The Automatic Processing of Multiword Expressions in Irish

Abigail Walsh

B.A.

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (PhD)

to the

Dublin City University
School of Computing

Supervisors:
Dr. Jennifer Foster
Dr. Teresa Lynn

January 2023

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, 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: 

(Candidate) ID No.: 17213757

Date: 10/01/2023

Contents

List of Figures

List of Tables

The Automatic Processing of Multiword Expressions in Irish

Abigail Walsh

Abstract

It is well-documented that Multiword Expressions (MWEs) pose a unique challenge to a variety of NLP tasks such as machine translation, parsing, information retrieval, and more. For low-resource languages such as Irish, these challenges can be exacerbated by the scarcity of data, and a lack of research in this topic. In order to improve handling of MWEs in various NLP tasks for Irish, this thesis will address both the lack of resources specifically targeting MWEs in Irish, and examine how these resources can be applied to said NLP tasks.

We report on the creation and analysis of a number of lexical resources as part of this PhD research. Ilfhocail, a lexicon of Irish MWEs, is created through extracting MWEs from other lexical resources such as dictionaries. A corpus annotated with verbal MWEs in Irish is created for the inclusion of Irish in the PARSEME Shared Task 1.2. Additionally, MWEs were tagged in a bilingual EN-GA corpus for inclusion in experiments in machine translation. For the purposes of annotation, a categorisation scheme for nine categories of MWEs in Irish is created, based on combining linguistic analysis on these types of constructions and cross-lingual frameworks for defining MWEs.

A case study in applying MWEs to NLP tasks is undertaken, with the exploration of incorporating MWE information while training Neural Machine Translation systems. Finally, the topic of automatic identification of Irish MWEs is explored, documenting the training of a system capable of automatically identifying Irish MWEs from a variety of categories, and the challenges associated with developing such a system.

This research contributes towards a greater understanding of Irish MWEs and their applications in NLP, and provides a foundation for future work in exploring other methods for the automatic discovery and identification of Irish MWEs, and further developing the MWE resources described above.

Acknowledgments

I am fond of quotes, and take inspiration from the familiar words of beloved and oft-revisited stories. This thesis contains a selection taken from the works of J.R.R. Tolkien, author of such classics of my childhood as The Hobbit, and The Lord of the Rings. I draw much comfort from these stories of perseverance, of hope, and of the power of even little acts of kindness to generate lasting change. When reflecting on my thesis journey, I cannot help but be reminded of that line delivered so poignantly by the most faithful of friends Samwise Gamgee: "I can't carry it for you, but I can carry you!" There are so many people I feel indebted to, for how they have carried me during these years of my life.

I would like to begin by acknowledging the incalculable work done by my supervisors, Dr. Teresa Lynn and Dr. Jennifer Foster. I could not have asked for more a dedicated, hard-working, supportive, and encouraging team of experts, assisting in everything from guiding me through my very first conference and academic publication, to providing excellent and expert advice on every conceivable aspect of my research, to furnishing me with the skills and expertise and networks necessary to becoming a fully-fledged researcher in my own right. Teresa, your insightful comments on everything from proper citation to effective communication to career advice have been of indescribable value to me in my academic journey, and Jennifer, your wise expertise and experience in so many fields has helped to broaden my own perspective on countless occasions. You have both served as role-models to me of dynamic and capable women in research, and I am inspired by your work-ethic, your breadth of knowledge, and your consistent striving for excellence in everything you do. This thesis is a testament to many, many hours of annotation, consultation, proof-reading, critique, revisions, re-submissions, shifting deadlines, and long video calls, interrupted by sometimes truly terrible internet connections. Thank you both for your patient perseverance through all of that and more.

My time at the ADAPT Centre was shaped by the contributions, insights, assistance and encouragements of so many of the people there, and the many wonderful friendships that developed. To name but a few, I would like to thank fellow researchers James Barry, Daria Dzendzik, Pintu Lohar, Yasufumi Moriya, Alan Cowap, Piyush Arora, Abhishek Kaushik, Alberto Poncelas, Filip Klubička, Cara Greene, and Joachim Wagner, for the many ways you have each contributed to my journey. Thank you to Sheila Castilho, for first introducing me to the wise practise of 'always write it down' and Andy Way, for convincing me to stay at ADAPT in the first place! Some members of ADAPT have moved on to new and exciting endeavours elsewhere, I would like to acknowledge and thank Carla Parra Escartín and Eva Vanmassenhove for your emotional support while struggling with PhD burnout. To Dimitar Shterinov, David Woods, Henry Elder, Longyue Wang, Koel Dutta Choudhary, Colm O'Hehir, and Jinhua Du (my very first supervisor at ADAPT), thank you for your role in making my time at ADAPT one of personal as well as academic growth.

Ba mhaith liom mo bhuíochas a chur in iúil, ar fheabhas a gcuid comhairle, comhrá, cuidiú agus cairdeas, don daoine ón 'GaelGang': comh-taighdóirí Lauren Ní Chasaide, Sarah McGuinness, Jason Phelan, Helen McHugh, Jane Dunne, Órla Ní Loinsigh, Séamus Lankford, John Judge, Carol O'Toole, Mícheál John Ó Meachair, agus Meghan Dowling. Go raibh míle maith agaibh as gach rud!

To the many researchers in my field whose work formed the foundation of my thesis, I am deeply indebted to you, thank you! I would like to say a special thank you to Katie Ní Loingsigh, for many interesting conversations about Irish idioms, and to Agata Savary and Carlos Ramisch, both of whom helped me first to navigate the exciting world of MWEs, and whose support and guidance during my PhD journey was greatly appreciated.

I have so many friends to thank for their encouragement during my PhD journey. Thank you to Jonny, Sam, Liam, Dan, Alex, Christie, Reii, Richard and Faolán for the help, guidance, expertise and inspiration! Thanks to the members of City Church

Dublin for being like a family away from home. Thank you to Abby and Nicole, for being the wonderful friends you are who helped me through some truly difficult times, you are both such blessings in my life. To Lizzie and T, thank you for putting up with me during those long years of PhD stress! To Ross, thank you for rooting for me, right to the end. To Jess, thank you so much for your support and friendship, you are an amazing human and I will never stop appreciating you. To my incredible family, I am so grateful for your kindness and help. Thank you Steven, Selwan, Philip and Katie for all your advice and kind words of encouragement. Thank you Chris and Iryna for being champion supporters, especially while writing this thesis! To Young, and especially to Lisa, you both helped me stay sane during the long, final stretch to finishing. Thank you for keeping me going!

Finally, I truly could not have finished this journey without the eternal, consistent and much-needed support of my parents, particularly in those final months of writing. You have both been my biggest cheerleaders, from my first early days of education, right through to what seems to be the end at last (for now, at least!) I love and appreciate you both so much. This thesis is dedicated to you, for it was you who first inspired my love of learning, and without you I would never have put these words to paper.

Chapter 1

Introduction

> *“It’s the job that’s never started as takes longest to finish.”*
>
> J.R.R. Tolkien, *The Fellowship of the Ring*

Multiword Expressions, or MWEs, pose a significant problem for NLP. They have been called a “pain in the neck” (Sag et al., 2002), a “hard nut” to crack (Villavicencio et al., 2005), and “hard going” (Rayson et al., 2010). This topic may be considered “much ado about nothing” (de Marneffe et al., 2009), or perhaps they are simply “rare birds that cause a stir” (Savary et al., 2019a).

Regardless of how this problem is presented, expressions such as ‘a piece of cake’, ‘kick the bucket’, ‘put your foot down’, ‘make out’, ‘put up with’, ‘beat around the bush’ have been a major bottleneck in NLP tasks (Sag et al., 2002) for many years now, in part due to just how varied such expressions can be. To compound the problem, they make up a considerable percentage of our lexicon (Jackendoff, 1997b) and are found prominently in written text. While interest in this topic has only increased in recent years, with several initiatives dedicated to processing MWEs, this research tends to be focused on MWEs in English, or other majority languages. For many minority and lesser-spoken languages, this research is limited in its application, whether this is due to the shortage of language resources such as

lexicons, or linguistic tools; or due to linguistic features of the language itself. The purpose of this thesis is to focus on the characteristics of Irish MWEs and to develop methods either novel or employed by other languages to automatically process these MWEs in text.

A Low-Resource Language

The position of the Irish language is somewhat unique. Its status as the official language of Ireland affords it official legislative and constitutional protection by the Irish government. Additionally, it was given status as an official language of the European Union in 2007. Despite these levels of recognition and protection, English is the dominant language used in Ireland, and the default language used for communication in most spaces. Recent census data (CSO, 2016) reports that 1.76 million people are recorded to be able to speak Irish (39.8% of the population), with only 73,800 (1.6% of the population) reporting to use Irish in their daily lives.

In terms of language resources, Irish is considered a low-density or low-resource language, and is not well supported in the digital sphere (Judge et al., 2012; Lynn, 2022). Language technology and support for the language have improved in recent years, with a growing body of research in the area of NLP for Irish, but still this support lags far behind when compared to English. An estimated 0.5% of Ireland-based `.ie` domains offer content through the medium of Irish, and the localisation of content or development of tools in Irish language technology has not been the priority of large technology companies, e.g. software developed by Google being dependent on volunteer translation efforts for the localisation of their content (Lynn, 2022).

This lack of digital support has been highlighted as putting Irish at risk of digital extinction, a phenomenon by which the lack of technological support contributes to the decline in language use. As technological support for majority languages, and English in particular, grows at an increasingly disproportionate rate compared to minority languages, there is a risk that these languages will be increasingly excluded from digital spaces (Judge et al., 2012; Pasikowska-Schnass, 2020; Lynn, 2022).

While support for Irish technology lags behind English, there is a growing body of resources online. Irish language users are active on websites such as Twitter, where over 4.5 million Tweets have been recorded in Irish.[1] An Vicipéid, or Irish Wikipedia, is a community fostered open source repository of knowledge, with over 56,000 articles published in Irish.[2]

Language resources for Irish have increased particularly in the past 10 years, thanks to the efforts of several data collection and corpus-creation initiatives, largely funded through EU projects for the development of machine translation resources. Specialised resources related to MWEs, such as the Irish Universal Dependencies Treebank (Lynn and Foster, 2016), and TwittIrish, the Irish UD Twitter Treebank (Cassidy et al., 2022), have been published, providing in-depth linguistic analysis of Irish text. These resources were aided by the development of specialised tools for the processing and parsing of Irish, such as the tokeniser and part-of-speech tagging system, developed by Uí Dhonnchadha (2009). Researchers at Fiontar & Scoil na Gaeilge, Dublin City University have developed lexical resources such as the Téarma.ie project, a national terminology database for Irish,[3] the Peadar Ó Laoghaire Idiom Collection (Ní Loingsigh, 2016) and the Irish placenames database Logainm (Měchura and Ó Raghallaigh, 2012). More information on these and other language resources is provided in Chapter 5.

Irish Machine Translation (MT) has seen developments in recent years, with the Tapadóir project (Dowling et al., 2015) providing a bespoke EN→GA MT system to the in-house translators of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media. Research on Neural MT for Irish has developed in recent years (Dowling et al., 2018; Defauw et al., 2019; Lankford et al., 2021). The topic of Neural MT for Irish is further explored in Chapter 6.

There has also been steady progress in the area of speech technology, with the development of tools such as the Irish text-to-speech system and speech synthe-

[1] http://indigenoustweets.com Figures as of June 2022.
[2] Figure from 30/03/2022.
[3] `https://www.tearma.ie/`

siser ABAIR (Ní Chasaide et al., 2017), which have seen application in assistive technology development (Ní Chasaide et al., 2022). Kevin Scannell, a researcher developing technology for minority languages and Irish in particular, was responsible for the creation of several tools for language use. These include the grammar checker and spell checker,[4] which have been incorporated alongside the ABAIR system in *An Scéalaí* (The Storyteller), the intelligent-Computer-Assisted Language Learning (iCALL) platform (Ní Chiaráin et al., 2022).

Despite these efforts, there is still a significant lack of technological support for many areas of Irish NLP, including automatic speech recognition, sentiment analysis, information retrieval, question answering, named-entity recognition, automatic subtitling, natural language generation, and other areas. This thesis is an attempt to fill the gap of dealing with MWEs, and the automatic processing of MWEs, particularly their identification.

1.1 Research Questions

Arising from an examination of the literature and research surrounding the topic of MWE identification in Irish, the following research questions were identified as areas that require addressing:

Research Question 1: Is it possible to classify Irish MWEs under a multilingual classification framework?

Research Question 2: How can existing Irish resources be leveraged to generate Irish MWE resources?

Research Question 3: What challenges exist for developing a system of automatic identification of Irish MWEs?

Research Question 4: Is knowledge of MWEs useful in an Irish↔English MT system? (Case study)

[4] `https://cadhan.com/lsg/index-en.html`

4a: How do we represent knowledge of MWEs in an MT system?

4b: How can we evaluate the effect of adding MWE information to an Irish↔English MT system?

RQ1 stems from the importance of aligning this research with similar efforts in other languages, in an attempt to leverage the resources available in those other languages. Similarly, RQ2 is concerned with leveraging existing resources for Irish, and combining the lexical entries with linguistic information gleaned from RQ1. Both of these research questions are explored in Chapters 2, 3, 4 and 5.

RQ3 attempts to define the problem of automatic classification of Irish MWEs thoroughly by examining what the challenges of this task are, considering the available resources explored in RQ1 and RQ2. This research question forms the basis of Chapters 7 and 8.

RQ4 focuses on the practical applications of this research. Chapter 6 is devoted to addressing this research question. RQ4a is addressed in Chapter 5, as we describe the creation of MWE-aware resources for training MT systems. RQ4b is addressed in Chapter 6, where we explore several metrics for evaluating the impact of including MWE information in our NMT models.

1.2 Thesis Structure

This first chapter introduces the problem of multiword expressions, as well as the importance and prevalence of language technology today, and the status of Irish in this respect. The rest of this thesis is structured as follows:

Chapter 2 explores the key concepts of this research work, defining what exactly we mean by multiword expressions, and how they are challenging to process automatically. We introduce two multilingual frameworks that have significant relevance in our research: PARSEME and Universal Dependencies. We also discuss the topic of language resources and how they are important to many NLP tasks, including

MWE processing. We present a number of existing language resources, many of which inform our own practices in developing MWE-specific resources for Irish.

Chapter 3 introduces the Irish language, providing an overview of its syntax, and highlighting certain features that are relevant to our understanding of MWEs in Irish. We also describe Irish language resources that we leverage in our research.

Chapter 4 presents a review of MWE discussion in Irish linguistics literature. We also describe the treatment of MWEs in the two multilingual frameworks introduced in Chapter 2: the PARSEME annotation guidelines, and the Universal Dependencies framework. We present our typology of Irish MWEs, based on a synergy between all of these sources. Some difficult cases are also discussed, along with the current decisions for such cases.

Chapter 5 describes the creation or enhancement of three language resources: the Ilfhocail lexicon of Irish MWEs, the PARSEME annotated corpus of Irish verbal MWEs, and an MWE-annotated parallel corpus for machine translation. We present a brief analysis of the MWE information included in each corpus.

Chapter 6 describes experiments on incorporating our MWE information into neural machine translation systems for GA↔EN machine translation. We present some background information on the field of machine translation, describing the architecture used for neural machine translation, and previous research on the incorporation of MWEs into machine translation systems. We then detail our own experiments in including MWEs as features in neural machine translation systems, evaluating our systems on two datasets, whose creation was described in Chapter 5.

Chapter 7 provides context to the task of identification of MWEs, one of the main subtasks of MWE processing, alongside methodologies commonly used to solve this task. We describe two shared tasks dedicated to the identification of MWEs, the SemEval 2016 DiMSUM task (Schneider et al., 2016b), and the PARSEME Shared

Task on the automatic identification of verbal MWEs (Savary et al., 2017). As one of the more widely adopted architectures in many NLP tasks in recent years, we explore the use of the Transformer model, and BERT, a pre-trained language model which has seen many applications for a wide range of NLP tasks (Devlin et al., 2019). This architecture is applied in our experiments in the identification of MWEs explored in Chapter 8. Some context is also provided as to how Irish was treated in the PARSEME shared task.

Chapter 8 details the development of a system for the automatic identification of Irish MWEs. We describe two series of experiments, aimed at investigating the abilities of two pre-trained language models (a multilingual language model mBERT,[5] and an Irish-language language model gaBERT (Barry et al., 2022)) to perform token-level classification of verbal MWEs when fine-tuned on a small dataset annotated with verbal MWEs. The models are evaluated, with the results compared to those achieved by systems submitted to the PARSEME shared task.

Chapter 9 summarises our main findings in our research, and the novel findings of this thesis. We discuss how each of the research questions were addressed in our work. Finally, we present some of the areas which remain to be explored in this topic, including planned future work.

Appendix A contains the annotation guidelines created for the manual annotation of Irish MWEs described in Chapter 5, based on the categories of MWEs described in Chapter 4.

Appendix B contains descriptions of some exploratory experiments in the task of MWE discovery, which were not the focus of this research, but constitute an area of future work, with applications in the extension of lexical resources for Irish MWEs.

[5] `https://huggingface.co/bert-base-multilingual-cased`

Appendix C contains a description of the creation of an annotate corpus of verbal MWEs for English, for inclusion in Edition 1.1 of the PARSEME shared task. This work helped to inform the author on the process of annotating MWEs, and served as an introduction to the PARSEME framework and annotation guidelines.

1.3 Publications

- Abigail Walsh, Claire Bonial, Kristina Geeraert, John P McCrae, Nathan Schneider, Clarissa Somers (2018): Constructing an Annotated Corpus of Verbal MWEs for English. In Proceedings of Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), At: COLING 2018, Sante Fe, USA, August 2018

  This publication details the creation of a annotated corpus of verbal MWEs for English, for submission to the PARSEME Shared Task 1.1.

- Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, Abigail Walsh (2018): Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), At: COLING 2018, Santa Fe, USA, August 2018

  This publication describes Edition 1.1 of the PARSEME shared task, which included new languages, one of which was English.

- Abigail Walsh (2019): Identifying Categories of Multiword Expressions in Irish for Automatic Processing [Slideshow]. At the 17th International Conference

on Minority Languages (ICML XVII), Leeuwarden, The Netherlands, May 2019

This presentation reported on our progress on developing a typology of MWEs in Irish, while also highlighting some open questions in the topic.

- Abigail Walsh, Teresa Lynn, Jennifer Foster (2019): Ilfhocail: A Lexicon of Irish MWEs. In Proceedings of Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, Italy, August 2019

  This publication describes the creation of the Ilfhocail lexicon, and includes a manual inspection of a sample of the lexicon.

- Abigail Walsh, Teresa Lynn and Jennifer Foster (2020): Annotating Verbal MWEs in Irish for the PARSEME Shared Task 1.2. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, At: COLING 2020, Barcelona, Spain (Online), December 2020

  This paper reports on the creation of the PARSEME annotated corpus for Irish verbal MWEs, for inclusion in Edition 1.2 of the shared tasks.

- Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, Hongzhi Xu (2020): Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, At: COLING 2020, Barcelona, Spain (Online), December 2020

  This publication describes Edition 1.2 of the PARSEME shared task, which included a new task focus, and several new languages, one of which was Irish.

- Sarah McGuinness, Jason Phelan, Abigail Walsh and Teresa Lynn, (2020): Annotating MWEs in the Irish UD Treebank. In Proceedings of the Fourth

Workshop on Universal Dependencies (UDW 2020), Barcelona, Spain (Online), December 2020

In this paper, we describe the treatment of MWEs in the Irish UD Treebank, including tests for determining MWEs of the fixed, flat and compound type.

- James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair and Jennifer Foster: gaBERT - an Irish Language Model. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), Marseille, France, June 2022

  This publication details the creation of an Irish monolingual pre-trained language model, gaBERT. Some experiments are described, including an experiment in MWE identification, intended to compare the performance of gaBERT to other multilingual pre-trained language models, with gaBERT generally outperforming these models.

- Abigail Walsh, Teresa Lynn and Jennifer Foster: A BERT's Eye View: Identification of Irish Multiword Expressions using Pre-trained Language Models. In Proceedings of the 18th Workshop on Multiword Expressions, At: LREC 2022, Marseille, France, June 2022

  This publication details experiments in fine-tuning pre-trained language models for the task of automatic identification of Irish MWEs. We compare our system with others submitted for the PARSEME Shared Task Edition 1.2, and find our model outperforms the top ranking model for this task.

- Abigail Walsh, Teresa Lynn, Jennifer Foster: Comparison of MWE-annotated corpora for the Irish language. In Proceedings of Europhras 2021 (EUROPHRAS 2021) - To Appear

  This presentation reports on an analysis of three MWE-aware resources for Irish, comparing the quality of two manually annotated corpora with an automatically tagged corpus.

Chapter 2

Background and Definitions

> *'Exactly!' said Thorin; 'but could you be more plain?'*
>
> J.R.R. Tolkien, *The Hobbit*

This chapter explores many of the concepts relating to our research on the processing of MWEs for Irish, and helps to put the rest of the research described in this thesis into context. Multiword expressions are a topic of interest for several fields (e.g. lexicography, syntax, corpus linguistics, language learning, etc.) However for the field of NLP, we are primarily interested in how to effectively and beneficially encode MWEs for their inclusion in a number of NLP tasks. To that end, we provide an overview of the properties displayed by MWEs, and how those properties might affect NLP performance. We also identify and discuss some of the main tasks within the topic of MWE processing, specifying which of those tasks are addressed in this thesis.

In order to address **RQ1**, we present two multilingual frameworks for the treatment of MWEs, with which we align our treatment of Irish MWEs. The importance of Language Resources (LRs) in this task is then highlighted. We provide some information on the forms that LRs can take, and include details of LRs that are useful to us in our task, whether that is through benefiting from the approaches

used in the creation or structuring of such resources, or for the purposes of creating multilingual resources for Irish and English.

2.1 Defining Multiword Expressions

Our first task is to provide a precise and sufficient definition of what we mean by a Multiword Expression (MWE), so that we can begin to understand how to process them automatically. This is not an easy task, as we shall see. However, it is crucially important for every aspect of MWE processing, from annotation judgements, and the collection and enhancement of MWE-aware resources, to the development of systems capable of accurate automatic processing of MWEs. Below, we explore some of the variations in how this phenomenon is defined, and provide the definition we will be applying for our research.

2.1.1 What is a Word?

A natural starting point when defining a Multiword Expression is to begin with the term itself, and to establish what is meant by a 'word', in order to decide what is signified by 'multiword'. Given how fundamental the concept of a 'word' is within the fields of linguistics and language, it is perhaps surprisingly problematic to define precisely what this term means. Let us take the word 'cat' as an example. Intuitively, we may understand the relationship between 'cat' and 'cats', with the latter being a variant of the former. In lexicography and linguistics, it is traditional to differentiate between *wordforms* and *lexemes*, i.e. 'cats' is a wordform of the lexeme 'cat', and can be derived using regular English rules of pluralisation. However, this does not help us in our definition of a 'word' – should we equate them with wordforms or with lexemes? Mel'čuk (1995) summarise this issue, saying "we know the restive character of the word *word*, which, until now, has escaped the attempts to circumscribe it with precision, although much has been written about this subject throughout the

decades".[1]

Di Sciullo and Williams (1987) attempt this challenging task and present four distinct definitions, three of which are applicable to written language. These definitions allow for words to exist as a listeme,[2] a morphological object, or a syntactic atom; each definition aligns with a domain or role that a word can play, in the semantic, morphological or syntactic context. Manning and Schütze (1999a) point out the difficulty such a series of definitions poses to the computational linguist, suggesting that a simple definition of a *graphic* word, or "contiguous alphanumeric characters with spaces on either side" is perhaps more practical, and yet not entirely sufficient, as show by strings such as '€22.50', 'www.cat-bounce.com', 'Ke$ha', 'can't', 'devil-may-care', and '1337ness'. Whether to treat each of these strings as words is debatable, however the latter example (an internet slang term for 'leet-ness' or 'elite-ness') demonstrates how such strings can exhibit word-like behaviours such as inflection.

The problem becomes even more thorny when we expand our perspective beyond the English language. Some language writing systems do not use whitespace characters between words, as with Chinese, Burmese, and Classical Latin, while others use characters other than whitespace to delimit words, as with Tibetan and Ge'ez script. As an analytic language,[3] English tends to be morphologically poor, however morphologically-rich languages such as the agglutinative language[4] Turkish or the polysynthetic language[5] Greenlandic form new words through the addition of many morphemes, in principle allowing for the creation of words of arbitrary length or complexity, e.g.

Nalunaarasuartaatee-raarannguallioqatigiiffissua-lioriataallaqqissupiloru-

[1]Translation of original text: *on connaît tout aussi bien le caratère rétif du mot mot, qui, jusqu'à présent, a échappé aux tentatives de le circonscirer avec précision et a fait couler beaucoup d'encre pendant des décennies*, as given in Ramisch (2015).

[2]i.e. a chunk of language that must be memorised as part of an internal list.

[3]An analytic language uses particles, prepositions and word order to convey relationships between words, rather than inflection.

[4]Agglutinative languages inflect through the concatenation of discrete morphemes.

[5]Polysynthetic languages allows for single words to be inflected with lexical morphemes, e.g. nouns, adjectives, etc.

jussuanngortartuinnaka-sinngortinniamisaalinn-guatsiaraluallaqqooqiga-minngamiaasiinngooq,

the so-called longest word in Greenlandic, and clocking in at 156 characters (Kyed, 2020).[6]

That said, it appears the humble word defies a simple definition. For our purposes then, we follow the example of Ramisch (2015) and adopt the flexible definition offered by Evert (2004), which considers a word as "an entirely generic term which may refer to any kind of lexical item, depending on the underlying theory or intended application." This allows us the freedom to consider "multiwords" as equivalent to "multilexemes", i.e. minimal linguistic units that convey a meaning. We maintain this view that MWEs should express some specific meaning (or meanings) in our research; a view which has bearings on our categorisation efforts and annotation decisions made in later chapters.

2.1.2 Definitions Abound

Similar to words, it can be a challenge to find a precise definition for MWEs as different definitions abound. Discriminant properties of multiword expressions can vary depending on which features are considered interesting or of importance (Constant et al., 2017). However, many definitions (Choueka, 1988; Manning and Schütze, 1999b; Moon, 1998; Sag et al., 2002; Baldwin and Kim, 2010) identify the following two conditions which must be met to qualify a phrase as a multiword expression:

1. The expression must consist of more than one word
2. The expression must exhibit some level of idiosyncrasy

The **first condition**, as already discussed, appears to be trivial but in practice can be challenging to determine. While most often MWEs appear as word units separated by whitespaces, sometimes MWEs may be composed of a single typographic

[6]The translation is given as 'There were reports that they apparently – God knows for how many times – once again had considered whatever I, my poor condition despite, still could be considered to be quite adept and resourceful as initiator to put a consortium together for the establishment of a range of tiny radio stations.'

unit, particularly in highly-productive languages, such as Basque (Alegria et al., 2004), or with non-segmenting languages, such as Chinese (Wang and Yu, 2010). We term these constructions as single-token MWEs.

The **second condition** is even more complex to determine, as there are many levels on which idiosyncrasy (or alternatively, idiomaticity) can be exhibited (Sag et al., 2002). We explore these levels of idiomaticity in Section 2.2.1 below.

At this point, it is also worth mentioning the variety of terms used to describe this phenomenon. Multiword expressions have also been referred to as idioms, collocations (Firth, 1957), multi-word (lexical) units (Cowie, 1981), phraseological units (Gläser, 1986), phrasal lexemes, neutral syntagmas (Lipka, 1992), phrasemes (Mel'čuk, 1995), and fixed expressions (Moon, 1998). Many of these terms overlap, but there are several distinctions across these terms, with certain properties more strongly associated with one term over another. For instance, an idiom (e.g. 'a piece of cake') is typically used to describe an MWE that displays idiomaticity on the semantic level; a collocation (e.g. 'salt and pepper') is usually applied to an MWE that displays idiomaticity on the statistical level; a fixed expression (e.g. 'more or less') is a term applied to an MWE that is syntactically fixed, and so on.

To illustrate how our established definition of 'multiwords' factors into our definition of 'multiword expression', we take an MWE such as 'cut the mustard' (to suffice/to be good enough) (borrowing the example from Findlay (2019)). A non-native speaker, who is familiar with the meaning of 'cut', 'mustard' and 'the' individually, would not be able to capture the meaning of the whole through the regular composition of words. As such, the MWE should form its own unique lexical entry (i.e. treated as a listeme). However, looking at Example 1 and 2, we can see that the items in this expression retain some of their original properties, the verb must inflect separately from the rest of the expression.

(1) That **cuts the mustard**.

(2) *That **cut the mustard**s.

This interplay between lexical item and syntactic construction is an important facet of MWEs, and is explored in some detail below, as one of the definitive properties and challenges exhibited by such linguistic phenomena.

2.2 Properties of MWEs

Just as the precise definition of MWEs can vary, so too can the characteristics that MWEs present. Many taxonomies in the literature make use of the various properties typical of MWEs to identify and organise them into classes. Rather than presenting an exhaustive taxonomy of MWEs, this section is intended as a brief summary of the various properties and classification characteristics that are common in discussion of MWEs for NLP. These properties are directly applicable to our own research, including the categorisation scheme outlined in Chapter 4.

2.2.1 Idiomaticity

We begin with a discussion of idiomaticity, perhaps one of the most commonly recognised properties that MWEs can display, and indeed included as a definitive feature of MWEs in several of the definitions presented above. Baldwin and Kim (2010) describe idiomaticity in the case of MWEs as being "a markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels". Table 2.1, adapted from a similar figure in their discussion, gives examples of some MWEs in English, marked with levels of idiosyncrasy displayed by each.

	Lexical	Syntactic	Semantic	Pragmatic	Statistic
Ad hoc	+	-	-	-	+
By and large	-	+	-	-	+
Be in touch	-	-	+	-	+
Throw in the towel	-	-	+	-	+
All aboard	-	-	-	+	+
Salt and pepper	-	-	-	-	+

Table 2.1: Levels of idiomaticity displayed by different examples of MWEs

Semantic idiomaticity, also known as **non-compositionality**, is the property whereby the meaning of the MWE as a whole cannot be determined compositionally from the individual words. This feature is commonly ascribed to idioms, a large subset of MWEs discussed in the literature. As with many of the other levels of idiomaticity, this ranges along a continuum of compositionality. MWEs can be completely compositional, such as *bus stop*, or completely opaque, such as *shoot the breeze* (to talk aimlessly of something), where the meaning is not easily derived from the words as they appear. Between these two opposing ends, there lie semi-compositional MWEs, whose meaning is partially composable, such as *to blow hot or cold*, meaning to be easily changeable on a subject (blow → *change*). Alternatively, the meaning of the expression may be possible to derive through *figuration*, or assigning a metaphorical or figurative meaning to the component words, such as with *spill the beans* meaning to tell a secret (tell → *spill*, secret → *beans*). Of note, Sag et al. (2002) remark that non-decomposable idioms are not subject to syntactic variability, such as internal modification (**shoot the coolest breeze*) or passivization (**the bucket was kicked*). Baldwin and Kim (2010) note a correlation observed by Keysar and Bly (1995) between the usage of an MWE and the level of semantic idiomaticity displayed, with higher-usage MWEs being perceived as more readily decomposable.

Syntactic idiomaticity occurs in MWEs where the syntactic label of the construction does not derive from the syntactic form of the components. Fillmore et al. (1988) refer to MWEs with this property as **extragrammatical idioms**, saying such constructions are unpredictable and may seem 'weird' to non-native speakers. This idiomaticity can be external, as with *kingdom come* (noun+verb), which appears to act as the NP object of a preposition (see Example 3).

(3) The bomb exploded, blowing the room to kingdom come.

Internal statistic idiomaticity demonstrates syntactic combinations that would not normally be considered grammatical, such as with *by and large*, which coordi-

nates a preposition and an adjective.

Lexical idiomaticity, a characteristic also referred to as **cranberry words** (Aronoff, 1976), describes MWEs which contain words that are limited to their usage within that expression, and never occur in the lexicon outside of the MWE. For example, the word 'amok' occurs almost exclusively within the construction *run amok*. This can also include MWE phrases which are borrowed wholesale from another language, as with *in situ*, although this is a case complicated by the complexity of phrasal borrows and code-switching, and such MWEs are perhaps best treated as single words, despite their orthography.

This type of idiomaticity inevitably results in semantic and syntactic idiomaticity, as there is no lexical knowledge associated with these 'words', making it one of the most clear-cut indicators of an MWE.

Statistical idiomaticity refers to words that occur in combination with a marked statistical frequency. As an illustration of statistical idiomaticity, we present a number of examples in Table 2.2, adapted from Baldwin and Kim (2010). The near-synonym adjectives presented (*flawless*, *immaculate*, *impeccable* and *spotless*) each show a preference for which nouns they pre-modify. The table indicates this preference with markers ('+' for positive lexical preference, '?' for neutral lexical preference and '-' for negative lexical preference). Of course, like other forms of idiomaticity, the lexical preference displayed by certain combinations lie on a spectrum.

Another example of statistical idiomaticity exists in binomials such as *black and white*, as in 'black and white photography', or *chicken and egg*, as in 'chicken and egg situation' (Benor and Levy, 2006). In both of these cases, reversing the order does not capture the same semantic meanings (*?white and black television*).

Sag et al. (2002) refer to MWEs demonstrating this property as **institutionalised phrases**, differentiating between **collocations**, which are simply statistically significant co-occurrences of words, and do not necessarily present idiomaticity. For

	flawless	immaculate	impeccable	spotless
condition	+	+	+	+
credentials	-	-	+	-
hair	-	+	?	-
house	?	+	?	+
logic	+	-	+	-
timing	?	-	+	-

Table 2.2: Statistical idiomaticity in the form of lexical preference displayed by a cluster of near synonyms and the nouns they pre-modify. Table slightly modified from similar table presented by Baldwin and Kim (2010), in turn adapted from Cruse (1986).

example, the words *sell* and *house* are considered collocations, given how the words co-occur with predictable frequency as a result of non-linguistic factors.

Pragmatic idiomaticity is captured in certain MWEs whose usage is associated with a particular context, such as with *good morning*, a greeting associated with that time of the day, and *all aboard*, a command associated with the imminent departure of a ship or train, and limited in usage to a train station or dock.

2.2.2 Variability

MWEs allow for varying degrees of flexibility in their formation, which poses a challenge in their processing (see Section 2.3.1). This variability can occur on the morphosyntactic or the syntactic level. For instance, words in the expression may be permitted to inflect for tense or person, as with Example 4, but other words must not be modified, as with Example 5. Syntactically, MWEs may permit internal modification, as with Example 6, or passivisation, as with Example 7.

(4) They all **kicked the bucket** last week.

(5) *It's awful to see them **kick the buckets** like that.

(6) This morning my coffee maker **kicked the** proverbial **bucket**, and I had to replace it.

(7) A **deal** was **struck** by members of the parliament.

This range of variability can be classified into three degrees of flexibility, as per Sag et al. (2002): fixed expression, semi-fixed expressions and flexible expressions. Of note, their descriptions only apply to **lexicalised phrases**, i.e. MWEs that are idiomatic on a lexical, semantic, pragmatic or syntactic level. Given that institutionalised phrases are marked due to their frequency of co-occurrence, they are as syntactically flexible as any non-MWE construction would be, so are not considered.

Syntactically fixed expressions do not undergo any modification or inflection (e.g. **by and larger*), and do not allow for any additional lexical components (e.g. **day by every day*). This is the general rule, though there may be exceptions where an additional lexical item is inserted for emphasis or to convey an ironic or other meaning, as with Example 6 above.

Semi-flexible expressions allow for morphosyntactic variation, such as inflection; however not all regular inflection is necessarily permitted (e.g. compare 'one hot dog' or 'thirty hot dog*s*', and '*hot*ter* dogs' or '*a hot *and tasty* dog'). This class includes compound nouns and named entities (e.g. *part of speech*, *Christmas Day*), as well as non-decomposable idioms (e.g. *shoot the breeze*), with the semi-flexible nature of such expressions attributed to their semantic opacity preventing the full syntactic flexibility of decomposable idioms. This characteristic of decomposability is further explored in Section 2.3.1.

Syntactically flexible expressions describe expressions containing several lexicalised components, but can exhibit the full extent of morphosyntactic variability and much more syntactic freedom. Expressions can show variable word order or include gaps (e.g. '**Check out** a book' vs. '**Check** *the book* **out**'), can be subjected to passivisation (e.g. '**strings** were **pulled** for the guest lecturer'), and the components may be inflected or elements may be inserted (e.g. 'He/they **gave**/*gives* a fascinating **lecture**'). Many verbal multiword expressions are flexible, as verbal components have a wide range of inflected forms, objects or modifiers may be

inserted, or constituents may be shifted around.

Another framework for capturing this variability is proposed by Fraser (1970), who presents six categories of MWEs, based on which family of syntactic behaviours can be exhibited by them. The categories form a hierarchy, from syntactically fixed to the most syntactically flexible, such that an MWE that displays a higher level of syntactic behaviour must also obey the patterns of behaviour exhibited by lower-ranking MWEs. This framework is intended for use within Transformational Grammar (Chomsky, 1965), and as such, we will not be applying it to our research.

2.2.3 Domain-specificity

Another property of MWEs, related to idiomaticity, is the domain-specificity of many MWEs. As illustrated by Smadja (1993), MWEs such as *The Dow Jones*, and its related variations of form would be difficult for a non-specialist to grasp and appropriately apply (see Example 8, and the ungrammaticality of Examples 9 and 10).

(8) **The Dow industrials** rose 26.28 points on Tuesday

(9) ***The Dow industrial** rose 26.28 points on Tuesday

(10) ***The Jones industrials** rose 26.28 points on Tuesday

These specialist domains, or **sublanguages**, are often densely populated with *terminology*, defined by Baldwin (2004)[7] as "a lexical unit consisting of one or more words which represents a concept inside a domain". While Sag et al. (2002) note that terminology overwhelming consists of MWEs, Baldwin and Kim (2010) note the distinction between research on MWEs and terminology in NLP, with the latter being both broader and narrower in scope, including simple lexemes where such are technical terms, but paying no attention to non-technical MWEs.

[7] as cited in Lyse and Andersen (2012).

To differentiate between terminology and non-technical MWEs, Savary et al. (2019b) propose the meta-categories of *sublanguage MWEs*, whose form-meaning association is usually determined by experts in a domain, and *general language MWEs*, which are coined by a much larger community of speakers via informal processes, such as institutionalisation. This property of MWEs, and the characteristics of sublanguage and general language MWEs are explored further particularly in Chapter 4, and also Chapter 5.

2.2.4 Morphosyntactic Classes

A property of MWEs that increases their challenge for processing is their **heterogeneity**, particularly on the level of morphosyntactic classes. Ramisch (2015) broadly categorises these classes as **nominal expressions**, **verbal expressions** and **adverbial and adjectival expressions**, although this classification does not cover all types of MWEs. Oftentimes, MWE classification is driven at least in part by their morphosyntactic class (e.g. Sag et al. (2002); Calzolari et al. (2002a); Baldwin and Kim (2010); Parra Escartín et al. (2018); Savary et al. (2017)), as this property is relatively trivial to determine automatically when compared with idiomaticity, variability or domain-specificity. There are also notable patterns in MWEs of different morphosyntactic classes, and the behaviours they display.

In our categorisation of Irish MWEs, we employ morphosyntactic classes as a means of distinguishing between MWE types. This is explored in Chapter 4.

2.3 MWEs in NLP

Within the field of NLP, MWEs have been recognised as a "pain in the neck" for at least the past twenty years (Sag et al., 2002). There have been enormous strides made in the field since then through the development of computational models, with improved stochastic methods, neural networks, and more recently, pre-trained language models compensating for—and in some cases rendering unnecessary—precise

linguistic analysis and hand-crafted language resources. However, the challenges posed by MWEs can not be so readily solved by computation power and effective architecture alone (Yu and Ettinger, 2020; Tayyar Madabushi et al., 2021).

In this section, we highlight the challenges posed by MWEs, based on the variable properties that they can display. We also describe the tasks inherent to the topic of automatic processing of MWEs for NLP, including how these tasks are impacted by the aforementioned challenges, and how this body of research tackles these tasks.

2.3.1 Challenges

Variability is the tendency for MWEs to vary in the degree of fixedness demonstrated (as discussed in Section 2.2.2). This is challenging for their processing as inflected forms (e.g. 'cat nap*s*', 'cut*ting* it close'), and MWEs with non-fixed word order (e.g. '**look up** the number' vs '**look** the number **up**') prevent a simple words-with-spaces treatment for the annotation of such MWEs. Furthermore, information regarding the degree of flexibility and the range of acceptable variations exhibited by each MWE must be included in the description of said MWE. Included in this are related properties such as **overlapping** or **nested** MWEs, as with Examples 11 (where the MWE 'make [a] decision' and 'make [a] change' are overlapping, with the same token 'make' applying to both MWEs) and 12 (where the MWE 'multiword expression' is nested inside the MWE 'look up'). Example 12 also illustrates the property of **discontiguity**, as the MWE 'look up' can include non-lexicalised components, such as 'the number', or 'the multiword expression', which can intervene within the MWE components. This property of variability adds to the difficulty of deciding which parts of the expression should be lexicalised.

(11) 'He tends to **make** the **decisions** and the **changes** around here'

(12) 'I **looked** the multiword expression **up**'

Productivity is related to the problem of variability, as some MWEs are formulaic in their composition, to varying degrees. Lexicalised components may be selected

from a syntactic or semantic class, as with the light verb 'take' in many light verb constructions ('take a walk/look/chance/note/message') or with the completive 'up' in verb particle constructions ('eat/finish/rest/fill/clean up'). As productive constructions such as these can combine to form potentially inexhaustive numbers of different MWEs, listing every MWE is not an effective, or indeed possible, approach for their processing.

Non-compositionality (also called **decomposability**) refers to the property of semantic idiomaticity exhibited by so many MWEs. This poses a challenge to the processing of MWEs for any task requiring semantic understanding or interpretation. For instance, the non-compositional MWE 'to cut a rug' in the sense of 'to dance well/enthusiastically' is semantically opaque from the meaning of each of the component words, and an attempt to generate this meaning from general language composition will fail. Overcoming this challenge requires both identifying the precise components of the MWE, and ascribing the correct meaning to the whole.

Language specificity in MWEs becomes a challenge in multilingual settings, as translating MWEs from one language to another in a word-by-word fashion often results in clunky, unnatural or meaningless translations. For example, the English MWE 'raining cats and dogs' directly translated into Irish (*ag cur cat agus madraí*) does not carry the idiomatic sense of the English expression. Likewise, the Irish MWE *ag cur sceana gréasaí* when directly translated into English ('raining cobbler's knives') sounds unnatural and lacks the idiomatic meaning of the original MWE. This challenge is closely related to the property of non-compositionality.

Ambiguity in MWEs generally refers to distinguishing between literal and idiomatic instances of MWEs. For instance, with the potential MWE *by the way*, it can be read as literal (Example 13) or idiomatic (Example 14).

(13) Anna was impressed **by the way** her sister bragged effortlessly.

(14) I saw that jacket you wanted, **by the way**, and it is expensive.

This challenging property of MWEs can affect how they are handled in running text, as not all instances of the same MWE components should be considered idiomatic. Syntactic processing (e.g. parsing) can aid with this ambiguity problem, as can the semantic analysis of the context in which such MWEs appear (e.g. text summarisation).

2.3.2 Automatic Processing of MWEs for NLP

In a comprehensive survey of this topic, Constant et al. (2017) discuss the task of automatic processing of MWEs, focusing on **discovery** and **identification** as the two main subtasks of this larger problem.

Discovery is concerned with finding **new** MWEs or types in text corpora, usually for the purposes of storing them for future use. The literature on this topic tends to focus on empirical strategies for discovery, i.e. automatically learning patterns of lexical information from textual data (Pearce, 2002; Villavicencio et al., 2007; Ramisch et al., 2008), over supervised approaches that require annotated data (Jacquemin, 1996; Farahmand and Martins, 2014). However, both these strategies are important and complementary to tackling the task of discovery, and can be combined in the creation of lexical resources (Rondon et al., 2015). The challenges posed by variability and discontiguity in particular increase the complexity of this task. In this thesis, we have not focused particularly on this task, but we report some exploratory experiments in MWE discovery in Appendix B.

Identification is the task of automatically finding MWEs in running text by associating them with known MWEs, at the token level. This task is often incorporated as a pre-processing step in some NLP applications. MWE tagging can help to reduce ambiguity for parsers or MT systems, and aids in other tasks such as semantic processing. Just as with MWE discovery, variability, including overlapping, nesting

and discontiguity increases the complexity of this task, along with the challenges provided by ambiguity and productivity. We investigate this task in greater detail in Chapters 7 and 8.

In addition to these two tasks highlighted, there are a number of related tasks within the field of MWE processing, some of which are addressed in Villavicencio et al. (2005) and Baldwin and Kim (2010). In our thesis work, we focus on the tasks of: **linguistic descriptions**, **lexical resources**, **applications**, and **evaluation**.

Linguistic descriptions of MWEs comprise much of the theoretical work on this topic. While many such studies on MWEs have been carried out from the lens of theoretical linguistics (Fraser, 1970; Bolinger, 1971; Fernando and Flavell, 1981) or corpus linguistics (Moon, 1998; Fellbaum, 2009) or from the perspective of psycholinguistics (Bobrow and Bell, 1973; Bhattasali et al., 2018), treatments of MWEs in NLP benefit from focused research that highlights the issues particular to that field. High-level overviews of the problem of processing MWEs for NLP, such as Sag et al. (2002), Baldwin and Kim (2010), Ramisch (2015), and Constant et al. (2017) serve to provide perspective on the major challenges in the field, as well as common methodologies and practises adopted to overcome these challenges. Linguistic taxonomies of various MWE types and behaviours are also offered, in these works and others (Moszczyński, 2007; Laporte, 2018), with some authors focusing on taxonomies for specific languages such as Spanish (Parra Escartín et al., 2018), Basque (Gurrutxaga and Alegria, 2013), Arabic (Hawwari et al., 2014), Croatian (Buljan and Šnajder, 2017), Hebrew (Al-Haj et al., 2013), and Japanese (Tanabe et al., 2014).[8] In this thesis, we benefit from the linguistic analysis performed on Irish and Irish MWEs when conducting our research, as demonstrated in Chapters 3 and 4.

[8]This linguistic analysis is often applied to the building of lexical resources, or other tasks in MWE processing.

Lexical resources are extremely useful for the accurate processing of MWEs. The development of such resources requires the extraction and sometimes processing of MWEs from textual data, with much of this work being the collection of terminology resources (Villavicencio et al., 2005; Lyse and Andersen, 2012; Logar et al., 2014), or lexicons (Kuiper et al., 2003; Bouamor et al., 2012a; Tanabe et al., 2014). Research on lexical resources has also been expanded through detailing methodologies and best practises for capturing MWEs across many levels of idiosyncrasy, with all the levels of variability that can be exhibited. Calzolari et al. (2002a); Copestake et al. (2002) and Villavicencio et al. (2004) provide recommendations for capturing MWEs and storing them in lexical resources. MWE properties such as variability, discontiguity, ambiguity and productivity all contribute to the difficulty of this task. We describe some of these resources in Section 2.5.1 below. In Chapter 5, we describe the development of several MWE-aware lexical resources for Irish.

Applications encompasses any NLP applications that can benefit from the addition of MWE information, that is, any application that requires some level of syntactic or semantic analysis. There have been many efforts in applying MWE-specific knowledge into tasks such as parsing (Waszczuk et al., 2017), machine translation (Barreiro et al., 2013), information retrieval (Acosta et al., 2011), question answering (Dowdall et al., 2003), and sentiment analysis (Williams et al., 2015), to name a few. Many MWE properties (e.g. semantic idiomaticity, variability, ambiguity, discontinuity, and language specificity) contribute to the complexity of processing MWEs for NLP applications, however some tasks are particularly impacted by certain properties above others (e.g. language specificity particularly poses a challenge for machine translation). Methodologies for incorporating MWEs in each of these tasks differ in whether MWE information is included in the pre-processing stage, the post-processing stage, or incorporated concurrent to the NLP task. In Chapter 6, we investigate the addition of MWEs into GA↔EN neural machine translation systems, namely by labelling MWEs as linguistic features in the training data.

Evaluation of the correct handling of MWEs can be considered a task in itself, whether it be measuring the success (i.e. precision and recall) of an MWE identification task, or measuring the accuracy of MWE translations in a machine translation system. Each task requires an evaluation process tailored to the specific requirements of the task, but usually a suitable lexical resource is required, whether that be gold-standard annotated corpora, lexicons of MWEs, or otherwise. The challenges inherent to this task are explored in greater detail in the rest of this thesis, as evaluation can be considered an important subtask for each of the other tasks we address.

2.4 Multilingual Frameworks for Capturing Linguistic Features

For low-resource languages, availing of multilingual methodologies and resources can be effective in combating the challenges posed by scarce language-specific resources. We describe here two initiatives that have developed a consistent cross-lingual approach to linguistic features, and from which we draw heavily in regards to our categorisation of MWEs in Irish (see Chapter 4).

2.4.1 PARSEME

The PARSEME (PARsing and Multiword Expressions) network arose from an IC1207 COST Action devoted to improving treatment of MWEs in parsing and linguistic resources (Savary et al., 2015). Their aims were (i) to focus on multilingualism in linguistic and technological studies, (ii) establish a cross-lingual research network in NLP, and (iii) bridge the gap between linguistic precision and computational efficiency in NLP applications.

There have been several outcomes of this initiative, including various surveys, papers and presentations. Two important outcomes are relevant to our work:

the **PARSEME Shared Tasks on the Automatic Identification of Verbal MWEs**, and the PARSEME ANNOTATION GUIDELINES.

2.4.1.1 PARSEME Shared Tasks on the Automatic Identification of Verbal MWEs

To date, PARSEME have organised three shared tasks focused on the automatic identification of verbal MWEs (vMWEs). Verbal MWEs were chosen as the focus of the shared tasks, as at that point, they were largely underrepresented in the research field. Savary et al. (2017) also highlight five properties of vMWEs that make them uniquely challenging in the task of MWE identification. Those properties being:

1. **Discontinuity** and **variable order** of the lexicalised elements
2. Lexicalised elements can occur in **idiomatic** or **literal** expressions
3. vMWEs may present **syntactically ambiguous** surface forms
4. vMWEs of different categories may present with the **same syntactic structure** or may select the **same lexical items**
5. vMWEs in different languages may be **treated differently** at the linguistic level

The first edition—Edition 1.0—included 18 languages and focused on addressing the above challenges presented by identifying vMWEs (Savary et al., 2017). The second edition—Edition 1.1—included 17 languages and built on the first task through the introduction of new evaluation settings (Ramisch et al., 2018).

