

Linguistic Analysis and Automatic Dependency Parsing of Tweets in Modern Irish

Lauren Cassidy

B.A.

A thesis submitted for the award of Doctor of Philosophy (PhD)

Dublin City University

School of Computing

Supervisors:

Dr. Jennifer Foster,

Dublin City University

Dr. Teresa Lynn,

Mohamed bin Zayed University of Artificial Intelligence

December, 2023

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Lauren Cassidy

A handwritten signature in black ink, appearing to read 'Lauren Cassidy', written in a cursive style.

ID No.: 19215029

Date: Thursday 21st December, 2023

Acknowledgements

Firstly, I thank my supervisors Jennifer Foster and Teresa Lynn. I feel extremely lucky to have, not one, but two fantastic mentors who have been so present and supportive throughout my PhD candidature. I am grateful for all of our conversations that have inspired and challenged me throughout this time. I also thank Kevin Scannell for providing the data for the TwittIrish treebank and sharing helpful feedback throughout the PhD. Thanks to Monica Ward and Brian Davis for examining my PhD Transfer and providing feedback and suggestions that inspired the questionnaire study. I am grateful to James Barry for the opportunity to collaborate on gaBERT and for sharing his dependency parsing knowledge. Thanks to Abigail Walsh for our interesting conversations and for all her helpful advice over the years. I am very grateful for the wonderful community and great friends I have made in the ADAPT lab and at Dublin City University. I extend my gratitude to all my co-authors and collaborators. It has been a pleasure to work with experts in many different areas and languages. I am grateful for the opportunities to attend and present my research at conferences such as ACL and LREC in 2022 where I got to learn about new developments in the field and connect with new and old friends. Thanks to Ailbhe Ní Chasaide who first encouraged me to study Computer Science and Language and then to all of my lecturers at Trinity who helped me realise how fascinating and exciting a topic it is. I thank all the researchers I have cited in this thesis. Thanks to everyone who took part in and helped to distribute the questionnaire study. I am grateful to Francis Tyers and Brian Davis for examining this thesis and to Irina Tal for chairing the viva. I also extend my gratitude to the Irish Government Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, who funded my research under the GaelTech Project at DCU.

I am so grateful to my family, Neil, Noreen, Féaron, and Kevin for their endless encouragement - and proofreading services! Thanks to all the Walshes for always being there for each other and all the craic at the Sunday gatherings in the Castle. I thank my grandmother, Nuala, who first inspired me to learn Irish and who spent countless joyful afternoons with me reading, chatting, and consulting the giant dictionary. I thank my friends for the love, music, laughs, and adventures. A special thanks to Inés and Fiona (the worms), Jill, Aidan, and Simon (the band), Megan and Cathy (Port Láirge), Aoife and Kev, Rosaleen and Cathleen, the Book Club, the Running Pals, the Spinsterz, and all my various housemates throughout this time. Finally, thanks to Tássia for the best four years.

Contents

1	Introduction	1
1.1	Thesis Topic	1
1.2	Research Questions	5
1.3	Thesis Structure	7
1.4	Publications	8
2	Background	11
2.1	The Irish Language	11
2.1.1	History	12
2.1.2	The Current Status of Irish	13
2.1.3	Linguistic Features of Irish	14
2.1.4	Irish-Language Technology	19
2.2	Dependency Parsing	20
2.2.1	Dependency Grammar	20
2.2.2	Dependency Trees	22
2.2.3	Treebanks	24
2.2.4	Universal Dependencies	25
2.2.5	Dependency Parsing Frameworks	26
2.2.6	Neural Networks in Dependency Parsing	30
2.3	User-generated Content	32
2.3.1	The Rise of User-generated Content	32
2.3.2	Processing User-generated Content	33
2.4	Language Contact	34
2.4.1	Terminology	34
2.4.2	Typologies and Frameworks	36
2.4.3	NLP for Code-switched Data	38
2.4.4	Code-Switching versus Borrowing	39
2.4.5	Language Contact in Irish	39
2.5	Research Gaps	41
2.6	Summary	42
3	TwittIrish Treebank Development	44
3.1	Data Curation	44
3.2	Data Preprocessing and Conversion	46
3.2.1	LTC Tokenisation Conversion	47
3.2.2	LTC Lemmatisation and POS Tag Conversion	47
3.2.3	NTC Tokenisation	49
3.2.4	NTC Lemmatisation and POS Tagging	50
3.2.5	Conversion to CoNLL-U Format	51
3.3	Syntactic Annotation	51
3.3.1	Annotation of Genre-specific Features	52

3.3.2	Bootstrap Annotation Cycle	53
3.4	Quality Review	54
3.5	Data Releases	56
3.6	Summary	57
4	Linguistic Analysis of Irish Language Tweets	58
4.1	Orthographic Variation in Irish Tweets	59
4.2	Morphological Variation in Irish Tweets	62
4.3	Lexical Variation in Irish Tweets	64
4.4	Syntactic Variation in Irish Tweets	67
4.5	Summary	70
5	Dependency Parsing Experiments	71
5.1	Experiment Setup	71
5.2	Establishing a Baseline	75
5.2.1	Baseline Parsing Results	75
5.2.2	Analysis of Baseline Results	76
5.3	Improving the Parser	83
5.3.1	Improved Results	83
5.3.2	Preliminary Analysis of Improved Results	84
5.4	Summary	87
6	Language Contact Questionnaire Study	89
6.1	Related Work	90
6.2	Questionnaire Design and Development	92
6.2.1	Design Choices	93
6.2.2	Language Contact Questions	95
6.3	Pilot Test	99
6.4	Data Collection	100
6.5	Results and Analysis	101
6.5.1	Demographic Distributions	101
6.5.2	Language Classification of BOR and CS Words	105
6.5.3	Usability of BOR and CS Words	106
6.5.4	Reflexive Thematic Analysis of Open-Ended Responses	107
6.5.5	Limitations	115
6.5.6	Summary	117
7	Conclusion	119
7.1	Contributions	119
7.2	Research Questions Revisited	120
7.3	Limitations	122
7.4	Future Work	123
7.5	Concluding Remarks	125
	Appendix	A1
A	TwittIrish Data Statement	A1
A.1	Header	A1
A.2	Executive Summary	A1
A.3	Curation Rationale	A1
A.4	Documentation for Source Data Sets	A2
A.5	Language Varieties	A2

A.6	Speaker Demographic	A2
A.7	Annotator Demographic	A3
A.8	Speech Situation and Text Characteristics	A3
A.9	Preprocessing and Data Formatting	A3
A.10	Capture Quality and Limitations	A4
A.11	Metadata	A4
B	TwittIrish Annotation Guidelines	B1
B.1	Segmentation	B1
B.2	Tokenisation	B1
B.3	Lemmatisation	B3
B.4	POS-tagging	B4
B.5	Dependency Relations	B6
B.6	Language Identification	B15
C	Ethical Approval for Language Contact Questionnaire Study	C1
D	Full Text of Language Contact Questionnaire Study	D1
E	Full Language Classification Results	E1
F	Full Phase 1 Codebook	F1

Acronyms

- API** application programming interface. 3, 122
- ARCOSG** Annotated Reference Corpus of Scottish Gaelic. 71, 84, 88
- BERT** bidirectional encoder representations from transformers. 6, 9, 31, 32, 53, 72, 73, 120, 125
- BiLSTM** bidirectional long short-term memory. 30, 31, 38, 72, 74–76, 83, 84, 87, 88
- CE** Common Era. 12, 21
- CMC** computer-mediated communication. 14, 32
- CoNLL-U** Conference on Computational Natural Language Learning (CoNLL) format for Universal Dependencies (U). iii, viii, 45, 50–52, B15
- ELMo** embeddings from language models. 31
- EM** exact match. 74
- IDT** Irish Dependency Treebank. 19
- IFST** Irish finite state tools. 19
- IUDT** Irish Universal Dependencies Treebank. viii, xi, 3–6, 19, 46, 47, 49, 50, 53, 58, 64, 71, 72, 75–81, 84–88
- LAS** labelled attachment score. viii, xi, 74–88, 120
- LOLI** lone other-language item. 6, 65
- LSTM** long short-term memory. 19, 30
- LTC** Lynn Twitter Corpus. iii, x, 20, 45–51, A4
- MLP** multilayer perceptron. 30, 73
- MST** maximum spanning tree. x, 27–29
- MWE** multiword expression. 17, 19, 32, 47, 86, A3, B13
- NEID** New English-Irish Dictionary. 53, 66, 90, 91, 98, 106, 117, B15
- NLP** natural language processing. 1–7, 11, 15, 17, 19, 20, 22, 24–26, 31–34, 37–39, 41, 42, 44, 51, 57–59, 70, 88–90, 117–119, 121–124, 126, 127, A1
- NLTK** Natural Language Toolkit. viii, 49, 50

NTC New Twitter Corpus. iii, x, 45, 46, 49–51, A2

POS Part-of-speech. iii, viii, 3, 4, 15, 16, 20, 25, 30, 38, 41, 45–51, 54, 71, 72, 77, 79, A2, B3

RNN recurrent neural network. 30, 38

RT retweet. 48

SVO subject-verb-object. 22

UAS unlabelled attachment score. 74–76, 84

UD Universal Dependencies. iii, viii, x, 2–5, 7–9, 11, 19, 20, 22, 25, 26, 34, 41, 42, 44, 46–49, 52, 54, 56, 57, 71, 75, 76, 79, 83, 84, 86, 119, 124, A3, B12

UGC user-generated content. viii, 3–7, 9, 11, 20, 32–34, 38, 41, 42, 44, 47, 49, 51, 52, 57, 60, 81–83, 85, 88, 98, 119–124, 126, A1

UPOS universal part-of-speech. ix, xi, 2, 6, 25, 26, 48, 50, 51, 77, 78, 85–87, B4

URL uniform resource locator. 17, 49, 88, B3

VSO verb-subject-object. 15, 17, 22

List of Tables

2.1	Inflections in Irish. Adapted from Uí Dhonnchadha (2002).	16
3.1	Metadata of TwittIrish source data.	45
3.2	Dataset sizes.	46
3.3	POS tag mapping. * Many-to-one relation † One-to-many relation	48
3.4	Example Irish tweet with LTC and corresponding universal POS tags.	49
3.5	Example Irish tweet with LTC and corresponding universal POS tags.	49
3.6	Example Irish tweet tokenised by UDPipe 1 trained on IUDT version 2.8 and NLTK TweetTokenizer.	50
3.7	Example conversion of Irish tweet from Morfette to CoNLL-U format ‘I will send her a DM’.	50
5.1	Treebank training sets used in parsing experiments	72
5.2	Chosen hyperparameters for the multitask parser and tagger (adapted from Barry et al. (2022)).	74
5.3	Baseline parsing results, median score over five random seed values using the biaffine parser of Dozat and Manning (2017), trained on the IUDT version 2.12.	75
5.4	Parsing results used for analysis. Biaffine w/ gaBERT refers to the biaffine dependency parser of Dozat and Manning (2017) with gaBERT encodings (Barry et al., 2022). The parser was trained on the IUDT version 2.8 and tested on the IUDT and TwittIrish test sets.	76
5.5	Confusion matrix of LAS by UPOS tag achieved by AllenNLP Biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.	79
5.6	Confusion matrix of LAS by dependency label achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.	80
5.7	Number of occurrences of UGC phenomena where easiest tweets refers to the 7 tweets that were parsed with LAS between 95 and 100 and hardest tweets refers to the 7 tweets (76 tokens) that were parsed with LAS between 0 and 5.	83
5.8	Full dependency parsing results using biaffine parser of Dozat and Manning (2017) on the TwittIrish test set using training data from Irish and Scottish Gaelic treebanks of UD version 2.12.	84
6.1	Borrowing criteria (Álvarez-Mellado and Lignos, 2022) adapted for Irish.	90
6.2	Specific research questions investigated in questionnaire study.	92
6.3	Demographic and language background variables.	95
6.4	Borrowing extracts from questionnaire study.	96
6.5	Code-switching extracts from questionnaire study.	97
6.6	Irish extracts from questionnaire study.	97
6.7	Ambiguous extracts from questionnaire study.	98
6.8	Response options for Part C questions considered in pilot study.	99

6.9	Phase 1 codes with the corresponding number of references, a definition and example (abbreviated here, full version in Appendix F).	109
6.10	Phase 4 candidate themes and descriptions.	111
6.11	Final themes and descriptions	113
B.1	UPOS tags with descriptions and Irish language examples.	B5
E.1	Language classification results from the language contact questionnaire study for GA category.	E1
E.2	Language classification results from the language contact questionnaire study for CS category.	E2
E.3	Language classification results from the language contact questionnaire study for BOR category.	E2
E.4	Language classification results from the language contact questionnaire study for AMBI category.	E3
F.1	Full phase 1 codebook.	F3

List of Figures

1.1	UD representation of an Irish sentence <i>Dheisigh Seán an rothar</i> ‘Seán fixed the bicycle’.	2
1.2	UD representation of English sentence ‘Seán fixed the bicycle’.	2
2.1	Relationship between Celtic languages. Image from Ó Siadhail (1989).	12
2.2	Phrase Structure tree ‘Mia saved a dog’.	21
2.3	Dependency tree ‘Mia saved a dog’.	21
2.4	Tokenisation of sentence <i>Fásfaidh crann mór</i> ‘A big tree will grow’.	23
2.5	Dependency tree ‘A big tree will grow’.	23
2.6	Non-projective sentence ‘I have friends now with an interest in computer games’.	24
2.7	Fully-connected, weighted dependency graph <i>Shábháil Mia madra</i> ‘Mia saved a dog’.	28
2.8	Chu-Liu-Edmonds algorithm for finding the MST of a weighted directed graph. Adapted from Jurafsky and Martin (2023).	29
3.1	The TwittIrish creation process. The corpora LTC and NTC are the sources of the treebank data.	45
3.2	Attachment of sentences within tweets via <code>parataxis:sentence</code> ‘@user is doing very well. The wind is very strong’.	52
3.3	Bootstrapping approach to semi-automated syntax annotation.	53
3.4	Parsed tweet with incorrect label and correct head corrected during review.	55
3.5	Incorrectly annotated tweet and corrected version.	56
4.1	Synthetic and analytic verb forms ‘I got 11’.	63
4.2	Username in a syntactic role ‘@user will be with you’.	67
4.3	Syntactic hashtag ‘Filming a new comedy series for #tg4’.	67
4.4	Contraction ‘I know’.	67
4.5	Over-splitting ‘I am not too sure’.	68
4.6	Code switching ‘Thank you for the follow’.	68
4.7	Code switching ‘as for hippy-dippy Irish speakers’.	68
4.8	Ellipsis ‘rain (is) here’.	69
4.9	Non-sentential tweet using emoji as punctuation ‘Bye :-) I hope you have a nice sleep’.	69
4.10	Grammatical variation ‘How to get a tonne of a mortgage’.	70
5.1	Biaffine dependency parser with BiLSTM encoder. Figure taken from Dozat and Manning (2017).	72
5.2	Biaffine dependency parser with gaBERT embeddings. Figure adapted from Dozat and Manning (2017).	73
5.3	Reference and system parses ‘Mia saved a dog’, resulting in an LAS of 1/3 and an UAS of 3/3.	75

5.4	LAS by number of tokens per tweet achieved by biaffine parser with gaBERT embeddings on the TwittIrish and IUDT test sets.	77
5.5	LAS by UPOS tag achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.	78
5.6	LAS achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets by dependency relation.	80
5.7	Correct and incorrect parse for phrase <i>míle maith agat</i> ‘[thanks] a million’.	81
5.8	Correct and incorrect parse for phrase <i>@user Blasta</i> ‘@user Tasty’.	82
5.9	Dependency parsing results showing median LAS over five random seed values on TwittIrish test set varying training data and encoder of biaffine parser (Dozat and Manning, 2017).	85
5.10	LAS by number of tokens per tweet achieved by biaffine parser with gaBERT embeddings trained on TwittIrish and IUDT version 2.12	86
5.11	Comparison of LAS by UPOS tag of our best parser against a reference. Our best parser is trained on TwittIrish and IUDT treebanks and tested on TwittIrish. The reference parser is trained and tested on the IUDT treebank.	87
5.12	Comparison of LAS by dependency relation of our best parser against a reference. Our best parser is trained on TwittIrish and IUDT treebanks and tested on TwittIrish. The reference parser is trained and tested on the IUDT treebank.	88
6.1	Diagram of the questionnaire study process.	90
6.2	Irish-language Wikipedia article with the headword ‘Twerking’.	92
6.3	Decision tree for classifying words as GA (Irish), CS (English code-switch), BOR (borrowing), or AMBI (ambiguous).	96
6.4	Count of participants in each age group.	101
6.5	Respondents’ self-reported level of Irish.	102
6.6	Participants’ level of Irish broken down by age group.	102
6.7	Respondents’ feelings about mixing Irish and English.	104
6.8	Respondents’ feelings about mixing Irish and English broken down by Irish proficiency. A score of 5 corresponds to ‘very positive’ and 1 corresponds to very negative.	105
6.9	Percentage of responses classifying BOR and CS words as English. The dark blue line shows the average BOR value: 52.3%. The light blue line shows the average CS value: 85.55%.	106
6.10	Mean ‘likelihood of use’ score for BOR and CS words where 5 is ‘very likely’ and 1 is ‘very unlikely’. The dark blue line represents the average score for BOR words. The light blue line represents the average score for CS words.	107
6.11	Overall mean ‘likelihood of use’ scores for CS (code-switched), BOR (borrowed), AMBI (ambiguous), and GA (Irish) words.	108
6.12	Map of candidate theme ‘Person’.	109
6.13	Map of candidate theme ‘Word’.	110
6.14	Map of candidate theme ‘Context’.	110
B.1	Verbal root with the nominal subject ‘he went’.	B6
B.2	Copular construction ‘it is good news’.	B6
B.3	Direct object of a finite verb ‘I got this’.	B6
B.4	Infinitive verb construction ‘to get information’.	B7
B.5	Question particle ‘Do you listen to audiobooks?’	B7
B.6	Clausal subject of a cleft construction ‘it’s there that it stays’.	B7
B.7	Clausal subject of a copular clause ‘maybe it is’.	B8
B.8	Clausal complement of a verb ‘I think [it] was’.	B8

B.9	Open clausal complement of a verb ‘She is working’.	B8
B.10	Predicate of the substantive verb ‘I was sick’.	B8
B.11	Oblique argument ‘I speak with people’.	B9
B.12	Personal preposition as oblique argument ‘I speak with them’.	B9
B.13	Dialogue participant directly addressed ‘any news, friend?’.	B9
B.14	Vocative mention ‘Thank you @user’.	B9
B.15	Adverbial modifier ‘Do it now’.	B10
B.16	Adverbial clause modifier ‘good if I say so myself’.	B10
B.17	Discourse marker ‘Oh don’t worry’.	B10
B.18	Pictogram ‘beautiful 🍷’.	B11
B.19	Nominal modifier ‘young musician of the year’.	B11
B.20	Appositional modifier ‘Did you enjoy the film #PerfectPitch’.	B11
B.21	Numeric modifier ‘Two days left’.	B11
B.22	Adjectival modifier ‘An interesting report’.	B12
B.23	Determiner ‘Praise the youth’.	B12
B.24	Case ‘in the office today’.	B12
B.25	Relative clause ‘A couple that doesn’t get along’.	B13
B.26	Coordinating conjunction ‘take it easy and take care’.	B13
B.27	Fixed ‘in a thousand years’.	B13
B.28	Flat ‘on the 9th February’.	B14
B.29	Compound ‘this weekend’.	B14
B.30	Parataxis and punctuation ‘Wonderful - good luck to him’.	B14
C.1	Dublin City University Ethical Approval for Questionnaire Study.	C1

Linguistic Analysis and Automatic Dependency Parsing of Tweets in Modern Irish

Lauren Cassidy

Abstract

Automatic syntactic parsing of user-generated content in Modern Irish poses significant challenges due to the language’s minority status and limited linguistic resources. In this thesis, we present TwittIrish, the first Universal Dependencies treebank of tweets in Irish, a linguistically-informed, genre-specific dataset developed via a cycle of automatic syntactic annotation and manual correction. We use this novel resource to document and quantify the linguistic differences between Irish tweets and standardised Irish text with regard to orthography, morphology, lexicon, and syntax. We provide examples of linguistic features observed in the tweets and describe how we have chosen to represent them within the Universal Dependencies framework. Furthermore, utilise the TwittIrish dataset to establish baseline parsing results and explore methods to increase parsing accuracy. We show that the use of monolingual Irish BERT embeddings provides a significant improvement over baseline results. Our error analysis shows that language contact phenomena constitute one of the greatest challenges associated with processing informal Irish text. We, therefore, extend our analysis of user-generated content to examine language contact in Irish-language tweets. Due to centuries of contact with English, code-switching, borrowing, and other language contact phenomena are frequent in informal Irish. We investigate the perceptions of Irish speakers with regard to language contact in the Irish-English language pair. Furthermore, we assess the advantages and disadvantages of distinguishing between code-switching and borrowing in the context of resource development for natural language processing. Our research contributes to language technology support for a low-resource language by providing a novel data set and facilitating more accurate dependency parsing of informal Irish. Additionally, the exploration of linguistic features of Irish-language tweets extends the impact of this research to linguistics, sociolinguistics, and the Irish-language community more broadly by enhancing the general understanding of the use of Irish on social media.

Chapter 1

Introduction

1.1 Thesis Topic

Irish is a low-resource language spoken mostly in small communities in Ireland called *Gaeltachtaí*. According to the most recent census, around 1.9 million people claim the ability to speak Irish but just around 72,000 (3.8% of speakers) use the language daily outside of the education system (CSO, 2022). Throughout history, Irish has been influenced by Latin, Old Norse, and English, experiencing periods of decline and revival. Despite its rich linguistic heritage and current official status in Ireland, Irish now faces many of the challenges of an endangered language and is at risk of digital extinction. To mitigate this risk, the development of language technology for Irish offers a promising avenue. Language technology has myriad applications in areas such as education, speech technology, social media integration, language learning, communication, and cultural preservation. In this way, language technology development for Irish can make the language more useful, relevant, and accessible for speakers and learners, by enabling the use of modern technology through Irish.

A fundamental natural language processing (NLP) task is **dependency parsing**, a type of syntactic analysis. Dependency parsing involves establishing grammatical relationships between words in a sentence, representing their syntactic structure as **dependency trees**. The dependency relations, illustrated by directed, labelled arcs from heads to dependents, can act as a loose approximation to the semantic connections between predicates and their arguments. The significance of dependency parsing extends beyond its inherent linguistic value. **Treebanks**, collections of parsed sentences annotated with syntactic de-

dependencies, have played an important role in advancing language technology. Treebanks provide annotated data that can be used to train and evaluate parsers, facilitating the improvement of NLP systems. For example, dependency representations have been leveraged to boost performance in tasks such as machine translation (Chen et al., 2017) and semantic role labelling (Strubell et al., 2018). Treebanks serve as rich datasets for various applications such as language-learning tools, lexicography, and linguistic research. Despite the advancements in end-to-end systems and the increasing prevalence of large pre-trained language models, treebanks remain essential for specific tasks in NLP. While end-to-end systems have shown remarkable capabilities, they may lack the nuanced understanding of syntactic intricacies that treebanks provide, especially in the low-resource context. Therefore, treebanks continue to be valuable assets in refining and validating the performance of language technology.

Universal Dependencies (UD) (Nivre et al., 2020) is a unified treebank annotation scheme that promotes consistency and interoperability across different languages, facilitating multilingual research and resource sharing. UD provides universal part-of-speech (UPOS) tags, dependency relations, and standard guidelines intended to be applicable to all languages. Being an open-source project, UD facilitates collaboration among contributors internationally. As of version 2.12, UD includes 245 treebanks in over 141 languages, many with detailed, language-specific annotation guidelines and examples.

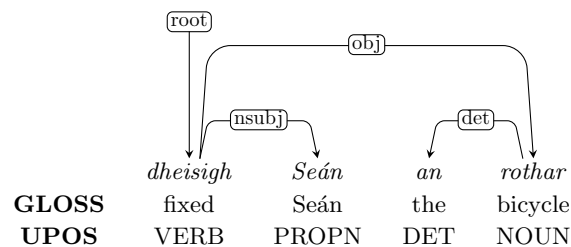


Figure 1.1: UD representation of an Irish sentence *Dheisigh Seán an rothar* ‘Seán fixed the bicycle’.

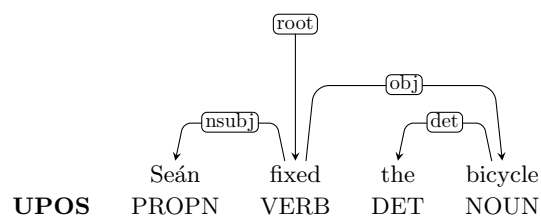


Figure 1.2: UD representation of English sentence ‘Seán fixed the bicycle’.

Figure 1.1 exemplifies a dependency tree from the Irish Universal Dependencies Treebank (IUDT) (Lynn and Foster, 2016). The root of the structure is assigned to the predicate or main verb of the sentence and the nominal subject and object arguments of the verb are attached via dependency labels that describe their relationship. Further, non-core elements such as determiners modify the core elements of the structure. Figure 1.2 shows the English translation of the phrase modelled in Figure 1.1. We observe that although the word order differs between Irish and English, the parts of speech (VERB, PROP, DET, NOUN) and relations (root, nsubj, obj, det) between them are represented similarly within UD, demonstrating the framework’s potential in multilingual applications. A more detailed explanation of dependency parsing is provided in Section 2.2 and a description of each Part-of-speech (POS) tag and dependency relation is provided in Appendix B.

User-generated content (UGC) is published information created by individuals, this includes social media text. In recent years, UGC has become increasingly popular as a data source for NLP research. The emergence of social media platforms, coupled with the growing volume and variety of UGC, has opened up exciting possibilities for analysing real-world language usage. The data we analyse in this thesis comes from the social media platform Twitter. Although Twitter has recently been renamed X, throughout the thesis, we refer to the platform as ‘Twitter’ and to its posts as ‘tweets’ to reflect the nomenclature of the time of data collection. As Plank (2016) points out, there are clear advantages to leveraging “fortuitous data” to develop more adaptable and robust language technology. While more recently, certain social media application programming interfaces (APIs) have introduced additional restrictions and limitations on accessing their data that restrict the availability and ease of access to UGC for research purposes, many social media platforms, host linguistic data that can provide valuable insights into the informal usage of the Irish language. It is important to recognise that UGC constitutes a distinct genre with unique features that distinguish it from both spoken language and the standardised written language commonly found in NLP corpora (Ferrara et al., 1991). These features include spelling and grammar variations, as well as language contact phenomena that occur when different languages interact. Online platforms allow users, including minority language and Irish speakers to communicate electronically from any location, rapidly reaching a broad audience without being bound by the conventional language norms upheld by publishers. By leveraging UGC, we gain the ability to document and analyse the rich linguistic

diversity prevalent in informal Irish discourse.

The accuracy of standard syntactic parsing tools tends to decline when evaluated on UGC data (Foster et al., 2011; Seddah et al., 2012). The decline in accuracy can be attributed to the inherent variations in spelling, vocabulary, and grammar often found in UGC in the context of Irish tweets. This variation also includes language contact outcomes such as code-switching with English (Lynn and Scannell, 2019). Overall, this means that existing NLP tools, such as parsers trained on IUDT, the UD treebank of standardised Irish, are not sufficient to accurately parse Irish. Additionally, Winata et al. (2023) point out that much of the research on code-switching utilises data from shared tasks, e.g. (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018; Patwa et al., 2020), and that there is a need for more open-source multilingual datasets. Therefore, the need arises for the creation of a dedicated treebank that encompasses up-to-date, real-world language data, enabling researchers to gain valuable insights into the everyday use of Irish and indeed its interactions with English. Such data is essential in order to analyse informal Irish, compare it to the prescriptive norms of standardised language observed in published Irish texts, and improve parsing accuracy for Irish UGC.

We address in this thesis the lack of linguistically-informed, genre-specific resources for accurate dependency parsing of Irish UGC. While highly accurate dependency parsers exist for well-resourced languages, low-resource languages like Irish face challenges due to the informal nature and language contact phenomena present in UGC. The absence of a UD treebank specifically designed for Irish UGC hinders the development and evaluation of dependency parsers for this genre of text. Domain adaptation or genre adaptation, in which data-driven NLP tools are trained on genre-specific data, has been shown to improve parser performance on English tweets (Kong et al., 2014) and POS tagging in Irish tweets (Lynn et al., 2015). The need, therefore, for genre-specific resources is clear in order to reliably process Irish UGC.

We present the following contributions:

1. A UD treebank of Irish tweets called **TwittIrish**
2. An analysis of the linguistic features of Irish tweets
3. Dependency parsing experiments
4. A questionnaire study

Overall, the work presented in this thesis contributes to improving the accuracy of dependency parsers for Irish tweets, enabling a better understanding of the syntactic structure of informal Irish, and facilitating the development of language technology for the Irish language community. This research contributes to the preservation and analysis of a minority language, promotes linguistic diversity in NLP, and supports the development of more effective NLP tools for low-resource languages.

1.2 Research Questions

Given the context of Irish as a low-resource language, the availability of informal textual data in the form of tweets, the multilingual applications of treebanks, and the active, growing community of UD de Marneffe et al. (2021), the overarching hypothesis of this research is that **the accuracy of dependency parsers for Irish tweets can be improved by linguistically-informed, genre-specific resource development**. To address this hypothesis, we specifically explore the following three research questions.

RQ1: How do Irish tweets differ from standard, edited Irish text? Social media has facilitated communication between Irish speakers online, resulting in a form of Irish text data different to the standardised edited text usually used in NLP. Expanding upon a previous work in which a collection of Irish Twitter text was annotated with lemmas and POS tags (Lynn, 2016), we present a linguistic examination of a dataset within this genre. To answer the research question, we examine the linguistic variation present in Irish tweets on the orthographic, morphological, lexical, and syntactic levels. We identify linguistic features present in the TwittIrish dataset and systematically compare them to the IUDT treebank of standard Irish, providing examples and explanations. This research contributes to the field of linguistics by expanding our understanding of language variation in contemporary Irish. The findings have the potential to inform language educators and curriculum developers, aiding in the adaptation of language instruction to reflect the evolving linguistic landscape. Furthermore, policymakers and cultural organisations can benefit from this research, utilising its insights to make informed decisions regarding the preservation and promotion of the Irish language. Additionally and importantly to this work, understanding the linguistic characteristics of Irish-language tweets facilitates the development and evaluation of NLP tools and models processing Irish UGC.

RQ2: What challenges are associated with parsing Irish tweets? As shown by Foster et al. (2011), parsing social media text involves various challenges. In the case of Irish, parsers trained on the IUDT (Lynn and Foster, 2016), which consists solely of standard Irish text without UGC, perform well on standardised text but, unsurprisingly, suffer a decline in accuracy when applied to Irish tweets. The test set of the TwittIrish treebank (Cassidy et al., 2022) enables us to quantify this decrease in parsing accuracy and establish baseline parsing results. Answering this research question allows us to explore various ways of enhancing parser performance. Considering the effectiveness of pre-trained, context-sensitive word encodings such as bidirectional encoder representations from transformers¹ (BERT) (Devlin et al., 2019) for many NLP tasks, we use the word representations of gaBERT (Barry et al., 2022), an Irish monolingual BERT model. We explore the impact of these contextualised token representations on parsing performance specifically for Irish tweets. We measure accuracy broken down by the length of the tweets, UPOS, and dependency relation. We answer RQ2 by performing error analysis on our preliminary parsing experiments, allowing us to identify the linguistic features of Irish-language tweets that pose the greatest challenges to dependency parsers. We then employ domain adaptation, a popular technique for resolving the challenges of parsing UGC, e.g. (Liu et al., 2018). This was done using the newly available genre-specific TwittIrish training and evaluation sets. Finally, we perform experiments combining the TwittIrish training data with data from treebanks of standardised text in Irish and Scottish Gaelic (Batchelor, 2019).

RQ3: How can language contact phenomena in Irish tweets be characterised?

As in previous work by Lynn and Scannell (2019) who performed a preliminary analysis of code-switching in Irish tweets, we find language contact to be a salient feature of this genre of Irish text. We recognise that, in order to accurately measure the frequency of language contact and its effects on downstream tasks such as dependency parsing, these phenomena must be categorised and annotated consistently. Given the lack of consensus in the academic literature on the distinction between types of language contact phenomena generally, we conduct a study to gather the opinions of Irish speakers on lone other-language items (LOLIs) observed in Irish tweets. We draw on previous research that examines English-Irish code-switching and borrowing in spontaneous spoken language (Hickey, 2009;

¹An explanation of the BERT architecture is found in Section 2.2.6.

Stenson, 1991). We also build on previous work that explores linguistic analysis of Irish tweets but does not distinguish between code-switching and borrowing (Lynn et al., 2015; Lynn and Scannell, 2019). We investigate the following subquestions:

- Do Irish speakers classify borrowed words as English less often than code-switched words?
- Do Irish speakers claim to be more likely to use borrowed words than code-switched words?
- What themes can be interpreted from Irish speakers’ explanations about word choice?

Ultimately, an accurate description of language contact phenomena is important for linguistic analysis, resource selection, data preprocessing, and technology development in NLP as well as potentially wider-reaching societal impacts on the perceptions of multilingualism and language diversity.

1.3 Thesis Structure

The thesis is structured as follows:

Chapter 2 provides a background to the various topics covered in this thesis, discussing the existing literature we draw from in order to address the research questions posed. First, we provide an overview of the Irish language by exploring the historical background, linguistic characteristics, current status of the language, and its NLP resources. We then provide background on dependency parsing, covering the topics of dependency grammar, annotated corpora and treebanks, the UD framework, and the ways in which automatic dependency parsing has evolved. Additionally, this chapter aims to place the current research within the broader scope of UGC and outline the opportunities and challenges associated with applying NLP techniques to this genre of data. Further, we examine the topic of language contact within the context of the Irish language and the field of NLP. Lastly, noteworthy research gaps in the existing literature are described.

Chapter 3 details the various stages of the TwittIrish treebank development. This dataset is the basis for the research in the chapters that follow. We describe the data

curation and the preprocessing and conversion steps carried out. We then outline the linguistic annotation process in line with the UD framework. By describing the methodology chronologically and explaining the challenges encountered, this chapter provides credence, transparency, and accountability while serving as a useful resource for any researcher undertaking a similar project.

Chapter 4 provides a linguistic analysis of Irish-language tweets as compared to standard Irish. Irish-language tweets are described at the levels of orthography, morphology, lexicon, and syntax. Extracts from Irish tweets are also analysed in this chapter. The purpose of this chapter is to answer RQ1 by explaining the differences between Irish-language tweets and standard Irish text.

Chapter 5 explores the task of parsing Irish-language tweets. First, baseline results for parsing Irish-language tweets are established. The impact of using pre-trained contextualised word embeddings from a monolingual Irish language model is also examined here. We provide insight into the strengths and weaknesses of the best-performing parser via error analysis, thus answering RQ2. Finally, we present improved results using parsers trained on the newly available *TwittrIrish* treebank in combination with standard Irish and Scottish Gaelic training data.

Chapter 6 presents a mixed methods questionnaire study investigating RQ3 by exploring how Irish speakers perceive and classify borrowed words and code-switched words in the context of Irish-language tweets. We also investigate criteria used to distinguish between code-switching and borrowing.

Chapter 7 provides a summary of the original contributions of the thesis, discusses the research findings and suggests future work in this area.

1.4 Publications

The work described in this thesis has been published in the following papers:

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster (2022). **TwittrIrish: A Universal Dependencies Treebank of Tweets in Modern Irish**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 6869-6884, Dublin, Ireland.

This paper describes the development of the TwittIrish treebank, the first UD treebank of UGC content in Irish. Linguistic analysis, parsing experiments, and baseline results using the novel resource are presented. This work is related to Chapters 3, 4, and 5 of this thesis.

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* (2022): 1-52.

This journal article provides an overview of the linguistic features of user-generated texts from the web and social media, compares existing treebanks of several languages, proposes annotation guidelines within the Universal Dependencies framework, and aims to establish a consistent framework for future research on UGC. The work of developing annotation guidelines was done as a group and the tasks of writing and editing were distributed among the authors. Our specific contribution to this article involved analysing Irish-specific examples of linguistic phenomena in UGC based on the TwittIrish treebank annotation and organising UGC linguistic phenomena into a taxonomy. This work was instrumental to the research described in this thesis. In particular, the linguistic annotation of the TwittIrish treebank documented in Chapter 3, the linguistic analysis of Chapter 4 and the annotation guidelines pertaining to UGC phenomena provided in Appendix B were largely informed by this work.

James Barry, Joachim Wagner, **Lauren Cassidy**, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, Jennifer Foster (2022). **gaBERT — an Irish Language Model**. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 4774–4788, Marseille, France. European Language Resources Association.

This paper describes the development of gaBERT, a monolingual BERT model for the

Irish language. The performance of gaBERT is evaluated against state-of-the-art multilingual models in various downstream tasks. Our contribution to this paper was evaluating the performance of each model with regard to Irish syntax using a Cloze test methodology. In Chapter 5, we describe the effects on parsing accuracy achieved by utilising the gaBERT embeddings.

Chapter 2

Background

This chapter explores previous work related to our research questions which pertain to the differences between standard Irish-language text and Irish-language tweets, the challenges of dependency parsing Irish language tweets, and the classification of language contact phenomena in Irish-language tweets.

Section 2.1 lays out the background of the Irish language by exploring its history, linguistic features, and current status, followed by an overview of the progress made with regard to NLP for Irish to date. Section 2.2 provides an overview of relevant work in the area of dependency parsing, highlighting pertinent research involving annotated corpora and treebanks, the UD framework, and current dependency parsing systems. Section 2.3 positions the current research in the history of UGC and outlines the opportunities and challenges associated with NLP for this genre of data. Section 2.4 examines the research topic of language contact as it applies to the context of Irish and the field of NLP. Finally, Section 2.5 identifies the research gaps that warrant further exploration.

2.1 The Irish Language

The Irish language, Gaelic, or *Gaeilge*, is a member of the Goidelic branch of the Celtic languages, along with Scottish Gaelic and Manx as shown in Figure 2.1. Irish is recognised as the first official language of the Republic of Ireland, a minority language in Northern Ireland, and is also an official language of the European Union. According to the 2016 Irish census (CSO, 2022), approximately 1.9 million people reported being able to speak Irish, with around 72,000 people speaking it daily outside of education. Irish has three

main dialects: Connacht, Munster, and Ulster named for the province of Ireland from which they originate.

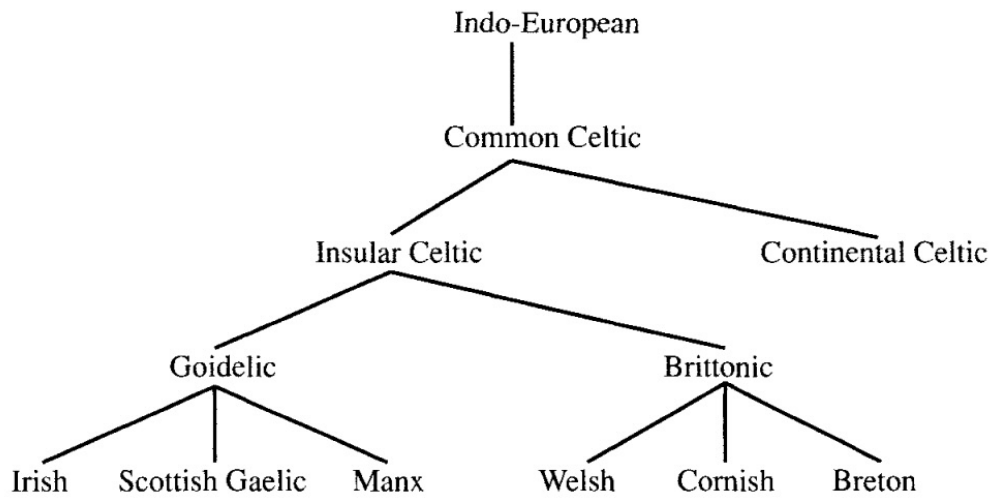


Figure 2.1: Relationship between Celtic languages. Image from Ó Siadhail (1989).

2.1.1 History

We draw from the works of Mac Giolla Chríost (2004) and Doyle (2015) to broadly delineate the history of Irish as context for our description of the current situation of the language. Latin arrived in Ireland with Christianity around 400 Common Era (CE), and by about 600 CE, Irish monks had adapted the Latin alphabet to write their native language, Old Irish. Irish invaders spread the language to Scotland and the Isle of Man where it would eventually evolve into Scottish Gaelic and Manx respectively. With the arrival of Vikings, some Old Norse borrowings were assimilated into Irish. The language of the period of 900 to 1200 CE is referred to as Middle Irish. In 1169, the Anglo-Normans invaded Ireland, bringing with them English and Norman French, but eventually adopting Irish themselves. Given the interaction between linguistic communities, the development of trading, and intellectual developments during this time, the Irish vocabulary was substantially enriched.

The language of the period 1200-1500 is referred to as Early Modern Irish and was more highly inflected than Modern Irish. During the Tudor and Stuart eras and the Late Modern period (1500-1800), the English language and culture spread throughout Ireland, increasingly influencing the Irish language. The government attempted to suppress the

Irish language and it gradually lost status in relation to English.

From 1800-1870, the increasing opportunities for education and mobility in Irish society accelerated the language shift from Irish to English. English became the language of aspiration, Catholicism, and the United States of America, which attracted millions of Irish emigrants. Despite these challenges, Irish continued to be spoken by a significant proportion of the population, particularly in rural areas. There was a revival of interest in Irish culture and language, which led to the formation of organisations such as *Conradh na Gaeilge* ‘the Gaelic League’ and the establishment of Irish-medium schools.

The 20th century then saw a number of important developments in the history of Irish, including the adoption of the language as the first official language of the Irish Free State in 1922, the establishment of the Gaeltacht regions and *An Caighdeán Oifigiúil* ‘The Official Standard’ (Rannóg an Aistriúcháin, 1958; Oireachtas, 2017). However, the decline of Irish continued throughout much of the 20th century, with many speakers switching to English in order to advance socially and economically.

