‘I should say!’

Initialism Multiword phrases are frequently represented by the initial letter of each of their constituent tokens. Example 9 shows *GRMA* ‘Thank you’ used to represent its expanded form *Go raibh maith agat*.

(9) *Scaip an scéal!* ***GRMA!***
‘Spread the word! Thank you!’

Pictogram Emojis, emoticons, etc. can be added to text to emulate gesture (Gawne and McCulloch, 2019) or they may play a syntactic role in a phrase, replacing a word as in Example 10, in which the symbol, 🖤, acts as the object of a verb. Pictograms tend not to have a one-to-one correspondence with natural language words.

(10) *Conas a deireann tú* 🖤*?*
‘How do you say 🖤’

Truncation Due to the current limit of 280 characters per tweet, the end of a tweet may be unnaturally attenuated, sometimes mid-sentence as in Example 11 or even mid-word.

(11) *Súil agam go bheas sé mar sin don…*
‘I hope it will be like that for the…’

Code-switching vs. borrowing 66.74% of tokens in the TwittIrish treebank are in Irish, 4.85% of tokens are in English and the remainder (consisting of punctuation, meta language tags, etc.)

Phenomenon	**Example**	**Standard form**	**Gloss**
Dialectal vocabulary	**fé**	*faoi*	'about'
Initialism	**BÁC**	*Baile Átha Cliath*	'Dublin'
Pictogram	*<3 mór*	*Grá mór*	'Lots of love'
Truncation	*thart fa' 53* **nó...**	*thart fa' 53 nóiméad*	'over 53 mi... (minutes)'
Code-switching vs. borrowing	*sa* **town** *amárach*	*sa bhaile amárach*	'in town tomorrow'
Other nonstandard lexical forms	**TochaltÓr**	*Tochaltóir óir*	'Gold-digger'

Table 2: Examples of lexical variation in Irish tweets.

are classified as neither, or indeed both in the case of intraword code-switching or nonce borrowing in which the morphologies of two languages are combined in a single word. In Example 12 the English verb root 'happen' is used instead of the Irish equivalent *tarlaigh*. Insertional code-switching (Muysken et al., 2000) and borrowing are common in informal Irish. 74.71% of the tweets in the TwittIrish treebank were considered to be entirely in Irish, the remaining 25.29% of tweets being considered bi- or multilingual. Example 13 shows a section of an Irish tweet utilising the English word 'Dubs', a nickname for 'Dubliners', and Example 14 shows the use of an eclipse and an acute accent applied to the foreign proper noun 'Barcelona'.

(12) *Eachtra i ndiaidh* ***Happenáil***
'An event (is) after happening'

(13) *Roimh na* ***Dubs***
'Before the Dubs'

(14) *Tá sin i* ***mBarcelóna***
'That is in Barcelona'

Other nonstandard lexical forms Other unfamiliar terms may occur in the form of hypercorrection and neologisms. Hypercorrection occurs when an autocorrection system is either not activated or available in a user's language of choice. As a result, their attempts to type a word are corrected to a word with a similar spelling in another language. Example 15 shows the Irish word *coicíse* rendered as 'concise' probably due to automatic English spelling correction software. It is often difficult to distinguish between hypercorrection, neologisms, typos, or other spelling variations. Example 16 shows *agus* (and) rendered as *agua* which may have occurred due to automatic hypercorrection as 'agua' (water) is a frequent token in other languages such as Portuguese and Spanish. However, it could also be a simple typo.

(15) *Mhúscail mé i mo leaba féin ar maidin i ndiaidh* ***concise***
'I woke up in my own bed after a fortnight'

(16) *tá an teanga ag fáil bháis* ***agua***
'the language is dying and'

4.3 Syntactic Variation

Grammatical phenomena observed in Irish tweets are described in this section. As these idiosyncrasies occur at the phrasal rather than token level, they may directly affect the structure of the parse tree. Some phenomena, such as contraction and over-splitting, cause difficulty during the tokenisation stage, potentially having a negative downstream effect on parsing. Table 3 exemplifies syntactic variation in Irish tweets.

Contraction Much like abbreviation at the token level, contraction is defined here as the fusion of several tokens for the purpose of brevity, sometimes mimicking spoken pronunciation. Figure 2 shows the phrase *go bhfuil siad* 'that they are' reduced to *gowil siad* tokenised incorrectly. Figure 3 shows the same contraction tokenised correctly.

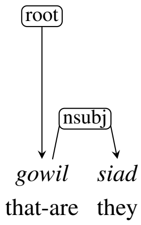

Figure 2: Incorrectly tokenised contraction 'that they are'.