The latest edition of this task—Edition 1.2—included 14 languages (Ramisch et al., 2020). This edition took into account the lessons learned from previous editions, as well as contemporary research efforts. One such development was the decreased emphasis on the task of disambiguating between literal and idiomatic occurrences of MWEs, given the infrequency that the literal form of these ambiguous

MWEs occurred. In their analysis of idiomatic and literal occurrences of MWEs (Examples 15 and 16), Savary et al. (2019a) found that literal occurrences such as Example 16, while grammatically correct, are nonetheless limited in language data, as speakers tend to avoid using them.

(15) Our waiter **pulled strings** to get us special treatment.

(16) I **pulled strings** to control the flight of the kite.

Other considerations include the Zipfian distribution of MWEs (few MWEs occur frequently in text, with a long tail of infrequently occurring MWEs), and the tendency for verbal MWEs to present idiosyncrasies at the level of types rather than tokens. Given these tendencies, it is difficult to identify new MWEs based on previously seen examples (Savary et al., 2019b). As such, the ability to generalise MWE identification based on supervised data is in practise applicable only to identifying new variations on previously seen MWEs. With these principles in mind, this edition of the PARSEME task focused on the identification of *unseen vMWEs*, as one of the most difficult tasks in MWE processing.

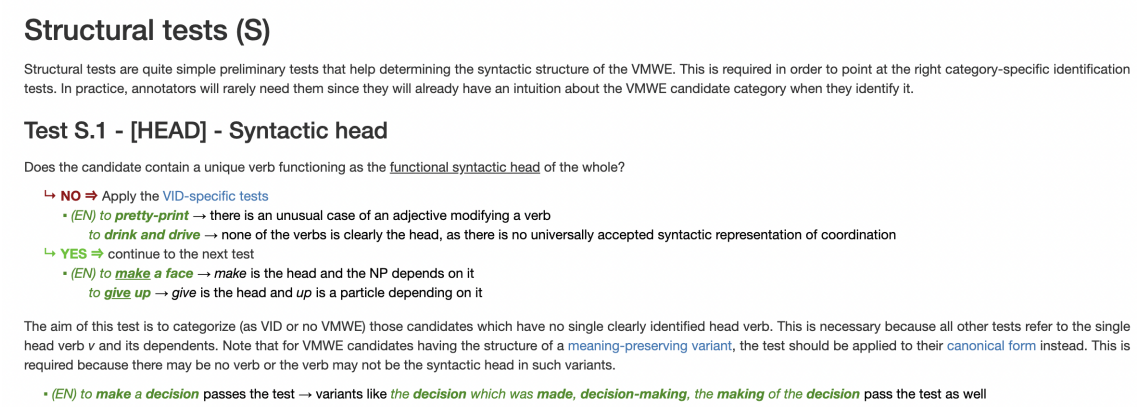
The results of this latest edition are explored in Chapter 7, with particular attention payed to the inclusion of Irish in this task.

2.4.2 PARSEME Annotation Guidelines

The annotation guidelines[9] developed as part of this shared task are currently available for 27 languages. The inclusion of a wide range of diverse languages allows for insights into the particular challenges inherent to the task of MWE identification that exist universally across the set of languages, as well as providing information on those challenges particular to individual languages themselves.

The annotation guidelines offer hierarchical structural and classification tests in the form of a decision tree (see a screenshot example in Figure 2.1). These tests confirm the candidates status as an MWE and assign it a class. Some language-

[9]Available at `https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/`



Structural tests (S)

Structural tests are quite simple preliminary tests that help determining the syntactic structure of the VMWE. This is required in order to point at the right category-specific identification tests. In practice, annotators will rarely need them since they will already have an intuition about the VMWE candidate category when they identify it.

Test S.1 - [HEAD] - Syntactic head

Does the candidate contain a unique verb functioning as the functional syntactic head of the whole?

↳ **NO ⇒** Apply the VID-specific tests
- *(EN) to **pretty-print*** → there is an unusual case of an adjective modifying a verb
  *to **drink and drive*** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination

↳ **YES ⇒** continue to the next test
- *(EN) to **make a face*** → *make* is the head and the NP depends on it
  *to **give up*** → *give* is the head and *up* is a particle depending on it

The aim of this test is to categorize (as VID or no VMWE) those candidates which have no single clearly identified head verb. This is necessary because all other tests refer to the single head verb *v* and its dependents. Note that for VMWE candidates having the structure of a meaning-preserving variant, the test should be applied to their canonical form instead. This is required because there may be no verb or the verb may not be the syntactic head in such variants.

- *(EN) to **make** a **decision*** passes the test → variants like *the **decision** which was **made, decision-making**, the **making** of the **decision*** pass the test as well

Figure 2.1: Screenshot of the structural tests used in the PARSEME annotation guidelines.

```
# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_Irish-IDT/ga_idt-ud-dev 805
# text = D'éirigh thar cionn léi.
1       D'      do      PART    Vb      PartType=Vb     2       mark:prt        _       SpaceAfter=No   *
2       éirigh  éirigh  VERB    VI      Form=Len|Mood=Ind|Tense=Past    0       root    _       _       1:VERB|MWE-IAV|_
3       thar    thar    ADP     Simp    _       4       case    _       _       *
4       cionn   ceann   NOUN    Noun    Case=Dat|Definite=Ind|Gender=Fem|Number=Sing    2       obl     _       _       *
5       léi     le      ADP     Prep    Gender=Fem|Number=Sing|Person=3 2       obl:prep        _       SpaceAfter=No   1
6       .       .       PUNCT   .       _       2       punct   _       _       *
```

Figure 2.2: Sample of PARSEME annotated data in CUPT format. The MWE *d'éirigh léi* 'she succeeded' is annotated here.

specific tests allow for determining certain language-specific categories (e.g. the test for Inherently Clitic Verbs in Italian), but the majority of the tests are cross-lingual. Figure 2.2 depicts a sample of the annotated data, as it appears in CUPT format.[10]

2.4.3 Universal Dependencies

The Universal Dependencies (UD) project[11] (Nivre et al., 2016; de Marneffe et al., 2021) is a framework to facilitate and improve cross-lingual learning and multilingual parsing. The project has grown substantially from the initial v.1 release in 2015, with new language treebanks and enhancements to the existing treebanks added with each six month release. The most recent edition v2.10 contains 228 treebanks for 130 languages.[12]

The basic tenants of UD are offering linguistic representation useful for mor-

[10] See Chapter 7 for a description of CUPT format.

[11] `https://universaldependencies.org/`

[12] This version was released on May 15, 2022, with the next release of v2.11 scheduled for November 15, 2022.

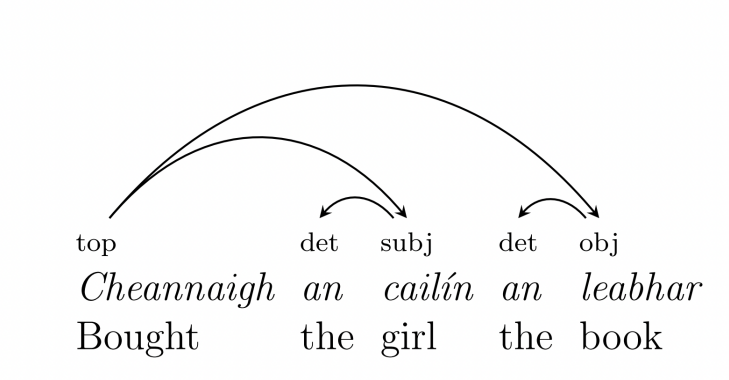


Figure 2.3: Example of a dependency tree represenation of the sentence *Cheannaigh an cailín an leabhar* 'The girl bought the book'. Example taken from Lynn (2016).

phosyntactic research, semantic interpretation, and practical cross-lingual NLP. To that end, simple surface representations are encouraged in order to maximise parallelisation between constructions across a wide range of languages. **Dependency grammar** (Tesnière, 1959) is used to capture the relationship between heads and dependents, with regard to the role played by different words in the sentence. Figure 2.3 displays an example of a **dependency tree** for the Irish sentence *Cheannaigh an cailín an leabhar* 'The girl bought the book'.

The UD framework is applied to dependency-parsed treebanks through the use of multilingual annotation guidelines. These guidelines are based on an evolution of the Stanford dependencies (de Marneffe et al., 2014), Google universal POS tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). Treebanks are described further in Section 2.5.1.3 below, and form a valuable linguistic resource for MWE processing. We also draw from the UD treatment of MWEs in our research, as discussed in Chapter 4.

2.5 Resources

Language resources come in many forms, and in a variety of structures, complexities, and sizes. They are necessary for many communities, including linguists, engineers, teachers, and language speakers, and naturally, those working in the field of NLP. Bird and Simons (2003) note that these resources are primarily a source of either

data, **tools** or **advice**. Data comprises any recorded information documenting or describing language. Tools encompass a variety of computational assets, including software, stylesheets, and even document type definitions. Advice includes descriptions or information of available resources, regarding usage of tools, and guidelines or best practises for creating new data, as per Bird and Simons (2002) and Calzolari et al. (2012).

There have been a number of initiatives set up in recent decades to assist with the development, distribution and organisation of Language Resources (LRs). The European Language Resource Association or ELRA[13] was established in 1995 with the goal of promoting LRs in the field of Human Language Technology (HLT). They accomplish this through the services of identification, collection, distribution, production, validation and standardisation of LRs, evaluation of systems, tools and products that relate to LRs, and promoting initiatives which accomplish any of these services. The biannual Language Resource and Evaluation Conference (LREC) was established for the purpose of fulfilling this latter service, with published conference proceedings promoting academic research in this field. Tasks such as distribution and evaluation are fulfilled through the Evaluations and Language resources Distribution Agency (ELDA), which was created as an operating body within ELRA to handle issues relating to LR distribution. One such contribution by this body was the introduction of an identification number system for LRs, the International Standard Language Resource Number[14] (ISLRN), which has been widely adopted to date (Mapelli et al., 2016). In addition to their catalogue of LRs (ELRA Catalogue[15]), they provide an R&D Catalogue dedicated to academic research, and a Universal Catalogue dedicated to LRs sourced from around the world. For the sharing and dissemination of LRs, the META-SHARE sharing and exchange network[16] was implemented in the META-NET Network of Excellence framework. For

[13] `http://www.elra.info/en/`
[14] `http://www.islrn.org/`
[15] `http://www.elra.info/en/catalogues/catalogue-language-resources/`
[16] `http://www.meta-share.org/`

the documentation and organisation of LRs, the LRE Map was created as a means of enhancing availability of information about new or existing LRs, and to enforce standards of documentation in the community (Calzolari et al., 2012).

Yet another initiative is the European Language Resource Coordination (ELRC). Their mission is to enable the sharing and submission of LRs to the European Commission for the purposes of improving their eTranslation platform, and they achieve this through development of central repositories for these LRs, such as their ELRC-SHARE repository.[17] The action provides support to public administrations, including services such as data cleaning, alignment, conversion, anonymisation, reformatting, among others. The European Language Resource Infrastructure (ELRI) project, which ran from October 2017 to October 2019, worked with bodies within each country to coordinate the submission of relevant LRs through the establishment of National Relay Stations (NRS). Dublin City University hosts the NRS for Ireland, and oversees the submission of many LRs.

Outside of European initiatives, there is the Linguistic Data Consortium (LDC), a US-based consortium of universities and research bodies dedicated to the collection, creation and distribution of LRs for speech and language technology. The LDC Catalog now contains hundreds of holdings for a range of language data, along with a metadata schema for categorizing these data (Jaquette et al., 2020). In addition to the catalog, a selection of open source tools are also supplied, alongside publications provided by LDC staff on related topics of research.

2.5.1 Resources for MWEs

Data resources make up a considerable portion of the resources for MWEs. In the domain of MWEs and their processing, three main domains of data resources are considered according to Constant et al. (2017). Those are: lexical resources, corpora, and treebanks. We explore each of these in turn. We also explore the uses of these lexical resources, including their use in evaluation, both intrinsic and extrinsic. For

[17] `https://elrc-share.eu/`

intrinsic evaluation, the resource is used as a gold standard, to be compared to the output of a model (e.g. predicting MWE types). For extrinsic evaluation, the output of the system is incorporated into another task, and the evaluation of that task's performance gives some indication of the effectiveness of the MWE input. Resources can be integrated into this pipeline also.

2.5.1.1 Lexical Resources

Lexical resources cover such items as dictionaries, lexical lists, lexicons and wordnets: any database that includes MWEs or MWE material. Lexical resources like these make up a substantial proportion of MWE-related resources.

There are a variety of ways that MWEs can be stored in lexicons, depending on the complexity and richness of information that is to be captured, and the intended uses of the lexicon. Factors such as the flexibility that MWEs can present, along with the complexity of interplay between lexicalised[18] and non-lexicalised elements, can make a thorough representation of MWEs in lexicons challenging. Calzolari et al. (2002a) address this need for a framework for representing MWEs in lexicons, focusing on MWEs that display productivity and regularities that can be generalized to classes of words with similar properties. In other words, they are concerned with MWEs that lie at the intersection of grammar and lexicon. They focus on two MWE types (support verbs/light verb constructions and noun compounds). Both of these types exhibit productivity and a regular syntactic pattern, along with a variability in lexicalisation and language-dependent variation, making them difficult to capture in multilingual lexicons. The authors explore both FrameNet (Fillmore, 2001) and SIMPLE (Bel et al., 2000) as frameworks for modelling these types of MWEs.

Lexical lists are a simpler method of capturing MWEs, with one such resource being the English Multiword Expression Lexicons, a collection of 9 lexicons compiled from various sources, and available for download online[19] under CC

[18]Lexicalised elements of a multiword expression are the compulsory arguments that are always present in the MWE.

[19]`https://www.cs.cmu.edu/~ark/LexSem/`

BY-SA 4.0 license. The lexicons include MWE entries extracted from English WordNet, Multiword chunks from SemCor (Miller et al., 1993), MWE entries from English Wikitionary,[20] the WikiMwe dataset mined from Wikipedia (Hartmann et al., 2012), named entities and other MWEs from the English side of the Prague Czech-English Dependency Treebank (Hajič et al., 2012), a dataset of verb-particle constructions by Baldwin (2008), a list of light-verb constructions provided by researcher Claire Bonial, idioms from Phrase.net (now Phrases.com),[21] and MWE entries from Oyz's compilation of dictionary entries for frequent English verbs.[22] Also included is a script for extracting MWEs from the SAID database of phrasal lexical idioms (Kuiper et al., 2003). Resources are in `json` format, and the entries vary from simply listing the MWEs in their canonical form, to providing extra information such as POS-tags, PMI scores, frequency counts, context examples, semantic relations, and MWE labels. The combined lexicons include light verb constructions, verb-particle constructions, idioms, inherently adpositional verbs, noun compounds, named entities, fixed expressions and others. These lexical resources were used to generate features in sequence-tagging MWE identification experiments (Schneider et al., 2014a). We also employ them in our research in incorporating MWEs in MT systems, as described in Chapters 5 and 6.

WordNet (Fellbaum, 1998) is a lexical resource of English words grouped into sets of cognitive synonyms called synsets, creating a network of conceptual-semantic and lexical relations. The resulting dataset resembles a thesaurus, except that the links are between word senses rather than the word strings, and the semantic links are labelled. The main relation between words are synonyms, i.e. words that have the same or similar meaning (e.g. 'car' and 'automobile'). These synonyms are collected into synsets, of which WordNet has 117,000. Synsets are related to other synsets through "conceptual relations", which are a limited set of relations that distinguish the role and relationship of words to one another (e.g. the super-

[20] `https://en.wiktionary.org`
[21] `https://www.phrases.com/`
[22] `http://home.postech.ac.kr/~oyz/doc/idiom.html` URL now defunct.

subordinate (hyperonymy) relation between 'furniture' and 'bed', or the part-whole relation (meronymy) between 'chair' and 'backrest'). Most of the relations connect words within the same POS group, but some cross-POS relations allow for semantically similar words with a shared stem (e.g. observe (verb), observant (adjective) observation, observatory (nouns)). There is a natural synergy between MWEs and WordNet entries, particularly for MWEs with an idiomatic semantic property. This natural synergy was explored in the Joint Workshop on Multiword Expressions and WordNet (Savary et al., 2019c). Topics included were the comparison of verbal MWEs across two language's wordnets (Barbu Mititelu et al., 2019), an analysis of the distribution and treatment of MWEs in the Bulgarian WordNet (Laskova et al., 2019), and enhancing MWEs from German WordNet with morphological information (Declerck et al., 2019). This last paper follows the example of others (McCrae et al., 2014) in linking WordNets with Lemon (LExicon Model for Ontologies).

Another method of representation is using the Head-driven Phrase Structure Grammar (HPSG) framework, as described in Sag et al. (2002). In this framework, MWEs are described in terms of their orthography and grammatical information such as semantic roles and morphosyntactic information. The LinGO English Resource Grammar (ERG) framework, described by Copestake and Flickinger (2000), is a broad coverage grammar of English with a lexical database structure. Copestake et al. (2002) describe the treatment of three varieties of MWEs in this framework: verb-particle constructions, compound nouns and idioms. This description is further developed by Villavicencio et al. (2004), taking into account the flexibility they can exhibit. While this framework allows for a richness and precision in MWE representation, and a database using this representation allows for integration into a grammar, the construction of such a lexicon is time-consuming, requiring expertise in linguistic analysis of MWEs. Depending on the task, simple lexical lists or similar structures, which can be automatically or semi-automatically generated, may be sufficiently rich resources.

MWE lexical resources play an important role in a myriad of NLP tasks and applications. For example, they can be used in MT during the pre-processing, post-processing, or translation steps. They can be used to delimit MWEs in the data, or replace them with equivalent single tokens or a translation equivalent before alignment takes place, or they can be used to annotate sense identifiers in the text. Bilingual lexicons can be used to improve Statistical Machine Translation systems, as demonstrated by Bouamor et al. (2012a) whereby adding a feature field to the phrase tables was found to improve the translation quality of their test sets. Lexicons are also singled out as important for the task of MWE identification by Savary et al. (2019b), who remark on the vitality of lexicons for a variety of identification methods, given the nature of MWEs and the challenges inherent in processing them, e.g. the issue of *unseen* MWEs. This topic is further explored in Chapter 7.

2.5.1.2 MWE Corpora

MWE corpora consist of bodies of text that have been annotated in some way, at minimum to indicate the presence of an MWE, or further information, such as MWE categories, can additionally be included. The boundary between a corpus and a treebank can become fuzzy when progressively more refined information is added, and in fact, some of the examples listed below include syntactic information typically indicative of a treebank. We present three examples of MWE corpora below.

The PARSEME CORPORA (Savary et al., 2017; Ramisch et al., 2018, 2020) are multilingual corpora annotated with verbal MWEs (vMWEs), developed for the PARSEME shared tasks described above in Section 2.4.1.1.[23] The corpora consist of texts in multiple languages that have been manually annotated for vMWEs, following the annotation guidelines (see Section 2.4.2). The creation and annotation of the Irish corpus for this task is described in Chapter 5.

STREUSLE (Supersense-Tagged Repository of English with a Unified Seman-

[23]The corpora are hosted on the LINDAT/CLARIAH-CZ Repository for LRs, and various editions of the corpora can be downloaded from the GitLab page for the project: `https://gitlab.com/parseme/corpora/-/wikis/home`

tics for Lexical Experessions) is a rich resource developed through many years of research (Schneider et al., 2014b, 2016b, 2018). The corpus is composed of the web reviews portion of the English Web Treebank (Bies et al., 2012) and comprehensively annotated for MWEs and semantic supersenses for lexical expressions. In addition to these supersenses and MWE types, the most recently released edition (4.4) features a hierachy for preposition and possessive supersenses (SNACS), as well as fine-grained categories for four vMWEs, based on the categories annotated for English in the PARSEME corpus (Walsh et al., 2018). The corpus is used for the task of MWE identification, which is explored in more depth in Chapter 7.

Wiki50 (Vincze et al., 2011) is a corpus composed of 50 Wikipedia articles annotated for MWEs and Named Entities (NEs). It is the first corpus of its kind to include both MWEs and NEs; the authors maintain that the similarities between these two constructions (with NEs also being considered by some as a specific type of MWEs (Jackendoff, 1997a)) makes their simultaneous treatment a logical operation. 50 articles (totalling 4,350 sentences) were randomly selected from Wikipedia and annotated by two annotators for the MWEs, NEs and sentence boundaries. Six categories of MWEs were annotated (nominal compounds, adjectival compounds, verb-particle constructions, light-verb constructions, idioms, and 'other'), and four categories of NEs were annotated (persons, organizations, locations, and miscellaneous). There were 3,861 MWEs annotated in total, and 8,976 NEs. Vincze et al. (2011) also conducted a number of experiments on the automatic identification of MWEs and NEs. One approach used a dictionary-based method, extracting the internal links of the Wikipedia articles as nominal compound and NE candidates. The annotated corpus was used to evaluate the results of this experiment. A machine learning method for identifying nominal compounds was also presented, with the Wiki50 corpus used as evaluation data for a Conditional Random Fields (CRF) (Lafferty et al., 2001) classifier. Training data was automatically generated using the dictionary based method described. This method of bootstrapping in order to generate more data is an effective method for low-resource languages such as Irish.

While this technique is not employed in this thesis, it provides a possible avenue for future work.

MWE corpora can also include parallel corpora, such as with the SzegedParalellFX English–Hungarian Parallel Corpus (Vincze, 2012). This corpus contains over 14,000 aligned sentences in English and Hungarian selected from theSzeged-Paralell English–Hungarian Parallel Corpus (Tóth et al., 2008), and is annotated by three annotators for light-verb constructions. The corpus can be used for training data in the task of MWE identification (specifically the identification of light-verb constructions), or in a number of downstream NLP tasks that will benefit from the inclusion of this MWE information, e.g. a machine translation system.

2.5.1.3 Treebanks

Treebanks differ from typical MWE corpora in that they are annotated with syntactic information denoting the relationships between words, as described in Section 2.4.3. In addition to syntactic relations, treebanks can also be annotated with MWEs. In terms of information, they are considered the richest of the three resources associated with MWEs, and a valuable resource for both data-driven parsing systems and syntax-aware MT systems. The practice of including MWE annotation or information in treebanks is debated in the field, with different opinions on whether syntactically regular but semantically idiomatic MWEs ought to be annotated in treebanks (e.g. 'spill the beans').

The advantage of including such information is in potentially improving awareness of MWEs in syntactic parsing, which has shown to have an improvement on parsing accuracy (Nivre and Nilsson, 2004; Seretan, 2011; Green et al., 2013; Candito and Constant, 2014; Savary et al., 2015). The idea being that treatment of certain constructions as syntactic units instead of individual tokens can improve a system's understanding and processing of sentences (e.g. *They tried to* ***hold up*** *a bank* vs *the container can* ***hold*** *up to 10 gallons*).

Analysis of MWEs has been carried out on a number of dependency treebanks

such as in Czech (Bejček and Straňák, 2010), Hungarian (Vincze et al., 2013) and Turkish (Eryiğit et al., 2015). As discussed in Section 2.4.1, the PARSEME network was established to improve cross-lingual treatment of MWEs, through establishing a multilingual framework for processing MWEs in applications such as machine translation and parsing (Savary et al., 2015). A survey on the treatment of MWEs in treebanks was performed by Rosén et al. (2015), and subsequent guidelines based on this survey were developed to aid the annotation of MWEs in both constituency and dependency treebanks (Rosén et al., 2016). One of the observations made in this research was that it should be possible to search for various types of MWEs based on their characteristics (e.g. compositional vs non-compositional).

2.5.1.4 Research

As with other areas in NLP, we have seen a increase in the creation, collection and distribution of MWE-aware resources in recent years. Despite such increased efforts, there is a deficit of such resources even in higher-resourced languages, and in particular, a lack of standardisation for such resources. Initiatives such as the XMELLT project (Calzolari et al., 2002b) and the ISLE (International Standards for Language Engineering)[24] project made strides towards developing practices for the treatment of MWEs in LRs, including a standardised description for MWE entries. Copestake et al. (2002) describe an architecture for lexical encoding of a large number of classes of MWEs, which takes into account the range of variability and flexibility that MWEs can present. This schema is further built upon by Villavicencio et al. (2004), extending it to be uniformly applied to all MWE types.

To consolidate research in this field, the Special Interest Group on the Lexicon Multiword Expressions (SIGLEX-MWE) was established as a SIG of the Association for Computational Lingusitics (ACL). Resources promoted by this section include the PARSEME Shared Tasks (see Section 2.4.1.1) and the multilingual corpora annotated for verbal MWEs (see Section 2.5.1.2). They also coordinate the MWE

[24] `http://www.ilc.cnr.it/EAGLES96/isle/`

workshop series[25] that have been important for spearheading multilingual research in this field.

In their survey paper, Losnegaard et al. (2016) explore existing resources for MWEs across the 35 languages represented in the responses. They highlight three major issues uncovered by this survey, namely: (i) metadata on MWE resources being inconsistent or incomplete, (ii) a lack of standardization in licensing, or a lack of license altogether, and (iii) the difficulty inherent in locating such resources from existing infrastructures, due to cataloging problems or missing metadata values to show that such resources contain MWEs.

The results of their survey show a bias in LRs available for majority languages, with the overwhelming number of resources available for English, followed at some distance by Italian, French, German, and Polish.[26] Figure 2.4 displays the number of language-dependent resources for each of the surveyed languages (excluding language-independent resources, such as tools).

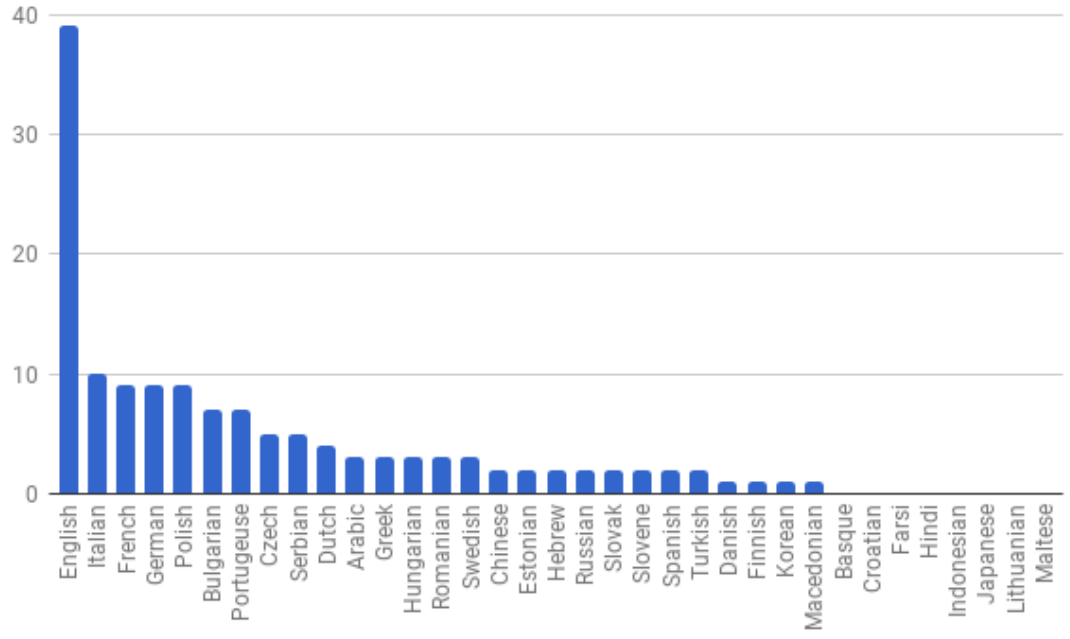


Figure 2.4: MWE Language Resources available for each of the surveyed languages.

This survey noticeably demonstrates a lack of resources for many European lan-

[25] `https://multiword.org/events/previousevents`
[26] Anonymized survey results are available online `https://goo.gl/P4To2f`.

guages, and does not report on any resources for Irish. Our aim in the next chapter is to describe existing Irish resources that are useful for tasks in the processing of MWEs.

2.6 Conclusion

In this chapter, we discuss many of the key concepts that exist in the field of MWE processing for NLP. We begin by presenting a definition of MWEs, drawing attention to the number of terms and analyses that have been previously offered for this linguistic phenomenon. We adopt a definition used by many authors in the field, focusing on two key characteristics of MWEs: 1) multiple words which 2) display idiosyncratic or idiomatic behaviour. We explore the properties exhibited by MWEs, including idiomaticity on the semantic, syntactic, lexical, statistical and pragmatic levels, as well as other traits such as variability, domain-specificity, and morphosyntactic heterogeneity.

Within the field of NLP, MWEs present a number of challenges, which correspond to the unique properties they display. These challenges increase the complexity of automatic processing of MWEs, with some challenging properties disproportionately affecting certain tasks in the topic.

We also summarise two multilingual frameworks, which factor into the decisions made during our analysis of Irish MWEs. The PARSEME framework includes a series of shared tasks on the identification of MWEs, which are discussed in more detail in Chapter 7. The UD framework is addressed again in Chapter 4, where we discuss how the treatment of MWEs in this framework influenced our categorisation efforts for Irish MWEs.

We discuss language resources and their uses, focusing on resources that can be useful for the purposes of MWE processing. In particular, we address lexical resources, corpora, treebanks, and research, providing an overview of many efforts in this field, from which we draw inspiration in the creation of our own resources

(see Chapter 5).

Next we will see how some of these key concepts apply to the Irish language.

Chapter 3

MWEs in Irish NLP

> *Now and again he spoke to those that served him and thanked them in their own language. They smiled at him and said laughing: 'Here is a jewel among hobbits!'*
>
> J.R.R. Tolkien, *The Fellowship of the Ring*

3.1 Introduction

While the previous chapter (Chapter 2) provided background information on the key concepts in the field of MWE processing for NLP, this chapter serves to ground that field of research in the Irish language. We provide an overview of the language, including an exploration of many morphosyntactic features of the language that are relevant to our discussion on idiosyncratic constructions. We also explore those language resources that are both available for Irish and applicable for our needs in the automatic processing of MWEs. In this chapter, we begin addressing **RQ2**, beginning with an investigation into the existing resources for Irish.

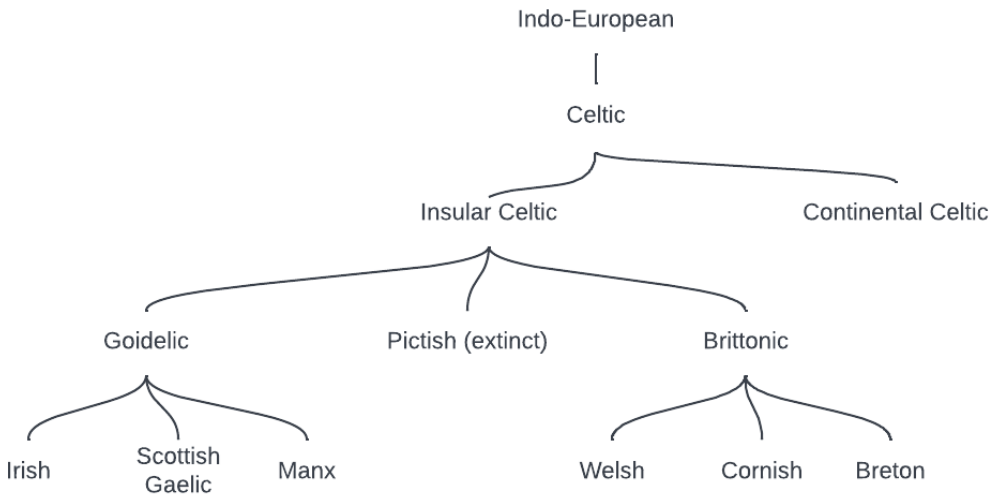


Figure 3.1: The Celtic Family of languages.

3.2 Irish Language

Irish, or *Gaeilge*, is the first official language of Ireland, the second official language being English. It belongs to the Goidelic branch of the Celtic languages, and as such, is closely related to both Scottish Gaelic and Manx Gaelic which are still spoken today (see Figure 3.1).[1]

The language is thought to have been introduced to Ireland in about 300 B.C. (Ó Siadhail, 1991). The language has been through several iterations, evolving from Old Irish (roughly 600–900 AD) to Middle Irish (900AD–1200AD) to Early Modern Irish (1200–1600), and finally Modern Irish, which is in use today (Lynn, 2016). While the language was widely spoken by the Irish population in the sixteenth century, a shift in language usage of the population of Ireland began in earnest during the seventeenth century with the English language being enforced by the ruling aristocracy.

By the nineteenth century, the effects of British colonialism, famine, emigration and social pressures to use English had relegated the usage of the Irish language to mainly rural areas in the west. A movement to revive the language came about

[1]Figure adapted from several sources (Stenson, 1981; Ó Siadhail, 1991; Tredinnick-Rowe, 2017).

at the turn of the last century, with the foundation of *Conradh na Gaeilge*, or the Gaelic League, which continues to promote and preserve the language to this day.

There are three major spoken dialects of Modern Irish, named for the geographic region each dialect is based in; Ulster Irish, spoken mostly in the north of the country, Connacht Irish, spoken in the west, and Munster Irish, spoken in the south. These dialects differ in pronunciation, and exhibit lexical and grammatical differences (Judge et al., 2012). However, we do not distinguish between these dialects for our work, and instead focus our research efforts on An Caighdeán Oifigiúl (The Official Standard), the standard written form of the language.

3.2.1 Syntax of Irish

For the purposes of identifying idiosyncratic behaviour displayed by Irish MWEs, we first establish some of the regular behaviours of the language. As we discussed in Chapter 2, idiomaticity in MWEs can be displayed on a semantic, syntactic, lexical, statistical or pragmatic level. Focusing on two of these levels, we describe the regular syntactic and lexical behaviour of Irish here below, focusing on those parts of speech that are most relevant to our understanding of MWEs, particularly the categories of MWEs discussed in Chapter 4.

The discussion below is a synthesis of discussions of Irish syntax by authors such as Christian Brothers (1999), Greene (1966), Stenson (1981), Ó Siadhail (1991), Bräsicke (2003), Uí Dhonnchadha (2009), and Lynn (2016), with the examples below largely taken from these works.

3.2.1.1 Word Order

Similar to the other languages of the Goidelic branch, Irish is a Verb-Subject-Object (VSO) language, with other constituents generally following these three (see Example 17). The order of these words is quite rigid for Irish, with only certain constructions altering this word order. There are few cases where an element can appear between the verb and the noun, such as the insertion of an adverb (see Example 18),

while prepositional phrases and adverbs can be inserted between the subject and the object (see Example 19). Some of the exceptions to this order are with narrative fronting (Example 20) and copular fronting (Example 21), sentence-initial adverbial phrases (Example 22), and semi-negative or 'only' constructions (Example 23).

(17) *Labhraíonn Abigail Gaeilge le Meghan go rialta*
speaks Abigail Irish to Meghan regularly
'Abigail regularly speaks Irish to Meghan'

(18) *Tá* ***ar ndóigh*** *daoine a chreideann...*
are of-course people REL believe...
'There are of course people who believe...'

(19) *Thug sé* ***dom inné*** *é*
gave he to-me yesterday it
'He gave it to me yesterday'

(20) ***Súil*** *ní rabhadar ag tógáil de*
eye not were-they at taking from-him
'An eye, they weren't taking off him'

(21) ***Is ise*** *a chonaic sibh ann*
is her REL saw you there
'It was she whom you saw there'

(22) ***Ar aon chuma****, níor bhac sí leis*
at any effect NEG interfered she with-it
'At any rate, she didn't bother with it'

(23) ***Níl ann ach*** *an bheirt againn*
is-NEG there but the two at-us
'There is only the two of us'

3.2.1.2 Nouns

Irish has been called a noun-centered language (Greene, 1966), and constructions of the noun and verb where the noun carries the semantic weight of the construction (Examples 24, 25, 26 and 27) occur commonly in Irish. Many of these constructions can be considered idiomatic, and are discussed further in Chapter 4.

(24) *Tá súil agam...*
is expectation at-me...
'I hope...'

(25) *Tá grá agam do...*
is love at-me for...
'I love...'

(26) *Bím ag rith gach Luan*
be-I at running every Monday
'I run every Monday'

(27) *Thug mé cic dó*
gave I kick to-it
'I kicked it'

Verbal nouns

The verbal noun is used frequently in Irish. The infinitive form of the verb is constructed using verbal nouns (Example 28), and it can also serve in the progressive aspectual form of the verb (Example 29), as well as nominalisations of the verb (Example 30). A verbal noun can be considered as something between noun and verb. Nominal traits include its ability to serve as the subject of the sentence, and its ability to inflect for case as with other nouns (Example 31). Verbal traits include its ability to be modified by adverbs or its ability to have direct objects (Example 32).

(28) *Thosaigh sí an scéal a insint*
started she the story PRT telling
'She started to tell the story'

(29) *Bhí sé ag rith*
was he at running
'He was running'

(30) *Cad é an bhaint leis*
what is the connection with-it
'What is the connection to it'

(31) *lucht foghlama na Gaeilge*
people learning-GEN the Irish-GEN
'Irish learners'

(32) *Ba mhaith liom é a fhéiceáil* ***go díreach***
COP good with-me him PRT see PRT directly
'I would like to to see him directly'

Noun forms that do not have a verbal equivalent can be expressed through verbal nouns, such as with *ag siopadóireacht* 'shopping' and *ag feadaíl* 'whistling', where verbs such as **siopadóireann sé* or **feadaíleann sé* do not exist.

3.2.1.3 Clitics

Clitic particles in Irish can be associated with nouns or verbs, and serve a variety of grammatical functions. Stenson (1981) advocates for treating these clitic particles as a class, since they share certain characteristics, such as never appearing in isolation and never being stressed.

Principal verbal clitics include the question particle *an*, the negative particle *ní*, the copula *is*, the complementizer *go*, the negative complementizer *nach*, the relative particle *a*, the negative relative particle *nach*, the verbal noun particle *a*, the conditional *má/dá*, the negative conditional *mura*, and the past tense particle *do*, which in contemporary Irish is realised as *d'* before vowel-initial verbs.

The clitic particle *go* can combine with adverbs to form adverbial constructions in Irish (e.g. *go háirithe* (PRT certain) 'especially'). Although these constructions appear similar to other MWEs such as fixed expressions (see Chapter 4), they are regular syntactic combinations of the language, and as such do not pass our criteria of idiosyncrasy.

3.2.1.4 Verbs

Modal and auxiliary verbs are constructed differently in Irish than in English. Auxiliary constructions such as progressive 'be ...ing', the perfective 'have ...en', and the expective 'going/about + INF' are instead formed in Irish using the substantive *bí* in combination with the verbal noun, with choice of preposition and word order distinguishing these forms. See Examples 33 (perfective), 34 (expective), and 35 (progressive).

(33) *Tá sí tar éis amhrán* ***a chanadh***
is she after song PRT sing

'She is after singing a song'

(34) *Tá sí le hamhrán* ***a chanadh***
is she with song PRT sing
'She is about to sing a song'

(35) *Tá sí* ***ag canadh*** *amhráin*
is she at singing song-GEN
'She is singing a song'

Modal verbs are, for the most part, similarly constructed using either the substantive *bí* or the copula *is* as the auxiliary verb, which inflect to show conditional/aspectual information. These auxiliary verbs combine with nouns (Example 36), adjectives (Example 37), adverbial (Example 38) and prepositional phrases (Example 39) to create modal verbs. There are a few exceptions, such as with the verbs *caith* 'must', *féad* 'can' and *glac* 'take/accept', which can theoretically fully inflect.

(36) *Is ceart dom*
COP right to-me
'I ought'

(37) *Tá mé ábalta*
is I able
'I am able'

(38) *Tá orm*
is on-me
'I must'

(39) *Tá mé in ann*
is me able
'I can'

3.2.1.5 The Copula

In Irish there are two verbs where the English has one verb 'to be'. The substantive verb *bí* inflects as any other verb in Irish, and uses separate particles for negation and interrogation. The copula *is* only has two tensed forms (present/future and past/conditional), and inflects for negation, interrogation and complementisation. The usage of the substantive verb and the copula are distinct. The copula is used

to connect subject and predicate in the cases of essential or inherent qualities, while the substantive verb is more often used in the cases of temporal qualities, existence, location, possession, and other similar qualities (Stenson, 1981).

Five distinct uses of the copula have been noted by Christian Brothers (1999): classification, identification, ownership, emphasis, and adjective predicates. To these uses, Ó Siadhail (1991) adds also uses with demonstratives pronouns *seo* 'this', *sin* 'that', *siúd* 'yon', etc. He also distinguishes between exclamatory use, equational use, comparative use, and use in prepositional phrases, which together are classed under ownership and adjective predicates.

Classification is the equation of a subject with a member of a class, such as in Example 40. Identification is the equation of two definite NPs, as with Example 41. Emphasis can be expressed through fronting, as illustrated in Example 21 above. This fronting structure can be applied to almost any element of the sentence, be it noun, adjective, adverb, prepositional phrase, etc. Example 42 demonstrates this structure with a prepositional phrase.

(40) *Is éan smólach*
COP bird thrush
'A thrush is a bird'

(41) *Is í* <u>*mo dheirfiúr*</u> <u>*an dochtúir*</u>
COP her <u>my sister</u> <u>the doctor</u>
'My sister is the doctor'

(42) *Is* ***ar an mbus*** *a bhuail mé léi*
COP on the bus REL met I with-her
'It's on the bus that I met her'

The exclamatory, equational and comparative usage are all uses of the copula linking the NP with an adjective. Examples 43, 44 and 45 illustrate these uses. The copula can be combined with prepositional phrases, some of which are idiomatic in nature. These categories are explored further in Chapter 4. Ownership is expressed as one of the number of idiomatic constructions that are formed with the copula; it is expressed through the combination of '*is* + *le* owner + object', as with Example

46. Other idiomatic combinations can be formed with *as* 'from' (Example 47), and *do* 'to' (Example 48). Finally, the usage of the copula with *seo* and other demonstrative pronouns is used somewhat similarly to the French *voici* 'here is', as shown in Example 49.

(43) *Nach deas é* !
COP-NEG nice it
'Isn't it nice!'

(44) *Is mar a chéile iad*
COP alike them
'They are alike'

(45) *Is sine Jason ná Gillian*
COP older Jason than Gillian
'Jason is older than Gillian'

(46) *Is le Jane an carr nua*
COP with Jane the car new
'The new car is Jane's'

(47) *Is as Ciarraí mé*
COP from Kerry me
'I am from Kerry'

(48) *Lauren is ainm dom*
Lauren COP name to-me
'Lauren is my name'

(49) *Seo leabhar*
this book
'This here is a book'

3.2.1.6 Prepositions

There are two features of prepositions in Irish that pose challenges for MWE processing. The first characteristic is the ability for simple prepositions in Irish to combine with personal pronouns to form "prepositional pronouns" or "conjugated prepositions". The 18 prepositions that can be combined in this way are given in Table 3.1, while Table 3.2 demonstrates a simple paradigm for this combination.

ag	at	*idir*	between
ar	on	*ionsar*	toward
as	out of	*le*	with
chuig	to	*ó*	from
de	of	*roimh*	before
do	of/for	*seach*	past
faoi	under	*thar*	past
fara	next to	*trí*	through
i	in	*um*	around

Table 3.1: Simple Irish prepositions and their English translations.

le + mé → liom	'with me'	*le + muid → linn*	'with us'
le + tú → leat	'with you'	*le + sibh → libh*	'with you'
le + é → leis	'with him, it'	*le + iad → leo*	'with them'
le + í → léi	'with her'		

Table 3.2: Paradigm of prepositional pronouns with *le*.

The second feature of prepositions that impacts on their processing with MWEs are compound prepositions. These are formed by combining a preposition and a noun to give rise to more complex prepositions that are treated as fixed expressions. Examples 50 and 51 below illustrate these constructions, and compound prepositions are discussed in more detail in Chapter 4.

(50) *os cionn*
over head
'Over/above'

(51) *le haghaidh*
with face
'For'

Prepositions are used heavily in Irish to alter or enhance the meaning of certain constructions. In our discussion of verbs (Section 3.2.1.4) and the copula (Section 3.2.1.5) we have already seen how prepositions can be vital in understanding certain contexts. The various combinations of preposition and verbs are discussed further in Chapter 4.

3.2.1.7 Adjectives

Adjectives typically follow the noun in Irish (with a few exceptions, e.g. *sean* 'old', *droch* 'bad', *dea* 'good') when used attributively (Example 52). They can also be used predicatively (Example 53), or alternatively, they can be turned into adverbs (Example 54).

(52) *An madra* ***cliste***
The dog clever
'The clever dog'

(53) *Tá an madra* ***cliste***
is the dog clever
'The dog is clever'

(54) *Tá an madra ag obair* ***go cliste***
is the dog at working PRT clever
'The dog is working cleverly'

Verbs can be formed into verbal adjectives, similar to the verbal noun. These verbal adjectives function as normal adjectives do, as attributive (Example 55) or predicative (Example 56). They cannot be formed into adverbs, however.

(55) *An buidéal* ***briste***
The bottle broken
'The broken bottle'

(56) *Tá an buidéal* ***briste***
is the bottle broken
'The bottle is broken'

3.2.1.8 Adverbs

The adverb can occur in many places in the sentence (Examples 57, 58, and 59). They fall under several categories. Directional or locational adverbs are used to provide spatial relations. This category of adverbs covers points of the compass, up and down, front and back, left and right, here and there, and in and out. Temporal adverbs include *inniu* 'today', *aréir* 'last night', *anois* 'now', and *anuraidh* 'last year'. Adverbs can also be formed from adjectives, as discussed in Section 3.2.1.7.

(57) *Chuala mé fuaim* ***laistigh***
heard I sound within
'I heard a sound within'

(58) *Tá sé caite* ***ansin*** *ina stolp*
is he thrown there in heap
'He is thrown there in a heap'

(59) ***Go tobann****, bhuail sé leo*
suddenly met he with-them
'Suddenly, he met with them'

3.3 Existing Resources Used in Irish MWE Research

In this section, we explore which resources in Irish are applicable to our tasks in the automatic processing of Irish MWEs. Following the same structure as in the previous chapter, we explore lexical resources, corpora, treebanks, and research that is relevant to our needs. These resources we describe serve as our data for the creation of categories of MWEs (see Chapter 4), as well as providing the foundation for the construction of several MWE resources, described in Chapter 5, which in turn is applied in our experiments exploring tasks in MWE processing in Chapters 6 and 8.

3.3.1 Lexical Resources

There are a number of lexical resources for Irish that are considered useful for the purposes of research into Irish MWEs, particularly in the context of NLP. In this section we describe some of these resources, which are used when building our first lexicon of Irish MWEs for NLP purposes (see Chapter 5).

***Foclóir Briathra Gaeilge*:** A valency dictionary for Irish verbs was created by Wigger (2008) and his team. *Foclóir Briathra Gaeilge*[2] is an effort in documenting

[2] `http://www.potafocal.com/fbg/_intro/`

differentiation in meaning of various constructions with Irish verbs. The dictionary is a data-driven documentation of how these verbs are used regularly in spoken and written Irish, together with a semantic review of the usual arguments that co-occur with the verb, expressed using an inventory of semantic roles. The dictionary includes 200 commonly used verbs selected from two digitized corpora: Caint Chonamara,[3] a corpus of spoken Irish recorded in 1964 by speakers in the Connemara region, and Corpas Náisiúnta na Gaeilge,[4] a corpus of written Irish compiled from a variety of genres and domains, published by Institúid Teangeolaíochta Éireann. The dictionary is available for download under the Open Database License.

Peadar Ó Laoghaire Idiom Collection: A database of manually annotated idioms in Irish, extracted from the published work of an tAthair Peadair Ó Laoghaire, was compiled and published online[5] for use in linguistic research (Ní Loingsigh, 2016). Several further publications document the creation of this database and analyse the idioms collected therein (Ní Loingsigh and Ó Raghallaigh, 2016, 2018; Ní Loingsigh, 2019, 2021). The collection consists of 420 entries that each include a baseform or canonical form of the idiom, along with a headword, a description of the idiom, a contextual example of the idiom taken from the corpus, usage notes where applicable, a cross-reference to idioms with a similar meaning, and additional information, including any additional examples. The database is hosted by the Gaois research group at Dublin City University.[6] While the resource is not currently available for download, a list of these idioms was made available to us for research purposes.

***An Bunachar Náisiúnta Téarmaíochta don Ghaeilge*:** 'The National Terminology Database for Irish' (Téarma) is a terminology database available for use online.[7] The database, previously known as Focal, is hosted by the Gaois research

[3] https://www.sksk.de/index.php/de/#conamara
[4] https://corpas.ie/ga
[5] https://www.gaois.ie/en/idioms/
[6] https://www.gaois.ie/en
[7] https://www.tearma.ie/

group in partnership with the Terminology Committee at Foras na Gaeilge. The Téarma database consists of about 186,000 terminological entries. A number of publications document the creation of this corpus, including Měchura and Ó Raghallaigh (2010), Nic Pháidín et al. (2010) and Bhreathnach and Nic Pháidín (2013). The entries are taken from a number of sources: (i) domain-specific dictionaries compiled by experts in the subject fields (e.g. 'Dictionary of Parliamentary Terms'), (ii) domain-specific lists compiled by the Terminology Committee for distribution to teachers and inspectors of these subjects or compiled in collaboration with third parties developing handbooks or domain-specific documentation, and (iii) miscellaneous terminology resulting from enquiries by the public. While the Téarma database is available for download either in `txt` or `tbx` format from their website,[8] the usage is limited to personal use only, and the copyright to this database prevents the contents from being republished in any form.