2.1.2 The Current Status of Irish

The Irish language holds a privileged position in terms of government support compared to other low-resource languages but still has many of the problems of a ‘regional’ or ‘minority’ language (Kelly-Holmes, 2006). Irish has the full status of a state language in Ireland, official status in the European Union, and is supported through government policies, such as the Official Languages Act and the 20-Year Strategy for the Irish Language 2010-2030 (Government of Ireland, 2010), aiming to enhance the provision of public services in Irish and promote its revival. Mainstream media outlets like TG4,¹ RTÉ,² and Raidió na Life³ also play a significant role in supporting Irish language users through television, online media, and radio broadcasting. While Irish is primarily spoken as a first language in a few rural regions in counties Cork, Donegal, Galway, Kerry, Mayo, Meath and Waterford, collectively known as the *Gaeltacht*, it is a mandatory subject in primary and secondary education. In the period 2017-2020, the number of Irish-medium schools in the education system, and the number of pupils in these schools, were found to be increasing (Government of Ireland, 2021).

¹<https://www.tg4.ie>

²<https://www.rte.ie>

³<https://www.raidionalife.ie>

Despite its privileged status and government support, Irish still faces numerous challenges, rendering it an endangered language. The digital presence of Irish is a critical aspect of its survival and integration into the mainstream. Caulfield (2013) presented a study of the online discourse of Irish language users showing that the majority of users were located outside the Gaeltacht and that users adapted the Irish language to text-based computer-mediated communication (CMC) through syntactic and morphological variations, acronyms, modality play, and code-switching. The use of the Irish language on social media platforms has experienced consistent growth (Lackaff and Moner, 2016) with an active community sharing information related to events, government policies, education, and language learning using the hashtag #Gaeilge (Nic Giolla Mhichíl et al., 2018). While the use of Irish on these platforms represents an evolving language, the digital resources and support for Irish are still limited (Lynn, 2023). The Digital Plan for Irish (Ní Chasaide et al., 2022) identifies several key areas of investment to enable the continuous development of Irish-language technologies. These include investing in skilled researchers with high levels of competence in Irish, establishing centres of excellence to host interdisciplinary teams, creating digital innovation hubs, and engaging with and involving the community. The plan also aims to serve as a model for other minority and lesser-spoken languages and to share knowledge and resources with other communities struggling to maintain their language in the digital age.

It is crucial to consider the needs of users and the sociolinguistic context when developing language technology. Questions, for example, of ownership and authenticity arise among new and native speakers of Irish (Fhlannchadha and Hickey, 2018). Understanding such dynamics can help to inform research priorities and develop effective methods to preserve and revitalise the language (Ní Chasaide et al., 2017).

2.1.3 Linguistic Features of Irish

This section outlines some of the linguistic features of Irish pertaining to orthography, morphology, lexicon, and syntax with the intention of facilitating comprehension of the following chapters, particularly Chapter 4, in which the linguistic variation observed in Irish-language tweets will be contrasted against the features of standard Irish briefly summarised here. Three salient features of the Irish language common to Insular Celtic languages are the initial mutation of words, prepositions that inflect for person and number,

and verb-subject-object (VSO) word order. These and other phenomena will be described and exemplified in the following sections.

Orthography As described in Hickey and Stenson (2011), Modern Irish uses 18 letters of the Latin alphabet, consisting of 5 vowels which can be short (a, e, i, o, u) or long (á, é, í, ó, ú), and 13 consonants (b, c, d, f, g, h, l, m, n, p, r, s, t). The letters j, k, q, v, w, x, y, and z are also used in loanwords. Vowels are classified as either broad (a, o, u) or slender (e, i) affecting the pronunciation of neighbouring consonants with regard to palatalisation. Vowel harmony within words consists of the nearest vowel on the right side of a consonant matching the nearest vowel to the left in terms of broad/slender quality.

In the process of standardising the Irish language, the spelling of many words was simplified in that letters were removed in cases where they were no longer pronounced in any dialect (Oireachtas, 1947). Many abbreviations are possible in standard Irish e.g. *agus* ‘and’ → *a’s*, *contae* ‘county’ → *co.*, etc.

Another salient feature of Irish, evident in its orthography is its system of initial mutation wherein the beginning of a word undergoes pronunciation changes such as lenition or eclipsis, indicating various morphological processes determined by properties such as gender, number, case, possession, adjective agreement, and tense.

Inflectional morphology Stenson (1981) and Uí Dhonnchadha (2002) both describe the inflectional morphology of Irish. Stenson (1981) provides a morphological description of the parts of speech of Irish, noting those that are common to Indo-European languages i.e. nouns, verbs, adjectives, adverbs, prepositions, conjunctions, along with several clitic particles. Uí Dhonnchadha (2002) describes the implementation of a rule-based analyser for Irish inflectional morphology (later extended to cover derivational morphology and other linguistic analysis (Uí Dhonnchadha, 2009)). The documentation of this work includes a discussion of the particularities of Irish inflectional morphology in the context of NLP. Table 2.1 provides an overview of the inflections in Irish according to POS. Adverbs can be formed using the particle *go* with an adjective, e.g. *mall* ‘slow’ → *go mall* ‘slowly’.

Derivational morphology New words can be formed in Irish via affixes such as the diminutive suffixes *-ín*, *-án*, *-óg*, the emphatic suffix *-sa/-se*, and prefixes such as *ath* ‘re-’, *frith* ‘anti-’. Prefixes sometimes require hyphenation and trigger lenition in the morpheme

Word class	Features reflected in inflection
Verb	tense, mood, aspect, voice, number, person
Noun	gender, case, number, definiteness, emphasis
Adjective	gender, case, number
Pronoun	gender, number, person
Article	gender, case, number
Prepositional pronoun	gender, number, person

Table 2.1: Inflections in Irish. Adapted from Uí Dhonnchadha (2002).

they modify, e.g. *óg* ‘young’ → *ró-óg* ‘too young’, *mór* ‘big’ → *rómhór* ‘too big’. Verbal adjectives and verbal nouns can be derived from verb stems e.g. *can* ‘sing’, *canta* ‘sung’, *canadh* ‘singing’. The verbal noun is the sole non-finite verb form, used to nominalise the verb or express gerundive constructions. Stenson (1981) notes its categorical ambivalence falling somewhere between a noun and verb in that its stem is deverbal but it acts as a noun phrase in some cases e.g. it induces the genitive case in a subsequent noun and can itself adopt the genitive case. Stenson (1981), Lynn and Scannell (2019), and Caomhánach (2022) all note that the suffix *-(e)áil* is particularly productive, frequently used with other language words to derive a verb or verbal noun e.g. *tvúit* ‘tweet’ → *tvúiteáil*. Adjectives can be derived from nouns using the suffix *-(e)ach* e.g. *cumas* ‘ability’ → *cumasach* ‘able’. In their linguistic analyses of Irish, Uí Dhonnchadha (2009) and Lynn (2016) opt to classify verbal nouns and verbal adjectives as nouns and adjectives respectively rather than verbs, a convention we also adopt in the current research.

Lynn et al. (2012) and (Bohnet et al., 2013) make observations about the parsing challenges that can arise due to lexical diversity in the context of a rich morphology as a single lemma or root word form can correspond to various surface forms, e.g. the singular noun *bliain* ‘year’ is rendered as *bliain*, *bhliain*, *bliana*, *mbliana* depending on the grammatical context. Awareness of these phenomena is necessary for accurate tokenisation and lemmatisation of Irish text, which has downstream effects on POS tagging and syntactic annotation.

Lexicon Words in Irish are generally separated by white space however several exceptions exist. In their respective resource development for Irish-language technology, Uí Dhonnchadha (2009) and Lynn (2016) provide full annotation guidelines, in which they explain the treatment of various phenomena such as multiword tokens and multitoken words in Irish. We follow these guidelines as closely as possible in order to maximise

resource compatibility. Consistent handling of such items is essential for accurate tokenisation or word segmentation. Further discussion of annotation decisions is included in Chapter 3 and full annotation guidelines are in Appendix B.

Multiword tokens are single orthographic tokens that correspond to multiple syntactic words e.g. *do + an* → *don* ‘for the’, *i + an* → *sa* ‘in the’, *le+a* → *lena* ‘with their’. Multiword tokens sometimes contain a word-internal apostrophe e.g. *do + ith* → *d’ith* ‘ate’, *ba + fhéidir* → *b’fhéidir* ‘maybe’, *mo + athair* → *m’athair* ‘my father’.

Simple prepositions in Irish can combine with personal pronouns to form prepositional pronouns or conjugated prepositions, e.g. *ag+mé* → *agam* ‘at me’, *le+tú* → *leat* ‘with you’. A compound preposition in Irish consists of a simple preposition combined with a noun e.g. *tar éis* ‘after’, *os cionn* ‘over’. Multitoken words constitute the inverse phenomenon in which multiple tokens function as a single syntactic unit. McGuinness et al. (2020) and Walsh (2023) provide an in-depth analysis of multiword expressions (MWEs) in Irish.

Uí Dhonnchadha (2009) opted to treat some MWEs as single units, joining them with an underscore. This decision was based on compositionality (whether the meaning of the entity can be inferred from the parts) e.g. *Baile Átha Cliath* ‘Dublin’. Though the string of words has an internal structure literally meaning ‘town of the hurdled ford’, it can be disadvantageous to represent these internal structures in NLP systems when the phrase is used as a unit to refer to a single named entity.

Uí Dhonnchadha (2009) notes several other phenomena that can cause issues for the tokenisation of Irish such as URLs, email addresses, and list items e.g. ‘(iii)’, ‘(B)’, typographical errors, dialectal variants, unseen words and named entities with the English possessive suffix, e.g. ‘Madigan’s’.

Syntax With regard to syntax, Irish has a verb-subject-object (VSO) word order as shown in Example 2.1. VSO word order is common to Insular Celtic languages but is only found in 9.2% of languages globally (Tomlin, 2014).

(2.1) *Scríobh sí leabhar*
wrote she a-book
VERB SUBJ OBJ
‘She wrote a book’

The English verb ‘to be’ corresponds in Irish both to a ‘substantive’ verb *bí* which inflects

like other verbs, as well as a copula with limited morphology. The word order in a substantive construction is verb-subject-predicate as exemplified in Example 2.2, where the predicate *suimiúil* ‘interesting’ is an adjective. Nominal predicates are not possible in the substantive construction (Carnie, 1995), as shown in Example 2.3. Where the predicate is a noun phrase, the copular construction is used as exemplified in Example 2.4.

(2.2) *Tá an leabhar suimiúil*
 is the book interesting
 VERB SUBJ PRED
 ‘The book is interesting’

(2.3) **Tá sí scríbhneoir*
 is she a-writer
 VERB SUBJ PRED
 ‘She is a writer’

The word order in a simple copular construction is copula-predicate-subject.

(2.4) *Is scríbhneoir í*
 is she a-writer
 COP PRED SUBJ
 ‘She is a writer’

A subpredicate or pronominal augment is needed in identificational copular constructions involving definite subjects as shown in Examples 2.5 and 2.6 where ‘AUG’ represents the pronominal augment.

(2.5) *Is scríbhneoir í Sally*
 is a-writer Sally
 COP PRED AUG SUBJ
 ‘Sally is a writer’

(2.6) *Is í Sally an scríbhneoir*
 COP AUG PRED SUBJ
 is her Sally the writer
 ‘Sally is the writer’

Doherty (1996) describes the pronominal augment as an extra morpheme orthographically represented by an accusative pronoun that agrees in person and number with the noun phrase immediately to its right.

Chapter 3 includes a description of the linguistic annotation carried out in the creation of the TwittIrish treebank and Appendix B contains the full annotation guidelines.

2.1.4 Irish-Language Technology

While some progress has been made with regard to the development of language technology for Irish, it is, along with Maltese, the only official language in the European Union considered to have weak or no support (Lynn, 2023). Irish has limited resources and, consequently, cannot fully benefit from the significant progress achieved in language technology for other languages that can leverage vast amounts of data. However, despite these challenges, notable advancements have been made in the field of Irish-language technology with regard to lexical resources, word embeddings, corpora, and tools, as well as their applications in areas such as computer-aided language learning, machine translation, and speech synthesis (Ní Chasaide et al., 2022; Lynn, 2023).

Automatic linguistic analysis of Irish In this section, we narrow our focus to the specific Irish-language technology resources relevant to this research. The XFST Finite State suite of tools for Irish (IFST), which comprises a tokeniser, lemmatiser, morphological analyser, POS tagger utilising the PAROLE tagset (Ó Cróinín and Uí Dhonnchadha, 1998), and partial constraint grammar parser represents a highly significant development in Irish NLP (Uí Dhonnchadha, 2009). These foundational rule-based tools have served as the basis for the development of various other tools and resources in the field, including the first Irish Dependency Treebank (IDT) (Lynn et al., 2012). Lynn (2016) then developed the IU DT by converting the IDT to the UD annotation scheme. As of UD version 2.12, the IU DT consists of 4,910 sentences with its data and annotation guidelines subject to frequent updates. These updates serve to promote consistency and linguistic accuracy. McGuinness et al. (2020), for example, describe an updated methodology for the annotation of MWEs in Irish within the UD framework.

The diversity of genres as well as languages on the UD platform continues to grow. Doyle et al. (2019) employed a character-level LSTM tokenisation approach for tokenisa-

tion of Early Irish texts allowing them to develop two UD treebanks⁴⁵. Scannell (2022) introduced a UD treebank of pre-standard Irish texts establishing baselines for lemmatisation, tagging, and dependency parsing using machine learning techniques. UD treebanks have also been developed for other Celtic languages related to Irish such as Breton (Tyers and Ravishankar, 2018), Scottish Gaelic (Batchelor, 2019), Welsh (Heinecke and Tyers, 2019), and Manx (Scannell, 2020). The consistency of annotation within UD across genres and languages facilitates interoperability and can be especially useful for low-resource languages.

Linguistic analysis of Irish UGC was first explored by Lynn et al. (2015) by lemmatising and POS tagging a corpus of 1,493 Irish-language tweets randomly sampled from 950k tweets by 8k users posted between 2006 and 2014. We refer to this corpus as the Lynn Twitter Corpus (LTC).⁶ This data set has been used to train a statistical POS tagging model that achieved a 10% accuracy improvement over the IFST rule-based tagger which was designed to process standard Irish text. Code-switching annotation was later added to the LTC (Lynn and Scannell, 2019).

2.2 Dependency Parsing

Dependency parsing is an NLP task whereby, for a given sentence, a dependency tree is generated to represent its syntactic structure. In this section, we outline the theoretical tradition of dependency grammar that underpins dependency parsing, explain the concepts of the dependency tree and treebanks, provide background on UD, the specific framework that we use in our syntactic analysis of Irish-language tweets, and, finally, describe methods for automatic dependency parsing with a focus on graph-based parsing.

2.2.1 Dependency Grammar

Dependency grammar is a linguistic framework in which syntactic structure is represented by the relationships between words in a sentence.

“Each word in a sentence is not isolated as it is in the dictionary. The mind perceives connections between a word and its neighbours. The totality of these connections forms the scaffold of the sentence.” (Tesnière, 1959)

⁴https://github.com/UniversalDependencies/UD_Old_Irish-DipSGG/tree/master

⁵https://github.com/UniversalDependencies/UD_Old_Irish-DipWBG/tree/master

⁶<https://github.com/tlynn747/IrishTwitterPOS>

Consider Examples 2.7 and 2.8 which consist of the same set of words, but differ in meaning. Although a certain amount of information is understood given the individual words of a sentence in isolation, the **interrelations** of the words are needed for full interpretation.

(2.7) *Shábháil Mia madra*
Mia saved a dog

(2.8) *Shábháil madra Mia*
A dog saved Mia

There exist many ways to represent a sentence structure. These approaches can be broadly grouped in to two main traditions: constituency grammars (exemplified in Figure 2.2) and dependency grammars (exemplified in Figure 2.3).

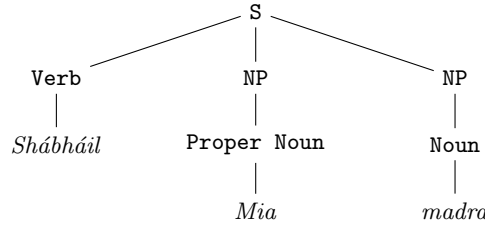


Figure 2.2: Phrase Structure tree ‘Mia saved a dog’.

Phrase Structure grammar describes how constituents are organised hierarchically in a sentence. This constituency-based approach was formalised in the seminal work of Chomsky (1956) and became a popular approach to syntactic analysis.

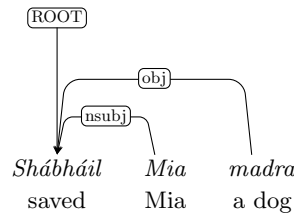


Figure 2.3: Dependency tree ‘Mia saved a dog’.

The roots of dependency grammar can be traced back to Pāṇini’s grammar of Sanskrit several centuries before the Common Era (CE). However, it was in the 20th century that modern dependency grammar emerged as a distinct linguistic theory. The ideas of Tesnière (1959) laid the groundwork for the development of modern dependency grammar, an alternative to the prevalent phrase structure-based theories of the time. Since then, various formulations and extensions of dependency grammar have been proposed by linguists and

researchers. These include Functional Generative Description (Sgall et al., 1986), Meaning-Text Theory (Mel’čuk et al., 1988), Word Grammar (Hudson, 2010), among others. Each of these theories offers its own perspectives and methodologies for analysing dependencies and syntactic structures in different languages.

Our motivations for opting to use a dependency-based approach for our analysis of Irish-language tweets are to enable ease of annotation with regard to the nature of Irish syntax and to maximise resource interoperability Uí Dhonnchadha (2009) and Lynn (2016) both note that theoretical questions arise when trying to represent Irish using phrase structure grammar, such as the nature of verb phrases and the possibility of discontinuous constituents, and that linguists disagree on these issues. In their development of NLP resources for Irish, Uí Dhonnchadha (2009) and Lynn (2016) recognised that dependency analysis allows for a more suitable representation of such constructions. Additionally, by adopting the UD framework (see Section 2.2.4), we aim to maximise compatibility, not only with other Irish-language technology resources, such as an existing treebank for standard Irish, but also cross-linguistically. In constituency-based approaches, the tree structure reflects the specific word order. Mel’čuk et al. (1988) argue that such approaches became popular in North America as they are better suited to languages with more fixed word orders such as English. Dependency trees, on the other hand, are not sensitive to word order. A dependency grammar is well suited to represent similarities in grammatical structure across many languages even when their word order differs, e.g. Irish is a VSO language and English is an SVO language. Additionally, dependency grammar can process languages with both flexible and strict word orders equally effectively.

2.2.2 Dependency Trees

In order to introduce the topic of dependency trees, we adopt a notation adapted from Kübler et al. (2009) and Jurafsky and Martin (2023). Figure 2.4 exemplifies an input sentence S consisting of a **sequence of tokens** $w_0w_1\dots w_n$. We consider tokenisation, the segmentation of a sentence into its component words, as a separate task and assume that tokenisation is known at the time of parsing. Each token w_i represents a word where i is an **index** referencing the position of every word in the sequence. The inclusion of an artificial ROOT node w_0 simplifies the formal definition and the computational execution of dependency structures (Kübler et al., 2009).

ROOT Fásfaidh crann mór
 w_0 w_1 w_2 w_3

Figure 2.4: Tokenisation of sentence *Fásfaidh crann mór* ‘A big tree will grow’.

A dependency tree represents the syntactic structure by directed, typed, binary relations between headwords and dependent words. For example, Figure 2.5 exemplifies a dependency tree in which *Fásfaidh* ‘will-grow’, as the main verb of the sentence, depends on the ROOT node. The token *crann* ‘tree’, is a dependent of the main verb via a nominal subject (**nsubj**) relation. Finally, the adjective *mór* ‘big’ is dependent on the noun which it modifies via the adjectival modifier relation (**amod**).

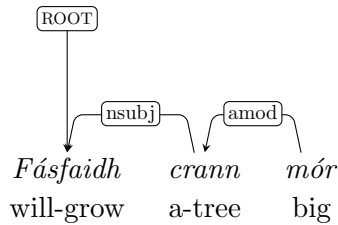


Figure 2.5: Dependency tree ‘A big tree will grow’.

A dependency tree can be conceptualised as a directed graph $G = (V, E)$ consisting of V the set of **vertices** representing the words of a sentence, and E the set of **edges** connecting the vertices. We use the words ‘edge’ and ‘arc’ interchangeably to describe these connections between vertices. $L = \{l_1, \dots, l_m\}$ is a set of **dependency relation labels** that can hold between any two words in a sentence, e.g. subject, object. The set of edges E represents the labelled dependency relations of the particular analysis G . Specifically, an edge (w_i, l, w_j) represents a dependency relation from head w_i to dependent w_j labelled with relation type l .

Given the example sentence of 2.4 and 2.5 *Fásfaidh crann mór* ‘A big tree will grow’, we can define equations (2.9) and (2.10)

$$V = \{\text{ROOT}, \text{Fásfaidh}, \text{crann}, \text{mór}\} \quad (2.9)$$

$$E = \{(\text{ROOT}, \text{root}, \text{Fásfaidh}), (\text{Fásfaidh}, \text{nsubj}, \text{crann}), (\text{crann}, \text{amod}, \text{mór})\} \quad (2.10)$$

A **dependency tree** is a specific type of dependency graph G for which the following constraints hold (Jurafsky and Martin, 2023):

1. G has a single ROOT with no incoming edges.
2. Except for the ROOT, each vertex of G has exactly one incoming edge.
3. There is a unique path from the ROOT to each vertex in G .

Thus the valid dependency tree is weakly connected, single-headed, and acyclic. V contains each token of the sentence as a node and E contains edges between pairs of words capturing a head-dependent grammatical function.

Another important property of dependency trees is **projectivity**. A tree is projective if none of its arcs cross one another. Non-projective trees occur in some languages especially those with more flexible word orders. Non-projective trees are not handled by all parsing algorithms. More details on how parsing algorithms handle non-projectivity are provided in Section 2.2.5. Non-projective trees are rare but possible in Irish. Figure 2.6 exemplifies a non-projective sentence in which the arc connecting the word *suim* ‘interest’ to *cairde* ‘friends’ intersects the arcs connecting the words *agam* ‘at-me’ and *anois* ‘now’ to *tá* ‘is’.

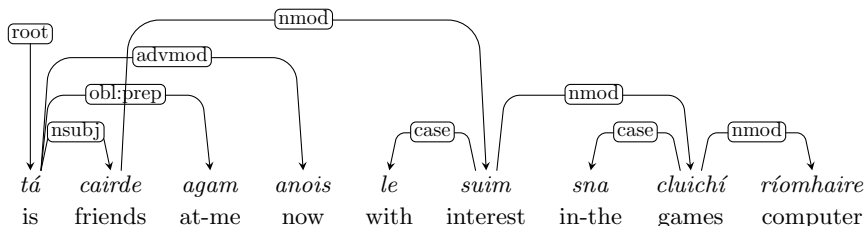


Figure 2.6: Non-projective sentence ‘I have friends now with an interest in computer games’.

2.2.3 Treebanks

A dependency treebank is a collection of texts annotated with their syntactic structures in the form of dependency trees. It is essentially a parsed corpus in which each sentence is represented as a dependency tree. Treebanks are important linguistic resources for research and analysis. They can be used for studying syntactic structures, linguistic phenomena, and language typology. Grammatical theories and formalisms can be developed and refined using quantitative analysis of annotated data in order to find language usage patterns, frequency distributions, and linguistic variation across different genres, registers, and time periods. In data-driven NLP, treebanks are also used as training and test data for automatic syntactic parsing.

Early treebank development used a constituency-based approach, the most well-known of which is the Penn Treebank (Marcus et al., 1993). Dependency treebanks were then developed for many languages including Czech (Hajič, 1998), Russian (Boguslavsky et al., 2000), Italian (Bosco et al., 2000), Dutch (Van der Beek et al., 2002), Arabic (Hajič et al., 2004), among others. Dependency annotation schemes varied greatly among these resources as they were developed for different languages and projects. This lack of standardisation made it difficult to compare and combine treebanks, impeding cross-linguistic research and multilingual NLP. With the increasing availability of dependency treebanks in various languages and growing interest in multilingual NLP, it became crucial to have a unified annotation scheme that could be applied to a wide range of languages.

This demand for standardisation of annotation schemes motivated the development of a reusable morphological feature tagset Zeman (2008), Google Universal POS tags (Petrov et al., 2012), a harmonised multi-language dependency treebank of Zeman et al. (2012), and the Universal Stanford dependencies of de Marneffe et al. (2014), each representing a vital step towards a more unified annotation scheme allowing consistency across data sets. McDonald et al. (2011) conducted a multilingual parsing experiment in which delexicalised parsing models across multiple source languages were evaluated on a variety of target languages, demonstrating that the target language with the highest parsing performance often did not align closely with the source language in terms of typology. However, subsequent research by McDonald et al. (2013) found that the potential advantage of using related training languages for improving parsing accuracy had been obscured by annotation discrepancies. Through the implementation of a standardised annotation scheme, the anticipated benefits of utilising related languages became evident in the parsing accuracy results. These efforts of unification culminated in the creation of UD, the framework we utilise in our research.

2.2.4 Universal Dependencies

UD (Nivre et al., 2016, 2020; de Marneffe et al., 2021) was launched as a cross-linguistically consistent annotation scheme aimed to facilitate the development of multilingual research and resource sharing. UD utilises 17 Universal POS (UPOS) tags which are derived from the Google Universal POS tagset (Petrov et al., 2012), a morphological feature set inspired by Zeman (2008), and 37 dependency relations derived from the Universal Stanford

dependencies (de Marneffe et al., 2014). The UPOS tags and dependency relations are described and exemplified in Appendix B.

Rather than a linguistic theory of universal grammar, UD is an evolving open-source framework developed collaboratively by an active community, to represent languages consistently in order to process them computationally, promoting the sharing of data, tools, and models. The UD framework is designed to uphold a balance of six criteria⁷:

1. Adequacy for individual language analysis
2. Usefulness for linguistic typology
3. Facility of efficient and consistent annotation
4. Suitability for accurate computer parsing
5. Comprehensibility and usability for non-linguists
6. Support for downstream NLP tasks

Manning’s Law states:

“It’s easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.” (Nivre et al., 2017)

Thus, trade-offs are made in UD with regard to its practical usefulness for downstream applications, such as question answering, named entity recognition, and consistency with linguistic theory (Osborne and Gerdes, 2019).

2.2.5 Dependency Parsing Frameworks

Automatic dependency parsing involves the use of computational methods to predict dependency representations. The two main approaches used are **transition-based** (Covington, 2001; Nivre, 2003; Yamada and Matsumoto, 2003) and **graph-based** parsing (McDonald et al., 2005).

Transition-based parsing is an incremental approach to sentence parsing that proceeds from the beginning to the end of the sentence. It utilises two data structures: a buffer, which holds unparsed words, and a stack, which stores words whose dependencies have

⁷<https://universaldependencies.org/introduction.html>

not been fully processed. Rather than directly predicting the parse tree edges, transition-based parsers predict the next action to be taken, such as shifting a word onto the stack, reducing words from the stack, or creating an arc between words. These predictions are based on the current state of the parser, which includes the configuration of the stack and the buffer.

Graph-based dependency parsing is another prominent approach for predicting dependency trees from input sentences and the approach we focus on in this thesis. Unlike transition-based parsers which make local decisions, graph-based parsing systems employ a different strategy, leveraging techniques from graph theory by searching through the space of possible dependency trees for an optimal solution. One advantage of graph-based parsers over transition-based is their ability to handle long-distance dependencies (McDonald and Nivre, 2011). By scoring entire trees instead of relying on local decisions, graph-based methods effectively address this challenge. Graph-based dependency parsers search through all possible parses for a given input, aiming to find the parse that maximises a specific score. Systems can be **first-order**, where the scoring function is based on a head-modifier relation, or **higher-order**, where sibling and grandparent relations are also included.

In the following sections, we will introduce the two main components of the graph-based parser: the **parsing algorithm** that finds the best parse tree given the scores of all potential edges and the **scoring model** that assigns a score to each edge.

Graph-based parsing algorithm The parsing algorithm is used to search the space of possible dependency trees \mathcal{G}_S and select the one that has the highest score according to the scoring model. The parsing problem can be stated mathematically as in equation (2.11).

$$\hat{T}(S) = \arg \max_{t \in \mathcal{G}_S} \text{Score}(t, S) \quad (2.11)$$

This problem can be solved using algorithms for finding the **maximum spanning tree** (MST). In a fully-connected graph $G = (V, E)$, a subgraph $T = (V, F)$ is considered a spanning tree if it has no cycles and each vertex, except the root, has exactly one incoming edge. If a spanning tree emanates from the ROOT then it is a valid parse. The MST, then, is the spanning tree with the highest score. Therefore, the optimal dependency parse of S is equivalent to finding the MST of G that emanates from the ROOT.

A fully-connected, weighted, directed graph G is created to represent the input sentence S . The vertices represent the words of S and the directed edges represent all possible head-dependent assignments. An additional ROOT node is added with outgoing edges directed at all of the other vertices. A weight is assigned to each edge in G reflecting its score as a possible head-dependent relation as determined by the scoring model. Figure 2.7 depicts such a graph with the desired parse, equivalent to the MST, shown in blue.

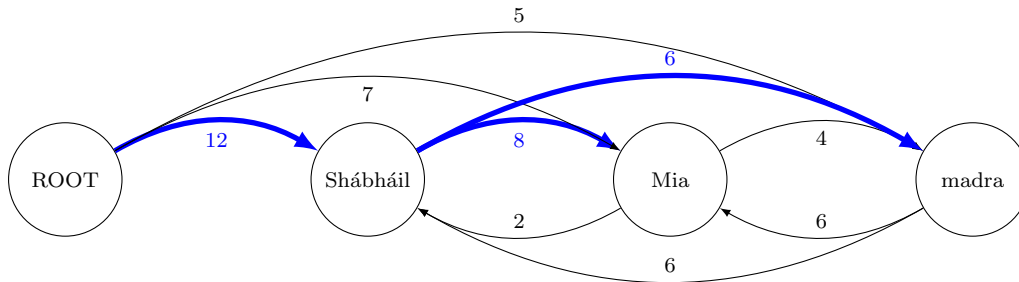


Figure 2.7: Fully-connected, weighted dependency graph *Shábháil Mia madra* ‘Mia saved a dog’.

Two notable approaches have been proposed for finding the MST in the context of dependency parsing. Eisner (1996) introduced a bottom-up, dynamic programming solution with $O(n^3)$ complexity for finding the MST in the space of possible projective trees. It is a generative model for performing second-order horizontal dependency parsing. McDonald et al. (2005) employed the Chu-Liu-Edmonds MST algorithm (Chu, 1965; Edmonds, 1967) to solve this problem. Their algorithm is greedy and recursive, providing a solution to the problem with $O(n^2)$ complexity while allowing for non-projective trees.

Figure 2.8 shows the Chu-Liu-Edmonds algorithm whereby for each vertex in G the incoming edge with the highest score is chosen. If the resulting set of edges produces a spanning tree, then that is the MST and is returned as the predicted parse tree. If the set of edges selected contains cycles, the cycles are eliminated using a recursive cleanup phase. All weights in the graph are scaled by subtracting the score of the maximum edge entering each vertex. A new graph is created by selecting a cycle and collapsing it into a single new node. This means that edges that entered or left the cycle now enter or leave the new node and the edges that were within the cycle are removed. The MST of this new graph is then found, indicating which edge can be deleted to eliminate the cycle. This can continue recursively as long as cycles are encountered. The collapsed node is expanded, restoring all the vertices and edges except the edge to be deleted.

```

function MAXSPANNINGTREE( $G = (V, E)$ ,  $root$ ,  $score$ )
   $F \leftarrow \emptyset$ 
   $T' \leftarrow \emptyset$ 
   $score' \leftarrow \emptyset$ 
  for each  $v \in V$  do
     $bestInEdge \leftarrow \arg \max_{e=(u,v) \in E} score[e]$ 
     $F \leftarrow F \cup \{bestInEdge\}$ 
    for each  $e = (u, v) \in E$  do
       $score'[e] \leftarrow score[e] - score[bestInEdge]$ 
  if  $T = (V, F)$  is a spanning tree then
    return  $T$ 
  else
     $C \leftarrow$  a cycle in  $F$ 
     $G' \leftarrow \text{Contract}(G, C)$ 
     $T' \leftarrow \text{MaxSpanningTree}(G', root, score')$ 
     $T \leftarrow \text{Expand}(T', C)$ 
  return  $T$ 

function CONTRACT( $G, C$ )
  Contract the cycle  $C$  into a single node in graph  $G$ 
  return contracted graph

function EXPAND( $T, C$ )
  Expand the contracted node in tree  $T$ 
  return expanded graph

```

Figure 2.8: Chu-Liu-Edmonds algorithm for finding the MST of a weighted directed graph. Adapted from Jurafsky and Martin (2023).

Scoring model In graph-based parsing, the scoring model is responsible for assigning scores or weights to different dependency trees based on their likelihood of being the correct parse. The score of an edge represents the probability of a dependency from the headword w_i to the modifier word w_j with the label l . In a first-order, **edge-factored** system, the score of a tree t representing sentence S is defined as the sum of the scores of each edge e of the tree as shown in equation (2.12).

$$\text{Score}(t, S) = \sum_{e \in t} \text{Score}(e) \quad (2.12)$$

An inference-based learning process is employed training the model to assign higher scores to correct parses and iteratively improve its performance. The initial step is parsing a sentence from the training data. During parsing, the model assigns a score to a parse tree using an initially random set of weights. The resulting parse is compared to the known, correct parse tree in the training data and the model parameters are updated accordingly.

2.2.6 Neural Networks in Dependency Parsing

In this section, we explain the motivation for the use of neural network architectures in the task of dependency parsing by first explaining previous popular approaches and their limitations. Prior to the widespread adoption of neural network approaches to dependency parsing, the scoring model described in the previous section relied on word representations in the form of manually engineered **features** consisting of lexical, syntactic, or contextual information. The scoring was based on the probability of dependency edges given sparse binary vectors encoding information such as wordforms, lemmas, POS tags, contexts before, after, and between the words, the dependency relation label, the length and direction of the edge, or the distance from the head to the dependent. Equation (2.13) shows the calculation of edge scores in a feature-based model using the weighted sum of features.

$$\text{Score}(S, e) = \sum_{i=1}^N w_i f_i(S, e) \quad (2.13)$$

The equation is simplified and made more computationally efficient using a dot product as in equation (2.14).

$$\text{Score}(S, e) = w \cdot f \quad (2.14)$$

Among the challenges of developing manually engineered features is the labour-intensive nature of the process and the need for linguistic expertise. Additionally, long-distance dependencies are difficult to capture using a feature-based approach.

Chen and Manning (2014) developed a **neural** parser to address the limitations of feature-based parsing and achieve improvements in speed and accuracy by utilising dense representations **learned** within the parsing task. The long short-term memory (LSTM), a type of recurrent neural network (RNN) that can capture sequential information, was then introduced to generate contextual representations for the stack and buffer in transition-based parsing (Dyer et al., 2015). Kiperwasser and Goldberg (2016) extended this approach by employing a bidirectional long short-term memory (BiLSTM) to create feature representations for individual tokens in both graph-based and transition-based parsing. Each word is encoded using its BiLSTM representation. The feature function comprises a concatenation of a small set of these encodings, which is subsequently fed into a non-linear scoring function in the form of a multilayer perceptron (MLP). By using a bidirectional LSTM, information from both preceding and succeeding words is considered, allowing the

model to capture the context of each word effectively. The BiLSTM is trained jointly with the parsing objective to encode an effective feature representation specific to the parsing task.

Pretrained contextualised word embeddings Representations of words learned by neural networks are known as **embeddings** or **vectors**. As touched on in the previous section, these representations reduce reliance on manual feature engineering. In this section, we describe the evolution of word embeddings. Advances in word embeddings have enhanced accuracy for dependency parsing and other NLP tasks by imbuing parsers with semantic context, improving the handling of out-of-vocabulary words, enabling cross-lingual generalisation, and capturing relevant contextual information. Early work in this area related to static embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). These representations are derived from unlabelled corpora and encode semantic similarity and linguistic relationships by embedding words in a vector space. Further improvement is achieved with dynamic contextual embeddings, such as embeddings from language models (ELMo) (Peters et al., 2018), whereby distinct embeddings for a word in its context are computed. ELMo uses a BiLSTM network to capture contextual information, generating a unique embedding for each occurrence of a word based on its surrounding context. ELMo embeddings are dynamic and context-dependent, allowing the model to capture nuances in meaning that static embeddings like Word2Vec or GloVe might miss. An additional performance boost became possible by using the encoder component of the transformer architecture (Vaswani et al., 2017) to train word representations. A notable example of this architecture is BERT (Devlin et al., 2019), a transformer network with 12 transformer blocks trained bidirectionally using masked language modelling, it learns to predict missing words within a sentence, and whether two sequences are contiguous, allowing the language model to capture intricate contextual nuances effectively. The key to BERT’s improvement over previous language models lies in its **attention mechanism**, which enables the model to weigh and attend to different parts of a sentence simultaneously, capturing rich contextual information across various positions.

gaBERT In our research, we leverage the innovation in the space of language models and word embeddings through the development of gaBERT (Barry et al., 2022), a mono-

lingual BERT model designed specifically for the Irish language. gaBERT was trained on a diverse set of data sources, including the CoNLL’17 Irish data (Ginter et al., 2017), the IMT collection (Dowling et al., 2020), the New Corpus for Ireland (Kilgarriff et al., 2006), the unshuffled Irish portion of the 2019 OSCAR corpus (Suárez et al., 2019), the Irish side of the ga-en bitext pair of ParaCrawl v7 (Bañón et al., 2020), and text from Irish Wikipedia,⁸ resulting in a total of 9.3 million sentences (171.3 million words) before applying sentence filtering. After applying a document-level filter to remove noisy text, 7.9 million sentences (161 million words) remained. Barry et al. (2022) explore different filtering criteria, vocabulary sizes, and subword tokenisation models and compare gaBERT against other multilingual and monolingual Irish BERT models, demonstrating that gaBERT provides superior representations for downstream parsing tasks. gaBERT was also assessed in a cloze test and MWE identification task. The gaBERT model and related code are publicly available.⁹

2.3 User-generated Content

In this section, we describe the history of UGC. We then explore its benefits and challenges as an NLP data source.

2.3.1 The Rise of User-generated Content

As the World Wide Web became publicly accessible in the early 1990s, a new form of communication emerged. CMC is ‘communication that takes place between human beings via the instrumentality of computers’ (Herring, 1996). Academic interest began in the specific style of language used on the internet, and terms ‘Electronic Language’, ‘E-grammar’, and ‘Netspeak’ were coined. The term ‘Web 2.0’ denotes the move from static to dynamic and interactive websites. The invention of social networks allowed users to create profiles and generate their own content. As the user base of the Internet expanded and diversified beyond just computer enthusiasts, businesses began to realise the value associated with the vast amounts of human data available online for marketing (Krumm et al., 2008). Naab and Sehl (2017) define UGC as consisting of published personal contributions created “outside the realm of a profession”. Tasks like sentiment analysis, topic modelling

⁸<https://dumps.wikimedia.org/gawiki/20210520/>

⁹<https://huggingface.co/DCU-NLP/bert-base-irish-cased-v1>

and named entity recognition using UGC have become popular in NLP research as social media platforms continue to evolve and gain popularity.

2.3.2 Processing User-generated Content

Benefits of processing User-generated Content Plank (2016) refers to UGC as a fortuitous data source for NLP researchers due to its availability. UGC is publicly accessible online, often in text format, making it easily obtainable for research purposes. This accessibility enables researchers to gather large amounts of data efficiently. Another benefit of UGC as a data source is that it exists in vast quantities on the Internet. This is particularly beneficial for NLP research involving languages with limited linguistic resources and relatively small speaker populations, like Irish. Collecting and studying social media data in these languages aids their preservation and promotion by contributing to the development of language-specific NLP tools, resources, and applications. Given that UGC is constantly being generated, it also provides researchers with corpora that are up-to-date, offering a unique window into current social, cultural, and linguistic phenomena. Through analysis of UGC, researchers can gain valuable insights into trends, topics, and opinions expressed by users. Another advantage of UGC is that it reflects natural language usage in informal registers. The conversational and unedited nature of UGC captures the everyday language employed by language users. Finally, UGC offers a diverse range of linguistic variation and contextual richness compared to other text genres such as newswire and fiction. With fewer barriers to entry, individuals from various backgrounds contribute to UGC, resulting in a broader representation of language use. This diversity is beneficial for NLP research as it allows researchers to analyse different language varieties. By incorporating and accounting for this diversity, language models and applications can become more inclusive and effective in addressing real-world linguistic needs.

Challenges of processing User-generated Content Several challenges are associated with processing UGC due to its linguistic variation (Foster, 2010). The ever-evolving nature of UGC also presents difficulties in maintaining effective domain adaptation. Eisenstein (2013) discusses two approaches to overcoming the challenges associated with processing social media text as opposed to standardised text: 1) Normalisation, whereby the text is adapted to suit the model, and 2) Domain/genre adaptation, whereby the model is

adapted to suit the text. We take the approach of genre adaptation by creating a dataset of UGC that can be reused for other tasks using data-driven techniques. In this way, no normalisation is needed, reducing the number of steps in the processing pipeline. Another obstacle in processing UGC is that data availability fluctuates. Furthermore, UGC as a genre, evolves rapidly, raising the possibility of dataset obsolescence over time.

Universal Dependencies for User-generated Content UGC, especially social media text, has recently become a popular focus in parsing and NLP research more broadly (Silveira et al., 2014; Luotolahti et al., 2015; Albogamy and Ramsay, 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 the UD framework as many of the linguistic phenomena common in UGC have not been covered in the UD annotation guidelines. In Sanguinetti et al. (2022), we provide a comprehensive overview of corpora, a discussion of the linguistic phenomena that cause difficulties in analysing user-generated texts, and unified recommendations for their treatment within the UD system of syntactic analysis.

2.4 Language Contact

Language contact refers to the ways in which languages influence one another, arising from multilingual interaction. The contact between the languages of English and Irish over centuries, and the dominance of English as a global language, has resulted in the vast majority of Irish-speakers being fluent in English (Stenson, 1993). It is, therefore, unsurprising that language contact phenomena would occur in this language pair, especially in an informal setting such as Twitter. The analysis of Irish-language tweets can provide us with valuable insights into the ways that the typologically different systems of English and Irish interact, helping us to understand the changing nature of contemporary informal Irish.