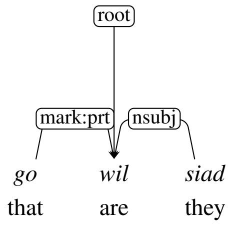

Figure 3: Correctly tokenised contraction 'that they are'.

Over-splitting The inclusion of extra white space within tokens is often observed in Irish tweets e.g. *Níl mé ró chinnte*. The prefix *ró-* ('too') is conventionally fused with the adjective it precedes in standardised text and so such tokens are annotated with the `goeswith` label as shown in Figure 4.

Phenomenon	Example	Standard form	Gloss
Contraction	*go **dtí'n***	*go dtí an*	'until the'
Over-splitting	***ana shuimiúil***	*an-suimiúil*	'very interesting'
Syntax-level code-switching	*Tá an **tweet machine** ró-tapa*	*Tá inneall na tvuíte ró-tapa*	'The tweet machine is too fast'
Dialectal grammar	*Ní **fhacthas***	*ní fhaca mé*	'I did not see'
Ellipsis	***jab iontach déanta aige***	*tá jab iontach déanta aige*	'he has done a wonderful job'
Meta language tags	***#sonas***	*sonas*	'happiness'
Non-sentential segmentation	***haha:) tá súil agam go raibh sé ann***	*ha ha! Tá súil agam go raibh sé ann.*	'haha:) I hope he was there.'
Other grammatical variation	***ce ata an athair?***	*cé hé an t-athair?*	'who is the father?'

Table 3: Examples of syntactic variation in Irish tweets.

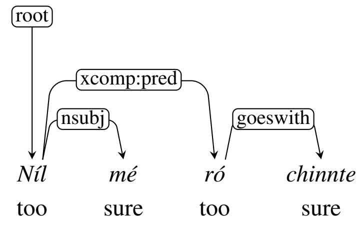

Figure 4: Over-splitting 'I am not too sure'.

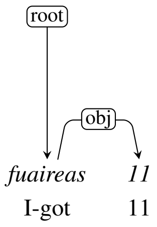

Figure 6: Synthetic verb form 'I got 11'.

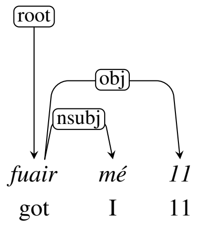

Figure 7: Analytic verb form 'I got 11'.

Syntax-level code-switching Alternational code-switching or congruent lexicalisation (Muysken et al., 2000) are likely to cause a change in the structure of the syntax tree, due to differing word orders of the languages involved, thus complicating the task of dependency parsing. In Irish, the adjectival modifier usually *follows* the noun it modifies whereas the inverse is true for English. Figure 5 exemplifies a case of congruent lexicalisation in which English adjective 'hippy-dippy' is positioned *before* an Irish noun rather than after as would be expected in 'classic' code-switching.

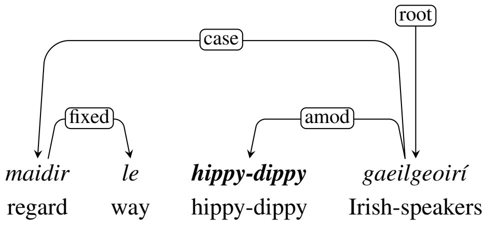

Figure 5: Congruent lexicalisation 'as for hippy-dippy Irish speakers'.

Dialectal grammar Figures 6 and 7 show semantically equivalent statements rendered using the synthetic, more common to the Munster dialect of Irish, and analytic verb forms respectively.

Ellipsis Example 17 shows a sentence fragment lacking a main verb. The probable inferred full phrase is *tá báisteach anseo* 'rain is here'.

(17) *báisteach anseo*
'rain here'

Meta language tags Hashtags are used in tweets to render a topic searchable and at-mentions are used to address or refer to another user. Either can play a syntactic role as exemplified in Figure 8.

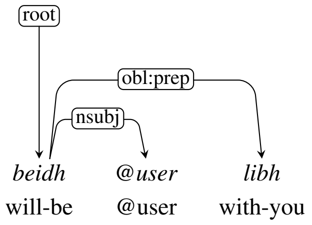

Figure 8: Syntactic meta language tag '@user will be with you'.

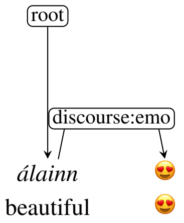

Figure 9: Non-sentential tweet using emoji in place of punctuation 'beautiful 😍'.

Non-sentential structure In tweets, the sentence is not an appropriate unit of segmentation as frequently non-standard punctuation, or none at all, is used. Figure 9 exemplifies a tweet utilising an emoji instead of punctuation.