***Líonra Séimeantach na Gaeilge*:** LSG is an Irish semantic network consisting of the semantic relationship between words in Irish (Scannell, 2003). This database is similar to the English WordNet, previously described in Chapter 2. As we have previously stated, wordnets are linguistically rich resources with many applications in NLP tasks. As well as creating semantic links between words, this resource links Irish words to their closest equivalent in English, adding further uses to this resource for improving machine translation. The database consists of 32,742 synsets, 36,262 headwords and 77,596 individual word senses, and includes terminology, literary terms, dialect forms and slang words. This resource is shared under a CC BY-SA licence.[9]

***Pota Focal Gluais Tí*:** 'Pot of Words House Glossary', or *Pota Focal*, is an Irish-to-English dictionary, developed as a resource for learners of Irish (Měchura, 2017). The dictionary contains over 6,000 terms of common vocabulary items. Entries

[8] `https://www.tearma.ie/ioslodail/`

[9] Resource can be downloaded in several formats from `https://cadhan.com/lsg/index-en.html`.

are listed under an Irish headword. Grammatical information such as POS tags, grammatical forms and morphological features are supplied, along with an example of usage of the word. The dictionary also offers some statistics on the frequency of each word. The dictionary is under the Creative Commons Attribution Non-Commercial Share-Alike licence.[10]

The English-Irish Dictionary: This dictionary (EID) was compiled and published in 1959 (de Bhaldraithe, 1959). Material was collected by the dictionary staff from written resources and a selection was published along with their English translations. The selection process chose words that were in usage by the contemporary Irish speakers of the Gaeltacht areas, and for which there existed an equivalent in common English. Terminology and technical terms included in the dictionary were added with consultation from specialists in the field, in order to choose just one single term. The entries are supplied under an English headword, with subsections for the different semantic uses of the word, and the different grammatical categories of the word in Irish. A searchable electronic version of the dictionary was published by Foras na Gaeilge on their Dictionary and Language Library website.[11]

As an updated version, the New English-Irish Dictionary (NEID) was launched as an online dictionary by Foras na Gaeilge in 2013[12] with revised entries and additional grammatical information, including standardised spelling matching the *Caighdeán Oifigiúil*/Official Standard. The initial version had over 7,000 headwords, with new entries being added regularly. A printed version of the dictionary was released (Ó Mianáin, 2020) containing 31,000 of the 48,000 entries. Entries are organised, similar to the EID, under an English headword, and includes different semantic and grammatical usages, and examples of the word in context.

***Foclóir Gaeilge-Béarla*:** The 'Irish-English Dictionary', or FGB, is an Irish-English dictionary first published in 1977 (Ó Dónaill, 1977), and published online

[10]Resource can be downloaded from `https://github.com/michmech/pota-focal-gluais/`.
[11]`https://www.teanglann.ie/en/eid/`
[12]`https://www.focloir.ie/en/`

by Foras na Gaeilge in 2013, on their Dictionary and Language Library website.[13] Included in the dictionary are some archaic or literary terms that are not commonly used by contemporary Irish speakers, but for which there lacks an equivalent word in common usage, as well as technical terms. The use of the dictionary is similar to the EID, with the aim of collecting and describing the meaning of common Irish words relative to English, however it is not limited to including words that have an equivalent word in common English. Some subsection entries require only a single English word to capture the meaning, while others require a short explanation of the meaning. The spelling of Irish words follows the *Caighdeán Oifigiúil*/Official Standard, with variant spellings and forms also provided, including archaic spellings where they exist. Grammatical information is provided for the Irish word, including POS information and inflected forms. Example uses of the Irish word are included in the subsection entries where necessary to demonstrate the difference in usage for English. Of the 59,700 entries, we found 48 multiword headwords, with most instances of idiomatic uses contained in the examples given in the subsection entries. Many of these subsection entries were added to our lexicon described in Chapter 5.

***An Foclóir Beag*:** 'The Small Dictionary', or FB, is a monolingual Irish dictionary first published in 1991 (Ó Dónaill and Ua Maoileoin, 1991), with a searchable electronic edition launched by Foras na Gaeilge on their Dictionary and Language Library website,[14] with some minor changes to the notation for clarity. The dictionary was intended to provide the most common meanings of many of the words in Irish frequently used by teenagers. The Irish headwords are combined with grammatical information, and brief subentries to describe the meaning.

Each of the four electronic dictionaries published by Foras na Gaeilge (EID, NEID, FGB and FB), as well as Téarma and the Peadar Ó Laoghaire Idiom Collection, were made available to us for the purposes of Irish NLP research. The files were

[13] `https://www.teanglann.ie/en/fgb/`
[14] `https://www.teanglann.ie/en/fb/`

provided in `xml` format or as a `csv` file. For any content that is under copyright, the entries that were extracted from these resources are not available for republishing. The other resources are open source, and can be downloaded at the links provided above.

3.3.2 Corpora

We avail of many open source parallel corpora in our experiments in developing datasets for MWE-aware MT systems, as described in Chapter 5. According to the ELRC White Paper country profile for Ireland, following promotional and educational activities organised by the ELRC, there has been increased communication and sharing of resources between a growing number of stakeholders and public bodies, including government departments, county councils, universities, the national broadcaster (RTÉ), dictionary publishers, and the language commissioner's office (Berzins et al., 2019). This activity was further increased with the ELRI project, which saw the launch of the National Relay Station (NRS),[15] a secure online platform for members of public institutions in Ireland to submit their data. More recently, the PRINCIPLE project (Way and Gaspari, 2019) saw the collection and processing of LRs, which were then uploaded to the ELRC-SHARE repository, a platform for the collection and sharing of many LRs that have been uploaded under an open licence.

Two resources we use from ELRC-SHARE include *Citizen's Information* and *Department of Culture, Heritage and the Gaeltacht* (*DCHG*). *Citizen's Information* is a dataset crawled from the organisation's national Irish website[16] containing Irish Public Sector Data. *DCHG* includes four datasets provided by the Irish Government Department of Culture, Heritage and the Gaeltacht, the department responsible for Irish language affairs.[17] The data was shared for the purposes of training MT

[15] `https://elri.dcu.ie/ga-ie/`

[16] `https://www.citizensinformation.ie/en/`

[17] As of 30 September 2020, this department has been renamed the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media.

models, and includes inter-departmental memos, emails and notices. Both *Citizen's Information* and an anonymised version of the *DCHG* dataset are available under a CC BY 4.0 licence.

Another source of Irish corpora are many of the datasets available on OPUS,[18] a repository for open source parallel language sentence-aligned corpora in a variety of formats (Tiedemann, 2012, 2016). 57 corpora are represented in over 700 language and language variants,[19] as well as a number of tools (Aulamo et al., 2020), and open translation models (Tiedemann and Thottingal, 2020).

These datasets include files from the Directorate-General for Translation (*DGT*), which is a source of data produced by the European Commission and comprises high-quality professionally translated parallel data based on translation memory files generated by DGT translators.[20] The OPUS website contains other datasets from European sources such as a parallel corpus of legal text collected from the European Constitution (*EUConst*), and documents from the EU bookshop (*EUbooks*) from the public admin domain. *Tatoeba* is a dataset of short learner-friendly sentences with simple vocabulary and grammar. *GNOME*, *KDE4* and *Ubuntu* are localized translations of the operation manual for these systems. The domain for each of these three corpora is technical and many of the sentences are translations of terminology or short phrases containing technical jargon.

Paracrawl,[21] an ongoing large-scale web-crawling project, is also available through the OPUS repository. The project used web crawling tools to gather parallel datasets across European languages. This process enables the creation of a large dataset, but the quality is not assured, and it can contain noise, such as non-post-edited MT output (Defauw et al., 2019). Other crawled data includes articles generated by *Teagasc*, which are published under PSI licence. In addition, there is a parallel dataset that has been crawled from bilingual websites (Dowling et al., 2018), using ISLP

[18] `http://opus.nlpl.eu/`

[19] Most recent numbers dating from 2020

[20] `https://ec.europa.eu/info/resources-partners/translation-and-drafting-resources/guidelines-translation-contractors/guidelines-contractors-translating-irish_en`

[21] `https://www.paracrawl.eu`

Focused Crawler (Papavassiliou et al., 2013). This data is not yet freely available as individual site owners have yet to confirm licensing and data re-use conditions.

Finally, some of the parallel corpora have been shared with us by the licence holders for research purposes, such as a collection of bilingual press releases from 2016 and 2017 generated by *Conradh na Gaeilge* (*Conradh*).

Monolingual corpora are a useful source of Irish data, and can be used in training language models in Statistical Machine Translation (SMT), or for use in filtering experiments. In Appendix B, we explore some of the uses of large monolingual corpora for the purposes of MWE discovery.

The New Corpus for Ireland (*NCI*) (Kilgarriff et al., 2006) contains a large corpus (approximately 30 million tokens) of written Irish created by Foras na Gaeilge while developing the NEID, and hosted on the Sketch Engine,[22] where members of the public can perform single- or multi-word term searches of the corpus. The corpus was compiled from a mixture of existing corpora, a collection of texts acquired from publishers, authors and newspapers, and data collected from the web. Due to this variety of sources, the domain is mixed, and includes works of fiction, informative text, news reports, and official documents.

Vicipéid[23] belongs to the Wikimedia collection of online resources as the Irish language Wikipedia, an open source publicly available encyclopedia written and maintained by a community of volunteers following a model of open collaboration. It consists of 56,268[24] articles, and currently ranks 93 out of 326 languages in terms of size. The open collaboration model allows for all members of the public to create and edit articles, fostering an environment of balanced, unbiased information. However, this open collaboration means the quality of the content cannot always be assured, and this can lead to noisy or inaccurate content, such as the case with Scots Wikipaedia, where a non-native speaker of the language inundated the Wikipedia with poor quality articles (Ultach, 2020).

[22] `https://www.sketchengine.eu/the-new-corpus-for-ireland/`
[23] `https://ga.wikipedia.org/`
[24] Accessed on 30/03/2022

These resources are valuable for a variety of NLP related research and downstream applications. However, these corpora do not include any MWE information, and are not useful for supervised tasks in MWE processing. As part of the scope of this research, we describe how we used these corpora in the development of MWE-aware resources, including the creation of an MWE-tagged GA-EN parallel corpus (See Chapter 5).

3.3.3 Treebanks

In the previous chapter (Chapter 2), we described the Universal Dependencies framework. The Irish UD Treebank (IUDT) (Lynn et al., 2017) has been part of the UD project since the initial release of v.1, and was converted from the original Irish Dependency Treebank (IDT)[25] (Lynn, 2016). However, a thorough and cohesive treatment of MWEs in the original IDT annotation scheme or the converted UD scheme was lacking, due to two factors: (i) both treebanks were the product of a PhD dissertation, and the coverage of MWEs was outside the scope of this work, and (ii) the limited nature of the research (both theoretical and applied) of Irish MWEs for NLP.

McGuinness et al. (2020) describe a treatment of MWEs in this treebank, which informs our own treatment of Irish MWEs. This topic is explored further in Chapter 4.

3.3.4 Theoretical Linguistic Research

While there have been research efforts covering a theoretical linguistic understanding of the Irish language, it is under-researched at the level of NLP (Lynn, 2022). This is particularly true for MWEs, which, although they have been treated by a number of researchers, have lacked a formalisation for a computational treatment until now. We present here a brief overview of the existing research on MWEs in Irish, which

[25] `https://github.com/tlynn747/IrishDependencyTreebank`

mainly focuses on a selection of semantically or syntactically idiomatic constructions. This research is explored in more depth in Chapter 4.

Compound nouns are discussed in Christian Brothers (1999), along with certain idiomatic constructions with the copula. Stenson (1981) also discusses idiomatic constructions with the copula, as well as verb+object constructions, many of which fall under the umbrella of support verb or light verb constructions. These types of constructions are also explored by Bloch-Trojnar (2009 and 2010) and Bayda (2015 and 2016). Idiomatic constructions with the substantive verb and prepositions are briefly touched on in Ó Siadhail (1991). These types of MWEs overlap with idiomatic verb + preposition constructions, such as those detailed by Ó Domhnalláin and Ó Baoill (1975). A treatment of idioms is given in Ní Loingsigh (2016) and Ní Loingsigh (2021). A concordance of idiomatic expressions in the writings of Séamus Ó Grianna is given in Ó Corráin (1989). Finally, a review of several types of verbal MWEs are given by Veselinović (2006) and Nic Niallais (2020).

3.4 Conclusion

In this chapter, we present an overview of the Irish Language, including the syntax and a brief explanation of many of the parts of speech of the language. This overview is relevant to the creation of MWE categories explored in Chapter 4, as well as the review of linguistic analysis presented in the same chapter.

We also describe the resources which were useful to us in the tasks in automatic processing of Irish MWEs. The lexical resources are used in the building of a lexicon of Irish MWEs, described in Chapter 5. The raw parallel corpora we describe are used in the creation of MWE-annotated corpora (see Chapter 5), which are then employed in our experiments in Machine Translation in Chapter 6. The treatment of MWEs in the Irish UD Treebank informs our understanding of Irish MWEs, as described in Chapter 4. Finally, the research efforts that explore MWEs for Irish are incorporated into our typology of Irish MWEs, presented in Chapter 4.

Chapter 4

Categorisation of Irish MWEs

> *'Never laugh at live dragons, Bilbo you fool!' he said to himself, and it became a favourite saying of his later, and passed into a proverb.*

J.R.R. Tolkien, *The Hobbit*

4.1 Introduction

In order to systematically approach the problem of incorporating MWEs in NLP technology and processing them, many researchers (e.g. Mel'čuk et al. (1995), Sag et al. (2002), Baldwin and Kim (2010), Antunes and Mendes (2013), Ramisch (2015), Parra Escartín et al. (2018)) have opted to develop taxonomies for the classification of MWEs. The taxonomies are typically based on properties such as those discussed in Chapter 2. Broadly, MWEs are often distinguished on the basis of: lexicalised phrases and institutionalised phrases; fixed expressions, semi-fixed expressions and syntactically flexible expressions; and syntactic parts of speech.

While such taxonomies allow for a more holistic approach to MWE processing, many such works tended to focus mainly or exclusively on English MWEs (Parra Escartín et al., 2018). Furthermore the taxonomies were oftentimes in disagreement,

with different interpretations given to the terms used in their classifications. As such, research on MWEs in languages other than English can struggle to adapt these taxonomies to the specific features of the languages.

To combat this focus on English language, many linguistically-informed frameworks for NLP tasks (e.g. parsing (de Marneffe et al., 2021)) have adopted a universal framework, one that attempts to simplify across cross-lingual differences. This has the unfortunate trade-off of losing some of the precision and language-specific features which represent important differences between language descriptions. This is particularly noticeable when it comes to MWEs, as these idiosyncratic constructions frequently break regular language rules, and are difficult to capture with general, language-agnostic criteria.

In this chapter, we address **RQ1**, by describing a typology of Irish MWEs, which attempts to align with the language-universal typologies that exist in the PARSEME framework and the Universal Dependencies framework, while also complying with linguistic analysis carried out on the Irish language. The categories listed in this chapter are not intended to be a comprehensive grammatical description of all idiomatic constructions in Irish. Instead, these categories represent MWEs that were found to be both frequent in Irish texts and lexical resources (see Chapter 3), as well as frequently discussed in both the Irish and multi-lingual discourse on idiosyncratic constructions.

The initial approach that we take is to broadly categorise Irish MWEs into *verbal* MWEs and *non-verbal* MWEs. The categories of verbal MWEs were chosen to align with the PARSEME Annotation Guidelines 1.2, as per the introduction of the Irish dataset into that shared task. Where we found the annotation guidelines disagreed with the Irish MWEs in the data, we highlighted areas where the guidelines may need to be extended, or where further research might reveal how Irish can be more closely integrated into this universal analysis. Non-verbal categories[1] have some

[1]The verbal MWE category of `compound:prt` in the UD treatment also overlaps with our definition of verb particle constructions.

overlap with those covered by the Universal Dependencies annotation guidelines, particularly with the compound constructions and fixed expressions, however, there are some deviations from this analysis.

Finally, we present some difficult cases, which challenge our understanding of how classification of Irish MWEs can overlap with multilingual frameworks. The decisions we have arrived at may be revisited again in future work.

4.2 Existing Analyses

Before we describe the typology of Irish MWEs applied in our research, we first examine which types have been discussed in the literature. Some of the terms for constructions vary depending on the body of research, and there is considerable overlap between different typologies.

4.2.1 MWEs in the Irish Linguistics Literature

MWEs remain an under-explored topic in Irish, and despite the growing body of work exploring the linguistic structures and features of the language, there remains a need for a comprehensive survey of the idiosyncratic structures that are particular to Irish. The focus of this section is to provide an overview of relevant literature, and present the types of MWEs that are discussed in that literature. The typologies introduced by other authors in the field of Irish NLP informed the decisions we made in our categorisation of Irish MWEs. However, each of the typologies or discussions listed below were created to serve the particular needs of the author and their work. As such, we choose not to attempt to synthesise all of these typologies, rather, we use this body of research as a supportive framework when constructing a typology for our particular needs, i.e. processing Irish MWEs automatically for NLP purposes.

One of the earlier explorations of Modern Irish grammar comes from the Christian Brothers, who published a reference grammar book covering topics of morphology and syntax (Christian Brothers, 1960), and an updated version some time later

(Christian Brothers, 1999). Of the discussion undertaken in their work, formulaic constructions with the copula are briefly discussed, such as the COP + *le* construction to denote ownership (*Is liomsa é* (COP with-me-EMP it) 'It is mine') and idiomatic psychological predicates formed by COP + adjective + preposition (often *le* 'with', as with *Is fearr liom tae* (COP good-COMP with-me tea) 'I prefer tea').

The phenomenon of compounding in Irish is also briefly described, mostly focusing on the construction of single-word compounds through the concatenation of two words or affixes. Some multi-word compounds are also listed (e.g. *bolg le gréin* 'sunbathe', *tabhairt suas* (give-VN up) 'surrendering', *uisce faoi thalamh* (water under ground) 'conspiracy'). These constructions are not analysed beyond a note regarding hyphenation, where such constructions can be hyphenated when using them in a suffix or adjective role (e.g. *uisce-faoi-thalamh baolach* 'dangerous conspiracy').

Ó Domhnalláin and Ó Baoill (1975) provide a collection of the more common verb + preposition combinations in Irish along with their meaning in usage. These constructions are each considered by the authors to be a unit with a specific meaning ("*saghas aonaid a mbíonn ciall faoi leith leis*" 'a type of unit that has a specific meaning with it'), a description that borders on idiomaticity. Although not every verb + preposition construction is strongly semantically idiomatic, the use of the preposition carries some semantic weight in each construction, and is considered an essential part of the construction. The discussion is largely limited to nine common verbs, based on the analysis of their use in Ó Huallacháin and Department of Education (1966): *bain* 'extract/dig out', *cuir* 'put', *déan* 'make/do', *gabh* 'take', *lean* 'follow', *lig* 'let/allow', *tabhair* 'give', *tar* 'come', and *téigh* 'go'. There also follows a brief list of 46 less-common verbs with the prepositions that were observed in Ó Huallacháin and Department of Education (1966). The constructions are not analysed on a macro level, instead, each construction is presented with the compulsory arguments (either lexicalised components or part-of-speech tag) with optional components included, and a brief discussion of the construction is presented with examples. This collection is a useful resource for building MWE-aware resources,

however it does not offer a comprehensive categorisation of such verb + preposition constructions.

A study of Irish syntax and features of the Irish language is presented by Stenson (1981), with some consideration made towards idiomatic constructions. The discussion includes a deeper exploration of the idiomatic copular constructions mentioned in Christian Brothers (1999) above, as well as certain frozen copula + predicate constructions like *Is maith liom (tae)* (COP good with-me tea) 'I like tea' and *Ba cheart duit (carr a cheannach)* (COP-PA right for-you car PART buy-VN) 'You ought to buy a car'.

Idiomatic constructions of a verb + object are also discussed. These constructions are formed from a small class of verbs (e.g. *cuir* 'put', *tabhair* 'give', *déan* 'do', *bí* 'be' (substantive)), along with a noun phrase or prepositional phrase, which carry most of the semantic weight.[2] Examples of constructions provided include *Thug Teresa cúnamh di* (give-PA Teresa help to-her) 'Teresa helped her', *Tá Jennifer in ann é a dhéanamh* (be Jennifer in there it PART do-VN) 'Jennifer is able to do it', and *Chuir Sheila ina luí air é a dhéanamh* (put-PA Sheila in-his lie-VN on-him it PART do-VN) 'Sheila persuaded him to do it'. It is argued that this frozeness displayed by such constructions varies, with some constructions displaying more syntactic freedom than others. Stenson notes the trend that the less semantically opaque the construction, the more freedom there is to insert various noun phrases in such constructions, i.e. they display higher degrees of productivity. Similarly, many of these idiomatic constructions do not display the same level of syntactic flexibility that is shown by constructions of similar non-idiomatic elements. See the grammaticality of Example 60 versus the questionable grammaticality of Example 61.

(60) *Bím i gcónaí in ann é a dhéanamh*
be-I always in there it PART do-VN
'I am always able to do it'

[2]This type of construction is often referred elsewhere as a 'light verb construction'.

(61) ? *Cuirim i gcónaí ina luí air é a dhéanamh*
put-I always in-his lie-VN on-him it PART do-VN
'I am always persuading him to do it.'

Another type of idiomatic construction discussed by Stenson are idiomatic constructions expressing psychological states with the substantive verb *bí* + NP_i + *ar* 'on' / *ag* 'at' + NP_j to indicate possession. For example, *Tá aiféala orm* (be sorrow on-me) 'I am sorry', and *Tá súil agam* (be hope at-me) 'I hope'.

Light verb constructions are also explored for Irish by Bloch-Trojnar (2009) and (2010). Examples include English constructions such as 'make a decision', and 'take a chance'. Irish examples provided by the authors include *tabhair tacaíocht* 'give support', and *déan obair* 'do work'. Bloch-Trojnar focuses on light verb constructions containing a verbal noun, such as *déan athrú* 'make a change' and *tabhair léim* 'give a jump'. Bloch-Trojnar examines the semantic constraints on such constructions, arguing that these constructions are a means of specifying telicity in Irish verbs (i.e. providing a specific endpoint to a verbal action).

Idiomatic constructions in Irish are also explored by Bayda (2015 and 2016), including copular constructions (e.g. *is maith le* 'is good with/like'), light verb constructions (e.g. *cuir iontas ar* 'surprise', *cuir i ngléas* 'repair/put in order') and verb + preposition constructions (e.g. *bain as* 'extract'). Bayda highlights how these constructions overlap, as copular constructions and light verbs both rely on the noun to carry much of the semantic weight in the construction, while also often selecting for a preposition that contributes semantically. Constructions *tabhair ar* and *tabhair do* can each be formed into light verb constructions, but the selection of preposition depends on the semantic roles of the noun, such as with the light verb constructions *tabhair aird ar* 'give attention on' and *tabhair aire do* 'give care to'.

Ní Loingsigh (2016) and (2021) provides an analysis of a number of idioms collected from the published works of Canon Peadar Ó Laoghaire, a prolific Irish author of the late nineteenth and early twentieth century. An analysis of the extracted idioms reveals these constructions belong to a variety of syntactic classes, with the

following syntactic constructions noted in the data:

1. Irreversible binomial idioms

   scéal ná duain (lit. neither story nor poem/song) 'neither tale nor tidings'

2. Idioms beginning with a compound preposition

   ar muin na muice (lit. on the pig's back) 'very happy/content'

3. Idioms beginning with a verbal noun

   ag crith ina chraiceann (lit. shaking in one's skin) 'very afraid'

4. Idioms containing open function slots

   ó bhaithis go bonn (lit. from the top of the head to the sole of the foot) 'completely covered'

Other features of these idioms noted by Ní Loingsigh are their inclusion of numerals (e.g. *ceithre shaol duine a fháil* (lit. to get four lives of a person) 'to live very long'), somatic idioms (e.g. *faoina fhiacla* (lit. under his teeth) 'mutter something'), idioms using colours (e.g. *ina chogadh dhearg* (lit. in its red war) 'a ferocious battle'), and fossilised idioms (e.g. *ag stracadh na bpíobán as a chéile* (lit. tearing the pipes out of each other) 'fighting each other').

Veselinović (2006) discusses many Irish verb constructions, including idiomatic verbal constructions, drawing comparisons between the development of such constructions from Early Modern Irish and similar developments in Early Middle English during the same period. A typology of verbal constructions in Irish is also presented, each formed from the verb *cuir* meaning 'put'. The following idiomatic constructions are presented:

1. Prepositional verbs

   Cuir (put) + *le* (with) = 'add to'

2. Phrasal verbs

   Cuir (put) + *amach* (out) = 'vomit'

3. Phrasal verbs with reduced valency

   Cuir (put) + *as* (out of) = 'put out/extinguish'

4. Phrasal prepositional verbs

   Cuir (put) + *suas* (up) + *le* (with) = 'put up with'

5. Prepositional support verb constructions

   Cuir (put) + *ar* (on) + *ceal* (cancellation) = 'cancel'

6. Prepositional support verb construction with object

   Cuir (put) + *ceist* (question) + *ar* (on) = 'question'

7. Support verb constructions with object

   Cuir (put) + *fuil* (blood) = 'bleed'

Finally, a classification of verbal MWEs in Irish is described by Nic Niallais (2020), who identifies six categories of verbal MWEs, based on syntactic constructions. Nic Niallais describes verbal MWEs, or *briathra frásacha* 'phrasal verbs', as a verb that is formed through a combination of a verb and at least one other word. This definition aligns with that given by Matthews (1997), namely "'any combination of two or more words that is treated as, or as equivalent to, a verb: e.g. take pictures of... might be seen as an equivalent to photograph". The six categories are as follows:

1. Verb + preposition

   Bain (take from) + *faoi* (under) = 'settle'

2. Verb + directional adverb + preposition

   Cuir (put) + *isteach* (in) + *ar* (on) = 'interrupt'

3. Verb + directional adverb

   Cuir (put) + *amach* (out) = 'guess'

4. Verb + noun + preposition

   Tabhair (give) + *iarradh* (attempt) + *ar* (on) = 'attempt/have a go at'

5. Verb + preposition + noun

   Cuir (put) + *faoi* (under) + *cos* (foot) = 'oppress/put down'

6. Verb + preposition + noun + preposition

   Teigh (go) + *i* (in) + *gleic* (grapple) + *leis* (with) = 'grapple with'

While there is considerable overlap in these MWE analyses offered above, it is evident certain authors have chosen to focus their analysis on particular aspects of such idiosyncratic constructions. Our proposed typology should be sufficiently detailed to encapsulate each of these idiosyncratic constructions described above, and sufficiently broad to allow for coverage of any MWEs encountered in the data. Following from this summary of MWEs explored in the Irish literature, we now turn our discussion of MWEs to their treatment on a multilingual level.

4.2.2 PARSEME Annotation Guidelines

The PARSEME network is described in Chapter 2. To restate, this was an initiative aimed at improving multilingual efforts in intelligent text processing (machine translation, information retrieval, question answering, text summarisation, etc.) An outcome of this network was the development of cross-lingual guidelines that facilitate the consistent annotation of verbal MWEs across many languages.

The following categories were used for Edition 1.1 of the PARSEME shared task in the automatic identification of verbal MWEs (Ramisch et al., 2018):

1. Universal Categories (these categories were found to be valid for all participating languages):

   (a) Light Verb Constructions (LVC), consisting of a head verb and a noun object, with the semantics of the construction contained in the object. This category is divided into two subcategories:

      i. Fully light verb constructions (LVC.full) where the semantics of the head verb is completely bleached of meaning (e.g. 'make a decision')

ii. Semi-light verb constructions (LVC.cause) where the head verb contributes only a causative meaning to the accompanying noun, i.e. the subject of the verb is the causative agent of the noun (e.g. 'the music **gives** me a **headache**')

(b) Verbal Idioms (e.g. 'A little birdie told me')

2. Quasi-Universal Categories (these categories were present in a majority of participating languages, but not all):

   (a) Inherently Reflexive Verbs (IRV) (e.g. *se laver* 'to wash oneself')

   (b) Verb Particle Constructions (VPC), divided into two subcategories:

      i. Fully non-compositional VPCs (VPC.full) where the particle completely changes the meaning of the verb (e.g. 'let on' (pretend))

      ii. Semi non-compositional VPCs (VPC.semi) where the particle causes a change in meaning to the verb that is not directional or spatial, but the new meaning is semi-predictable (e.g. 'bring up')

   (c) Multi-Verb Constructions (MVC) (e.g. 'make do')

3. Language Specific Categories (only one language specific category was introduced in the latest version of the shared task, for Italian):

   (a) Inherently Clitic Verbs (LS.ICV) (e.g. *entrarci* 'to be relevant to something')

4. Experimental Optional Category

   (a) Inherently Adpositional Verbs (IAV) (e.g. 'put up with')

We apply these categories to Irish where appropriate, as discussed in Section 4.3 below.

4.2.3 Categories of MWEs in the Irish UD Treebank

The MWEs considered in the UD framework (`compound`, `fixed`, and `flat`) are those which most clearly demonstrate syntactic idiomaticity. Given how the primary analysis of a treebank is syntactic, it is appropriate to focus on this level of idiomaticity. However it is impossible to completely disregard semantic idiomaticity, as there is an overlap of the two levels for many MWEs, with the majority of MWEs displaying some level of semantic idiomaticity, and indeed for certain categories of MWEs annotated in UD, semantic idiomaticity is used as a distinguishing criteria to determine just how closely some components attach to others.

We briefly summarise the MWE labels applied to the Irish UD Treebank and the criteria for their annotation, as presented by McGuinness et al. (2020).

4.2.3.1 Compounds

The `compound` label is intended to be applied to systematic compounding constructions, with the intention that each language determine how best to apply this label. McGuinness et al. (2020) performed a survey of the use of this label in four languages (French, Spanish, Chinese and English). This survey revealed that the usage varies not only between languages, but also between differing treebanks, with some languages choosing to apply the label in very rare cases, such as with hyphenated words that had been split during tokenisation in the UD-French-Spoken treebank (e.g. *outre-mer* 'overseas').

The subtype label `compound:prt` is intended for use with particle verbs to attach the particle to the head verb. There are a number of languages in the UD project which do not demonstrate particle verbs (e.g. French), however this label is applicable to Irish. The guidelines do not offer comprehensive tests to determine whether a dependent is a particle, instead the criteria of semantic idiomaticity is applied in the IUDT guidelines, where a particle is considered an integral part of the construction in order to conserve the meaning. Examples include *tabhair faoi* (give under) 'undertake', and *cuir suas* (put up) 'provide'.

Semantic idiomaticity is also used as one of the criteria for determining usage of the `compound` relation. In Irish, compounding occurs when either a word or an affix attaches to another word to generate a new word or multiword construction with a shift in meaning. Multiword compounds, which are pertinent to our work, can be formed through the addition of an adjective or a noun, and include certain non-compositional constructions such as *mac tíre* (son of land) 'wolf' and *mac léinn* (son of learning) 'student'.

To distinguish between these compound nouns and compositional noun phrases, three tests were devised by McGuinness et al.:

1. If the definite article is omitted from the construction, it may indicate that the construction is not merely a compositional construction (compare *deireadh seachtaine* 'weekend' with *deireadh* ***na*** *seachtaine* 'the end of the week')

2. If the construction contains a cranberry word[3] it is a strong indicator that the construction is a compound. Constructions containing *déag* meaning 'teen' are almost exclusively numeral phrase, as with *cúig déag* (five teen) 'fifteen', and are annotated as compound.

3. If the meaning of either noun is determined to have significantly changed (i.e. semantic idiomaticity), this is an indicator that the construction is a compound.

This final test has been noted to be challenging to apply, as compositionality tends to occur on a spectrum, and words can adopt new meanings in different contexts. To demonstrate this, a poll was conducted between six annotators, who were supplied with a range of candidate compounds and asked to vote whether the construction was fully-compositional (not semantically idiomatic), semi-compositional or non-compositional (semantically idiomatic). Where more than half the annotators voted for a candidate to be fully-compositional, the candidate was annotated

[3]A cranberry word is defined in Chapter 2 and refers to a word that occurs only in a specific construction.

using the standard `nmod` relation, otherwise, the `compound` relation was applied.

4.2.3.2 Fixed

The `fixed` label is intended to capture immutable constructions that behave as function words or short adverbials and have no internal syntax. They are used in the IUDT to capture compound prepositions (e.g. *in aice* 'beside', *le haghaidh* 'for'), adverbial modifiers (e.g. *ar ais* 'back', *chomh maith* 'as well'), and determiners (e.g. *seo caite* 'last') among others.

4.2.3.3 Flat

The `flat` relation is intended to capture headless, semi-fixed constructions with no internal syntax, such as titles, foreign strings, and dates. The two subtype labels are `flat:name`, intended for capturing personal names, including honorifics (e.g. *Pádraig Mac Piarais* 'Patrick Pearse', and *An tUasal Way* 'Mr Way'); and `flat:foreign`, intended for the annotation of linear sequences of foreign text incorporated within the sentence (e.g. *go raibh sé* **cut off with a shilling** 'that he was cut off with a shilling').

The `flat` relation is also used to capture days of the week (e.g. *Dé Luain*, 'Monday') and months (e.g. *Deireadh Fómhair* 'October'). All names, regardless of origin, are annotated with `flat:name` (e.g. *an tUasal* Durkan, Robert de Niro, etc.), while titles in a foreign language are annotated with `flat:foreign` (e.g. '*Tristan Und Isolde*', 'Entering Jerusalem'). MWEs in a foreign language are also annotated with `flat:foreign` (e.g. *vice versa*).

Of note, McGuinness et al. (2020) cite their use of this label to include all proper noun strings regardless of whether such strings displayed a flat structure or could be syntactically parsed (e.g. *Choiste Turasóireachta na Gaillimhe* 'Galway Tourism Board'). This use has been modified as of version 2.7 of the IUDT, where such constructions were instead tagged with a Named Entity feature. This decision is discussed further in Section 4.4.3, as it has ramifications for our own categorisation

efforts.

4.3 Verbal MWE Categories in Irish

Following from the analyses presented above, we now take steps to creating the first typology of Irish MWEs for the purposes of applications in NLP. We begin with a discussion of verbal MWEs.

Many of the categories for verbal MWEs originated from categories annotated in the PARSEME annotated corpus of Irish verbal MWEs. One difference we highlight in our approach to categorisation is our use of the hierarchical tests in the PARSEME annotation guidelines, which include structural tests to identify which category-specific tests to apply. We discuss how our approach differs from that of PARSEME in Section 4.3.6.

Of the categories discussed in Section 4.2.2, five were identified as existing in Irish, and were applied in the annotation work described in Chapter 5. These five categories are Light Verb Constructions (e.g. *déan dearmad ar* 'to forget'), Verbal Idioms (e.g. *caith i ndiaidh* 'to hanker after'), Verb-Particle Constructions (e.g. *lig amach* 'to let out'), Inherently Reflexive Verbs (e.g. *iompair féin* 'to behave oneself') and Inherently Adpositional Verbs (e.g. *cuir as* 'to deprive').

Upon further investigation, the category of Inherently Reflexive Verbs was determined to be too rare in the data for consideration as a category in our typology, and was removed. Additionally, MWEs involving idiomatic constructions with the copula are sometimes annotated in the PARSEME guidelines as Verbal Idioms, however, we have elected to define a category for these constructions separately.

4.3.1 Light Verb Constructions

Light verb constructions (LVCs) have been introduced in Section 4.2.1. We adopt the definition offered by PARSEME for these vMWEs: constructions formed by a verb, v, and a (single or compound) noun, n, which contributes most of the semantic

information in the construction. Importantly, the noun must be a predicative noun (i.e. a noun with semantic arguments) that refers to an event or a state, and the semantic arguments must be shared by the head verb. For example, the noun *coir* meaning 'crime' is predicative, as the concept entails a do-er of the crime, and refers to an event. The noun *crosfhocal* meaning 'crossword' is not predicative, as there is no semantic agent implied. As such, we consider *déan coir* to be a `LVC`, but not *déan crosfhocal*. The noun n may also directly depend on v or can be introduced by a preposition, as with *tabhair* ***faoi*** *deara* (lit. take under notice) 'notice', and *cuir* ***le*** *cuthach* (lit. put to rage) 'enrage'.

PARSEME offers a further distinction between full `LVCs` and causative `LVCs`. Constructions where v's syntactic subject is n's semantic argument are full `LVCs` and annotated as `LVC.full`, while constructions where the subject of v is the cause or source of the event or state expressed by n are annotated as `LVC.cause`. Examples 62 and 63 illustrate full `LVCs` in Irish, and Examples 64 and 65 illustrate causative `LVCs` in Irish. However, we do not distinguish between these types of `LVCs` in our categorisation of this type of MWE.

(62) *Caith vóta*
cast vote
'cast a vote'

(63) *Déan géarleanúint ar*
do persecution on
'persecute'

(64) *Déan dochar*
do harm
'harm'

(65) *Cuir tús le*
put start to
'start/put a start to'

As can be seen from the examples above, the noun n may require an additional preposition (Examples 63 and 65) or may not (Examples 62 and 64). This additional preposition is discussed further in Section 4.3.3.

4.3.2 Verb Particle Constructions

Verb particle constructions (VPCs) consist of a verb, and a dependent intransitive particle where the particle causes a significant shift in meaning in the verb. In Irish, this particle is often homographic with directional adverbs (e.g. *amach* 'out' or *suas* 'up'), but sometimes it can look like a preposition (e.g. *as* 'off/from'). The important distinction between an adverb and a particle is that the addition of said particle lends a non-compositional meaning to the verb.

In the PARSEME annotation scheme, two levels of non-compositionality are considered. If the addition of the particle lends a meaning that is fully non-compositional, it is annotated as `VPC.full`, (Examples 66 and 67), while a meaning that is semi-non-compositional is annotated as `VPC.semi` (Examples 68 and 69).

(66) *Tabhair amach*
give out
'complain'

(67) *Caith anuas ar*
throw down upon
'condescend to/belittle'

(68) *Glan suas*
clean up
'clean up'

(69) *Cuir isteach ar*
put in on
'Apply'

As with the `LVC` category before, some instances of `VPCs` in Irish require an additional preposition to give this particular sense (Examples 67 and 69, while others do not (Examples 66 and 68). The addition of this extra preposition is discussed in Section 4.3.3.

4.3.3 Inherently Adpositional Verbs

Inherently adpositional verbs (IAVs), otherwise known as prepositional verbs, consist of a verb and a dependent prepositional phrase, with the preposition considered an integral (lexicalised) component of the construction, i.e. "it cannot be omitted without markedly altering the meaning of the verb" (as per the PARSEME annotation guidelines). This construction occurs frequently in Irish, as reflected in the extensive collection of such constructions by Ó Domhnalláin and Ó Baoill (1975). Examples 70 and 71[4] illustrate these constructions.

(70) *Buail le*
hit with
'meet'

(71) *Lig faoi*
let under
'settle oneself'

In the PARSEME guidelines `IAV` is considered an optional and experimental category, and the guidelines require this construction to be annotated only as a final step. The reason being that vMWEs can themselves form part of the `IAV` construction, as with the `IAV` *cuir suas le* (put up with) 'endure', which contains a `VPC` *cuir suas* (put up) 'put up', and with the `LVC` *tabhair tacaíocht do* (give support to) 'support' which contains a `LVC` *tabhair tacaíocht* 'give support'.

4.3.4 Copular Constructions

Copular Constructions (CCs) are not considered a category in the PARSEME annotation guidelines. According to the UD annotation guidelines, the head of a copular construction is not the copula, but the predicate. This means that the canonical form of a copular construction does not have a verbal head, and thus it cannot be qualified as a verbal MWE. However, we decided to include these constructions

[4]This expression bears some similarities to the controversial category of Inherently Reflexive Verbs (IRVs) which is discussed further in Section 4.5.

in our categorisation, as they feature frequently in the Irish data, where frozen idiomatic constructions can be formed from the copula and a substantive noun (Example 72), the copula and a preposition (Example 73),[5] or the copula and a noun and a prepositional phrase (Example 74).

(72) ***B'fhéidir*** *go* *bhfuil sé te*
COP possible PART is it hot
'It's possible that it's hot'

(73) ***Is*** ***le*** *James an cupán*
COP with James the cup
'The cup belongs to James'

(74) ***Is*** ***maith leat*** *tae*
COP good with-you tea
'You like tea'

Discussion regarding the inclusion of such constructions in future editions of the PARSEME annotation guidelines are underway at the time of this research work, and may be added in the near future to the annotation scheme.

4.3.5 Verbal Idioms

Verbal idioms (VIDs) are idiomatic constructions with at least two lexicalised components, including a verbal head and at least one dependent. In cases where the construction has only one dependent, it is necessary to ensure the construction is not belonging to a different MWE category. However, constructions with more than one dependent can only be considered `VID`. Dependents can vary by POS category, such as nominal dependents (Example 75), prepositional phrases (Example 76) and adjectival phrases (Example 77). In cases where the POS is nominal or prepositional, tests must be applied to ensure the construction is not a LVC or IAV construction.

(75) *Déan* ***seacht míle*** ***dícheall***
make seven thousand best efforts
'Do one's utmost'

[5]In cases such as Example 73, the predicate is selected by the construction, but it is not a lexicalised component of the construction.

(76) *Snámh* ***in aghaidh easa***
swim against a waterfall
'Do something that's a waste of time'

(77) *Bheith* ***dubh dóite*** *de [rud]*
be black burned of [thing]
'be sick of something'

Dependents can serve various sentential roles, such as a subject (Example 78), a direct object (Example 79), or both (Example 80). Furthermore, the number of dependents can vary (See Example 81, which has one dependent (*i mo* 'in my') and Example 82 which has three dependents (*do lámh* 'your hand', *i mbéal an mhadra* 'in the dog's mouth', *agat* (at-you) 'have')). While constructions with just one dependent may be considered as another category, constructions with more than one dependent are always considered VIDs.

(78) *Bí* ***seacht gcúraimí an tsléibhe*** *ar [duine]*
be seven cares the mountain-GEN on [person]
'Be seven cares of the mountain on someone/be very busy'

(79) *Déan* ***cat is dhá eirbeall*** *air*
make a cat and two tails on him
'Make a cat with two tails/do miracles'

(80) *Sceitheann* *fíon* *fírinne*
spews wine truth
'Wine reveals the truth'

(81) *bí* *i mo* [gairm]
be in my [profession]
'be a [profession]'

(82) *Ná bíodh* *do lámh* *i mbéal an mhadra* *agat*
NEG be-HABITUAL your hand in mouth the dog-GEN at-you
'Don't be having your hand in the dog's mouth/Don't invite trouble'

Sentential expressions with no open slots (oftentimes proverbs) are also considered VIDs. Such VIDs include Examples 83 and 84. Some constructions which have no clear verbal head, as with coordinated verbs, may be considered VID if there is some level of idiosyncrasy (Example 85).

(83) *Maireann croí éadrom i bhfad.*
lives heart light in length
'A light heart lives long.'

(84) *Is fearr glas ná amhras.*
COP better a lock than doubt
'Better to be safe than sorry.'

(85) *Déan mórán agus can beagán.*
do much and say little
'Actions speak louder than words.'

This category is particularly challenging to identify and annotate given the extent of variability in the structure.

4.3.6 Observations on Verbal MWEs

PARSEME makes use of hierarchical tests in their annotation guidelines. However, we do not employ such hierarchical tests, instead focusing on attributes displayed by Irish MWEs to distinguish between the categories. The exception to this general rule is the decision to annotate `IAV` constructions after annotating `LVC` and `VPC` type MWEs. Additionally, one of the criteria for `VID`-type MWEs is that it does not pass the test for any other category of MWE, so structurally, tests for `VID` should be applied only after all other tests for verbal MWEs have been applied.

As a note on terminology, the term 'phrasal verb' is often used synonymously with the VPC category described above (Section 4.3.2), as in Stenson (1981); Veselinović (2006) and Lynn (2016). However, this term has also been used by Uí Dhonnchadha (2009) in the same sense as the IAV category described in this section, while in Nic Niallais (2020), the term is applied to a large number of verbal constructions, including the LVC category described in Section 4.3.1. It is important to note which sense is being applied to the term in linguistic descriptions of these constructions.

4.4 Non-Verbal MWEs

The non-verbal categories we include in our typology are partly informed by the categories of the UD treebank (i.e. `fixed`, `compound` and `flat` relations, as discussed in Section 4.2.3), and draw heavily from the work of McGuinness et al. (2020). As a result, these categories are defined in large part because of their syntactic idiosyncrasy. Lexical, statistical, and semantic idiosyncrasy play a role in defining these categories also, however these criteria are often fuzzy, and this contributes towards the challenge of capturing these expressions.

4.4.1 Fixed Expressions

Fixed expressions (FEs) are a type of MWE commonly discussed in the literature, and are described in Sag et al. (2002) as a class of immutable expressions that "are fully lexicalised and undergo neither morphosyntactic variation... nor internal modification". As discussed in Chapter 2, MWEs can be broadly categorised through this feature of morphosyntactic fixedness, resulting in the distinctions between fixed, semi-fixed and flexible expressions. However, in order to distinguish constructions in this MWE category, we consider whether the fixed expression appears to be a unit of language: i.e. each token of the expression is a member of the same constituent, and the components are not productive and cannot be readily replaced by other tokens from the same semantic class. This definition of a unit of language is largely influenced by the concept of fixed expressions in the UD guidelines,[6] which was intended to capture certain idiosyncratic constructions or "fixed grammaticized expressions that behave like function words or short adverbials". Our definition is intended to rule out constructions with fixed syntactic properties but that are not considered a language unit, such as *ar an* 'on the', or syntactically fixed content constructions, such as *Oíche Shamhna* 'Halloween'.

Compound prepositions, which are described in Chapter 2 are examples of fixed

[6] `https://universaldependencies.org/ga/dep/fixed.html`

expressions in Irish, for instance, Examples 86 and 87.

(86) *i ndiaidh*
in after
'after'

(87) *os comhair*
over present
'in front of'

Fixed adverbial phrases are also considered fixed expressions, as shown by the semi-non-compositional Example 88. Fully semantically opaque expressions, such as Examples 89 and 90 are also considered fixed expressions, whereas idioms containing a verbal head are classified as VID.

(88) *Ceart go leor*
right PRT enough
'Ok'

(89) *Ar nós na gaoithe*
on manner the wind
'Like the wind/Quickly'

(90) *Idir dhá thine Bhealtaine*
between two fire May-GEN
'Between two May fires/Between a rock and a hard place'

Other MWEs in the fixed expression category include fixed question constructions such as Example 91, fixed foreign constructions 92, and certain idiomatic prepositional phrases, such as Example 93.

(91) *Cén fáth*
what reason
'Why'

(92) *De facto*
de facto
'De facto'

(93) *Faoin gcéad*
under hundred
'Percent'

4.4.2 Nominal Compounds

Nominal Compounds (NCs) are constructions that consist of a head noun and at least one dependent that forms a noun phrase with the head noun. This definition aligns with that of the `compound` label described in McGuinness et al. (2020). The dependent can be a noun (Example 94), an adjective (Example 95) or a prepositional phrase (Example 96).

(94) *Garrán préachán*
grove rooks-GEN
'Rookery'

(95) *Toradh leagtha*
fruit knocked-down
'Windfall'

(96) *Cur ar aghaidh*
put-VN forward
'Promotion'

Our use of the compound category varies slightly from that in the UD, as we also extend our coverage to fully compositional terms that are considered specialist language. As discussed in Chapter 2, such specialist language MWEs are associated with a particular domain, and have a precise meaning ascribed to them which may not be intuitive for a non-expert. In this way, we consider such terms to be pragmatically idiomatic, even if the meaning can be considered transparent. Examples 97 and 98 illustrate these specialist terminology MWEs.

(97) *aigéad sulfarach*
acid sulfuric
'sulfuric acid'

(98) *Cuardach le cuidiú gutha*
search with help of-voice
'Voice enabled search'

4.4.3 Named Entities

Named Entities (NEs) are a special category of nominal compounds that refer to the names of individual, recognised entities. This category also includes foreign expressions, although they differ from `FEs` in that the foreign expressions used must take the role of a noun phrase. The category includes the names of people (Example 99), months or dates (Example 100), place names (Example 101) and organisations or titles (Examples 102, 103 and 104).

(99) *An Taoiseach Micheál Martin*
the prime minister Micheál Martin
'The Taoiseach Micheál Martin'

(100) *Deireadh Fómhair*
end autumn-GEN
'October'

(101) *Baile Átha Cliath*
town ford-GEN wattled/hurdled
'Town of the wattled/hurdled ford/Dublin'

(102) *An Roinn Dlí agus Cirt*
the department law-GEN and right-GEN
'Department of Justice'

(103) *Cumann na nGalfairí Gairmiúla*
association the golfers-GEN professional
'Professional Golfers' Association'

(104) *GNU General Public License*
GNU general public license
'GNU General Public License'

As mentioned in Section 4.2.3, the IUDT has pared back the use of the `flat` label, where previously it was applied to proper noun strings with syntactic structure such as *Uachtarán na hÉireann* 'President of Ireland'. Aligning with this use of the `flat` label, we have opted to apply the `NE` label to constructions that do not have a flat structure, such as with Examples 101-104 above, so as to capture named entity information.

4.4.4 Institutionalised Phrases

Institutionalised Phrases (IPs) are described in Sag et al. (2002) as expressions that are statistically idiosyncratic. As discussed in Chapter 2, `IPs` are distinct from collocations in that `IPs` discount compositional phrases that are predictably frequent for non-linguistic reasons. While these expressions are not semantically idiomatic (non-compositional), their frequency in language creates a strong association between the concept and the expression. For instance, some `IP` constructions occur more regularly in language than a grammatical and equivalent construction (e.g. Example 105 vs Example 106), displaying syntactic idiosyncrasy.

(105) *Aire agus forcamás*
care and attention
'Care and attention'

(106) *#Forcamás agus aire*
attention and care
'#Attention and care'

With such expressions there is a difficulty in distinguishing between statistical idiosyncrasy and regular language composition. The concept of "lexical bundles" is introduced by Biber et al. (1999), as "sequences of words that commonly go together in natural discourse". Such items are largely formulaic constructions that occur frequently in language, and may intuitively seem like MWEs due to their strong association. In order to determine when to classify such construction as `IP`, we consider whether any of the elements of the construction can be substituted with another of the same meaning, and whether there is a tendency in general language use to prefer one construction over the other. This criteria is difficult to apply consistently however, and it may prove better to apply this category of MWE on the basis of statistical association measures (see Appendix B for more information on these measures).

4.5 Difficult Cases

Here we describe some of the difficulties encountered during the development and application of these categories, along with the practice that has been currently adopted in the annotation of such constructions. This section can also be considered a future work section, as some or many of these issues may need to be revisited in the future when more data is available.

4.5.1 Annotating IAVs

`IAV` type MWEs are challenging to define, particularly in distinguishing between idiomatic usage and the regular valency of the verb. For this reason, they were included as an optional category in the PARSEME annotation guidelines and many language teams chose not to annotate them when creating the annotated corpora for their language. As demonstrated in the collection of verb + preposition constructions by Ó Domhnalláin and Ó Baoill (1975), there are a number of common verbs in Irish that can adopt a wide range of meanings depending on context. The preposition in such constructions may be integral to the syntax of the construction, but it is difficult to ascertain for each construction whether the preposition is causing the shift in meaning, or whether the meaning of the verb is determined by its context.