2.4.1 Terminology

Linguistic outcomes of language contact refer to the various changes that occur in languages when they come into contact with one another (Sankoff, 2004). The following language contact outcomes are common in Irish-language tweets.

- **Transliteration** is the spelling of a lexical item from one language using the orthographic conventions of another language. Transliteration is common in placenames of Ireland, which tend to be anglicised rather than translated. e.g. *Baile Beag*, literally meaning ‘little town’, is known as ‘Ballybeg’ in English. Words of English or other origins may also be Gaelicised. The language and grammar supplement of the Concise English-Irish Dictionary (Ó Mianáin, 2020) offers the examples *súisí* ‘sushi’, *cáirióice* ‘karaoke’ and *truip* ‘trip’.
- **Borrowing** can be defined as the transfer of a lexical item from one language into another e.g. *séipéal* ‘chapel’ is a loanword or borrowing derived from the Latin ‘chapele’ (Hickey, 2014). In this case, the spelling and pronunciation of the loanword have been adapted from the writing and sound system of the donor language, Latin, to suit that of the recipient language, Irish. This adaptation process is called **integration** or **assimilation**. The Irish word *craic* ‘fun’ (Lomas, 2017) is borrowed into Irish English, the dialect English spoken in Ireland of (Hickey, 2007)
- **Code-switching** is defined as alternation between languages within a single utterance e.g. *ag caint fúmsa, I suppose?* ‘talking about me, I suppose?’ (Ní Laoire, 2016). The first part of the phrase is in Irish and the second part is in English. In this case, no integration is observed. Each language maintains its spelling, morphology, and grammar in its respective phrase. The term ‘code-mixing’ can also be used to describe this phenomenon (Muysken, 2000) however some scholars, such as Bokamba (1989), distinguish between them.
- **Calquing** occurs when the structure of one language is directly translated into another. This can lead to the adoption of new expressions that mirror the syntax and structure of the source language but are used in the context of the borrowing language. The language and grammar supplement of the Concise English-Irish Dictionary (Ó Mianáin, 2020) provides an example of a literal translation from English: *an tuáille a chaitheamh isteach* ‘to throw in the towel’. This expression is now used in both languages to mean ‘to give up’. Hickey (2007) offers examples of grammatical structures derived from the Irish language such as ‘He’s **after** breaking the glass’ meaning ‘He has broken the glass’ and ‘He **does be** mending cars’ meaning ‘He (habitually) mends cars’.

2.4.2 Typologies and Frameworks

The field of contact linguistics has evolved over the decades through the development of various frameworks and typologies to model language contact from various perspectives such as grammar, sociolinguistics, and psychology.

Grammatical approaches Early research on the grammar of language contact includes the work of Whitney (1881) who suggested a hierarchy of borrowability whereby content words are more readily borrowed than function words. Haugen (1950) built on this foundation, developing an early typology using the term ‘substitution’ to refer to integrated forms of borrowing and ‘importation’ for unintegrated forms. Since Labov (1971) highlighted the idiosyncratic nature of code-switching and suggested that it may not conform to conventional sociolinguistic regularities, various language contact frameworks have been proposed to explain the systematic and patterned aspects of code-switching. For example, Pfaff (1979) indicated the avoidance of structural conflict between the grammatical systems of the languages in question. The seminal work of Poplack (1980) introduced the concepts of equivalence and free morpheme constraints, while also distinguishing between tag-switching and intrasentential code-switching. Myers-Scotton (1989) contributed the matrix language frame, a framework that highlights the asymmetric nature of code-switching and assigns the roles of ‘matrix language’ and ‘embedded language’. It is now agreed that code-switching is systematic rather than random, in that certain points in a clause can be identified by bilinguals as valid switch sites while other sites are deemed ungrammatical. However, the specific constraints that govern this phenomenon are disputed.

Sociolinguistic approaches Other research has approached language contact from a sociolinguistic perspective by focusing on variables influencing speech communities like population demographics, social functions, duration of language contact, linguistic profiles of the languages in question, as well as historical, cultural, socio-economic, and political factors. Blom and Gumperz (1972) have categorised code-switching as either situational or metaphorical. Similar to the concept of diglossia, situational code-switching refers to a change in language corresponding to a change in the social setting of the speaker, whereas metaphorical code-switching refers to a change in language which does not appear

to correspond to any change in social setting. Auer (1984) approached code-switching analysis in the context of the conversation in which it occurs. Various attitudes towards language contact phenomena tend to co-exist in a bilingual community. They can be seen as a signal of the deterioration or evolution of a language on a societal level, a useful tool to fill a linguistic gap, an indication of a lack of proficiency, an expression of creativity, a gesture of social inclusion/exclusion, or a demonstration of group membership (Gal, 1988; Blommaert, 1992). The interaction between the two languages involved in a switch can often be described in terms of the power and status associated with each language. Thus the choice of one language over another in a given context, subconscious or otherwise, communicates extra-linguistic information. In this way, a language may be considered unmarked where the choice of language is appropriate to the setting or marked when the language choice does not fit the situation (Myers-Scotton, 1995).

Psychological approaches Weinreich (1953) referred to language contact phenomena in general as ‘interference’ whereby adult learners process a new language through knowledge of their primary language. Giles et al. (1973) found that accommodation, a process whereby speakers adjust their speech to accommodate their audience, plays a role in word choice and in shaping interpersonal interactions, perceptions, and attitudes within a bilingual society. (Gollan and Ferreira, 2009) show that accessibility, how easily information is retrieved from long-term memory, is a factor that influences bilinguals’ choices to switch languages in spontaneous conversations. Bilinguals switch languages voluntarily when it does not significantly delay their response time, especially when switching to the non-dominant language is easy and does not compromise accuracy. The theory of translanguaging suggests that individuals, including bilinguals and multilinguals, draw from a single linguistic repertoire to communicate, rejecting the idea of distinct language systems (Vogel and García, 2017). This concept underpins the multilingual parsing experiment presented in Sections 5.3 and to multilingual NLP systems in general whereby a single model is capable of processing more than one language. The questionnaire study we present in Chapter 6 touches on some of these themes in an investigation of Irish speakers’ perceptions of word choice in the context of language contact.

By outlining the historical development of contact linguistics, from grammatical, so-

ciolinguistic, and psychological perspectives, we lay the theoretical foundation for the subsequent investigation, our questionnaire study focused on language contact phenomena in Irish-language tweets, presented in Chapter 6. This understanding of the evolution of contact linguistics background not only informs the motivation, design, methodology, and interpretation of the questionnaire study but also situates our research within its historical context and allows the reader to appreciate its multifaceted and interdisciplinary nature of language contact research.

2.4.3 NLP for Code-switched Data

According to Winata et al. (2023), research in the area of NLP for multilingual data is driven by three key factors: The majority of the world’s population being able to speak more than one language (Tucker, 2001), the need to process multilingual content from social media platforms, and demand for multilingual interaction with voice assistants, applications, etc.

NLP tasks for multilingual data such as language identification and code-switch point prediction have been approached using various methods, some based on linguistic constraints and statistical machine learning methods (Solorio and Liu, 2008; Li and Fung, 2012) and some more recent neural approaches using RNNs and BiLSTMs complemented by pretrained embeddings (Samih et al., 2016; Winata et al., 2019).

As a largely informal linguistic genre, language contact is frequent on social media (Bali et al., 2014). Mitigating the necessity for the recording and transcription of informal speech, UGC has recently become a popular data source for code-switched text in NLP tasks such as POS tagging (Jamatia et al., 2015), classification and visualisation of multilingual corpora (Guzmán et al., 2017), and dependency parsing (Bhat et al., 2017).

Various approaches have been taken with regard to the classification of language contact outcomes. Barman et al. (2014), for example, developed a dataset of Facebook posts and comments exhibiting language contact between Bengali, English, and Hindi. The data was labelled the token level using categories for each language, as well as categories for “mixed”, “universal”, and “undefined” words. However, they also note a degree of ambiguity in the annotation in that English words were sometimes labelled as Hindi or Bengali. Maharjan et al. (2015) and Çetinoğlu (2016) each perform language identification on a corpus of code-switched tweets at the token level using categories for each language

and categories for “named entities”, “ambiguous”, “mixed”, and “other”. Çetinoğlu et al. (2016) demonstrate the difficulty of the language identification task even for humans. They note that words could be considered by some to be in a foreign language while others believe the same word to be already integrated into the recipient language. Álvarez-Mellado and Lignos (2022) point out that not all other-language items are code-switches and introduce a methodology for language identification that includes a label for lexical borrowing which they apply to a corpus of Spanish tweets. In Chapter 6, we experiment with this methodology in the context of Irish-language tweets.

2.4.4 Code-Switching versus Borrowing

The motivation for distinguishing between code-switching and borrowing is evident in several areas. For example, the ability to describe and compare the frequency of code-switching in different contexts is useful for sociolinguistic research. The development of lexical resources can be enriched with an understanding of which lexical items are code-switched most often. In data-driven NLP, the ability to estimate how much of a multilingual data set is in a given language can facilitate the data selection and curation process. For these tasks to be conducted accurately, a methodology to classify code-switching is required. Without such a methodology, all instances of borrowing are likely to be considered code-switching, inflating and invalidating claims about the frequency of code-switching.

Various criteria have been used to distinguish borrowing from code-switching, however a lack of consensus persists on this topic. Whitney (1881) described loanwords as being assimilated into the borrowing tongue. Another potential property of loanwords is that they are recurrent and widespread (Poplack et al., 1989). A further possible criterion for identifying loanwords is ‘listedness’ (Muysken, 2000), the presence of the word in an established dictionary of the recipient language.

2.4.5 Language Contact in Irish

Studies have approached the contact linguistics of Irish from various perspectives. Bisagni (2014) and Stam (2017) have analysed historical written code-switching of Irish and Latin. From a sociolinguistic point of view, Atkinson and Kelly-Holmes (2011) analysed instances of English-Irish code-switching in a comedy radio show concluding that language

use reveals ambivalent attitudes towards language ownership and identity in Ireland. Fh-lannchadha and Hickey (2018) also explored the theme of ownership as well as authority in surveys with native and L2 Irish speakers.

Many studies of language contact in Irish have tended to focus on the grammatical aspects of spontaneous, naturally occurring speech. O'Malley-Madec (2007) examined intrasentential code-switching in two Irish-speaking communities. Treating all lone other-language items as borrowings, over 66% were found to be discourse markers and were 30% nouns. (Hickey, 2009) reported on the frequency and type of language contact among a group of adult native Irish speakers who were leaders of Irish-language preschools in Irish-speaking communities. The frequency of code-switching varied from 2.3% to 19.3% depending on whether the leaders were addressing children from monolingual Irish or bilingual home settings. English discourse markers (e.g. 'but', 'because', 'sure') were categorised as code-switches and borrowings based on their frequency and diffusion relative to their Irish equivalents. Moal et al. (2018) analysed linguistic features in the speech of presenters on an Irish-language radio programme *RíRá ar RnaG*, RTÉ Raidió na Gaeltachta. Despite the informal nature of the programme, they found that there was very limited code-switching. They consider that presenters may adhere to a more traditional variety of Irish on air than in casual speech in order to adhere to the perceived prescriptive linguistic stance of the broadcaster.

McCloskey (2017) provides the following Irish examples of nonce borrowings: *miss-áil*, *enjoy-áil*, *bother-áil*. Such constructions, which involve an English verb with an Irish gerund suffix, have elsewhere been classified as intra-word code-switching (Lynn and Scannell, 2019). Stenson (1991) acknowledges that this particular kind of construction is grammatically integrated but only minimally, and concludes that grammatical assimilation is insufficient as a diagnostic. Stenson (1993) concludes that distinguishing between code-switching and borrowing in modern Irish based on integration is challenging, as borrowings retain English phonological features due to universal bilingualism. She also notes that code-switching is increasingly prevalent among speakers of all ages and explains the difficulty, specifically within the language pair of Irish and English, of using phonological, morphological or syntactic assimilation as a measure of loanword integration into the recipient language.

2.5 Research Gaps

Having provided in the previous sections of this chapter background to the main topics of the thesis, in this section, we identify the research gaps to be explored in the subsequent chapters.

Irish-language technology resource development While acknowledging the limitations of existing resources, Section 2.1.4 describes progress in the development of Irish-language technology. We identify a research gap in the potential for the creation of new resources and the expansion of those already available. For example, resources such as word embeddings and datasets can be enriched with increased coverage of different domains such as UGC. Such research could enable the language to leverage developments in language technology more effectively.

Unified recommendations for UGC treatment in NLP In Section 2.3.2, we highlight the popularity of social media text as a data source in NLP research, and a recent attempt to unify the representation of UGC within the UD framework (Sanguinetti et al., 2022). The continued development of standardised guidelines for annotating and processing UGC constitutes an important research gap. Work in this area contributes to the relevance of resources and promotes consistency and comparability across studies.

NLP capabilities for Irish UGC As discussed in Section 2.1.4, experiments have been carried out involving POS tagging for Irish UGC (Lynn et al., 2015; Lynn and Scannell, 2019). We highlight a research gap in that many NLP tasks such as dependency parsing and named-entity recognition have not been implemented in the context of Irish language UGC. Research might focus on refining models and tools to better handle diverse linguistic characteristics present in UGC or it may investigate the robustness of language models like `gaBERT` to such linguistic variations and propose techniques to improve their performance.

Language Contact A salient feature of Irish language UGC and a challenging aspect of NLP, Section 2.4 provides background on the topic of language contact as it relates to the current research. One research gap in this area relates to developing a robust methodology or set of criteria for distinguishing between code-switching and borrowing in the context of the language pair of Irish and English. Such a methodology would enable a more nuanced

analysis of the frequency and patterns of interactions between Irish and English. The development of such a methodology may need to take into consideration sociolinguistic attitudes towards language contact as held by different language communities of Irish speakers, including native Irish speakers and those with varying levels of proficiency.

2.6 Summary

Despite the privileged position that Irish holds relative to many minority languages and the resources developed for Irish-language technology, Irish remains an endangered language and could benefit from linguistic resources among other interventions. While we have shown that there has been significant progress towards developing language technologies for Irish, we highlight a research gap in the form of Irish-language technology resource development. This gap forms the primary motivation for our work.

We have introduced the NLP task of dependency parsing, describing the theoretical tradition of dependency grammar and explaining concepts such as dependency trees and treebanks. We have explained the motivations for using a UD-based approach in analysing Irish-language tweets, highlighting the advantages of representing Irish syntax and maximising interoperability and resource-sharing. We have justified our use of a graph-based neural parsing architecture and we have explained the important role of contextualised word embeddings in our research.

As the data we work with falls under the broad category of UGC, we described the various opportunities and challenges that UGC poses for NLP. We refer to the growing body of research that has explored this and specifically highlight evidence to support the demand for unified recommendations for UGC treatment in NLP and the improvement of NLP capabilities for UGC.

We provide background on the topic of language contact, particularly in the context of Irish and English, and suggest analysis of Irish-language tweets as a valuable avenue of research. We introduce the terminology of language contact outcomes relevant to Irish-language tweets and the typologies and frameworks in the field of contact linguistics that provide lenses through which language contact has been viewed. Given the frequency of language contact globally, we highlight the demand for NLP for multilingual data, particularly in the context of UGC. Additionally, the distinction between code-switching and

borrowing is addressed, emphasising the importance of developing methodologies for accurate classification. Finally, we outline the research gaps to be addressed in the following chapters and highlight the implications of this work.

Chapter 3

TwittIrish Treebank Development

This chapter details the procedural methodology employed in the creation of TwittIrish (Cassidy et al., 2022), a novel resource for both NLP and linguistic research. TwittIrish is a UD treebank comprising 2,596 Irish-language tweets (47,790 tokens). Motivated by the lack of an accurate parser for Irish UGC, TwittIrish is a valuable, genre-specific resource that can be used to enhance parsing accuracy for Irish UGC, to facilitate experimentation with other NLP tasks, and to enable in-depth linguistic analysis. Given the distinct linguistic features of UGC compared to standard text, and the lack of a universally accepted annotation scheme for these features, our work in Sanguinetti et al. (2022) involves proposing such annotation guidelines. These guidelines were closely adhered to during the creation of TwittIrish, and are outlined throughout the chapter. The full TwittIrish guidelines are provided in Appendix B. Each section of this chapter corresponds to a phase of the treebank development pipeline. Figure 3.1 illustrates the methodological trajectory. Furthermore, Appendix A provides a comprehensive data statement for TwittIrish.

Section 3.1 describes the curation of the tweets used in TwittIrish. Section 3.2 details the preprocessing and conversion steps carried out to prepare the data for syntactic annotation. Section 3.3 describes the syntactic annotation cycle. Finally, Section 3.4 details the quality-check phase of the treebank development.

3.1 Data Curation

All of the tweets in the TwittIrish treebank were sourced via Indigenous Tweets a project that compiles statistics on social media data of 185 minority and indigenous languages

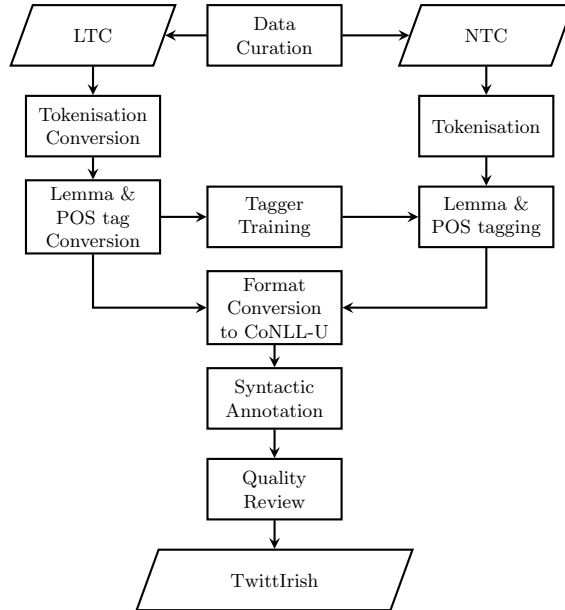


Figure 3.1: The TwittIrish creation process. The corpora LTC and NTC are the sources of the treebank data.

including Irish.¹ At the time of writing (August 2023), Indigenous Tweets has identified approximately 4.9 million tweets in Irish. Table 3.1 provides an overview of the Lynn Twitter Corpus (LTC) and the New Twitter Corpus (NTC), the two corpora of Irish-language tweets that we used as source data for the TwittIrish treebank.

Dataset	Date	Tokenised	Lemmatised	POS tagged	Parsed
LTC	2009-2014	✓	✓	✓	×
NTC	2010-2019	×	×	×	×

Table 3.1: Metadata of TwittIrish source data.

In order to leverage existing resources, we included 1,299 tweets from a corpus of 1,493 tweets that had previously been tokenised, lemmatised, POS tagged (Lynn et al., 2015) with a specialised POS tag set for Irish-language tweets based on that of Gimpel et al. (2011) for English language tweets. The LTC was also later annotated with code-switching information (Lynn and Scannell, 2019). The LTC tweets, randomly sampled from tweets by 8,000 users who had tweeted in Irish, were posted between the years 2009 and 2014.

Additionally, we included more recent tweets sampled from more users to make the treebank more diverse and up-to-date. We refer to this newer Twitter corpus as NTC. The 1,297 tweets in NTC, randomly sampled from 25,000 tweets by 14,111 users, were posted between 2010 and 2019. The specific number of tweets included in the final dataset

¹<http://indigenoustweets.com/>

was not chosen specially but is a result of our aim of including as many tweets as possible within a given time frame and balancing the amount that came from each source dataset. Usually, a training set would be larger than the test and development set and so, in Section 3.5, we provide an explanation as to why that is not the case in this dataset.

Set	LTC	NTC	Total
Test	700	166	866
Development	100	764	864
Training	499	367	866
Total	1299	1297	2596

Table 3.2: Dataset sizes.

Table 3.2 shows the number of tweets from LTC and NTC in the final TwittIrish test, development, and training sets. Any duplicate or non-Irish tweets were excluded from the final datasets.

In our data curation, we attempted to mitigate bias by using a random sample of tweets, however, we acknowledge that some users are overrepresented in the dataset due to “Participation Inequality” (Duval and Ochoa, 2008), whereby users generate content disproportionately. For example, based on a sample of 2,596 tweets from the NTC data for which we have user IDs, most users in the dataset contributed a single tweet, whereas less than 3% of users contributed 10 tweets or more and just two users contributed 100 tweets or more.

3.2 Data Preprocessing and Conversion

The LTC had been previously tokenised, lemmatised, and POS tagged. As such, a conversion process was required to map the data to UD conventions and Irish-specific conventions as defined by Lynn and Foster (2016) for the IUDT. Irish-language examples and detailed annotation guidelines detail how to apply the general framework of UD to the specific context of Irish.² This conversion process involved both automatic and manual adjustments. The NTC consists of a dataset of tweets in plain text format. As such, preprocessing was required in the form of tokenisation, lemmatisation, and POS tagging using tools trained on the converted LTC data.

²<https://universaldependencies.org/ga/index.html>

3.2.1 LTC Tokenisation Conversion

In Sanguinetti et al. (2022), we discuss the challenges of tokenising informal text. For example, contractions like ‘gonna’, representing the two words ‘going to’, should not be split. Similarly, acronyms like ‘TL;DR’ in which each character represents a separate word of the phrase ‘too long; didn’t read’ should not be split. However, conventionally separate tokens that seem to have been merged accidentally can be split, e.g. ‘goingto’ should be split into the tokens ‘going’ and ‘to’. Such decisions were made by taking into consideration treebank consistency, accurate linguistic representation, and annotation effort.

When converting the LTC data to be compatible with UD and the recommendations of Sanguinetti et al. (2022), the most notable difference was in the treatment of MWEs. In LTC, the individual tokens of MWEs were fused with an underscore. Such an approach is not permitted in UD which keeps tokens separate on a tokenisation level but connects them on a syntactic level.³ Several minor differences were also observed between the two tokenisation schemes such as whether or not certain symbols, abbreviations or punctuation marks should be merged with 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, while in LTC they are combined. We resolved to manually separate such occurrences in the TwittIrish tokenisation scheme.

3.2.2 LTC Lemmatisation and POS Tag Conversion

The lemmatisation of user-generated text is typically guided by UD guidelines related to morphology, which can be straightforward to apply. Only minor manual adjustments were required for lemmatisation to ensure alignment with the IUdT.

For the various tokens and symbols associated with UGC, we adhere to the suggestions of Sanguinetti et al. (2022).

- **At-mentions, handles, or usernames** are tagged as PROPN.
- **Hashtags** are tagged with the tag they would otherwise have without the hashtag symbol. e.g. because *madra* ‘dog’ is tagged as NOUN, *#madra* should also be tagged as NOUN. Multiword hashtags are kept as a single token and assigned the POS tag of the head word.

³<https://universaldependencies.org/v2/mwe.html>

- **Pictograms, emojis, and smileys** are tagged as **SYM**.
- **retweet (RT) symbols** are tagged as **SYM**.
- **URLs** are tagged as **SYM**.

Recommended annotation strategies vary based on each element’s syntactic, semantic, and contextual properties, ensuring consistency and clarity in UGC annotation. Finally, the POS tagset used in the LTC was converted to the UD POS (UPOS) tagset as shown in Table 3.3. 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 (i.e. **SCONJ**, **CCONJ**, **VERB**, **AUX**) automatic identification and manual correction were performed.

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	any †
EN	any †
G	any †

Table 3.3: POS tag mapping.

* Many-to-one relation

† One-to-many relation

Table 3.4 demonstrates the mapping of a sample tweet from the LTC to the UD scheme. As all English language tokens were annotated with a single tag ‘EN’ in the LTC scheme, these tags were converted to the appropriate UPOS tags in TwittIrish.

Table 3.5 shows that using the LTC POS tagset, all verbs are tagged **V**. As previously described in Section 2.1.3, Irish has two verbs corresponding to the English verb ‘to be’. According to UD, the Irish copula (e.g. *is* ‘is’, *ní* ‘is not’) should be tagged as **AUX** distinguishing it from the substantive verb (e.g. *tá* ‘is’, *níl* ‘is not’) which are tagged **VERB**.

Surface	LPOS	UPOS
@user	@	PROP
#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
‘@user #cutie where is the love you ask’

Table 3.4: Example Irish tweet with LTC and corresponding universal POS tags.

Surface	LPOS	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 cáilúil é ach táim bródúil #Grá
‘He is not a celebrity but I’m proud #Love ’

Table 3.5: Example Irish tweet with LTC and corresponding universal POS tags.

Where conflicts between the annotation scheme of LTC and IUdT were observed, consistency with the IUdT was preferred as the IUdT annotation scheme has been in development since the first Irish UD treebank was introduced and so the data and annotation guidelines are regularly updated and debugged. This consistency between TwittIrish and IUdT was necessary to leverage the IUdT as training data in the initial stages of syntactic annotation (detailed in Section 3.3).

3.2.3 NTC Tokenisation

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),⁴ a language-agnostic trainable pipeline for tokenisation, tagging, lemmatisation and dependency parsing, and TweetTokenizer⁵ from NLTK (Bird et al., 2009), a rule-based tokeniser designed for noisy UGC. The TweetTokenizer is specifically tailored for tokenising textual content from social media, with a focus on tweets by employing a combination of regular expression patterns to handle genre-specific features encountered in the context of tweets, including URLs, emoticons, hashtags, and user mentions. We carried out a brief examination of the outputs of both systems in order to compare them. As exemplified in Table 3.6, NLTK TweetTokenizer was more effective at tokenising the UGC phenomena such as emoticons, URLs and meta-language tags that are frequent in tweets. We chose to tokenise the NTC tweets using the TweetTokenizer for this reason. Manual corrections were then applied in order to adhere to the Irish-specific tokenisation scheme within current UD guidelines. Table 3.6 provides

⁴Trained on IUdT v2.8 with no pre-trained embeddings.

⁵<https://www.nltk.org/api/nltk.tokenize.html>

an example of tokenisation by UDPipe 1 trained on IUDT v2.8 compared to the NLTK TweetTokenizer.

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

Dé Céadaoin #Midweek #Beagnachann :) :)
 ‘Wednesday #Midweek #Almostthere :) :)’

Table 3.6: Example Irish tweet tokenised by UDPipe 1 trained on IUDT version 2.8 and NLTK TweetTokenizer.

3.2.4 NTC Lemmatisation and POS Tagging

To establish the best system to use for automatic lemmatisation and POS tagging, we tested two tools, Morfette (Chrupała et al., 2008), a probabilistic lemma and POS tagger that uses a Maximum Entropy classifier for supervised learning of inflectional morphology, and UDPipe (Straka et al., 2016), a lemmatiser and POS tagger using MorphoDiTa, a supervised averaged perceptron neural network that utilises a rich feature set. Both systems were trained on a merged dataset of the converted LTC training data and the entire IUDT and tested on the converted LTC test data. Morfette achieved a lemmatising accuracy of 88.87% and 93.24% for POS tagging, outperforming UDPipe which achieved 88.41% for lemmatising and 87.68% for POS tagging. Therefore Morfette was used to lemmatise and POS tag the NTC tweets.

CoNLL-U									
# sent_id = X									
# text = Cuirfidh mé DM chuici									
CoNLL-U	Morfette			CoNLL-U					
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Cuirfidh	cuir	VERB	-	-	-	-	-	-
2	mé	mé	PRON	-	-	-	-	-	-
3	DM	DM	NOUN	-	-	-	-	-	-
4	chuici	chuig	ADP	-	-	-	-	-	-

Table 3.7: Example conversion of Irish tweet from Morfette to CoNLL-U format ‘I will send her a DM’.

3.2.5 Conversion to CoNLL-U Format

Both the LTC and NTC were converted automatically from the 3-column Morfette format, consisting of the token, lemma, and POS tag, to the 10-column Conference on Computational Natural Language Learning (CoNLL) format for Universal Dependencies (U) (CoNLL-U) format, as demonstrated in Table 3.7. CoNLL-U, widely used in NLP consists of a plain-text representation of a sentence’s dependency tree structure, where each token is described on a separate line with various linguistic attributes, i.e. a token id (ID), word form or token (FORM), lemma or base dictionary form (LEMMA), universal part-of-speech (UPOS), language-specific part-of-speech tag (XPOS), morphological features (FEATS), head of the current word (HEAD), dependency relation between the token and its head (DEPREL), an enhanced dependency graph as described in Section 2.2.4 (DEPS), and any other miscellaneous information annotators wish to capture (MISC).

In order to make optimum use of the time spent by the sole 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 also requires a sentence ID and the original raw text to be included preceding the annotation.

3.3 Syntactic Annotation

Segmentation In the syntactic analysis of standard text, the conventional unit of analysis is the sentence. However, segmenting text into sentences poses a challenge in the context of diverse linguistic sources such as spoken language transcriptions and UGC on social media platforms. In conventional written texts, sentence boundaries are typically determined by punctuation, but this approach fails when applied to non-standard text. In spoken language, the concept of a sentence is debatable and human-performed segmentation is inconsistently applied across text types. When considering UGC from social media, the inconsistent use of punctuation poses difficulties for both manual and automatic segmentation. In the work Sanguinetti et al. (2022), we highlight distinct segmentation approaches employed in various treebanks and suggest segmenting data into sentences where feasible using a subtype of the `parataxis` label, `parataxis:sentence`, as demonstrated in Figure 3.2. This approach aligns with annotation practices for stan-

dard written language and facilitates any cross-dataset comparisons, indicating potential sentence boundaries within tweets. This approach aids in identifying segmentation points and distinguishing them from other forms of parataxis, ensuring connectivity between multiple sentential units within a tweet.

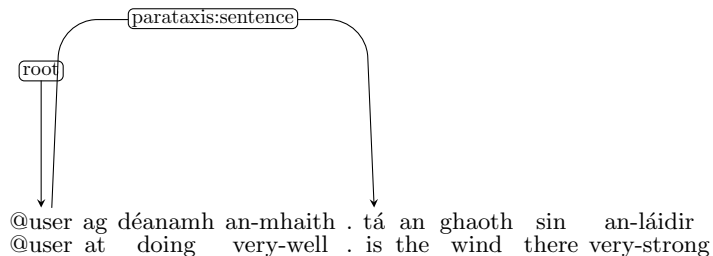


Figure 3.2: Attachment of sentences within tweets via `parataxis:sentence` ‘@user is doing very well. The wind is very strong’.

3.3.1 Annotation of Genre-specific Features

The task of syntactic annotation involved defining binary relations between head tokens and their dependents, using a fixed set of relations prescribed by UD, as described in Section 2.2. We also employ subtypes of the UD dependency relations for the specific case of Irish-language annotation (Lynn and Foster, 2016) and further subtypes for the case of UGC as recommended by Sanguinetti et al. (2022):

- **At-mentions, handles, or usernames** are attached via `vocative:mention`
- **Hashtags** are attached to the head of the relevant phrase via `parataxis:hashtag`.⁶
- **Pictograms** are attached to the head of the relevant phrase via `discourse:emo`.
- **Retweet markers** are attached to the root of the tweet via `parataxis:rt`.
- **URLs** are attached to the head of the relevant phrase via `parataxis:url`.

In the case that any of the above items plays a syntactic role in the sentence, it should instead be attached to appropriately represent that role. Language identification was also performed at the token level. In the 10th (miscellaneous) column of the CoNLL-U format, the annotation `Lang=ga` was used for Irish words and `Lang=en` was used for English words. Proper nouns, metalanguage tags, and punctuation received no language

⁶An earlier version of these guidelines recommended that all hashtags be tagged as X. At the time of writing, this update has yet to be applied to TwittIrish.

annotation. When the language of a word was perceived as ambiguous by the annotator, it was annotated as Irish only if it was listed in the New English-Irish Dictionary (NEID). Further exploration of language identification and language contact between English and Irish is provided in Chapter 6.

3.3.2 Bootstrap Annotation Cycle

As a method shown to reduce manual annotation efforts in this task (Judge et al., 2006; Seraji et al., 2012), we carry out a bootstrapping approach to syntactic annotation as recommended by UD.⁷ This process is illustrated in Figure 3.3.

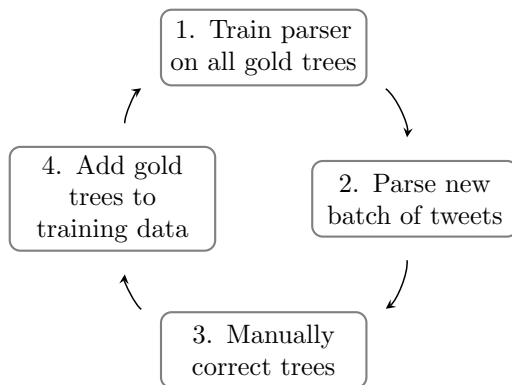


Figure 3.3: Bootstrapping approach to semi-automated syntax annotation.

Manual annotation of a small set of 166 tweets was carried out so that the annotator could learn the process and establish a seed training set. The annotator was in regular contact with experienced annotators so that issues could be resolved in the initial stages, preventing error propagation. This manually annotated data and the IUDT were used to initiate the bootstrapping cycle. During this process, the biaffine parser (Dozat and Manning, 2017) was tested with different encoders such as Multilingual BERT (Devlin et al., 2019) and wikiBERT. Ultimately, we chose monolingual Irish embeddings of gaBERT based on findings by Barry et al. (2022) that they outperform multilingual embeddings when tested on IUDT.

Step 1 A parsing model was trained on IUDT in combination with the newly annotated tweets.

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

Step 2 The parsing model was used to automatically annotate the next batch of 100 POS-tagged tweets with syntactic information.

Step 3 These parsed tweets were then manually corrected by the sole annotator. Any bugs or inconsistencies identified by the annotator were discussed and corrected where possible.

Step 4 The corrected tweets were then added to the training data.

Steps 1 to 4 were repeated until the deadline of the UD version 2.8 data freeze (1 May 2021), ensuring that the dataset remained consistent and stable for its release. At this point, 866 tweets (15,433 tokens) were fully parsed.

3.4 Quality Review

In order to assess the accuracy of the dependency annotation by the sole annotator, a randomly selected subset of the annotated data, consisting of 46 trees (773 tokens), was reviewed for errors by another Irish speaker experienced in linguistic annotation. The task of the reviewer was to flag potential errors in the form of a token with an incorrect head and/or label. 46 potential errors were identified by the reviewer. The potential errors were then discussed by a team of two expert annotators to confirm whether the potential errors were true errors. 32 potential errors were confirmed as true errors and the other 14 that had been flagged as potential errors were determined to be correct but highlighted areas where improvements could be made to the annotation guidelines to clarify or disambiguate details.

The overall error rate per token of the treebank annotation can be estimated as 0.004 by dividing the number of incorrectly annotated tokens by the total number of tokens in the review, as shown in equation (3.1). This means that approximately 4% of tokens in the review were annotated incorrectly. Following the methodology outlined by Mikulová and Štěpánek (2009), we also calculated the error rate per tree (or tweet) of the annotation. The tweet error rate can be estimated as 0.7 by dividing the number of incorrectly annotated tokens by the total number of tweets in the review, as shown in equation (3.2). This means that approximately 70% of the tweets in the review contained an annotation

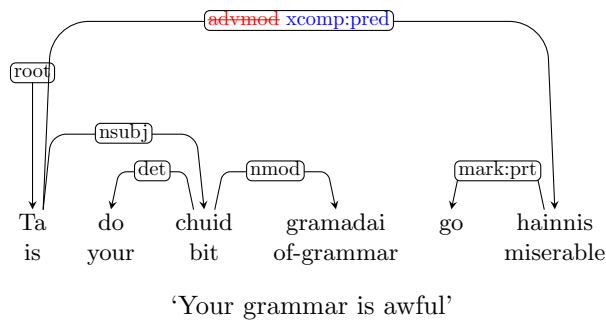


Figure 3.4: Parsed tweet with incorrect label and correct head corrected during review.

error.

$$\text{Token Error Rate} = \frac{\text{Number of Errors}}{\text{Number of Tokens in the Review}} \quad (3.1)$$

$$\text{Tweet Error Rate} = \frac{\text{Number of Errors}}{\text{Number of Tweets in the Review}} \quad (3.2)$$

This was a useful process for establishing both common errors made by the parsing model that had been missed by the annotator, and human errors. The annotation guidelines were then refined based on this information.

Correction type 1: Incorrect label, correct head 16 tokens (2.07% of all tokens in the review) had an incorrect label and correct head. Figure 3.4 exemplifies one such correction. *Go* is a common particle in Irish, which can precede an adjective to create an adverb. When used for this function it is roughly equivalent to the suffix ‘-ly’ in English, e.g. *ainnis* (‘miserable’), *go hainnis* (‘miserably’). For this reason, a parser is likely to annotate this construction as `advmod`. However, these constructions also appear as the complement of the substantive Irish verb *bí* ‘to be’ and in this case, they should be considered as `xcomp:pred`, as shown in Figure 3.4.

Correction type 2: Correct label, incorrect 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 incorrect head and label.

Correction type 3: Incorrect label and head Figure 3.5 shows the phrase *maith sibh* (‘good on you’) incorrectly annotated with *sibh* as the `root` and *maith* as an `amod` (adjectival modifier). It was identified in the review that *maith* should be considered the



Figure 3.5: Incorrectly annotated tweet and corrected version.

adjective predicate of an elided copula (Stenson, 2019). The full phrase is thought to be *is maith sibh* and the corrected annotation is also shown in Figure 3.5.

3.5 Data Releases

By the UD version 2.8 release deadline (15 May 2021), 866 tweets had been fully parsed. The parsed tweets were then validated as required by UD, using the UD validation script which highlights any automatically-detectable errors. Manual corrections were applied to the data until it passed all the checks, ensuring that the data met the standards of UD. The validated tweets were released as part of UD version 2.8 as a test set as recommended by UD when a treebank contains less than 20,000 words. At this point, the parsing model described in Section 3.3 had reached sufficient accuracy that it was not necessary to retrain the parser as frequently.

Four more bootstrapped iterations of the parser were used in the remainder of the annotation process. This work continued intermittently until the UD version 2.12 data freeze (1 May 2022) at which point a total of 2,598 tweets were fully parsed. All tweets were then anonymised, i.e. usernames, email addresses, and phone numbers were replaced with anonymous strings so that nobody would be identifiable. During the validation process, two tweets were removed because one contained the exact same text as another tweet in the dataset, and the other had no Irish-language words. The remaining 2,596 tweets were released with UD version 2.12, keeping the original test set of 866 tweets as a test set, while adding a development set of 864 tweets, and a training set of 866 tweets.

3.6 Summary

This chapter has described the development of TwittIrish, a UD treebank for Irish-language UGC. We have explained our annotation methodology for the linguistic features of UGC, such as hashtags, at-mentions, and emojis. We have also detailed the systematic and iterative annotation process that can serve as a blueprint for the creation of similar resources for other lesser-resourced languages. Through manual annotation, automation, and iterative improvement, a high-quality treebank was generated and released. TwittIrish provides up-to-date insight into Irish use in an informal context and has several potential applications. In the context of a low-resource language like Irish, it is especially important to be able to leverage existing data. In this sense, TwittIrish can be used for future language technology evaluation and development that harnesses UGC. TwittIrish also has potential applications in linguistic research, language documentation and revitalisation offering a rich source of data to explore language variation, syntactic structures, cross-linguistic comparisons, and language evolution over time. Ultimately, as the first treebank of Irish-language UGC, TwittIrish is a valuable resource with applications in NLP and linguistics.

Chapter 4

Linguistic Analysis of Irish Language Tweets

This chapter addresses RQ2 ‘How do Irish tweets differ from standard edited Irish text?’ by exploring the linguistic features of Irish-language tweets from the TwittIrish treebank in comparison to standard Irish text from the IUDT treebank. We consider standard Irish text as following *An Caighdeán Oifigiúil* ‘The Official Standard’ (Oireachtas, 2017). The motivation for our analyses is to facilitate investigation into the challenges of parsing Irish social media text, as explored in Chapter 5.

Despite the recent advancements in NLP that have reduced the need for in-depth domain knowledge and manual feature engineering, it is still valuable to have a grasp of the specific linguistic genre being worked with. Understanding the linguistic nuances of the genre helps to better interpret the results of “black box” models. Having this understanding ensures that the outcomes are correctly understood and prevents misinterpretation of the results that might occur if solely relying on automated metrics. Linguistic analysis facilitates in anticipating challenges, understanding the context and biases of the data, and refining the model for better performance, ultimately enhancing its quality and reliability.

Section 4.1 details the orthographic or spelling variation often present in Irish-language tweets. Section 4.2 examines the morphological variation observed in Irish-language tweets. Section 4.3 describes the differences in the lexicon or vocabulary of Irish-language tweets as compared to that of standard Irish. Section 4.4 explores the grammatical variation common to Irish-language tweets. Each section provides examples and a discussion of the linguistic phenomena mentioned.

4.1 Orthographic Variation in Irish Tweets

The examples described in this section pertain to tokens that are part of the standard Irish lexicon but that we have observed to deviate from the conventional spelling system of the language in the context of tweets. This variation can affect the lemmatisation of a token in an NLP pipeline, potentially affecting other downstream tasks. In our TwittIrish sample, 2.5% of tokens contained some such orthographic variation. We classify these occurrences into the following categories: Diacritic variation, abbreviation, lengthening, nonstandard capitalisation, punctuation variation, hypercorrection, and other spelling variation.

Diacritic variation The acute accent or *síneadh fada* is used in Irish to indicate a long vowel and is necessary to disambiguate between certain words. Diacritic marks are often omitted or incorrectly added to tweets. Example 4.1 shows the most probable intended word *léacht* ‘lecture’ rendered as *leacht* ‘liquid’. Example 4.2 shows a diacritic mark incorrectly added to the word *am* ‘time’ resulting in the meaningless token **ám*.