Other grammatical variation Grammatical variation can also occur via unintentional deviation from conventional spelling or grammar by an L2 Irish speaker. Example 18 shows a grammatically incorrect phrase roughly translating to 'I have to *going'. In such cases, though the annotator may be able to infer the intended phrase *Caithfidh mé dul* 'I have to go', no corrections are made by the annotator to the surface form, however this information can be represented in the annotation via the label `CorrectForm` as described by Sanguinetti et al. (2022). Additionally, Irish tweets contain extremely unconventional constructions. This can occur in the form of unnatural phrases that have been machine-translated or generated by bots. Example 19 shows an ungrammatical construction that appears to have been translated automatically word by word. A more natural construction might be *conas tonna morgáiste a fháil* 'How to get a tonne of mortgage'. Some examples of this variety are easy to identify from surrounding context such as links to websites with similar content however, tweets may consist of text alone making it difficult to infer whether the author is human or machine.

(18) *Caithfidh mé* ***ag*** *dul*
'I have to *going'

(19) *Conas a Faigh tonna de Morgáiste*
'*How to get a tonne of mortgage'

5 Parsing Experiments

We compare the performance of two widely used neural dependency parsers on the TwittIrish test set, and examine the effect of using pre-trained contextualised word embeddings from a monolingual Irish BERT model (gaBERT). We report parsing performance broken down by sentence/tweet length, UPOS tags, and dependency labels and carry out a manual error analysis. Further information is detailed in Appendix B.

5.1 Parser Comparison

We experiment with two neural dependency parsing architectures: UDPipe (Straka et al., 2016), an NLP pipeline that includes a transition-based non-projective parser, and AllenNLP (Gardner et al., 2018), a biaffine dependency parser with a BiLSTM encoder (Dozat and Manning, 2017). Both systems are trained on IUDT version 2.8[11] and tested on the IUDT and TwittIrish test sets for comparison. Gold standard tokenisation is provided to the models which then predict UPOS tags and dependency relations. As the TwittIrish test set is the only gold annotated treebank of Irish UGC, no UGC is used as training or development data in these experiments. We opt to preserve it as a test set so that our results and results of future research in this area will be comparable.

System	LAS IUDT	LAS TwittIrish
UDPipe v1	70.58	47.33
AllenNLP	71.56	48.73
AllenNLP + gaBERT	**84.25**	**59.34**

Table 4: Comparison of parsing systems UDPipe v1, AllenNLP: Biaffine dependency parser (Dozat and Manning, 2017) with BiLSTM encoder, and AllenNLP + gaBERT: Biaffine dependency parser where BiLSTM is replaced with pretrained Irish BERT model (Barry et al., 2021). All were trained on IUDT version 2.8 and tested on the IUDT and TwittIrish test sets.

To leverage the substantial advances in accuracy achieved in dependency parsing by the use of pre-trained contexualised word representations (Che et al., 2018; Kondratyuk and Straka, 2019; Kulmizev et al., 2019), we use AllenNLP with token representations obtained from the last hidden layer of the gaBERT model (Barry et al., 2021) which are then passed to the biaffine parsing component.

Table 4 shows that, when tested on the IUDT version 2.8 test set, UDPipe achieves 70.58 labelled attachment score (LAS). In comparison, UDPipe achieves a much lower LAS of 47.33 on the TwittIrish test set. Similarly to UDPipe, AllenNLP achieves 71.56 LAS on the IUDT test set with a similar decrease of 22.83 points on the TwittIrish test set. The highest accuracy of 84.25 LAS is achieved by gaBERT with a difference of 24.91 points when tested on the TwittIrish test set. The lower accuracy obtained by parsers on the TwittIrish test set is unsurprising given the linguistic differences between the training and test sets. The 10+ LAS improvement provided by the gaBERT embeddings is seen in both test sets.

5.2 Analysis

Analysis was carried out on the AllenNLP parser with gaBERT embeddings using Dependable (Choi et al., 2015).

LAS by Number of Tokens per Sentence/Tweet The mean sentence length of the IUDT is 23.5 tokens, whereas the mean tweet length in TwittIrish is 17.8. Figure 10 shows that, when tested on the IUDT, parsing accuracy decreases as the length of the sentence increases. The highest accuracy of 87.92 LAS is associated with sentences of 10

[11] Models were trained with and without XPOS and feature annotation. The results shown here are without XPOS and features. The addition of XPOS and features constituted a difference of approximately +/-1 LAS.