Our current practice is to annotate IAVs where the meaning of the verb appears to be less dependent on particular pragmatic context of its usage, such as with *cuir* 'put' + *ar* 'on' → *cuir ar* 'bother', or *bain* 'take' + *le* 'with' → *bain le* 'relate to', both of which can be used in a very general sense with a wide range of noun phrases.

4.5.2 Extending MWEs with IAVs

`LVCs` often select for a specific preposition, with the construction never occurring without this preposition. It is common for these prepositions to be included as an integral part of the construction when they are being discussed or analysed in

the literature (Stenson, 1981; Bloch-Trojnar, 2009; Bayda, 2015), even when the required preposition does not lend an idiomatic meaning. Examples of this include Example 107 and 108.

(107) *Déan taighde* ***ar***
make research on
'Do research **on**/research'

(108) *Bain triail* ***as***
take test from
'Try'

This phenomenon also occurs with `VPCs`, as with Example 109 below.

(109) *Cuir isteach* ***ar***
put in on
'Bother'

When annotating these constructions it was challenging to decide whether to extend the label with an `IAV` categorisation. This aligns with the issue of `IAV` annotation mentioned in Section 4.5.1. We ultimately made the decision to annotate such constructions with both labels, i.e. for Example 107, we annotated *déan taighde* 'do research' as `LVC` and the entire construction *déan taighde ar* as `IAV`. This decision may be revisited in future versions of the corpus.

Addition of IA-LVC Category

The decision made above to annotate such constructions with both the `LVC` label and the `IAV` label was found to be insufficient. This approach led to applying the `LVC` label to constructions which are not full MWEs themselves, such as *Cuir teannadh* in Example 110, and *Bain sult* in Example 111,[7] as such constructions cannot occur without the preposition.

(110) ***Cuir*** *teannadh leis an dath*
put emphasis with the colour

[7]Note that *bain sult* can be used as an imperative statement without an oblique argument, as with the English version 'Enjoy!'. However, this can be treated as an ellipsis, where the preposition has been dropped.

'Emphasise the colour'

(111) ***Bain sult*** *as*
take enjoyment from
'Enjoy it'

On the other hand, labelling the entire construction with `IAV` loses the information that the noun is the carrying the semantic weight of the construction, in the same way a `LVC` does.

Our solution is to propose a category of Inherently Adpositional Light Verb Constructions (IA-LVCs), which aligns with the categories of **verb + noun + preposition** and **verb + preposition + noun + preposition** noted by Nic Niallais (2020) in Section 4.2.1.

4.5.3 Identifying Particles in VPCs

One of the challenges encountered was the discrepancy between the behaviours of prepositional particles and adverbial particles. We require consistent rules for deciding when a word is a particle or not for our annotation, and so examine the different behaviours of each here.

In English, particles are often homonymous with prepositions (though not always: e.g. *back, through*), although their behaviour is markedly different (Jackendoff, 2002). In terms of identification, particles of intransitive verbs are relatively simple to spot, as they can function as the only complement of the verb ('the spaceship blew **up**'). For transitive verbs, it is possible to attempt to restructure the sentence in order to identify whether the candidate particle forms a constituent with the verb or with the noun phrase object ('Katie looked **up** the answer/Katie looked the answer **up**' vs 'Katie walked **up** the lane/*Katie walked the lane **up**').

When addressing such constructions (termed 'phrasal verbs') in Irish, Uí Dhonnchadha (2009) suggests testing which words can be inserted between the verb and particle (in this case a preposition) to determine whether the particle is bound to the verb or not. For instance, Uí Dhonnchadha notes that the addition of an adverb

between the verb and particle in the idiomatic construction *d'éirigh le* (rise with) 'succeed' (Example 112) is acceptable (Example 113), but the addition of an NP subject (Example 114) or a prepositional adverb (Example 115) is not grammatical, and the construction loses its idiomatic meaning.

(112) *D'**éirigh** **le**is an mac léinn sa scrúdú*
rose with it the student in the exam
'The student succeeded in the exam'

(113) *D'**éirigh** go maith **le**is an mac léinn sa scrúdú*
rose well with it the student in the exam
'The student succeeded well in the exam'

(114) **D'**éirigh** an mac léinn **le**is sa scrúdú*
rose the student with it in the exam
? 'The student rose with it in the exam'

(115) **D'**éirigh** sa scrúdú **le**is an mac léinn*
rose in the exam with it the student
*'Rose in the exam with it the student'

When the particle is an adverb these insertion rules differ, such as with the directional adverb *amach* in the intransitive construction *tabhair amach* 'give out/complain' (see the ungrammaticality of Example 116) or in the transitive construction *leag amach* 'lay out' (see the ungrammaticality of Example 117).

(116) ***Thug** sé go feargach **amach***
gave he angrily out
*'He gave angrily out'

(117) ***Leag** sí go maith **amach** an bord*
lay she well out the table
*'She laid well out the table'

Applying the same intransitive test as with English particles to the construction *tabhair amach* results in the perfectly grammatical *Thug sé amach* 'He gave out/He complained'. Similarly, the particle of a transitive verb can move in Irish as well as English, such as *leag amach*, meaning both *leag sí amach an bord* and *leag sí an bord amach* are grammatical.

Given that the category of VPC was originally created for annotating verbal MWEs in the Irish corpus for inclusion in the PARSEME Shared Task, we have decided to adopt these tests as our metrics for determining when a word is a particle. As such, we limit our definition of particles to directional adverbs (e.g. *suas* 'up', *siar* 'westwards') or prepositions which follow the tests employed by PARSEME (e.g. *as* 'out', *faoi* 'under') in our categorisation of this MWE type.

4.5.4 Inherently Reflexive Verbs

Inherently reflexive verbs (IRVs) are a quasi-universal category that have been annotated in the PARSEME corpus of annotated vMWEs for Irish. An `IRV` consists of a verb v and a reflexive clitic $RCLI$ where either v never occurs without $RCLI$, or the omission of $RCLI$ causes a significant change in meaning. The reflexive pronoun in Irish is formed through the combination of personal pronoun + *féin*. There are very few constructions where the addition of this reflexive pronoun appears to cause a shift in meaning, such as *d'iompair mé* 'I carried' vs. *d'iompair mé mé féin* (carried I me self) 'I behaved myself'. Given how rarely this category occurs in the Irish data, we decided to exclude the category from our categorisation following the release of the PARSEME corpus of annotated vMWEs for Irish.

Of interest are certain verb + preposition constructions, where an inflected preposition implies reflexivity, such as with *bhailigh + sé + leis* (gathered he with-him) 'he removed himself/he left', or *bhain mé fúm ann* (took I under-me there) 'I settled myself there'. Our current practice is to annotate such constructions as `IAV`. However this remains an interesting question to explore for future work.

4.6 Conclusion

This chapter presents a typology of MWEs in Irish, the first such typology that attempts to categorise MWEs for the purposes of NLP. Our typology features categories selected from two universal frameworks for MWE categorisation: the

PARSEME annotation guidelines and the Universal Dependencies guidelines.

The categories fall under verbal and non-verbal types. The five verbal categories considered include four categories from the PARSEME annotation guidelines (Light verb constructions, verb-particle constructions, inherently adpositional verbs, and verbal idioms), as well as an extra category (copular constructions) which is not currently considered in the PARSEME annotation guidelines. The four non-verbal categories include two categories that are approximately equivalent to categories of MWEs in the UD guidelines (nominal compounds and fixed expressions), and one category that overlaps with the `flat` category of the UD guidelines (named entities), with some notable differences. We also introduce a new category (institutionalised phrases).

Some of the challenging aspects of applying these labels are addressed in the section on difficult cases. These highlighted issues have been flagged as areas of future work, the result of which may require updating and enhancing the current practice for the annotation of Irish MWEs. Our understanding of such constructions is likely to evolve as further annotation work reveals patterns in the frequency of such constructions, and how the capturing of such constructions impact downstream NLP applications.

Chapter 5

Resources for Irish MWEs

> *There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.*
>
> J.R.R. Tolkien, *The Hobbit*

As stated in Chapter 2, adequate language resources are essential for any task in NLP. The automatic processing of MWEs similarly require language resources that are sufficient for the task.

As discussed in Chapter 3, Irish is a minority and low-resource language, and as such, data and other underlying NLP resources are scarce, although there does exist a body of resources applicable to the task of MWE processing for Irish. In this chapter, we describe three resources explicitly created as part of our research for use in the processing of Irish MWEs, addressing once more **RQ2**.

5.1 Ilfhocail

In Chapter 3, we described a number of lexical resources for Irish. While many of these lexical resources are rich, extensive, and useful for linguistic analysis, language

speakers, or teachers, the format of these lexicons makes them less convenient for the purposes of NLP, and particularly for the task of MWE processing. In order to proceed with this task, we created a lexicon of Irish MWEs expressly for NLP purposes, by parsing the lexical resources at our disposal, and extracting multiword entries.

5.1.1 Extraction

In building this resource, we extracted MWEs from the lexical resources described in Chapter 3: Peadar Ó Laoighaire Idiom Collection, *An Bunachar Náisiúnta Téarmaíochta don Ghailge* (Téarma), *Líonra Séimeantach na Gailge* (LSG), *Pota Focal Gluais Tí* (Pota Focal), the English-Irish Dictionary (EID), the New English-Irish Dictionary (NEID), *Foclóir Gaeilge-Béarla* (FGB), and *An Foclóir Beag* (FB). The lexical resources come in a variety of formats, mostly `XML` documents. The exceptions are the Peadar Ó Laoghaire Idiom Collection, which was shared with us in `csv` format, and LSG, which was shared with us in `po` format. Téarma multiword entries had previously been extracted in related research work, and so we concatenated the text files containing n-grams of 2 or more words.

The extracted entries were then concatenated into a `csv` with the following columns: `GA-Head`, `GA`, `POS`, `EN`, `Source` and `ID`. `GA-Head` is a headword, included to facilitate searchability. This corresponds to the word that the dictionary entry was stored under, or the first word of the Irish MWE entry, where the lexical term was not stored under a single Irish word (e.g. the English-Irish Dictionary which is sorted using English headwords). The `GA` column contains the Irish MWE entry. The `Source` column consists of a string indicating the lexical source of the entry, so as to allowing filtering entries. `ID` is created by concatenating the source code with an integer to generate a unique string for each entry.

These latter four columns were each populated by either the extracted lexical item or generated, as with the `GA-Head` or `ID`. The two remaining columns, `POS` and `EN` are optionally populated columns, containing the POS information and the

English translation of the lexical entry where this information is included. The POS information that was extracted varied from some resources containing no POS label (Téarma and the Peadar Ó Laoghaire Idiom Collection), to broad level POS information (NEID) to including information such as transitivity, gender, or number (Pota Focal, English-Irish Dictionary, LSG, Foclóir Beag, FGB). English translations were extracted from all resources except LSG, the Peadar Ó Laoghaire Idiom Collection and FB.

Table 5.1 reports on the number of MWE entries extracted from each resource.

Resource	**Info**	**# MWEs**
New English-Irish Dictionary	GA POS EN	90,140
Foclóir Gaeilge-Béarla	GA POS EN	38,823
English-Irish Dictionary	GA POS EN	15,218
Pota Focal	GA POS EN	375
Líonra Séimeantach na Gaeilge	GA POS	8,995
Foclóir Beag	GA POS	771
Téarma	GA EN	137,944
Peadar Ó Laoghaire Idiom Collection	GA	420

Table 5.1: Resources used to build our lexicon, what information was extracted, and the number (#) of MWEs extracted from each resource. GA = Irish entry included; POS = POS entry included; EN = English entry included.

5.1.2 Cleaning

There were many instances of redundant or duplicate entries in the lexicon, due to duplicated MWEs in the resources. Additionally, the same lexical entry was duplicated within individual resources where the English translation or POS information differed. For example, the Irish verb+preposition construction *tabhair faoi* is translated in many ways in the NEID, including 'attack', 'take on', 'pursue', and 'attempt', while the NP *Cósta Ríceach* 'Costa Rican' is entered as both a noun, and an adjective. To reduce redundancy, we employed a deduplication method.

Entries that were distinguished on POS information were kept as separate entries (Example 118). MWE entries with different English translations were combined, as were MWEs from different sources. Example 119 shows how these English trans-

lations were combined. In cases where the English translation for an MWE was a sub-string of another English translation of the same MWE, the former translation was subsumed by the latter (Example 120).

(118) *Cósta Ríceach* (**ADJ**) 'Costa Rican'

Cósta Ríceach (**NOUN**) 'Costa Rican'

(119) *An beag is an mór* 'great and small'

An beag is an mór 'young and old'

An beag is an mór '**great and small; young and old**'

(120) *Breithlá Sona!* 'birthday'

Breithlá Sona! 'birthday (Happy Birthday!)'

Breithlá Sona! '**birthday (Happy Birthday!)**'

Following these cleaning steps, the lexicon was reduced from 292,686 entries to 201,795 entries.

GA-Head	GA	POS	EN	Source	ID
cú	*cú allta*	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	*muc mhara*	s.	porpoise	eid	141920eid
min	*min sáibh*	noun	UNK	lsg	140450lsg
min	*min sáibh*	UNK	sawdust	tearma	140451tearma

Table 5.2: Sample entries from the Ilfhocail lexicon displayed under their respective headings. The table shows issues that require addressing, such as non-unified POS tags, duplicate entries, and missing fields (UNK tokens inserted).

5.1.3 Manually Annotated Sample

To evaluate the quality of the lexicon, we semi-randomly selected,[1] examined and annotated 720 entries with 3 levels of MWE category: level of fixedness, 'generalness', and MWE category. Each of the sources was represented at least once in the random selection, with the exception of the Peadar Ó Laoghaire Idiom Collection.

[1]A random sample of entries were extracted, and then examined in batches of 20-50 entries at a time. In order to assess the quality of each of the sources that were present in the random sample, and to broaden the range of MWEs present, we preferred batches containing MWEs originating from sources other than Téarma during our manual inspection

This latter resource is a collection of idioms, so there is no need to assess the quality of the automatic extraction method for MWEs from this resource, as all of the entries are known MWEs.

The level of fixedness of the MWE refers to the potential for syntactic and morphological flexibility of the MWE, and includes fixed expressions (labelled *f*), semi-fixed expressions (labelled *s*) or non-fixed or flexible expressions (labelled *n*). The 'generalness' or domain-specificity category, as we described in Chapter 2, reflects the division of MWEs into *sublanguage MWEs* (or SL-MWEs) and *general language MWEs* (or GL-MWEs), based on the work of Savary et al. (2019b). SL-MWEs includes terminology and named entities that are specific to a domain, and whose precise meaning is defined by experts in that domain. GL-MWEs are MWEs that are used by the larger language community, and whose idiomaticity is not dependent on the context of a certain domain. The boundary between these two categories can be sometimes difficult to delimit. Finally, the MWE category assigns one of nine types of MWE (or a 'non-mwe' label) to the lexical entry, based on our typology of Irish MWEs outlined in Chapter 4.

5.1.3.1 Categories

The MWEs sampled from the corpus were roughly proportional with the MWEs extracted from each source, with about half (366/720) of the MWEs coming from Téarma, even with our selection bias for batches containing MWEs from sources other than Téarma. As such, terminology and NEs were highly represented in the sample. We decided that entries originating from the Téarma database would not be rejected as MWEs, even if considered non-idiomatic, due to the difficulty of delineating between some general and sub-language expressions.[2] During our manual inspection, we found 74 of the 720 entries were non-MWEs and were annotated as such. For purposes of comparison in the categories of fixedness and generalness, the non-MWEs entries were removed, and the remaining 646 entries are compared.

[2]We may revise this decision in future versions and analyses of the lexicon.

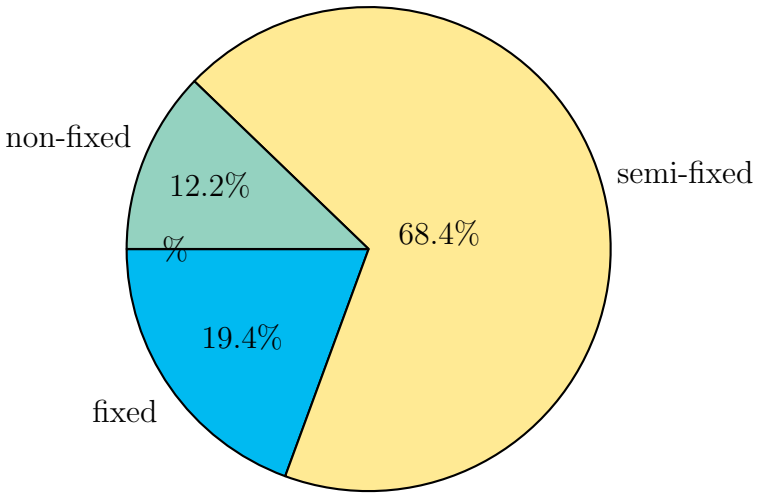


Figure 5.1: Breakdown of MWEs in annotated sample by level of fixedness.

Fixed, semi-fixed and non-fixed MWEs are described in Chapter 2. In brief, fixed MWEs cannot display any inflection or variation. We annotate morphologically flexible but syntactically fixed MWEs as semi-fixed, while MWEs that display syntactic flexibility are annotated as non-fixed. The breakdown for these three categories in the annotated sample is displayed in Figure 5.1. The highest proportion of MWEs were annotated as semi-fixed, which is not surprising, as the vast majority of lexical entries taken from Téarma are semi-fixed nominal compounds (NCs). Many NEs are annotated as fixed, as they do not allow for any morphological changes and should be treated as fixed units (e.g. *Muir Aidria* 'Sea of Adria', *commedia erudita*, *Comhairle Contae Thiobraid Árann Theas* 'South Tipperary County Council').

The breakdown of SL-MWEs vs GL-MWEs is shown in Figure 5.2. The large proportion of terminology is again reflected in the substantial number of SL-MWEs, with this number perhaps inflated due to our decision to retain all entries from the Téarma database as MWEs, while non-idiomatic entries from other sources were rejected.

Table 5.3 contains the number of MWEs per category in the manual sample,[3] with Figure 5.3 displaying the relative proportion of these categories. Named En-

[3]The total number of entries categorised actually sums to greater than 720, as 19 entries were found to have embedded MWEs, and both MWE categories were recorded in this table.

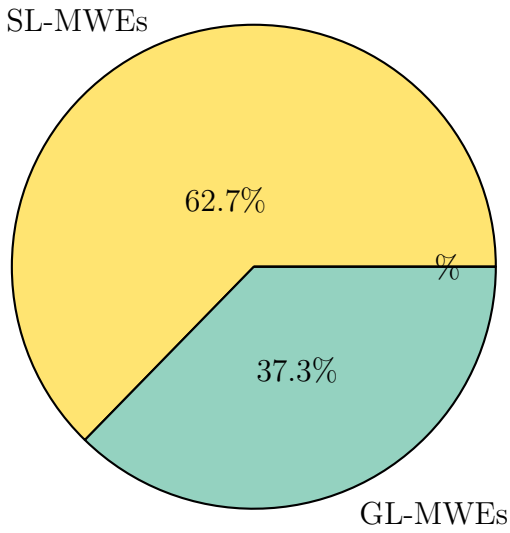


Figure 5.2: SL-MWEs vs GL-MWEs in annotated sample

tities and Nominal Compounds together make up 75% of the entries (excluding non-MWE entries). Notably, each of the nine categories explored in Chapter 4 is represented at least once in this sample.

The annotation process revealed some difficult-to-categorise entries. Some of these difficulties are described in our discussion of issues below.

5.1.3.2 Quality Issues

During the manual inspection, a number of issues with the lexicon were observed, and are listed below. These issues will be addressed in future versions of this lexicon.

Non-MWEs As shown in both Table 5.3 and Figure 5.3, there are a significant portion of non-MWEs in the lexicon (74 entries). These non-MWEs came from the NEID (15), the EID (21), and the FGB (36). There were also two instances from the LSG, and one entry from Téarma[4] that were not annotated as MWEs. Example 121 from EID is a single-token entry (*suaimhnithe* 'of reassurance'), which includes in the entry a non-idiomatic example of its use. Example 122 from FGB is an example of usage of the word *leasc* 'sluggish', which shows no idiomatic behaviour. Example 123 from the NEID was extracted as a multiword translation of the English entry

[4]The entry we rejected as an MWE was the verb + noun construction *éirigh torrach* 'become pregnant', as it did not fit any of our MWE categories, failing as both an LVC and a VID.

Tag	Category	Example	# MWEs
CCs	Copular Constructions	*is líth le* 'it pleases'	1
FEs	Fixed Expression	*ar nós an diabhail* 'like the devil'	24
IAVs	Inherently Adpositional Verbs	*maith do* 'forgive'	24
IPs	Institutionalised Phrases	*gruth agus meadhg* 'curds and whey'	24
LVCs	Light Verb Constructions	*déan fead* 'whistle'	55
NEs	Named Entity	*an Teach Bán* 'The White House'	55
NCs	Nominal Compounds	*bun cluaise* 'earlobe'	448
	Non-MWEs	*fear ard* 'a tall man'	74
VIDs	Verbal Idioms	*as an obair a thagann an fhoghlaim* 'practise makes perfect'	30
VPCs	Verb-Particle Constructions	*athraigh síos* 'shift'	4

Table 5.3: Categorisation of 720 MWEs, including non-MWEs.

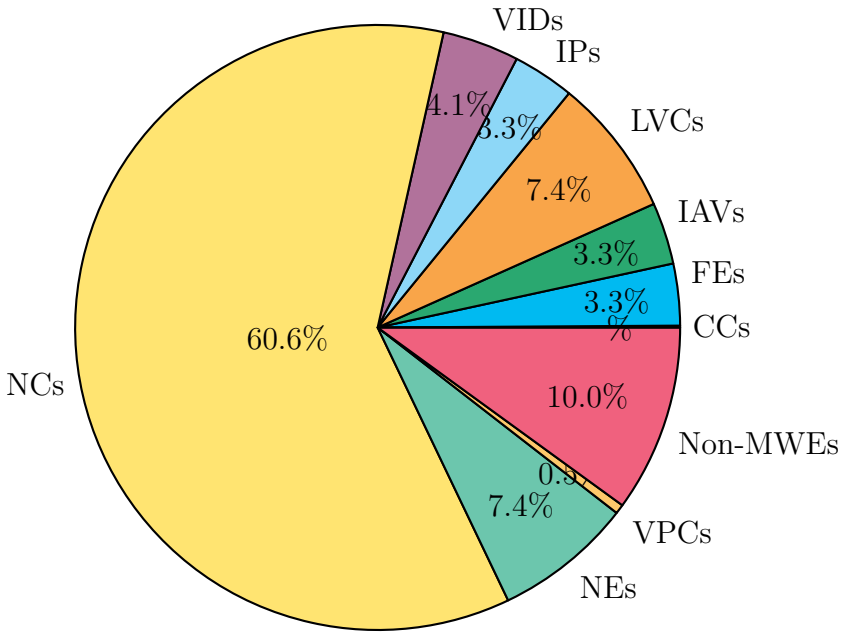


Figure 5.3: Breakdown of categories in annotated sample.

'giddily', but the translation in Irish is a regular construction of *go* + adj to form an adverb.

(121) *(scéala) suaimhnithe* '(news) of reassurance'

(122) *Bheith leasc chun rud a dhéanamh* 'Be slow to do something'

(123) *go héaganta* 'giddily'

Canonical Form and Lexicalised Elements Ideally, each MWE in the lexicon would be stored in its canonical form,[5] and only lexicalised elements would be included. However, many of the entries include spurious tokens, particularly those entries from FGB which are intended as example usages of a single word or MWE (Examples 124 and 125).

(124) *Dá mbeadh cosúlacht ar bith orthu* 'if they showed any promise'
canonical: *bí + cosúlacht + ar + (duine/rud)* and *ar + bith*

(125) *líonaim (soitheach) go béal* 'I fill (a vessel) to the brim'
canonical: *líon (rud) go béal*

Some entries, particularly those extracted from the English-Irish Dictionary, included non-lexicalised elements as part of the entry, however the non-lexicalised elements were members of a relatively small semantic class of words, which provide information on how the term would be used (Example 126). These entries lie on the border between *lexical item* and *description of usage*.

(126) *gearr* 'cut'
lexicon entry: *gearraim (pionós, fíneáil, dualgas)* 'I impose (a penalty, a duty, a fine)'

The best practice according to Calzolari et al. (2002a) is to include both these lexical items listed, as well as the complete description of the behaviour of MWEs

[5]The PARSEME annotation guidelines recommend the canonical form of the MWE be the least marked form of the expression that it can occur in, e.g. a verbal phrase in active voice whose head verb is in a finite form and whose other lexicalized components depend either on the verb or on another lexicalized component

both syntactically and semantically. However, generating and storing this information requires considerable manual effort.

Headwords The headword information, as described in Section 5.1.1, was extracted from only one lexical resource, FGB. Entries from the other resources had a headword that was automatically generated from the Irish entry. The automatic selection results in some headwords that are unhelpful (Examples 127 and 128). Headword information from the FGB were also inconsistent, as the headword selected may not be the head of the expression (see Example 129, where the headword is *caobh*, and the head of the expression is *cara*).

(127) ***an***, *an mhainistir* 'the cloister'

(128) ***a***, *a ardaíonn agus a íslíonn go rianúil* 'raise and lower regularly/arcing'

(129) ***caobh***, *cara caobh* 'gentle friend'

It is questionable how much information the headword is capable of providing, and aside from a means of filtering expressions with the same headword, does not appear to serve any function. Moreover, if the headword information provided by sources such as FGB is inconsistent, it may be preferable to always select the first word of the expression to maintain consistency. In this case, the same filtering effect can be achieved by sorting the lexicon based on the Irish MWE entry.

POS tags The POS information would ideally provide the POS of the head of the MWE. As already discussed, the headword information does not reliably give the head of the MWE, and, as shown in Table 5.1, some sources did not include POS information. Where POS information was included, it varied in what information was provided.

We see in Example 130 that POS information extracted from FB is in Irish (*ainmfhocal* 'noun'), and moreover, the POS information does not apply to the MWE, as *ina steillbheatha* 'in the flesh' should be classified as a prepositional phrase. Example 131 shows the POS information extracted from FGB, including the gender

of the headword *gort* 'field', and also includes extraneous tags intended to provide the inflected forms of the noun for genitive single, nominative plural and genitive plural. Examples 132 and 133, both MWEs extracted from EID, show semantic information beyond a broad-grained POS tag. Example 132 classifies *An Aisiria* as both a proper noun, and a geographical item, while Example 133 indicates the verb is transitive. In contrast, Example 134, an MWE extracted from Téarma, is simply maked as 'verb', while Example 135 shows an MWE extracted from NEID, whose POS information (phrasal verb) refers to the POS of the English entry, not the Irish translation, which is categorised as a Light-Verb Construction.

(130) *ina steillbheatha* (ainmfh) 'as large as life'

(131) *an gort amuigh* (m. gs. npl. gpl) 'the outfield'

(132) *an Aisiria* (Pr.n. Geog) 'Assyria'

(133) *déanaim soiscéalaíocht do* (v.tr.) 'evangelize'

(134) *bain tátal as* (verb) 'draw a conclusion from'

(135) *faigh ar iasacht* (phr_v) 'take out'

Licensing restrictions Due to the various licences and copyrights of the lexical sources these MWEs were extracted from, releasing the entire lexicon is not currently possible. Instead, an open-source version of this lexicon must be filtered for the entries that can be republished. A potential avenue for future work is to revisit these licences with copyright holders and explore the possibility of publishing some or all of this lexicon under an open-source licence.

Despite the aforementioned issues, this lexicon is a good starting point for the creation of MWE categories in Irish, which are fully explored in Chapter 4. The verbal categories found during the manual annotation became the foundation for the annotation of Verbal MWEs, discussed in Section 5.2, as well as allowing for the creation of the MWE-Annotated Parallel Corpus, discussed in Section 5.3.

5.2 PARSEME Annotated Corpus of Irish vMWEs

The PARSEME shared task is discussed in Chapter 2. The latest edition (Edition 1.2) saw the inclusion of Irish for the first time. In this section, we describe the creation of this corpus and provide some analysis of its quality and content.

5.2.1 Corpus Creation

As we described in Chapter 2, there was a shift of focus in the identification task for Edition 1.2, with a new focus on *unseen* vMWEs. As such, both a labelled and an unlabelled corpus were created for this edition of the shared task.

The supervised training of MWE identification in previous editions of the shared task relied on corpora that had been annotated for POS information, morphological tags, and dependency trees, as well as manually annotated with MWE information. The unlabelled corpus was automatically tokenised, lemmatised and parsed using UDPipe (Straka and Straková, 2017), to make it useful for unsupervised training.[6]

Both the labelled corpus and the unlabelled corpus described here are useful resources for the purposes of processing MWEs, particularly the labelled corpus. The labelled corpus is used in our experiments in MWE identification, as described in Chapter 7. In addition, the discussions arising from the annotation process provided much insight into the categories of vMWEs that exist for Irish, and was instrumental to developing the annotation guidelines for Irish MWEs described in the previous chapter (Chapter 4).

5.2.1.1 Labelled Corpus

1,700 sentences were taken from v2.5 of the Irish Universal Dependency Treebank (Lynn and Foster, 2016), containing gold-standard POS-information, morphological

[6]Note that the unlabelled corpus only includes MWEs of the kind annotated by UD (i.e. `flat`, `fixed`, `compound` and `compound:prt`), and not the categories we are trying to classify.

information, and dependency relations. The annotation was carried out by three annotators. Annotator A was the primary annotator, having had extensive experience with annotation of both English and Irish MWEs, as well as experience with using the PARSEME annotation guidelines. Annotator B and Annotator C were secondary annotators, both experts in Irish linguistics and experienced in syntactic annotation. A pilot annotation task was first carried out by Annotator A, where 100 sentences were annotated to verify the following categories to be used: `LVC.full`, `LVC.cause`, `VPC.full`, `VPC.semi`, `VID`, `IAV`, and `IRV` (see Chapter 4). The guidelines for these categories were expanded for use when annotating Irish vMWEs, with some adjustments made for differences in syntax and lexicon.

Following this step, Annotators B and C selected 600 sentences between the two of them for annotation, while Annotator A annotated 1,000 sentences, with all annotators using the modified annotation guidelines. Discussion took place among the annotators during and following this process, and where necessary, points in the guidelines were clarified and edited to be more applicable to the Irish data. Finally, Annotator A performed a review on all 1,700 sentences, including the 100 pilot sentences, to ensure decisions were applied consistently and to resolve disagreements between annotators B and C during annotation, based on decisions reached after discussion. These discussions were basis for the decisions made in our categorisation described in Chapter 4, with some of the more challenging questions being highlighted as difficult decisions.

To test annotation consistency, 800 sentences that had been annotated early in the process were selected to be annotated a second time, following the end of the annotation. 312 vMWEs were annotated in the first pass, while 270 were annotated during the second pass. *Intra*-annotator agreement[7] was calculated for these 800 sentences, using F_{measure} (an optimistic measure that ignores agreement due to chance), κ (an estimated Cohen's κ that measures the rate of agreement of annotation for all verbs in the corpus allowing for chance), and κ_{cat} (a score which takes

[7]Intra-annotator agreement was used as the same annotator had annotated all sentences.

into account only those vMWEs where both passes agreed on the span). The F_{measure} was 0.71. The κ was 0.66, i.e. substantial agreement, and the κ_{cat} score was 0.84, i.e. almost perfect agreement, according to Landis and Koch (1977).

5.2.1.2 Unlabelled Corpus

Our so-called unlabelled corpus was not manually annotated for vMWEs, and so was provided for unsupervised learning in this task. To construct the unlabelled corpus, 1,379,824 sentences were collected from the sources listed in Table 5.4.[8] UDPipe trained on v2.5 of the Irish UD treebank was used to perform automatic tokenisation, POS-tagging, lemmatisation, morphological analysis, and dependency parsing. As an upper bound on parsing accuracy, UDPipe achieves an Unlabelled Attachment Score (UAS) of 0.85 and a Labelled Attachment Score (LAS) of 0.78 on the v2.5 test set. A pre-processing step was introduced where a period was added at the end of each line where it did not already exist, which appears to have improved sentence splitting. A manual inspection of 100 sentences from each source was performed to asses the quality of the corpus, with some issues noted. Quality issues include lemmatisation (e.g. *dtagraíonn* lemmatised to *tagraigh* when it should be *tagair*; surface form *n-oibrítí* has both initial mutation and is in habitual past autonomous form), tokenisation (*d'imir* should be tokenised into *d'* and *imir*) and POS-tagging (*is* tagged as `AUX Cop` when it should be `CCONJ Coord`). These issues may have affected parsing accuracy.

Source	**Size**	**License**
Paracrawl (OPUS)	782,769	Creative Commons CC0 Licence
Vicipéid	302,838	GNU Free Documentation License (GFDL)
EU Bookshop (OPUS)	113,363	open-source (particular license not specified)
Citizen's Information website	10,297	CC BY 4.0
Tatoeba (OPUS)	1,894	CC–BY 2.0 FR

Table 5.4: Sources of unlabelled data, size in # sentences, and licence of the source.

[8]Text from Vicipéid Irish Wikipedia accessed 1/11/2019 and text from OPUS accessed at `http://opus.nlpl.eu/`.

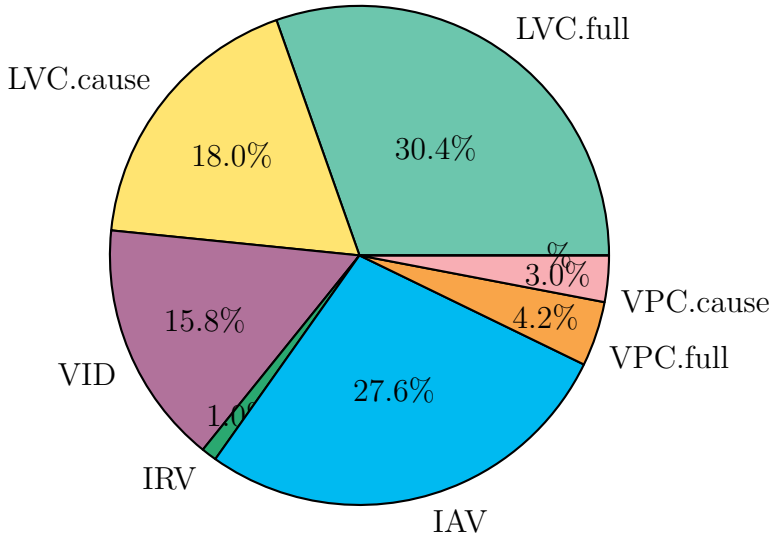


Figure 5.4: Breakdown of vMWE categories annotated in labelled corpus.

5.2.2 Analysis of Labelled Corpus

Our labelled corpus has a total of 662 vMWEs. Figure 5.4 shows a breakdown of the categories represented in this corpus, with the precise numbers shown in Table 5.5. `LVC.full` was the most commonly applied label in the corpus, with almost a third of the MWEs being annotated with this label. Combined with `LVC.cause`, almost half (48.3%) of the vMWEs were annotated as `LVC`. This is reflected in the results of the manually annotated sample of the Ilfhocail lexicon (Figure 5.3 and Table 5.3), which saw `LVC` as the most frequently occurring MWE with a verbal head. `IAV` was the second most commonly applied label, closely behind `LVC.full`. The least frequent category was `IRV`, with only 6 instances.

Category	**#Annotations**
LVC.full	201
IAV	183
LVC.cause	119
VID	105
VPC.full	28
VPC.semi	20
IRV	6
Total	662

Table 5.5: Number of annotations per category.

As the only Celtic language in the PARSEME shared task, Irish does not have

a close language neighbour for comparison. There is some evidence for English and Irish treatment of VPCs and IAVs overlapping (Stenson, 1997; Veselinović, 2006), and historically English has shared some lexical items with Irish through loanwords. In comparison to the English annotated corpus released for Edition 1.1 of the shared task (Walsh et al., 2018), there appears to be a higher density of vMWEs, with roughly 1 vMWE per 2.6 sentences, compared to 1 per 8.9 sentences in English. By estimating verb phrases using POS information in released `cupt` files, we see roughly 1 out of every 8 verb phrases contains a vMWE in Irish, while in English that becomes 1 out of every 47.8 verb phrases. The categories annotated in the English dataset were similar to those annotated in Irish, with the addition of `MVCs` (Multi-Verb Constructions) and without `IRV` vMWEs.

Table 5.6 shows the languages submitted to Edition 1.2. of the PARSEME shared task. Of these submitted languages, Hindi was the language whose corpus was the closest in size to Irish (1,684 sentences). Only four categories have been annotated in Hindi: `LVC.full`, `LVC.cause`, `VID` and `MVC`. Compared to Irish, vMWEs appear to be more dense, with 1 vMWE per 1.6 sentences, or 1 out of every 3.2 verb phrases. With the focus of this edition on the identification of *unseen* vMWEs,[9] the split of each corpus into training, development and test datasets was balanced using a minimum number of unseen vMWEs in the test (300) and development (100) data. When the ratio of unseen vMWEs is compared across languages, Irish has the highest rate of unseen vMWEs (0.69), with Hebrew showing the second highest rate (0.60). Compared to Irish, there were four categories annotated in Hebrew: `VID`, `LVC.full`, `LVC.cause` and `VPC.full`. The density of vMWEs annotated in the Hebrew corpus is closer to English than Irish (1 vMWE per 7.6 sentences, or 1 vMWE per 23.9 verb phrases).

[9] *Unseen* vMWEs are vMWEs that occur in the test data and not in the training or development data, and as such were not seen during training time. This is discussed further in Chapter 7.

Language	#Sents	#vMWEs	#Cats
German (DE)	8,996	4,041	6
Greek (EL)	21,447	7,444	5
Basque (EU)	11,158	4,246	3
French (FR)	20,961	5,654	5
Irish (GA)	1,700	662	7
Hebrew (HE)	19,200	2,533	4
Hindi (HI)	1,684	1,034	4
Italian (IT)	15,728	4,210	9
Polish (PL)	23,547	7,186	4
Brazilian Portuguese (PT)	32,117	6,437	5
Romanian (RO)	56,703	6,171	4
Swedish (SV)	4,304	1,991	6
Turkish (TR)	22,311	7,730	3
Chinese (ZH)	39,929	9,164	5

Table 5.6: Comparison of labelled corpora for each language submitted to Edition 1.2 of the PARSEME shared task. #Sents refers to the size of the corpus in sentences, #vMWEs refers to the number of annotated vMWEs in total per corpus, and #Cats refers to the total number of categories of vMWEs that are annotated per corpus.

5.3 MWE-Annotated Parallel Corpus

The importance of parallel corpora is discussed in Chapter 2, as well as how such corpora can be annotated with MWE information. We listed several sources of parallel data in Chapter 2, which became the data we used to build an MWE-aware parallel dataset. To build this dataset, we tagged both the English and Irish side of the dataset with MWE information, using the Ilfhocail lexicon we described in Section 5.1, as well as the English Multiword Expression Lexicons we described in Chapter 2.

This method of automatic tagging resulted in an automatically MWE-tagged parallel corpus. A portion of this data was manually annotated with gold MWE labels, and used as a focused MWE test set in our experiments. We describe in this section the methodology used for building both of these resources, and perform some evaluation of the automatic tagging method used. Additionally, we examine the following uses of these resources:

- Our automatically MWE-tagged parallel corpus is used as a resource for train-

ing MWE-aware MT systems, as described in Chapter 6.

- The manually MWE-annotated sample corpora are useful for the evaluation of MT systems, particularly in the evaluation of MWEs, given the density of MWEs annotated in this sample.

- By comparing the annotations of the manually MWE-annotated sample with the automatically MWE-tagged portion of the corpus, we can gauge the efficiency of the automatic tagging process, and the value of the lexicon of Irish MWEs described above.

5.3.1 Datasets Used

We first performed a manual inspection of the parallel corpora listed in Chapter 2, which revealed some issues with the data. As some of these datasets were automatically scraped from websites quality issues like this are to be expected. Some cleaning steps were undertaken, including manually realigning data where misalignments had occurred, and running a cleaning script to remove noisy tokens (e.g. '&amp'), redundant punctuation ('. ,'), and tokenisation errors ('mygaelic.com').

We categorised the datasets into one of four domains: `technical`, `legal`, `general`, and `crawled` data. `Technical` domain includes open source documentation (e.g *GNOME* localization files). `Legal` data includes EU and Irish national legislation (e.g. *EUconst*, a parallel corpus collected from the EU constitution). `General` domain makes up the majority of our resources, and includes datasets generated by public administration (e.g. *Conradh*, bilingual press releases compiled by Conradh na Gaeilge). `Crawled` includes data automatically scraped from online websites (e.g. *ParaCrawl*, parallel corpora collected in the ParaCrawl project). This data was split into training, tuning and test datasets, with each domain represented for each dataset. Table 5.7 shows the resources used, and the size and domain of each of these resources, and how these resources were split into training, tuning and test datasets.

Dataset	**# of sentences**	**domain**
Train		
Citizen's Information	10.3k	General
Conradh	1.4k	General
Crawl data	3.3k	Crawled
DAHG	59.7k	General
DGT	66.2k	General
EU Bookshop	108.9k	General
EUconst	10k	Legal
Europe	29.4k	General
GNOME	85k	Technical
KDE4	114.5k	Technical
Paracrawl	214.4k	Crawled
Paradocs	96.5k	Legal
Teagasc	1.5k	General
Tatoeba	0.5k	General
TOTAL	827.3k	
Tuning		
Crawl data	750	Crawled
DCHG	750	General
GNOME	750	Technical
Paradocs	750	Legal
TOTAL	2424	
Test		
Gov. Memos	1516	General
KDE4	1200	Technical
Paracrawl	1500	Crawled
Paradocs	1500	Legal
Ubuntu	381	Technical
TOTAL	6097	

Table 5.7: Statistics showing dataset name, size and proportion of data.

5.3.2 Annotating the Data

Multiword entries were extracted from the English and Irish resources into separate lists. As the Irish lexicon was automatically curated, and the English lexicon had been reviewed, we performed a trial tagging to assess the lexicons. 2,000 sentences were extracted from both the Irish and English side of each corpus. This smaller dataset was tagged with the list of MWEs in the corresponding language, and manually reviewed. MWEs that occurred more than 5 times per sample-corpus were printed to a list, and the list was inspected for so-called "nuisance MWEs", i.e. MWEs that were overly context dependent or whose literal occurrences were more frequent than idiomatic usage (e.g. EN: 'in this', GA: *na daoine* 'the people/the population'), collocations or non-idiomatic entries (e.g. EN: 'is added', GA: *an chéad* 'the first'), or otherwise deemed to not be useful to tag (e.g. EN: 'etc .', GA: *go hálainn* 'lovely'). These nuisance-MWEs were removed from the MWE-list for each language, along with other unhelpful entries, such as numerical entries or other noisy entries (e.g. EN: '1 200', EN: ', , , c. itoh , , , ,').

Following this cleaning step, the training, tuning and test datasets were automatically tagged using these MWE lists. Tokens were tagged with either an `MWE` label or a `NONE` label. Tagging was performed using a script to check each word in the dataset for a matching headword in the MWE lists, and to append the `MWE` label to each word in the expression, if it matched an MWE entry. The special character '|' is used in OpenNMT to add additional features to the data. There were three approaches used for identifying MWEs.

Fixed MWEs were annotated using a words-with-spaces approach, where the candidate expression and MWE entry had to exactly match to be annotated, with the intent that fixed continuous expressions (i.e. with no inflected elements or gaps or non-lexicalised elements interleaving) were annotated in this pass (e.g. EN: 'Ace of Clubs', GA: *trína chéile* 'mixed up').

Semi-fixed MWEs were annotated after first using the UDPipe tool to lemmatise the corpus. The MWEs lists were also lemmatised, and these lemma forms were used to tag MWEs in the datasets. This allows for tagging MWEs that would otherwise be ignored due to inflection (e.g. ***dhéanann*** *tagairt do* 'makes a reference to/refers to', 'petroleum **oils**').

A joint combination of these MWEs were annotated by combining the tags from the fixed and semi-fixed MWE tagged datasets.

Example 136 shows a sentence in the Irish corpus tagged with MWEs, while Example 137 shows the same sentence in the English corpus.

(136) GA: *grúpa*|NONE *gan*|NONE *Ghaeilge*|NONE *ar*|MWE *bith*|MWE

(137) EN: a|NONE group|NONE without|NONE any|NONE Irish|NONE at|MWE all|MWE

5.3.3 Manually Annotated Portion

To serve as a gold standard for evaluation purposes, 25 sentences were selected from each of the four domains (`technical`, `legal`, `general` and `crawled`), and manually examined and annotated for the nine categories of MWEs described in Chapter 4, using annotation guidelines (see Appendix A). These sentences were selected using a filtering method to count the number of MWEs in each sentence and output sentences with at least 5 MWE-tagged tokens.[10] Some domains failed to yield 25 sentences with at least 5 MWE tokens, so less MWE-dense sentences were chosen from the corpus. By using this filtering method, we ensured that these manually annotated samples were dense with examples of MWEs. The English sample (which we call 'Gold_100_EN') contained 274 MWEs, or an average of 2.7 MWEs per sentence, while the Irish sample (which we call 'Gold_100_GA') contained 393 MWEs, or an average of 3.9 MWEs per sentence.

[10]As the mode MWE was 2 tokens in length, this number was selected to aim at 2 MWEs per sentence.

There were three annotators, Annotators A, B and C, who performed the annotation on the two corpora, each of which underwent three rounds of annotation. The first round was conducted by the primary annotator A, who annotated the sentences, applying the first version of the annotation guidelines for the nine categories of Irish MWEs discussed in Chapter 4. Following this, the Irish corpus was passed to Annotator B, while the English corpus was given to Annotator C. These two annotators reviewed the annotations, consulting the annotation guidelines, and these reviews were then discussed with Annotator A. Inconsistencies in the labelling of the corpora were discussed, and the guidelines were refined to clarify certain points. Finally, Annotator A applied these updated guidelines to the sentences and addressed any inconsistencies.

5.3.3.1 Comparison and Analysis of Parallel Corpora

As an evaluation of the automatic tagging process for the Irish corpus, we compared the gold-standard manually annotated corpus for Irish ('Gold_100_GA') with the automatically tagged portion of the Irish side of the parallel corpus (joint combination), which we call 'Tagged_100_GA'. Each of the MWEs tagged in Tagged_100_GA was compared to the MWEs annotated in Gold_100_GA, and the MWE was assigned an evaluation label depending on how it compared. If the tagged MWE matched completely with the gold-annotated MWE, it was considered a *full match* ('F'). A *partial match* ('P') MWE was one that includes some or all matching tokens from the gold-annotated MWE, but had either too few or too many tokens. It differed from an *overlapping* ('O') MWE in that a 'P' MWE appeared to be tagging the same MWE as the gold-annotated MWE, while an 'O' MWE was tagging an MWE that shared tokens with the gold-annotated MWE but did not appear to be referring to the same MWE, or was tagging an MWE of a different category. This included embedded MWEs, where the inner MWE was captured but not the outer MWE. *Incorrectly-tagged* ('I') MWEs were those tagged by the automatic tagging that were not considered MWEs in the gold-annotations. Finally, *missed* ('M') MWEs were

Category	Gold_100_GA	Tagged_100_GA
(F)ull match	*Ceart go leor*	*Ceart go leor*
(P)artial match	*i gcomparáid*	*i gcomparáid le*
(O)verlapping	*in aimsir*	*aimsir chogaidh*
(I)ncorrectly-tagged		*ar an Domhnach*
(M)issed	*teacht le*	

Table 5.8: Examples of MWEs from each corpus and the evaluation label assigned them.

Category	F	P	O	M	Total
CC	10 (47.6%)	3 (14.3%)	1 (4.8%)	7 (33.3%)	21 (5.3%)
FE	113 (80.7%)	8 (5.7%)	11 (7.9%)	8 (5.7%)	140 (35.6%)
IAV	10 (30.3%)	7 (21.2%)	8 (24.3%)	8 (24.2%)	33 (8.4%)
IP	3 (60%)	0	1 (20%)	1 (20%)	5 (1.3%)
LVC	2 (5%)	2 (5%)	26 (65%)	10 (25%)	40 (10.2%)
NC	35 (63.6%)	1 (1.8%)	5 (9.1%)	14 (25.5%)	55 (14.0%)
NE	18 (21.2%)	8 (9.4%)	19 (22.3%)	40 (47.1%)	85 (21.6%)
VID	3 (37.5%)	0	5 (62.5%)	0	8 (2.1%)
VPC	6 (100%)	0	0	0	6 (1.5%)

Table 5.9: Evaluation of MWE categories annotated in Gold_100_GA with regards to their automatic tagging in Tagged_100_GA.

those that were annotated in Gold_100_GA that were not tagged at all by the automatic tagging. Table 5.8 displays examples of an MWE in the Gold_100_GA and Tagged_100_GA corpus for each of these evaluation labels. 393 MWEs were annotated in Gold_100_GA, while 403 were tagged in Tagged_100_GA. Figure 5.5 displays the number of MWEs for each of the five evaluation labels. In total, 49.0% of the tagged MWEs were correct (F), and 57.6% of the tagged MWEs were correct or almost correct (F+P). As an estimate of accuracy, we assigned a score of 1.0 (full credit) to full matches, 0.8 (almost full credit) to partial matches, and 0.5 (partial credit) to overlapping MWEs, to give us an estimated tagging accuracy score of 65.0%. The breakdown of these labels for each MWE category is shown in Table 5.9.