(4.1) *Leacht faoi stair*
Léacht faoi stair
‘A lecture about history’

(4.2) *Ag an ám seo den oíche*
ag an am seo den oíche
‘at this time of night’

Abbreviation Predictable shorthand forms can occur in standard Irish texts, e.g. *lch* is an abbreviated form of *leathanach* ‘page’. These and other more unconventional, and thus less predictable, abbreviations are observed in Irish tweets. Example 4.3 shows the word *seachtain* ‘week’ shortened to *seacht* ‘seven’. Example 4.4 shows the words *fhoireann* ‘team’ and *hÉireann* ‘Ireland’ shortened to *fhoir* and *hÉir* respectively. Abbreviations are more common in tweets than standard text as the character limit and real-time posting nature of the platform encourage the user to be efficient with respect to time and space.

(4.3) *Bím de ghnáth ach sa bhaile an tseacht seo*
Bím de ghnáth ach sa bhaile an tseachtain seo
‘I am usually but home this week’

(4.4) *ar fhoir rugbaí na hÉir*
ar fhoireann rugbaí na hÉireann

‘on the Irish rugby team’

Lengthening The converse phenomenon of abbreviation, whereby a token is elongated by repeating one or more characters, is also a salient feature in Twitter text. This can be considered an encoding of sociophonetic information (Tatman, 2015). Despite incentives to save time and space while tweeting, users often elongate certain words for expressive purposes (Brody and Diakopoulos, 2011). Example 4.5 shows elongation of the word *buí* ‘yellow’. Similarly, Example 4.6 demonstrates the lengthening of the word *mór* ‘big’. The repetition of characters is used to encode information about the emphasis and rhythm of the spoken utterance.

(4.5) *tá siad go léir buuuuuuúí*
tá siad go léir buí
‘they are all yellow’

(4.6) *ag gáire go mórrrrr*
ag gáire go mór
‘laughing a lot’

Nonstandard capitalisation 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. Example 4.7 shows the phrase *ar domhain* ‘on earth’. Similarly, Example 4.8 shows the capitalisation of the word *breá* ‘lovely’. This kind of formatting is used to emphasise the words in uppercase.

(4.7) *Níl todhchaí na Gaeilge sa Ghaeltacht, ach in aon áit **AR DOMHAIN***
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’

(4.8) *is **BREÁ** le daoine áirithe é*
is breá le daoine áirithe é
‘certain people love it’

Punctuation variation Punctuation is often 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. Example 4.9 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, indeed, a smiley e.g. ‘:-)’). Example 4.10 demonstrates punctuation being used to add style or formatting to the word *folúntas* ‘vacancy’ perhaps as a way to make the text stand out.

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

(4.10) ***folúntas***
 folúntas
 ‘vacancy’

Transliteration Common to language contact situations, transliteration occurs when a word in one language is rendered using the writing system of another. An instance of transliteration, in which an Irish word is rendered using the writing system of English is shown in Example 4.11. The Irish word *raibh* ‘was’ is replaced with *rev*. Similarly, in Example 4.12, the Irish word *bhfuil* is replaced with *wil*. Both cases reflect the pronunciations of the words following the English spelling system.

(4.11) An *rev* foireann acu
 An raibh foireann acu
 ‘Did they have a team’

(4.12) *Déarfainn go wil*
 Déarfainn go bhfuil
 ‘I’d say there is’

Hypercorrection Orthography is sometimes corrupted by hypercorrection when auto-correct software is enabled in a language other than the user’s language of choice. As a result, attempts to type a word are corrected to a token with a similar spelling in another language. Example 4.13 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 4.14 shows *agus* ‘and’ rendered as *agua* which may have occurred due to hypercorrection as *agua* ‘water’ is a frequent token in other languages such as Portuguese and Spanish. However, it could also be a simple typo.

(4.13) *Mhúscail mé i mo leaba féin ar maidin i ndiaidh **concise** mór*
Mhúscail mé i mo leaba féin ar maidin i ndiaidh coicíse mór
 ‘I woke up in my own bed after a big fortnight’

(4.14) *tá an teanga ag fáil bháis **agua** níl ach uaireanta*
tá an teanga ag fáil bháis agus níl ach uaireanta
 ‘the language is dying and there are only hours’

Other spelling variation Any form of orthographic variation that cannot be classified in the above categories is considered here. These are mostly slight variations very close to the intended word and may occur due to typographical errors. 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 4.15 shows *sraith* ‘season’ rendered as **stait*. 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. Example 4.16 shows the vowels of the word *bhuel* ‘well’ transposed. This is another common variety of typographical error.

(4.15) **stait** 6 de Imeall
sraith 6 de Imeall
 ‘season 6 of Imeall’

(4.16) **bheul** gan dabht
 bhuel gan dabht
 ‘well no doubt’

4.2 Morphological Variation in Irish Tweets

In this section, we explore morphological variation in Irish-language tweets. Where the previous section on orthographic variation examined the arrangement of **characters** in a word, here we analyse the arrangement of **morphemes**, the smallest meaningful units of words. Specifically, we focus on morphological variation due to language contact (mixed-language tokens) and differences in Irish dialect (dialectal morphology).

Mixed-language tokens 66.74% of tokens in our TwittIrish sample are in Irish, 4.85% of tokens are in English, and the remainder (consisting of punctuation, meta language tags, etc.) are classified as neither, or indeed both in the case of mixed-language tokens. Both

Caomhánach (2022) and Lynn and Scannell (2019) note the propensity of Irish speakers to conjugate an English language verb with the Irish gerund suffix *áil*. Example 4.17 demonstrates an Irish utterance that uses the English verb root ‘happen’ instead of the Irish equivalent *tarlaigh*. Such mixed-language tokens constitute a point of controversy in the language contact literature as described in Section 2.4.4. McArthur (1998) would classify our Example 4.17 as intra-word code-switching whereby “a change occurs within a word boundary”. Poplack et al. (1988), however, might call Example 4.17 a nonce borrowing given that it behaves like a borrowing insofar as it is morphosyntactically integrated into the host language but is not an established loanword. Example 4.18 shows the Irish word *leaid* ‘lad’ with the English plural suffix ‘-s’. Whether indeed *leaid* should be classified as an Irish word in this case, as opposed to a transliteration of the English word ‘lad’ using the Irish spelling system, is a topic further explored in Chapter 6.

(4.17) *Eachtra i ndiaidh **Happenáil** i nGaoth Dobhair*
*Eachtra i ndiaidh **tarlú** i nGaoth Dobhair*
 ‘An event after happening in Gweedore’

(4.18) ***leaid***
leaideanna
 ‘lads’

Dialectal morphology Figure 4.1 shows semantically equivalent statements rendered using the synthetic and analytic verb forms. The synthetic verb form, more common to the Munster dialect, incorporates the subject in the verb ending. In Figure 4.1 the verb is conjugated with the first person singular synthetic verb form ending *-(e)as* whereas in the analytic construction, the subject pronoun *mé* appears as a separate token, resulting in two different tree structures.



Figure 4.1: Synthetic and analytic verb forms ‘I got 11’.

4.3 Lexical Variation in Irish Tweets

Parsers trained on standard text often encounter unfamiliar tokens when processing Twitter data due to its heterogeneous vocabulary. Based on our sample, we estimate that just 38.32% of the set of unique lemmata that make up the vocabulary of our TwittIrish sample occur in the IUdT training data.

Dialectal vocabulary Irish has three major dialects; Connaught, Munster, and Ulster. The visibility of the distinctive characteristics of spelling and grammar unique to each dialect have been somewhat diminished by the standardisation of Irish (Hickey, 2011). However, distinctive features of these dialects in the form of lexical variation are still evident in spoken language and informal text such as tweets. Example 4.19 shows the use of *domh*, the Ulster variant of *dom* meaning ‘to me’. Example 4.20 shows the word *arís* ‘again’ rendered as *aríst*, a variation more common to the Munster dialect.

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

(4.20) *caithfidh mé fanacht **aríst***
caithfidh mé fanacht arís
‘I have to wait again’

Initialism In Irish tweets, multiword phrases are frequently represented by the initial letter of each of their constituent tokens for the sake of brevity. Example 4.21 shows *GRMA* ‘Thank you’ used to represent its expanded form *Go raibh maith agat*. Example 4.22 represents the phrase *buíochas le Dia* ‘thank God’ as *BLD*.

(4.21) *Scaip an scéal! **GRMA!***
*Scaip an scéal! **Go raibh maith agat!***
‘Spread the word! Thank you!’

(4.22) *tirim ar maidin i gConamara **BLD***
tirim ar maidin i gConamara buíochas le Dia
‘dry this morning in Connemara thank God’

Pictogram Emojis, emoticons, etc. can be added to a text to emulate gestures (Gawne and McCulloch, 2019) or they may play a syntactic role in a phrase, replacing a word

as in Example 4.23, in which the symbol ‘♥’ acts as the object of a verb. Pictograms tend not to have a clear one-to-one correspondence with natural language words. While in Example 4.23, ♥ acts as the object of a verb, such symbols could be employed to stand in for the verbs ‘like’, ‘love’, etc. In Example 4.24 the symbol \mathfrak{J} represents the word *grá* ‘love’. In Example 4.25, the smiley :) does not play any syntactic role in the sentence but is appended to an utterance to clarify the intended tone or emotion.

(4.23) *Conas a deireann tú ♥?*
Conas a deir tú “croí”
 ‘How do you say “heart”?’

(4.24) \mathfrak{J} *mór*
Grá mór
 ‘Lots of love’

(4.25) *Tá tusa gnóthach!! :)*
Tá tusa gnóthach!!
 ‘You are busy!!’

Truncation Due to the current character limit of a tweet, the end of a tweet may be unnaturally attenuated mid-sentence and sometimes even mid-word as in Examples 4.26 and 4.27. In these examples, we can guess what word was intended but it is not possible to infer any words that are entirely missing, thus some syntactic structure may be lost.

(4.26) thart fa’ 53 nó. . .
thart fa’ 53 nóiméad
 ‘over 53 minutes’

(4.27) *as an gcomhairlea. . .*
as an gcomhairleamh
 ‘out of the running’

Lone other-language items Based on our analysis of the TwittIrish sample, 25.29% of the TwittIrish tweets were found to be bi- or multilingual with the vast majority of other-language tokens being English words. There is a lack of agreement among the research community on how best to classify instances of language contact that occur as single words. Here, we refer to these occurrences as lone other-language items (LOLIs) (Poplack and Meechan, 1998) and revisit this issue in Chapter 6. Example 4.28 shows an Irish tweet with the English word ‘Dubs’, a nickname for ‘Dubliners’. This English token is used in an

otherwise Irish-language context and has undergone no orthographic nor morphosyntactic assimilation in that the English plural suffix ‘-s’ is used rather than an Irish plural suffix.

- (4.28) *Roimh na **Dubs***
Roimh lucht Bhaile Átha Cliath
‘Before the Dubs’

Example 4.29 shows a reference to the city ‘Barcelona’ in an Irish-language tweet. In this instance, the foreign proper noun has been assimilated to the orthographic and morphosyntactic frame of the Irish sentence by prepending an eclipse *m* to indicate the case of the noun and applying an acute accent to the ‘o’ to indicate the long vowel pronunciation.

- (4.29) *Tá sin i **mBarcelóna***
Tá sin in Barcelona
‘That is in Barcelona’

In our TwittIrish sample, the English language phrase ‘fair play’ occurs twice while variations ‘fair plé’, as shown in Example 4.30 and ‘féar plé’ occur once each. Interestingly, at the time of writing no variation of this phrase is listed in the Irish side of the NEID (Ó Mianáin, 2020).

- (4.30) ***Fair plé** daoibh’*
maith sibh
‘Fair play to you’

Words and phrases may be recurrent and/or diffuse and still not listed in a dictionary especially if it is a new term in an online, informal space rather than in edited publications. Another aspect to consider is whether or not the donor language term has an equivalent in the recipient language. For example, the English word ‘like’ is regularly used as a noun in the context of social media. No such nominal equivalent exists in Irish and so the phrase *is maith liom* ‘I like’ is used by many social media platforms. For the sake of clarity and brevity, the term ‘like’ is regularly borrowed.

- (4.31) *Tabhair **like** dúinn*
tabhair ‘is maith liom’ dúinn
‘Give us a like’

Meta-language tags Hashtags are used in tweets to render a topic searchable and at-mentions or handles are used to address or refer to another user. Both can play syntactic roles as shown in Figures 4.2 and 4.3.

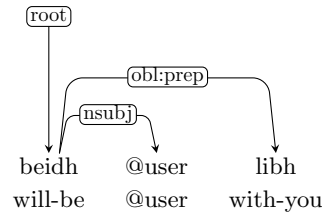


Figure 4.2: Username in a syntactic role ‘@user will be with you’.

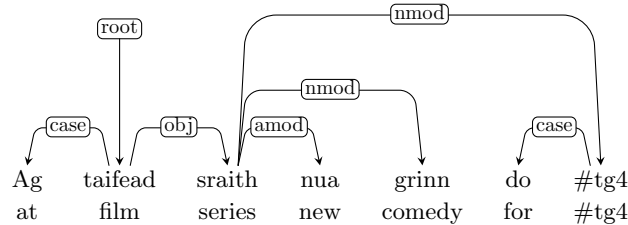


Figure 4.3: Syntactic hashtag ‘Filming a new comedy series for #tg4’.

4.4 Syntactic Variation in Irish Tweets

Grammatical phenomena observed in Irish tweets are described in this section. As these idiosyncrasies occur at the phrasal rather than token level, their effect is observed on the structure of the parse trees.

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 4.4 shows the syntactic annotation of the standard phrase *tá a fhios agam* and a contracted variation *tá’s agam*.

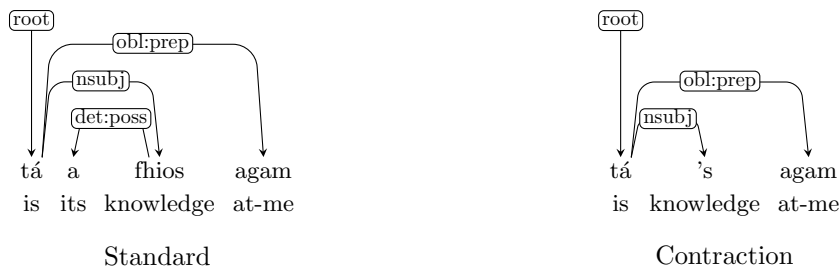


Figure 4.4: Contraction ‘I know’.

Over-splitting The inclusion of extra white space within tokens is also often observed in Irish tweets. This is exemplified in Figure 4.5. The prefix *ró-* (‘too’) is conventionally fused with the adjective it precedes in standardised text. Over-split tokens are annotated with the *goeswith* label as shown in Figure 4.5.

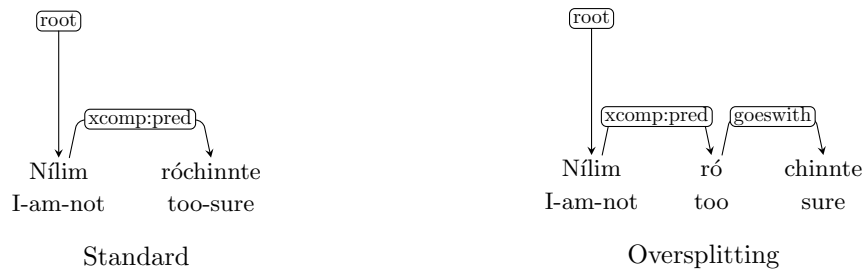


Figure 4.5: Over-splitting ‘I am not too sure’.

Code-switching Alternating languages within a tweet can alter the structure of the syntax tree, due to differing word orders of the languages involved, thus complicating the task of dependency parsing. Figure 4.6 shows an example in which an English phrase

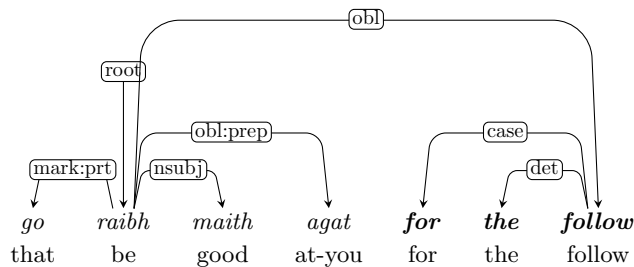


Figure 4.6: Code switching ‘Thank you for the follow’.

‘for the follow’ is inserted into an otherwise Irish sentence. This kind of code-switching follows the equivalence constraint of Poplack (1980) (see Section 2.4.2) in that, due to the location of the language switch in the sentence, the grammatical structure of both languages remains intact. Figure 4.7 provides a counter-example to the equivalence con-

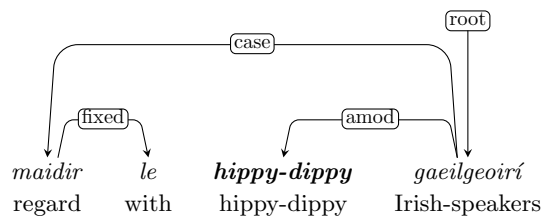


Figure 4.7: Code switching ‘as for hippy-dippy Irish speakers’.

straint as the language switch occurs at a point in the sentence where it is not possible to follow the grammar of both languages simultaneously. Poplack and Meechan (1998) refer to such points as ‘conflict sites’. In Irish, the adjectival modifier usually **follows** the noun it modifies whereas the inverse is true for English. In Figure 4.7 the English adjective ‘hippy-dippy’ is positioned **before** an Irish noun rather than after.

Ellipsis Figure 4.8 shows a sentence fragment lacking a main verb. The probable inferred full phrase is *tá báisteach anseo* ‘rain is here’. When the head of a phrase is elided, one of its dependents is promoted to the role of the head. In Figure 4.8, the nominal subject *báisteach* ‘rain’ is promoted.

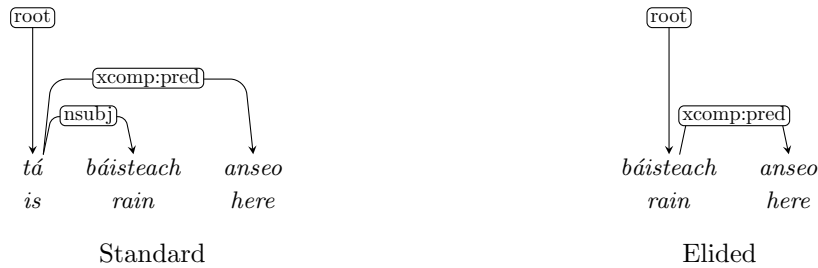


Figure 4.8: Ellipsis ‘rain (is) here’.

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 4.9 exemplifies a tweet utilising a smiley instead of punctuation.

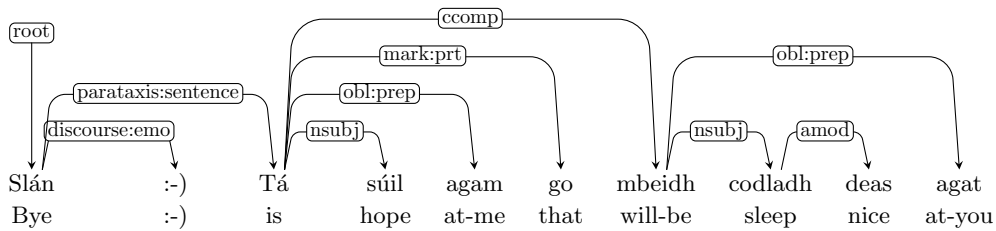


Figure 4.9: Non-sentential tweet using emoji as punctuation ‘Bye :-). I hope you have a nice sleep’.

Other syntactic variation Grammatical variation can also occur via unintentional deviation from conventional syntax by learners of Irish. Additionally, tweets can contain extremely unconventional constructions that have been machine-translated or generated by bots. Figure 4.10 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 a mortgage’. In the usual syntactic structure of a nonfinite clause in Irish, the object **precedes** the nonfinite verb. The non-standard variation in Figure 4.10 is a structure observed in our TwittIrish dataset in which the object **follows** the nonfinite verb, possibly mimicking the structure of English due to the introduction of the English word. As such structures do not follow the normal structure

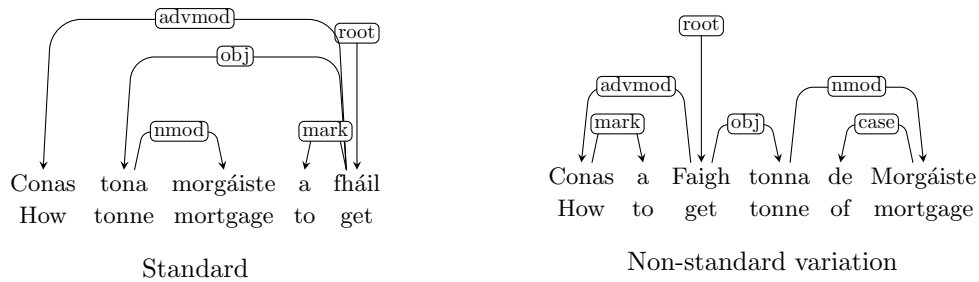


Figure 4.10: Grammatical variation ‘How to get a tonne of a mortgage’.

of Irish grammar, they pose a particular challenge for dependency parsing and produce unexpected results.

4.5 Summary

Our investigation into linguistic variation in Irish tweets has explored and categorised differences in Irish tweets as compared to standard Irish text. Understanding these variations is important for accurate language processing and analysis and it offers insights into the evolving nature of language in digital communication contexts. The variation we have described occurs for several reasons such as maximisation of tweets’ limited available space, dialect differences, self-expression, errors, automatic text generation or translation, and language contact. The resulting linguistic outcomes can create data sparsity and thus parsing challenges. While we acknowledge that machine learning techniques for NLP no longer require in-depth linguistic knowledge for feature engineering, we argue that understanding the domain allows for interoperability with other linguistic resources and efficient, reliable technology development that meets the needs of the users.

Chapter 5

Dependency Parsing Experiments

In this chapter, we explore the task of parsing Irish-language tweets. In Section 5.1, we describe the methodology used in our parsing experiments. In Section 5.2, we establish baseline results, examine the effect of using pre-trained contextualised word embeddings, and perform automatic and manual analysis of the results obtained. Finally, in Section 5.3, we present improved results from parsers developed using the newly available TwittIrish training and development sets, released in UD version 2.12.

5.1 Experiment Setup

Training Data Table 5.1 lists the treebank training sets used in our parsing experiments. The TwittIrish training set of 866 tweets, newly released with UD version 2.12, is currently the only genre-specific resource available. At the time of writing, the IUDT is the only other Irish-language treebank that includes a training set. The IUDT training set contains 4,005 sentences in the domains of fiction, government, legal, news, and web. Further background on the IUDT is provided in Section 2.1.4. Given that Scottish Gaelic is closely related to Irish, we also utilise the training data of the Annotated Reference Corpus of Scottish Gaelic ARCOSG treebank consisting of 3,541 sentences in the genres fiction, news, nonfiction, and spoken.

Biaffine Parser Our experiments are carried out using the state-of-the-art, graph-based biaffine dependency parser (Dozat and Manning, 2017). We implement this parser using AllenNLP (Gardner et al., 2018), a library for deep learning built on PyTorch. The parser is a multitask model that takes tokenised text as input and predicts POS tags and

Treebank	Tokens	Sentences/ Tweets	Language	Genre
TwittIrish	15,777	866	Irish	social media
IUDT	95,881	4,005	Irish	fiction, government, legal, news, web
ARCOSG	65,721	3,541	Scottish Gaelic	fiction, news, nonfiction, spoken

Table 5.1: Treebank training sets used in parsing experiments

dependency relations. Out of the box, the biaffine parser utilises contextual representations generated by a BiLSTM encoder. Alternatively, pre-trained language models like BERT can be used to generate contextual representations. We experiment with two encoder configurations in order to evaluate the effect of pre-trained word embeddings on parsing accuracy.

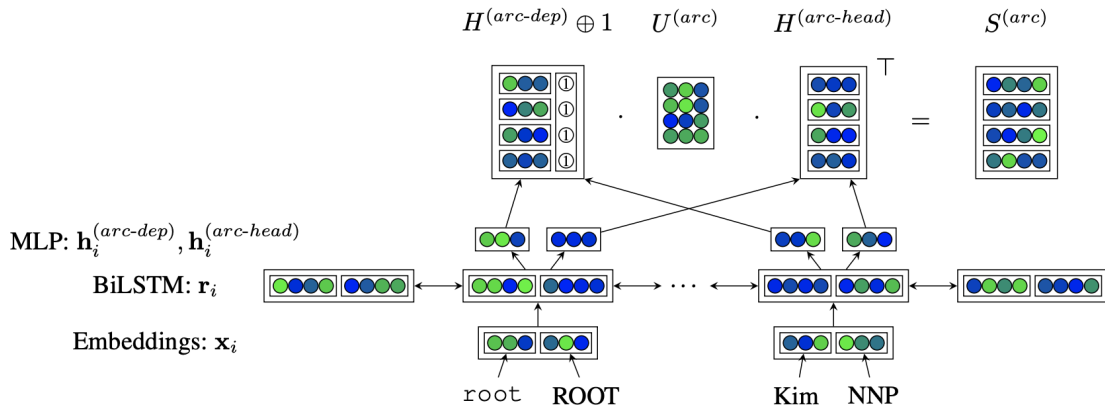


Figure 5.1: Biaffine dependency parser with BiLSTM encoder. Figure taken from Dozat and Manning (2017).

Biaffine Parser with BiLSTM Encoder Figure 5.1 shows the architecture of the biaffine parser using a BiLSTM encoder. The following steps outline the system’s parsing process.

1. An input word w_i and its POS tag t_i are each initialised as unique embeddings $v_i^{(word)}$ and $v_i^{(tag)}$.
2. Embedding x_i is created by concatenating the initial vectors $v_i^{(word)}$ and $v_i^{(tag)}$.
3. Three BiLSTM layers generate r_i , a context-aware representation, capturing contextual information from the training data.

4. The MLP layers generate two distinct representations for each word in the sentence. These representations correspond to two different perspectives of the word’s role in the dependency tree. $h_i^{(arc-dep)}$ represents the word when it is viewed as a dependent in an arc seeking a head and $h_i^{(arc-head)}$ represents the word when it is viewed as a head seeking all its dependents.
5. The biaffine attention mechanism calculates arc scores using the $h_i^{(arc-dep)}$ and $h_i^{(arc-head)}$ representations and the weight vector, $U^{(arc)}$. This attention mechanism efficiently captures the relationships between words, enabling the model to predict the most likely dependency arcs in the sentence.
6. The arc scores obtained from the biaffine attention mechanism are used as input to the Chu-Liu/Edmonds algorithm (Chu, 1965; Edmonds, 1967) to find the maximum spanning tree in the sentence, representing the most probable dependency parse tree.
7. Given the predicted dependency parse tree, a separate biaffine classifier predicts the dependency labels for each arc.

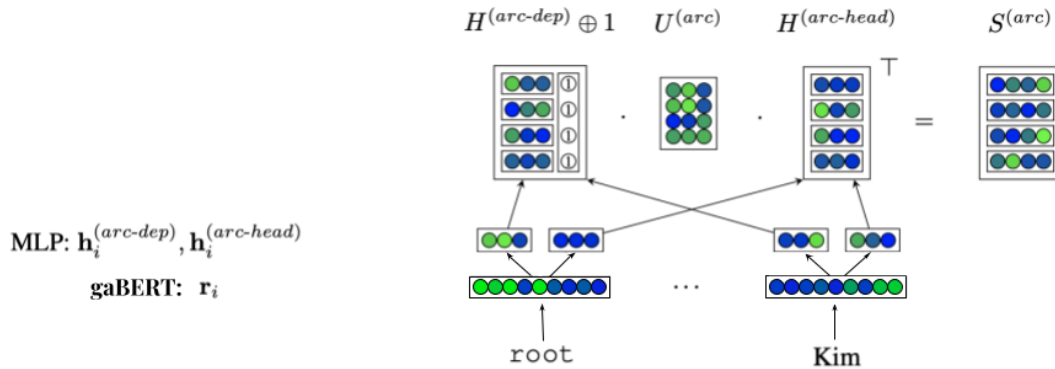


Figure 5.2: Biaffine dependency parser with gaBERT embeddings. Figure adapted from Dozat and Manning (2017).

Biaffine parser with gaBERT Encoder In order to leverage the substantial advances in accuracy achieved in dependency parsing by the use of pre-trained contextualised word representations (Che et al., 2018; Kondratyuk and Straka, 2019; Kulmizev et al., 2019), we also configure the biaffine parser with a gaBERT encoder.¹ In this case, we pass token representations obtained from the last hidden layer of a monolingual Irish BERT model

¹This model and its associated code are available on request.

(gaBERT) (Barry et al., 2022) to the parser. As discussed in Section 2.2.6, gaBERT embeddings have the potential to enhance the accuracy of dependency parsing for the Irish language as they have been trained on a diverse dataset of approximately 7.9 million sentences. As contextual information is already embedded within the gaBERT representations, the need for the BiLSTM layers is negated.

As shown in Figure 5.2, steps 1-3 of the previous section are replaced by initialising r_i as the gaBERT representation of the first WordPiece token of w_1 , e.g. if the word *málaí* ‘bags’ is tokenised as [má1, ##a1], then r_i is set to the gaBERT contextual representation for má1. Steps 4-7 apply as in the previous section. Table 5.2 shows the hyperparameters of the multitask tagging and parsing model.

Encoder	
Word-piece embedding size	768
Word-piece type	average
Dropout	0.33
Tagger	
MLP size	200
Dropout MLP	0.33
Nonlinear act. (MLP)	ELU
Parser	
Arc MLP size	500
Label MLP size	100
Dropout MLP	0.33
Nonlinear act. (MLP)	ELU
Optimiser and Training	
Optimiser	AdamW
Learning rate	3×10^{-4}
β_1	0.9
β_2	0.999
Num. epochs	50
Patience	10
Batch size	16

Table 5.2: Chosen hyperparameters for the multitask parser and tagger (adapted from Barry et al. (2022)).

Evaluation In order to evaluate the accuracy of dependency parsers, we compare their output on test data to a reference or gold standard version of the same data. An exact match (EM) metric determines how many trees are parsed entirely correctly. However, EM tends to be overly pessimistic and lacks the granularity needed to guide the development process. As a result, more refined metrics are employed, namely the labelled attachment score (LAS) and unlabelled attachment score (UAS). The LAS (Nivre et al., 2004) is a standard evaluation metric that we use in our dependency parsing experiments to measure

accuracy by calculating the percentage of words that are assigned the correct dependency label in the predicted dependency tree. UAS focuses solely on the correctness of the assigned heads, disregarding the labels. These metrics quantify the percentage of words in an input that receive the correct head and label assignments. Figure 5.3 exemplifies evaluation with UAS and LAS. In all of the following experiments we report the evaluation



Figure 5.3: Reference and system parses ‘Mia saved a dog’, resulting in an LAS of 1/3 and an UAS of 3/3.

metrics UAS LAS, as produced by the official CoNLL 2018 evaluation script.²

5.2 Establishing a Baseline

In order to establish baseline results we trained a biaffine dependency parser on the IUdT. Prior to the development of TwittIrish, this was the sole Irish-language UD treebank.

5.2.1 Baseline Parsing Results

Encoder	Training data	dev		test	
		UAS	LAS	UAS	LAS
BiLSTM	IUdT	58.26	48.5	57.79	46.96
gaBERT	IUdT	69.38	61.81	67.43	58.76

Table 5.3: Baseline parsing results, median score over five random seed values using the biaffine parser of Dozat and Manning (2017), trained on the IUdT version 2.12.

Table 5.3 presents our baseline results comparing the performance of the biaffine parser firstly using the BiLSTM encoder and secondly using the gaBERT encoder. The parsers were trained on the IUdT version 2.12 dataset. For the first configuration (Biaffine with BiLSTM encoder), the results on the development set show a UAS of 58.26 and an LAS of 48.5. On the test set, the UAS is 57.79, and the LAS is 46.96. In contrast, the second configuration (Biaffine with gaBERT encoder) outperforms the first significantly.

²https://github.com/ufal/conll2018/blob/master/evaluation_script/conll18_ud_eval.py

Parser	test LAS	
	IUDT	TwittIrish
Biaffine w/ gaBERT	84.25	59.34

Table 5.4: Parsing results used for analysis. Biaffine w/ gaBERT refers to the biaffine dependency parser of Dozat and Manning (2017) with gaBERT encodings (Barry et al., 2022). The parser was trained on the IUDT version 2.8 and tested on the IUDT and TwittIrish test sets.

On the development set, it achieves a UAS of 69.38 and an LAS of 61.81. On the test set, the UAS achieved is 67.43, and the LAS is 58.76. These results demonstrate that utilising the gaBERT encoder instead of the BiLSTM encoder substantially improves the performance of the biaffine parser across the board by about 11 LAS. As a reference point, the same biaffine parser with gaBERT encoder achieves about 84 LAS on IUDT test data, as shown by Cassidy et al. (2022) and Barry et al. (2022) suggesting potential for further improvement in the parser’s accuracy on Irish-language tweets.

5.2.2 Analysis of Baseline Results

We use Dependable (Choi et al., 2015), a web-based dependency parsing evaluation tool, to automatically break down the baseline parsing results, to gain a better understanding of the parser’s strengths and weaknesses with regard to standard Irish text and Irish-language tweets. We also perform manual error analysis on the most and least accurate parses to identify the most challenging aspects of parsing Irish-language tweets. The results shown in the previous section have been updated to show scores associated with datasets from UD version 2.12, for the purpose of compatibility. However, the following analysis was performed on the results earlier iteration of this experiment which used data from UD version 2.8, shown in Table 5.4. Updates to the datasets between these two versions mean that the results vary by approximately 1 LAS.

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 5.4 shows that, when tested on the IUDT, the parsing accuracy decreases as the length of the sentence increases. The highest accuracy of 87.92 LAS is associated with sentences of 10 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

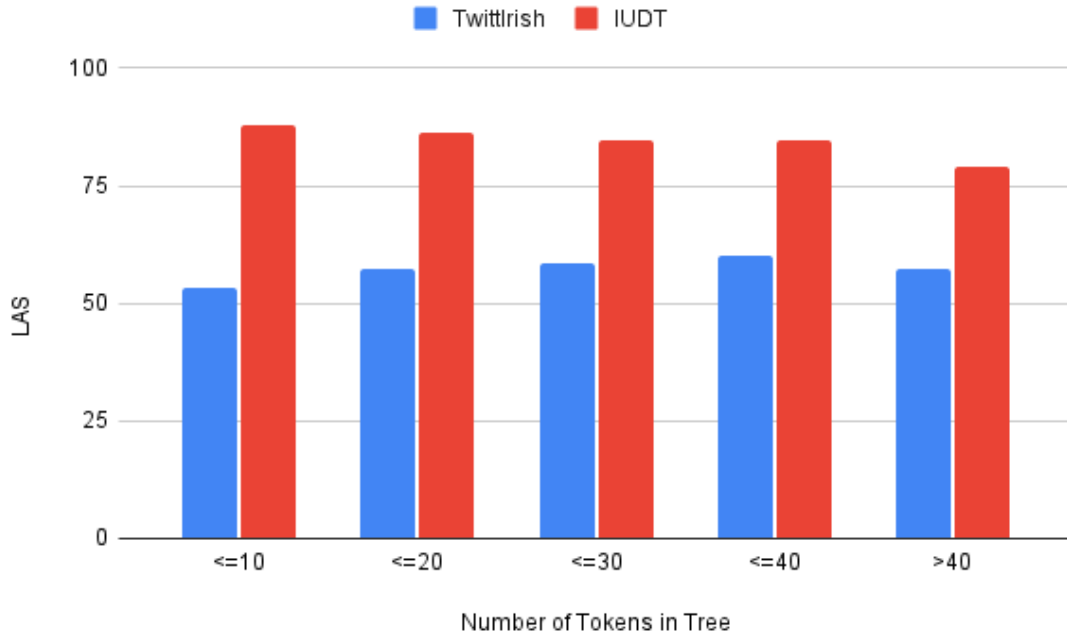


Figure 5.4: LAS by number of tokens per tweet achieved by biaffine parser with gaBERT embeddings on the TwittIrish and IUdT test sets.

dependency distances and more complex constructions within a sentence. While the range of scores is smaller and the 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 is associated with tweets of 31 to 40 tokens in length and the lowest accuracy is associated with tweets of 10 tokens or less. Kulmizev et al. (2019) found that deep contextualised word representations improve parsing accuracy for longer sentences, both for transition-based and graph-based parsers. However, in our experiments, higher accuracy for longer tweets is also observed when gaBERT representations are not used, suggesting that, in this case, deep contextualised word embeddings do not cause this effect. From manual inspection of the data, we observe that the genre-specific phenomena which challenge the parser such as ellipsis, metalanguage tags, and URLs, occur in higher proportions in shorter tweets making them harder to parse, whereas longer tweets tend to more closely resemble standardised language making them easier to parse.

LAS by UPOS To facilitate the interpretation of LAS broken down by POS tags, it is important to consider that the distribution of POS tags varies between standard text and Twitter text. For example, our analysis of the distribution of POS tags in the TwittIrish treebank reflects the observations of Rehbein et al. (2019), who developed a treebank of

German-language tweets, in that there is a larger proportion of symbols and punctuation in tweets as compared to standardised text which contains a higher proportion of the tags NOUN, DET, and ADP. Figure 5.5 shows the LAS associated with each UPOS tag when tested

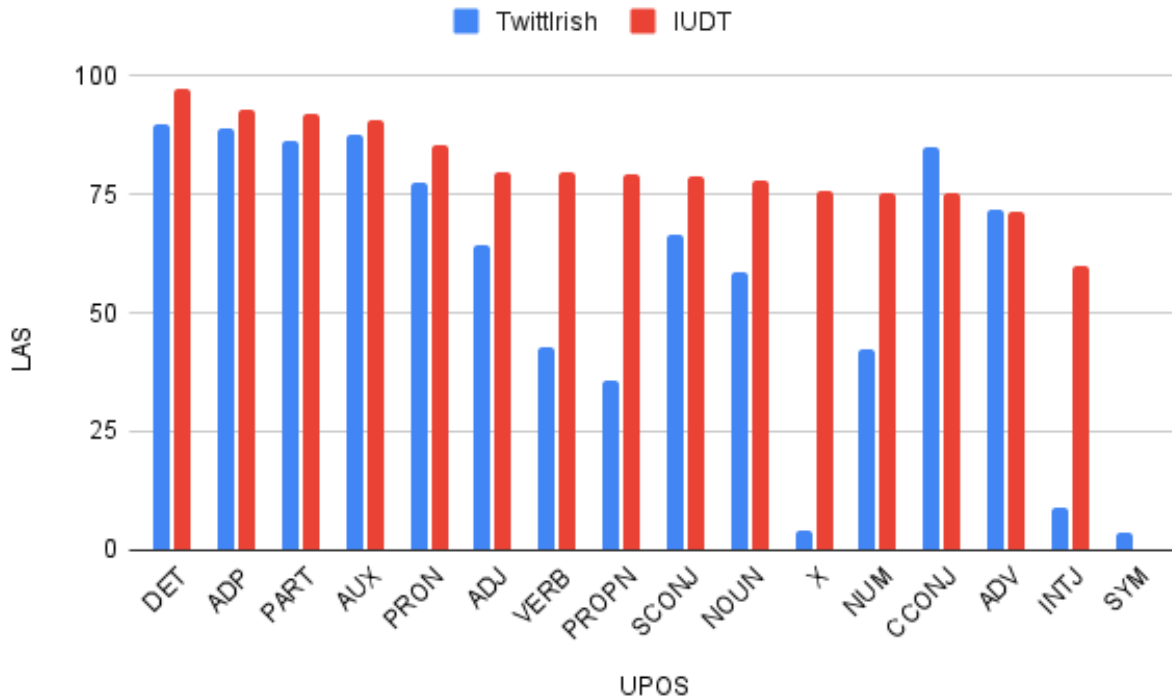


Figure 5.5: LAS by UPOS tag achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

on the IUDT and TwittIrish. As the parser is trained only on standard Irish, it obtains a higher LAS 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), and 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. In TwittIrish, however, non-Irish words are annotated with their true UPOS tag. With regard to the tag INTJ, in IUDT it occurs very rarely. However, due to the informal nature of TwittIrish, colloquial interjections, rare in standard text, are frequent.

Table 5.5 shows which UPOS tags are associated with higher- or lower-than-average LAS in both test sets. High accuracy is correlated with tokens that occur frequently

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 5.5: Confusion matrix of LAS by UPOS tag achieved by AllenNLP Biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

and have low lexical diversity. Lexical diversity refers to the number of different inflected forms that can be derived from a single underlying lemma. 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 a closed set of lemmata and thus has a low lexical diversity 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 could be due to variations in the usage of these tags in formal versus informal contexts. 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, just 0.02% of tokens have the UPOS tag SYM and the lexical diversity is high making it difficult for a parser to learn patterns.

LAS by dependency relation As with POS tags, the distribution of dependency relations varies between standard text and Twitter text. For example, when we compare the dependency relation distribution in the IUDT to TwittIrish, we find that the labels `case`, `det`, and `nmod` are more common in the IUDT and `parataxis`, `vocative`, and `advmod` are more frequent in TwittIrish. A cursory comparison of standard and Twitter UD treebanks in English, German, and Italian shows that this same effect is present in these languages. These differences could be due to the conversational nature and character limitations of tweets which can lead to different syntactic structures and dependency relations compared to more formal, standardised language.

Figure 5.6 shows parsing accuracy broken down by dependency relation. The chart orders dependency relations from left to right by their LAS on the IUDT test set. 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 accuracy of the two test sets are associated with the labels `root`, `vocative`, `obl:tmod`, `csubj:cleft`, `conj`, and `punct`.