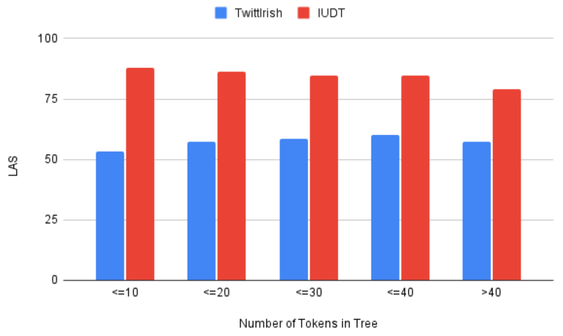

Figure 10: LAS broken down by number of tokens per tree achieved by AllenNLP Parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

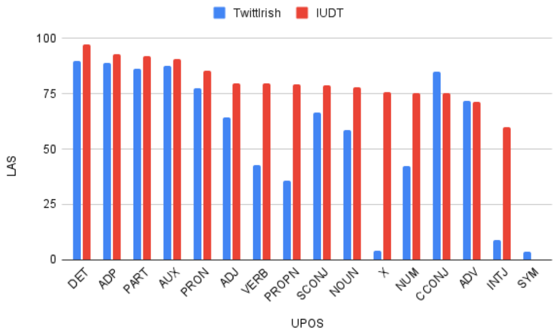

Figure 11: LAS broken down by UPOS tag achieved by AllenNLP Parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

tokens or fewer, and the lowest accuracy is observed in sentences of 40 tokens or more. This is an unsurprising trend as a higher number of tokens increases the probability of longer dependency distances and more complex constructions within a sentence. While the range of scores is smaller and trend less pronounced, the opposite effect is observed when the same parser is tested on TwittIrish, whereby LAS tends to increase as the length of the tweet increases. The highest LAS of 59.97 is associated with tweets of 31 to 40 tokens in length and the lowest accuracy of 53.47 LAS is associated with tweets of 10 tokens or less. This trend is also observed when gaBERT representations are not used, suggesting that, in this case, deep contextualised word embeddings do not cause this effect as observed in (Kulmizev et al., 2019). From manual inspection of the data, we observe that the genre-specific phenomena which challenge the parser such as ellipsis, meta language tags, and URLs, occur in higher proportions in shorter tweets, which would explain this trend.

LAS by UPOS and dependency relation We observe a larger proportion of PROPN, SYM, and PUNCT tags in Irish tweets in comparison to standardised Irish text, which contains a higher proportion of NOUN, DET, and ADP tags. This reflects the observations of Rehbein et al. (2019), who compare the distribution of POS tags in four German treebanks. Additionally, we compare the POS tag distribution in treebanks of English (Liu et al., 2018) and Italian (Sanguinetti et al., 2018) tweets to treebanks of standard text in those languages. We similarly observe that symbols, punctuation, and pronouns are more frequent in tweets and that nouns, determiners, and prepositions are more frequent in standard text for both languages.

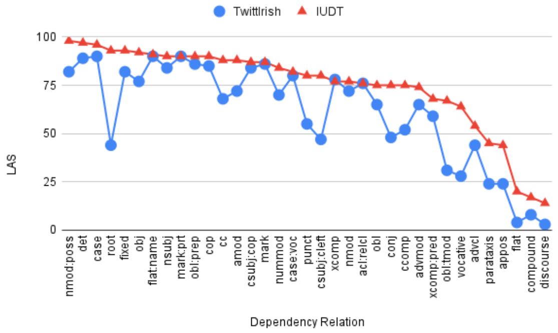

Figure 12: LAS broken down by dependency relation achieved by AllenNLP Parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

Figure 11 shows LAS associated with each UPOS tag when tested on the IUDT and TwittIrish. LAS is higher when tested on the IUDT for all UPOS tags except CCONJ, ADV, and SYM and in these cases the difference is small (<10 LAS). The most notable differences are X (71.6 LAS), INTJ (51.3 LAS), PROPN (43.5 LAS). These differences are due to 1) the divergent genres of the treebanks e.g. in the TwittIrish treebank the UPOS tag X is used for all non-syntactic hashtags, and PROPN is used for all at-mentions, neither of which occur in the IUDT and 2) differing annotation conventions e.g. in the IUDT, the tag X is used mostly for foreign-language tokens, whereas, in TwittIrish, due to the high proportion of English language tokens, non-Irish words are annotated with their true UPOS tag where the language is known to the annotator. The tag INTJ occurs very rarely in IUDT. However, due to the conversational nature of tweets, phatic expressions and emotional signifiers (not normally present in standard text) are frequent.

Our analysis of the dependency relation distribution of standard English, German, and Italian text compared to that of tweets in those languages reveals that the `parataxis`, `vocative`, and `advmod` relations are more frequent in tweets and that the `case`, `det`, and `nmod` relations are more frequent in standard text. We observe that this same effect is present in Irish tweets.