The measure of accuracy we use to evaluate this method of tagging MWEs is not directly comparable with most systems for MWE identification, where token-based F1 scores are preferred for analysing accuracy. This metric is explored in greater detail in Chapters 7 and 8. However, by examining the figures in Table

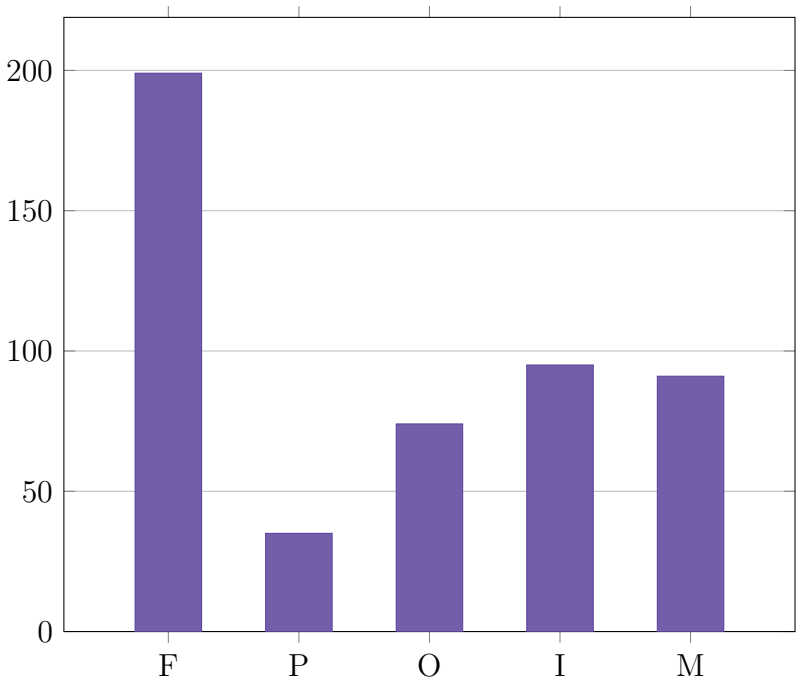


Figure 5.5: Comparing automatically tagged and gold-annotated MWEs in Tagged_100_GA and Gold_100_GA.

5.9, we see that some categories of MWEs are easier to tag than others. Fixed expressions in particular were captured well, with 80.7% of MWEs deemed correct, and 86.4% correct or almost correct, which is intuitive given how the method is only concerned with syntactically fixed or semi-fixed expressions. On the other end of the spectrum, verbal MWEs such as LVCs and VIDs appear to be more difficult to accurately target with this method. While the VPC category appears to have been tagged with a high degree of accuracy, this is a relatively rare category in the data, so a larger sample may demonstrate lower accuracy in tagging.

This token-matching approach is simple to apply, however it has two major drawbacks. The first drawback is that the accuracy of this method depends entirely on the quality of the external lexicon. Automatically generated lexicons, such as Ilfhocail, offer the benefit of wide coverage; however, this is offset by the potential inclusion of noisy or incorrect data. The second drawback of this method is that it does not attempt to distinguish between idiomatic and literal uses of MWEs, a recognised challenge in the problem of MWE identification discussed in Chapter 2. We explore a more sophisticated system for the identification of MWEs in Chapter

8.

5.4 Conclusion

This chapter describes a number of MWE-aware resources we have compiled for use in NLP tasks. The contributions of this chapter are the following:

- We report on the compilation of a lexicon of Irish MWEs (Ilfhocail), and perform a manual inspection of a sample of the data, highlighting the quality issues discovered, as well as the breakdown of MWE categories therein.
- We describe the manual annotation of a corpus of verbal MWEs for Irish for inclusion in the PARSEME shared task on the automatic identification of verbal MWEs. We perform an analysis of this corpus, comparing it with those of other languages, and examining the breakdown of MWE categories that are annotated.
- We also describe the creation of a bilingual parallel corpus that has been automatically annotated with MWEs from the Ilfhocail lexicon. We perform a manual inspection and annotation of a small sample of sentences in both Irish and English, creating two parallel corpora annotated on the source side with MWEs. We perform a comparison of the Irish manually annotated corpus with the automatically tagged portion to determine the quality of the automatic annotation.

These resources are used both in our experiments on incorporating MWEs into GA→EN MT, described in Chapter 6, as well as our experiments on MWE identification described in Chapter 8.

Chapter 6

Incorporating MWEs in MT systems

> *'What do you mean?' he said. 'Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?'*
>
> J.R.R. Tolkien, *The Hobbit*

6.1 Introduction

In Chapter 2, we highlight the nature of MWEs as a linguistic phenomenon that impacts language on both the lexical and grammatical level. Given this aspect of MWEs, it is not surprising that machine translation systems are a commonly cited downstream application of MWE identification and processing. Machine translation (MT) systems are designed to automatically translate text in some source language to a target language while retaining the meaning of the original text, and ensuring that the output is fluid and grammatical. With the prevalence of MWEs in language, it is clear that such constructions need to be addressed in this context.

Given its official language status in both Ireland and the EU, the Irish language would greatly benefit from reliable MT systems. Yet one of the ongoing challenges for Irish (GA) ↔ English (EN) MT has been the lack of sufficient available parallel data to train NMT systems (Judge et al., 2012; Dowling et al., 2019). This chapter explores methods of overcoming quality issues in MT systems through leveraging other existing NLP resources for Irish, in this case, MWEs. The experiments described in this chapter investigate the extent to which current ongoing research in the automatic processing of MWEs in Irish can help to improve upon the current state-of-the-art (SOTA) GA↔EN translation models.

Our approach involves automatic tagging of source and target parallel data for the GA↔EN language pair using a lexicon-based look-up tagger. The experiments are then run for both EN→GA and GA→EN MT using the open source OpenNMT toolkit, which allows the MWE tokens to be included as word features on the source side. An evaluation of the output is performed, using both automatic evaluation metrics and manual inspection of the output. In this chapter, we focus on addressing **RQ4**, first by investigating methods of incorporating MWE information in MT systems, drawing from the existing research in this field (**RQ4a**), and secondly through exploring different evaluation metrics for our analysis of the results (**RQ4b**).

6.2 Background

Historically, machine translation systems were **Rule-Based MT** (RBMT) systems which used large lexicons and rule bases to systematically translate words and phrases into the target language (Varile and Lau, 1988; Senellart et al., 2001). These RBMT systems were time consuming and labour intensive to create, and coverage was difficult to attain. **Example-Based MT** (EBMT) and later, **Statistical MT** (SMT) offered a new paradigm of automatic translation. EBMT and SMT are both based on the idea that given sufficient training data in both the source and target language, a computer was capable of implictly learning grammar rules and lexical

translations automatically. EBMT matches fragments of text from the bilingual data to produce a translation in the target language. SMT systems required a large amount of parallel text to build their translation models, and a body of monolingual text in the target language to build their language model. These systems are largely evaluated through automatic metrics such as BLEU score (Papineni et al., 2002), which measures the n-gram overlap between the MT-generated translation and a human translation for the same source sentence.[1] **Neural MT** (NMT) systems have become the new state-of-the-art approach for automatic translation in many well-resourced language pairs (Wu et al., 2016; Bojar et al., 2016). Informally, they may be considered as more sophisticated SMT models which use encoders and decoders to transform the input text into vectors. NMT is on par or has surpassed SMT in terms of automatic metric scores for many high-resource languages (Junczys-Dowmunt et al., 2016), although NMT has been shown to struggle with small datasets (Lohar et al., 2019). Although low-resource languages are thus at a disadvantage with such methods when compared to resource-rich languages such as English (Koehn and Knowles, 2017), experiments have shown that NMT matches or outperforms SMT for Irish (Dowling et al., 2018; Defauw et al., 2019; Lankford et al., 2021).

6.2.1 MWEs in MT

The challenge posed by MWEs in the field of NLP, including MT, have given rise to several research initiatives devoted to focusing on these issues. The PARSEME project, described in Chapter 2, has its origins in a working group created to improve the automatic processing and identification of MWEs in a highly multilingual context, with machine translation noted as one of the most prominent use cases in MWE processing (Savary et al., 2015).

Following on from both the findings and the research network built from a working group in PARSEME, a workshop on Multi-word Units in Machine Translation

[1]BLEU and other automatic metrics are explored further in Section 6.4.2.

and Translation Technology was formed. The biannual workshop is devoted in particular to the challenges associated with automatic translation of MWEs, which remains an open problem today. The workshop explores a multitude of topics, such as development of multilingual MWE resources, identification and acquisition of MWEs and their variant forms, word alignment techniques, MWEs in term extraction, and evaluation of machine translated MWEs, and many others.

There has been substantial research on the topic of MWEs in SMT (Carpuat and Diab, 2010; Bouamor et al., 2012b; Tan and Pal, 2014; Cholakov and Kordoni, 2014; Skadina, 2016) and RBMT (Deksne et al., 2008; Monti et al., 2011), and to a lesser extent within the context of EBMT (Anastasiou, 2010; Kim et al., 2010); however research on integrating MWEs in NMT systems remains sparse, despite NMT systems being recognised as state-of-the-art (Sennrich et al., 2016; Wu et al., 2016).

The process of integrating MWEs in each of these systems varies depending on the system; for instance, phrase-based SMT systems in principle can capture MWE information during the initial phase of building phrase tables, though this syntax-agnostic approach can lead to issues with structuring (Barreiro et al., 2013). In order to combat this, one approach is to treat MWEs in the training data as a single unit or one-word token, i.e. static integration. An alternative approach is to integrate the MWE information dynamically, as a feature following the generation of phrase tables. Both of these methods show an increase in MT performance (Carpuat and Diab, 2010; Simova and Kordoni, 2013). Cholakov and Kordoni (2014) demonstrated that additional linguistic information can further improve translation of MWEs (i.e. phrasal verbs) in SMT, indicating MWE handling benefits from linguistically-aware methods.

NMT, however, is structurally different to its predecessors, and strategies that improve SMT systems may not be successful for NMT. To investigate this, Sennrich and Haddow (2016) examined whether linguistic information could improve NMT or whether the strong learning capabilities of NMT rendered linguistic input redundant.

They found that including linguistic features such as morphological features, part-of-speech tags, and syntactic dependency labels caused an improvement in NMT systems across three metrics (BLEU, CHRF and perplexity).

NMT systems represent a sentence as a high-dimensional vector, and reproducing MWEs in this structure may be challenging, particularly if such constructions are sparse in the data. Rikters and Bojar (2017) experimented with two methods of integrating MWEs in NMT: (1) through automatically generating a parallel corpus of pairs of MWEs and including this parallel corpus as additional training data, and (2) using a parallel corpus of the sentences containing the MWEs extracted in the first step as training data. The results showed a small increase in BLEU score for both methods, and a manual inspection of the translation of certain MWE types revealed that the inclusion of extra MWE data improved these translations.

Drawing inspiration from both these experiments, Zaninello and Birch (2020) employed two methods of integrating MWEs in NMT. Firstly they augment the training data both with a parallel corpus of MWEs extracted from a bilingual dictionary, and a backtranslated parallel corpus containing MWEs in the target data. The MWEs in the data were then annotated using a words-with-spaces approach, and an IOB (inside, outside, beginning) approach.[2] They found all of these methods outperformed a baseline NMT system output in terms of automatic metrics, and a human evaluation of the systems confirmed this.

6.2.2 Description of NMT Model Architectures

As mentioned previously, NMT models are based on the use of encoder-decoder architectures. The encoder maps text into a continuous vector representation using a Recurrent Neural Network (RNN) or a Transformer. This vector acts as input to the decoder, which generates output using another RNN to map the vector representation into text in the target language. Figure 6.1 shows a simplified illustration of this architecture.

[2]The IOB labelling approach is discussed further in Chapter 8.

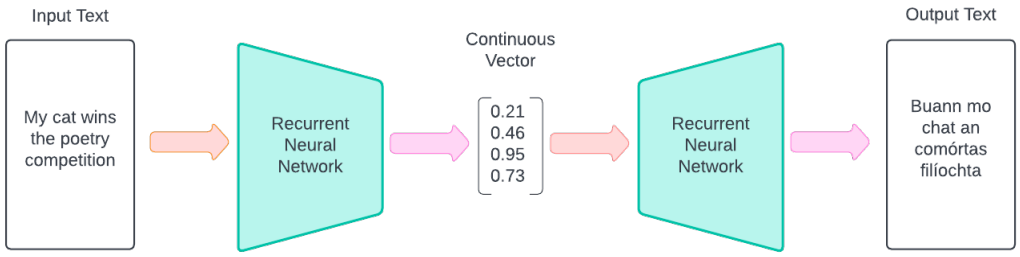


Figure 6.1: Simplified Encoder-Decoder model.

To explain simply, RNNs are neural networks used to predict some output following a sequence. Like a feed-forward network, the input is passed to a number of hidden states, where it is transformed into a numerical representation before being used to predict an output. However, unlike a traditional feed-forward network, RNNs contain a looping mechanism that allows prior information to be passed forward. This means that the final item in the sequence is represented in an encoding that contains information from all the previous items. This process makes the RNN suitable for processing sequences such as words in a sentence, as it enables the model to remember all the previously seen words. Figure 6.2 visualises how input is processed sequentially in an RNN.[3]

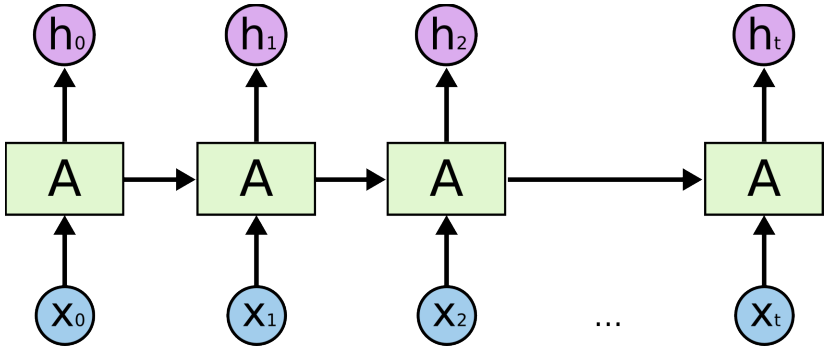


Figure 6.2: Sequential processing in Recurrent Neural Networks: x_n represents an input, A represents a neural network module, and h_n represents an output at that module.

However, one of the issues with this model is short-term memory. As the RNN model processes more and more steps (or words, in the case of NLP), the information

[3]Diagram from `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

contained in the earlier steps becomes less and less important to the representation of the later steps. This is because as the model updates the weights during back-propagation, it makes the adjustment at each layer based on the adjustments made to the previous layer, meaning that a small adjustment will become exponentially smaller as the model back-propagates down the layers, a phenomenon know as vanishing gradient. As each layer represents the learning of a step (or word) in the sequence, this means earlier steps will not learn as the rest of the model does, and the weights given to these earlier layers will not be trained up. As such, for the representation of a sentence, long distance dependencies between words are not well captured by the RNN model, a problem known as short-term memory. The opposite effect can also occur, where adjustments to the layers becomes exponentially larger as the model back-propagates, which is called an exploding gradient.

In order to address this problem, two specialised RNN architectures were devised, Long Short-Term Memory (LSTM) RNNs, and Gated Recurring Unit (GRU) RNNs. We describe LSTMs in greater detail here, as these are used in our experiments described below. LSTMs have a more intricate structure to the looping mechanism than simple RNNs, which allows them to remember information for a long time. Rather than a single layer to connect each cell as in a simple RNN, LSTMs have four interactive layers connecting each cell. These layers contain gate mechanisms that decide which information should be passed along to the next cell, and which information should be forgotten. Figures 6.3 and 6.4 illustrate the different cell structures in a simple RNN and an LSTM RNN model.[4] During back-propagation, the gradient is controlled by these gate mechanisms and so is regulated to avoid vanishing or exploding.

Both simple RNNs and LSTM RNN architectures are used in encoder-decoder NMT systems. In each of these models, the input is passed sequentially through each hidden state, until the final item or step is reached. At this point, the vector representation of the final hidden state is passed to the decoder, where the items are

[4]Diagrams both from `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

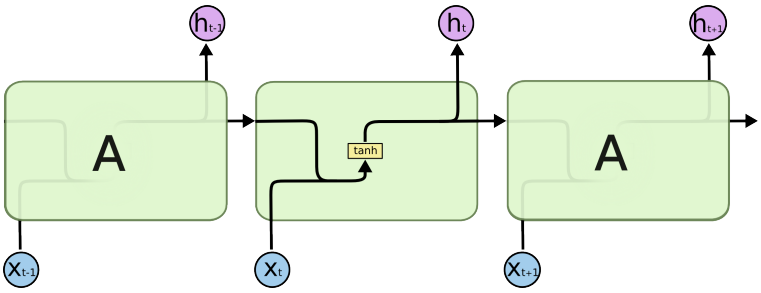


Figure 6.3: Cell structure within a simple RNN cell.

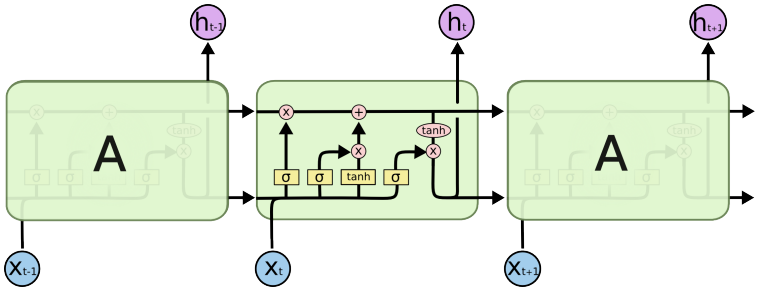


Figure 6.4: Cell structure within a LSTM RNN cell.

predicted sequentially. Similarly, each predicted step uses both the encoder vector and the hidden states of previously predicted items to generate the next hidden state.

Transformer models are another type of encoder-decoder architecture employed in sequence-to-sequence modeling. This architecture is explained in more detail in Chapter 7, as the Transformer architecture is also used in training pre-trained language models. To provide a brief explanation: Transformer models differ from RNN models in that they do not process the input sequentially (i.e. word-by-word), instead the entire input is processed in parallel. Positional-encoding for each item is used to represent the order in which the items occur. An attention mechanism is used to calculate how items in the sequence should relate to each other.

A more detailed explanation of many of these concepts can be found in the literature, such as RNNs (Sutskever et al., 2014), vanishing gradient problem (Bengio et al., 1994), LSTMs (Hochreiter and Schmidhuber, 1997), and Transformers (Vaswani et al., 2017). For understanding the field more generally, several blog posts have been written exploring the details of many of these architectures (Olah, 2015; Phi, 2018). Forcada (2017) offers an accessible explanation of the terms and

processes of NMT for readers without a technical background.

6.3 Experiments

The experiment outlined in this chapter is partly informed by the work of Sennrich and Haddow (2016) and Zaninello and Birch (2020). Through appending MWE information as features to the source data, we investigate whether this additional information leads to better translations of MWEs for both Irish and English.

6.3.1 Data

In training the NMT systems, we use a parallel corpus of Irish and English text, compiled from various resources. MWEs were tagged as linguistic features in this corpus, following from the approach of Sennrich and Haddow (2016), and Vanmassenhove and Way (2018). The collection and automatic tagging of this corpus is described in detail in Chapter 5. Briefly, the MWEs were identified and tagged using a lexical lookup tool, with the MWE tag appended to each token of the MWE as with a linguistic label (see Example 138). In total, four versions of the data were created for Irish and English each: `baseline`, which had no MWEs tagged; `fixed`, which had fixed MWEs tagged; `semi-fixed`, which had semi-fixed MWEs tagged; and `joined`, which was a concatenation of the labels for fixed and semi-fixed MWEs.

(138) tuairisceáin|NONE a|NONE chur|NONE **ar**|MWE **fáil**|MWE
'to make reports available'

6.3.2 Data Processing

Data for machine translation systems require tokenisation in order to segment the sentences into chunks that can be encoded, in our encoder-decoder models. One method of tokenisation is to split sentences into words. There are drawbacks to this method however. Languages with a rich morphology will generate a very large

vocabulary size, and out-of-vocabulary words are highly likely. To address this, Sennrich et al. (2016) propose a technique of word segmentation called Byte-Pair Encoding (BPE). This technique allows for rare words to be captured in a segment or subword, rather than as a whole. This technique is particularly effective for languages with extensive morphology, and particularly, a large number of prefixes and suffixes. Example 139 demonstrates BPE for a short construction (*Péint nó vearnais*), while Example 140 shows an MWE (*Laíon adhmaid*) segmented into subwords. The "@@" symbol depicts where a word has been broken into subwords.

(139) pé@@|NONE int|NONE nó|NONE v@@|NONE earn@@|NONE ais|NONE
'Paint or varnish'

(140) La@@|MWE íon|MWE adhmaid|MWE
'Wood pulp'

We use the Subword-NMT tool[5] to apply BPE to our data. The tool builds a vocabulary list of subwords for each language first from the training data, which is segmented using this list. The vocabulary is built by first splitting input sentences into characters, and then merging frequently co-occuring characters into subword units. These subword units can be concatenated with other characters or subword units, until a predefined number of merge operations have been carried out. The larger the number of merge operations specified, the larger the subword units will be. Generally in NMT this number is set somewhere between 1,000 and 100,000 operations (Morishita et al., 2018), with certain numbers occurring frequently in the literature (e.g. 30k, 32k, 89k). For this experiment, we opt to use 32k merges. Baseline data could be segmented directly using these vocabulary files. However, for the MWE-tagged files, a conversion script was applied to append MWE-tags and NONE-tags to each subword unit as appropriate.

[5] `https://github.com/rsennrich/subword-nmt`

6.3.3 Models

OpenNMT (Klein et al., 2017) is an open source toolkit for building NMT systems. It can be implemented using the open source machine learning framework PyTorch[6], which implementation allows for capturing linguistic features in the source data. We use two model architectures for our experiments, both of them implemented through the PyTorch framework.

Our first model architecture is an LSTM RNN model, which is one of the two default architectures offered by OpenNMT for NMT models (the other being a GRU RNN). We use the default settings for this model, which is a LSTM of size 500, with 2 layers in both the encoder and decoder, a dropout of 0.3, a batch size of 160 sentences, a learning rate of 1.0, and a Stochastic Gradient Descent (SGD) optimiser.

Our second model is a Transformer model, with the following settings: a model size of 512, with 6 layers (blocks) for both the encoder and the decoder, 8 attention heads, a dropout of 0.1, a batch size of 4096 tokens, a learning rate of 2.0, and an Adam optimiser with beta2 of 0.998. The settings chosen were a combination of recommended settings given in Vaswani et al. (2017), as well as recommended settings based on current research in Irish NMT (Lankford et al., 2021).

Both models were trained on Nvidia GPUs. The RNN model was trained for 100,000 steps, while the Transformer model was trained for 200,000 steps.

6.4 Results

The models trained in each direction were tested on four domains of data, reflecting the four domains that make up the training data: legal, technical, general and crawled data. The content of these test datasets were chosen to be of high quality, although the nature of the crawled data is such that quality cannot be assured.

[6] `https://pytorch.org/`

6.4.1 Evaluation of MWEs in MT

Evaluating the integration of MWEs into MT systems can be challenging, as with data sparsity, the measurable effect of adding MWE information may be missed. Two aspects normally used to judge the quality of machine translation output are **adequacy** i.e. how well the MT system preserves the meaning when translating from the target language to the source language, and **fluency** i.e. how well the MT system creates a translation in the source language that is grammatically correct and uses natural language. MWEs affect both these measures; poor handling or a literal word-to-word translation can result in translations that are clunky and ungrammatical, as well as a loss of meaning.

Barreiro et al. (2013) cite fragmentation as the most common source of errors in multiword expression processing, meaning that MT systems struggle to address the non-compositionality of some MWEs. As MWEs tagged in our corpus are always contiguous, the issue with fragmentation may be mitigated. However, when analysing the types of MWEs that occurred in their training data, Barreiro et al. noted that compound nouns represented the type of MWE that contained the greatest number of errors, which are captured in our experiments. This type of error is difficult to recognise through automatic metrics, however, so a manual inspection of the data is important.

Monti and Todirascu (2015) highlight the lack of benchmarking resources (e.g. parallel corpora annotated with MWEs), and a lack of agreed-upon methodologies and procedures for assessing MWE translation quality in this field. Linguistic resources often must be manually created and are time-consuming and labour intensive to develop. In an attempt to mitigate this problem, we have created two test sets for evaluation: a larger test corpus that has been automatically tagged with MWEs in the same method as the training data, and a smaller corpus that has been manually annotated for MWEs. Both of these corpora are described in more detail in Chapter 5.

6.4.2 Automatic Metrics

Automatic metrics are widely used to evaluate MT systems due to being fast, cheap and easy to apply without the need for trained human experts.

BLEU Scores

The most commonly used metric for this task is BLEU (BiLingual Evaluation Understudy), introduced by Papineni et al. (2002), which calculates an n-gram overlap between the machine translated output and a human translated reference. The simplicity of the algorithm combined with its applicability to any language with segmented tokens make it a popular choice for reporting MT systems results. However, it has received extensive criticism in the field for a number of reasons (Callison-Burch et al., 2006).

One of these criticisms it that BLEU is a measure of translation similarity, which does not equate to translation quality. Translation is not a binary outcome, and there are many possible fluent and adequate translations of a single sentence. As such, an MT translation could have a BLEU score of 0.0 (no similarity to the reference translation) and still be a perfect translation. For instance, Example 141 demonstrates two valid translations of the English phrase 'I would like another chance', neither of which contain any words in common.

(141) Source: 'I would like another chance.'
Reference: *Ba mhaith liom deis eile.*
Target: *Tá seans sa bhreis uaim.*

Furthermore, these words need to be identical to be scored as a match, so partially matching words are not considered in the score. This unfairly penalises languages with a rich or complex morphology, such as Irish.

CHRF Scores

CHRF (Popović, 2015) is a character based n-gram F-score, intended to combat the problems posed by BLEU scores. As it scores matching characters and not words, it is a more lenient scoring method for languages, particularly those with rich morphologies, and is recommended for use instead of BLEU as an automatic metric (Kocmi et al., 2021). Both BLEU and CHRF metrics can be calculated using the SacreBLEU tool (Post, 2018), which we used in our evaluation.

Score_mwe

An MWE-aware automatic metric was proposed by Zaninello and Birch (2020), called *score_mwe*. The calculation returns an average of the levenshtein distance[7] between all reference MWE translations and the system hypothesis translations in the data. We adopt this as a focused metric to evaluate how MWEs in particular are treated in our MT systems, applying it to the gold manually annotated MWEs in our small test set.

Statistical significance tests

Statistical significance tests are performed to determine the validity of the results. This metric indicates the likelihood of the observed results occurring due to random chance. The confidence interval used for our tests is 95%, i.e. we can assume the difference in the metric achieved by our baseline data and our modified data is not due to random chance where the p-value is less than 0.05. In the event that the p-value is greater than this instance, we cannot confidently say that the observed results did not occur due to random permutations of the data. SacreBLEU includes an option for calculating statistical significance using paired-bootstrap resampling, similar to that used by Koehn (2004). This is a method to sample with replacement from the test dataset a number of times (e.g. 1000) and compute the evaluation

[7]Levenshtein distance between two words is the minimum number of edits to single characters (insertions, deletions or substitutions) necessary to change one word into the other.

metric for each system at each sample. If one system is superior to the other greater than 95% of the time, we can conclude that this result is statistically significant.

While we have described how BLEU scores are not ideal for measuring MT quality, comparing BLEU scores between different systems trained on the same data can be helpful for finding patterns in the translation decisions made by different systems. In Table 6.1, we report the BLEU scores, broken down by domain for each system translating EN→GA. Table 6.2 shows these results for each system translating GA→EN. We can see immediately from both tables that the Transformer-based model results in a higher BLEU score for both EN→GA and GA→EN MT. Moreover, across all systems, the BLEU score for each domain is relatively consistent, with crawl data consistently yielding the highest BLEU score, and technical data consistently yielding the lowest. We inspect these domains more closely in our manual inspection in Section 6.4.3.

The results in Table 6.1 indicate that the addition of MWE labels does not have a large impact on the BLEU score of EN→GA MT, for either the LSTM-RNN or the Transformer models. However, we see a slight overall improvement in the BLEU score for the LSTM-RNN models, and it appears adding MWEs improves BLEU scores for every domain except technical. The improvement is particularly noticeable with the legal data, where BLEU improved by +0.84 points with fixed MWEs tagged, and +1.25 with semi-fixed and fixed MWEs tagged. The Transformer model showed a slight decrease in BLEU score overall with the addition of fixed or semi-fixed MWEs, although both the technical and the legal data show improvements in BLEU when fixed, semi-fixed, or both MWEs are tagged.

For GA→EN MT, Table 6.2 shows overall BLEU scores are similarly not greatly impacted by adding MWEs, for both the LSTM-RNN and Transformer models. The most interesting changes in BLEU score for the LSTM-RNN models occur in the technical and legal data once more. The technical domain shows a greater decrease in BLEU when MWEs are added, particularly with the addition of semi-fixed MWEs,

Model	Overall	Crawl	Technical	Legal	General
LSTM-RNN baseline	46.9668	66.4232	28.3495	47.7262	34.8168
LSTM-RNN fixed	**47.7898***	67.4611*	**28.6594**	48.5665*	**35.5912***
LSTM-RNN semi-fixed	47.5545*	**67.4679***	28.2682	48.8106*	34.8576
LSTM-RNN joined	47.4443*	66.8321	27.1197	**48.9780***	35.1320
Transformer baseline	56.8470	76.5594	33.3514	57.8874	**45.1700**
Transformer fixed	56.5696	76.0044*	33.3826	58.0655	44.7325*
Transformer semi-fixed	**56.8652**	76.4098	**34.8958**	**58.1493**	44.905
Transformer joined	56.7718	**76.7354**	34.6634	58.0232	44.5068

Table 6.1: EN→GA BLEU scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

which shows a decrease of 2.43 BLEU points. The legal domain, however, shows an increase in BLEU score with the addition of MWEs, particularly with the addition of semi-fixed MWEs, which shows an increase of 1.49 BLEU points. The Transformer models show very little change in BLEU score with the addition of MWEs, however, the technical data shows a greater decrease in BLEU score with the addition of semi-fixed MWEs (-3.76 points), and a slight increase in BLEU score with the addition of both fixed and semi-fixed MWEs (+0.81 points).

Model	Overall	Crawl	Technical	Legal	General
LSTM-RNN baseline	52.9694	73.2454	**32.3495**	52.9904	40.9148
LSTM-RNN fixed	52.9657	73.0244	30.8239*	53.4689	40.9818
LSTM-RNN semi-fixed	**53.5982***	73.4574	29.9244*	**54.4811***	**41.7676***
LSTM-RNN joined	53.3488*	**73.6475**	31.4069	53.1372	41.6037*
Transformer baseline	62.4658	81.6737	38.1029	63.6806	51.0163
Transformer fixed	62.3404	81.2478	36.3014*	63.1241	**51.5935***
Transformer semi-fixed	62.1166	**81.7347**	34.3434*	63.0541	50.8094
Transformer joined	**62.7060**	81.5721	**38.9161**	**63.9820**	51.3343

Table 6.2: GA→EN BLEU scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

Tables 6.3 and 6.4 show the CHRF scores for each of these models, broken down by domain, for EN→GA and GA→EN respectively. Similar to BLEU scores, CHRF scores do not indicate MT quality, merely the similarity between the hypothesis translation and the reference translation. As such, we are primarily interested in

patterns of differences in these scores with the addition of MWE information. Generally, these scores are higher than the BLEU scores for the equivalent system, which intuitively is logical, as the metric is more lenient. We see this difference is more extreme for EN→GA MT, for both the LSTM-RNN models (baseline models show increase of +17.5 vs +15.1 CHRF points to BLEU points) and Transformer models (baseline models show increase of +14.7 vs +12.5 CHRF points to BLEU points). This aligns with the previous assertion that BLEU scores penalise Irish hypothesis translations more severely than English hypothesis translations. Comparing the difference between the BLEU and CHRF scores across domains also reveals a pattern, with the General domain showing a larger gap between BLEU and CHRF scores across both system architectures and both languages.[8] To understand this difference, we compared the datasets for general domain, and crawl domain, which appeared to have the least difference between BLEU and CHRF scores. Crawl domain contains data scraped from websites, and includes many sentence fragments with repetitive content, and notably, verbs are often omitted from sentences (e.g. 'Online debate (chat) session no. 8'). This lack of verbs, and the inflection that they can display, may be the reason the CHRF score does not vary so much from the BLEU score in this domain, as BLEU score penalises improper inflection. Likewise, the general domain appears to contain fewer sentence fragments, and as such, the CHRF score would more fairly evaluate inflection errors made by the translation of verbs. A human evaluation study could confirm this intuition, or reveal other patterns in the data and the translation decisions of the models.

Table 6.3 reports CHRF scores for EN→GA MT. We see a similar pattern as with the BLEU scores for the LSTM-RNN models, as both technical and legal data are the most impacted by the addition of MWE data. Technical data shows a decrease in CHRF, particularly when both fixed and semi-fixed MWEs are added (-

[8]With differences in BLEU and CHRF scores averaged across the four models within each domain, architecture, and language direction group, we find CHRF scores for the general domain are consistently higher than the Overall domain, with this difference varying from +7.25 points for Transformer models in the GA→EN direction, to +8.82 points for LSTM models in the EN→GA direction.

1.71 CHRF points). Legal data shows a similar level of increase in CHRF scores with the addition of MWE information, with the greatest increase (+1.33) resulting from adding semi-fixed MWEs. The Transformer models do not reveal any particular pattern of increase or decrease of CHRF scores, with MWE information improving CHRF scores for some domains (technical and legal), but diminishing CHRF scores in other domains (crawl and general).

Model	**Overall**	**Crawl**	**Technical**	**Legal**	**General**
LSTM-RNN baseline	64.4688	78.0791	**46.2322**	61.9792	61.1520
LSTM-RNN fixed	**64.9339***	**78.6509***	45.6821	63.1088*	**61.3694**
LSTM-RNN semi-fixed	64.7355*	78.3560	45.8298	**63.3136***	60.9441
LSTM-RNN joined	64.7310	78.2696	44.5184*	63.1877*	61.3071
Transformer baseline	71.5535	85.3334	50.2967	71.3903	67.4505
Transformer fixed	71.3275	84.8405	50.8932	71.6267	66.9564
Transformer semi-fixed	**71.6636**	85.3065	51.0917	71.4381	**67.5440**
Transformer joined	71.5971	**85.4839**	**51.7588**	**71.6811**	67.0203

Table 6.3: EN→GA CHRF scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

For the other direction, as shown in Table 6.4, patterns are hard to discern, with some domains showing an increase in CHRF scores with the addition of MWE information, either fixed, semi-fixed, or both. However these improvements are always small, and, bar the legal and general domain for the LSTM-RNN and Transformer models, the baseline CHRF score is never lower than the score of all three of the MWE-informed systems.

Tables 6.5 and 6.6 display the results for calculating the `score_mwe` for each system, broken down by each domain, for EN→GA and GA→EN MT respectively. These scores were calculated by running the `score_mwe` tool described above on the gold annotated 100 sentences test set described in Chapter 5. Given the very small test size, the patterns observed in these `score_mwe` scores may not be reflective of more general patterns in the data, however, they provide helpful starting points for manual inspection of the data.

The `score_mwe` scores for EN→GA MT displayed in Table 6.5 show a higher

Model	Overall	Crawl	Technical	Legal	General
LSTM-RNN baseline	68.1253	82.0093	**48.0593**	66.3506	64.1276
LSTM-RNN fixed	68.1093	81.6814	47.6292	66.4461	64.3397
LSTM-RNN semi-fixed	**68.7102***	81.9075	47.2532*	**67.5929***	**65.0556***
LSTM-RNN joined	68.4379*	**82.1221**	47.8693	66.5635	64.7369*
Transformer baseline	74.9288	87.6912	52.4872	75.5227	70.7752
Transformer fixed	74.9585	87.6730	52.0300	74.9893	**71.2129***
Transformer semi-fixed	74.6904	87.9118	50.9375*	74.8466*	70.8293
Transformer joined	**75.1909***	**87.9193**	**52.8869**	**75.7661**	71.0348

Table 6.4: GA→EN CHRF scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

variability in scores across domains and systems than with either BLEU or CHRF scores. The LSTM-RNN models actually shows a decrease in `score_mwe` scores for models trained on some MWE-aware datasets in crawl, technical, and general data, with these results particularly diverse in technical data, where adding fixed MWEs results in -7.44 `score_mwe` points. Legal data, however, appears to improve with the addition of all MWE information, and adding both fixed and semi-fixed MWEs results in +7.93 `score_mwe` points. The Transformer models see an overall improvement in `score_mwe` scores with the addition of MWE information, however, both the technical and legal domains show a diminished `score_mwe` score for some MWE-aware models, while the `score_mwe` scores for the general domain show a slight decrease for all three MWE-aware models. The crawl domain shows an increase in all three MWE-aware models, and adding both fixed and semi-fixed MWEs improves the `score_mwe` by +1.64.

In Table 6.6, we see the `score_mwe` scores for GA→EN MT, with the scores generally diminishing with the addition of MWE information. The LSTM-RNN models show an overall decrease in `score_mwe` scores for fixed, semi-fixed and both MWEs added. Both technical and legal domain show an increase in `score_mwe` scores when fixed MWEs are added, but adding semi-fixed MWEs or both fixed and semi-fixed MWEs results in an overall `score_mwe` decrease. Crawl domain shows an increase in `score_mwe` for all three MWE-aware models, and general domain shows a

Model	Overall	Crawl	Technical	Legal	General
LSTM-RNN baseline	70.3699	**73.0337**	**59.1286**	68.8358	**73.6713**
LSTM-RNN fixed	70.6822	72.8692	51.6861*	75.5606*	72.7785
LSTM-RNN semi-fixed	69.4669	72.6062	52.8305	75.1943*	68.7250*
LSTM-RNN joined	**71.4200**	71.5677	56.7834	**76.7614***	72.6212
Transformer baseline	73.4852	73.9112	58.3156	77.1764	**76.3567**
Transformer fixed	**75.5654**	74.8032	53.8065	76.1768	75.5389
Transformer semi-fixed	74.7020	75.4041	**61.1382**	79.5937	75.2649
Transformer joined	74.3126	**75.5469**	55.6367	**80.9426***	75.5390

Table 6.5: EN→GA `score_mwe` scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

decrease in `score_mwe` for all three MWE-aware models. Patterns are more difficult to discern with the Transformer models, with some MWE-aware models increasing `score_mwe` scores in some domains, and decreasing `score_mwe` scores in others.

Model	Overall	Crawl	Technical	Legal	General
LSTM-RNN baseline	**71.6465**	72.4156	57.6219	74.3550	**72.4778**
LSTM-RNN fixed	70.7771	73.2541	**58.0612**	**75.1125**	68.4457*
LSTM-RNN semi-fixed	70.6226	**74.3735**	56.1961	72.9155	69.3661*
LSTM-RNN joined	70.2185	73.5512	57.3458	71.8411	69.0517*
Transformer baseline	75.2568	74.0565	67.4649	76.1630	**76.5781**
Transformer fixed	74.3125	73.1689	70.4785	75.6920	73.6611
Transformer semi-fixed	73.6649	**76.6333***	64.3673	74.4352	72.1652*
Transformer joined	**76.1143**	74.9672	**72.7149**	**76.5970**	75.8954

Table 6.6: GA→EN `score_mwe` scores per model broken down for four domains. Scores marked with an asterisk * show a statistically significant difference from the baseline with a threshold of 0.05.

6.4.3 Manual Inspection

The patterns in the automatic metrics reported above seem to indicate that adding MWE information has a larger effect on the translation of technical and legal domains of data, while general and crawl data appear to be less affected. To investigate this further, we examine more closely the content of each of these domains. Table 6.7 displays some statistics for these domains in both the Irish and English test

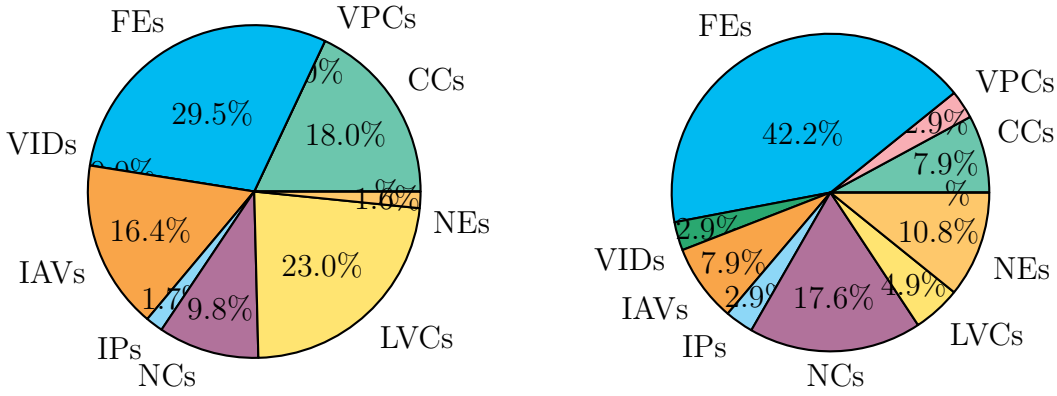
corpora.

Irish		**All**	**Crawl**	**Tech**	**Legal**	**Gen**
# MWEs	CC	21	2	11	8	0
	FE	140	46	18	43	33
	IAV	33	4	10	8	11
	IP	5	1	1	3	0
	LVC	40	6	14	5	15
	NC	55	17	6	18	14
	NE	85	34	1	11	39
	VID	8	1	0	3	4
	VPC	6	1	0	3	2
Total # MWEs		393	112	61	102	118
Total # Tokens		2904	854	358	740	952
Avg Sent Len		29.04	34.16	13.77	29.60	38.08
English		**All**	**Crawl**	**Tech**	**Legal**	**Gen**
# MWEs	FE	44	17	3	14	10
	IAV	18	5	3	5	5
	IP	20	6	2	9	3
	LVC	20	3	4	1	12
	NC	64	21	16	15	12
	NE	78	20	3	14	41
	VID	7	2	0	2	3
	VPC	22	2	3	8	9
Total # MWEs		273	76	34	68	95
Total # Tokens		2481	656	287	596	942
Avg Sent Len		24.81	26.24	11.48	23.84	37.68

Table 6.7: Summary of statistics in Irish and English gold annotated 100 sentences for each domain.

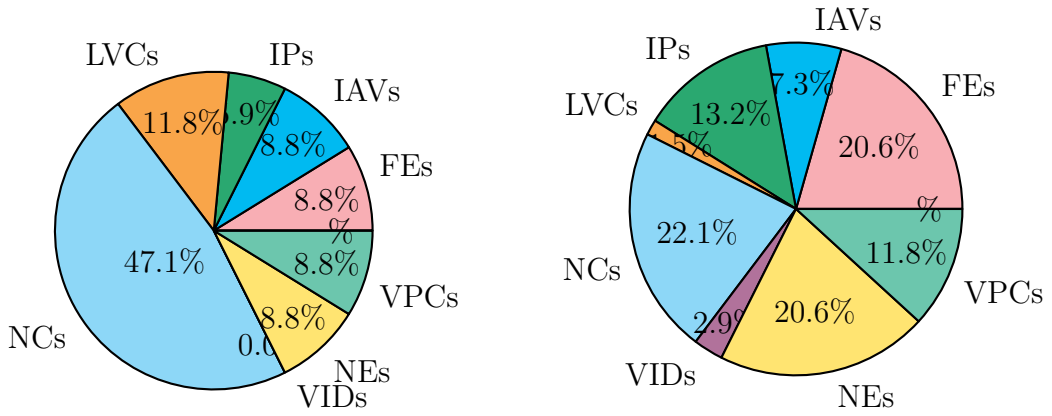
The proportion of MWE categories in the technical and legal domain is provided for the Irish corpus in Figure 6.5 and for the English corpus in Figure 6.6. Notably, the legal domain contains a higher proportion of fixed expressions (FEs) when compared to the technical domain, for both Irish and English corpora. Named entities (NEs) are also more represented in the legal domain, a category which are often syntactically fixed as discussed in Chapter 2. Conversely, the technical domain contains a higher proportion of verbal MWEs, and light verb constructions (LVCs) in particular, which are more syntactically flexible.

To investigate some of these translations closer, we use the CompareMT tool (Neubig et al., 2019), which can be used to compare the outputs of two different



(a) Technical (b) Legal

Figure 6.5: Proportion of MWEs per category in technical and legal domains for Irish small test corpus.



(a) Technical (b) Legal

Figure 6.6: Proportion of MWEs per category in technical and crawl domains for English small test corpus.

MT systems, presenting the salient differences between the hypothesis translations for both systems. The tool allows for comparing BLEU scores across both systems for sentences of differing lengths, and words of different frequencies, as well as highlighting n-grams that were translated better in one system over the other.

Reference	**Baseline LSTM**	**MWE-tagged LSTM**
a lucht féachana nó éisteachta a fhás	*a gcuid scileanna a fhás*	*a gcuid lucht féachana a fhás*
Deontas báis	*Deontais báis*	*Deontas báis*
photo Album	photography	photo Album
hydrofluoric acid	hydroflufluential acid	hydrofluoric acid
Reference	**Baseline Trans**	**MWE-tagged Trans**
Foras na Gaeilge	*Gaeilge*	*Foras na Gaeilge*
_ *Barra Uirlisí*	*barra na nUirlisí*	_ *Barra Uirlisí*
a person of limited means	a small person	a person of limited means

Table 6.8: Comparison of MWE translations from baseline models and MWE-tagged models (either fixed, semi-fixed or joined), where MWE tagging was perceived to have a positive impact on translation.

By comparing some of the output from LSTM-RNN models trained on the baseline data with the models trained on MWE-tagged data (either fixed, semi-fixed or joined), for both EN→GA and GA→EN MT, we can see minor differences in system predictions. Table 6.8 displays some of these examples. Some constructions containing or consisting of MWEs appear to be translated better in systems where the MWE is tagged, such as *go dtí seo* 'until now', which contains the fixed expression *go dtí* 'until', and *lucht féachana* 'spectators', which is annotated as an MWE in all three MWE-tagged corpora. In the English translations, we see some examples of this occurring, such as with the nominal compound 'photo album', and the MWE 'hydrofluoric acid'. In cases where the baseline model outperformed the MWE-tagged models, certain n-grams included MWEs that had not been tagged, such as *is fearr* 'best', and *Séala Oifigiúl* 'official seal' in the Irish translations. The English baseline translations that outperform the MWE-tagged translations also demonstrate this phenomenon (e.g. 'personal takings of', where 'personal takings' was not tagged), although it also includes the MWE 'progress report', which is tagged in all three MWE-tagged corpora.

Examining the output from the Transformer models, we see that the models trained on MWE-tagged data often have lower BLEU scores than the baseline model. MWEs tagged in the training data appeared to help with their translation in some cases (e.g. *Foras na Gaeilge*, an NE for the body for the promotion of the Irish language), however, others are translated in the baseline model where they are not translated properly in the MWE-tagged models, despite being tagged as MWEs in these training datasets (e.g. *cuir i bhfeidhm* 'put into effect', *Comhairle Contae Dhún na nGall* 'Donegal County Council'). One example of MWE tagging improving the English translation is the phrase *duine ar bheagán airgid* which should be correctly translated as 'a person of limited means'. This correct translation is generated by the Transformer model trained on data tagged with both fixed and semi-fixed MWEs, while the baseline model translates this as 'a small person', which does not capture the same meaning. However, as with the other MT models, there were some instances of MWEs which were correctly translated in the baseline model but not in the MWE-aware models, despite being tagged in the training data (e.g. 'Union Square', 'go to', 'National Parks and Wildlife Service').

Applying the CompareMT tool to the gold annotated 100 test sentences, we find it is difficult to ascertain just how the translations of MWEs are affected by tagging them in the training data. Looking at the output of the LSTM-RNN models, it appears some of the MWEs are better translated by models that have been trained on MWE-tagged data (e.g. *ar feadh* 'during', *deontas báis* 'death certificate', 'sodium hydroxide solution', 'National Archives'). However, there are also many instances of MWEs that are translated better in the models trained on baseline data (e.g. *aigéad sulfarach* 'sulfuric acid', 'Address book', 'dealing with'). The Transformer models show similar patterns of MWEs whose translation is helped by the addition of MWE tags (*is gaire* 'nearest', *le haghaidh* 'for', 'Security Council', 'per cent.'), and MWEs whose translation appears to be more accurate in the baseline models (*i gceist* 'in question', *ar intinn* 'in mind', 'according to'). One particular case is the VID *sa lá atá inniu* 'nowadays'. A portion of this VID (*sa lá* 'in the day') has

been automatically tagged in the MWE-tagged corpora, despite not being an MWE. Even though the whole MWE is not captured, it appears to aid with the translation of the MWE for the Transformer models trained on that data.

Overall, the effect of tagging MWEs automatically is difficult to track in translation quality, with some Automatic Evaluation Metrics (AEMs) indicating an improvement in overall translation, while others indicate a decline. Examining individual translations of MWEs reveals some cases where the additional MWE information appears to help, and some cases where tagging the MWEs may add further noise to the training dataset. There does appear to be patterns in the effect of adding MWE information to certain domains, and future work should be conducted towards investigating the particular types of MWEs, if any, that best benefit from this tagging method.

6.5 Conclusion

The experiments described in this chapter explore the impact of including MWE information in both EN→GA and GA→EN MT, a topic which has not been explored to date. We describe our experiment methodology, and investigate some of the metrics for estimating the effect of adding MWE information in our MT systems.

While the effect of adding MWE information is quite small, and it is difficult to determine whether the overall impact is positive for each of the four domains included in the data, there do appear to be some instances where adding MWE information in the form of linguistic tags can help with the automatic translation of these MWEs. NMT is known for its sometimes unpredictable behaviour, and so experiments such as this one must account for a certain level of unpredictability in the model performance which may not be entirely due to the effects of the additional information alone. Furthermore, there is only so much that can be concluded from such small test sets.

Such experiments would benefit significantly from conducting human evaluation

to investigate whether trends observed in the AEMs and the manual inspection of our small test data hold for larger amounts of data. This is particularly the case for MWEs that cannot be translated literally (e.g. *mac tíre* (son of land) 'wolf').

Future work for these experiments would be to fully analyse the types of MWEs that are impacted by the inclusion of MWE labels in the MT systems, and to investigate whether legal domain text contains more of these types of MWEs than the other domains, and whether this might account for the perceived impact of adding MWE information in this domain above the other domains.

Another direction to investigate is the translation of discontiguous MWEs, which are not explored in this experiment. Such MWEs may demonstrate a greater disparity between the LSTM-RNN and Transformer models, as the former translate inputs sequentially, and as such may find gappy MWEs more challenging to identify and translate.