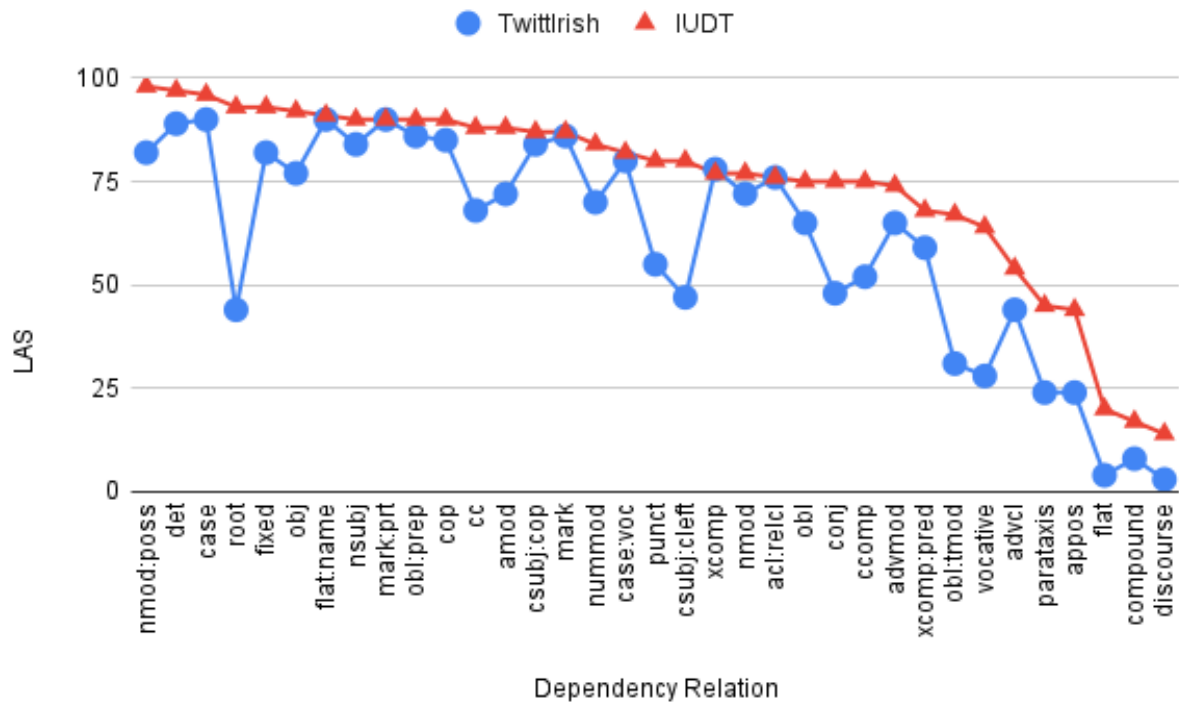


Figure 5.6: LAS achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets by dependency relation.

LAS	TwittIrish High	TwittIrish Low
IUDT High	nmod:poss, det, case, fixed, obj, flat:name, nsubj, mark:pnt, 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 5.6: Confusion matrix of LAS by dependency label achieved by the biaffine parser with gaBERT embeddings on the IUDT and TwittIrish test sets.

Table 5.6 shows dependency relations associated with higher or lower than their average LAS. High accuracy is seen in both test sets for dependency relations that apply to function words e.g. `det`, `case`. Function words tend to be part of a closed set and are

therefore likely to have been represented in the IUDT training data, resulting in higher accuracy. `root`, `csubj:cleft` and, `punct` are associated with higher-than-average LAS in the IUDT test sets but lower-than-average accuracy in the TwittIrish test set. This is likely due to the sentence segmentation differences in the IUDT and TwittIrish as described in Section 4.4. `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. As standard text, the IUDT contains more complex sentence structures and longer dependency distances. This complexity may have led to lower accuracy for these dependency relations. The relations `discourse`, `parataxis`, and `vocative` are not common in IUDT and are used in entirely new ways in TwittIrish, for emoji, hashtags, and usernames respectively. Therefore it is unsurprising that accuracy is low in both treebanks for these labels.

Error analysis In order to assess the effect on LAS of the UGC phenomena present in Irish-language tweets, we analyse the most and least accurate parses. Seven tweets (76 tokens) were parsed with LAS between 0 and 5. Examples 5.1 and 5.2 are two of these least accurately parsed tweets.

(5.1) *@user míle maith agat! :-)*
 ‘@user [thanks] a million :-)’

(5.2) *@user Blasta, ach beagán trom?*
 @user Tasty, but a bit heavy?

The tweet in Example 5.1 implies the full phrase *go raibh míle maith agat* ‘thanks a million’ however the main verb is ellided. The parser therefore incorrectly identifies the `root` as *míle*, which can also mean ‘mile’. Figure 5.7 illustrates this error.

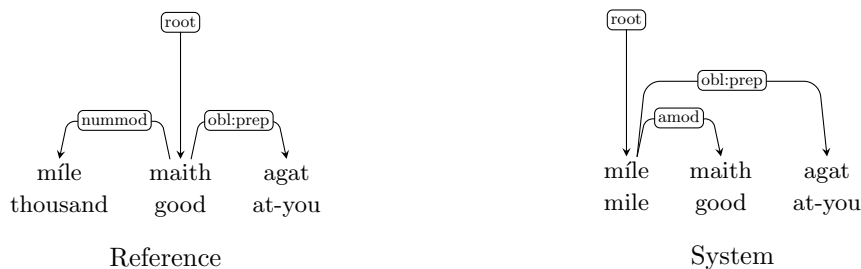


Figure 5.7: Correct and incorrect parse for phrase *míle maith agat* ‘[thanks] a million’.

Figure 5.8 also illustrates incorrect `root` identification degrading the accuracy of every

attachment in the parse tree. A parser trained on standardised language might expect a full sentence such as ‘It is tasty, but a bit heavy’ however, when the main verb of the sentence is elided, it is likely to misinterpret the syntactic structure implied.



Figure 5.8: Correct and incorrect parse for phrase *@user Blasta* ‘@user Tasty’.

Seven tweets (89 tokens) were parsed with an accuracy between 95 and 100 LAS. All of these were grammatical, well-formed sentences as exemplified in Examples 5.3 and 5.4. There were three usernames and one hashtag all of which were syntactically integrated and so they were parsed correctly. There was one occurrence of insertional single-word code-switching 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.

(5.3) *Is mise Arnaut, grá agam don ghaoth, a théann i ndiaidh an giorria leis an damh agus a théann ag snámh in aghaidh an easa - Dante*

‘I am Arnaut who loves the wind, who chases the hare with the ox, and swims against the current - Dante’

(5.4) *Beidh mé ar chlár @user anocht ag labhairt leis @user faoi #neknominations má tá fonn oraibh mo ghuth binn a chloisteáil.*

‘I will be on @user’s programme talking with @user about #neknominations if you want to hear my sweet voice.’

Table 5.7 shows the counts of UGC phenomena present in the most and least accurate parses. There were fifteen occurrences of emojis which were most commonly incorrectly labelled **punct**. The ten occurrences of code-switching were most commonly incorrectly attached via **flat:foreign**. The nine (two syntactic) occurrences of usernames were most commonly incorrectly labelled 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 was observed in the form of diacritic omission wherein the word *ár* ‘our’ was rendered as *ar* ‘on’ causing the parser to misinterpret the dependency label. From these results, it is evident that UGC phenomena and language contact are associated with lower parsing accuracy.

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

Table 5.7: Number of occurrences of UGC phenomena where easiest tweets refers to the 7 tweets that were parsed with LAS between 95 and 100 and hardest tweets refers to the 7 tweets (76 tokens) that were parsed with LAS between 0 and 5.

5.3 Improving the Parser

When the TwittIrish training and development sets were complete as of UD version 2.12, we carried out our final parsing experiments using the biaffine parser with BiLSTM and gaBERT encodings. We trained the parsing model on three variations of training data. In an ideal situation, we would use a large, gold-standard training set in the same language and genre as the test data. Such data is not yet available for Irish, as is the case for many low-resource languages. For this reason, we experiment with the training data of different genres currently available via UD in Irish and Scottish Gaelic, a Celtic language closely related to Irish. This demonstrates the benefit of the standardised UD annotation scheme allowing us to leverage other treebanks, even if the data is not genre- or language-specific.

5.3.1 Improved Results

Table 5.8 presents the full results of our dependency parsing experiments on the TwittIrish test set. The first two rows show the baseline results. We observe that using the TwittIrish training data offers an improvement in test LAS of about 13 points over the baseline for the parser with the BiLSTM and about 18 points for the parser with gaBERT encodings.

Encoder	Training data	Dev		Test	
		UAS	LAS	UAS	LAS
BiLSTM	IUDT	58.26	48.5	57.79	46.96
BiLSTM	TwittIrish	73.62	62.64	72.27	60.1
BiLSTM	IUDT + TwittIrish	81.07	73.22	79.14	70.32
BiLSTM	ARCOSG + IUDT + TwittIrish	80.87	73.05	78.83	70.13
gaBERT	IUDT	69.38	61.81	67.43	58.76
gaBERT	TwittIrish	84.83	79.70	82.95	76.88
gaBERT	IUDT + TwittIrish	88.58	84.10	85.41	79.71
gaBERT	ARCOSG + IUDT + TwittIrish	88.62	84.07	85.54	79.47

Table 5.8: Full dependency parsing results using biaffine parser of Dozat and Manning (2017) on the TwittIrish test set using training data from Irish and Scottish Gaelic treebanks of UD version 2.12.

The addition of the IUDT to the training data further boosts the test LAS by about 10 points for the parser with BiLSTM encodings and by 3 points for the parser with gaBERT encodings resulting in the highest test LAS of 79.71. The difference in the boost offered by the additional training data to the parsers can be explained by the different strengths and weaknesses of the encodings. The BiLSTM encoding might have benefited more from the additional data due to its ability to capture sequential patterns effectively. On the other hand, the gaBERT encoding, being a transformer-based model, might have already captured some relevant linguistic information from the large-scale pre-training on diverse corpora, making the gains from the IUDT dataset less substantial. The final addition of the ARCOSG data to the training set resulted in a slightly higher UAS than the combination of IUDT and TwittIrish when the gaBERT encoder was used but slightly lower LAS overall. In this configuration of the experiment, wherein the training data of each treebank was simply concatenated, the best results are achieved by the combination of IUDT and TwittIrish. However, more sophisticated techniques for multilingual parsing may yield different results.

Figure 5.9 illustrates the median test LAS of all model configurations across five random seed values. It is evident that incorporating additional training data from various treebanks and using the gaBERT encoder leads to improved dependency parsing results. This performance gain demonstrates the effectiveness of using UD resources for low-resource languages like Irish and Scottish Gaelic.

5.3.2 Preliminary Analysis of Improved Results

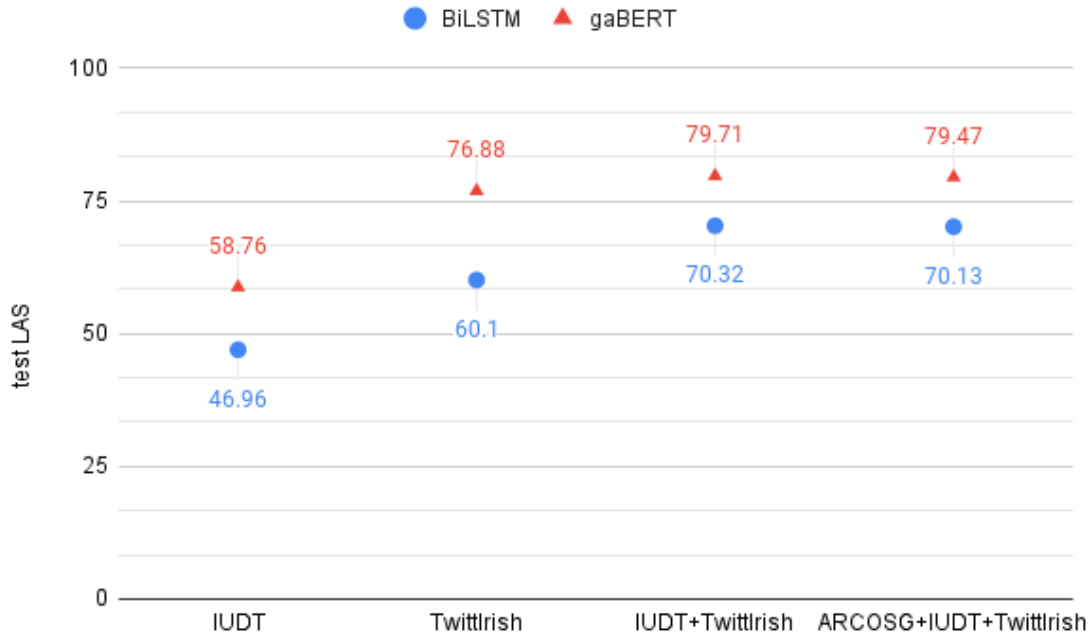


Figure 5.9: Dependency parsing results showing median LAS over five random seed values on TwittIrish test set varying training data and encoder of biaffine parser (Dozat and Manning, 2017).

LAS by number of tokens per tweet Figure 5.10 shows the LAS of our best-performing model, biaffine parser with gaBERT embeddings trained on TwittIrish and IUDT version 2.12 broken down by the number of tokens in the tree. In comparison to our baseline results illustrated in Section 5.4, LAS has increased for every segmentation. Further, the trend of shorter tweets being more difficult to parse has been reversed, i.e. we now observe that parsing accuracy decreases as the length of the tweet increases. This implies that our improved parser has reduced the challenges caused by UGC phenomena being disproportionately present in shorter tweets.

LAS by UPOS tag Figure 5.11 shows the LAS broken down by UPOS tag of our best parser compared to that of a parser with the same architecture trained and tested on the IUDT treebank. We use this comparison to demonstrate that for many UPOS tags, our best parser achieves a similar performance to the reference due to the addition of genre-specific training data. Indeed, our parser outperforms the reference for the UPOS tags SYM, CCONJ, and SCONJ. However, our parser is still less accurate for several tags, notably INTJ and NUM. This may be due to varying usage of these parts-of-speech in the tweets e.g. numbers are common in tweets to report scores in live sport.

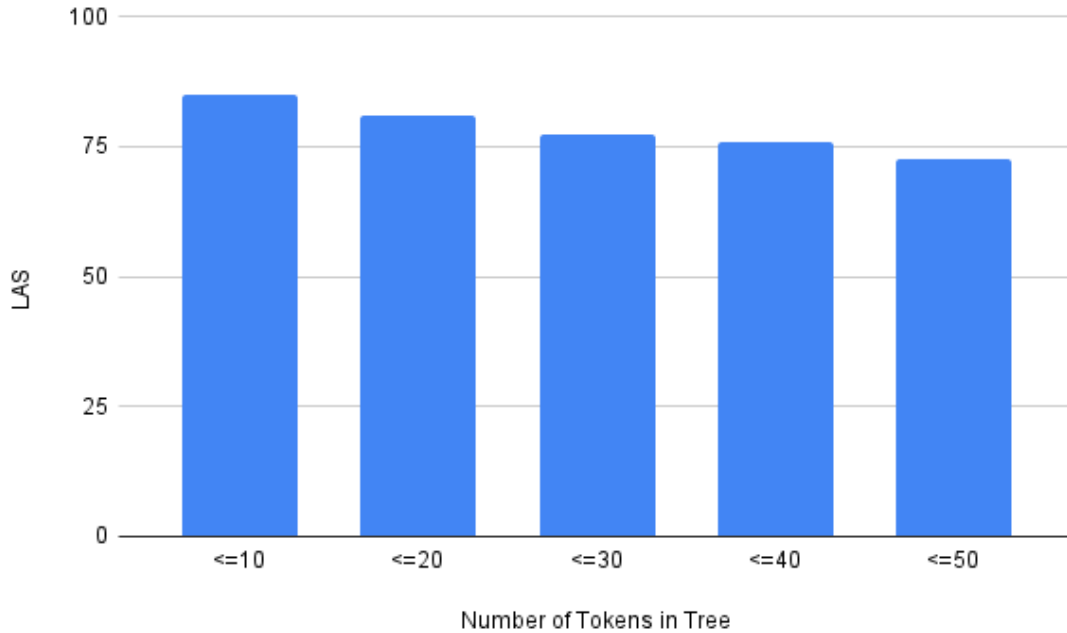


Figure 5.10: LAS by number of tokens per tweet achieved by biaffine parser with gaBERT embeddings trained on TwittIrish and IUDT version 2.12

LAS by dependency relation Figure 5.12 shows the LAS categorised by dependency relation for our top-performing parser, in comparison to a parser with the same architecture trained and tested on the IUDT treebank. As we have shown in the previous paragraph in the case of UPOS tags, our best parser also achieves comparable performance to the reference for most dependency relations. Notably, substantially higher performance is observed for dependency relations `discourse` and `case:voc`. This is likely due to the conversational nature of Twitter text in which users are more likely to address one another directly. Thus, when training and testing parsers within the social media genre, the model has a greater opportunity to learn these syntactic structures. Some dependency relations with which our parser still struggles are `flat`, `compound`, and `compound:prt`. These dependency relations are used for annotating MWEs and pose a challenge for parsers and annotators largely due to the labels being applied inconsistently within UD treebanks. Though McGuinness et al. (2020) have explored the annotation of MWEs in the context of Irish, these results imply that further work is needed to align the annotation conventions among treebanks. There is also a substantial dip in the accuracy of the relations `vocative` and `csubj:cleft` which could be attributed to these syntactic structures being relatively infrequent in tweets. With regard to `vocative`, in tweets, the subtype `vocative:mention`

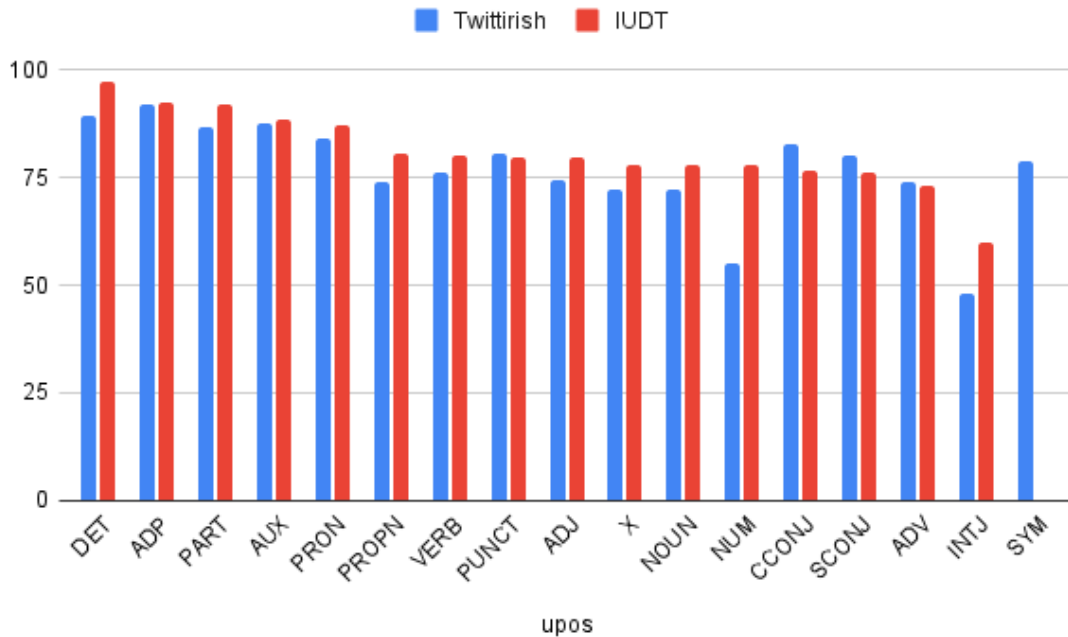


Figure 5.11: Comparison of LAS by UPOS tag of our best parser against a reference. Our best parser is trained on TwittIrish and IUDT treebanks and tested on TwittIrish. The reference parser is trained and tested on the IUDT treebank.

will usually be used to address another user by their handle rather than a `vocative` which would be associated with a regular noun or proper noun. The relation `csubj:cleft` is also infrequent in tweets as compared to more formal or literary text.

5.4 Summary

This chapter has presented dependency parsing experiments for Irish-language tweets using a biaffine dependency parser tested on the TwittIrish treebank. Two encoder configurations, BiLSTM and gaBERT, were used to evaluate the effect of pre-trained contextualised word embeddings on parsing accuracy. Our baseline parsers were trained on the IUDT, an Irish-language treebank of standardised text. Even without genre-specific training data, the use of gaBERT embeddings led to an improvement of 12 LAS over the BiLSTM encoder in our baseline experiments. The introduction of the TwittIrish training data led to a substantial increase in test LAS, with a 13-point improvement over the baseline for the parser with the BiLSTM encoder and an 18-point improvement for the gaBERT encoder. The dependency parsing experiments on the TwittIrish test set have shown significant improvements over the baseline results. Further, the addition of the

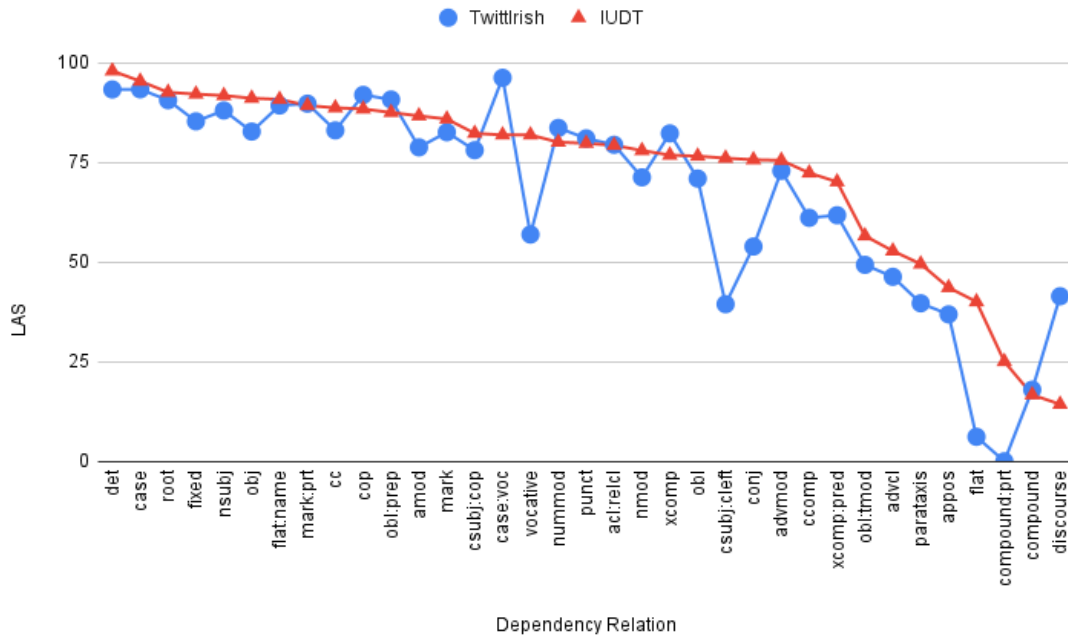


Figure 5.12: Comparison of LAS by dependency relation of our best parser against a reference. Our best parser is trained on TwittIrish and IUDT treebanks and tested on TwittIrish. The reference parser is trained and tested on the IUDT treebank.

IUDT dataset to the training data resulted in an additional boost in accuracy of 10 LAS for the parser with the BiLSTM encoder and an increase of 3 LAS for the parser with the gaBERT encoder. This improvement has greatly reduced the challenges associated with parsing UGC-specific features such as URLs, emoticons, usernames, and hashtags in Irish. We find that the addition of Scottish Gaelic training data from the ARCOSG treebank did not improve parsing accuracy in this case but many possibilities exist for future multilingual experimentation. We have also provided error analysis of our baseline and top-performing models allowing us to better understand the genre of social media text. This can inform researchers’ NLP model selection when working with social media text in Irish and other low-resource languages. Error analysis also helps treebank developers to enhance dataset quality by prompting the resolution of differences in annotation conventions. Ultimately, the introduction of genre-specific training data and gaBERT encodings has increased parsing accuracy for Irish-language tweets. These findings contribute to the advancement of NLP for lesser-resourced languages and facilitate the development of more accurate NLP models for social media content.

Chapter 6

Language Contact Questionnaire

Study

Language contact with English is one of the most salient features of Irish-language tweets, as discussed in Chapter 4. The curation and preprocessing of data for NLP often requires language identification (Lui and Baldwin, 2012; Jauhiainen et al., 2019) which can be a challenging task in language contact situations. This chapter explores the theoretical linguistic distinction between code-switching and borrowing in the context of Irish-language tweets. We aim to answer RQ3: ‘How should language contact phenomena in Irish tweets be classified?’ by presenting the results of an anonymous, internet-based, mixed method questionnaire study on language contact in Irish-language tweets. The participants were 256 adult Irish speakers of all levels. The goal of the study was to investigate the perceptions of Irish speakers regarding code-switching and borrowing. It was hypothesised that borrowed words would be considered ‘less English’ than code-switched words and that Irish speakers would be more likely to use words borrowed from English rather than code-switched English words in an Irish-language context. The results support both hypotheses. Qualitative analysis of the data further revealed the key themes of **clarity**, **convenience**, **conformity**, and **language contact** as factors influencing word choice among Irish speakers. As such, we conclude that it is important to distinguish between code-switching and borrowing in the development of NLP resources while acknowledging the lack of consensus within the research community on how best to approach this task. These results contribute to a better understanding of language contact and variation, as well as potential practical applications in language education, policy, and social integration

among linguistic communities.

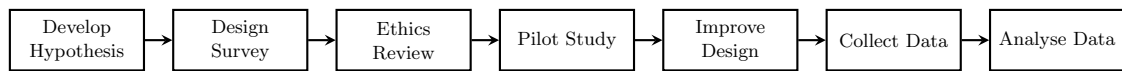


Figure 6.1: Diagram of the questionnaire study process.

Figure 6.1 outlines the steps involved in the questionnaire study process. Each step is described in detail in the following sections of this chapter.

6.1 Related Work

As described in Section 2.4, many frameworks and typologies have been proposed to describe language contact (e.g., Weinreich (1953); Muysken (1997)); yet they often contradict one another. One school of thought argues that borrowing and code-switching exist on a continuum (Myers-Scotton, 1992; Boztepe, 2003) while others propose methods to distinguish between them (Poplack and Meechan, 1998; Lipski, 2005). This theoretical dissonance has led to a grey area in the management and processing of bilingual data in NLP. The simplest solution to this problem is to label every occurrence of an English word as code-switching. However, this approach works off the assumption that any token from another language is a code-switch, excluding the possibility of borrowed words, named entities, and items with no particular language e.g. usernames, URLs, etc.

Álvarez-Mellado and Lignos (2022) recommend a more nuanced solution and lay out a useful set of criteria for identifying borrowed words in Spanish tweets. We aim to investigate whether or not these criteria for distinguishing between code-switching and borrowing work well in the context of Irish-language tweets.

Criterion	Description
C1	English words related to Twitter terminology: such as ‘tweet’, ‘follower’, etc.
C2	Technology words: ‘server’, ‘hosting’, ‘user’, ‘post’, ‘blog’, etc.
C3	English words that are already registered in the New English-Irish Dictionary ^a (NEID), e.g. <i>bus</i> ‘bus’
C4	English words that are the headword of an entry in Vicipéid ^b (Irish Wikipedia), such as music styles, genres and other cultural things.

Table 6.1: Borrowing criteria (Álvarez-Mellado and Lignos, 2022) adapted for Irish.

^a<https://www.focloir.ie/en/>

^b<https://ga.wikipedia.org/>

Table 6.1 shows the criteria developed by Álvarez-Mellado and Lignos (2022) adapted

here for the case of Irish-language tweets. We use these criteria to classify the words tested in the questionnaire study. C1 states that if an English word is a Twitter-related term, it should be classified as a borrowing. C2 states that if an English word pertains to technology, it should be classified as a borrowing. C1 and C2 were directly transferable to the case of Irish-language text and so were not adapted.

C3 states that if an English word is already registered in a particular dictionary of the target language, it should be classified as a borrowing. Álvarez-Mellado and Lignos (2022) refer to *Diccionario de la Lengua Española* (Real Academia Española, 2021), the general dictionary of standard Spanish compiled by the Royal Spanish Academy. We adapt C3 by referring instead to NEID, the online, searchable version of the English-Irish Dictionary (de Bhaldraithe, 1959; Ó Mianáin, 2020). We chose this dictionary as it is representative of contemporary Irish, reflects different dialects, and strikes a balance between formal and informal registers (Ó Murchadha and Kavanagh, 2022).

C4 states that if an English word is the headword of an entry in the Wikipedia of the target language, it should be classified as a borrowing. Where Álvarez-Mellado and Lignos (2022) refer to Spanish Wikipedia, we refer to Vicipéid ¹, the Irish-language Wikipedia. In Figure 6.2, we provide an image of the Irish-language Wikipedia article with the headword ‘Twerking’.

Álvarez-Mellado and Lignos (2022) include two additional criteria which we do not employ. The first is “English words that are already registered in *Diccionario de Americanismos* [...] a specialised dictionary that covers the vocabulary spoken in American Spanish and that has a rich representation of well-established lexical borrowings from English used in Latin America”. There is no such equivalent for the case of Irish. The second is “words that have English origin but were used following Spanish grammatical structure, such as noun-adjective word order (*mensajes offline, rating online*)”. We, however, did not use this criterion in the current study when classifying code-switching and borrowing as it deals with syntax whereas our intention is to investigate code-switching and borrowing on a purely lexical level. Our reason for this is that language contact which causes differences in syntax will generally be considered code-switching rather than borrowing Deuchar (2020). These cases are easier to detect automatically as they will affect the structure of a parse tree.

¹<https://ga.wikipedia.org>

Twerking

33 languages

Alt Plé

Léigh

Cuir in eagar

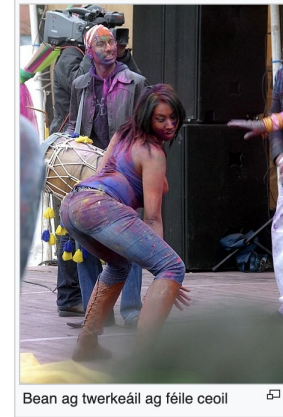
Cuir foinsé in eagar

Féach ar stair

Uirlisí

Ón Vicipéid, an chlicpéid shaor.

Nós **damhsa Meiriceánach** é **twerking** (/ˈtwɜːrkɪŋ/), damhsa ina bhfuil duine ag rince i stíl ghnéasach le gluaiseacht na cromán, agus sé nó sí cromtha go híseal. Cé nach bhfuil a fhios ag éinne cárbh as a tháinig an focal seo, deirtear go coitianta go bhfuil aicearra ar an bhfocal **Béarla** “footwork” ann, agus **portmanta** ar na focail “twist” agus “jerk” i gceist. Áfach, tá fianaise ann ó agallaimh gur tháinig an frása ó sráideanna **New Orleans** nuair a thosaigh an stíl **rapcheol** darbh ainm “Bounce”. Bhí baint idir ceol “Bounce” agus twerking ó na 1990í i leith, agus scaipeadh í trí mheán fhíseanna rapcheoil agus suíomhanna físeáin ar líne ó lár na 2000í. Tá sé scríofa as Gaeilge (go neamhoifigiúil) mar **twerkeáil** nó **twerkáil** uaireanta.



Bean ag twerkeáil ag féile ceoil

Figure 6.2: Irish-language Wikipedia article with the headword ‘Twerking’.

The words selected as examples of code-switching and borrowing were therefore classified based on the four criteria C1-C4. Where a lone English-origin word was detected in an otherwise Irish-language tweet, if any of C1-C4 applied, borrowing was assumed. Alternatively, if none of C1-C4 applied then code-switching was assumed. The specific examples used in the study are discussed in Section 6.2.2

6.2 Questionnaire Design and Development

Research Question	Variable	Analysis
RQ3.1: Do Irish speakers classify borrowed words as English less often than code-switched words?	Language classifications	Quantitative
RQ3.2: Do Irish speakers claim to be more likely to use borrowed words than code-switched words?	‘Likelihood of use’ score	Quantitative
RQ3.3: What themes can be interpreted from Irish speakers’ explanations about word choice?	Explanation of ‘likelihood of use’ score	Qualitative

Table 6.2: Specific research questions investigated in questionnaire study.

We break RQ3 down into three subquestions (see Table 6.2) addressed through the questionnaire. RQ3.1 explores whether or not Irish speakers consciously perceive a difference with regard to language membership between words classified as borrowed or code-switched words by the criteria C1-C4. We hypothesise that borrowed words are classified

as ‘English’ by respondents less often than code-switched words. We aim to establish this via quantitative data analysis on the questionnaire responses to the language classification task. If the results obtained are not significantly different between these two groups, then we could infer that code-switched words and borrowed words are perceived in the same way or that some words have been misclassified by the borrowing criteria used.

RQ3.2 investigates any difference in the willingness of Irish speakers to use code-switched versus borrowed words. We aim to compare, using statistical methods, the responses in which participants rate how likely they would be to use a given word in a given context. We hypothesise that the perception of a word as ‘English’ will be associated with a lower ‘likelihood of use’ rating.

To answer RQ3.3, we use reflexive thematic analysis (Braun and Clarke, 2006) to examine the open-ended responses in which respondents explain their reasoning around why they would (not) use a particular word in a given context.

6.2.1 Design Choices

The full text of the questionnaire is provided in Appendix D. We follow the guide of Dörnyei and Dewaele (2022) for producing and using a self-completed questionnaire with a clear layout and simple natural language for reliable and valid research.

English language The questionnaire was available in English as opposed to Irish in order for it to be accessible to Irish speakers of all levels. An Irish-language version of the survey was not made available.

Anonymous study We decided to make the questionnaire anonymous to encourage honest responses that would not be linked to the respondents’ identities. In this way, participants do not need to fear any adverse consequences for expressing unpopular views or for difficulties they may have with the language.

Internet-based study We chose to conduct the questionnaire online as it was convenient, cost-effective, and efficient. The questionnaire was hosted on Google Forms, where a spreadsheet of responses securely stored in the Dublin City University Google Drive was automatically populated, eliminating any other potential data-processing steps such as transcribing or collating files.

Mixed methods study The questionnaire included a mix of closed- and open-ended questions to collect data on participants' background, language usage, judgements and opinions on 36 words sampled from Irish-language tweets. The closed-ended responses from all participants were analysed quantitatively using statistical methods. 36 open-ended responses per 36 sampled words were analysed using the qualitative method of reflexive thematic analysis. By combining quantitative and qualitative methods, we achieved richer results and a more in-depth and nuanced understanding of language contact phenomena in the context of informal Irish. While quantitative data has the advantage of determining statistical significance, qualitative data can provide insights into the experiences and perspectives of participants.

Questionnaire length We aimed to keep the duration of the questionnaire under 20 minutes, considering the trade-off between the number and variety of questions, the level of detail in responses, and the overall response rate. The questionnaire consisted of three sections: Plain Language Statement and Informed Consent, Your Language Background, and Name the Language.

Ethical considerations A Dublin City University ethical review was carried out prior to launching the pilot study. The ethical approval is provided in Appendix C. We took care in forming each element of the questionnaire. The first section of the questionnaire contained a detailed plain language statement and 10 statements of informed consent to ensure that participants were over eighteen years old and were fully aware of what was involved in participating in the study. The participant could only proceed with the questionnaire having agreed to all statements by ticking a box. We did not expect this study to involve any risks to the participant nor did this project include any procedure which is beyond already established and accepted techniques. Only questions that would be relevant to the data to be studied were included in the second and third sections of the questionnaire.

Theoretical framework The reflexive thematic analysis method was utilised for conducting the qualitative analysis. This approach involved deriving themes from the data in an inductive manner, as opposed to relying on pre-existing linguistic theory. The analysis was approached from a constructionist perspective, which emphasises the role of language

in shaping meaning and understanding (Burr, 2015). Moreover, our analysis identified themes at the latent rather than semantic level i.e. themes not immediately apparent from the surface-level data. Further explanation of latent analysis is provided in Section 6.5.4.

Variable	Values
Age	18-24, 25-34, 35-44, 45-54, 55-64, 65+
Level of Irish	Beginner (A1-A2), Intermediate (B1-B2), Advanced (C1), Fluent (C2), Native
Dialect	Connacht, Munster, Ulster, a mix, other
Number of Proficient Languages	1, 2, 3, 4+
Frequency of formal Irish use	daily, weekly, monthly, less than once a month
Frequency of Irish-English mixing in formal context	1 (never), 2, 3, 4, 5 (always)
Frequency of informal Irish use	daily, weekly, monthly, less than once a month
Frequency of Irish-English mixing in informal context	1 (never), 2, 3, 4, 5 (always)
Feelings on mixing Irish and English	1 (very negative), 2, 3, 4, 5 (very positive)

Table 6.3: Demographic and language background variables.

Demographic and language-background questions We included demographic and language-background questions in order to assess whether we had achieved broad coverage of the Irish-speaking population and to stratify the results based on these responses. This was important to help us to understand if any of these factors may had an effect on the judgements of respondents with regard to language classification and willingness to use certain words. The demographic variables are shown in Table 6.3

6.2.2 Language Contact Questions

The third and final section contained 36 extracts from Irish-language tweets, each with a single highlighted word. Respondents were asked three questions regarding each extract:

Part A This multiple-choice, required question asked participants to identify the language of the given word. Participants could select one of the following options: ‘Irish’, ‘English’, ‘The word exists in both languages’, ‘The word is a mix of Irish and English’, or ‘Neither’.

Part B This 5-point Likert scale, required question asked participants to rate, from ‘very unlikely’ to ‘very likely’, how likely they would be to use the given word in the given

context.

Part C This open-ended, optional question asked participants to explain their answer to Part B.

The 36 highlighted words were selected to comprise four groups: ‘Borrowed’, ‘Code-switched’, ‘Irish’, and ‘Ambiguous’ words. Each group consisted of nine words. Classifications were made using criteria C1-C4 as shown in Figure 6.3. The following section explains these classifications in more detail.

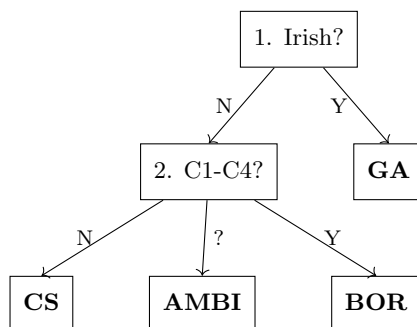


Figure 6.3: Decision tree for classifying words as GA (Irish), CS (English code-switch), BOR (borrowing), or AMBI (ambiguous).

Extract	Borrowing Criteria
<i>Cuirfidh mé DM chuici</i> ‘I will send her a DM ’	C1
<i>níl haon ionadh orm go bhfuil na hits a méadú</i> ‘it’s no surprise that the hits are increasing’	C2
<i>Tá’n blag ag lorg scríbhneoir faisean</i> ‘The blog is looking for a fashion writer’	C2, C3
<i>Cuireann an twerking sin isteach orm</i> ‘That twerking annoys me’	C4
<i>Tá keyboards beag an deachair</i> ‘Small keyboards are very difficult’	C2
<i>níl mé ach tar éis tweet a léamh</i> ‘I just read a tweet ’	C1
<i>#Gaeilge mar rogha ar aip agus ATM</i> ‘#Irish as an option on an app and an ATM’	C2, C3
<i>Ag déanamh meaitseáil ar an ríomhaire</i> ‘ matching on the computer’	C3
<i>Tá físeáin haiceanna, cláir agus stíleanna gruaige ann</i> ‘There are hack videos, programmes and hairstyles’	C2, C3

Table 6.4: Borrowing extracts from questionnaire study.

Borrowed words Table 6.4 lists the extracts in the ‘borrowed words’ group. Each highlighted word meets at least one criterion of C1-C4. e.g. ‘DM’ refers to a direct message which is common terminology on Twitter and other social media platforms so we conclude that it meets criterion C1. ‘Hits’ can have several meanings, one of which refers to the number of times a website has been visited. Therefore, we conclude that it meets criterion C2 while also noting that this is subject to interpretation.

Extract	Translation
<i>50 bliain idir na pics seo</i>	‘50 years between these pics ’
<i>Ranganna yoga trí Ghaeilge anocht</i>	‘ yoga classes through Irish tonight’
<i>Roimh na Dubs</i>	‘Before the Dubs ’
<i>Wish nach raibh aon obair le déanamh agam</i>	‘ Wish I didn’t have work to do’
<i>Just samhlaigh an racht feirge a mhothaímse</i>	‘ Just imagine the surge of anger I feel’
<i>D’ioslodail me an album nua</i>	‘I downloaded the new album ’
<i>deochanna le mo kinda col ceathar</i>	‘drinks with my kinda cousin’
<i>absolutely álainn</i>	‘ absolutely beautiful’
<i>Tá tú an-mhaith ag an housework inniu</i>	‘you are very good at the housework today’

Table 6.5: Code-switching extracts from questionnaire study.

Code-switched words Table 6.5 lists the extracts in the ‘code-switched’ group along with their English translations. They all consist of a single English-language highlighted word in an otherwise Irish context. We classify each of them as code-switching as they meet none of the criteria C1-C4 for borrowed words.

Extract	Translation
<i>faigh réidh leis an riail sin</i>	‘get rid of that rule ’
<i>Ní féasta go rósta is ní céasta go pósta</i>	‘no feast until a roast , no torture until marriage’
<i>Grma a chroí 😊</i>	‘ thanks love 😊’
<i>Tá an fhoireann ar fad go hálainn.</i>	‘The whole team is lovely’
<i>Tá siad fós ag imirt 😊</i>	‘they are still playing 😊’
<i>Samplaí anseo de logainmneacha</i>	‘ Examples here of placenames’
<i>Beidh muid ag plé an ábhair seo</i>	‘We will be discussing this subject’
<i>Drámaí deasa inniu</i>	‘ Nice plays today’
<i>Tá fáilte roimh gach duine</i>	‘Everyone is welcome ’

Table 6.6: Irish extracts from questionnaire study.

Extract	Translation
<i>lmao!! Rud ar bith tusa ?</i>	‘ lmao !! Nothing you ?’
<i>Amhrán pop an lae</i>	‘ pop song of the day’
<i>Anois a chonaic mé é seo!!!! Wtf!</i>	‘Now I saw this!!!! Wtf! ’
<i>Raight. Shlog mé an t-iomlán</i>	‘ Right. I gulped it all’
<i>anois am réiteach don dioscó! 🕺🎉</i>	‘now it’s time to get ready for the disco! 🕺🎉’
<i>Comhghairdeas leis na leáids</i>	‘Congratulations to the lads ’
<i>bei 2 ag partyáil</i>	‘you will be partying ’
<i>An féidir leat rt an ocáid seo</i>	‘Can you retweet this event ’
<i>tá arán banana agam</i>	‘I have banana bread’

Table 6.7: Ambiguous extracts from questionnaire study.

Irish words Table 6.6 lists the Irish-only language extracts included in the questionnaire study and their translations. All of the highlighted words were present in NEID except *grma*. This is a very common initialism in informal Irish, standing for *go raibh maith agat*, meaning ‘thank you’.