Figure 12 shows LAS broken down by dependency relation. The parser obtains higher scores on the IUDT for all dependency relations except `xcomp` for which it is just one point higher when tested on TwittIrish. The largest differences between the parsing performance on the two test sets are associated with the labels `root`, `vocative`, `obl:tmod`, `csubj:cleft`, `conj`, and `punct`. As regards `root` and `punct`, the difference in accuracy could be attributed to the non-sentential nature of tweets. In the IUDT each tree consists of a single sentence, whereas tweets may consist of sentence fragments or indeed several sentences, making root identification and establishing punctuation attachment more complex. `csubj:cleft` tends to be mislabelled in the absence of the copula which is often elided in standard text. This copula drop occurs even more frequently in tweets, negatively impacting on parsing accuracy. With regard to `conj`, both nonstandard forms of coordinating conjunctions (e.g. 'and', '+', misspellings etc.) and differing annotation styles between IUDT and TwittIrish lead to attachment errors. As regards `obl:tmod` and `vocative`, the respective differences in accuracy are due to the infrequent occurrences in the IUDT of a speaker or author directly addressing someone in the text and references to time (e.g. 5pm), both of which are common occurrences in tweets.

Error Analysis In order to assess the effect of the UGC phenomena present in Irish tweets, we analyse the most and least accurate parses as shown in Table 5. Seven tweets (76 tokens) were parsed with LAS between 0 and 5. On investigation, we observed fifteen occurrences of emojis that were most commonly incorrectly labelled `punct`. The ten English tokens were most commonly attached incorrectly via `flat:foreign`. The nine (two syntactic) usernames were most commonly mislabelled as `root`. There were five occurrences of ellipsis in the form of verb omission obfuscating the task of root selection. The three hashtags were most commonly mislabelled as `nmod` as were the three URLs. One occurrence of spelling variation in the form of diacritic omission caused the parser to misinterpret the token *ár* 'our' as *ar* 'on' meaning it was mislabelled as `case` instead of `nmod:poss` . Seven tweets (89 tokens) were parsed with an accuracy between 95 and 100 LAS. All of these were grammatical, well-formed sentences. There were three usernames and one hashtag all of which were syntactically integrated and so they were parsed correctly. There was one of insertional single-word code-switch which was accurately parsed. There were two occurrences of spelling variation, both in the form of diacritic omission but, as these do not resemble any other words, they were parsed correctly.

Phenomenon	Easiest Tweets	Hardest Tweets
Emoji	0	15
English Token	1	9
Username	3	10
Ellipsis	2	5
Hashtag	1	3
RT	0	3
URL	0	3
Spelling variation	2	2

Table 5: Number of occurrences of UGC phenomena where 'Easiest Tweets' refers to the 7 tweets that were parsed well with LAS between 95 and 100 and 'Hardest Tweets' refers to the 7 tweets (76 tokens) that were badly parsed with LAS between 0 and 5.

6 Conclusion

Presented in this paper is the novel resource, TwittIrish, the first Universal Dependencies treebank for Irish UGC. Analysis of this linguistic genre and anonymised examples of Irish tweets are presented. This research facilitates the development of NLP tools such as dependency parsers for Irish by providing a test set on which future Irish language technology can be tested. Future work will involve both further annotation and exploration of semi-supervised techniques.

Acknowledgements

We warmly thank the anonymous reviewers, as well as Steven Bird and Kevin Scannell for their valuable feedback. This work was funded by the Irish Government Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media under the GaelTech Project, and also supported by Science Foundation Ireland in the ADAPT Centre (Grant No. 13/RC/2106) at Dublin City University.

References

Fahad Albogamy and Allan Ramsay. 2017. Universal Dependencies for Arabic tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 46–51.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J Ó Meachair, and Jennifer Foster. 2021. gaBERT–an Irish language model. *arXiv preprint arXiv:2107.12930*.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396, Beijing, China. Association for Computational Linguistics.

Grzegorz Chrupala, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRO-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197.

CSO. 2016. Census of Population 2016 – Profile 10 Education, Skills and the Irish Language. Publisher: Central Statistics Office.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. # hardtoparse: POS tagging and parsing the twitterverse. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, pages 20–25.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Lauren Gawne and Gretchen McCulloch. 2019. Emoji as digital gestures. *Language@Internet*, 17(2).

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Heath. 2021. NO NEED TO YELL: A Prosodic Analysis of Writing in All Caps. *University of Pennsylvania Working Papers in Linguistics*, 27(1).

John Judge, Aoife Cahill, and Josef Van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 497–504.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A Dependency Parser for Tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano, and Joakim Nivre. 2019. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2755–2768, Hong Kong, China. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A Smith. 2018. Parsing tweets into universal dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards Universal Web Parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220. Uppsala University, Uppsala, Sweden.