Chapter 7

MWE Identification

> *'Go back?' he thought. 'No good at all! Go sideways? Impossible! Go forward? Only thing to do! On we go!'*
>
> J.R.R. Tolkien, *The Hobbit*

7.1 Introduction

In their survey paper, Constant et al. (2017) discuss the two main sub-tasks of MWE processing: discovery and identification, as briefly discussed in Chapter 2. In this chapter, we examine the task of MWE identification; that is, the task of *finding* MWEs in running text, and applying a *known label* to these MWEs.[1]

The field of MWE identification has seen a dramatic change in recent years, with shared tasks (such as the two discussed in this chapter) offering new perspectives on the challenges inherent to this task. Additionally, thanks to the creation of cross-lingually consistent manually annotated corpora (Savary et al., 2017; Ramisch et al., 2018, 2020), and the development of multilingual methodologies and evaluation schemes, a unified framework for the identification of MWEs across languages

[1]Identification handles MWEs on the level of tokens, rather than the task of discovery, which handles them on the level of types.

is emerging. Alongside these developing multilingual frameworks, monolingual resources such as lexicons have been earmarked as being of great importance in this task (Savary et al., 2019b). While low-resource languages can benefit from the shared knowledge inherent to multilingual approaches, they may lag behind when it comes to developing sufficiently rich language specific resources. This chapter addresses **RQ3**, by describing the task of MWE identification, providing background knowledge that is necessary to understanding the experiments described in Chapter 8. We define the evaluation metrics used to measure the performance of systems on this task. We also describe the architecture of the Transformer model, and the pretrained language models built on this architecture, which are used in our system for automatic identification of Irish MWEs described in Chapter 8. Finally, we present the results of Edition 1.2 of the PARSEME shared task, focusing on the treatment of Irish in this task, and highlighting potential challenges for Irish.

7.2 Background of MWE Identification

As mentioned, *identification* forms one of the two subtasks in the subfield of MWE processing. Unlike *discovery*, which is concerned about the properties and features of canonical forms of expressions and their potential to be MWEs (otherwise known as type-based extraction), MWE identification is centered on distinguishing the idiomatic usage of MWEs in text from literal or non-MWEs, i.e. token-based identification. Early work in this field was limited, in part due to lack of adequate resources, including MWE-annotated corpora and lexicons. Early efforts such as Lin (1999) and Bouma (2010) relied on statistical methods for identifying idiomatic constructions. Hybrid systems were also introduced, incorporating linguistic features with statistical measures such as in Evert and Kermes (2003) and in Bannard (2007). However for the most part, these methods were concerned with the identification of `types` of MWEs, rather than differentiating between literal and idiomatic occurrences of MWEs. In this respect, these efforts were closer to the task of *dis-*

covery than *identification* (see Appendix B for experiments exploring some of these methods). Cook et al. (2007) focus on developing unsupervised methods for token classification that differentiates between idiomatic and literal usage, a task that is much closer to the identification task we attempt in the following chapter.

Beyond the difficulty of defining the task and securing adequate resources, this task is understood to be challenging in part because of the idiosyncratic behaviours of MWEs that were discussed in Chapter 2. For instance, discontiguity in MWEs can make it difficult to relate tokens together that are not adjacent. Rohanian et al. (2019) explore the challenge of discontinuity through two types of neural architecture, Graph Convolution Network, and multi-head self attention. Other challenges that MWE identification present, as listed in Constant et al. (2017), are non-compositionality, nested or overlapping MWEs, ambiguity and variability (e.g. syntactic variability).

Constant et al. (2017) highlight common methodologies for MWE identification, such as rule-based matching, supervised classification, parsing and sequence tagging. Briefly, these methodologies are described as follows:

Rule-based matching involves applying a series of rules of varying complexity to the data, with the goal of extracting MWEs from the text based on the candidate expressions' adherence to the rules. Some typical approaches include the use of finite-state transducers to apply pattern matching algorithms to the data based on dictionary entries, or the application of rules to text following a morphological analysis of the text. Examples of this type of methodology are the Maximum Forward Matching algorithm used by Ghoneim and Diab (2013), and the lexical lookup strategy we employed in our experiments described in Chapter 6.

Classifiers using word sense disambiguation (WSD) treat MWE identification as a specialised in-context classification task. Normally these classifiers are dependent on another process or system for identification of idiosyncratic candidates, and are instead focused on detecting which of these candidates are true MWEs, us-

ing features such as the surrounding tokens, POS-information, lemmas, distribution, etc. An example of this method is the work of Hashimoto and Kawahara (2008), who employ a WSD framework in combination with idiom-specific features to construct a corpus annotated with Japanese idioms.

Sequence tagging uses stochastic models such as conditional random fields, structured perceptron, or structured support vector machines to predict token-level MWE labels. These models can use token-level features, local context and external resources to assist in this task. The multi-head self-attention method used by Rohanian et al. (2019) is an example of this approach. This is also the methodology that we apply in our experiments outlined in Chapter 8.

Parsing is another method that can be used to generate MWE labels, sometimes as a by-product of the process, as MWEs tend to form full syntactic constituents. For instance, Green et al. (2013) predict MWEs as a by-product of the parsing process. Identification of MWEs can occur before, during and after parsing, with each method having different applications and benefits. For instance, Kong et al. (2014) first annotate their data with MWEs before parsing takes place, Constant and Nivre (2016) extend the classical arc-standard parser to jointly predict MWEs with syntactic parsing, while Vincze et al. (2013) apply a classifier to predict MWEs after a parsing step. MWE identification can have positive effects on parsing, as it can help to address tricky syntactic structures in the text, and issues such as ambiguity. For example, the noun phrase 'green card office' can be parsed in two different ways, however, identifying 'green card' as a linguistic unit reduces the ambiguity of this phrase.

In addition to parsing, MWE identification is shown to be of benefit to MT systems (see Chapter 6), and can be considered a pre-processing step for both tasks, often using a lexicon and string lookup. MWE identification can also be treated in data post-processing, or form the basis of a joint-learning task. Depending on the

method of integration, this gives rise to a variety of architectures and methodologies that best work for this task. Other NLP tasks that can benefit from MWE identification include many that require some level of semantic processing, such as information retrieval (Acosta et al., 2011), question answering (Dowdall et al., 2003), and sentiment analysis (Williams et al., 2015).

Two shared tasks that focus on the automatic identification of MWEs are the DiMSUM shared task and the PARSEME shared tasks (the PARSEME shared tasks have been previously described in Chapter 2). These shared tasks provide many insights into the challenges inherent in this type of task. We explore these shared task initiatives below, in order to highlight the various approaches and evaluation methods used in this field.

Both the shared tasks made use of standard Precision, Recall and $F1$ scores for evaluation purposes. Section 7.3 describes these metrics in detail, while the different methods for calculating these scores used by both shared tasks is discussed in Section 7.3.1.

7.2.1 SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings

The SemEval-2016 shared task on Detecting Minimal Semantic Units and their Meanings (DiMSUM) (Schneider et al., 2016a) was an initiative in labelling English sentences with MWE tags and noun and verb supersense tags. The goal of this task was to move away from the paradigm of fine-grained word sense inventories such as WordNet (Fellbaum, 1998), which are difficult to annotate in corpora, can result in data sparsity, and are not easily generalizable to cover out-of-vocabulary words. Instead, the task aims to create a paradigm of broad-coverage, coarse-grained lexical semantic analysis.

The tagset for this task includes 'strong' MWEs and supersenses. 'Strong' MWEs are described by Schneider et al. (2014b) as MWEs that are mostly non-

Class	Supersense	Denotes	Example
Noun	Artifact	Man-made objects	'bridge'
Noun	Group	Groupings of people or objects	'family'
Noun	Relation	Relations between things, people, ideas	'contrast'
Verb	Competition	Fighting, athletic activities	'wrestle'
Verb	Possession	Buying, selling, owning	'tax'

Table 7.1: A number of supersenses for both nouns and verbs.

compositional in meaning. The supersense tagset consists of 41 broad semantic categories (Ciaramita and Johnson, 2003) based on the 'lexicographer class' labels used in WordNet (Fellbaum, 1998). These 26 noun classes and 15 verb classes offer a middle ground between the fine-grained senses used in WordNet, which are challenging to capture automatically, and the restricted coverage offered by the more broad-grained Named Entity Recognition classes. The supersense tagset provides human-interpretable high-level clustering across the word sense in WordNet. Table 7.1 displays a select number of examples of both noun and verb supersenses.

This task chooses to treat supersenses and MWEs at the same time, as the authors recognise that the supersense of an MWE as a whole may differ greatly from the supersenses of each of its individual components, particularly when the non-compositionality and semantic opacity often displayed by 'strong' MWEs is taken into account. Examples 142 and 143 demonstrate the supersense tagging of the individual components of 'spill the beans' (treating 'spill' and 'beans' as separate entities), versus tagging the expression as a unit.

(142) spill [Verb:Contact] the beans [Noun:Food]

(143) spill the beans [Verb:Communication]

The data for this task comprised a portion of the STREUSLE corpus (Schneider et al., 2014b; Schneider and Smith, 2015), a collection of Twitter data, and transcripts taken from TED Talks. This data was annotated with MWEs and supersenses by a single annotator. Submissions could be in one of three conditions: *open condition*, where participants were free to make use of any available resources

for the task; *closed condition*, where participants were limited to the resources provided for the task; and *semi-supervised closed condition*, where participants could make use of a large unlabelled corpus provided by the authors. Systems were evaluated on precision, recall and $F1$ scores, for both the MWE and supersense labels, as well as an averaged combined score for both tasks.

A total of six teams participated in the shared task, and nine systems were submitted. These systems used a variety of methods, including pattern-matching, a single- and double-chained Conditional Random Field (CRF), a structured perceptron, maximum entropy classification, and a neural network. The results indicate that the models best performing at this task were the single- and double-chained CRF method, and the structured perceptron which leveraged word embeddings. Both of these methods made use of the feature set defined by the AMALGraM sequence tagger (Schneider and Smith, 2015).

The task of MWE identification appears more challenging than supersense tagging, based on the results of the above task. For the task of supersense tagging, of the nine systems and across three domains of text, 74% of the $F1$ scores were above 50, with an average $F1$ score of 50.5, while for the task of MWE identification, only 41% of the $F1$ scores were above 50, with an average $F1$ score of 36.5. In an attempt to determine whether a better system could be built by combining systems, the authors compared each of the systems with the others to identify which cases were caught by one system but not the other. By performing an error analysis on the model output, they found that the best performing systems tended to capture the same knowledge or subsets of knowledge from the data. This result aligns with the intuition that certain MWEs are easier to identify than others.

7.2.2 The PARSEME Shared Tasks on Automatic Identification of Verbal Multiword Expressions

The PARSEME shared tasks, as described in Chapter 2, are a series of shared tasks focused on developing an automatic treatment of verbal MWEs that is consistent across languages. The latest edition (Edition 1.2) saw 14 languages take part, including Irish (Ramisch et al., 2020). The inclusion of Irish in this shared task is explored in Section 7.5.

Seven teams participated in this shared task, with nine systems submitted, two to the closed track (no external resources allowed) and seven to the open track (external resources allowed). A variety of architectures and resources were used by participants in this task. Of the nine systems participating in the shared task, five systems made use of neural networks: MultiVitaminBooster (Gombert and Bartsch, 2020), TRAVIS-mono and TRAVIS-multi (Kurfalı, 2020), MTLB-STRUCT (Taslimipoor et al., 2020) and ERMI (Yirmibeşoğlu and Güngör, 2020). Three used methods based on filtering using association measures: HMSid (Colson, 2020), Seen2Seen (Pasquer et al., 2020b) and Seen2Unseen (Pasquer et al., 2020a), while one system used a rule-based joint parsing and MWE identification system: FipsCo. Of the systems using neural networks, four of them included the use of pre-trained language models, those being multilingual BERT, monolingual BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020). Systems using such architectures performed particularly well on the task of identifying *unseen* MWEs (for the languages that were captured by each system), indicating that the context-aware word embeddings of BERT seem to capture some elements of semantic or syntactic idiomaticity.

7.3 Evaluation of MWE Identification

The three metrics commonly used to evaluate a task such as MWE-identification are precision, recall and $F1$. In general, these evaluation metrics are given as follows:

Precision is a measure of the number of correctly labeled instances normalised over the total number of system predictions.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall is the measure of the number of correctly labelled instances that have been identified by the system, normalised over the total number of instances in the gold dataset.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 score, also known as 'F-score', 'F-measure', or even 'balanced F1-score', is a calculated balance between the system's precision and recall, through assigning a value to a weight α to represent the weight given to the precision and recall scores respectively. It is commonly calculated as a harmonic mean between the two other scores, as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Huggingface supports several evaluation metrics in their datasets library,[2] which can be easily integrated into the training process. *seqeval*, a Python framework for evaluating sequence labelling (Nakayama, 2018), is commonly used for evaluating chunking tasks such as Named-Entity Recognition and POS-tagging, and reports precision, recall and $F1$ scores, among others. The framework supports several tagsets, including IOB2 tagsets, the modified version of which we use for our dataset. If a predicted MWE sequence does not fully match the gold MWE sequence, it is not counted as a match. In other words, this method of evaluation does not award partial credit for overlapping predicted and gold labels.

seqeval allows for reporting on different averages across the labels for these three

[2] `https://huggingface.co/metrics`

metrics, micro-averages, macro-averages and weighted averages. Micro-averages are calculated by counting the global total true positives, false negatives and false positives and averaging across these global values. Macro-averages are calculated per metric, and these values are then averaged. This calculation does not take label imbalance into account. Weighted-averages, in contrast, do take into account label imbalance by factoring in the number of occurrences of each label in the dataset. This can result in an $F1$ score that is not between precision and recall.

7.3.1 Alternative Evaluation Metrics

Both the PARSEME and the DiMSUM shared task recognised that the standard evaluation metrics described above do not perfectly measure the ability of systems to accurately label MWEs. Instead, they have proposed slight changes to the standard metrics.

Schneider et al. (2014b) propose a precision/recall measure based on the coreference scoring method proposed by Vilain et al. (1995), which measures the precision/recall of links between tokens of the MWE. The authors link consecutive tokens in the MWE together, so the order of tokens must be maintained for a match, but gaps in tokens are allowed. This method of scoring MWE matches harmonises with the *gappy 1-level* annotation scheme[3] used in the DiMSUM task, and allows for partial credit where gold and predicted MWE elements overlap. The precision is calculated as the proportion of predicted links where both tokens belong to the same expression in the gold standard. The recall is the proportion of gold links that are captured in predicted links. The $F1$ score is calculated as a harmonic mean of the precision and recall. Figure 7.1 illustrates how the gold standard (top) and predicted MWE (bottom) links are used to calculate precision, recall and $F1$.

In their first shared task paper, Savary et al. (2017) describe the evaluation metrics they use for evaluating vMWE identification. They consider each instance of vMWE to be indivisible, and all lexicalised tokens must be annotated as part of the

[3]This scheme and other labelling schemes are discussed in more detail in Chapter 8.

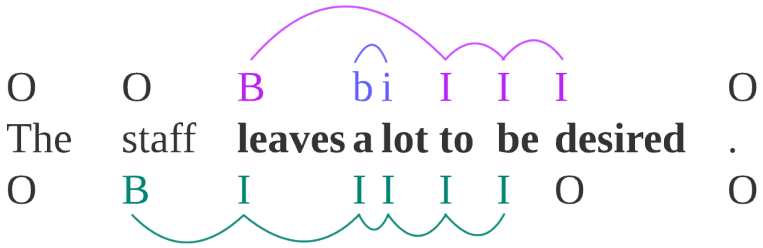


Figure 7.1: Example of an MWE annotated with gold standard (top) and prediction labels (bottom). Links between consecutive tokens are matched from the predicted labels to the gold standard labels. The precision of the bottom annotation relative to the top is 2/5. The recall of the top annotation relative to the bottom is 3/4. The $F1$ is the harmonic mean of the precision and the recall, working out to 12/23.

vMWE to be scored. This per-vMWE scoring is strict and can unfairly penalise vMWEs with many lexicalised tokens, or those with less-stringent lexicalised components, such as those including indefinite or definite articles (e.g. 'answer a/the call of nature'), or long vMWEs.

In order to report on partially matched vMWE candidates, they describe a per-token score. Unlike the evaluation method used in the DiMSUM paper, they do not consider vMWEs as necessarily formed of paired tokens, given that single-token vMWEs are considered valid in the annotation guidelines (e.g. 'to **pretty-print**' in English, ***kinyír*** (out.cut) 'to kill' in Hungarian, and *a* ***se-ndura*** 'to have heart' in Romanian[4]). Instead, all possible bijections between the vMWEs in the gold and predicted datasets are calculated, and the matches that maximise the number of correct token predictions are selected as the most optimistic measure of vMWE matching.

To illustrate this, we adopt the toy corpus example given in Savary et al. (2017). Table 7.2 contains this toy corpus with three tokens `t1, t2, t3`, and two vMWE labels `1` and `2`. The three system predicted labels are also provided. Given these predictions, the following statements can be assumed (G=Gold, S1=System1, etc.):[5]

[4]Examples taken from the PARSEME Annotation Guidelines 1.1

[5]If A is a given set, $|A|$ is the size of set A, and $||A||$ is the sum of the sizes of each set in set A

- $G = \{\{t1, t2\}, \{t3\}\}$

  $|G| = 2$

  $||G|| = 3$

- $S1 = \{\{t1\}, \{t2, t3\}\}$

  $|S1| = 2$

  $||S1|| = 3$

- $S2 = \{\{t1\}, \{t2\}, \{t3\}\}$

  $|S2| = 3$

  $||S2|| = 3$

- $S3 = \{\{t1\}, \{t2\}, \{t3\}, \{t1, t3\}\}$

  $|S3| = 4$

  $||S3|| = 5$

Token	**Gold**	**System1**	**System2**	**System3**
t1	1	1	1	1;4
t2	1	2	3	3
t3	2	2	2	2;4

Table 7.2: Toy corpus consisting of three tokens, two gold vMWE labels, and three system predicted labels.

The strict **per-vMWE** score would consider each instance of vMWE as indivisable, and only an exact match is considered. As such, the systems would be scored as follows (TP = True Positive; R = Recall; P = Precision):

- $TP1 = |G \cap S1| = |\emptyset| = 0$

  $R = TP1/|G| = 0/2 = 0.00$

  $P = TP1/|S1| = 0/2 = 0.00$

- $TP2 = |G \cap S2| = |\{\{t3\}\}| = 1$

  $R = TP2/|G| = 1/2 = 0.55$

  $P = TP2/|S2| = 1/3 = 0.33$

- $TP3 = |G \cap S3| = |\{\{t3\}\}| = 1$

  $R = TP3/|G| = 1/2 = 0.50$

  $P = TP3/|S3| = 1/4 = 0.25$

However, a more lenient **per-Token** score would instead select the matching that maximises the correct token predictions. With the above toy corpus, the **per-Token** scoring would be as follows (TP_{max} = True Positive score assuming a matching that maximises the number of correct tokens):

- $TP1_{max} = |\{t1, t2\} \cap \{t1\}| + |\{t3\} \cap \{t2, t3\}| = 2$

  $R = TP1_{max}/||G|| = 2/3 = 0.66$

  $P = TP1_{max}/||S1|| = 2/3 = 0.66$

- $TP2_{max} = |\{t1, t2\} \cap \{t1\}| + |\{t3\} \cap \{t3\}| + |\emptyset \cap \{t2\}| = 2$

  $R = TP2_{max}/||G|| = 2/3 = 0.66$

  $P = TP2_{max}/||S2|| = 2/3 = 0.66$

- $TP3_{max} = |\{t1, t2\} \cap \{t1\}| + |\{t3\} \cap \{t3\}| + |\emptyset \cap \{t2\}| + |\emptyset \cap \{t1, t3\}| = 2$

  $R = TP3_{max}/||G|| = 2/3 = 0.66$

  $P = TP3_{max}/||S3|| = 2/5 = 0.40$

It should be noted that these improved metrics are not perfect methods of capturing the performance of MWE identification models, given that they place equal weight on every element of the MWE, regardless of whether it forms a truly integral part of the expression or not. Constant et al. (2017) highlight the need to further investigate this issue in the future, and develop evaluation schemes that use a weighting algorithm to assess the importance of tokens in an MWE and award more credit for essential tokens.

7.4 Transformer-based Models in NLP

Since the seminal work of Vaswani et al. (2017), there has been a dramatic evolution of neural network-based methods, with Transformer-based architectures consis-

tently outperforming previously dominant models at a variety of tasks in NLP. Ruder (2018) summarises the major milestone approaches in the past 20 years, often building on previous milestones, and culminating in the development of high-performing Transformer-based models, which quickly became the standard for most NLP tasks. This is also the case for MWE identification, as discussed in Section 7.4.5.

7.4.1 Encoder and Decoder

The encoder-decoder model was developed with the rise in popularity of sequence-to-sequence neural architectures over Deep Neural Networks (DNNs). Sutskever et al. (2014) highlight the disadvantages of DNNs, namely that DNNs can only be applied to problems where the dimensions of both the input and output sequences are known, which makes sequence modelling difficult. Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models are described in Chapter 6, but to briefly restate, these models, when following the encoder-decoder structure, first encode input sequentially into a vector of fixed dimensionality. This vector is then used as input to the decoder, which would use it to predict an output sequence.

The Transformer encoder-decoder architecture has some similarities to that of the sequence-to-sequence model; it consists of an encoder stack that accepts input data, and a decoder stack, which generates output. The difference lies in the use of attention blocks without the need for using RNNs. The use of attention allows for greater parallelization than its predecessor models, and the Transformer architecture allows for training on large amounts of data in less time, paving the way for pre-trained language models such as BERT. The diagram shown in Figure 7.2 displays the encoder stack on the left and the decoder stack on the right.[6]

Each encoder has a self-attention layer and a feed forward layer, which are repeated N times (six in the vanilla Transformer model proposed by Vaswani et al. (2017)). In the first encoder block of the stack, the input tokens (i.e. words or subwords) are first passed into a embedding vector of a fixed size (e.g. 512 di-

[6]Taken from Vaswani et al. (2017).

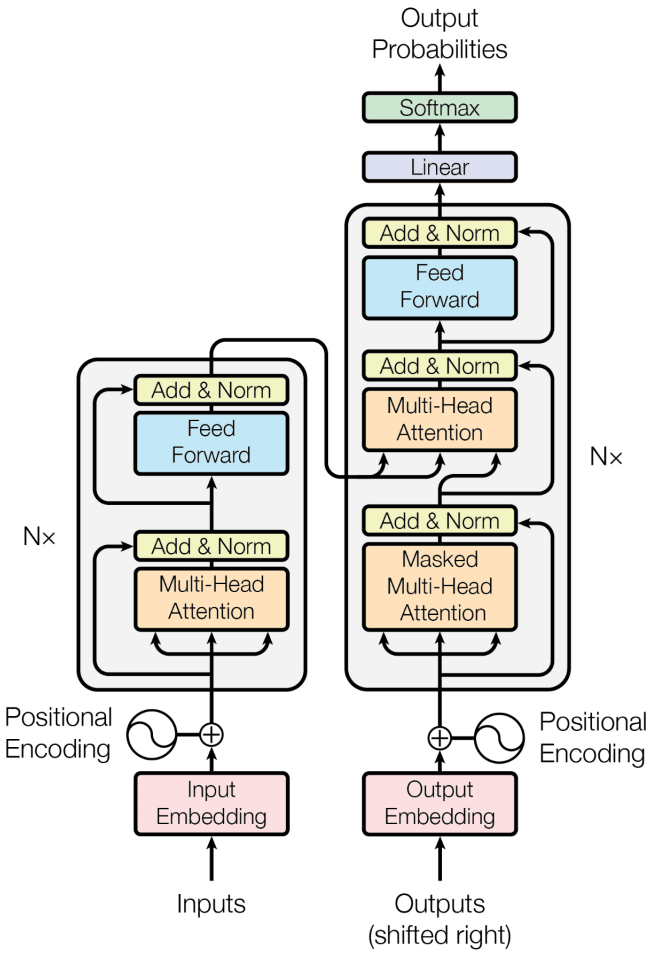


Figure 7.2: Diagram showing the encoder and decoder stacks in the Transformer model.

mensions). Each subsequent encoder block accepts as input the output embedding vectors of the previous encoder block. Additionally, both the attention layer and the feed-forward layer in each encoder block have layer normalization applied to their outputs. The output vectors of the last encoder on the block are transformed into a set of attention vectors. These vectors are passed to the decoder block.

Also visible in Figure 7.2 is the positional encoding operation applied to the embedding vector. The positional encoding generates a vector which allows a model to learn the distance between different tokens in the sequence, which it can use once the vectors are projected into their respective query (Q), key (K) and value (V) vectors in the attention layers. The attention layer, shown in Figure 7.3,[7] is used to

[7]Diagram taken from Vaswani et al. (2017)

relate parts of the sequence with each other by calculating a score between $token_a$ (query) and $token_b$ (key) representing the relative importance of $token_b$ to $token_a$. Multi-head attention splits the token vectors into fixed chunks, which allows for the model to attend to different kinds of dependencies, allowing for capturing many linguistic features and giving the model greater representation power.

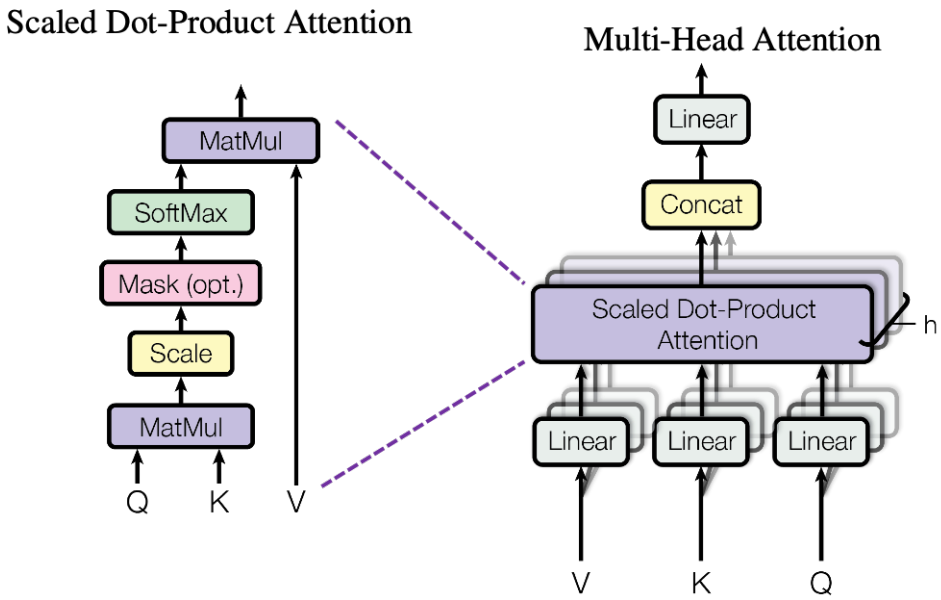


Figure 7.3: Scaled dot-product and multi-head attention in the self-attention function. The dot-product between Q and K is calculated (MatMul), giving a relative attention weight to each token from the token at that position. These attention weights are first subjected to a scaling factor (Scale), with an optional mask step applied to prevent leftward information flow (Mask) and then normalized with a softmax function (SoftMax), and then finally multiplied by V, the value vector (MatMul). These weighted values are summed together to achieve a weighted vector representation of the sequence, which is passed to subsequent layers.

Sequence-to-sequence tasks require a decoder block, which contains much of the same architecture and components as the encoder block, although there are some key differences. As Figure 7.2 shows, there are two attention layers in the decoder block. The first attention layer is a self-attention layer, as with the encoder block. It functions slightly differently, however, as this layer can only attend to tokens in the sequence that it has already seen before. All tokens in the sequence after this current point are masked. The second attention layer is an encoder-decoder attention layer which is used to help the decoder focus on certain tokens in the input sequence.

The figure shows this layer receiving input from the encoder block in the form of attention vectors K and V, which are used in the place of the key and value vectors in the attention function.

The last elements to discuss are the linear layer and softmax layer which generate the output. The linear layer is a connected neural network that projects the output vector of fixed size into a logits vector which is the same size as the vocabulary learned from the training data. Each cell in the logits vector corresponds to a token in the vocabulary, and the linear layer gives each token a certain score based on the output vector. The softmax layer converts those scores into probabilities, and the token with the highest probability is selected as the output token for this step of the time sequence. The decoder block receives the output from the previous time step in the form of an input sequence, which is then transformed into an embedding vector and a positional vector as with the encoding block. This process continues until a special end-of-sequence symbol is produced, indicating that the decoder has finished outputting tokens.

The decoder block as described here is employed in sequence-to-sequence tasks, such as the MT models described in Chapter 6. For classification tasks, such as the MWE identification task described in this and the next chapter, only the encoder block is used, and a classification layer is applied to the final output vector.

7.4.2 Pre-trained Language Models

As mentioned above, pre-trained neural language models have taken the NLP community by storm, being used for more and more tasks in NLP, and often outperforming previous SOTA results for these tasks (Radford et al., 2018). These models trained on large quantities of unannotated data can then be fine-tuned on a smaller, task-specific dataset for downstream tasks such as Named Entity Recognition or question answering. The most prominent of these language models is the BERT model and its multilingual version mBERT, which both make use of bidirectional Transformer encoders (Devlin et al., 2019).

Pre-trained word embeddings have their roots in early neural language models, such as that of Bengio et al. (2001), where each input was a vector representation of the previous words, which were looked up in a table. In 2013 *word2vec* was introduced (Mikolov et al., 2013b,a), which was a more efficient vector representation of words than what had come previously. This new method for representing words as vectors allowed for large-scale training of word vectors, which could be used in many downstream tasks. These pre-trained word embeddings could even represent semantic relationships between words, which made them viable in few-shot and zero-shot[8] cross-lingual transfer, by projecting word embeddings from different languages into the same learning space (Conneau et al., 2017; Artetxe et al., 2018).

These earlier pre-trained word embeddings were not dependent on context however. For example, the vector representations for the word "break" in each of the following sentences would be the same.

(144) It's important to take a **break** sometimes.

(145) He didn't expect the chair to **break** when he sat on it.

(146) I finally got my lucky **break**.

(147) **Break** ups are always hard.

Following the release of the Transformer, research began on developing the architecture for pre-trained language models that could be used in classification tasks, similar to how pre-trained word embeddings improved performance on many NLP tasks. The key difference between how these words representations were generated lies in how BERT embeddings are dynamically generated for each word in its own context, while *word2vec* and similar word embeddings create an embedding for the word based on the context seen in training, and the same word embedding updated when a word is encountered in a new context. With the release of ELMo (Peters et al., 2017), OpenAI Transformer Generalised Pre-trained Transformer (GPT)

[8]Zero-shot learning takes place when the model is forced to generalize on a category of samples that have not been previously seen before.

(Radford et al., 2018) and GPT-2 (Radford et al., 2019), language models were shown to be effective in many downstreams tasks, such as POS-tagging, NER, textual entailment, question answering, and other language modelling tasks. In order to ensure that words did not see themselves in the bidirectional context, vectors had to be forward generated only, meaning that context beyond the word in question could not be attended to. While ELMo used a bidirectional LSTM to allow for backwards generation to take place, the process of forward and backward generation had to take place separately and could not be generated at the same time. Enter BERT.

7.4.3 BERT Architecture

Bidirectional Encoder Representations from Transformers (BERT) is trained in two steps: an unsupervised *pre-training* step where unlabeled text data is used to train on two tasks, and a supervised *fine-tuning* step, where downstream tasks are trained using labelled data. The two pre-training tasks BERT is trained on are masked language modelling (masked LM), and next-sentence prediction (NSP), which trains BERT in recognising the relationships between sentences. Unlike the LSTM architecture of ELMo, BERT uses masking of tokens to prevent words from "seeing themselves" in the context of the other words.

BERT's architecture is similar to the encoder stack discussed in Section 7.4.1, however, with larger parameters than the vanilla Transformer model. BERT has no need for a decoder stack, as it is not predicting tokens. The input sequence begins with a special classification `[CLS]` token. Each word passed through the encoder block generates an output vector of fixed size (e.g. 768 in `Bert Base`). The output vector for the first classification token is then used in many classification tasks as input, such as sentence classification or sentence-pair classification tasks. For a sequence tagging task such as NER or MWE identification, the output vector for each token after the classification token is passed through a classification layer, and the labels are predicted from these output encodings. Figure 7.4 represents the architecture for this sequence tagging task, showing how the tokenised input is

encoded into embeddings, passed into the encoder block, and the output generated from this block is passed into the classification layer, where labels are predicted.

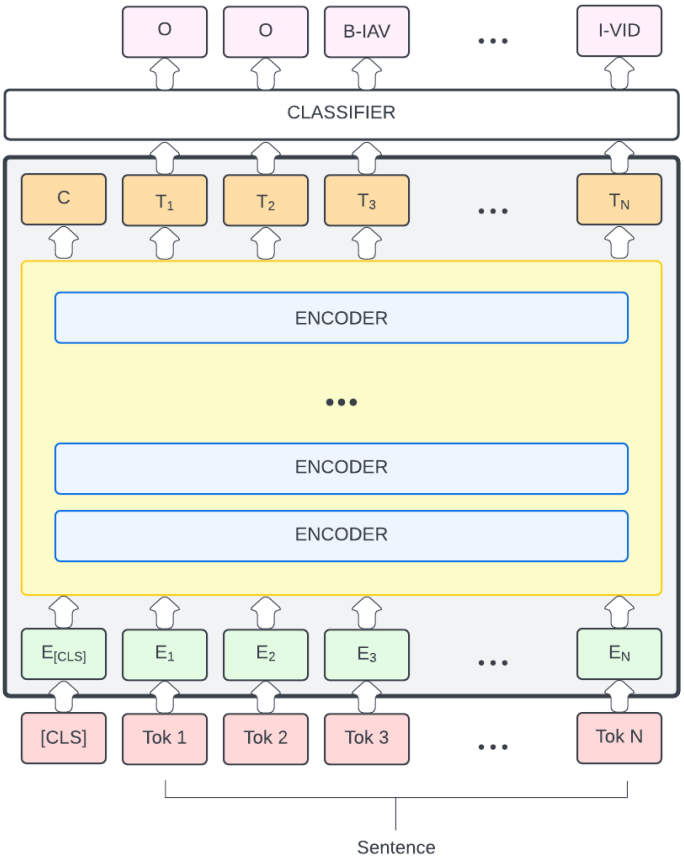


Figure 7.4: Diagram of BERT model with linear classification layer.

Performance-wise, BERT outperforms previous SOTA results (Devlin et al., 2019) on a number of NLP tasks in the GLUE benchmark (Wang et al., 2018), and indeed appears to capture a number of features of language such as part-of-speech, morphology and syntactic relations (Hewitt and Manning, 2019; Jawahar et al., 2019), although the claim that these models "know syntax" is challenged in recent research (Sinha et al., 2021). Despite these disputed claims, Pires et al. (2019) demonstrate that this model performs surprisingly well in zero-shot learning, generalizing information across languages, even those written in different scripts, which consequently have no lexical overlap.

7.4.4 GaBERT

For lower-resourced languages such as Irish, multilingual models like mBERT allow for training on NLP tasks even with limited amount of data (Wu and Dredze, 2020). However, models such as the French model CamemBERT (Martin et al., 2020) or the Dutch model BERTje (de Vries et al., 2019) demonstrate the benefits of training such Transformer-based *monolingual* languages models for many NLP tasks.

gaBERT (Barry et al., 2022) is a monolingual BERT model for Irish trained on a selection of Irish texts amounting to approximately 7.9 million sentences. Training followed largely the same process and hyperparameters as BERT, using a smaller batch size due to the smaller size of available memory. It was evaluated on three downstream tasks: dependency parsing, a cloze test, and MWE identification (the latter task is the focus of Chapter 8). The results for these tasks were compared with mBERT and a monolingual Irish model WikiBERT (trained on Irish Wikipedia data). It was found that for parsing accuracy and a masked token prediction task, gaBERT was more effective than an off-the-shelf version of either of the other two models. However, when compared to an mBERT model that had been subjected to continued pre-training on the same data used to train gaBERT, this version of mBERT slightly outperformed gaBERT at the token prediction task. We leave discussion on the performance of the gaBERT model and the mBERT model on the task of MWE identification for the next chapter.

7.4.5 Pre-trained Language Models for the Task of MWE Identification

It is clear that pre-trained language models are continuing to take the NLP world by storm, and the task of MWE identification is no different. In the most recent edition of the PARSEME shared task, for identifying unseen MWEs, the best performing system relied on BERT pre-trained language models (Taslimipoor et al., 2020), integrating dependency parse information as a jointly learned task to render the system

semi-supervised. This system proved effective at capturing MWE information in a manner similar to other linguistic information apparently captured by the language model. Similar to this task, Peters et al. (2017) employed bi-directional language models with a sequence tagging model, and evaluated the model on NER and a chunking task, which significantly outperformed previous SOTA models.

With this in mind, we conduct our own series of experiments, investigating the use of pre-trained language models for the task of identification of Irish MWEs. These experiments are presented in the next chapter. We use the HuggingFace Transformers library (Wolf et al., 2020) for our experiments. Both the mBERT[9] and gaBERT[10] models are available on the HuggingFace repository. To evaluate our models, we make use of both the evaluation library provided by *seqeval*, as well as the evaluation algorithm used in the PARSEME Shared Task, which was described in Section 7.3.1.

7.5 Irish vMWE Identification in the PARSEME Shared Task

Irish was included for the first time in Edition 1.2 of the PARSEME Shared Task, with the creation of the Irish PARSEME Corpora (Walsh et al., 2020). Of the nine submitted systems for the task of vMWE identification, six systems reported evaluation scores on Irish. The systems that attempted this task on the Irish corpus included both the closed-track systems: ERMI (Yirmibeşoğlu and Güngör, 2020) and Seen2Seen (Pasquer et al., 2020b), and also four of the open-track systems: MTLB-STRUCT (Taslimipoor et al., 2020), Seen2Unseen (Pasquer et al., 2020a), TRAVIS-multi (Kurfalı, 2020), and MultiVitaminBooster (Gombert and Bartsch, 2020).

[9] `https://huggingface.co/bert-base-multilingual-cased`
[10] `https://huggingface.co/DCU-NLP/bert-base-irish-cased-v1`

Architecturally, both MTLB-STRUCT and TRAVIS-multi make use of pre-trained multilingual BERT models, like those discussed in Section 7.4.3. MTLB-STRUCT takes a multi-task learning approach, and jointly learns vMWE-identification as well as dependency parsing. TRAVIS-multi is one of the two systems submitted by the same team, the other being TRAVIS-mono, with both systems using BERT models with a connected linear layer to perform token classification. TRAVIS-multi uses BERT multilingual model mBERT, while TRAVIS-mono use monolingual BERT models that have been trained for the participating languages. At the time of this shared task, there was no monolingual model trained for Irish, so the TRAVIS team only attempted this task with TRAVIS-multi. MultiVitaminBooster likewise utilises pre-trained contextual word embeddings, combining the word embeddings of another language model XML-RoBERTa with a set of linguistic features.

Seen2Unseen relies on the closed-track system Seen2Seen, which identifies vMWEs previously seen during training by using relatively simple extraction and filtering techniques, based on 8 interpretable binary parameters. In combination with the Seen2Seen model, Seen2Unseen uses lexical replacement, translation and statistical ranking to identify vMWE candidates that did not occur during model training. Finally, ERMI (Embedding-Rich Multiword expression Identification) uses a bidirectional LSTM-CRF architecture, which uses as input the embedding of the word, its POS tag, dependency relation and head word.

7.5.1 Results of Systems on Irish Dataset

The results for Irish are captured in Figure 7.5,[11] which reports the precision, recall and $F1$ scores in three sub-categories of the task.

For unseen vMWEs, the system with the highest $F1$ score of 19.54 was MTLB-STRUCT. This is considerably lower than the cross-lingual macro-average unseen

[11]Table taken from PARSEME Results page `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_50_Shared_task_results`.

System	Track	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
ERMI	closed	14.79	6.98	**9.48**	1	32.62	13.99	**19.58**	2	69.71	21.15	**32.45**	1
Seen2Seen	closed	100	0.33	**0.66**	2	77.17	16.28	**26.89**	1	80.1	16.3	**27.09**	2
MTLB-STRUCT	open	23.08	16.94	**19.54**	1	37.72	25	**30.07**	2	65.02	33.79	**44.47**	1
Seen2Unseen	open	21.74	9.97	**13.67**	2	44.16	23.39	**30.58**	1	50.41	24.11	**32.62**	2
TRAVIS-multi	open	3.75	1.99	**2.6**	3	12.36	5.05	**7.17**	3	65.48	16.3	**26.11**	3
MultiVitaminBooster	open	0	0	**0**	n/a	0	0	**0**	n/a	0	0	**0**	n/a

Figure 7.5: Precision, Recall and $F1$ scores for each sub-category in the PARSEME Shared Task 1.2 for Irish.

vMWE $F1$ score of 38.53 for the same system. MTLB-STRUCT also had the highest $F1$ score (44.47) when calculated using the Token-based metric, but was just outperformed by Seen2Unseen $F1$ score, calculated using the MWE-based metric (30.07 vs 30.58) (see Section 7.3.1 for specifics on the differences between these metrics). TRAVIS-multi generated the highest Token-based precision score (65.48), but its recall and precision scores (and thus, its $F1$ scores) for the MWE-based metric and the unseen MWE-based metric were considerably lower than both MTLB-STRUCT and Seen2Unseen. MultiVitaminBooster did not achieve any score in this task for Irish, which may be related to issues we later discuss in Chapter 8.

Both of the closed track models achieve Token-based and MWE-based $F1$ scores that are comparable to the open-track models, and indeed the Token-based precision scores are higher than the best performing open track models. We can see from the precision of both the global MWE-based and Token-based scores of Seen2Seen that this language-agnostic model appears to perform well at the task of correctly identifying vMWEs in Irish that have been previously seen. Seen2Seen notably performs poorly at the task of identifying unseen vMWEs, which is inherent to the model's design.

In comparison with the other languages in this shared task, the systems performed the most poorly on the Irish dataset. In contrast, systems performed better on the Hindi dataset, which was actually smaller than the Irish dataset (1684 sen-

tences in Hindi, and 1700 sentences in Irish). As shown in Table 7.6,[12] the highest unseen MWE-based $F1$ score for Hindi was 53.11, a difference of 33.57 points over Irish, and the highest Token-based $F1$ was 79.17, a difference of 34.70 points.

System	Track	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
ERMI	closed	37.09	41.67	**39.25**	1	63.48	56.32	**59.69**	1	79.48	62	**69.66**	1
Seen2Seen	closed	0	0	**0**	n/a	86.56	39.23	**53.99**	2	90.53	39.25	**54.76**	2
MTLB-STRUCT	open	48.75	58.33	**53.11**	1	72.25	75.04	**73.62**	1	81.2	77.24	**79.17**	1
Seen2Unseen	open	43.71	41.67	**42.66**	2	67.63	58.99	**63.02**	2	73.46	60.68	**66.46**	2
TRAVIS-multi	open	29.21	43.33	**34.9**	3	51.43	50.82	**51.12**	3	69.08	57.06	**62.5**	3
MultiVitaminBooster	open	0.72	1	**0.84**	4	0.96	0.59	**0.74**	5	10.94	7.79	**9.1**	4
TRAVIS-mono	open	8.33	0.33	**0.64**	5	60.71	2.53	**4.85**	4	86.96	2.78	**5.39**	5

Figure 7.6: Precision, Recall and $F1$ scores for each sub-category in the PARSEME Shared Task 1.2 for Hindi.

As a point of comparison, Hebrew proved to be the most similar to Irish in terms of system performance. As shown in Table 7.7,[13] the highest unseen MWE-based $F1$ score was 19.59, a difference of only 0.05 points over Irish, and the highest Token-based $F1$ being 52.99, a difference of 8.52 points.

System	Track	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
ERMI	closed	11.49	6.62	**8.4**	1	41.81	24.85	**31.17**	2	46.72	26.22	**33.59**	2
Seen2Seen	closed	0	0	**0**	n/a	65.84	31.81	**42.9**	1	68.37	30.93	**42.6**	1
MTLB-STRUCT	open	25.53	15.89	**19.59**	1	56.2	42.35	**48.3**	1	64.81	44.82	**52.99**	1
TRAVIS-multi	open	21.14	12.25	**15.51**	2	51.9	35.39	**42.08**	2	61.1	37.28	**46.3**	2
Seen2Unseen	open	3.88	1.66	**2.32**	3	44.36	33.6	**38.24**	3	47.51	32.73	**38.76**	3

Figure 7.7: Precision, Recall and $F1$ scores for each sub-category in the PARSEME Shared Task 1.2 for Hebrew.

There are a number of possible reasons for the poor performance shown by all systems on the Irish dataset. An obvious issue is the relatively small number of vMWEs present in the data: the Irish dataset has a total number of 662 annotated

[12]See footnote 11
[13]See footnote 11

vMWEs, with a density of 0.4 vMWEs per sentence, compared to 1,034 vMWEs annotated in Hindi, with a density of 0.6 vMWEs per sentence. However, the Hebrew dataset contains more annotated vMWE examples than Hindi, with 2,533 annotated vMWEs (although the density was significantly lower, with an average of 0.1 vMWEs per sentence), and yet systems performed better on the Hindi dataset.

Another point of interest is the rate of unseen vMWEs with regards to the training and development set. The highest rate for any language was Irish, at 0.69, followed by Hebrew, at 0.60. The average rate across all languages was 0.33, meaning Irish had a rate of unseen vMWEs that was over double the average rate. The high rate of unseen vMWEs is likely due to the small number of MWEs annotated in total, but also may be due to variability in Irish MWEs themselves. In particular, the category of verbal idioms makes up 16% of the annotated examples (e.g. *bheith chomh ramhar le taobh an fhalla* (lit. be as thick as the side of the wall) 'be furious'). Such constructions tend to present the most syntactic and lexical diversity for any category. Additionally, the fact that the Irish language makes frequent use of verbal nouns for both the infinitive form and the continuous tense, may lead to further variability in the annotation of vMWEs in both verbal and nominal forms. For example, the light verb *déan iarracht* 'make an attempt' may be present in the data in a variety of ways:

(148) **Rinne** mé **iarracht** mór inné .
[Verb] [Noun]
I **made** an big **attempt** yesterday.

(149) Is gá di **iarracht** a **dhéanamh** .
[Noun] [Noun: Infinitive-form]
There is a need for her to **make** an **attempt**.

(150) Bíonn tú ag **déanamh** **iarrachta** gach lá .
[Noun: Verbal-noun] [Noun]
You do be **making** an **attempt** every day.

(151) Tá sár-**iarracht** **déanta** aige .
[Noun] [Adjective: Participle-form]
He has **made** a great **attempt**.

The high rate of unseen vMWEs may also be a product of the IUDT (Chapter 5), which is a balanced corpus with a mixture of domains and genres. As a result of this variety in domain and genre, there is likely to be greater diversity in the types of MWEs represented. While such diversity gives a valuable representation of a wide range of Irish text, when combined with the smaller size of the data, this can lead to data sparsity. With all of these factors combined, it is perhaps unsurprising that the systems struggled to perform as well on the the Irish dataset as on other, larger languages, with lower rates of unseen vMWEs, and less data sparsity in vMWE types represented also.

Based on the results of this shared task, we decided to follow the approach described by TRAVIS, using both the multilingual mBERT model as well as the newly trained monolingual Irish language model gaBERT, and attempt to improve on the baseline performance they report. Our experiments investigate whether it is possible to achieve competitive results with the other models submitted for this task, notably MTLB-STRUCT, by exploring model parameters, and adjusting the data.

7.6 Summary

In this chapter, we introduce the task of MWE identification, and its core concepts. We outline some of the challenges that make this task difficult, and sketch out some of the methods of approaching this problem. Additionally, we examine in detail two shared tasks which are focused on MWE identification, and draw insight from the results of these shared tasks.

We describe how the task of MWE identification is evaluated, noting the shortcomings in the traditional precision, recall and $F1$ measures used to evaluate other sequence tagging tasks. We compare the evaluation metrics used by both the DiMSUM and the PARSEME shared tasks.

We discuss the Transformer-based pre-trained language models, and their rele-

vance to many tasks in the field of NLP. We explain the architecture of the Transformer model, and in particular, the attention mechanism. We introduce the BERT model and the ways in which it is superior to previously used pre-trained language models. We also introduce gaBERT, a monolingual BERT model for Irish.

Finally, we examine the most recent edition of the PARSEME shared task (Edition 1.2), which saw the inclusion of Irish for the first time. The systems that attempted this task for the Irish dataset are described, and the results of these systems' performance on the Irish test set are considered. The performance on the Irish dataset was worse than on similarly sized datasets for both Hindi and Hebrew. Some potential reasons for the performance demonstrated by the systems on the Irish dataset are presented.

In our next chapter, we will describe the series of experiments we conducted investigating the use of both the multilingual BERT model (mBERT) as well as the monolingual Irish BERT model (gaBERT), comparing how these models perform against the state-of-the-art in MWE identification for Irish.

Chapter 8

A System for Automatic Identification of Irish MWEs

> *There are no safe paths in this part of the world. Remember you are over the Edge of the Wild now, and in for all sorts of fun wherever you go.*
>
> J.R.R. Tolkien, *The Hobbit*

8.1 Introduction

In our previous chapter, we explored the task of MWE identification, and in particular, the results on the Irish dataset for the systems submitted to the PARSEME shared task Edition 1.2. Drawing from insights in this shared task, and employing the pre-trained language models previously described, in this chapter we report on our experiments in building our own system for the automatic identification of Irish MWEs.

8.2 MWE Fine-tuning Experiments

Following the example of two well performing systems in Edition 1.2 of the PARSEME shared task, we explore the task of identification of verbal MWEs by applying fine-tuned pre-trained language models, and investigate how the systems' performance varies with certain design decisions. Particular consideration was given to how lower-resourced languages can maximise on smaller datasets and exploit existing resources to achieve better results in this task.

8.2.1 Data

As per the PARSEME shared task, we use the annotated portion of the Irish corpus of verbal MWEs, described in Chapter 5. With the dataset consisting of 1700 sentences, and only 662 MWEs annotated, the dataset can be considered a low-resource language for this task, when compared with the dataset sizes of other participating languages.

The data was split into training, development and testing datasets by the PARSEME shared task organisers, according to their two prerequisites: 1) to have a sufficient number of unseen vMWE examples in each test set, and 2) to adapt their split strategy for languages that had no new data with regards to previous editions of the shared task (Ramisch et al., 2020). Their process split the corpora so as to have at least 300 MWEs in the test set that did not appear in either the training or the development set, and at least 100 MWEs in the development set that did not occur in the training data. According to these metrics, the Irish corpus was split with 257 examples in the training data (and 100 MWE examples), a development set of 322 examples (and 126 MWE examples), and a test set of 1120 examples (and 433 MWE examples). While this split allows for fair comparison with other participating languages, it only generates a very small training corpus for this task for Irish.[1]

[1]One sentence was excluded from the data as its length excluded the maximum size for an input embedding.