Ambiguous words Table 6.7 lists the ambiguous language extracts included in the questionnaire study and their translations. We attempted to classify as ‘CS’ and ‘BOR’ selected extracts from tweets that contained a single English origin word in an Irish-language context by applying the criteria C1-C4. However, some words were more difficult to classify so another category ‘AMBI’ was created in order to investigate the status of these more ambiguous words. The popular internet initialisms ‘lmao’ and ‘wtf’ were included in this category. Álvarez-Mellado and Lignos (2022) do not specifically mention such terms in their borrowing criteria. In fact, they considered such words to be part of the internet jargon and annotated them as Spanish, however, we strongly suspected that Irish speakers would not perceive them as Irish words. Another interesting case is that of cognates that are rendered with the same spelling in both English and Irish. We included the words *pop* and *banana* in the ‘AMBI’ category as we found them more difficult to classify. We also included the words *raight* ‘right’, *leáids* ‘lads’, and *dioscó* ‘disco’, as we found that these could be interpreted as transliterations of English words or as borrowed words that have been adapted into Irish. Finally, we included *ocáid* ‘event’ because of its subtle spelling error. The standard spelling is *ócáid*. The full phrase *An féidir leat rt an **ocáid** seo* ‘can you retweet this **event**’ also has a nonstandard grammatical structure. As such variation is common in UGC, we did not correct any errors in the extracts.

6.3 Pilot Test

This section describes the lessons learned through testing the survey and improvements made via feedback from test participants. Two iterations of the questionnaire were each pilot tested on five participants from the target population to ensure that the questions were clear and easy to understand and that the data collected would be useful for the research questions. Several small changes were made to the questionnaire in light of feedback received during these tests.

Reason to use word
It is adequate/works/makes sense
It is handy/easy to use
It is correct
I cannot think of suitable equivalent
It is widely recognised/universal/easy to understand
I has a specific sense that would be lost if translated
I like it
It is humorous
It is more efficient than its translation
I would use this word or its translation interchangeably
Reason not to use word
I would prefer a term with the same meaning in the same language
I would prefer a term with the same meaning in another language
It is incorrect
I do not know it
I do not like it

Table 6.8: Response options for Part C questions considered in pilot study.

Open- vs. closed-ended question We wanted to gather information on what factors influenced whether or not a participant would use particular code-switched or borrowed words. We considered whether to use a quantitative approach to capture this information via closed-ended questions. In the pilot study, we used open-ended questions to get an idea of how much variation would appear in the responses. Table 6.8 lists the options we considered for the closed-ended version of Part C questions in which they explain why they would (not) use the highlighted word in the provided extract. These options were derived from the responses to the pilot test but we concluded that we would lose a lot of the richness of the data by prompting the participants with predetermined options. We ultimately decided to use open-ended questions. This allowed us to utilise qualitative methods to gain a deeper understanding of the perceptions of Irish speakers in their own words.

Rephrasing ambiguous question The pilot test brought our attention to the inadequate wording of the questions in Part B. Specifically, the original question in the pilot test was as follows: “Would you be likely to phrase the sentence 1a using the word in bold in an informal Irish-language setting?” Understandably, this led participants to interpret the question as asking about the likelihood of encountering such a phrase in their everyday speech. However, our actual intention was to investigate their opinion on the specific word highlighted in the phrase and whether they would use that precise word in the given context. To clarify the intended meaning and focus the question on the desired data, we rephrased it as follows: “If you were to phrase sentence 1a in an informal Irish context, how likely is it that you would include the highlighted word?” Although we acknowledge that the revised question is still lengthier than desired, we observed that, for the most part, the intended meaning was effectively conveyed in the results.

Additional information in Language Proficiency Question Participants of the pilot study found it difficult to place themselves within the language proficiency options provided. To solve this issue, we added a description of each language proficiency level.

Clearer ‘both’ options The pilot study gave participants four language options in each Part A question: ‘Irish’, ‘English’, ‘Both’, ‘Neither’. We realised that ‘Both’ was ambiguous as it could imply that the word was a full member of both Irish and English or that one part of the word was Irish and another part was English. We therefore added a fifth option and specified the difference in meaning, resulting in the following options: ‘Irish’, ‘English’, ‘the word exists in both Irish and English’, ‘the word is a mix of Irish and English’, and ‘Neither’.

6.4 Data Collection

The population targeted in the study is self-reported Irish-language speakers over the age of eighteen of all levels and dialects. The survey was distributed over a period of two months, February and March 2023, to as many Irish speakers as possible via social media, email, and physical posters, targeting Irish-language centres, universities, and libraries in particular. We aimed to reach a convenient sample of the Irish-speaking population. The high-end estimate of the number of Irish speakers at the time of survey creation was 1.7

million (CSO, 2016), therefore we aimed to reach 385 respondents for a margin of error of 5% and a confidence level of 95%. Ultimately, we received 256 responses meaning that the margin of error for our results is 6.13%.

6.5 Results and Analysis

This section describes the demographic characteristics of the participants and presents the findings of the study in relation to the research questions.

6.5.1 Demographic Distributions

What is your age? Figure 6.4 shows the number of participants in each age group. The age distribution among participants revealed an overrepresentation of the 25-34 age group and an underrepresentation of the 65+ age group. Specifically, 69 participants (26.95%) fell within the 25-34 age range. Considering the Irish population over the age of 18, we would expect this age group to constitute approximately 11% of the sample. In contrast, only 8 participants (3.13%) belonged to the 65+ age bracket, whereas an expected distribution of 18% was anticipated. The remaining age groups were distributed within a 5% range of what was expected. This discrepancy can be attributed to several factors. Firstly, the mode of distribution primarily relied on internet-based platforms, which may have attracted a larger proportion of participants from the younger age range. Additionally, the age composition of the author's social circle could have influenced the recruitment of participants, leading to a higher representation of individuals within the author's age demographic.

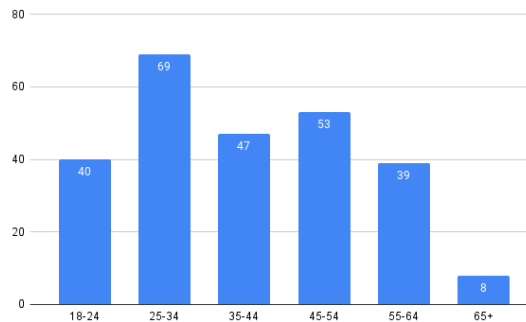


Figure 6.4: Count of participants in each age group.

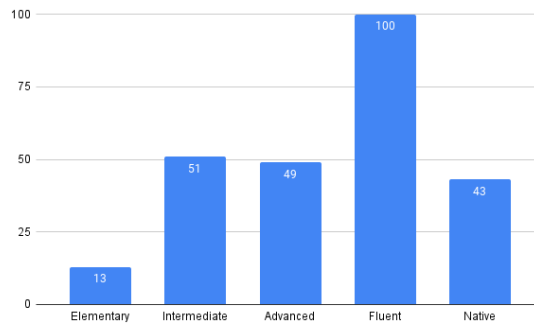


Figure 6.5: Respondents' self-reported level of Irish.

What is your level of Irish? Figure 6.5 shows the number of participants at each level of Irish-language proficiency. Among the 256 respondents, 13 individuals (5.08%) reported their level as elementary, while 43 (16.8%) indicated they were native speakers. There were 49 respondents (19.14%) who identified their level as advanced, followed by 51 (19.92%) who considered themselves at an intermediate level. The largest group consisted of 100 individuals (39.06%) who reported being fluent in Irish. Figure 6.6 shows the number and distribution of participants' levels of Irish among the various age groups. Notably, the survey had no respondents at the elementary level in the 65+ age group.

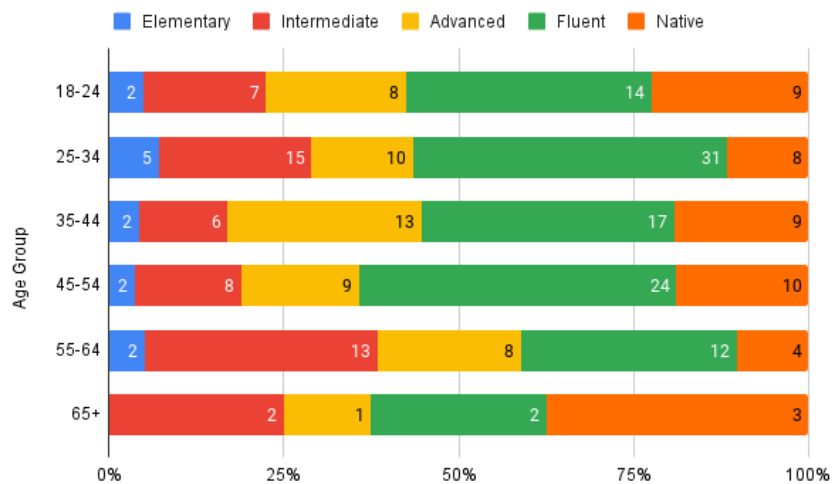


Figure 6.6: Participants' level of Irish broken down by age group.

Which dialect(s) of Irish do you identify with? Each dialect was well represented with the largest group claiming to have a mix of dialects. Seven respondents made use of the 'Other' category to give more information on their dialect. Three of these respondents identified with a Leinster/Dublin dialect of Irish. Due to the density of the population,

Dublin is an Irish-language hub in Ireland, however, this region has not been traditionally included in the dialects of Irish.

How many languages are you proficient in? We observed that the majority of respondents reported being proficient in multiple languages. The highest number (49.61%) of respondents reported proficiency in two languages, followed by proficiency in three languages (30.08%). Just 10.94% of respondents were proficient in only one language. Finally, 9.38% of respondents demonstrated a high level of multilingualism by reporting proficiency in 4 or more languages. These results highlight the prevalence of multilingualism in the surveyed group.

How regularly do you use Irish in a formal context? e.g. professional email 47.66% of respondents reported using Irish in a formal context daily, indicating that they incorporate the language regularly into their professional communications. A significant portion of respondents, 28.13%, indicated using Irish in a formal context less than once a month, suggesting infrequent or sporadic usage. Approximately 8.59% reported using Irish in a formal context on a monthly basis, indicating occasional utilisation of the language in professional communication. Finally, 15.63% of respondents reported using Irish in a formal context weekly, suggesting a moderate level of frequency in incorporating the language into their professional interactions.

Do you mix Irish and English in formal contexts? e.g. professional email The largest group of respondents, accounting for 39.84%, indicated that they never mix Irish and English in formal contexts. 29.69% of respondents reported rarely mixing the two languages, while 17.19% indicated they sometimes engage in code-mixing. Just 7.81% and 5.47% responded with 'often' and 'always' respectively. These results indicate an aversion to code-switching in a formal setting.

How regularly do you use Irish in an informal context? e.g. texting, chatting The majority of respondents (63.67%) reported using Irish in informal contexts on a daily basis, indicating a high frequency of incorporating the language into their informal communication. A smaller portion of respondents (24.61%) indicated using Irish in an informal context on a weekly basis, suggesting a regular and consistent usage pattern.

A small percentage of respondents reported using Irish in an informal context less frequently, with 7.42% indicating a monthly usage and 4.3% reporting usage less than once a month. The data reveal a strong presence of daily and weekly usage of Irish in informal contexts, highlighting a consistent and active engagement with the language in day-to-day conversations.

Do you mix Irish and English in informal contexts? e.g. texting, chatting

The largest proportion of respondents (30.86%) indicated that they sometimes mix Irish and English in informal contexts. A significant portion of respondents (28.52%) reported mixing Irish and English often. Additionally, 18.36% of respondents expressed a tendency to mix the languages to some extent, while 17.97% reported consistent code-mixing in informal contexts. A small proportion of respondents (4.3%) indicated that they never mix Irish and English in informal contexts. Overall, the results suggest a higher level of code-mixing in informal rather than formal communication.

How do you feel about mixing Irish and English? As shown in Figure 6.7, respondents mostly felt neutral with regard to mixing Irish and English, with the second most popular category being ‘very positive’.

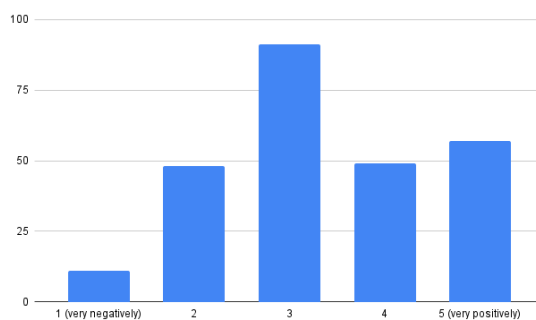


Figure 6.7: Respondents’ feelings about mixing Irish and English.

Figure 6.8 illustrates the average scores provided by participants when asked about their sentiments towards mixing Irish and English, categorised according to their self-reported level of Irish proficiency. The average scores for individuals at the elementary, intermediate, and advanced levels fall within the “moderately positive” range. Participants at the fluent level, on the other hand, tend to have a “neutral” sentiment. Lastly, native speakers’ average score indicates a “slightly negative” sentiment.

These results suggest that individuals with lower levels of language proficiency exhibit

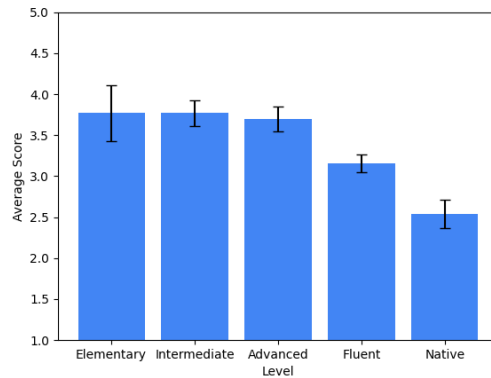


Figure 6.8: Respondents’ feelings about mixing Irish and English broken down by Irish proficiency. A score of 5 corresponds to ‘very positive’ and 1 corresponds to very negative.

a more positive attitude towards mixing Irish and English. This could indicate that language-mixing is used as a tool for language learning.

Average ratings are relatively consistent across age groups, with slight variations, except for the 65+ age group, which shows a lower average rating and higher standard deviation compared to other age groups. This, however, can be attributed to the low number of respondents over the age of 65.

6.5.2 Language Classification of BOR and CS Words

In order to answer RQ3.1 ‘Do Irish speakers classify borrowed words as English less often than code-switched words?’, we analysed the responses to the instruction ‘please name the language of the highlighted word’. Figure 6.9 shows the percentage of responses that classified each BOR and CS word as English. An unpaired t-test resulted in a P value of 0.0571 indicating that there is a 5.71% chance of observing the observed data or more extreme results under the assumption that there is no significant difference between the compared groups.

Very few respondents considered *meaitseáil* ‘matching’, *haiceanna* ‘hacks’, *blag* ‘blog’, and *aip* ‘app’ to be English-language words suggesting a level of integration to the Irish language. The other BOR words and indeed all of the CS words were considered by the majority of respondents to be in English. Though the average BOR value at 52.3% is much lower than the average CS value of 85.55% as hypothesised, these results raise the question of whether *tweet*, *DM*, *hits*, *twerking*, and *keyboards* would be more accurately classified as code-switching rather than borrowed words. Referring back to Table 6.4, we

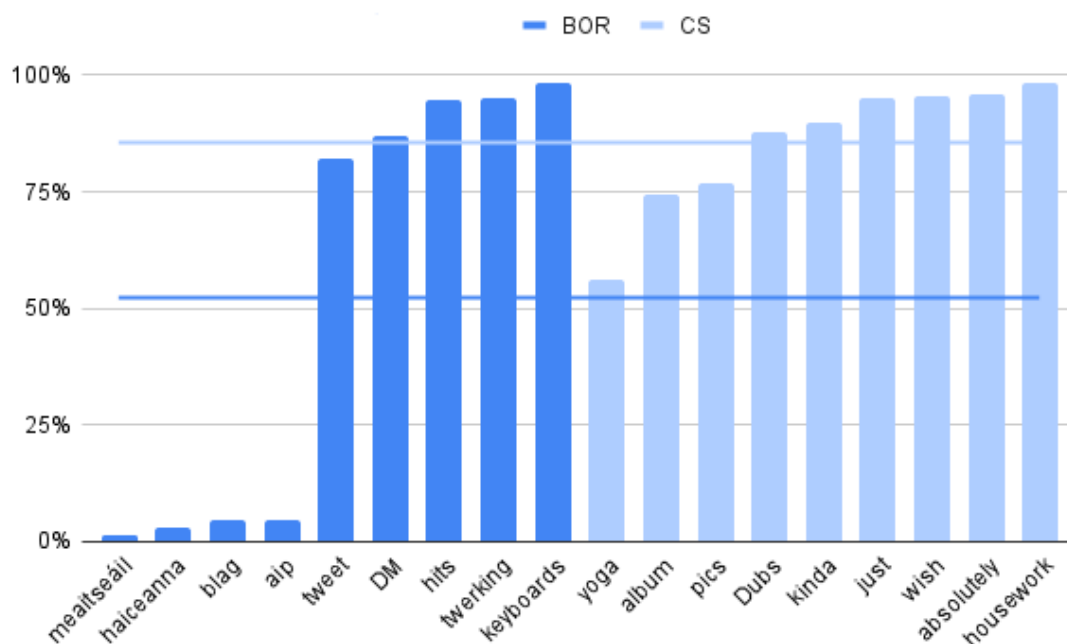


Figure 6.9: Percentage of responses classifying BOR and CS words as English. The dark blue line shows the average BOR value: 52.3%. The light blue line shows the average CS value: 85.55%.

note that C3 (listedness in NEID) is successful as a criterion for classifying borrowings, insofar as it aligns with the perceptions of respondents. The criteria C1, C2, and C4, however, did not align with participants' opinions. Words selected using the criteria C1, C2, and C4 were most likely to be classified as English words by participants indicating that these words were not perceived to be integrated into the Irish language. Full language classification results are provided in Appendix E.

6.5.3 Usability of BOR and CS Words

In order to answer RQ3.2 'Do Irish speakers claim to be more likely to use borrowed words than code-switched words?', respondents were prompted to assign a 'likelihood of use' score to certain words. Figure 6.10 shows the mean scores given by respondents when asked if they would use these BOR and CS words in a given context. A score of 5 corresponds to 'very likely' and 1 to 'very unlikely'. Overall, the average score for BOR words was 3.16 and the average score for CS words was 2.65. The unpaired t-test produced a P value of 0.137, suggesting that there is a 13.7% probability of obtaining the observed data or more extreme results if there is no significant difference between the groups being compared. Figure 6.11 shows how these scores compare to the overall mean scores for all word groups.

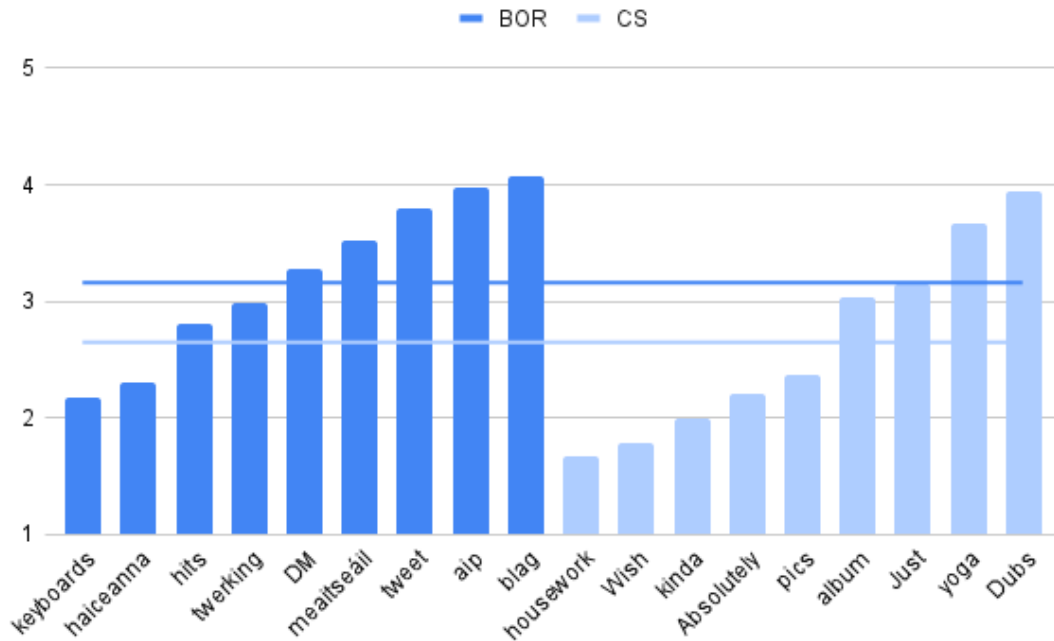


Figure 6.10: Mean ‘likelihood of use’ score for BOR and CS words where 5 is ‘very likely’ and 1 is ‘very unlikely’. The dark blue line represents the average score for BOR words. The light blue line represents the average score for CS words.

GA had the highest ‘likelihood of use’ score at 4.58.

6.5.4 Reflexive Thematic Analysis of Open-Ended Responses

To answer RQ3.3 ‘What themes can be interpreted from Irish speakers’ explanations about word choice?’, we used the six phases of reflexive thematic analysis (Braun and Clarke, 2006, 2021) to examine the open-ended questions: 1. Data familiarisation and writing familiarisation notes; 2. Systematic data coding; 3. Generating initial themes from coded and collated data; 4. Developing and reviewing themes; 5. Refining, defining and naming themes; 6. Writing the report.

Phase 1: Data familiarisation and writing familiarisation notes The first step was to carefully read and re-read the subset of open-ended questionnaire responses to become acquainted with the data. During this process, we noted that many of the responses were in Irish despite the survey being in English. We interpret this as a personal preference on the part of the respondent and possibly a way of asserting their linguistic identity. We also noted a variety of tones and differing opinions in the data, e.g. where some respondents found a phrase ‘fun’ or ‘humorous’ others found the same phrase to be

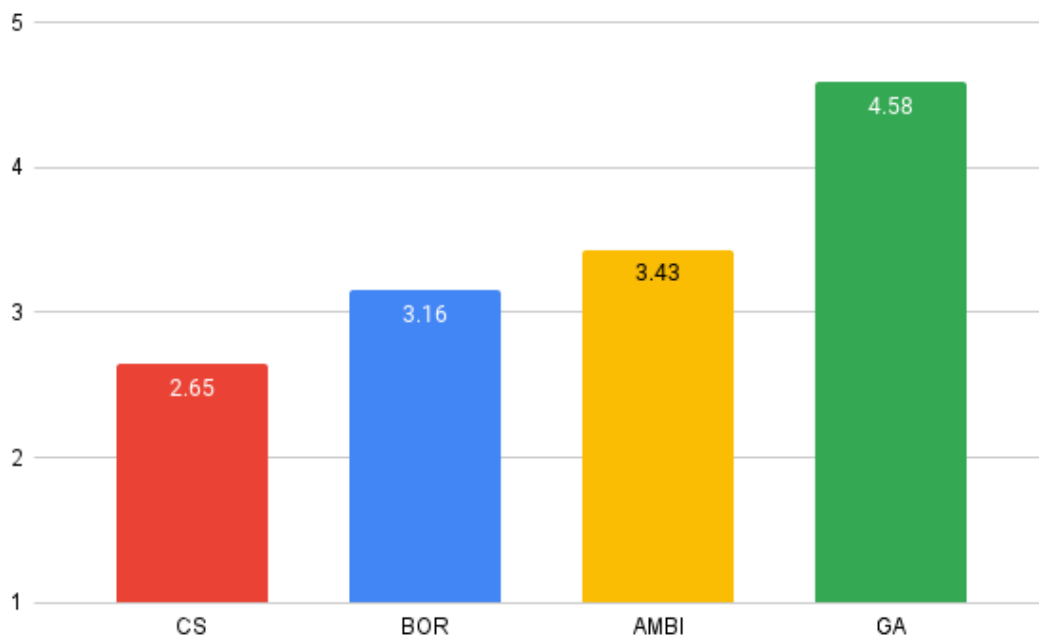


Figure 6.11: Overall mean ‘likelihood of use’ scores for CS (code-switched), BOR (borrowed), AMBI (ambiguous), and GA (Irish) words.

‘wrong’ or ‘awkward’. The data were explored thoroughly giving equal time and attention to each response. We then imported the data into the NVivo data management tool.²

Phase 2: Systematic data coding Noteworthy features of the data were categorised systematically into succinct, descriptive initial codes. Codes can be conceptualised as the elemental units that form subsequent themes. Our initial codes were non-hierarchical and were derived inductively from the words of the participants rather than from predefined codes based on theory. Significant and informative aspects of the data items were identified to facilitate theme development. A single response often corresponded to more than one code. Gradually, throughout this process, a description of each code was created to elucidate the shared characteristics among constituent data items. Table 6.9 shows the ten most frequent codes with a definition and example as they appeared at the end of this phase. The initial codes reflect various aspects of the responses, such as the speakers’ proficiency and attitudes towards Irish, the context of language usage, and the social and cultural identity associated with particular words.

²<https://lumivero.com/products/nvivo/>

Code	Refs	Definition	Example
Irish	213	Irish language or Irishness	A 'new' word that may not sound Irish enough for me
Preferred term	189	Suggestion of another way to express the same concept	Would use cineál instead
Vocabulary	130	Set of available words at the level of the individual or the language community	No Irish word that I know explains it
Personal taste	119	Expression of like, dislike, individual opinion	It's just an ugly word.
Usage	116	Use of term	It's normal among Irish speakers
Understandability	103	Comprehensibility	Accessible phrase for people at all stages
English	81	English language or Englishness	Too Englishy
Naturalness	58	Intuition	Donna about that one ... just doesn't flow.
Familiarity	56	Recognition, knowledge	The Irish I grew up with.
Ease	54	Accessibility	Always use the word pics when im writing an email in Irish - its easy and quick

Table 6.9: Phase 1 codes with the corresponding number of references, a definition and example (abbreviated here, full version in Appendix F).

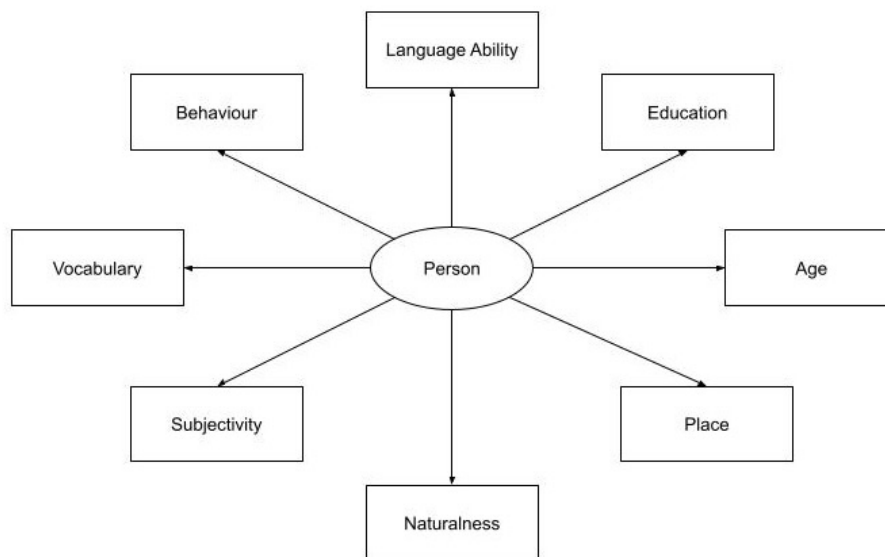


Figure 6.12: Map of candidate theme 'Person'.

Phase 3: Generating initial themes from coded and collated data With all relevant data items coded, codes were revised and further analysed in order to identify potential themes. This involved merging and grouping codes that had overlapping or related

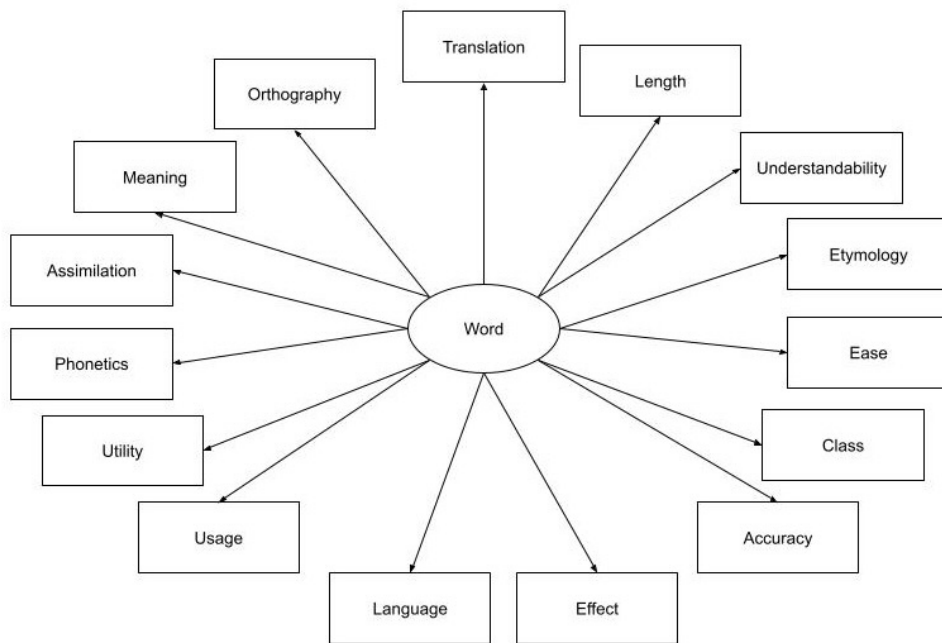


Figure 6.13: Map of candidate theme 'Word'.

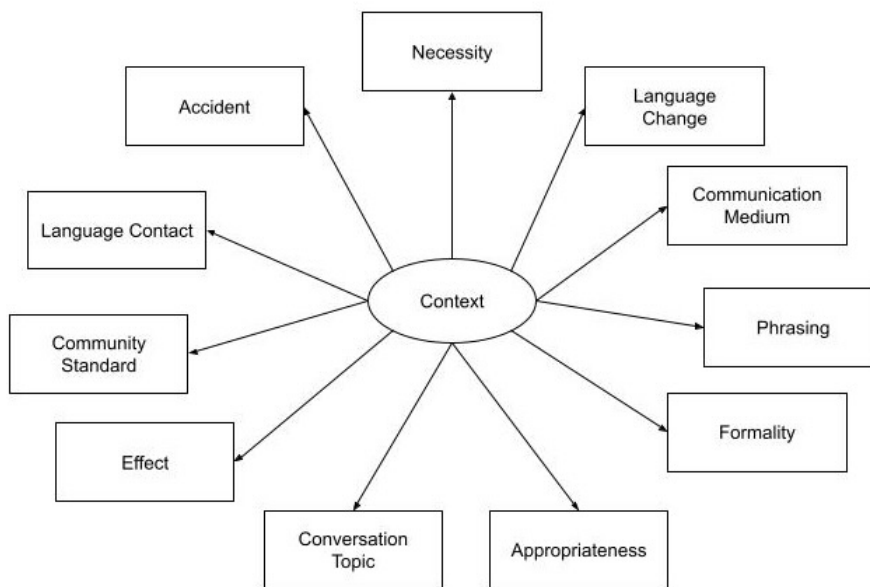


Figure 6.14: Map of candidate theme 'Context'.

meanings, e.g. the codes 'clarity' and 'understandability' were grouped together during this phase. Codes were split into multiple codes where distinct meanings were observed

within a code e.g. the ‘age’ code was used sometimes to refer to the age of a respondent and sometimes to refer to the age of a term so they were split into separate themes. Hierarchical structures were also implemented where one code encompassed another, e.g. codes ‘social media’, ‘email’, ‘texting’, and ‘speaking’ were grouped under ‘communication medium’. These themes were then organised into a framework guided by RQ3.3. The importance of a theme was not determined by the number of data items that supported it. Rather, themes were included where a pattern communicates something meaningful with regard to the research question (Braun and Clarke, 2006). At this stage, any prospective themes irrelevant to the research question were discarded. For example, the initial code ‘Preferred term’ was used to describe responses that provided a term that they preferred to the word in question. ‘Preferred term’ was discarded as a potential theme as it doesn’t bring us any closer to answering the research question which pertains to the reasons behind such preferences. During this phase, the 80 initial codes were distilled to 34 candidate themes, grouped by descriptive categories ‘person’, ‘word’, and ‘context’. The goal was to strike a balance between the number of themes and the depth of analysis conducted for each. Figures 6.12, 6.13, and 6.14 show the development of candidate themes in this phase and the first attempt at hierarchical organisation.

Candidate Theme	Description
Desire to be understood	Preference for clear and easily understandable language communication.
Desire for convenience	Preference for language choices based on convenience or practicality.
Desire to fit in with social norms	Motivation to adhere to societal expectations and norms in language usage.
Preference for Irish over English	Desire to prioritise using Irish language instead of English.
Limited by lack of proficiency	Constraints due to insufficient language skills or limited exposure to Irish.
Desire for creativity/fun	Interest in expressing creativity and enjoyment through language usage.

Table 6.10: Phase 4 candidate themes and descriptions.

Phase 4: Developing and reviewing themes The candidate themes, shown in Table 6.10, were revised during this stage to ensure that the data within each theme cohered together meaningfully. Iterations of the thematic map were generated to check whether the themes worked in relation to the coded extracts. At this point, we moved from

the semantic description of the responses to the interpretation of the latent meaning by inferring underlying thoughts or desires motivating the particular responses. For example, when asked for reasons they would (not) use a word, one respondent said: “Occasionally slips out but I’d be a little more likely to use ‘... ar fad’ or similar”. On the surface level, the respondent’s statement could be simply coded under ‘preferred term’, however, deeper analysis is needed to understand the underlying meaning, emotions, and attitudes expressed by the participant. The respondent’s use of the phrase “occasionally slips out” implies that they may use the word in question unintentionally. However, their preference for using ‘ar fad’ or a similar phrase suggests a conscious choice to express themselves differently. This subtle preference might indicate a desire to conform to certain linguistic patterns or cultural norms. By considering the response at a latent level, the analysis aims to uncover the aspects that may not be explicitly stated. It involves looking beyond the literal level of the words and exploring the motivations, emotions, and attitudes that may shape the participant’s response. The candidate themes were distilled down to six candidates for key themes ‘preference for Irish over English’, ‘desire for creativity/fun’, ‘desire to be understood’, ‘limited by lack of proficiency’, ‘desire to fit in with social norms’, and ‘desire for convenience’.

Phase 5: Refining, defining and naming themes In this phase, the candidate themes were further distilled so that the essence of each theme could be identified. Table 6.11 shows the final themes and their descriptions. The candidate theme ‘preference for Irish over English’ was renamed ‘language contact’ because it better encompassed the pattern of responses expressing opinions on English-origin words in Irish-language contexts. The candidate themes ‘desire to be understood’, ‘desire for convenience’, and ‘desire to fit in with social norms’ were renamed ‘clarity’, ‘convenience’, and ‘conformity’ respectively to convey the same concepts more concisely. Finally, the candidate themes ‘desire or creativity/fun’ and ‘limited by lack of proficiency’ were discarded at this point as they did not comprise a prevalent pattern in the responses.

Phase 6: Writing the report In this phase, we aim to tell the story of the data by describing each theme and selecting extracts that help us to understand the perspectives of Irish speakers regarding word choice in the context of language contact.

Theme	Description
Clarity	The preference for commonly understood abbreviations, acronyms, or familiar words in a conversation or communication rather than unfamiliar or ambiguous terms, to ensure clarity and facilitate understanding.
Convenience	The inclination towards English words or abbreviations in Irish-language contexts due to lack of familiarity with the Irish equivalent, ease of use, absence of a direct translation, and the perception of English terms as more widely understood.
Conformity	The desire for communication to align with norms of language usage.
Language Contact	The preference for Irish equivalents or creating new Irish words rather than English loan words and mixing Irish with English.

Table 6.11: Final themes and descriptions

Clarity The theme of clarity is evident in the responses where participants emphasise the importance of using clear and easily-understood language. We observed a preference for familiar words to prioritise effective communication and mutual understanding between speakers. In the following quote, a fluent Connacht Irish speaker aged 18-24 describes their choice to use the word ‘tweet’ in order to minimise confusion.

“tweet is almost like a brand name and to gaeilge-ise it would cause confusion”

Many respondents also took into consideration the recipient’s language proficiency, claiming to adjust their choice of words accordingly. In summary, clarity is a key factor affecting word choice. While many respondents felt an aversion to the use of English words in an Irish-language context, many expressed a willingness to adapt their vocabulary to incorporate English words to avoid ambiguity or misunderstanding.

Convenience The theme of convenience is evident in the data where participants highlighted their tendency to opt for words that are easier or quicker to use. Many individuals mention their lack of knowledge or familiarity with the Irish equivalents of certain words, leading them to choose the English terms instead. Some refer to the convenience of using abbreviations, acronyms, or informal language, as they provide a shorter and more efficient way to communicate. The desire for simplicity and ease is apparent throughout. For example, a native Munster-Irish speaker aged 45-54 claimed:

“b’fhéarr liom leaideanna a úsáid ach uaireanta cioraím é go leaids”

(I would prefer to use ‘leaideanna’ but sometimes I shorten it to ‘leaids’.)

This neatly summarises a conflict of values that was present throughout the data. The desire to avoid language contact is sometimes outweighed by the desire for quick com-

munication. Simple language is seen by some participants as a good thing and by some as implying a laziness as exemplified by the following quote from a fluent Munster-Irish speaker aged 45-54 in reference to the word ‘kinda’.

“English word, lazy”

This tension between convenience and linguistic purity emerged as a recurring theme in the data, highlighting the ongoing negotiation between individuals’ desire for quick and efficient communication and their commitment to preserving the Irish language. Ultimately, the theme of convenience permeated the data, with participants acknowledging their tendency to choose words that were easier or quicker to use. However, the acceptance or rejection of simplified language varied among individuals, reflecting a broader conflict of values and attitudes towards language use.

Conformity We interpret the theme of conformity in respondents’ claims of their word choice being influenced by whether a term is “widely used” or “commonly accepted”. Such statements demonstrate a desire to conform to linguistic norms. For example, a fluent Irish speaker aged 18-24 of no particular dialect, referring to the word ‘just’, stated:

“Códmháirt arís. Ceapaim gur cleachtas coitianta é, tá sé le cloisteáil go han-mhinic ach nach focal Gaeilge é dá bharr sin. Más cleachtas coitianta é go ginearálta, is cinnte go mbainfinn úsáid as dá mbeadh abairt le cumadh go neamhfhoirmiúil.”

(Code-switching again. I think it is common practice, one hears it regularly but that doesn’t make it an Irish word. If it is generally common practice, I’m sure I would use it to compose an informal sentence.)

This response indicates that such language conformity could also happen subconsciously. Ultimately, the data suggest a strong inclination to conform to established linguistic practices and maintain consistency in communication.

Language contact This theme captures participants’ negative and positive views about language contact phenomena. Negative feelings about language contact include the idea that English words are not acceptable in an Irish-language context. For instance, a native Connacht-Irish speaker between the ages of 35 and 44, referring to the word *just* expressed their viewpoint:

“Úsáidim sa chaint é ach drochnós atá ann”

(I use it when talking but it is a bad habit)

On the other hand, we found that language contact was deemed more acceptable if it was perceived to be employed by native speakers. For instance, an advanced Connacht-Irish speaker aged 25-34, referring to the word *pics* stated:

“It’s a shortening based on the English plural as opposed to a loan word, though it’s probably still used among native Gaeltacht speakers, so it’d be fine in general.”

We also observed that positivity about language contact was associated with creativity and fun, whereby respondents claimed to enjoy mixed language neologisms. For example, an elementary-level speaker of Connacht Irish stated:

“I like this Irish-ifying of English words”

This indicates that language contact is also seen as a creative way to make Irish more accessible to learners. In conclusion, language contact influences word choice in a variety of ways such as aversion to using English words in Irish-language contexts, acceptance of certain terms once they have reached some level of integration, and enjoyment of neologism creation via language contact.

6.5.5 Limitations

The following limitations should be taken into account when interpreting the results.

Sample size As mentioned in Section 6.4, the survey was completed by a limited sample of 256 respondents, meaning that the margin of error for our results is 6.13%, based on a total population of 1.7 million Irish speakers in Ireland. In our survey, the majority of respondents were advanced, fluent or native speakers of Irish which is not representative of the population of Irish speakers. According to CSO (2022), of the total population of Irish speakers, 10% spoke the language very well, 32% spoke it well, and 55% of people who indicated that they spoke Irish did not speak the language well. Based on feedback from respondents, we conclude that disparity is likely due to learners of Irish not identifying as an Irish speaker and not feeling confident in their ability to complete the survey. CSO (2022) does not provide information on the distribution of dialects of Irish speakers. Future studies aiming to collect data from a more representative sample may target more learners of Irish at beginner and intermediate levels.

Confidence of Irish speakers Explicitly inquiring about participants' confidence levels with regard to responding to questions in an Irish-language context could have contributed valuable insights to our analysis. While we didn't incorporate 'confidence' as a conclusive theme, due to insufficient evidence, we noticed that even some fluent speakers and native Irish speakers expressed uncertainty about their responses.

Questionnaire format The format of the study as a questionnaire is likely to have affected the quality of data. As Agheyisi and Fishman (1970) note, open questions are less successful in questionnaires than in interviews because the effort required to write the answers may lead to the respondent providing a lack of detail. In an interview or focus group, questions can also be adapted to elicit more information. Additionally, in spoken surveys or focus groups, pauses, tones, and inflections can be useful for analysis. Another limitation of the questionnaire format is the amount of data that can be tested. We limited the focus of our investigation to 36 words so as not to induce fatigue, frustration, or boredom among the respondents.

Self-reported data We recognise that our study was based on self-reported data, which may not reflect the actual behaviour or attitudes of the respondents. Self-reported data can be influenced by social desirability bias, memory bias, or lack of awareness.

Anonymity The anonymous nature of the questionnaire also meant that we had no ability to verify the accuracy of the responses.

Bias We note the possibility of a volunteer bias affecting the findings, as the individuals who opted to participate in the study may not accurately represent the overall population. Moreover, online questionnaires may not reach people who do not have access to the internet or who are not comfortable with technology, which can also introduce a selection bias. We acknowledge the age imbalance in our sample may be due to distributing the survey via the internet.