Teresa Lynn and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Proceedings of the 2nd Celtic Language Technology Workshop*, Paris, France.

Teresa Lynn and Kevin Scannell. 2019. Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the 3rd Celtic Language Technology Workshop*.

Teresa Lynn, Kevin Scannell, and Eimear Maguire. 2015. Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8, Beijing, China. Association for Computational Linguistics.

Pieter Muysken, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *Bochumer Linguistische Arbeitsberichte*, page 13.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. WikiBERT models: Deep transfer learning for many languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ines Rehbein, Josef Ruppenhofer, and Bich-Ngoc Do. 2019. tweeDe – A Universal Dependencies treebank for German tweets. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 100–108, Paris, France. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French social media bank: A treebank of noisy user generated content. In *COLING 2012-24th International Conference on Computational Linguistics*.

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. 2012. Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7(18).

Natalia Silveira, Timothy Dozat, Marie Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 2897–2904. ELRA.

Nancy Stenson. 2019. *Modern Irish: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Taylor & Francis.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Rachael Tatman. 2015. #go awn: Sociophonetic Variation in Variant Spellings on Twitter. *Working Papers of the Linguistics Circle*, 25(2):97–108. Number: 2.

Rob Van Der Goot and Gertjan van Noord. 2018. Modeling Input Uncertainty in Neural Network Dependency Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4991.

Hongmin Wang, Yue Zhang, Guang Yong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies Parsing for Colloquial Singaporean English. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pages 1732–1744.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

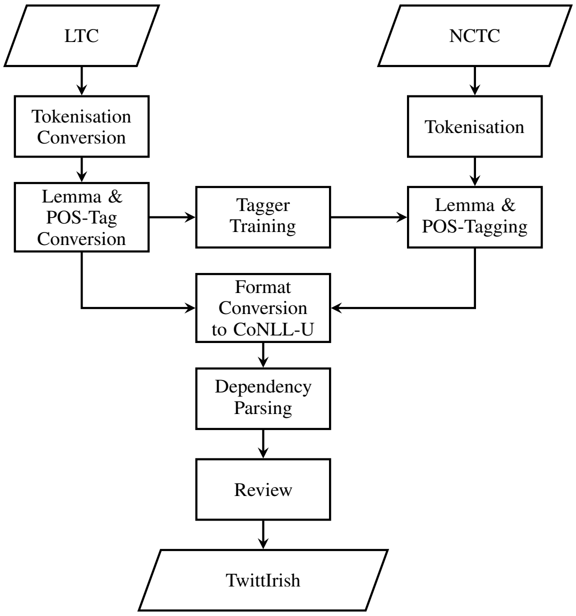

Figure 13: Diagram of the TwittIrish development process where LTC and NCTC refer to corpora of tweets.

A TwittIrish Development

Figure 13 outlines the stages of the TwittIrish treebank development.

A.1 LTC Tokenisation Conversion

The most notable difference in the tokenisation approach of LTC as compared to that of UD, was in the treatment of multi-word expressions (MWEs). In LTC, the individual tokens of MWEs are fused with an underscore whereas words with spaces are not allowed in UD. [12] Several minor differences were also observed between the two tokenisation schemes such as whether or not certain symbols, abbreviations, or punctuation marks should be attached to the token they follow or considered as a separate token. e.g. 5%, ama..., 1-0, 10pm. UD tends to favour the approach of separating such combinations[13] therefore we resolved to manually separate such occurrences in the TwittIrish tokenisation scheme.

A.2 LTC POS-tag Conversion

Table 6 shows the mapping of LTC POS to UPOS. LTC POS tags were automatically converted to the corresponding UPOS tag where a one-to-one or many-to-one mapping existed. In the case of one-to-many relationships, automatic identification and manual correction was performed. [14]

LTC POS	UPOS
N, VN	NOUN *
^, @	PROPN *
O	PRON
V	VERB, AUX †
A	ADJ
R	ADV
D	DET
P	ADP
T	PART
,	PUNCT
&	CCONJ, SCONJ †
$	NUM
!	INTJ
U, ~, E	SYM *
#, #MWE	X *
EN	any †
G	any †

Table 6: POS tag Mapping
* Many-to-one relation
† One-to-many relation

Surface	LTC POS	UPOS
@user	@	PROPN
#cutie	#	X
ca	R	ADV
bhfuil	V	VERB
an	D	DET
ghra	N	NOUN
you	EN	PRON
ask	EN	VERB

@user #cutie ca bhfuil an ghra you ask [15]
'@user #cutie where is the love you ask'

Table 7: Example Irish tweet with LTC and corresponding universal POS tags.