One of the design questions during the initial planning of this experiment was whether to compare model performance based on the performance of the development set or the test set. This is a question of note among the NLP community, particularly with the recently released series of "We need to talk about" papers (van der Goot, 2021), which challenge conventions in the field of using either the development set or the test set for comparing model performance, increasing the risk of "bias from research design" (Hovy and Prabhumoye, 2021).

We use the development dataset to compare model performance during the hyperparameter tuning of Series 1 of the experiments, and once a model had been selected, the test data was used to report on model performance to compare it to the other systems from the PARSEME Shared Task, as well as to compare models trained on different datasets in Series 2 of the experiments.

8.2.1.1 Data Format and Pre-processing

The PARSEME corpora are annotated for verbal MWEs using `cupt` format,[2] which merges CoNLL-U[3] and parseme-tsv[4] format. `cupt` follows the guidelines of CoNLL-U Plus format, with the differences firstly of a mandatory metadata field `source_sentence_id`, which is optional in CoNLL-U Plus, and secondly of an optional metadata field `sent_id`, which is mandatory in CoNLL-U Plus.

The experiments required converting the `cupt` format into a `json` file of tokens with their corresponding MWE labels. The labelling scheme used was a modified *IOB2* labelling scheme, and so the `cupt` annotations were converted using a script. One long sentence consisting of 311 tokens was removed from the training data. Tokenisation was performed using BertTokenizerFast for the mBERT model, which is in turn based on WordPiece. The gaBERT model uses SentencePiece tokenisation (Kudo and Richardson, 2018). Where the input token is split into many sub-tokens,

[2] `https://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_45_Format_specification`

[3] `https://universaldependencies.org/format.html`

[4] `https://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation`

we wish to pass only the encoding of the first sub-token to the linear layer classifier, with the predicted class standing for the entire input token. We achieve this by assigning the non-initial sub-tokens a special label to be ignored during the classification step.

Several processing steps were performed to alter the datasets for the second series of experiments (see Section 8.2.3.1 for more details).

When evaluating the output of the MT models, the `json` format was converted back to `cupt` format. For converting the predicted data back to a `cupt` format for the PARSEME evaluation, a script was devised to interpret the MWE labels used for the training data into the labelling used in `cupt` format, a step which required some slight interpretation of the predicted labels, particularly in the case of single-token predictions, i.e. the system predicts an MWE label on a single token, with no other tokens of the same label within the sentence. Any MWE tokens not tagged as the initial token in the group were tagged as belonging to the closest group of the same category with a tagged initial token. Where a token was predicted as belonging to an MWE group but there was no clear MWE group for it to belong to, we assumed it was the initial MWE of a new group.

In addition to this step, a post-processing script was optionally applied to remove single-token predictions, which are more likely to be noisy labels. This script is described further in Section 8.2.4.4.

8.2.2 Known Issues Influencing Experiment Design

8.2.2.1 Model Performance Instability

One of the issues faced in these experiments was the instability that models can display during training, in particular, the tendency for some models to learn to predict only one class. Table 8.1 shows the impact on precision, recall and $F1$ scores displayed by varying just the random seed when fine-tuning an mBERT model for 20 epochs. The table shows that certain random seeds will result in a system that

fails to predict any MWE labels at all, resulting in scores of zero for precision, recall and $F1$.

Run	Precision	Recall	$F1$
1	0.3288	0.2330	0.2727
2	0.3158	0.2330	0.2682
3	**0.0**	**0.0**	**0.0**
4	0.2870	0.1602	0.2056
5	0.3401	0.2427	0.2833
6	0.2566	0.1408	0.1818
7	**0.0**	**0.0**	**0.0**
8	0.2727	0.1602	0.2018
9	0.3008	0.1942	0.2360
10	0.2966	0.1699	0.2160

Table 8.1: Variable performance on fine-tuning mBERT, with different random seed variables

This behaviour is likely exacerbated by the data used in this experiment, in which MWE labels are scarce. Of the 38,909 tokens total, only 1,499 or 3.9% of them were labelled with any MWE label. However, this tendency for instability is a known issue in the topic of training Neural Networks in NLP (Devlin et al., 2019; Dodge et al., 2020; Mosbach et al., 2020), and the choice of random seed used to initialise the training weights, particularly with a small sized dataset, can lead to a significant difference in model performance. Given that lower resourced languages are likely to have a small training data size, it is important to consider the effects of such behaviours.

As further evidence of the challenges a smaller dataset presents, we ran a number of experiments using the the French annotated corpus of verbal MWEs[5] in order to assess whether this is a general problem or a language specific problem. Table 8.2 displays the results for six experiments performed on this dataset, fine-tuning an mBERT model. For each run of the experiment, we decreased the number of training examples used, adding these examples to the tuning dataset instead.

There appears to be a minimum size of training dataset, below which the model

[5] `https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2/FR`

No. training examples	Precision	Recall	$F1$
2302	0.7500	0.600	0.6667
1860	0.6821	0.5568	0.6131
1395	0.4893	0.5016	0.4954
930	0.5532	0.4517	0.4973
465	0.3712	0.1414	0.2048
232	0.000	0.000	0.000

Table 8.2: Precision, recall and $F1$ scores for models trained on decreasing amounts of training data for French dataset.

is unlikely to predict any MWE labels. Considering that the original split used for the Irish corpus in the PARSEME experiments had a training set size of 257, the chances of the model failing to learn to predict any MWE could likely be reduced with a slightly larger training set.

8.2.2.2 Labelling Schemes for Sequence Labelling Task

Sequence labelling tasks require each token to be tagged with a single label. This can pose a problem in tasks where the token may have more than one label, as with certain cases of MWEs. Typically for chunking tasks such as NER, a system of tagging known as Inside-Outside (IO) or Inside-Outside-Beginning (IOB) tagging is used (Ramshaw and Marcus, 1995).

Any element that is inside the chunk is tagged with 'I', while outside elements are tagged with 'O'. For example, the vMWE 'let out' is tagged as:

(152) **Let out** a scream
`I I O O`

In cases where a second MWE chunk immediately follows the first, the 'B' tag indicates a new chunk is beginning. An alternative tagging scheme *IOB2* uses the 'B' tag for the beginning of each new chunk, regardless of which tokens proceed it. Examples 153 and 154 demonstrate *IOB* and *IOB2* tagging respectively.

(153) She **takes up water skiing**
`O I I B I`

(154) She **gives up hot dogs**
`O B I B I`

IOB, *IOB2* and other variants of this tagging scheme are commonly used in NER (Ratinov and Roth, 2009), as well as related chunking tasks such as terminology extaction and shallow parsing. However, more so than with other sequence labelling tasks, MWEs present difficulties such as discontinuity and nesting (see Chapter 2). Tagging schemes such as *IOB* and its variants are not ideally suited to this property of MWEs. Additionally, the PARSEME annotation guidelines allow for annotation of tokens that have more than one MWE label (Example 155).

(155) '**Let** the cat **out** of the bag'
`VID;VPC.semi VID VID VID;VPC.semi VID VID VID`

Schneider et al. (2014a) discuss this issue with *IOB* tags, proposing modified versions of this tagset that allow for capturing gaps, nested expressions, and distinctions between strong and weak MWEs. Their proposed *gappy 1-level* scheme uses 6 labels: `B,I,O,b,i,o`. The lowercase tags represent tokens occurring within a gap (in the case of `o`) or nested MWEs (in the case of `b` and `i`). Examples 156 and 157 demonstrate how discontinuity in MWEs is captured in this annotation scheme.

(156) **Give** something your **best shot**
`B o o I I`

(157) I **looked** the **worst-case scenario up**
`O B o b i I`

However, this method does not account for instances of overlapping or discontinuity in nested MWEs, as with, for example, 'he **kept** the **clean sweep** his mother had **made in mind**', where the expression 'keep in mind' contains a nested discontinuous expression 'make {a} clean sweep'. This issue is addressed by Berk et al. (2019), who introduce the modified scheme *bigappy-unicrossy*. The tagset for this scheme uses the same labels as with *gappy 1-level* (`B,I,O,b,i,o`), however the use of the lower-case labels is changed slightly. This tagset allows for one level of nested MWEs, two levels of discontinuity of MWEs (including nested discontinuous MWEs), and one level of crossing MWEs.

`B` is used for the beginning of a chunk, including single-token chunks. `I` is used for the inside of a chunk, where `I` follows `B`. `O` is used for tokens outside of the chunks. The `o` tag is used for tokens that occur between tokens of a chunk, or all *gaps* between chunks, whether that gap occurs at the MWE level or the nested MWE level. Similar to their corresponding uppercase tags, `b` is used for tokens at the beginning of a nested chunk, and `i` is used for tokens of the chunk following a `b` tag. The above expression is annotated in Example 158 according to *bigappy-unicrossy*.

(158) He **kept** the **clean sweep** his mother had **made in mind**
`O B o b i o o o i I I`

This annotation scheme only partially solves the challenge of overlapping MWEs, as it cannot handle overlapping MWEs with a shared token. For example, the sentence 'I made not only changes but also additions' could be annotated in two ways. In Example 159, the MWE chunk 'made additions' is omitted, while with Example 160, the MWE chunk 'made changes' is omitted. Similarly, in our previous example 'let the cat out of the bag', the secondary MWE 'let out' cannot be captured.

(159) I **made not only changes but also additions**
`O B b i I i i O`

(160) I **made not only changes but also additions**
`O B b i o i i I`

To resolve this issue, we propose a modified version of both the *IOB2* and *bigappy-unicrossy* tagging schemes, which we call *IOB2-double* or *IOB2-d*, and *bigappy-unicrossy-double* or *bigappy-unicrossy-d*. This modification is described further in Section 8.2.3.2.

8.2.3 Experiment Design

These experiments are organised into two series, in order to focus on two different streams of tuning: (i) hyperparameter tuning, and (ii) data optimisation. These experiments are intended to be exploratory investigations into the effect of modifying hyperparameters for training models, and altering data structure and labelling.

As such, we focus our analysis on any patterns that emerge, and suggesting best practises for attempting this task, particularly for low-resourced languages.

We use the HuggingFace Transformers library (Wolf et al., 2020), which provides both the mBERT[6] and gaBERT[7] models, and allows for integration with their tokenising library.

8.2.3.1 Series 1: Model Hyperparameters

Hyperparameters are configurations of a model that can be altered to optimize the model to train on certain data and for certain tasks, or, as Bengio (2012) puts it: "annoying knobs to be adjusted". These settings differ from model parameters that are tuned during the training process. Instead, hyperparameters are determined prior to training, and may be continuous values, such as setting the learning rate of the model or its weight decay, or discrete values, such as the number of epochs trained for, the number of layers in the model, or selecting the model itself.

Practically, the number of hyperparameters that may be adjusted and the ranges in which these hyperparameters best affect the model's performance vary depending on the task and the data at hand. The search space for the optimal settings of these hyperparameters increases exponentially with each additional feature that needs to be tuned. As such, users of machine learning algorithms may simply opt to use default hyperparameter settings that have been shown to be effective in a number of other tasks. Probst et al. (2019) assess the importance of various hyperparameter settings across many NLP tasks, and across different Machine Learning algorithms. They provide optimised default values for many hyperparameters in this space. Likewise, Bengio (2012) list several hyperparameters relevant for training gradient-based models, highlighting the **initial learning rate**, **mini-batch size**, and **number of training iterations**, among others, as key hyperparameters to tune. Some hyperparameters are seen to have a direct impact on the performance on the training data, while others impact the training process more, causing longer training times

[6]Available at `https://huggingface.co/bert-base-multilingual-cased`.
[7]Available at `https://huggingface.co/DCU-NLP/bert-base-irish-cased-v1`.

and taking up more resources. Bengio (2012) also differentiate between **model hyperparameters**, which define the structure of the model itself, and **optimization hyperparameters**, which relate to how the model learns patterns in the data. Data hyperparameters are considered another type of hyperparameter that can be tuned through various pre-processing steps, but we do not consider these optimisation steps here.

During the fine-tuning of BERT, Devlin et al. (2019) select three commonly-preferred optimisation hyperparameters for tuning. Those are the **learning rate**, the **batch size** and the **number of epochs**, noting that smaller datasets appear more sensitive to hyperparameter variation than large ones. In addition to these three hyperparameters, we consider three model hyperparameters: the **model**, the **number of layers fine-tuned on**, and the **initial random seed value**. We examine each of these hyperparameters further, highlighting their importance in this experiment.

Learning rate: This hyperparameter in particular is recommended by Bengio (2012) as perhaps the single most important hyperparameter to tune. At each step in the training, the model calculates a loss, or a calculation of the error rate of the model compared to the training data. The learning rate defines the step size towards the minimum loss or amount the model changes at each stage of the search process, usually a positive float varying between 0.0 and 0.1.

A lower learning rate (e.g. $2e-5$) can help prevent catastrophic forgetting (McCloskey and Cohen, 1989), a phenomenon by which previously learned information is erased during learning of new knowledge (see Chapter 6). However, too small of a learning rate will slow down the learning of a model, and it may get stuck at a sub-optimal point, unable to achieve a lower error rate.

Number of epochs and batch size: These hyperparameters are somewhat linked, as both hyperparameters affect the training time of the model. The number of epochs determines the number of times the learning algorithm will have worked

through the entire training data. Within each epoch, the model is updated after each batch of data is processed, known as an iteration. The batch size determines the number of samples contained in each batch. Batch size can range from 1 up to the size of the entire training dataset.

During training, we expect the loss to decrease through repeated exposure to the data. The longer we train for, i.e. the higher the number of epochs, we can expect the loss to continue to decrease as the accuracy of the model increases. However, there must reach a point where both the loss and accuracy stabilise. Selecting the number of epochs is important to avoid either under-fitting, where the minimum loss is not reached, as well as over-fitting, where the model has learned the data well but has trouble generalizing outside of this specific dataset. This is also called the bias-variance tradeoff.

The batch size will affect computational speed and affect how quickly the algorithm converges. A larger batch size means the algorithm sees more samples at each stage of training, and fewer iterations are necessary to see all the training samples. This increase in computation speed comes at the cost of an increase in memory used. A smaller batch size means more noise is introduced at each stage of training, as the samples viewed are smaller. However, the benefit of this is that noisy data can help the algorithm to "jump out" of a bad local minimum.

The selection of model: This is itself a hyperparameter, and as mentioned in our description of the pre-trained language models in Chapter 7, there is a noticeable difference in the performance of monolingual and multilingual language models. We compare the performance of the mBERT and gaBERT models on this task, and in particular, note which patterns of hyperparameter tuning are consistent across both models, and where the models differ in performance.

In series 2 of the experiments, we repeat all the experiments with the mBERT model in order to demonstrate how these techniques can be applied and an improvement in model performance can be achieved, regardless of whether a low-resource

language has access to a pre-trained monolingual language model or not.

Number of layers fine-tuned: This hyperparameter can be specified beforehand. Both mBERT and gaBERT models have 12 layers in total. These layers can each be "frozen" during the fine-tuning phase, that is, the weights of these layers are not updated during backpropagation calculations. This can help to speed up model training, as there are fewer computations to make at each step. Tasks that are similar to the original task that was used to train the language model may not require fine-tuning on many, or indeed any of the layers, and can achieve good results with fine-tuning only the classification layer. However, the more dissimilar a task is to the original one that the model was trained on, the more likely it is the layers of the language model, particularly the top layers, will need further tuning to solve the task.

Random seed variable: This hyperparameter is used to initialise a pseudo-random number generator, which is in turn used to set model parameters such as the initial weights of the nodes. Unlike other hyperparameters discussed, there is no intuitive link between one random seed value and the performance of the model it initialises. However, it has been shown to have a significant effect on model performance, and particularly with fine-tuned language models such as BERT. As such, it is a hyperparameter to be tuned as any other. This effect of the random seed variable on model performance can be seen with the instability issue discussed in Section 8.2.2.1.

These hyperparameters can be adjusted using a variety of tuning strategies, which vary in complexity from manually setting variables, to a simple automated random search, to more optimized methods such as Bayesian optimisation or Population based training. Despite research exploring these complex optimisation algorithms, simple methods such as manual search or grid search remain popular, as they allow for a degree of insight into the tuning process, and have a low technical

barrier to optimisation. Bergstra and Bengio (2012) consider random search to be superior to grid search, in that it offers many of the same benefits, but is much more efficient when tuning in a large dimension search space.

For our experiments, we begin with a manual tuning in order to analyse the effect of each setting on the model performance. We employ a random search algorithm in the same search space to verify the results we find.

When selecting the range of values for each hyperparameter, we first consider the range of values used by Devlin et al. (2019). They found the following range of values worked well across a number of tasks:

1. Number of epochs: 3, 4
2. Batch size: 16, 34
3. Learning rate: $5e-5$, $3e-5$, $2e-5$

This is quite limited search space of hyperparameters, but offers some guidance as to where to begin our search. Given the limited computational resources available to us, a selection of batch sizes in a lower range seems appropriate. To compensate for the smaller size of the dataset, we expand the number of epochs trained for. We also look at a larger range of learning rates. This gives us the following search space for these three parameters:

1. Number of epochs: 5, 10, 15, 20, 25, 30, 35, 40
2. Batch size: 1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20
3. Learning rate: $1e-6$, $2e-6$, $1e-5$, $2e-5$, $1e-4$, $2e-4$, $1e-3$, $2e-3$, $1e-2$, $2e-2$, 0.1, 0.2

In terms of number of layers, we opted for three settings: (i) freezing all 12 layers, (ii) freezing layers 1-8 and training on layers 9-12, (iii) training on all layers. For random seeds, we train across 20 values. As there is no relationship between size of the random seed and the performance of the model, we opted to select multiples

Hyperparameter	Default Value
Number of epochs	20
Batch size	8
Learning rate	$2e-5$
Random seed	10

Table 8.3: Default values used for each hyperparameter when not being tuned.

of 5. However, we postponed tuning on the random seed variable until the other parameters had been tuned, as there are no discernible patterns to observe in the model's behaviour.

As we opted to tune each hyperparameter individually, it became necessary to select a default value for each of the hyperparameters not being tuned. We selected our default values somewhat randomly, tending towards a value close to the median, and balanced against the concerns highlighted above. The default values for each hyperparameter set at the beginning of the experiment is given in Table 8.3. The results of these experiments are explored in Section 8.2.4.1.

8.2.3.2 Series 2: Data Structuring and Optimisation

As a low-resourced language, the availability of data greatly impacts on a number of tasks in Irish NLP, including the task of MWE identification. As such, this series of experiments is focused on data structuring and optimisation steps, with the goal of understanding where potential issues may lie, and how the data can be used to its full potential in order to achieve competitive results with other languages.

Section 7.5.1 analysed the performance of systems on the Irish dataset, and posited a number of potential reasons for the poor performance seen across these systems. Those issues can be summarised broadly as issues of a **large tagset**, **complexity in the data**, and a **scarcity of training samples**. We address these three issues in our second series of experiments.

The large tagset: We address this issue by merging the subcategories for both the 'LVC' and 'VPC' vMWE types. By replacing the 'LVC.full' and 'LVC.cause'

labels with 'LVC', and replacing the 'VPC.full' and 'VPC.semi' labels with 'VPC', we reduce the number of tags in the tagset from 7 to 5. We investigate whether this reduction in complexity improves the evaluation scores, particularly the precision of the models.

We perform a second experiment addressing this issue with a more aggressive merging of labels, where each category is replaced with a simple 'MWE' label, reducing the number of tags in the tagset to 1.

Complexity of the data: We address data complexity through the removal of two difficult labels. The first label is the contentious inherently reflexive verbs (IRV) label, which was included in the construction of the Irish corpus for Edition 1.2 of the PARSEME shared task, but the appropriateness of this label was later called into question in Walsh et al. (2020). This label occurs only six times and only in the test set, yet removing the label from the tagset further simplifies the task, reducing the tag space. The second label we remove is the verbal idioms (VID) label, the most syntactically and lexically diverse category of vMWE. The 'VID' label occurs 14 times in the training dataset, and 22 times in the development dataset, making it the fourth most frequently occurring label, behind 'IAV', 'LVC.full', and 'LVC.semi'. Of those occurrences, 75% (27) of the 'VIDs' occurred only once in the training and development datasets. The vMWE with the largest number of tokens per vMWE (six tokens) also belongs to the 'VID category'.

Size of the dataset: An obvious obstacle to optimising the performance of systems on this task is the small size of the dataset. This decreases the number of *seen vMWE* examples and subsequently increases the number of more challenging *unseen vMWE* examples to be identified in the test set. The rate of unseen vMWEs in the training data may be the largest contributing factor for the reduced capability of systems performing the task of vMWE identification on Irish data. Additionally, the smaller dataset contributes to the issue of instability mentioned in Section 8.2.2.1, as well as increasing the model's sensitivity to hyperparameter values shifting, and

thus the dramatic range in performance seen in experiment Series 1.

We devised the following experiments to address these issues:

- Experiment 2A reduces the number of tags through first merging the two fine-grained labels ('LVC.full' and 'LVC.cause' → 'LVC'; 'VPC.full' and 'VPC.semi' → 'VPC')
- Experiment 2B merges all tags into a single 'MWE' tag
- Experiment 3 reduces the complexity of the data through removing two challenging vMWE labels ('IRV' and 'VID')
- Experiment 4 increases the size of the training and development datasets through re-splitting these datasets into datasets of about equal number of examples, with 216 vMWEs annotated in the training data (+116 vMWEs), 213 vMWEs annotated in development data (+87 vMWEs) and 230 vMWEs in the test data (-203 vMWEs)[8]

Choice of labelling scheme: The final variable we examine in this series of experiments is the labelling scheme used. As discussed in Section 8.2.2.2, Example 161 remains an issue, as the problem of double-tagged vMWEs is not addressed by the other proposed labelling schemes.

(161) '**Let** the cat **out** of the bag'
`VID;VPC.semi VID VID VID;VPC.semi VID VID VID`

For Series 1 of our experiments, we adopted a version of *IOB2* tagging, using the 'O' tag to annotate intervening tokens in discontinuous vMWEs. In the case of double tagging, as with example 161, the first tag applied to the token was maintained. For Series 2 of our experiments, we employ an alternative version of *IOB2* tagging that we call *IOB2-double*. This tagging scheme uses the same tags as *IOB2* (`B, I, O`), however the 'B' tag may now also be used on the second token of

[8]As sentences contained unequal numbers of MWEs per sentence, splitting the data into datasets of exactly equal number of MWEs proved impossible without reordering the data.

MWEs that share an initial token with another MWE. While this does not perfectly capture the embedded vMWE it does allow for both labels to be applied in the case of crossing vMWEs sharing a token, by treating the second token in the embedded vMWE as a single-token vMWE whose verbal head is implied. We apply this same strategy to the *bigappy-unicrossy* scheme proposed by Berk et al. (2019), creating a version *bigappy-unicrossy-double*. Examples 162, 163 and 164 demonstrate how the embedded vMWE 'let out' in 'let the cat out of the bag' is annotated in each of these annotation schemes.

(162) '**Let** the cat **out** of the bag'
`B-VID I-VID I-VID I-VID I-VID I-VID I-VID`
IOB2 tagging

(163) '**Let** the cat **out** of the bag'
`B-VID I-VID I-VID B-VPC.semi I-VID I-VID I-VID`
IOB2-double tagging

(164) '**Let** the cat **out** of the bag'
`B-VID i-VID i-VID b-VPC.semi I-VID I-VID I-VID`
bigappy-unicrossy-double tagging

We compare the performance of models trained on data tagged using these three tagging schemes across each of the experiments highlighted above.

8.2.4 Results and Analysis

This section looks at the results of the experiments described above, reporting on the evaluation metrics used in Chapter 7.

8.2.4.1 Experiment Series 1: Hyperparameter Optimisation

We plot the results of the initial hyperparameter tuning in Figures 8.1, 8.2 and 8.3. For each chart, we plot the value of the hyperparameter being tuned against the $F1$ score achieved on the development set for that model. Given the scarcity of vMWE labels in the data, the $F1$ score generally gives a better measurement of performance on this task than accuracy. As mentioned in Section 8.2.3, for each language model,

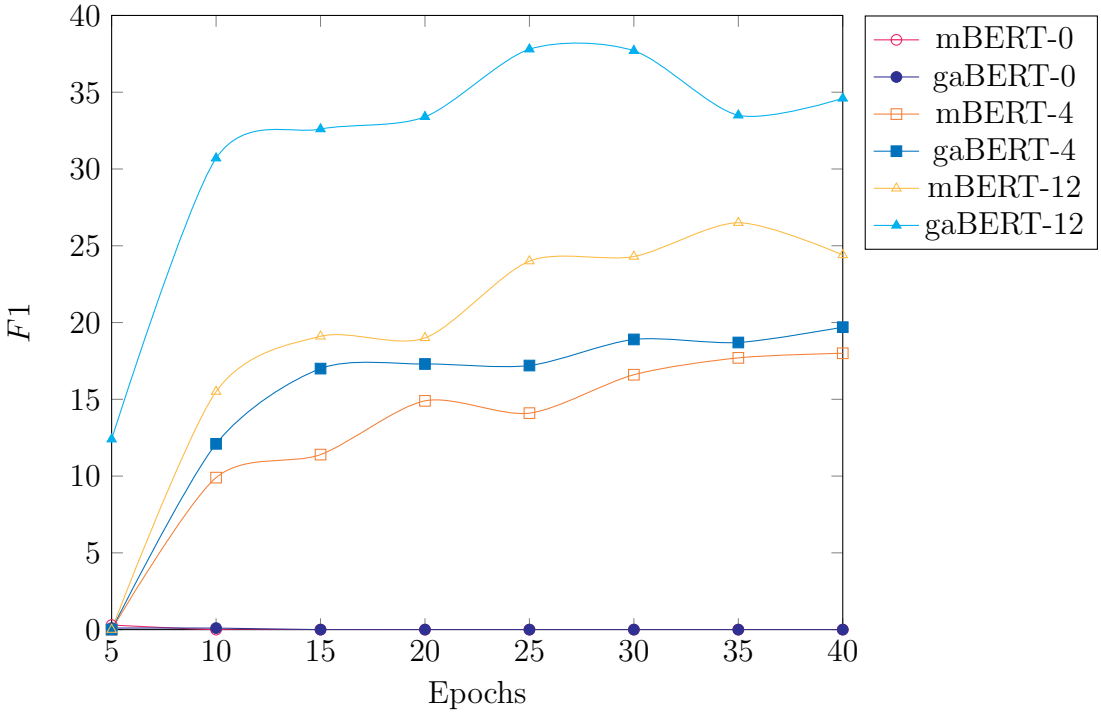


Figure 8.1: $F1$ scores of all models when trained on different numbers of epochs.

we trained on three layer settings, denoted here as mBERT-0 and gaBERT-0 (all layers frozen), mBERT-4 and gaBERT-4 (layers 1-8 frozen), and mBERT-12 and gaBERT-12 (trained on all 12 layers). From the plots, patterns in hyperparameter tuning become evident.

It is immediately evident from these three plots, and in particular the results shown by mBERT-0 and gaBERT-0, that freezing all 12 layers is not helpful for model performance. In fact, the best results for both language models are achieved when all 12 layers are fine-tuned (mBERT-12 and gaBERT-12). The exception to the trend of poor model performance when freezing all layers occurs when the learning rate is increased to between $1e$-3 and $2e$-1. In this range, both language models achieve an $F1$ score that is more competitive with the scores achieved by the language models when fine-tuning on most or all of the layers.

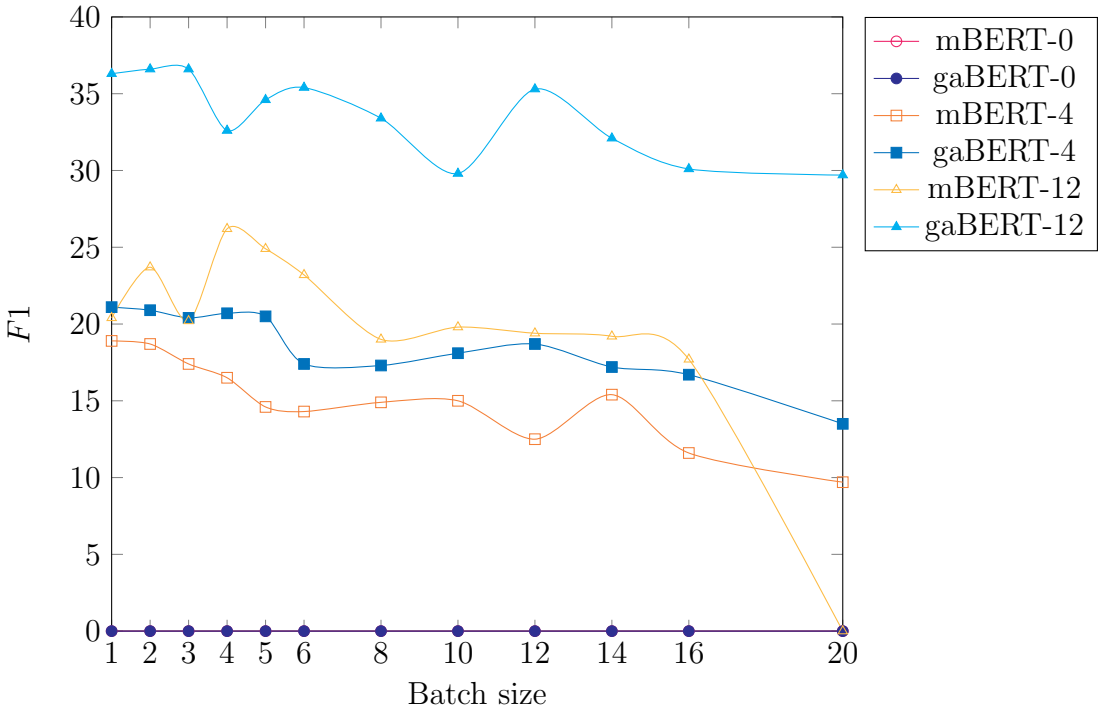


Figure 8.2: $F1$ scores of all models when trained with different batch sizes.

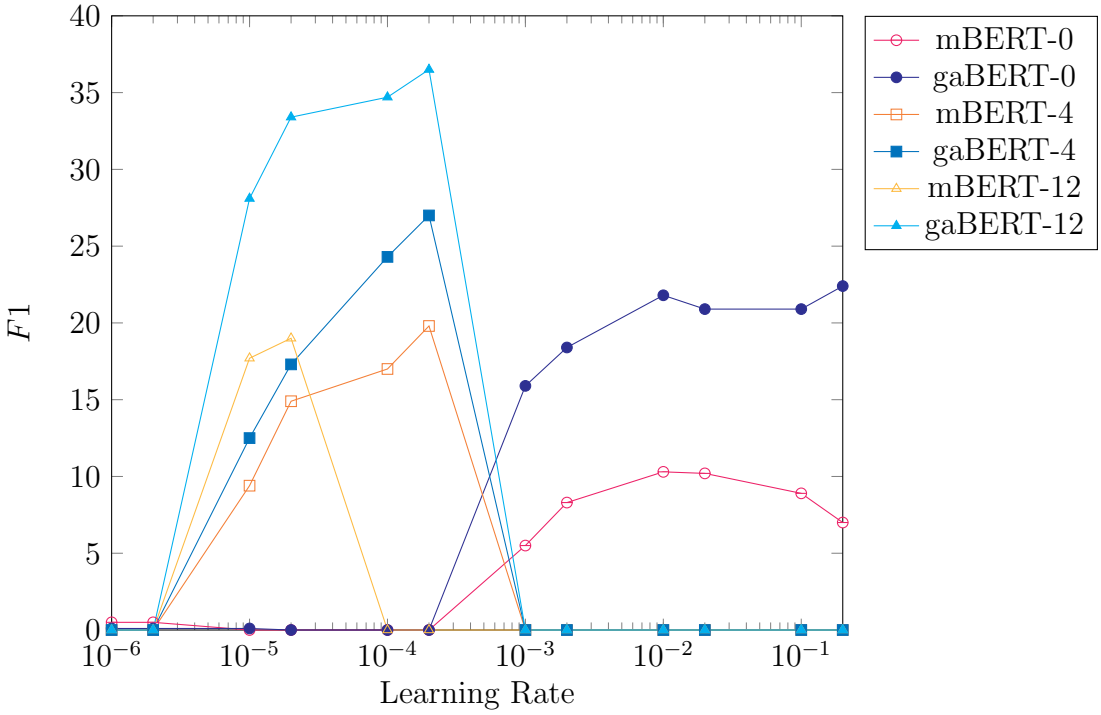


Figure 8.3: $F1$ scores of all models when trained with different learning rates.

It appears from Figure 8.1 that as the number of epochs increases, the $F1$ score increases also. However, the increase appears to begin stabilising at approximately 25 epochs, after which the increase in performance must be balanced against the increase in resources needed for longer training. Figure 8.2 demonstrates how increasing the batch size beyond a certain point actually decreases the $F1$ score, with the highest $F1$ score for each model being with a batch size between 1-4. Of note is the drop of the $F1$ score to 0.0 when mBERT-12 is trained with a batch size of 20. This model did not finish training as the GPU ran out of memory at this point, indicating that further increasing the batch size would likely be impossible without adequate resources.

The data in Figure 8.3 was taken from a number of sample learning rates across a logarithmic scale. Due to the limitations of tuning a continuous value using discrete samples, the plot is an estimated curve along the few data points provided. Given the limited range of values plotted here, a second tuning experiment was devised, expanding the values of the learning rate where the highest $F1$ scores appear to be concentrated. Figure 8.4 shows how the $F1$ scores of mBERT-0 and gaBERT-0 vary as the sample learning rates increase from 0.002 to 0.8. Figure 8.5 shows the $F1$ scores of mBERT-4, gaBERT-4, mBERT-12 and gaBERT-12 at sample learning rates from $2e$-5 to $8e$-4.

From the above results, it appears that the best performing language model for both mBERT and gaBERT is achieved when training on all 12 layers. For each experiment, we selected the hyperparameter settings that yielded the best model and trained an "optimised" mBERT and gaBERT model, varying the random seed with this experiment. However, when fine-tuning the gaBERT model on these new hyperparameters, we found that while the batch size of 2 performed well with the learning rate of $2e$-5, and the learning rate of $2e$-4 performed well with the batch size of 8, setting the batch size at 2 and the learning rate at $2e$-4 yielded a model

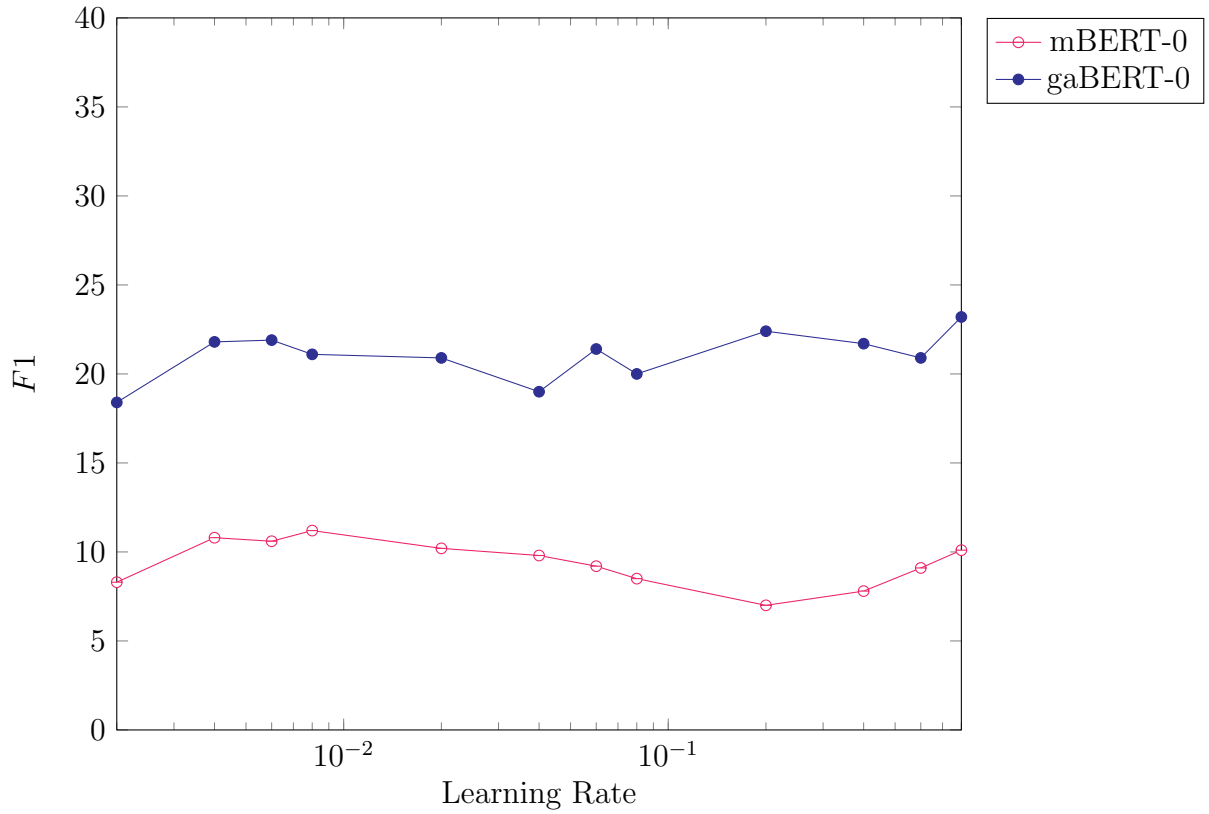


Figure 8.4: $F1$ scores of mBERT-0 and gaBERT-0 with new range of learning rates.

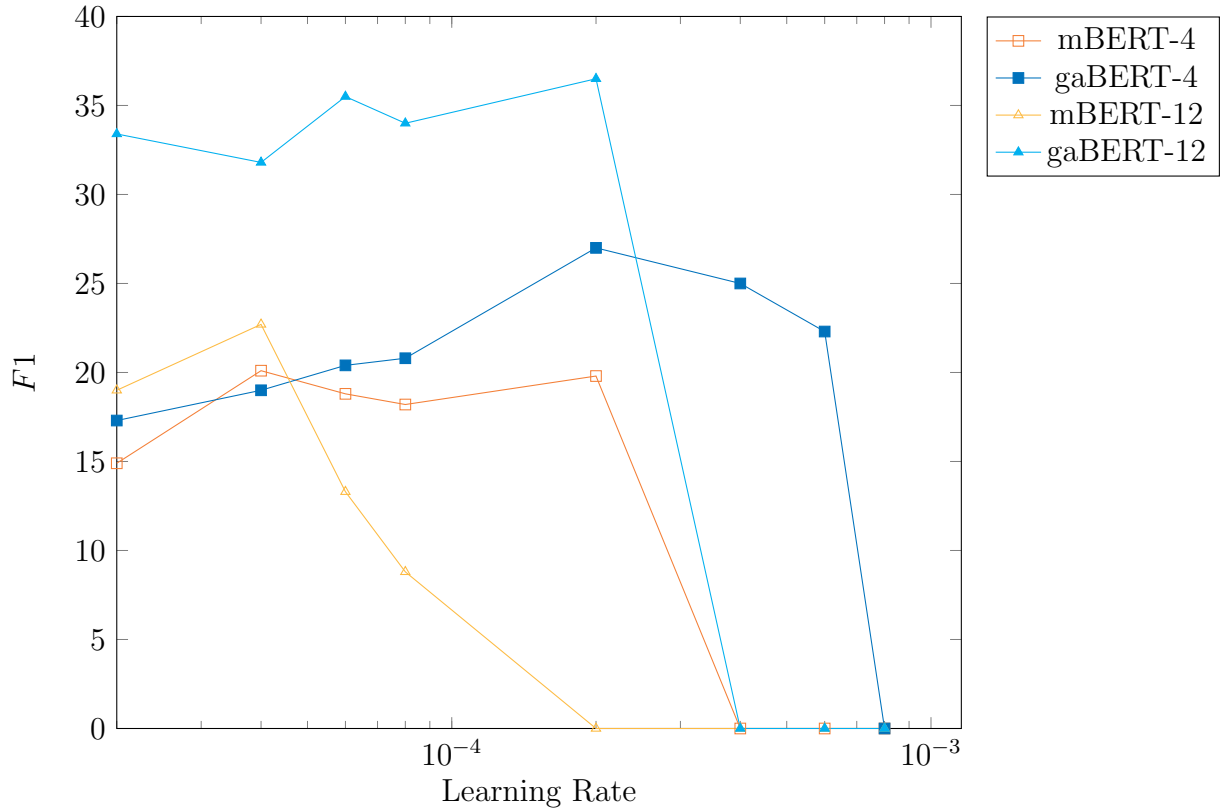


Figure 8.5: $F1$ scores of mBERT-4, gaBERT-4, mBERT-12 and gaBERT-12 with new range of learning rates.

Parameter	mBERT-12	gaBERT-12-rate	gaBERT-12-batch
Number of epochs	30	30	30
Batch size	4	8	2
Learning rate	4e-5	2e-4	2e-5

Table 8.4: Hyperparameter settings for random seed optimisation.

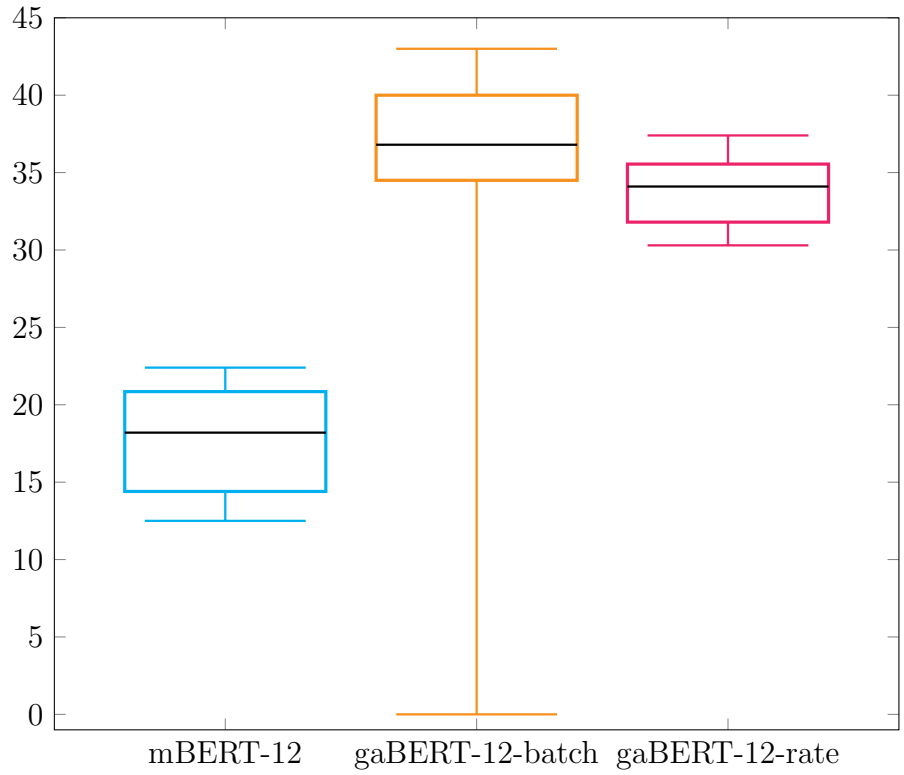


Figure 8.6: Box plot of $F1$ scores generated by mBERT-12, gaBERT-12-batch and gaBERT-12-rate models trained across 20 random seed values.

that failed to predict any vMWEs at all.

Following this, we opted to fine-tune two gaBERT models, one with the best performing batch size of 2, and the default learning rate of $2e$-5 (gaBERT-12-batch), and one with the best performing learning rate of $2e$-4 and the default batch size of 8 (gaBERT-12-rate). Table 8.4 lists the hyperparameters used for these three models. We compare the $F1$ scores across these random seeds for both models, as well as the mBERT-12 model in Figure 8.6.

The results show that the gaBERT model generally outperforms the mBERT

model. From Figure 8.6, we can see the best performing gaBERT model was achieved by gaBERT-12-batch, however this model demonstrates once again the instability problem, with one of the random seeds resulting in a model that failed to predict any MWEs. The F1 scores for both gaBERT models are much higher than for the mBERT-12 model, and the distribution of F1 scores is narrower. On average, the gaBERT-12-batch performs better than gaBERT-12-rate. We select gaBERT-12-batch at random seed 10 for our "optimised" gaBERT model, and compare the results with mBERT-12 at random seed 75. The full results as generated by the *seqeval* evaluation metrics are given in Tables 8.5 (mBERT) and 8.6 (gaBERT).

Category	Precision	Recall	$F1$	#Samples
Micro	0.3192	0.2021	0.2475	668
Macro	0.2394	0.1613	0.1897	668
Weighted	0.3015	0.2021	0.2386	668
IAV	0.3238	0.2112	0.2556	161
IRV	0.0000	0.0000	0.0000	7
LVC.cause	0.3733	0.2500	0.2995	112
LVC.full	0.3098	0.2426	0.2721	235
VID	0.2188	0.0654	0.1007	107
VPC.full	0.4500	0.3600	0.4000	25
VPC.semi	0.0000	0.0000	0.0000	21

Table 8.5: Precision, recall and $F1$ scores for optimised mBERT model, giving the micro, macro and weighted average, and a breakdown per category. Support refers to # of MWEs in gold test dataset.

Category	Precision	Recall	$F1$	Support
Micro	0.4952	0.3862	0.4340	668
Macro	0.3767	0.2556	0.2832	668
Weighted	0.4573	0.3862	0.4029	668
IAV	0.4940	0.2547	0.3361	165
IRV	0.0000	0.0000	0.0000	7
LVC.cause	0.5968	0.6607	0.6271	112
LVC.full	0.5000	0.5404	0.5194	235
VID	0.1887	0.0935	0.1250	107
VPC.full	0.8571	0.2400	0.3750	25
VPC.semi	0.0000	0.0000	0.0000	21

Table 8.6: Precision, recall and $F1$ scores for optimised gaBERT model, giving the micro, macro and weighted average, and a breakdown per category. Support refers to # of MWEs in gold test dataset.

For a comparison to the results of the shared task explored in Chapter 7, we use the evaluation script for the PARSEME shared task to evaluate our system output. Table 8.7 lists both the MWE-based and Token-based evaluation metrics for the optimised **mBERT** model, while Table 8.8 displays these results for the optimised **gaBERT** model.

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.1609	0.1293	**0.1434**	433	348
Token-based	0.5661	0.2896	**0.3831**	1005	514
IAV (MWE-based)	0.2250	0.1552	0.1837	116	80
IAV (Token-based)	0.4194	0.2194	0.2881	237	124
IRV (MWE-based)	0.0000	0.0000	0.0000	6	0
IRV (Token-based)	0.0000	0.0000	0.0000	14	0
LVC.cause (MWE-based)	0.1515	0.1370	0.1439	73	66
LVC.cause (Token-based)	0.5046	0.2865	0.3654	192	109
LVC.full (MWE-based)	0.0676	0.0735	0.0704	136	148
LVC.full (Token-based)	0.4080	0.2867	0.3654	286	201
VID (MWE-based)	0.1667	0.0725	0.1010	69	30
VID (Token-based)	0.4583	0.1048	0.1705	210	48
VPC.full (MWE-based)	0.3529	0.3000	0.3243	20	17
VPC.full (Token-based)	0.6400	0.4000	0.4923	40	25
VPC.semi (MWE-based)	0.0000	0.0000	0.0000	13	7
VPC.semi (Token-based)	0.1429	0.0385	0.0606	26	7
Unseen MWE-based	0.0731	0.0736	**0.0733**	299	301

Table 8.7: Precision, recall and $F1$ scores for optimised **mBERT** model, giving their MWE- and Token-based scores, and a breakdown per category. Unseen MWE-based scores refer to the prediction of MWEs that did not occur in either the training or developmental data. # Gold refers to the number of MWEs or MWE-Tokens in the gold test data, while # Pred refers to the number of MWEs or MWE-Tokens predicted by the system.

Comparing the results of both the PARSEME and *seqeval* evaluation metrics, we can see immediately that the number of MWE samples counted in the gold test dataset differs, with the *seqeval* metrics counting 668 MWEs in total, and the PARSEME evaluation metrics counting just 433. This discrepancy is due to discontinuous MWEs which are counted as separate MWE groups by the *seqeval* metrics, while the PARSEME metrics count all MWEs of the same group regardless if there are interleaving tokens, or gaps. As such, the PARSEME evaluation metrics

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.4167	0.3580	**0.3851**	433	372
Token-based	0.6813	0.4786	**0.5622**	1005	706
IAV (MWE-based)	0.3099	0.1897	0.2353	116	71
IAV (Token-based)	0.6019	0.2616	0.3647	237	103
IRV (MWE-based)	0.0000	0.0000	0.0000	6	0
IRV (Token-based)	0.0000	0.0000	0.0000	14	0
LVC.cause (MWE-based)	0.5055	0.6301	0.5610	73	91
LVC.cause (Token-based)	0.6288	0.7500	0.6841	192	229
LVC.full (MWE-based)	0.3210	0.3824	0.3490	136	162
LVC.full (Token-based)	0.5235	0.5070	0.5151	286	277
VID (MWE-based)	0.1190	0.0725	0.0901	69	42
VID (Token-based)	0.3765	0.1524	0.2169	210	85
VPC.full (MWE-based)	0.8333	0.2500	0.3846	20	6
VPC.full (Token-based)	0.8333	0.2500	0.3846	40	12
VPC.semi (MWE-based)	0.0000	0.0000	0.0000	13	0
VPC.semi (Token-based)	0.0000	0.0000	0.0000	26	0
Unseen MWE-based	0.3201	0.3244	**0.3223**	299	303

Table 8.8: Precision, recall and $F1$ scores for optimised **gaBERT** model, giving their MWE- and Token-based scores, and a breakdown per category. Unseen MWE-based scores refer to the prediction of MWEs that did not occur in either the training or developmental data. # Gold refers to the number of MWEs or MWE-Tokens in the gold test data, while # Pred refers to the number of MWEs or MWE-Tokens predicted by the system.

are likely a better measure of model performance for this task, and we prioritise reporting these results for Series 2.

When these results are compared with those reported by the PARSEME shared task in Chapter 7, we can see that our mBERT model ranks third for both the Unseen MWE-based score and the Global MWE-based score. It ranks second for the Global Token-based score. Our gaBERT model outperforms the other systems for all three metrics, beating the $F1$ score of the best system by 12.69 points. The improved ability of our gaBERT model to predict unseen MWEs when compared to our mBERT model ($F1$ score of 32.23 vs 7.33) is particularly interesting when comparing these models.