Missing demographic information Finally, the omission of certain demographic information, such as gender and education level, could have provided valuable insights and helped measure the representation of different cohorts.

6.5.6 Summary

This chapter has reported on a questionnaire study of language contact phenomena in Irish-language tweets, particularly the distinction between code-switching and borrowing. Recognising the lack of consensus in the research community on how best to represent language contact phenomena in NLP resources, we have described the anonymous online questionnaire conducted among 256 adult Irish speakers to investigate their perceptions and usage of code-switched and borrowed words. We outlined the background work that we have drawn from in the design of our study and the research which informed our initial classification of code-switching and borrowing for the extracts used in the study. We have then described the specific research questions investigated in the study as well as the variables and methods of analysis that correspond to each. Additionally, we have explained the rationale behind the design choices made during the development of the questionnaire study. We have also described the example text from tweets and questions used in the survey. We have documented the lessons learnt in the pilot test which led to improvements in the survey design. We have then described the method of questionnaire dissemination.

Moving on to the findings, we have reported the breakdown of respondent demographics, presented a statistical analysis of the quantitative survey responses, and documented the phases of reflexive thematic analysis of the qualitative survey responses. The results supported the hypotheses that borrowed words were considered ‘less English’ than code-switched words, and Irish speakers were more likely to use borrowed words in an Irish-language context. However, neither result is statistically significant. It is clear that the criteria used to distinguish between code-switching and borrowing in this study did not align with the perceptions of Irish speakers. If we were to reclassify all words based on the results of the study, the criterion C3 (listedness in NEID) appears to be the strongest indicator of loanword status. We hope that future work in NLP development for Irish can build on these insights, incorporating an awareness of the prevalence of language contact.

Through qualitative analysis, we have identified four key themes of **clarity**, **convenience**, **conformity**, and **language contact** that influence word choice among Irish speakers. This study has also highlighted the stigma surrounding the mixing of Irish and English, as well as a lack of confidence among individuals regarding how and by whom the language should be spoken. These social factors can complicate language revitalisation

efforts. Also evident in our analysis was how closely intertwined the Irish and English languages are in today's Irish society. Our findings may help to mitigate negative attitudes towards Irish as a result of an often polarised view on the separation of the language communities. Furthermore, we have acknowledged the limitations of the study including the small sample size, potential volunteer bias, self-reporting discrepancies, the potential influence of question-wording, age imbalance due to online survey distribution, and the omission of important demographic information, which should be considered when interpreting the findings. Ultimately, these findings contribute to a better understanding of language contact and variation and have potential applications in NLP, language education, policy, and social integration among linguistic communities.

Chapter 7

Conclusion

In this concluding chapter, we reflect on the key findings and insights of the research. In Section 7.1 we summarise the preceding chapters by highlighting the contributions of our research, namely, we presented the novel resource, TwittIrish, the first UD treebank for Irish UGC, a linguistic analysis of Irish-language tweets, parsing experiments and error analysis using the TwittIrish dataset, and a questionnaire study exploring best practices with regard to the categorisation of language contact phenomena in Irish tweets. In Section 7.2, we revisit the research questions posed in Section 1.2. In Section 7.3, we acknowledge the limitations of the research. In Section 7.4, we describe potential avenues for future inquiry. Finally, in Section 7.5, we underscore the significance of our research highlighting their implications in the field of NLP and beyond.

7.1 Contributions

The contributions of this research are as follows:

1. The primary contribution of this research is **the TwittIrish treebank**. This is the first syntactically annotated corpus of Irish-language UGC. The treebank is available for download through the UD GitHub repository¹. We have documented the methodology employed in creating TwittIrish in Chapter 3. The TwittIrish annotation guidelines are provided in Appendix B and Irish-specific annotation guidelines are maintained on the UD website². These are valuable references for researchers undertaking similar projects.

¹https://github.com/UniversalDependencies/UD_Irish-TwittIrish/tree/master

²<https://universaldependencies.org/ga/index.html>

2. We present a **linguistic analysis of Irish language tweets** in Chapter 4. We have presented several key insights on the levels of orthography, morphology, lexicon, and syntax. By providing this in-depth exploration of the linguistic genre of Irish UGC, we facilitate the development of efficient technologies that are compatible with existing resources and cater to user requirements.
3. Using TwittIrish as training data, we have built the first UGC-based parsing model for Irish, achieving **parsing accuracy of almost 80 LAS on Irish-language tweets**. As documented in Chapter 5, we first established baseline parsing results for Irish-language tweets of $\tilde{48}$ LAS using a parser trained on a treebank of standard Irish. We improved this baseline by ~ 11 LAS through the use of encodings from a monolingual Irish BERT model. Finally, our highest scores were achieved using TwittIrish training data in combination with training data from the treebank of standard Irish.
4. We have presented the results and analysis of a **questionnaire study of language contact in Irish** in Chapter 6. We adapted the criteria of Álvarez-Mellado and Lignos (2022) for identifying borrowing and code-switching. We label these words as BOR and CS respectively. Our results show that, in the context of informal Irish, BOR words were classified as English less often than CS words and that BOR words were more likely to be used than CS words. This study has found that the themes of **clarity, convenience, conformity**, and **language contact** inform preferences with regard to word choice.

7.2 Research Questions Revisited

RQ1: How do Irish tweets differ from standard edited Irish text? Arising from factors such as character limits imposed by platforms like Twitter, regional dialect differences, self-expression, errors, automatic text generation or translation, and language contact, we observe linguistic variation in Irish tweets on the levels of orthography, morphology, lexicon, and syntax. We have classified the orthographic variation that we have observed in Irish-language tweets into the following categories: diacritic, abbreviation, lengthening, nonstandard capitalisation, punctuation variation, transliteration, and hypercorrection. Our analysis of morphological variation focused on mixed-language tokens,

where English and Irish elements coexist within a single word and morphological differences across dialects are explored. Our analysis also explores lexical variation in Irish tweets. The vocabulary of Irish-language tweets differs from standard Irish text, with regard to dialectal differences, initialisms, pictograms, truncation, and lone other-language items. Finally, our analysis explored the syntactic characteristics of Irish tweets as compared to standard text. We have classified the syntactic variation as contraction, over-splitting, code-switching, ellipsis, and non-sentential structure. These findings describe the specific linguistic phenomena that characterise Irish-language tweets and distinguish them from standard Irish text. Understanding these differences is crucial for developing accurate NLP tools for Irish social media content.

RQ2: What challenges are associated with parsing Irish tweets? In our error analysis, we have found that parsing Irish-language tweets presents various challenges due to the linguistic characteristics phenomena inherent in Irish UGC texts. We list here the linguistic features of Irish tweets that we have found to be associated with low parsing accuracy. Ellipsis, a common phenomenon in Irish-language tweets can result in incorrect root identification by the parser, leading to parsing errors. Code-switching, where multiple languages are used within a tweet, can disrupt the syntactic structure. Parsers may struggle to correctly attach foreign language elements, leading to errors in the dependency parse. Pictograms, such as emojis and smileys, which are common in tweets, can be incorrectly labelled as punctuation by parsers, impacting the accuracy of the parse tree. Non-standard punctuation usage in tweets challenges parsers' ability to segment sentences and determine sentence boundaries accurately. Usernames and hashtags are frequently integrated into tweets. Parser mislabelling of usernames and hashtags, such as attaching them as root or assigning incorrect dependency labels, can lead to inaccuracies in parsing. Spelling variation, such as diacritic omission, can lead to misinterpretation by parsers. Even small spelling differences can affect the dependency relationships and attachment points in the parse tree. Parsers might struggle to handle mixed-language tokens, impacting the accuracy of parsing results. Ultimately, we have found that the task of parsing becomes more challenging when tweets exhibit these phenomena, which diverge from the grammatical norms that parsers are traditionally trained on. Based on the large increase in parsing accuracy that have achieved in our experiments, we find

these challenges can be effectively addressed by utilising genre-specific training data and transformer-based word embeddings.

RQ3: How can language contact phenomena in Irish tweets be characterised?

Given the lack of consensus among researchers of linguistics and NLP on how to best distinguish between code-switching and borrowing, we tested a methodology proposed by Álvarez-Mellado and Lignos (2022) using a questionnaire study. Our findings indicate a clear difference between code-switching and borrowing, however, they also suggest that not all of the borrowing criteria of Álvarez-Mellado and Lignos (2022) align with the perceptions of Irish speakers. We observe that the criterion of ‘listedness’, whereby the word is listed in a dictionary, was the strongest predictor of a potentially borrowed or code-switched word being perceived as Irish suggesting that it has achieved some level of assimilation and is therefore borrowed rather than code-switched. Given the frequency of language contact in informal Irish and many other languages, and the demand for multilingual NLP, the insights this study has provided will be useful for future research and resource development.

7.3 Limitations

In this section we discuss several limitations that should be considered in the interpretation of the findings we have presented in this thesis.

User-generated content The use of UGC for research purposes raises some ethical and privacy concerns, imposing certain limitations on the study. UGC may contain sensitive information such as names, locations, contact details, and controversial opinions. This limitation impacted the research in that the data had to go through an extra step of anonymisation to prevent the identification of individuals in order to protect the privacy of users. A further limitation of this genre is that the availability of UGC data tends to fluctuate over time due to shifts in the popularity of platforms and changes to API access policies. Additionally, UGC as a genre evolves rapidly, potentially rendering the dataset outdated as the characteristics of UGC change over time.

TwittIrish data curation The TwittIrish treebank contains data from Twitter only, restricting the scope of this study. In lieu of other varieties of UGC in our dataset, more generalised conclusions cannot be drawn. Another important point to note is that the tweets in our treebank may not represent the diversity of the population. Some demographics and individual users are overrepresented in the dataset due to “Participation Inequality” (Duval and Ochoa, 2008) as described in Section 3.1.

TwittIrish annotation Annotation of the TwittIrish treebank was performed by a single annotator rather than many. However, efforts were made to mitigate bias and errors resulting from this limitation. Regular meetings were held with expert annotators at the beginning of the annotation process to discuss and correct any issues. Additionally, a quality review was conducted before the release of the TwittIrish test set as described in Section 3.4. Nonetheless, the absence of multiple annotators limits the ability to assess the reliability and consistency of the annotations.

Dependency parsing experiments It is important to acknowledge that our experiments have focused exclusively on the task of dependency parsing. No downstream NLP applications have been tested in our study.

Questionnaire data collection As described in Section 6.5.5, the questionnaire study was subject to some limitations. The sample size of 256 respondents is relatively small and potentially skewed by volunteer bias and accessibility issues. Thus it might not be reflective of the broader population of Irish speakers. The questionnaire format may have yielded less detailed responses compared to interviews. Finally, the reliance on self-reported data could have introduced a social desirability bias.

7.4 Future Work

The current study could be built upon in several ways such as expanding, enhancing or performing other kinds of analysis on the TwittIrish treebank. This data can also be used for further parsing experimentation, language model development, or in the creation or evaluation of downstream NLP applications. This section suggests potential areas for future research.

TwittIrish expansion The TwittIrish treebank described in Chapter 3 could be enhanced by incorporating further linguistic annotation such as named entity annotation. Moreover, expanding the size of the treebank by including additional data and more recent content, would further enrich its utility as a linguistic resource. These measures collectively could increase the value of the TwittIrish treebank. The annotation of Twitter data has opened the door for dependency parsing for other sources of Irish UGC such as reviews, transcripts of videos, other social media text, etc. Rather than starting from scratch with the development of a treebank for other kinds of UGC, the parsers described in Chapter 5 could be used to automatically parse new target data. The annotation guidelines of Appendix B could be adapted to other kinds of UGC and a bootstrapping process of automatic parsing and manual correction like the one described in Section 3.3.2 could be employed for rapid and efficient resource development.

Longitudinal study The linguistic analysis presented in Chapter 4 could be extended in various ways. For example, language variation could be tracked over time in a longitudinal study allowing the evolution of language patterns to be analysed. Such research would provide valuable insights for linguists and NLP developers by deepening our understanding of the evolution of Irish, supporting cultural preservation efforts, and facilitating the development of more effective language technologies for Irish speakers and learners.

Semi-supervised learning Another idea to consider is to experiment with semi-supervised techniques such as self-, co-, and tri-training which have been shown to improve accuracy in pre-neural parsing systems, in both in-domain (McClosky et al., 2006; Huang and Harper, 2009; Søgaard and Rishøj, 2010) and out-of-domain scenarios (Petrov et al., 2010; Sagae, 2010). Lynn et al. (2013) experimented with self- and co-training on an early pre-UD version of the Irish treebank. Wagner and Foster (2021) have shown that such techniques can still be effective in low-resource dependency parsing, even in the presence of contextualised word embeddings. Such techniques could provide further boosts to dependency parsing accuracy in the context of Irish UGC.

Multilingual and cross-lingual dependency parsing In Section 5.3, we experimented with simply concatenating Scottish Gaelic dependency trees to our training data. There are many ways to extend this research using more advanced multilingual depen-

dependency parsing techniques. Further experimentation could include the implementation of multilingual word embeddings such as multilingual BERT (Devlin et al., 2019). Language embeddings (Ammar et al., 2016), where information on the target language is incorporated into each word embedding, could also improve multilingual parsing performance. Another possible experiment could utilise static treebank vectors which encode treebank information (Stymne et al., 2018) or dynamic treebank vectors that the parsing model dynamically interpolates based on the characteristics of the test set (Wagner et al., 2020). The research could also be expanded to include the generation of synthetic treebanks to facilitate cross-lingual dependency parsing (Tyers et al., 2018). Improvements in multilingual parsing could provide valuable resources and benefits even beyond the context of Irish, particularly with related Celtic languages.

Downstream applications The TwittIrish treebank is open-source and can be used to develop interactive tools and applications to assist researchers, linguists, data analysts, language teachers, and language learners in understanding the grammatical structure and syntactic relationships within Irish-language social media text.

Language contact experimentation Many possibilities exist for experimentation with language contact data. One avenue that could be explored in the current research is to directly measure the effect of language contact on dependency parsing accuracy. In Chapter 5 we found an association between language contact and lower parsing accuracy. However we did not establish causation or quantify the effect. This could be done by developing a test set of Irish tweets in which each tweet contains some form of language contact. A copy of this dataset in which the language contact phenomena have been translated into Irish could then be created in order to compare dependency parsing results across the two test sets. Alternatively, different methodologies for classifying code-switching and borrowing could be tested against the perceptions of speakers.

7.5 Concluding Remarks

The future of the Irish language is by no means certain. As one of the oldest European languages, Irish has survived many challenges, with the most pressing being the dominance of English as a global language. Although the number of fluent speakers may be

limited, Irish is fortunate to receive government support and general goodwill. However, in the digital age where communication heavily relies on technology, any efforts towards language revitalisation must embrace modern technology in order to succeed. While recent advances in NLP and artificial intelligence have capitalised on the abundant data available on the internet, unfortunately, low-resource languages like Irish face a higher risk of digital extinction due to the scarcity of available data.

Our research endeavoured to address this issue by increasing the number of linguistic resources specifically tailored for Irish. In pursuit of this goal, we focused on creating a comprehensive dataset consisting of linguistically annotated informal Irish text sourced from tweets. This dataset enabled us to closely examine unedited Irish language usage in a contemporary context and it can also facilitate future resource development for processing UGC, which only continues to grow and evolve as a genre.

While analysing data of this variety has allowed us to gain insight into unedited Irish language usage, the specific linguistic features of UGC brought with it a particular set of challenges. One of the most notable features of the data was the eclipsing presence of English in many Irish-language tweets. While addressing the implications of this on parser development, we encountered conflicting ideas in the literature regarding best practices for processing language contact phenomena. This led us to investigate the attitudes of Irish speakers, recognising that they are the potential users of any language technology developed for Irish. We observed in our questionnaire study that some learners of Irish suffer from a lack of confidence with regard to using the language and worry sometimes about the stigma of mixing languages. We also found that a word being in English or Irish was not always the most important factor affecting word choice as it is part of a trade-off with clarity, convenience, and conformity. Considering these findings could be helpful in informing the development of effective language policy and revitalisation strategies that encourage language use, foster confidence, and celebrate linguistic diversity.

While only a small part of solving the larger problem faced by many low-resource languages, the present research represents a step towards preserving Irish. Even in the unfortunate event of Irish language extinction, the documentation of the language would serve as an invaluable historical resource. Additionally, it could potentially contribute to future language revival efforts, aiding in the reconstruction and revitalisation of the language.

In conclusion, despite the uncertain future of the Irish language, with government support, positive attitudes towards the language, and the integration of modern technology, there is hope for its preservation. Our research aims to contribute to this cause, through the advancement of both NLP technology and linguistic understanding.

Bibliography

- Agheyisi, R. and Fishman, J. A. (1970). Language attitude studies: A brief survey of methodological approaches. *Anthropological Linguistics*, 12(5):137–157.
- Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., and Solorio, T. (2018). Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia.
- Albogamy, F. and Ramsay, A. (2017). Universal Dependencies for Arabic tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 46–51, Varna, Bulgaria.
- Álvarez-Mellado, E. and Lignos, C. (2022). Borrowing or codeswitching? Annotating for finer-grained distinctions in language mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201, Marseille, France.
- Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Atkinson, D. and Kelly-Holmes, H. (2011). Codeswitching, identity and ownership in Irish radio comedy. *Journal of Pragmatics*, 43(1):251–260.
- Auer, P. (1984). *Bilingual conversation*. Pragmatics & beyond: An interdisciplinary series of language studies, 5:8. J. Benjamins Pub. Co., Amsterdam.
- Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). “I am borrowing ya mixing ?” an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Work-*

- shop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.
- Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., et al. (2020). Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online.
- Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.
- Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Meachair, M. J. Ó., and Foster, J. (2022). gaBERT—an Irish Language Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France.
- Batchelor, C. (2019). Universal Dependencies for Scottish Gaelic: Syntax. In *Proceedings of the Celtic Language Technology Workshop*, pages 7–15, Dublin, Ireland.
- Bhat, I., Bhat, R. A., Shrivastava, M., and Sharma, D. (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.
- Bhat, I. A., Bhat, R. A., Shrivastava, M., Sharma, D. M., and LTRC, I.-H. (2017). Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. *EACL 2017*, page 324.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O’Reilly Media.
- Bisagni, J. (2014). Prolegomena to the study of code-switching in the old irish glosses. *Peritia*, 24:1–58.
- Blodgett, S. L., Wei, J., and O’Connor, B. (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*

- Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Burr, V. (2015). *Social constructionism / Vivien Burr*. Routledge, Hove, East Sussex, third edition.
- Caomhánach, C. (2022). Códswitcheáil agus códmixeáil: iniúchadh ginearálta ar an gcódmheascadh Gaeilge-Béarla ar na meáin shóisialta. *Léann Teanga: An Reiviú*.
- Carnie, A. H. (1995). *Non-verbal predication and head-movement*. PhD thesis, Massachusetts Institute of Technology.
- Cassidy, L., Lynn, T., Barry, J., and Foster, J. (2022). TwittIrish: A Universal Dependencies treebank of tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884.
- Caulfield, J. (2013). *A social network analysis of Irish language use in social media*. PhD thesis, Cardiff University.
- Çetinoğlu, Ö. (2016). A turkish-german code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4215–4220, Portorož, Slovenia.
- Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (2016). Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11.
- Che, W., Liu, Y., Wang, Y., Zheng, B., and Liu, T. (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 (CoNLL 2018)*, pages 55–64, Brussels, Belgium.
- Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, Doha, Qatar.
- Chen, K., Wang, R., Utiyama, M., Liu, L., Tamura, A., Sumita, E., and Zhao, T. (2017). Neural machine translation with source dependency representation. In *Proceedings of*

the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2846–2852, Copenhagen, Denmark.

Choi, J. D., Tetreault, J., and Stent, A. (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.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakesh, Morocco.

Chu, Y.-J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Cignarella, A. T., Bosco, C., and Rosso, P. (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, Paris, France.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, volume 1. Athens, GA.

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

CSO (2022). Census of population 2022 - summary results: Education and Irish Language. CSO statistical publication.

de Bhaldraithe, T. (1959). *English-Irish Dictionary*. Oifig an tSoláthair, Baile átha Cliath.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, volume 14, pages 4585–4592, Reykjavik, Iceland.

- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Deuchar, M. (2020). Code-switching in linguistics: A position paper. *Languages*, 5(2):22.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.
- Doherty, C. (1996). Clausal structure and the modern Irish copula. *Natural Language & Linguistic Theory*, 14(1):1–46.
- Dörnyei, Z. and Dewaele, J.-M. (2022). *Questionnaires in second language research: Construction, administration, and processing*. Taylor & Francis.
- Dowling, M., Castilho, S., Moorkens, J., Lynn, T., and Way, A. (2020). A human evaluation of english-irish statistical and neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 431–440, Lisboa, Portugal.
- Doyle, A. (2015). *A History of the Irish Language*. Oxford University Press.
- Doyle, A., McCrae, J. P., and Downey, C. (2019). A character-level LSTM network model for tokenizing the Old Irish text of the Würzburg glosses on the Pauline epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.
- Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France*.
- Duval, E. and Ochoa, X. (2008). Quantitative analysis of user-generated content on the web. *Web Science*, 4:22.
- Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. 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 334–343, Beijing, China. Association for Computational Linguistics.
- Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.
- Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.
- Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: an exploration. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 340–345, Copenhagen, Denmark.
- Ferrara, K., Brunner, H., and Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written communication*, 8(1):8–34.
- Fhlannchadha, S. N. and Hickey, T. M. (2018). Minority language ownership and authority: Perspectives of native speakers and new speakers. *International Journal of Bilingual Education and Bilingualism*, 21(1):38–53.
- Foster, J. (2010). “cba to check the spelling”: Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384.
- Foster, J., Çetinoğlu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and Van Genabith, J. (2011). # hardtoparse: POS tagging and parsing the twitterverse. In *Proceedings of the 5th AACL Conference on Analyzing Microtext*, pages 20–25, San Francisco, California.
- Gal, S. (1988). The political economy of code choice. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:245–64.
- Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (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.
- Gawne, L. and McCulloch, G. (2019). Emoji as digital gestures. *Language@Internet*, 17(2).
- Giles, H., Taylor, D. M., and Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some Canadian data. *Language in society*, 2(2):177–192.
- Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (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.
- Ginter, F., Hajic, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). Conll 2017 shared task-automatically annotated raw texts and word embeddings. *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University*.
- Gollan, T. H. and Ferreira, V. S. (2009). Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):640.
- Government of Ireland (2010). *20-Year Strategy for the Irish Language 2010–2030*. Stationery Office, Dublin.
- Government of Ireland (2021). Preliminary report to inform the development of a policy for the Irish-medium sector outside of the gaeltacht.
- Guzmán, G. A., Ricard, J., Serigos, J., Bullock, B. E., and Toribio, A. J. (2017). Metrics for modeling code-switching across corpora. In *INTERSPEECH*, pages 67–71.
- Hajič, J. (1998). Building a syntactically annotated corpus: The Prague dependency treebank. In *Issues in Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Charles University Press, Praha, Karolinum.

- Hajič, J., Smrz, O., Zemánek, P., Šnaidauf, J., and Beška, E. (2004). Prague Arabic dependency treebank: Development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, volume 1, Cairo, Egypt.
- Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2):210–231.
- Heath, M. (2021). NO NEED TO YELL: A Prosodic Analysis of Writing in All Caps. *University of Pennsylvania Working Papers in Linguistics*, 27(1).
- Heinecke, J. and Tyers, F. (2019). Development of a Universal Dependencies treebank for Welsh. In *Proceedings of the Celtic Language Technology Workshop*, pages 21–31, Dublin, Ireland.
- Herring, S. C., editor (1996). *Computer-Mediated Communication*, volume 39 of *Pragmatics and Beyond New Ser.* John Benjamins Publishing Company, first edition.
- Hickey, R. (2007). *Irish English: History and present-day forms*. Cambridge University Press.
- Hickey, R. (2011). *The Dialects of Irish: Study of a Changing Landscape*. Walter de Gruyter GmbH, Berlin/Boston, Germany.
- Hickey, R. (2014). *The sound structure of Modern Irish*. De Gruyter Mouton.
- Hickey, T. (2009). Code-switching and Borrowing in Irish. *Journal of Sociolinguistics*, 13(5):670–688.
- Hickey, T. and Stenson, N. (2011). Irish orthography: What do teachers and learners need to know about it, and why? *Language, Culture and Curriculum*, 24(1):23–46.
- Huang, Z. and Harper, M. (2009). Self-training PCFG grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 832–841, Singapore. Association for Computational Linguistics.
- Hudson, R. (2010). *An Introduction to Word Grammar*. Cambridge Textbooks in Linguistics. Cambridge University Press.

- Jamatia, A., Gambäck, B., and Das, A. (2015). Part-of-speech tagging for code-mixed English-Hindi Twitter and facebook chat messages. In *Proceedings of the international conference recent advances in natural language processing*, pages 239–248.
- Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.
- Judge, J., Cahill, A., and Van Genabith, J. (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, Sydney, Australia.
- Jurafsky, D. and Martin, J. H. (2023). Speech and language processing. Draft of January 7, 2023. Retrieved from <https://web.stanford.edu/~jurafsky/slp3/18.pdf>.
- Kelly-Holmes, H. (2006). Irish on the world wide web: Searches and sites. *Journal of Language and Politics*, 5(2):217–238.
- Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E. (2006). Efficient corpus development for lexicography: building the new corpus for ireland. *Language resources and evaluation*, 40:127–152.
- Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.
- Kondratyuk, D. and Straka, M. (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.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (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.

- Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.
- Kübler, S., McDonald, R., and Nivre, J. (2009). Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.
- Kulmizev, A., de Lhoneux, M., Gontrum, J., Fano, E., and Nivre, J. (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.
- Labov, W. (1971). The notion of ‘system’ in Creole studies. *Pidginization and creolization of languages*, 447:472.
- Lackaff, D. and Moner, W. J. (2016). Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th ACM International Conference on the Design of Communication*, pages 1–9, Silver Spring, Maryland.
- Li, Y. and Fung, P. (2012). Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*, pages 1671–1680, Mumbai, India.
- Lipski, J. M. (2005). Code-switching or borrowing? “no sé so no puedo decir,” you know. In *Selected proceedings of the Second Workshop on Spanish Sociolinguistics*, pages 1–15, Somerville, Massachusetts. Cascadilla Proceedings Project.
- Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (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, New Orleans, Louisiana.
- Lomas, T. (2017). The spectrum of positive affect: A cross-cultural lexical analysis. *International Journal of Wellbeing*, 7(3):1–18.
- Lui, M. and Baldwin, T. (2012). langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

- Luotolahti, J., Kanerva, J., Laippala, V., Pyysalo, S., and Ginter, F. (2015). Towards Universal Web Parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220. Uppsala University, Uppsala, Sweden.
- Lynn, T. (2016). *Irish dependency treebanking and parsing*. PhD thesis, Dublin City University and Macquarie University, Sydney.
- Lynn, T. (2023). *Language Report Irish*, pages 163–166. Springer International Publishing, Cham.
- Lynn, T., Çetinoğlu, Ö., Foster, J., Dhonnchadha, E. U., Dras, M., and van Genabith, J. (2012). Irish treebanking and parsing: A preliminary evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1939–1946, İstanbul, Turkey.
- Lynn, T. and Foster, J. (2016). Universal Dependencies for Irish. In *Proceedings of the 2nd Celtic Language Technology Workshop*, Paris, France.
- Lynn, T., Foster, J., and Dras, M. (2013). Working with a small dataset - semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA. Association for Computational Linguistics.
- Lynn, T. and Scannell, K. (2019). Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the 3rd Celtic Language Technology Workshop*, Dublin, Ireland.
- Lynn, T., Scannell, K., and Maguire, E. (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.
- Mac Giolla Chríost, D. (2004). *The Irish language in Ireland: from Goídel to globalisation*, volume 3. Routledge.
- Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado.

- Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.
- McArthur, T. (1998). Code-mixing and code-switching. <https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/code-mixing-and-code-switching>. Accessed on August 24, 2023.
- McCloskey, J. (2017). Object positions (in Irish). *A Schrift to Fest Kyle Johnson*, page 255.
- McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.
- McDonald, R. and Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.
- McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria.
- McDonald, R., Pereira, F., Ribarov, K., and Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 523–530.
- McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 62–72.
- McGuinness, S., Phelan, J., Walsh, A., and Lynn, T. (2020). Annotating MWEs in the Irish UD treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 126–139, Barcelona, Spain (Online).
- Mel’čuk, I. A. et al. (1988). *Dependency syntax: Theory and practice*. SUNY press.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Mikulová, M. and Štěpánek, J. (2009). Annotation quality checking and its implications for design of treebank (in building the prague czech-english dependency treebank). In *Eighth International Workshop on Treebanks and Linguistic Theories*, page 137, Milan, Italy.
- Moal, S., Ó Murchadha, N. P., and Walsh, J. (2018). New speakers and language in the media: Audience design in Breton and Irish broadcast media. In *New speakers of minority languages: Linguistic ideologies and practices*, pages 189–212. Palgrave Macmillan, London.
- Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas.
- Muysken, P. (1997). Code-switching processes: Alternation, insertion, congruent lexicalization. *Language choices: Conditions, constraints, and consequences*, 20:361–380.
- Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
- Myers-Scotton, C. (1989). Codeswitching with English: types of switching, types of communities. *World Englishes*, 8(3):333–346.
- Myers-Scotton, C. (1992). Comparing codeswitching and borrowing. *Journal of Multilingual & Multicultural Development*, 13(1-2):19–39.
- Myers-Scotton, C. (1995). *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.
- Naab, T. K. and Sehl, A. (2017). Studies of user-generated content: A systematic review. *Journalism*, 18(10):1256–1273.

- Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A., and Gobl, C. (2017). The ABAIR initiative: Bringing spoken Irish into the digital space. In *Proceedings of Interspeech*, Stockholm, Sweden.
- Ní Laoire, S. (2016). *Irish-English Code-switching: a Sociolinguistic Perspective*, pages 81–106. Palgrave Macmillan UK, London.
- Nic Giolla Mhichíl, M., Lynn, T., and Rosati, P. (2018). Twitter and the Irish language, #Gaeilge – agents and activities: exploring a data set with micro-implementers in social media. *Journal of Multilingual and Multicultural Development*, 39(10):868–881.
- Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the eighth international conference on parsing technologies*, pages 149–160, Nancy, France.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, Portorož, Slovenia.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (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.
- Nivre, J., Hall, J., and Nilsson, J. (2004). Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 49–56.
- Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain. Association for Computational Linguistics. Tutorial presentation.

- Ní Chasaide, A., Ní Chiarán, N., Uí Dhonnchadha, E., Lynn, T., and Judge, J. (2022). Digital plan for the Irish language.
- Oireachtas (1947). *Litriú na Gaeilge: Lámhleabhar an chaighdeáin oifigiúil*. Oifig an tSoláthair, Baile Átha Cliath.
- Oireachtas (2017). *Gramadach Na Gaeilge: An Caighdeán Oifigiúil: An Treoir le haghaidh Scríbhneoireacht sa Ghaeilge*. Arna Fhoilsiú ag Seirbhís Thithe an Oireachtais.
- O'Malley-Madec, M. (2007). How one word borrows another: The process of language-contact in two Irish-speaking communities. *International Journal of Bilingual Education and Bilingualism*, 10(4):494–509.
- Osborne, T. and Gerdes, K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A journal of general linguistics*, 4:17.
- Patwa, P., Aguilar, G., Kar, S., Pandey, S., Pykl, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 774–790, Barcelona, Spain.
- Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Petrov, S., Chang, P.-C., Ringgaard, M., and Alshawi, H. (2010). Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA. Association for Computational Linguistics.

- Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, İstanbul, Turkey. European Language Resources Association (ELRA).
- Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55(2):291–318.
- Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. *Bochumer Linguistische Arbeitsberichte*, page 13.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: Toward a typology of code-switching. *Linguistics*, 18.
- Poplack, S. and Meechan, M. (1998). How languages fit together in codemixing. *International journal of bilingualism*, 2(2):127–138.
- Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26:47–104.
- Poplack, S., Wheeler, S., and Westwood, A. (1989). Distinguishing language contact phenomena: evidence from finnish-english bilingualism. *World Englishes*, 8(3):389–406.
- Rannóg an Aistriúcháin (1958). *Gramadach na Gaeilge agus litriú na Gaeilge: An caighdeán oifigiúil*. Oifig an tSoláthair, Baile Átha Cliath.
- Real Academia Española (2021). Diccionario de la lengua española. <https://dle.rae.es/>.
- Rehbein, I., Ruppenhofer, J., and Do, B.-N. (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.
- Sagae, K. (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44, Uppsala, Sweden. Association for Computational Linguistics.

- Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L., and Solorio, T. (2016). Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the second workshop on computational approaches to code switching*, pages 50–59, Austin, Texas.
- Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2022). Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, pages 1–52.
- Sankoff, G. (2004). Linguistic outcomes of language contact. *The handbook of language variation and change*, pages 638–668.
- Scannell, K. (2011). Indigenous Tweets: Welcome/Fáilte! <http://indigenoustweets.blogspot.com/2011/03/welcomefailte.html>.
- Scannell, K. (2020). Universal Dependencies for Manx Gaelic. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 152–157.
- Scannell, K. (2022). Diachronic parsing of pre-standard Irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC 2022*, pages 7–13, Marseille, France.
- Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Ortiz Suárez, P. J., Sagot, B., and Srivastava, A. (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.
- Seddah, D., Sagot, B., Candito, M., Moulleron, V., and Combet, V. (2012). The French social media bank: A treebank of noisy user generated content. In *COLING 2012-24th International Conference on Computational Linguistics*, Mumbai, India.
- Seraji, M., Megyesi, B., and Nivre, J. (2012). Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7(18).
- Sgall, P., Hajicová, E., and Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

- Silveira, N., Dozat, T., de Marneffe, M. C., Bowman, S. R., Connor, M., Bauer, J., and Manning, C. D. (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, Reykjavik, Iceland. ELRA.
- Søgaard, A. and Rishøj, C. (2010). Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1065–1073, Beijing, China. COLING 2010 Organizing Committee.
- Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar.
- Solorio, T. and Liu, Y. (2008). Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.
- Stam, N. (2017). *A typology of code-switching in the Commentary to the Féilire Óengusso*. Netherlands Graduate School of Linguistics.
- Stenson, N. (1981). *Studies in Irish Syntax*. Ars linguistica. Tübingen: Narr.
- Stenson, N. (1991). Code-switching vs. borrowing in modern Irish. In *Language contact in the British Isles*, pages 559–580. Max Niemeyer Verlag.
- Stenson, N. (1993). *Variation in phonological assimilation of Irish loanwords*, volume 98. John Benjamins Publishing.
- Stenson, N. (2019). *Modern Irish: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Taylor & Francis.
- Straka, M., Hajic, J., and Straková, J. (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, Portorož, Slovenia.

- Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. (2018). Linguistically-informed self-attention for semantic role labeling. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.
- Stymne, S., de Lhoneux, M., Smith, A., and Nivre, J. (2018). Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.
- Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.
- Tatman, R. (2015). #go awn: Sociophonetic Variation in Variant Spellings on Twitter. *Working Papers of the Linguistics Circle*, 25(2):97–108. Number: 2.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Editions Klincksieck, Paris.
- Tomlin, R. S. (2014). *Basic Word Order (RLE Linguistics B: Grammar): Functional Principles*. Routledge.
- Tucker, G. R. (2001). A global perspective on bilingualism and bilingual education. *Georgetown University Round Table on Languages and Linguistics 1999*, page 332.
- Tyers, F., Sheyanova, M., Martynova, A., Stepachev, P., and Vinogorodskiy, K. (2018). Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150.
- Tyers, F. M. and Ravishankar, V. (2018). A prototype dependency treebank for Breton. In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, pages 197–204.
- Uí Dhonnchadha, E. (2002). An analyser and generator for Irish inflectional morphology using finite-state transducers. Master’s thesis, Dublin City University.

- Uí Dhonnchadha, E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. PhD thesis, Dublin City University.
- Van der Beek, L., Bouma, G., Malouf, R., and Van Noord, G. (2002). The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pages 8–22. Brill.
- Van Der Goot, R. and van Noord, G. (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, Brussels, Belgium.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Vogel, S. and García, O. (2017). Translanguaging. In Noblit, G. and Moll, L., editors, *Oxford Research Encyclopedia of Education*. Oxford University Press.
- Wagner, J., Barry, J., and Foster, J. (2020). Treebank embedding vectors for out-of-domain dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8812–8818, Online. Association for Computational Linguistics.
- Wagner, J. and Foster, J. (2021). Revisiting tri-training of dependency parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9457–9473, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Walsh, A. (2023). *The Automatic Processing of Multiword Expressions in Irish*. PhD thesis, Dublin City University.
- Wang, H., Zhang, Y., Leonard Chan, G. Y., Yang, J., and Chieu, H. L. (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.
- Weinreich, U. (1953). *Languages in contact: Findings and problems*. Mouton, The Hague.

- Whitney, W. D. (1881). On mixture in language. *Transactions of the American Philological Association (1869-1896)*, 12:5–26.
- Winata, G., Aji, A. F., Yong, Z. X., and Solorio, T. (2023). The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.
- Winata, G. I., Lin, Z., and Fung, P. (2019). Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186, Florence, Italy. Association for Computational Linguistics.
- Yamada, H. and Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 195–206, Nancy, France.
- Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.
- Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*, volume 2008, pages 28–30, Marrakesh, Morocco.
- Zeman, D., Marecek, D., Popel, M., Ramasamy, L., Stepánek, J., Zabokrtský, Z., and Hajic, J. (2012). Hamlet: To parse or not to parse? In *Proceedings of LREC*, pages 2735–2741, İstanbul, Turkey.
- Ó Cróinín, D. and Uí Dhonnchadha, E. (1998). Le-parole and corpus náisiúnta na gaeilge. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Granada, Spain.
- Ó Mianáin, P. (2020). *Concise English-Irish Dictionary*. An Gúm.
- Ó Murchadha, N. and Kavanagh, L. (2022). Language ideologies in a minority context: An experimental study of teachers’ responses to variation in Irish. *Journal of Sociolinguistics*, 26(2):197–220.
- Ó Siadhail, M. (1989). *Modern Irish: grammatical structure and dialectal variation*. Cambridge University Press, Cambridge.

Appendix A

TwittIrish Data Statement

A.1 Header

Dataset Title: TwittIrish

Dataset Curator(s): Lauren Cassidy, Teresa Lynn, Jennifer Foster, Sarah McGuinness

Curator(s) Affiliation(s): ADAPT Centre, Dublin City University

Dataset Version: UD version 2.12

Dataset Citation: (Cassidy et al., 2022)

Data Statement Author(s): Lauren Cassidy

Data Statement Author(s) Affiliation: ADAPT Centre, Dublin City University

Data Statement Version: 1.0

A.2 Executive Summary

The TwittIrish Treebank (Cassidy et al., 2022) was created to address the lack of resources for Irish UGC by providing a resource for training parsers, facilitating experimentation with NLP tasks on informal Irish and enabling linguistic analysis of Irish UGC. The dataset consists of Irish-language tweets and contains some language contact with English. The dataset consists of 2,596 tweets (47,790 tokens).

A.3 Curation Rationale

The TwittIrish Treebank was curated to address the lack of an accurate parser for Irish UGC, particularly in the context of social media platforms. The dataset, which comprises

2,596 Irish-language tweets annotated with linguistic information, was created to serve as a valuable resource for both NLP and linguistic research. The dataset has several potential uses. Firstly, it can be used as training data for parsers, to improve parsing accuracy for informal and non-standard language such as that found in UGC. The dataset also enables experimentation with various NLP tasks, beyond just parsing. The diverse linguistic features present in TwittIrish offer a testbed for evaluating the performance of NLP models on such data. Beyond NLP applications, TwittIrish can also facilitate in-depth linguistic analysis of Irish UGC.

A.4 Documentation for Source Data Sets

The TwittIrish treebank’s tweets were gathered through the Indigenous Tweets (IT) project (Scannell, 2011), which compiles social media data from 185 minority and indigenous languages, including Irish. The treebank’s source data encompassed two Irish language tweet corpora, namely the LTC (Lynn et al., 2015; Lynn and Scannell, 2019) and the NTC. 1,299 tweets were included from the LTC. The LTC tweets were from 2009 to 2014 and had previously undergone linguistic processing (tokenisation, lemmatisation, POS tagging and code-switching annotation). 1,297 more recent tweets were added from the NTC. The NTC tweets were from 2010 to 2019 and in plain text format. Duplicates and non-Irish tweets were excluded.

A.5 Language Varieties

The language variety of the TwittIrish treebank is that of informal Irish. All three main dialects (Connacht, Munster, and Ulster) are represented and language contact with English is frequent in the data.

A.6 Speaker Demographic

The speaker demographic of the TwittIrish dataset is Twitter users with various levels of proficiency in Irish. We estimate a wide range of ages with older speakers being underrepresented based on the demographic of social media users more generally. We do not have information about the gender or socioeconomic status of the speakers. We also do not have

information on the race/ethnicity of speakers but we estimate a large majority to come from Ireland based on the demographic of Irish speakers more generally. We estimate that the first language of most speakers is English. We estimate that approximately 1,260 speakers are represented in the dataset with most speakers contributing a single tweet. We base this estimate on a sample of 2,596 tweets from the NTC data for which we have user IDs.

A.7 Annotator Demographic

At the time of annotation, the annotator was aged 26 to 29 years. Their gender is female, and their ethnic background is Irish. Their first language is English. The time of annotation spans from 2020 to 2023. The annotator is fluent in Irish, the language of the data being annotated. The annotation was performed by a single annotator. The annotator’s training consisted of extensive engagement with expert annotators throughout the course of their doctoral studies.

A.8 Speech Situation and Text Characteristics

The tweets of the TwittIrish dataset, collected between the years 2009 and 2019, offer a snapshot of linguistic activity during this period. The modality of the tweets is typed, representing spontaneous rather than elicited expressions within an asynchronous conversation, taking place on the Twitter platform. The intended audience encompasses Twitter users and a broader online community, reflecting the public nature of this digital discourse. The genre of these texts, classified as social media, encompasses a diverse range of topics, each contributing to the varied vocabulary and structural characteristics of the tweets.