Table 7 demonstrates the mapping of a sample tweet from one scheme to the other. As all English language tokens were annotated with a single tag 'EN' in the LTC POS scheme, these tags were converted to the appropriate UPOS tag in the TwittIrish treebank.

Table 8 shows that, using the LTC POS tagset, all verbs are tagged `V`. According to UD, the Irish copula (e.g. *is*, *ní*) is tagged as `AUX` distinguishing it from other verbs (e.g. *tá*, *níl*) which are tagged `VERB`.

A.3 Preprocessing of newly-crawled tweets

Table 9 shows that hashtags and emoticons were not correctly handled by the UDPipe tokenizer trained

[12] https://universaldependencies.org/v2/mwe.html
[13] Not all treebanks apply this consistently.
[14] Both the Gimpel et al. (2011) and UD tagsets derived from the Google Universal POS tagset (Petrov et al., 2012)

Surface	LTC POS	UPOS
Ní	V	AUX
duine	N	NOUN
cáilúil	A	ADJ
é	O	PRON
ach	&	CCONJ
táim	V	VERB
bródúil	A	ADJ
#Grá	#	X

Ní duine cáiliúil é ach táim bródúil #Grá
'He is not a celebrity but I'm proud #Love '

Table 8: Example Irish tweet with LTC and corresponding universal POS tags.

UDPipe (IUDT)	NLTK Tweettokenizer
Dé	Dé
Céadaoin	Céadaoin
# Midweek	#Midweek
# Beagnachann	#Beagnachann
:)	:)
:)	:)

Dé Céadaoin #Midweek #Beagnachann :) :)
'Wednesday #Midweek #Almostthere :) :)'

Table 9: Example Irish tweet with UDPipe and NLTK tokenization

on the IUDT. Despite being trained on Irish data, Twitter-specific features such as meta language tags are not present in its training data.

A.4 Conversion to CoNLL-U format

Table 10 shows that the Morfette format is a subset of the CoNLL-U format used by UDPipe. The LTC and NCTC (newly-crawled tweets) were thus converted automatically from the 3-column Morfette format, consisting of the token, lemma, and POS-tag to the 10-column CoNLL-U format. CoNLL-U enables additional token-level annotation i.e. a token id, language-specific part-of-speech tags (XPOS), morphological features, the head of the current word, the dependency relation, an enhanced dependency graph in the form of a list of head-deprel pairs, and any other miscellaneous annotation.[16] CoNLL-U also requires a sentence ID and the original raw text to be included preceding the annotation. Further, in the miscellaneous column, the label 'SpaceAfter=No' encodes information about which tokens have a space after them in the original text for detokenisation purposes enabling automatic conversion from raw text to tree and vice versa.

A.5 Review

In order to assess the accuracy of the dependency annotation, a subset of the annotated data, consisting of 46 trees (773 tokens), was reviewed for errors by another Irish speaker trained in linguistic annotation. The task of the reviewer was to flag possible errors in the form of a token with an incorrect head and/or label. 46 possible errors were identified by the reviewer. The possible errors were then discussed by a team of two expert annotators to confirm whether the possible errors were true errors. 32 possible errors were confirmed as true errors. The overall accuracy of the treebank annotation can be estimated as 95.86% by dividing the number of correctly annotated tokens by the total number of tokens in the review. 16 tokens (2.07% of all tokens in the review) had an incorrect label and correct head. 12 tokens (1.55% of all tokens in the review) had an incorrect head and correct label. The most common error (5 instances) was incorrect punctuation attachment. Only 4 tokens (0.52%) were identified as having both an incorrect head and label. Figure 14 shows the phrase *maith sibh* ('good on you') incorrectly annotated with *sibh* as the `root` and *maith* as its `adjectival modifier`. It was identified in the review that *maith* should be considered the adjective predicate of an elided copula (Stenson, 2019). The full phrase is thought to be *is maith sibh* and the corrected annotation is shown in Figure 15.

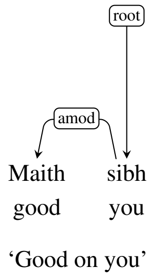

Figure 14: Example of tweet with incorrect head and label.

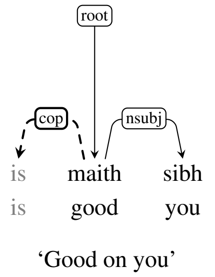

Figure 15: Reviewed tweet with corrected head and label.

[16] In order to make optimum use of the time spent by the annotator, language-specific part-of-speech tags, morphological features, and enhanced dependency annotation were not included in this version of the TwittIrish dataset. These elements can be automatically added in later versions of the treebank.