A.9 Preprocessing and Data Formatting

The development of the TwittIrish Treebank involved data preprocessing and conversion procedures. The LTC, having been tokenised, lemmatised, and assigned POS tags, required conversion to adhere to UD and Irish-specific conventions outlined by Lynn (2016). This process involved automated and manual adjustments, particularly regarding MWEs. The NTC, being in plain text format, necessitated preprocessing via tokenisation, lem-

matisation, and POS tagging. These tasks were completed using NLTK TweetTokenizer (Bird et al., 2009) and Morfette (Chrupała et al., 2008) trained on the converted LTC data. All the tagged tweets were then converted to the CoNLL-U format, and annotated with syntactic information in a bootstrapping cycle of automatic dependency parsing using the biaffine parser (Dozat and Manning, 2017) and manual corrections. Language IDs were assigned to all English and Irish tokens. Finally, all usernames and contact information were anonymised.

A.10 Capture Quality and Limitations

With regard to the quality of the dataset, we acknowledge varying proficiency in the Irish language among speakers. We also note the possibility of a small number of tweets being generated by bots or machine-translated. In terms of the quality of the annotation, a limited review of 46 trees (773 tokens) was conducted to assess the quality. This review found just 32 errors. We acknowledge the likelihood of the dataset containing some bugs and biases due to only having a single annotator, however, overall, the dataset passes the validation script mandated by UD and has been shown to improve parsing accuracy in the genre of Irish tweets in our experiments.

A.11 Metadata

License: CC BY-SA 4.0

Annotation Guidelines: See Section B and the Irish-specific UD guidelines¹

Dataset: The UD GitHub repository for TwittIrish²

¹<https://universaldependencies.org/ga/>

²https://github.com/UniversalDependencies/UD_Irish-TwittIrish

Appendix B

TwittIrish Annotation Guidelines

B.1 Segmentation

Each tweet is considered a unit of analysis with a unique identifier. When more than one sentence occurs within a tweet, they are attached via the relation `parataxis:sentence` (See Section B.5)

B.2 Tokenisation

In the CoNLL-U format, each token will have an ID number within the tree. The unit of annotation is a syntactic word. Word boundaries in Irish are generally denoted by whitespace or punctuation however some exceptions exist for which we list the following guidelines.

Contractions Contractions, which involve combining two words into one by omitting one or more letters and replacing them with an apostrophe, are considered separate tokens.

We use the ‘+’ symbol here to denote the separation of tokens.

(B.1) *b'fhéidir* → *b' + fhéidir*

(B.2) *n'fheadar* → *n' + fheadar*

MWEs Although sometimes acting as a single syntactic unit, where the constituents of MWEs are separated by whitespace, they should be considered separate tokens.

(B.3) *Baile Átha Cliath* → *Baile + Átha+ Cliath*

(B.4) *Raidió na Gaeltachta* → *Raidió + na + Gaeltachta*

Incorrectly fused words Tokens that are usually separated by whitespace but appear fused, are tokenised as separate words.

(B.5) *arais* → *ar* + *ais*

(B.6) *ArdMhacha* → *Ard* + *Mhacha*

Incorrectly separated words Morphemes that are usually considered single tokens are not fused during tokenisation. Rather, they are later attached by the dependency relation *goeswith* (See Section B.5)

(B.7) *ró chinnte* → *ró* + *chinnte*

(B.8) *an deacair* → *an* + *deacair*

English Clitics Clitics, morphemes that are phonologically dependent on a nearby word, but syntactically independent, are tokenised as separate words.

(B.9) *Madigan's* → *Madigan* + 's

(B.10) *don't* → *do* + *n't*

Standard punctuation and symbols Punctuation (excluding hyphenation) and symbols, such as mathematical operators, are considered separate tokens even when attached to another token.

(B.11) *2-13* → *2* + *-* + *13*

(B.12) *mhaith/Good* → *mhaith* + */* + *Good*

Time indicators and measurements

(B.13) *8pm* → *8* + *pm*

(B.14) *5i.n.* → *5* + *i.n.*

(B.15) *8KM* → *8* + *KM*

Prepositional pronouns Prepositions with inflected affixes corresponding to pronouns are not separated into their constituents, e.g. *agam*, *ort*.

Abbreviations and initialisms Abbreviated representations of phrases are not separated into their constituents, e.g. *srl*, *wtf*, *dr.*

Hyphenation Tokens that contain a hyphen for grammatical reasons are considered single tokens, e.g. *n-athair*, *t-aonad*. Tokens that include a prefix attached with hyphenations are also considered single tokens, e.g. *fo-alt*, *an-gheit*. Compound words connected via a hyphen are also considered single words, e.g. *Cipirigh-Gréigeacha*, *ceard-cumannachas*.

Phone numbers, times and dates Phone numbers, timestamps, and date strings are tokenised as single words.

URLs, Hashtags, Usernames, pictograms, email addresses URLs, Hashtags, Usernames, pictograms, email addresses are all tokenised as single units. e.g. *:)*, *#sonas*.

Nonstandard or repeated punctuation Emphatic or stylistic punctuation attached to a word is considered a single token e.g. ***folúntas***, *!!!!!!!*. Apostrophes used instead of diacritics are considered part of the token e.g. *la'*.

B.3 Lemmatisation

As per the guidelines of UD, the lemma should be the canonical or base form of the word that is commonly present in dictionaries. It should not have any inflectional suffixes and should have only one form for each POS paradigm. The lemma should be in the positive form. It should not eliminate the derivational morphology, meaning that the lemma for *eagraíochtaí* ‘organisations’ should be *eagraíocht* ‘organisation’ instead of *eagraigh* ‘organise’.

Inflection Any inflection should be removed.

(B.16) *téann* → *téigh*

(B.17) *bhoird* → *bord*

(B.18) *seachtaine* → *seachtain*

Orthographic variation Orthographic variation like nonstandard use of diacritics, and typos, should be removed in the lemma.¹ The lemma should be denoted in lowercase except where a token is case sensitive e.g. proper nouns, usernames, or pictograms, where the case should be preserved in the lemma column.

¹<https://universaldependencies.org/u/overview/morphology.html>

(B.19) *reasunta* → *réasúnta*

(B.20) *bheul* → *bhuel*

Capitalisation

(B.21) *BREÁ* → *breá*

(B.22) *Leabhar* → *leabhar*

(B.23) *:D* → *:D*

(B.24) *Gaeilge* → *Gaeilge*

B.4 POS-tagging

Table B.1 provides descriptions of each UPOS tag as well as Irish language examples.

UPOS tag	Description	Examples
ADJ	An adjective can modify nouns or act as a predicate. Ordinal numbers and verbal adjectives are considered adjectives.	<i>mór, maith, céad, 3ú, céanna, déanta</i>
ADP	An adposition forms a structure with a complement noun phrase. In compound adpositions, the constituent words are tagged according to their basic use e.g. in the phrase <i>i gcoinne</i> , <i>i</i> is tagged as ADP, while <i>gcoinne</i> is tagged as NOUN.	<i>i, ar, le, go dtí, tar éis</i>
ADV	An adverb typically modifies a verb, adjective or adverb and may form part of a phrasal verb, e.g. <i>cur isteach</i> .	<i>riamh, anois, cá, arís</i>
AUX	An auxiliary expresses grammatical distinctions not carried by the lexical verb of a phrase. AUX is used only for the copula in Irish as opposed to the substantive verb ‘to be’	<i>is, ba, ní, nach</i>
CCONJ	A coordinating conjunction joins constituent words or phrases in a syntactic relationship in which no constituent is subordinate to another.	<i>agus, nó, ach, é</i>
DET	A determiner modifies and expresses the reference of a noun phrase.	<i>an, seo, aon, mo</i>
INTJ	An interjection is usually all or part of an exclamation or feedback particle.	<i>bhuel, ambaiste, haló, psst</i>

UPOS tag	Description	Examples
NOUN	A noun refers to a person, place, thing, idea, or concept. Pronominal quantifiers, verbal nouns, abstract nouns, and abbreviated nouns are tagged as NOUN .	<i>duine, cúpla, cur, leor, réir, Dr.</i>
NUM	A numeral expresses a relation to a number such as quantity, fraction or order, e.g. cardinal numbers in the form of digits and words, Roman numerals, list items, and numerals that form part of named entities.	<i>dó, 11,000</i>
PROPN	A proper noun is a subset of nouns which is the name or part of the name of a specific individual, place, or object.	<i>Gaeilge, Éire, RTÉ</i>
PART	A particle is a function word that is dependent on another word or phrase. Particles in Irish do not inflect but may add grammatical information, such as negation, mood, or tense, to the clause.	<i>a, go, níos</i>
PRON	A pronoun functions as a substitute for a noun or noun phrase e.g. personal pronouns, interrogative pronouns, relative pronouns, indefinite pronouns, demonstrative pronouns.	<i>sí, iad, cad</i>
PUNCT	Punctuation is a non-alphabetical character or characters marking sentence or clause boundaries.	<i>..., !, ;</i>
SCONJ	A subordinate conjunction joins clauses by making one a constituent of another.	<i>nuair, má, mura</i>
SYM	A symbol is a word-like element of alphanumeric and/or special characters. It can usually be substituted by a normal word or words e.g. mathematical operators, currency symbols, emojis, URLs, and email addresses.	<i>%, €, :)</i>
VERB	A verb typically signals an event or action in a clause. A verb usually governs the number and types of other constituents which occur in the clause. Irish verbs often inflect to indicate grammatical categories such as aspect, mood, tense and voice.	<i>bí, déan, téigh</i>
X	X is used restrictively when no real part-of-speech category can be assigned.	<i>wkdwdj</i>

Table B.1: UPOS tags with descriptions and Irish language examples.

B.5 Dependency Relations

root The dependency relation **root** is used to indicate the head which governs the structure of the sentence. In the simple verbal constructions, this is the main verb as exemplified in Figure B.1.

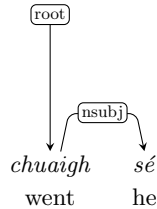


Figure B.1: Verbal root with the nominal subject ‘he went’.

nsubj The nominal subject is a noun or pronoun that acts as the syntactic subject or proto-agent of a verb. In Irish the nominal subject usually directly follows the verb it depends on, as illustrated in Figure B.1.

cop In simple copular constructions in Irish, the copula is followed by the predicate and then the subject, as described in Section 2.1.3. As there is no true verb in such constructions, the predicate is annotated as the root as illustrated in Figure B.2.

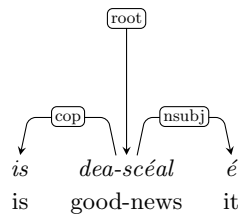


Figure B.2: Copular construction ‘it is good news’.

obj In Irish, the object of a finite verb usually follows the subject as illustrated in Figure B.3

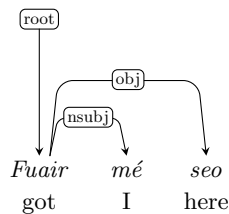


Figure B.3: Direct object of a finite verb ‘I got this’.

mark The **mark** label is used for infinitive markers and subordinate conjunctions. In infinitive constructions, the object and an infinitive marker precede the verb as illustrated in Figure B.4.

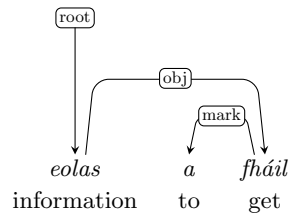


Figure B.4: Infinitive verb construction ‘to get information’.

mark:prt For particles, such as the question particle exemplified in Figure B.5, the subtype **mark:prt** is used.

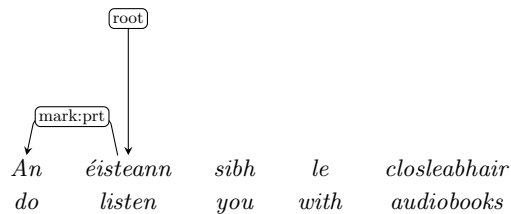


Figure B.5: Question particle ‘Do you listen to audiobooks?’

csubj A clausal subject functions much the same way as a nominal subject except that it consists a clause rather than a single word. Clausal subjects in Irish appear in two subtypes.

csubj:cleft For cleft constructions in which an element is fronted to the predicate position, the label **csubj:cleft** is used. In Figure B.6, the adverb *ansin* ‘there’ is fronted.

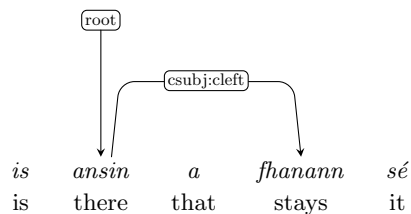


Figure B.6: Clausal subject of a cleft construction ‘it’s there that it stays’.

csubj:cop the label **csubj:cop** is used for a clause that acts as the subject of a copular clause. In Figure B.7, the clause *go bhfuil sé* ‘that it is’ acts as the subject of the outer copular construction.

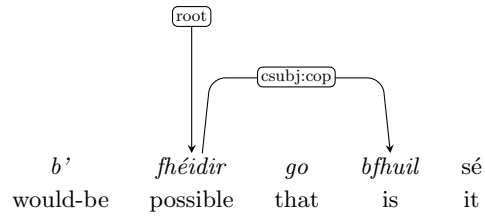


Figure B.7: Clausal subject of a copular clause ‘maybe it is’.

ccomp A clausal complement is a subordinate clause with its own subject as demonstrated in Figure B.8.

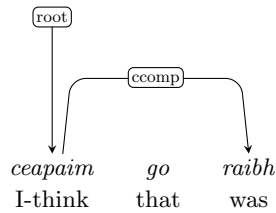


Figure B.8: Clausal complement of a verb ‘I think [it] was’.

xcomp An open clausal complement exemplified in Figure B.9 is a predicative or clausal complement that lacks an independent subject. The subject’s identity is shared with the main clause.

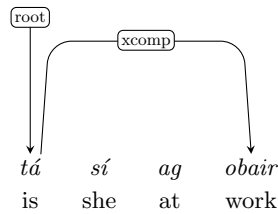


Figure B.9: Open clausal complement of a verb ‘She is working’.

xcomp:pred The subtype **xcomp:pred** is used to indicate complements the substantive verb *bí* ‘to be’ as illustrated in Figure B.10

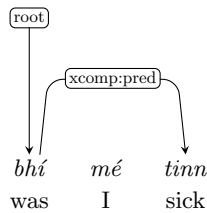


Figure B.10: Predicate of the substantive verb ‘I was sick’.

obl The **obl** label is used for a nominal functioning as an oblique argument.

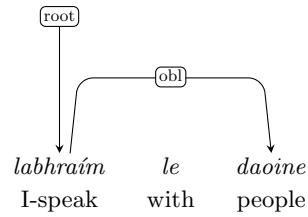


Figure B.11: Oblique argument ‘I speak with people’.

obl:prep In instances where the noun is altered due to being part of a pronominal preposition, the subtype **obl:prep** is used as illustrated in Figure B.12.

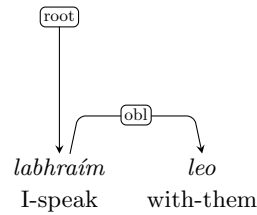


Figure B.12: Personal preposition as oblique argument ‘I speak with them’.

vocative The **vocative** label is used when a participant in a conversation is directly addressed as exemplified in Figure B.13.

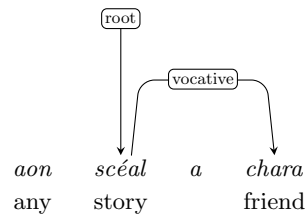


Figure B.13: Dialogue participant directly addressed ‘any news, friend?’.

vocative:mention When a user is addressed directly the label **vocative:mention** is used, as demonstrated in Figure B.14. Usernames can alternatively play other syntactic roles.

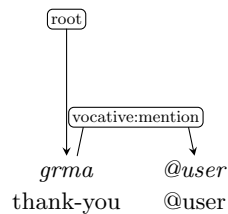


Figure B.14: Vocative mention ‘Thank you @user’.

advmod An adverbial modifier refers to a non-clausal adverb as demonstrated in Figure B.15

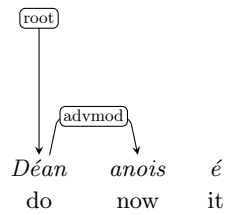


Figure B.15: Adverbial modifier ‘Do it now’.

advcl An adverbial clause modifier, exemplified in Figure B.16, is a clause that functions like an adverb. The head of the adverbial clause is a dependent on the main predicate of the clause it modifies. It serves as an adjunct, meaning it can be removed from the sentence without causing grammatical issues.

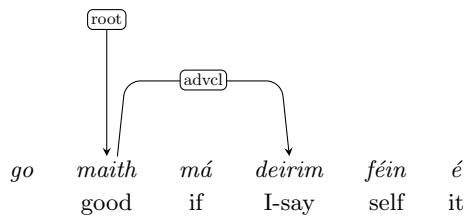


Figure B.16: Adverbial clause modifier ‘good if I say so myself’.

discourse The **discourse** label is used to link interjections and discourse particles to the syntactic structure as illustrated in Figure B.17.

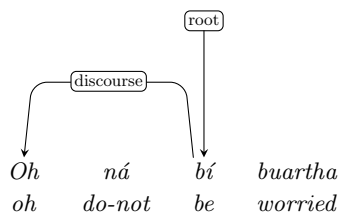


Figure B.17: Discourse marker ‘Oh don’t worry’.

discourse:emo The label **discourse:emo** is used to connect pictograms to the syntactic structure as demonstrated in Figure B.18. Pictograms can also play other syntactic roles.

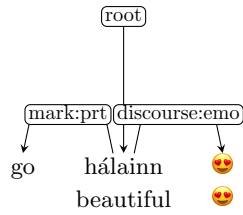


Figure B.18: Pictogram ‘beautiful 😊’.

nmod The **nmod** label is used to indicate nominal modifiers of nouns or clausal predicates as exemplified in Figure B.19.

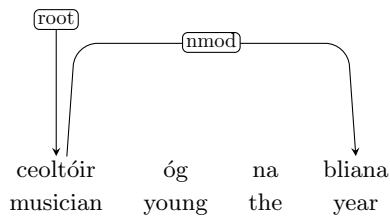


Figure B.19: Nominal modifier ‘young musician of the year’.

appos An appositional modifier, exemplified in Figure B.20, is noun phrase directly following another noun phrase which it modifies. Its purpose is to provide a definition or name for the first noun phrase.

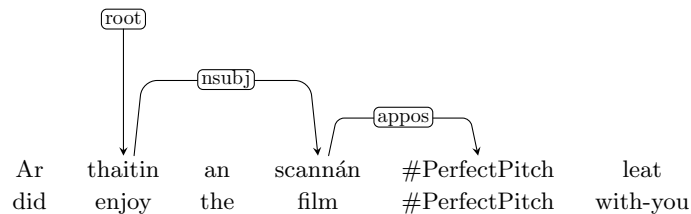


Figure B.20: Appositional modifier ‘Did you enjoy the film #PerfectPitch’.

nummod A numeric modifier associated with noun phrase is labeled with the dependency type **nummod**, as exemplified in Figure B.21.

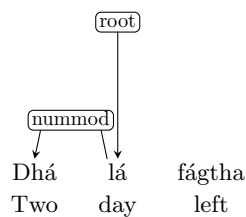


Figure B.21: Numeric modifier ‘Two days left’.

amod Adjectival modifiers of noun phrases are attached via the label **amod** as illustrated in Figure B.22.

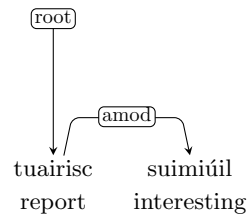


Figure B.22: Adjectival modifier ‘An interesting report’.

det The label **det** connects a determiner to the noun that governs it as illustrated in Figure B.23.

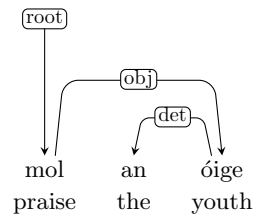


Figure B.23: Determiner ‘Praise the youth’.

case In UD, prepositions are treated as dependents on nominals and are connected via the label **case**, as illustrated in Figure B.24.

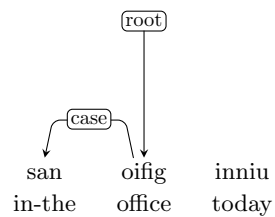


Figure B.24: Case ‘in the office today’.

acl The label **acl** is employed for both finite and non-finite clauses that function as modifiers for a noun.

acl:relcl Relative clauses are labelled with the subtype **acl:relcl** as illustrated in Figure B.25

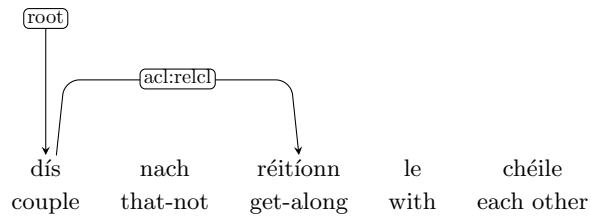


Figure B.25: Relative clause ‘A couple that doesn’t get along’.

conj In UD coordination, the initial conjunct serves as the governing element of the coordinated phrase, and all subsequent conjuncts are considered dependents, attached via the **conj** relation. This is exemplified in Figure B.26.

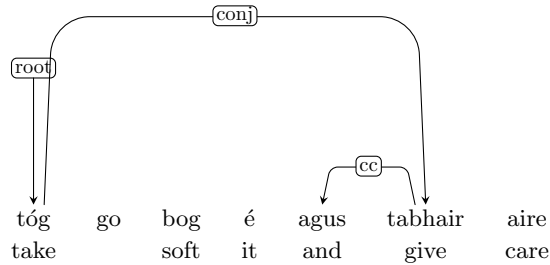


Figure B.26: Coordinating conjunction ‘take it easy and take care’.

cc The **cc** label, shown in Figure B.26, denotes the connection between the coordinating conjunction and the non-initial conjunct.

fixed The label **fixed** is one of the three relations used to represent MWEs, the other two being **flat** and **compound**. It is used for establishing grammatical expressions that function collectively as a single function word. The first token is considered the head of the fixed unit, and each following component of the MWE is connected to the head via the **fixed** label. Compound prepositions in Irish are represented as such, as exemplified in Figure B.27.

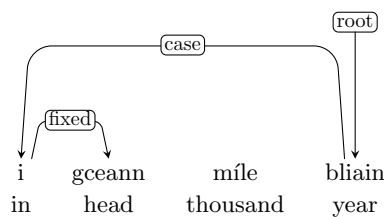


Figure B.27: Fixed ‘in a thousand years’.

flat The label **flat** is used in the context of proper nouns consisting of multiple nominal elements and is commonly used for dates as shown in Figure B.28.

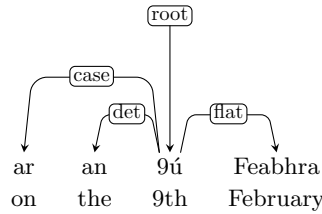


Figure B.28: Flat ‘on the 9th February’.

compound The label **compound**, exemplified in Figure B.29, is used to signify noun compounding where two or more nouns are combined to describe a distinct entity. The compound noun should possess a meaning that differs from or is more specific than the individual components combined.

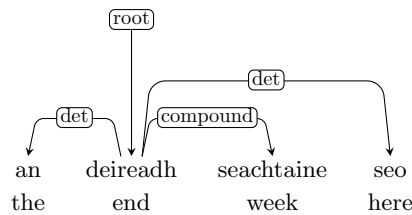


Figure B.29: Compound ‘this weekend’.

parataxis The **parataxis** relation involves a connection between clauses sentences juxtaposed without any clear coordination or subordination.

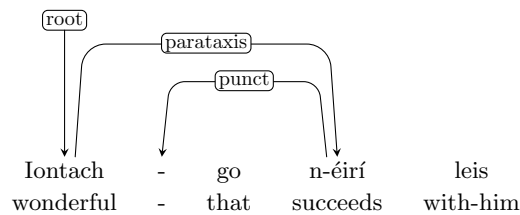


Figure B.30: Parataxis and punctuation ‘Wonderful - good luck to him’.

Where a tweet has more than one sentence, the head of the first sentence is labelled as **root** and the head of each subsequent sentence is be labelled as **parataxis:sentence** and attach to the head of the previous sentence. Non-syntactic URLs, hashtags and usernames are labelled as **parataxis:hashtag**, **parataxis:URL** and **vocative:mention** respectively. They attach to the head of the sentence they are associated with.

punct Using the label **punct**, sentence-final punctuation is attached to the head of the sentence. Coordinating punctuation is attached to the head of the non-initial coordinated clause, as shown in Figure B.30.


B.6 Language Identification

In the miscellaneous (10th) column of the CoNLL-U format, Irish words are marked with the annotation **Lang=ga**, while English words are marked with **Lang=en**. Proper nouns, metalanguage tags, and punctuation marks are not assigned any language annotation. In the case of uncertainty about the language of a token, it is labelled as Irish only if it appears in the NEID.

Appendix C

Ethical Approval for Language Contact Questionnaire Study

Oliscóil Chathair Bhaile Átha Cliath
Dublin City University



Ms Lauren Cassidy
Computing/ADAPT Centre

21st December 2022

REC Reference: DCUREC/2022/225
Proposal Title: Questionnaire on Language Contact in Irish
Applicant(s): Dr Jennifer Foster, Dr Teresa Lynn


Dear Colleagues,

Thank you for your application to DCU Research Ethics Committee (REC). Further to notification review, DCU REC is pleased to issue approval for this research proposal.


DCU REC's consideration of all ethics applications is dependent upon the information supplied by the researcher. This information is expected to be truthful and accurate. Researchers are responsible for ensuring that their research is carried out in accordance with the information provided in their ethics application.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee. Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,



Dr. Melrona Kirrane
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacaíocht
Oliscóil Chathair Bhaile Átha Cliath,
Báile Átha Cliath, Éire
Research & Innovation Support
Dublin City University,
Dublin 9, Ireland
T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie

Note: Please retain this approval letter for future publication purposes (for research students, this includes incorporating the letter within their thesis appendices).

Figure C.1: Dublin City University Ethical Approval for Questionnaire Study.

Appendix D

Full Text of Language Contact Questionnaire Study

Plain Language Statement

Research Study Title: “Questionnaire on Language Contact in Irish”

Principal Investigators: Lauren Cassidy, Dr. Teresa Lynn, Dr. Jennifer Foster

DCU, School of Computing

DCU, ADAPT Centre

Contact Details: xxx

This research looks at the way Irish interacts with English. The aim of the study is to better understand Irish speakers’ perceptions and intuitions about various forms of language contact observed in Irish language tweets and to investigate whether they align with linguistic theories of language contact.

Participation will involve the completion of a questionnaire that will take approximately 20-30 minutes. The questionnaire contains three sections 1) Plain Language Statement & Informed Consent, 2) Language Background, and 3) Name the Language.

The survey will be anonymous and no personal or identifying information will be collected. Responses will be encrypted and stored securely for a maximum of four years. Responses will then be deleted. In the event that personal information is unintentionally shared with us, it must be noted that protection of this data is subject to legal limitations. It is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions. There are no notable risks/benefits to taking part in the study. You will be able to withdraw from the study at any time by not submitting the questionnaire form. Your response will not be saved unless you submit the form. The results of this study will be disseminated in academic conferences, research papers, and a PhD thesis.

This research is part of the GaelTech project and funded by the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media.

If you have any queries and would like to contact the researchers, please contact: xxx

If you have concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support,
Dublin City University, Dublin 9. Tel: 01-7008000, Email: rec@dcu.ie

Informed Consent

- I have read the Plain Language Statement.
- I understand the information provided.
- I know how to contact the organisers of the study if I have any questions.
- If I had questions about the study, I have received satisfactory answers.
- I understand the information provided about data protection.
- I understand I may withdraw from the research study at any point.
- I understand the effort made to protect the confidentiality of the data, and that the confidentiality is subject to legal limitations.
- I consent to participate in this research study.
- I am over the age of 18.

I agree with the above statements.*

Your Language Background What is your age?*

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 and over

What is your level of Irish?*

- **Beginner/Elementary (A1-A2):** can talk about basic, familiar topics.
- **Intermediate (B1-B2):** can talk about events, experiences and plans, can understand the main points of a conversation when the official standard of Irish is used (i.e. An Caighdeán Oifigiúil as taught in schools).
- **Advanced (C1):** can understand long, challenging texts, can have spontaneous conversations in social and professional situations.
- **Fluent/Proficient (C2):** can understand almost everything they hear or read, can talk about any topic without making noticeable errors.
- **Native:** first language, acquired from birth.

Which dialect(s) of Irish do you identify with?*

- Connacht

- Munster
- Ulster
- A mix
- None
- Other...

How many languages are you proficient in?*

- 1
- 2
- 3
- 4+

How regularly do you use Irish in a formal context? e.g. professional email*

- Daily
- Weekly
- Monthly
- Less than once a month

Do you mix Irish and English in formal contexts? e.g. professional email*

- 1 (Never)
- 2
- 3
- 4
- 5 (Always)

How regularly do you use Irish in an informal context? e.g. texting, chatting*

- Daily
- Weekly
- Monthly
- Less than once a month

Do you mix Irish and English in informal contexts? e.g. texting, chatting*

- 1 (Never)
- 2
- 3
- 4
- 5 (Always)

How do you feel about mixing Irish and English?*

- 1 (Very negatively)
- 2
- 3
- 4
- 5 (Very positively)

Name the Language The following are 36 example Irish tweets. For each example, please name the language of the highlighted word. You will then be asked whether you would choose to use the highlighted word if you were to phrase the example yourself. You can then choose to explain this choice. In the case that you aren't familiar with any word, there is no need to look it up, please answer the questions as best you can and you will then have the option to provide further context for your choices in the c part of each question.

1a) *faigh réidh leis an **riail** sin**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

1b) If you were to phrase sentence 1a in an informal Irish context, how likely is it that you would include the highlighted word?*

- 1 (Very unlikely)
- 2
- 3
- 4
- 5 (Very likely)

1c) Briefly explain why

2a) *Cuirfidh mé **DM** chuici**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

2b) If you were to phrase sentence 2a in an informal Irish context, how likely is it that you would include the highlighted word?*

- 1 (Very unlikely)
- 2
- 3
- 4
- 5 (Very likely)

2c) Briefly explain why

3a) *50 bliain idir na **pics** seo**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

3b) If you were to phrase sentence 3a in an informal Irish context, how likely is it that you would include the highlighted word?

- 1 (Very unlikely)
- 2
- 3
- 4
- 5 (Very likely)

3c) Briefly explain why

4a) *emphRanganna **yoga** trí Ghaeilge anocht*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

4b) If you were to phrase sentence 4a in an informal Irish context, how likely is it that you would include the highlighted word?

- 1 (Very unlikely)
- 2
- 3
- 4

- 5 (Very likely)

4c) Briefly explain why

5a) *Imao!! Rud ar bith tusa ?*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

5b) If you were to phrase sentence 5a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

5c) Briefly explain why

6a) *níl haon ionadh orm go bhfuil na **hits** a méadú*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

6b) If you were to phrase sentence 6a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

6c) Briefly explain why

7a) *Amhrán **pop** an lae*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

7b) If you were to phrase sentence 7a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

7c) Briefly explain why

8a) *Tá'n **blag** ag lorg scríbhneoir faisean*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

8b) If you were to phrase sentence 8a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

8c) Briefly explain why

9a) *Anois a chonaic mé é seo!!!! **Wtf!***

- Irish

- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

9b) If you were to phrase sentence 9a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

9c) Briefly explain why

10a) *Cuireann an **twerking** sin isteach orm*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

10b) If you were to phrase sentence 10a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

10c) Briefly explain why

11a) *Ní féasta go **rósta** is ní céasta go pósta*

- Irish
- English
- The word exists both in Irish and English

- The word is a mix of Irish and English
- Neither

11b) If you were to phrase sentence 11a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

11c) Briefly explain why

12a) ***Raíght.** Shlog mé an t-iomlán*

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

12b) If you were to phrase sentence 12a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

12c) Briefly explain why

13a) *Roimh na **Dubs****

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

13b) If you were to phrase sentence 13a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

13c) Briefly explain why

14a) *Grma a chroí**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

14b) If you were to phrase sentence 14a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

14c) Briefly explain why

15a) *Tá keyboards beag an deachair**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

15b) If you were to phrase sentence 15a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

15c) Briefly explain why

16a) **Wish** nach raibh aon obair le déanamh agam

- Irish
- English
- Neither
- The word exists both in Irish and English
- The word is a mix of Irish and English

16b) If you were to phrase sentence 16a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

16c) Briefly explain why

17a) *Tá an **fhoireann** ar fad go hálainn**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

17b) If you were to phrase sentence 17a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1

- 2
- 3
- 4
- 5 (Very likely)

17c) Briefly explain why

18a) *níl mé ach tar éis **tweet** a léamh**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

18b) If you were to phrase sentence 18a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

18c) Briefly explain why

19a) ***Just** samhlaigh an racht feirge a mhothaímse**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

19b) If you were to phrase sentence 19a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3

- 4
- 5 (Very likely)

19c) Briefly explain why

20a) *anois am réiteach don **dioscó!****

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

20b) If you were to phrase sentence 20a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

20c) Briefly explain why

21a) *Tá siad fós ag **imirt****

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

21b) If you were to phrase sentence 21a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

21c) Briefly explain why

22a) *D'ioslodail me an **album** nua**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

22b) If you were to phrase sentence 22a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

22c) Briefly explain why

23a) ***Samplaí** anseo de logainmneacha**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

23b) If you were to phrase sentence 23a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

23c) Briefly explain why

24a) *Comhghairdeas leis na **leaid**s**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

24b) If you were to phrase sentence 24a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

24c) Briefly explain why

25a) *#Gaeilge mar rogha ar **aip** agus ATM**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

25b) If you were to phrase sentence 25a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

25c) Briefly explain why

26a) *deochanna le mo **kinda** col ceathrar**

- Irish
- English

- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

26b) If you were to phrase sentence 26a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

26c) Briefly explain why

27a) *beir 2 ag **partyáil****

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

27b) If you were to phrase sentence 27a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

27c) Briefly explain why

28a) *Beidh muid ag **plé** an ábhair seo**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English

- Neither

28b) If you were to phrase sentence 28a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

28c) Briefly explain why

29a) *Ag déanamh meaitseáil ar an ríomhaire**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

29b) If you were to phrase sentence 29a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

29c) Briefly explain why

30a) *Drámaí deasa inniu**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

30b) If you were to phrase sentence 30a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

30c) Briefly explain why

31a) ***Absolutely** álainn.**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

31b) If you were to phrase sentence 31a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

31c) Briefly explain why

32a) *Tá tú an-mhaith ag an **housework** inniu.**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

32b) If you were to phrase sentence 32a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1

- 2
- 3
- 4
- 5 (Very likely)

32c) Briefly explain why

33a) *Tá físeáin **haiceanna**, cláir agus stíleanna gruaige ann**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

33b) If you were to phrase sentence 33a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

33c) Briefly explain why

34a) *Tá **fáilte** roimh gach duine**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

34b) If you were to phrase sentence 34a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3

- 4
- 5 (Very likely)

34c) Briefly explain why

35a) *An féidir leat rt an **ocáid** seo**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

35b) If you were to phrase sentence 35a in an informal Irish context, how likely is it that you would include the highlighted word?*

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

35c) Briefly explain why

36a) *Tá arán **banana** agam.**

- Irish
- English
- The word exists both in Irish and English
- The word is a mix of Irish and English
- Neither

36b) If you were to phrase sentence 36a in an informal Irish context, how likely is it that you would include the highlighted word?

- Very unlikely
- 1
- 2
- 3
- 4
- 5 (Very likely)

36c) Briefly explain why

Appendix E

Full Language Classification

Results

Tables E.1, E.2, E.3, and E.4 show the full language classification results from the language contact questionnaire study for GA, CS, BOR, and AMBI categories respectively.

Word	Irish	English	Mix	Both	Neither
<i>riail</i>	93.75%	0.00%	1.95%	3.13%	1.17%
<i>rósta</i>	88.28%	0.39%	5.08%	4.30%	1.95%
<i>Grma</i>	96.48%	0.39%	0.39%	0.00%	2.73%
<i>fhoireann</i>	99.22%	0.39%	0.00%	0.39%	0.00%
<i>imirt</i>	99.22%	0.00%	0.39%	0.39%	0.00%
<i>Samplaí</i>	89.45%	0.00%	6.25%	4.30%	0.00%
<i>plé</i>	96.88%	0.00%	1.95%	1.17%	0.00%
<i>deasa</i>	97.27%	0.00%	0.39%	0.39%	1.95%
<i>fáilte</i>	98.44%	0.00%	0.39%	1.17%	0.00%

Table E.1: Language classification results from the language contact questionnaire study for GA category.

Word	Irish	English	Mix	Both	Neither
<i>pics</i>	0.78%	76.95%	10.55%	9.38%	2.34%
<i>yoga</i>	1.56%	56.25%	1.56%	21.48%	19.14%
<i>Dubs</i>	0.39%	87.89%	2.34%	8.20%	1.17%
<i>Wish</i>	0.78%	95.31%	1.95%	1.17%	0.78%
<i>Just</i>	1.56%	94.92%	0.78%	2.73%	0.00%
<i>album</i>	5.08%	74.22%	2.34%	17.58%	0.78%
<i>kinda</i>	1.95%	89.84%	1.17%	0.78%	6.25%
<i>Absolutely</i>	1.56%	96.09%	0.78%	1.17%	0.39%
<i>housework</i>	0.39%	98.44%	0.39%	0.39%	0.39%

Table E.2: Language classification results from the language contact questionnaire study for CS category.

Word	Irish	English	Mix	Both	Neither
<i>DM</i>	1.56%	87.11%	2.34%	5.47%	3.52%
<i>hits</i>	1.56%	94.53%	1.17%	2.34%	0.39%
<i>blag</i>	69.14%	4.69%	14.45%	8.59%	3.13%
<i>twerking</i>	0.39%	94.92%	0.78%	1.95%	1.95%
<i>keyboards</i>	0.00%	98.44%	0.78%	0.00%	0.78%
<i>tweet</i>	0.39%	82.03%	1.17%	15.23%	1.17%
<i>aip</i>	58.20%	4.69%	17.19%	13.67%	6.25%
<i>meaitseáil</i>	59.38%	1.17%	31.25%	5.47%	2.73%
<i>haiceanna</i>	41.41%	3.13%	37.50%	2.73%	15.23%

Table E.3: Language classification results from the language contact questionnaire study for BOR category.

Word	Irish	English	Mix	Both	Neither
<i>lmao</i>	0.78%	84.42%	1.17%	1.17%	14.45%
<i>pop</i>	1.95%	33.20%	1.95%	62.11%	0.78%
<i>Wtf</i>	0.39%	92.19%	1.56%	0.39%	5.47%
<i>Raight</i>	13.67%	17.19%	51.17%	8.59%	9.38%
<i>diosc6</i>	54.69%	2.73%	19.14%	23.05%	0.39%
<i>leaid</i>	24.22%	3.52%	49.61%	19.53%	3.13%
<i>party6il</i>	4.30%	7.42%	76.95%	4.30%	7.03%
<i>oc6id</i>	89.06%	5.47%	1.56%	0.78%	3.13%
<i>banana</i>	11.72%	13.67%	0.78%	71.48%	2.34%

Table E.4: Language classification results from the language contact questionnaire study for AMBI category.

Appendix F

Full Phase 1 Codebook

Code	Refs	Definition
Irish	213	Irish language or Irishness
Preferred term	189	Suggestion of another way to express the same concept
Vocabulary	130	Set of available words at the level of the individual or the language community
Personal taste	119	Expression of like, dislike, individual opinion
Usage	116	Use of term
Understandability	103	Comprehensibility
English	81	English language or Englishness
Naturalness	58	Intuition
Familiarity	56	Recognition, knowledge
Ease	54	Accessibility
Context	52	Setting or situation
Uncertainty	49	Controversial
Borrowing	48	Loanword
Phrasing	43	Syntax, grammar, structure
Formality	40	Register
Code-mixing	37	Combining languages
Normality	35	Ordinariness
Abbreviation	34	Shorthand
Accuracy	33	Correctness
Length	30	Number of characters
Effect	28	Impact
Spelling	27	Orthography
Age of term	23	Time of term existence
Semantics	23	Meaning

Code	Refs	Definition
Speech	23	Vocal communication as opposed to writing
Reasonableness	22	Extent to which term follows logic, makes sense
Proficiency	22	Person's language ability
Interchangablity	21	Exchangablity of terms
Clarity	20	Clearness
Origin	19	Etymology
Translation similar- ity	19	Degree to which term is like translation
Gaeltacht	18	Irish-speaking region
Appropriacy	17	Suitability
Person age	16	How old a person is
Quality	16	Measurable standard
Shame	16	Emotion related to deficiency or inappropriate behaviour
Writing	15	Communication via text as opposed to speech
Translation	14	Conversion to another language
Effort	13	Amount of work
Multilingualism	13	Relating to more than one language
Neccessity	13	Need
Fit	11	Suitability
Phonetics	11	Sound
Speed	11	Rapidity
Named entity	10	Specific object, person, location, Organisation, or other unique and identifiable entity
Initialism	10	An abbreviation consisting of initial letters e.g. acronym
Social media	10	Web-based communication platforms
Acceptance	9	Act of deeming valid or adequate
Assimilation	9	Integration or incorporation
Utility	8	Usefulness
Simplicity	7	Straightforwardness
Texting	7	SMS
Education	6	Learning
Habit	5	Pattern of behaviour
Importance	5	Significance
Memory	5	Recall
Technology	5	Inventions such as electronics, computers, telecommunications etc.
Accident	4	Without intention
Interjection	4	Exclamation
Sport	4	Games or athletic activity

Code	Refs	Definition
Development	3	Growth or evolution
Identity	3	Sense of self
Problem	3	Issue or obstacle
Adequate	2	Satisfactory
Community	2	Group of people with commonality
Discourse marker	2	Linking phrase
Email	2	Electronic mail, digital message
Feeling	2	Subjective experience
Hiberno English	2	Irish variety of English
Invention	2	Creation
Not acceptable	2	Inadequate
Transliteration	2	Writing words of one language using the spelling system of another
Action	1	Verb
Damage	1	Harm
Exclusion	1	Omission
Fatigue	1	Tiredness
Idea	1	Concept
Lockdown	1	Confinement
Spectrum	1	Non-binary classification
Status	1	Social rank

Table F.1: Full phase 1 codebook.