CoNLL-U	Morfette			CoNLL-U					
ID	**FORM**	**LEMMA**	**UPOS**	**XPOS**	**FEATS**	**HEAD**	**DEPREL**	**DEPS**	**MISC**
1	Cuirfidh	cuir	VERB	_	_	0	root	_	_
2	mé	mé	PRON	_	_	1	nsubj	_	_
3	DM	DM	NOUN	_	_	1	obj	_	_
4	chuici	chuig	ADP	_	_	1	obl:prep	_	_

‘Cuirfidh mé DM chuici’
‘I will send her a DM’

Table 10: Example conversion of Irish tweet from Morfette to CoNLL-U format

B Parsing Experiments

B.1 Parser Hyperparameters

Biaffine Parser Details	
AllenNLP	
Word embedding	100
Character embedding	32
Char-BiLSTM layers	3
Char-BiLSTM size	64
BiLSTM layers	2
BiLSTM size	200
AllenNLP + gaBERT	
BERT word-piece embedding size	768
BERT word-piece type	average
Parser	
Arc MLP size	500
Label MLP size	100
Dropout LSTMs	0.33
Dropout MLP	0.33
Dropout embeddings	0.33
Nonlinear act. (MLP)	ELU
Optimiser and Training Details	
Optimizer	AdamW
Learning rate	3e-4
beta1	0.9
beta2	0.999
Num. epochs	50
Patience	10
Batch size	16

Table 11: Chosen hyperparameters for the AllenNLP and the AllenNLP + gaBERT parsers. In the AllenNLP parser, a character- and word-level BiLSTM is used. In the gaBERT variation, these components are replaced by the Transformer model. The parsing module and training setup is the same for both parsers.

LAS	**TwittIrish High**	**TwittIrish Low**
IUDT High	DET, ADP, PART, AUX, PRON, SCONJ	VERB, PROPN, PUNCT, X, INTJ
IUDT Low	ADJ, CCONJ, ADV	NOUN, NUM, SYM

Table 12: Confusion matrix of LAS by UPOS tag achieved by AllenNLP Parser with gaBERT embeddings on the IUDT and TwittIrish test sets

B.2 LAS by UPOS

Table 12 shows which UPOS tags are associated with higher or lower than average LAS in both test sets. High accuracy is correlated with tokens which occur frequently and have low variation.[17]

UPOS tags DET, ADP, PART, AUX, PRON, and SCONJ are associated with higher than average LAS in both the TwittIrish and IUDT test sets. In the IUDT, a high proportion, 8.87%, of tokens have the UPOS tag DET. As is common with function words, DET comprises of a closed set of lemmata and thus has the low variation of 0.21%.

The tags ADJ, CCONJ, and ADV are associated with higher than average LAS in the TwittIrish test set but lower than average LAS in the IUDT. This might be because these tags are more likely to be involved in more complex, ambiguous, or long-distance attachments.

The tags VERB, PROPN, PUNCT, X, and INTJ are associated with higher than average LAS in the IUDT test set but lower than average LAS in TwittIrish. In the case of VERB and PUNCT, this can be attributed to the non-sentential nature of tweets.

UPOS tags NOUN, NUM, and SYM are associated with lower than average LAS in both the TwittIrish and IUDT test sets. In the IUDT, a low proportion, 0.02%, of tokens have the UPOS tag SYM. The variation is high (83.33%).

[17] Variation is calculated by dividing the number of occurrences by then number of unique lemmata

B.3 LAS by Dependency Relation

LAS	TwittIrish High	TwittIrish Low
IUDT High	nmod:poss, det, case, fixed, obj, flat:name, nsubj, mark:prt, obl:prep, cop, cc, amod, csubj:cop, mark, nummod, case:voc	root, csubj:cleft, punct
IUDT Low	xcomp:pred, advmod, obl, acl:relcl, nmod, xcomp	discourse, compound, flat, appos, parataxis, advcl, vocative, obl:tmod, ccomp, conj

Table 13: Confusion matrix of LAS by dependency label achieved by AllenNLP Parser with gaBERT embeddings on the IUDT and TwittIrish test sets

Table 13 shows that high accuracy is associated with dependency relations nmod:poss, det, case, fixed, obj, flat:name, nsubj, mark:prt, obl:prep, cop, cc, amod, csubj:cop, mark, nummod, case:voc in both the IUDT and TwittIrish. root, csubj:cleft and, punct are associated with higher than average LAS in the IUDT test set but lower than average in the TwittIrish set. xcomp:pred, advmod, obl, acl:relcl, nmod, and xcomp are associated with higher than average LAS in the TwittIrish test set but lower than average LAS in the IUDT. discourse, compound, flat, appos, parataxis, advcl, vocative, obl:tmod, ccomp, and conj are associated with lower than average LAS in both the TwittIrish and IUDT test sets.