8.2.4.2 Random Search Optimisation Test

Following these experiments, we performed a sanity check using a hyperparameter optimisation tool Ray Tune (Liaw et al., 2018), in order to investigate alternative combinations of the hyperparameters tuned. Tune is a hyperparameter tuning library built on Ray, and can be integrated into Python code to help with the laborious process of hyperparameter tuning. We ran 20 trials on each model, selecting our values in the follow ranges:

1. Number of epochs: random value from all number of epochs
2. Batch sizes: random value from all batch sizes
3. Learning rate: random float sampled in log space $2e-6$–$4e-4$
4. Random seeds: random int sampled in range 1–1000

We plot the $F1$ scores for each model trained on random hyperparameters selected from these ranges, displaying these results in Figure 8.7. We can see that the results fall within the $F1$ scores generated from the manual tuning above, with some of the trials generating models that performed very close to the optimised models reported above. In particular, trial #17 generated a gaBERT model that achieved an $F1$ score of 41.4 on the test data, compared to the optimised model which achieved an $F1$ score of 43.4.

The best performing mBERT model generated by random search optimisation achieved an $F1$ score of 20.4, compared to our manually optimised model, which achieved an $F1$ score of 24.8.

8.2.4.3 Experiment Series 2: Data Optimisation

Series 2 of experiments are organised into 3 categories of experiments. Experiment 1 compares the results of the best performing model from Series 1 when trained

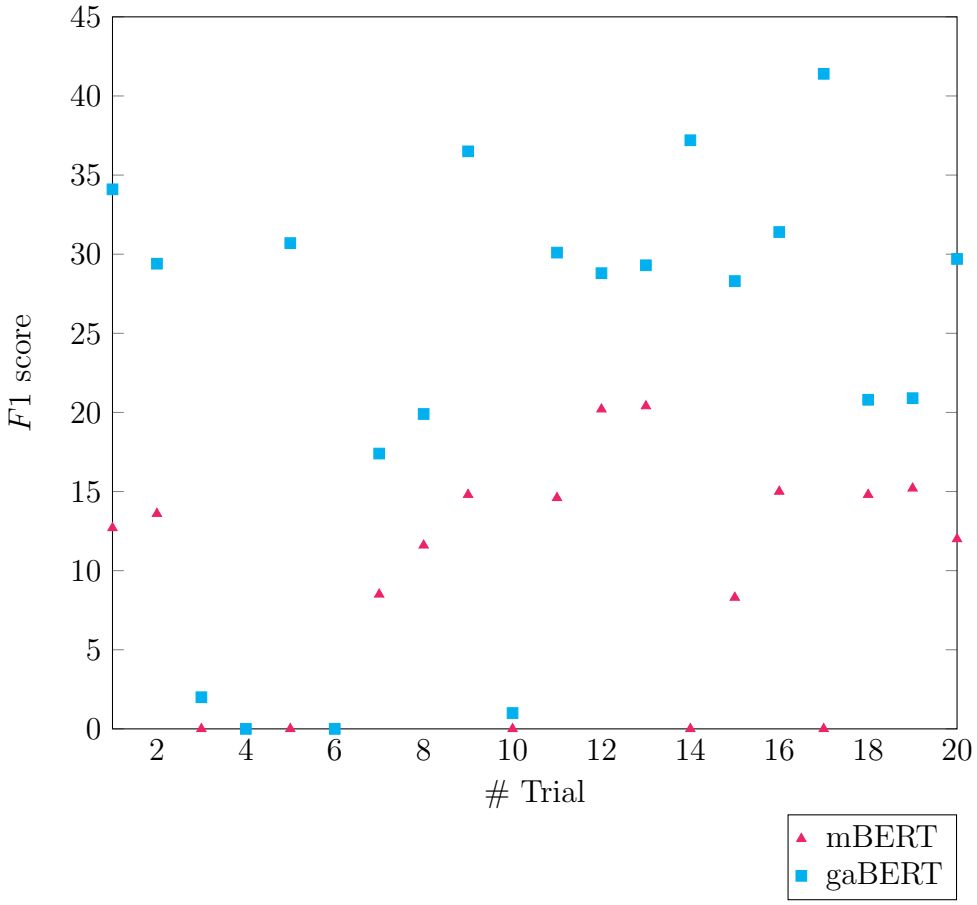


Figure 8.7: $F1$ scores of mBERT and gaBERT models generated by random search optimisation test. Each trial represents a model trained on a random combination of hyperparameters from the specified range.

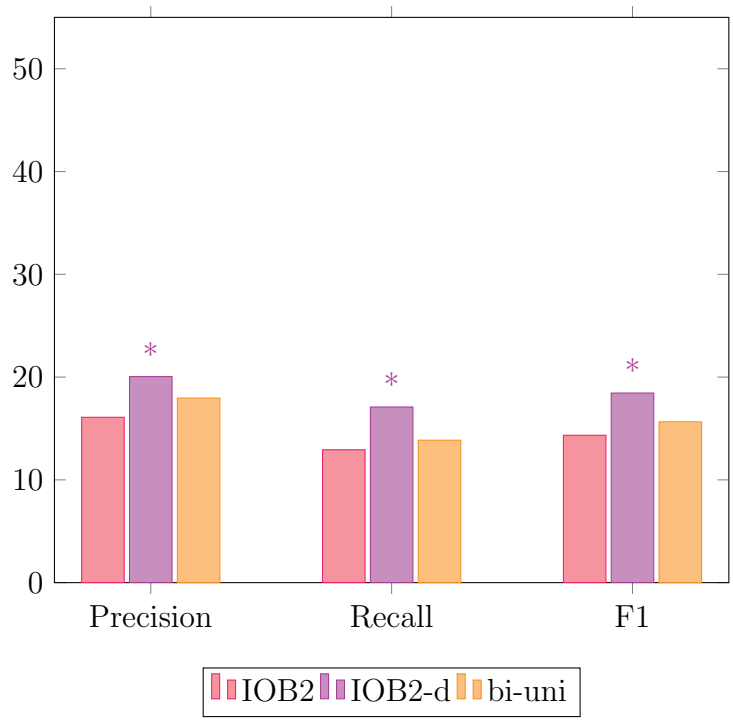


Figure 8.8: Exp 1: Precision, Recall and F1 score for mBERT model for the three labelling schemes. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

on the three labelling schemes: IOB2, IOB2-double, and bigappy-unicrossy-double. Experiments 2A and 2B investigate how reducing the tagset affects the results of these models. Experiment 3 investigates how reducing complexity in the data affects the performance of these models on the task of vMWE identification.

The MWE-based Precision, Recall and $F1$ scores of the "optimised" mBERT model from Series 1 when trained on the three different datasets is shown in Figure 8.8. According to the results, *IOB2* tagging yields lower results than *IOB2-d* tagging or *bigappy-unicrossy-d* tagging, possibly as these tagging systems may more accurately capture instances of 'double-tagged' MWEs, or MWEs sharing a common token. However, the results shown in Figure 8.9 for the same experiment on the "optimised" gaBERT model show that the *IOB2* tagging gives a better performing model, particularly with regards to recall. One potential issue with the *bigappy-unicrossy-d* tagging scheme is the introduction of more labels leading to a larger tagset to select from.

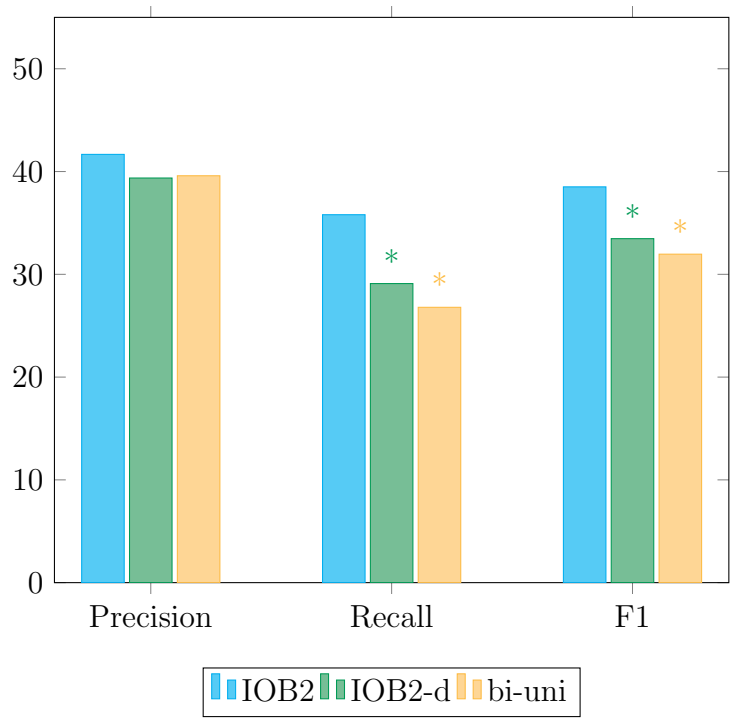


Figure 8.9: Exp 1: Precision, Recall and F1 score for gaBERT model for the three labelling schemes. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

For Experiments 2A, we explored simplifying the tagsets, to note how this affected model performance. We can see from the results captured in Tables 8.7 and 8.8 that both the 'VPC.semi' label and the 'LVC.full' label were more difficult to capture than their counterparts 'VPC.full' and 'LVC.cause', which is perhaps unsurprising, as the latter more often tend to present as productive MWEs. Merging these tags reduces the tagset to 6 for *IOB2* and *IOB2-d* tagged dataset, and 12 for *bigappy-unicrossy-d* tagging. The results are captured for the mBERT model in Figure 8.10, and the gaBERT model in Figure 8.11. On these plots we have included a baseline indicator for the MWE-based $F1$ score of the "optimised" models reported in Series 1.

Looking at Figure 8.11, we see that when using the *IOB2-d* tagging for this merged-tag dataset, the model failed to predict any MWE tags, indicating that, once again, these models are highly susceptible to instability and sensitive to changes in the data. From the plot, we see this merge-tagging system decreased model

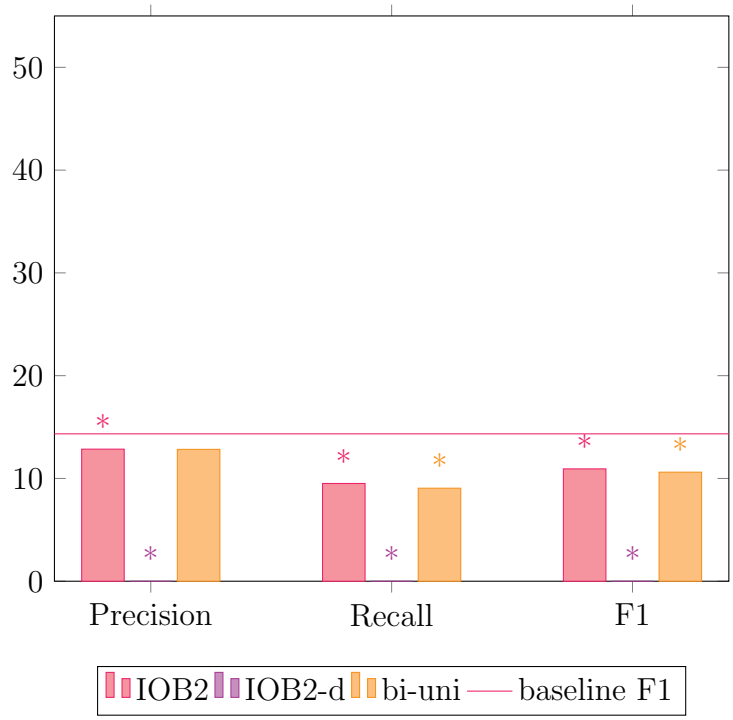


Figure 8.10: Exp 2A: Precision, Recall and F1 score for mBERT model trained on merged 'LVC' and 'VPC' tags. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

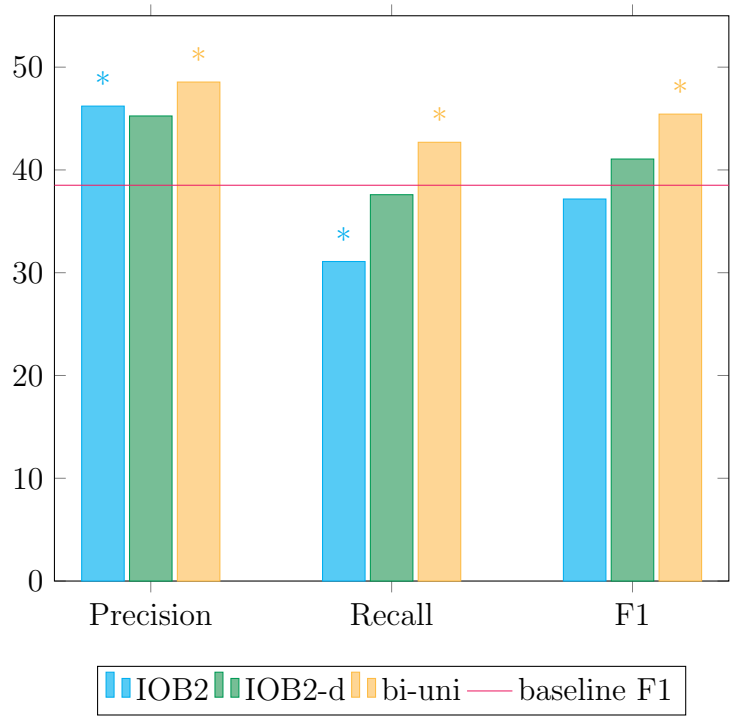


Figure 8.11: Exp 2A: Precision, Recall and F1 score for gaBERT model trained on merged 'LVC' and 'VPC' tags. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

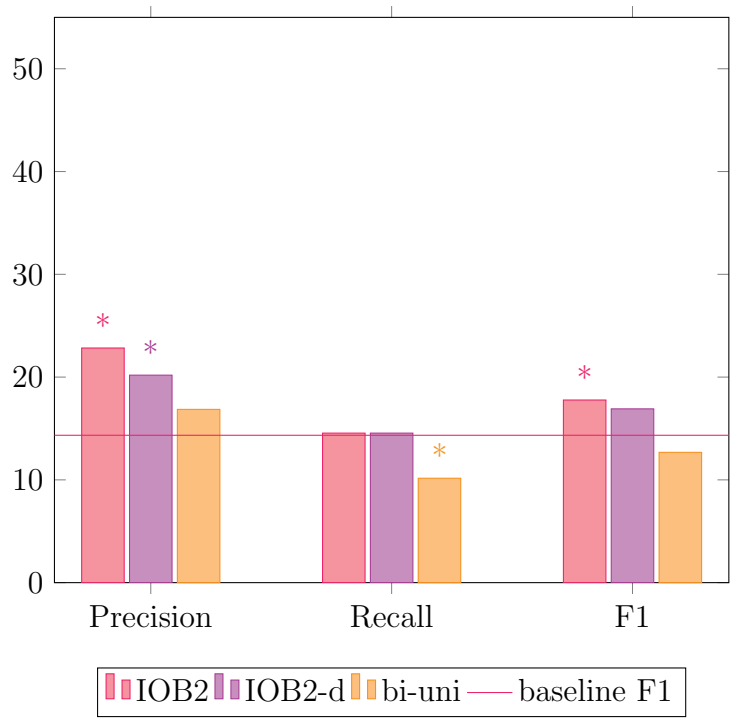


Figure 8.12: Exp 2B: Precision, Recall and F1 score for mBERT model with single MWE tag. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

performance for the mBERT model. There was some improvement on the baseline for the gaBERT model, but only when using the *IOB2-d* and *bigappy-unicrossy-d* tagging. It is difficult to draw any conclusions from these results about the effect of either the tagging scheme used, or the simplification of the tagset.

Experiment 2B further reduced the tagset to a single 'MWE' label. While this greatly reduces the difficulty of assigning the correct label to the task, given that the correct label does not affect the evaluation metrics, this may not be an issue. On the other hand, reducing the tagset in this way decreases productivity of certain labels, such as the 'VPC.semi' or 'LVC.cause' labels. Figures 8.12 and 8.13 display the results of this experiment.

Figure 8.12 shows a slight improvement on the baseline $F1$ score when using a single MWE tag versus the many categories, however, any advantages offered by using *IOB2-d* or *bigappy-unicrossy-d* tagging are not evident here, as the *IOB2* tagging proves more useful in predicting MWEs. Figure 8.13 likewise shows the decrease in ability for the models to predict MWEs when using single MWE tagging

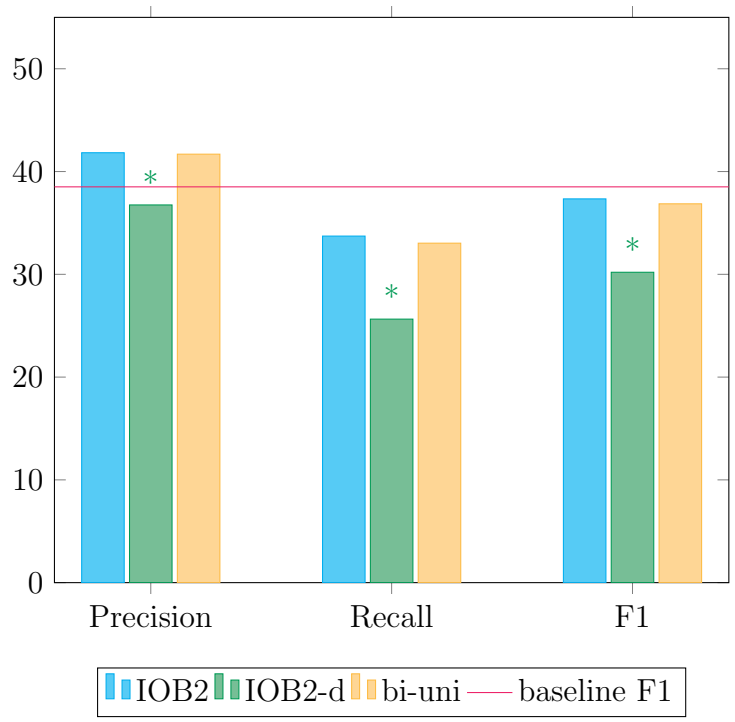


Figure 8.13: Exp 2B: Precision, Recall and F1 score for gaBERT model with single MWE tag. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

in the data. It appears that while reducing the tagset may help with the complexity of applying the correct category, the level of syntactic and lexical variability and diversity across all of the categories may be too complex of a pattern to learn with only one label.

Experiment 3 removes the more complex constructions from the dataset, to observe how this decrease in task complexity is reflected in the results, with the expectation that this will result in an increase in model performance. We present the results of this experiment in Figures 8.14 and 8.15.

Surprisingly, we see from Figure 8.14 that there was no such increase in performance, while Figure 8.15 shows only a slight increase in performance when using *bigappy-unicrossy-d* tagging.

The results from these series of experiments show little pattern in tagging schemes used or simplification of either the tagset or the data. This indicates that the hyperparameters selected in Series 1 are in fact highly sensitive to the dataset used, and variations in the data, even slight, can result in significant changes in model

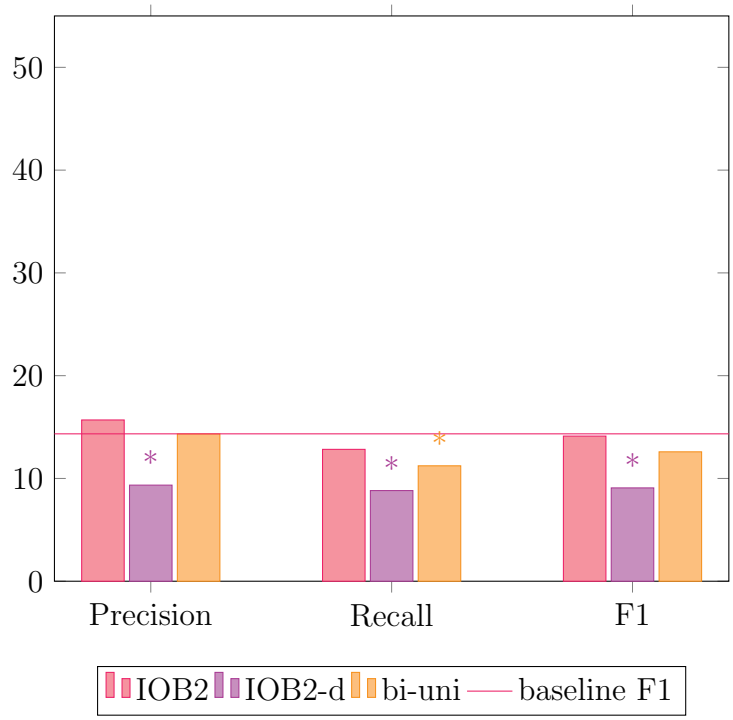


Figure 8.14: Exp 3: Precision, Recall and F1 score for mBERT model with more complex MWEs removed. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

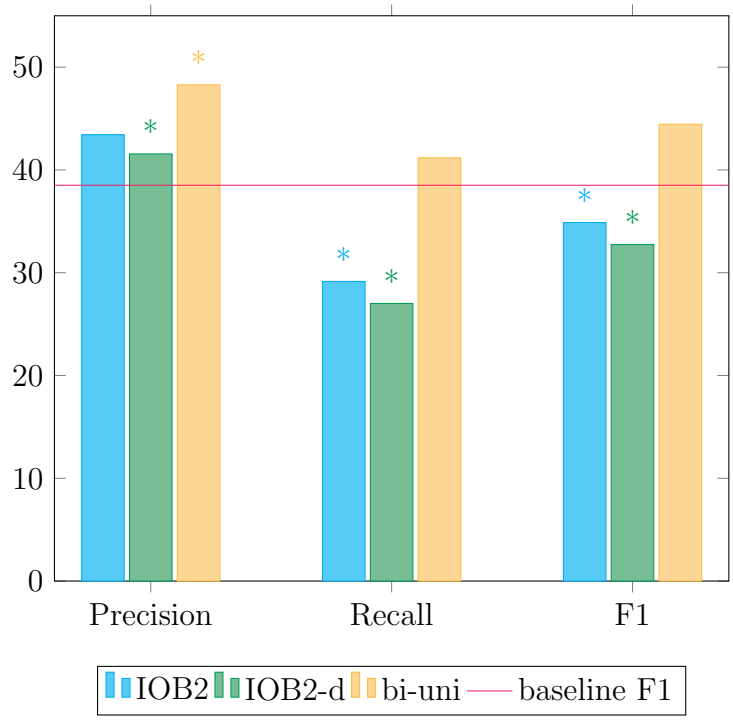


Figure 8.15: Exp 3: Precision, Recall and F1 score for gaBERT model with more complex MWEs removed. Results annotated with * were found to be statistically significant with respect to the baseline results found in Series 1, with a threshold p-value of 0.05.

performance (see the failure of the mBERT model in Experiment 2A to predict any MWE labels).

8.2.4.4 Manual Exploration of Data in Series 1

A manual inspection of the data from the experiments in Series 1 revealed many single-token MWEs predicted. Given single-token MWEs are not considered valid MWEs in the Irish data, we removed these predicted labels in a post-processing step, and evaluated the predicted output at this stage.[9] Although this step negatively affected the recall and thus the Token-based $F1$ scores suffered, overall this was shown to have a positive impact on the systems' precision scores, and thus the MWE-based $F1$ scores increased for both the mBERT and gaBERT models (see Table 8.9 and 8.10).

The MWE-based $F1$ score for unseen MWEs for the mBERT model rose from 7.33 to 11.46, an improvement of 4.13 points, and the score for the gaBERT model rose from 32.23 to 40.33, an improvement of 8.1 points.

8.2.4.5 Manual Exploration of Data in Series 2

We performed a manual inspection of the predicted output of all models generated in Series 2 of experiments, and noted some patterns that occurred in the predicted labels across all systems. For instance, almost all of the predicted 'VPC.full' (or 'VPC') labels were applied to the combination *bain + amach* 'extract out' or any of its variations (*baint amach*, *bainte amach*), with some systems also recognising '*bris amach*' 'break out'. This trend can be traced to the training data, where only two examples of 'VPC.full' are labelled, both of which are *bain+amach* constructions, while the development data has another eight such labels, three of which are this construction also. 'IAV' tended to be correctly applied to verbs combined with

[9]While single-token vMWEs did occur in our data as a result of converting from doubly-annotated tokens (see Section 8.2.3.2), these are relatively rare occurrences, and will only ever occur in combination with a multi-token vMWE. In contrast, the predicted single-token vMWEs would often occur with no other vMWE in context.

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.43	0.13	**0.20**	433	129
Token-based	0.66	0.19	**0.30**	1005	295
IAV (MWE-based)	0.46	0.16	0.23	116	39
IAV (Token-based)	0.55	0.19	0.29	237	83
IRV (MWE-based)	0.00	0.00	0.00	6	0
IRV (Token-based)	0.00	0.00	0.00	14	0
LVC.cause (MWE-based)	0.34	0.14	0.20	73	29
LVC.cause (Token-based)	0.61	0.23	0.33	192	72
LVC.full (MWE-based)	0.24	0.07	0.11	136	41
LVC.full (Token-based)	0.46	0.15	0.23	286	94
VID (MWE-based)	0.42	0.07	0.12	69	12
VID (Token-based)	0.63	0.09	0.16	210	30
VPC.full (MWE-based)	0.75	0.30	0.43	20	8
VPC.full (Token-based)	0.75	0.30	0.43	40	16
VPC.semi (MWE-based)	0.00	0.00	0.00	13	0
VPC.semi (Token-based)	0.00	0.00	0.00	26	0
Unseen MWE-based	0.26	0.07	**0.11**	299	85

Table 8.9: Precision, recall and $F1$ scores for the optimised **mBERT** model, after removing single-token predictions.

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.63	0.36	**0.46**	433	246
Token-based	0.74	0.43	**0.54**	1005	580
IAV (MWE-based)	0.71	0.19	0.30	116	31
IAV (Token-based)	0.73	0.19	0.31	237	63
IRV (MWE-based)	0.00	0.00	0.00	6	0
IRV (Token-based)	0.00	0.00	0.00	14	0
LVC.cause (MWE-based)	0.57	0.63	0.60	73	81
LVC.cause (Token-based)	0.65	0.74	0.69	192	219
LVC.full (MWE-based)	0.50	0.38	0.44	136	103
LVC.full (Token-based)	0.59	0.45	0.51	286	218
VID (MWE-based)	0.20	0.07	0.11	69	25
VID (Token-based)	0.41	0.13	0.20	210	68
VPC.full (MWE-based)	0.83	0.25	0.38	20	6
VPC.full (Token-based)	0.83	0.25	0.38	40	12
VPC.semi (MWE-based)	0.00	0.00	0.00	13	0
VPC.semi (Token-based)	0.00	0.00	0.00	26	0
Unseen MWE-based	0.53	0.32	**0.40**	299	182

Table 8.10: Precision, recall and $F1$ scores for the optimised **gaBERT** model, after removing single-token predictions.

the *le* 'with' preposition such as *glac le* 'take with', *tarraing le* 'draw along', *bain le* 'go with/relate' or *éirigh le* 'succeed'. There was a strong pattern of applying the 'LVC.cause' (or 'LVC') label to constructions consisting of *cuir* (put) + *i/ar* (in/on) + abstract noun, such as with *cuir in iúl* 'put in knowledge/inform' but also incorrectly *cuir i gcuimhne* 'put in recollection/remind', and *cuir ar fáil* but also incorrectly *cuir ar bun* 'set up'. Verbs *tabhair* 'give' and *déan* 'make' were frequently associated with the 'LVC.full' (or 'LVC') label. 'VID' labels showed perhaps the most variation across the systems, with some combinations consistently recognised across systems (e.g. *dar le* 'according to'), and some combinations occurring frequently across systems (e.g. various lexical items from the VID *an lá atá inniu ann* 'nowadays').

Between the two systems, it appears single-token predictions occur more regularly in the mBERT models than the gaBERT models, regardless of tagging scheme or dataset used (e.g. *beir* 'give birth to' annotated as 'IAV', *lámh* 'hand' annotated as 'LVC'). The rate of single-token to multi-token MWE predictions is almost double for the mBERT models versus the gaBERT models, across all labelling schemes. Additionally, generating a bag-of-words of the predicted tokens of both models shows gaBERT-based models predict labels attached to a wider variety of tokens than mBERT-based models, particularly for 'LVC' type vMWEs.Taking a concatenation of all the predicted outputs for both mBERT and gaBERT models in series 2 of our experiments, Table 8.11 shows the top twenty most frequent tokens labelled with any of the MWE labels, while Table 8.12 shows the top twenty most frequent single-token predictions labelled with any of the MWE labels. The two tables clearly demonstrate a large overlap in the predicted tokens.

Looking at the experiments individually, it was not immediately obvious whether a discernible pattern was present in the tagging schemes used or the method taken to optimise the data, or what conclusions could be drawn from these experiments. However some observations are listed below.

mBERT	Avr. Freq	gaBERT	Avr. Freq
cuir	37.9	*cuir*	43.7
le	32.5	*le*	42.9
cur	29.9	*cur*	38.1
déan	24.5	*i*	31.4
i	19.9	*déan*	24.5
déanamh	16.0	*ar*	23.5
tabhair	15.5	*tabhair*	22.6
baint	13.0	*déanamh*	18.8
bain	12.3	*bain*	14.3
amach	12.2	*baint*	12.3
éirigh	11.0	*éirigh*	11.7
ar	10.1	*tabhairt*	11.2
tabhairt	7.9	*faoi*	8.3
deara	5.8	*amach*	8.1
chun	5.4	*fáil*	7.8
caith	5.2	*chun*	6.8
faoi	4.6	*tar*	5.7
iarracht	4.6	*deara*	5.7
síos	4.5	*iúl*	4.5
ceann	4.2	*téigh*	4.5

Table 8.11: Table showing 20 most frequently labelled tokens for mBERT and gaBERT models, including single-token predictions. Avg. Freq is calculated by concatenating number of tokens tagged across the 15 systems for each model, and averaging the total.

Experiment 1: Most of the model predictions were similar to those mentioned above, however, the *bigappy-unicrossy-d* tagged dataset showed a significant decrease in prediction ability of 'VID' labels in the gaBERT model. Of the 26 predicted MWEs, 14 of them were single-token predictions, and only 4 of the predicted MWEs contained a verb, much less than the predictions of the other tagging schemes for either model. Also of note is that the gaBERT model did not predict any 'VPC.semi' type MWEs, while the mBERT model generated only single-token predictions across the three tagging schemes.

Experiment 2A: Aside from the failure of the mBERT model using *IOB2-d* tagged data to predict any labels, we can see the general decrease in ability to predict labels for the mBERT model when the tags were merged. Interestingly, we

mBERT	Avr. Freq	gaBERT	Avr. Freq
cuir	21.3	*le*	9.3
déan	17.1	*déan*	7.3
cur	14.3	*tabhair*	5.2
tabhair	9.1	*déanamh*	5.1
déanamh	8.4	*cuir*	4.7
baint	5.9	*cur*	4.5
le	5.5	*ar*	3.7
amach	5.4	*éirigh*	3.7
tabhairt	4.3	*caith*	2.8
bain	4.1	*tabhairt*	2.7
caith	3.5	*tar*	2.3
éirigh	3.5	*bain*	2.1
déanta	3.4	*i*	2.1
iarracht	3.1	*baint*	1.3
i	2.6	*déanta*	1.3
ar	2.5	*bí*	1.2
síos	2.5	*glac*	1.1
glac	2.3	*tit*	1.1
chuig	1.7	*chun*	1.1
bí	1.5	*dein*	0.9

Table 8.12: Table showing 20 most frequently labelled tokens for mBERT and gaBERT models, examining only single-token predictions. Avg. Freq is calculated by concatenating number of tokens tagged across the 15 systems for each model, and averaging the total.

see particularly a decrease in the performance of predicting 'LVC' type MWEs, with many occurrences of single-token predictions being applied to the verbs '*déan*', '*cuir*' and '*tabhair*'. The gaBERT model, however, shows a slight increase in performance when predicting 'LVC' type MWEs, with fewer single-token predictions generated. Also of note is a notable improvement in the prediction of 'VID' type MWEs for the gaBERT model trained on *bigappy-unicrossy-d* tagged data. The application of 'VPC' labels remains quite conservative across the gaBERT models; out of the 18 tags predicted by the three systems cumulatively, only one was not some variant of '*bain amach*'.

Experiment 2B: The MWEs predicted by these models are all of the same type, so patterns are difficult to observe across MWE categories. It does appear that

certain verbs are regularly given an MWE label, with the verbs '*bain*', '*cuir*', '*déan*', '*éirigh*', '*tabhair*', and to a lesser extent, '*glac*', and '*tar*' being very common in mBERT model predictions across all three tagging schemes. The gaBERT models seem to have slightly more variety in the verbs annotated with MWE labels, but the bulk of the predictions follow this pattern.

Experiment 3: In the predicted labels of experiment 3, there doesn't appear to be a notable difference in the tokens annotated with MWE labels from Experiment 1, with the exception of MWEs assigned 'LVC.full', which appears to have increased single-token predictions. In the gaBERT models, there was a decrease in performance when predicting 'LVC.cause' type MWEs consistent across the three tagging schemes.

8.2.4.6 Experiment 4: Resizing of Dataset

Our final experiment explores the remaining issue of the limited size of the training and development datasets for this task. We reshuffle the dataset to redistribute the MWEs more equally, with 559 examples (and 216 MWE examples) in the training data, 546 examples (and 213 MWE examples) in the development data, and 594 examples (and 230 MWE examples) in the test data. We use the same hyperparameter settings for both the mBERT and gaBERT models as with the "optimised" models of Series 1, and compare the results across the three tagging schemes, as we did for previous experiments in Series 2. The results are shown in Figures 8.16 and 8.17, and compared to our baseline optimised models.

We can see an increase in performance from the baseline, particularly with the mBERT model. The *bigappy-unicrossy-d* tagging scheme appears to yield the best results for both models, although these results are not robust enough across the other experiments to draw conclusions on this being a superior tagging scheme. This redistribution of data does not make for a fair comparison with the other results, but as a hypothetical best case scenario, we present in Tables 8.13 and

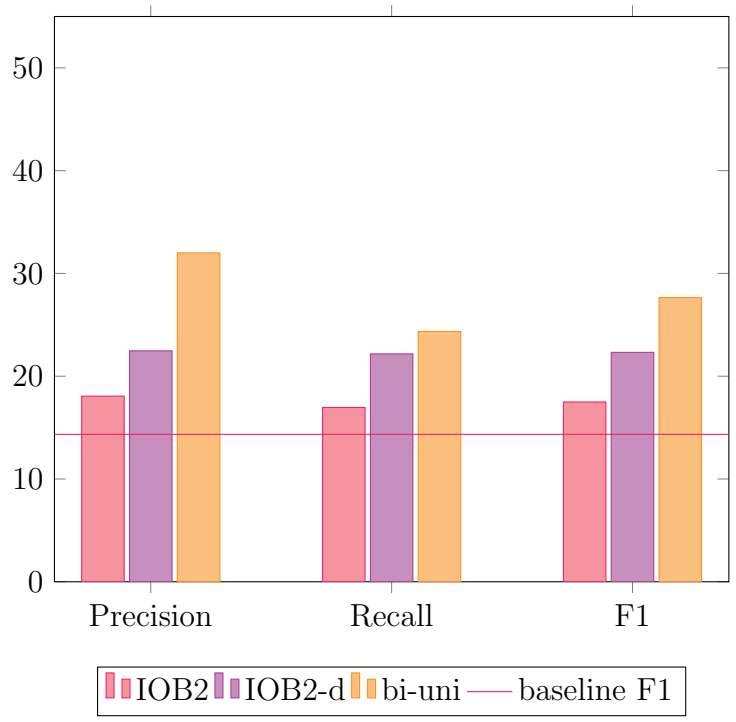


Figure 8.16: Exp 4: Precision, Recall and F1 score for mBERT model with reshuffled data.

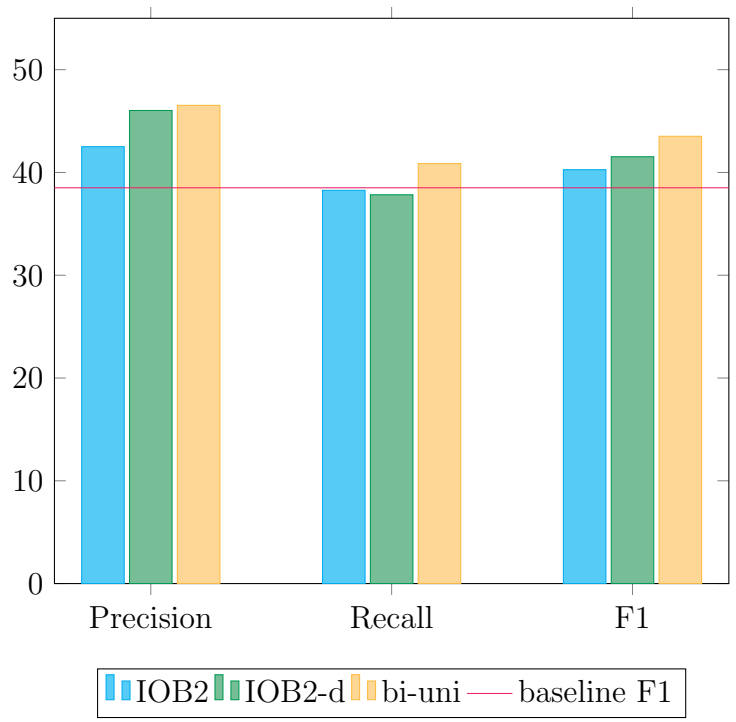


Figure 8.17: Exp 4: Precision, Recall and F1 score for gaBERT model with reshuffled data.

8.14 the results of these mBERT and gaBERT models when trained on the larger dataset using *bigappy-unicrossy-d* tagging, after applying the post-processing script to improve precision. Results are compared with the results in Table 8.9 and 8.10 from the optimised models from Series 1, with the scores coloured in green where there is an increase, and red to indicate a decrease.

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.57(+0.14)	0.24(+0.11)	**0.34(+0.14)**	230	99
Token-based	0.70(+0.04)	0.30(+0.11)	**0.42(+0.12)**	521	225
IAV (MWE-based)	0.58(+0.12)	0.26(+0.1)	0.36(+0.13)	68	31
IAV (Token-based)	0.62(+0.07)	0.29(+0.1)	0.40(+0.11)	136	65
IRV (MWE-based)	0.00	0.00	0.00	1	2
IRV (Token-based)	0.00	0.00	0.00	2	4
LVC.cause (MWE-based)	0.33(-0.01)	0.16(+0.02)	0.21(+0.01)	32	15
LVC.cause (Token-based)	0.49(-0.12)	0.23	0.31(-0.02)	80	37
LVC.full (MWE-based)	0.57(+0.33)	0.24(+0.17)	0.34(+0.23)	88	37
LVC.full (Token-based)	0.70(+0.24)	0.31(+0.16)	0.43(+0.2)	189	83
VID (MWE-based)	0.09(-0.37)	0.04(-0.03)	0.06(-0.06)	25	11
VID (Token-based)	0.20(-0.43)	0.07(-0.02)	0.11(-0.05)	82	30
VPC.full (MWE-based)	0.33(-0.42)	0.10(-0.2)	0.15(-0.28)	10	3
VPC.full (Token-based)	0.33(-0.42)	0.10(-0.2)	0.15(-0.28)	20	6
VPC.semi (MWE-based)	0.00	0.00	0.00	6	0
VPC.semi (Token-based)	0.00	0.00	0.00	12	0
Unseen MWE-based	0.37(+0.11)	0.14(+0.07)	**0.21(+0.1)**	146	57

Table 8.13: Precision, recall and $F1$ scores for the optimised mBERT model trained on reshuffled data tagged with *bigappy-unicrossy-d* after removing single-token predictions.

We see from Tables 8.13 and 8.14 that there is an overall increase in MWE-based precision, recall and $F1$ scores for these models, but the breakdown of individual categories shows not every type of MWE was similarly impacted by this increase in data. Across both models, it appears that 'LVC.full' type MWEs were the most impacted, with the MWE-based $F1$ score increasing by 23 points for the mBERT model, and 20 points for the gaBERT model. Likewise, both models saw a decrease in scores predicting 'VID' and 'VPC.full' type MWEs, with the mBERT model MWE-based $F1$ scores decreasing by 6 points and 28 points respectively, and the

Category	Precision	Recall	$F1$	# Gold	# Pred
MWE-based	0.63	0.41(+0.05)	**0.47(+0.01)**	230	149
Token-based	0.70(-0.04)	0.44(+0.01)	**0.54**	521	329
IAV (MWE-based)	0.50(-0.21)	0.32(+0.13)	0.39(+0.09)	68	44
IAV (Token-based)	0.51(-0.22)	0.35(+0.16)	0.41(+0.1)	136	92
IRV (MWE-based)	0.00	0.00	0.00	1	0
IRV (Token-based)	0.00	0.00	0.00	2	0
LVC.cause (MWE-based)	0.53(-0.04)	0.31(-0.32)	0.39(-0.21)	32	19
LVC.cause (Token-based)	0.57(-0.08)	0.33(-0.41)	0.41(-0.28)	80	46
LVC.full (MWE-based)	0.74(+0.24)	0.56(+0.18)	0.64(+0.2)	88	66
LVC.full (Token-based)	0.77(+0.18)	0.56(+0.11)	0.64(+0.13)	189	137
VID (MWE-based)	0.06(-0.14)	0.04(-0.03)	0.05(-0.06)	25	16
VID (Token-based)	0.20(-0.21)	0.11(-0.02)	0.14(-0.06)	82	46
VPC.full (MWE-based)	0.50(-0.33)	0.20(-0.05)	0.29(-0.09)	10	4
VPC.full (Token-based)	0.50(-0.33)	0.20(-0.05)	0.29(-0.09)	20	8
VPC.semi (MWE-based)	0.00	0.00	0.00	6	0
VPC.semi (Token-based)	0.00	0.00	0.00	12	0
Unseen MWE-based	0.49(-0.04)	0.32	**0.39(-0.01)**	230	149

Table 8.14: Precision, recall and $F1$ scores for the optimised gaBERT model trained on reshuffled data tagged with *bigappy-unicrossy-d* after removing single-token predictions.

gaBERT model MWE-based $F1$ scores decreasing by 6 points and 9 points respectively.

Manually inspecting the predicted labels of these models, we see the same patterns in the scores, with a noticeable improvement in 'LVC.full' type MWEs predicted in both the mBERT and gaBERT output. However, the number of 'VID' and 'VPC.full' labels predicted has fallen, and the predicted labels are applied to a wider variety of tokens. This is likely due to the greater exposure to these labels in the new datasets, which saw an increase of 8 additional 'VPC.full' labels and 26 additional 'VID' labels in the training data, and 2 additional 'VPC.full' and 18 additional 'VID' labels in the development data. The 'VPC.full' labels predicted by the mBERT and gaBERT models show a decrease in the '*bain + amach*' constructions favoured by the mBERT and gaBERT models of Experiment Series 1. The increase in exposure (22 additional labels in training, 26 additional labels in development) to the 'LVC.full' type MWE seems to have improved the systems' ability to predict this label. This is perhaps due to the more syntactically systematic nature of this

construction when compared to the 'VID' labels.

8.2.5 Relevance to Low-resource Languages

These results demonstrate the value of **monolingual language models** in such tasks. Our gaBERT-based models outperformed the mBERT-based models in almost all experiments conducted, barring some models which failed to predict any MWEs at all. This significant increase in performance is particularly reflected in the case of unseen vMWEs, which by their nature, present a great challenge to low-resource languages, as they are likely to be more prevalent where there is a scarcity of data/resources. Our experiments show how even models trained on a very small dataset and fine-tuned properly can yield results similar to other languages in the Shared Task with much larger datasets (e.g. models trained on Portuguese, which had 6,437 annotated vMWEs, almost 10 times the number annotated in the Irish dataset, gave similar $F1$ scores).

As demonstrated in Chapter 7, such monolingual language models are expensive to train, both in language resources and in hardware required, and may be a challenge for lower-resource languages to build. However, our experiments show that multilingual models such as mBERT show promising capabilities to capture even unseen vMWES, and even small additions to the data can dramatically improve these results. These experiments also highlighted the importance of careful **hyperparameter tuning**, as the manual explorations of the hyperparameter space resulted in an improvement of 4.73 (8.86 after single-tokens were removed) in the unseen MWE-based $F1$ score compared to the mBERT-based system submitted by TRAVIS-multi.

Our experiments confirm the susceptibility of Transformer-based models to **instability**, where even small variations in the data or in the hyperparameters selected (particularly the varying of the random seed variable) can result in a model that fails to predict any labels whatsoever. This problem seems to be exacerbated by a small-sized training dataset. However, our experiments indicate that the issue

can be combatted by increasing the number of epochs trained for, and by varying the learning rate. This finding of ours parallels the work of Mosbach et al. (2020) who, upon investigating the topic of instability in fine-tuning BERT, recommend using small learning rates with bias correction to avoid vanishing gradients early in training, and increasing the number of iterations considerably and training to near zero training loss. However, as we see in Section 8.2.4.1, some combinations of hyperparameters may result in unexpected model behaviour during training. As such, a random search hyperparameter tuning approach may be the most effective, as there is little guarantee that a well-performing hyperparameter setting will still perform well when combined with a different well-performing hyperparameter.

We also investigated the potential for **alternative sequence labelling schemes** that more accurately capture the vMWE labels. Our experiments on this topic are inconclusive, as there is no guarantee that the results we found are consistent when applied to a model trained on different hyperparameter settings. However, these alternative labelling schemes do allow for capturing doubly-annotated tokens, which previously would have been lost when using a traditional *IOB2* labelling scheme.

8.3 Summary and Conclusion

This chapter presents the task of MWE identification for the Irish language, which we consider low-resourced in terms of data for this task.

We describe our first attempt at building a system for the automatic identification of Irish MWEs, based on similar approaches taken by other systems. In particular, we follow the example of the TRAVIS-multi and TRAVIS-mono systems, by fine-tuning a multilingual BERT (mBERT) and monolingual BERT (gaBERT) language model on the vMWE annotated Irish treebank. We highlight two known issues in this process, namely the instability shown by BERT models, particularly when fine-tuned on a small dataset, as well as the issue of accurately labelling the data for a token sequence labelling task such as this. We compare the labelling tasks

traditionally used for this task, and investigate their shortcomings.

We explore two series of experiments: (i) in Series 1, we attempt to optimise hyperparameter settings to improve the performance of mBERT- and gaBERT-based models on this task, and (ii) in Series 2, we explore the optimisation of the data and attempt to address the potential issues with the Irish dataset outlined in the previous chapter, presenting four experiments to investigate these issues, and simultaneously compare three labelling schemes on these four experiments.

From our experiments in this chapter, we conclude that this task is indeed a challenging one for lower-resourced languages such as Irish. The small size of the dataset exacerbates issues such as model instability and increased model sensitivity to hyperparameter settings, which in turn makes it difficult to explore fully the impact of data optimisation steps such as those in experiment Series 2. While there are potential conclusions to be drawn regarding the relative difficulty in identification of certain vMWE categories, or the effect of certain tagging-schemes on a models ability to identify vMWEs, more experimentation is necessary to confirm these findings, particularly with regards to the impact of alternative labelling schemes.

This exploration marks the first thorough investigation into this task for the Irish language, and some best practises can be gleaned from these experiments. It is clear that hyperparameter settings significantly impact the model performance in such a task, particularly with a small dataset, and as such, the tuning of these hyperparameters should be conducted after dataset optimisation steps have taken place. There are also considerations to be made as to how the data will be tagged, and how to deal with the case of double-tagged MWEs. The solution we present here may be beneficial to another language seeking to attempt this task. Finally, the results of these experiments show the benefits of using a monolingual language model in this task, which may be a motivation for prioritising the building of such a model for other low-resource languages.

Chapter 9

Conclusion

> *I wish life was not so short. Languages take such a time, and so do all the things one wants to know about.*
>
> J.R.R. Tolkien, *The Lost Road and Other Writings*

The problems posed by MWEs represent some of the inherent challenges of the field of NLP. Language as a means of communication is both a system and an art; it functions through the application of structures, rules, and conventions, but also nuance, complexity, and creativity. Languages have consistently evolved throughout history to express in new and novel ways the same human emotions, experiences, and expressions that have always been a part of humanity. For machines to achieve the same level of language understanding and processing as humans, we must fully embrace the idiosyncratic, the odd, and the contradictory aspects of language. Such qualities are as intrinsic to languages as their grammar rules and conventions.

Research into such nebulous concepts as MWEs is tricky; definitions always appear to be lacking in some regard, and rules can never capture every exception. However, we explore these topics in order to refine that amorphous and ephemeral line between convention and contrariness. Irish is a language rich with metaphors, and, as with any language, has its own particular patterns of speech and structure.

Where language technology does not well capture these idiosyncrasies and aspects of the language that make it unique, it runs the risk of homogenising languages into something overly simplistic, or perhaps worse, forcing a language to conform to the conventions of more widely researched dominant languages, such as English.

This research contributes towards an understanding of how idiomatic aspects of the Irish language should be treated in the field of NLP. We hope that the resources developed within, as well as the insights gleaned from the methodologies applied can be used in a wide variety of NLP tasks and the development of more tools for Irish, and possibly for other low-resource languages.

9.1 Summary and Contributions

This thesis addresses many of the challenges inherent to the automatic processing of MWEs for NLP. Applying this research to the Irish language requires an understanding of the unique challenges presented by this language. Through exploring this topic, we have made the following contributions to the field:

1. **Creation of an annotated corpus of verbal MWEs for English:** As part of our initial MWE classification efforts, our analysis of the verbal MWEs present in English contributed towards the development of the PARSEME annotation guidelines, which were updated to reflect discussions and decisions made for certain constructions in English. We also contributed towards the creation of an English corpus annotated with verbal MWEs for inclusion in Edition 1.1 of the PARSEME shared task, which is detailed in Appendix C.

2. **Categorisation scheme for English MWEs:** Our analysis of multilingual MWE types contributed towards the development of annotation guidelines for the annotation of MWEs in English. This categorisation scheme and guidelines align with categorisation of English verbal MWEs (Walsh et al., 2018), while also closely aligning with categorisation of Irish MWEs, allowing for

the development of bilingual MWE-enhanced language resources, such as the annotated test corpora described in Chapter 5.

3. **Enhancement of IUDT MWE labels:** Our analysis of the behaviours of Irish MWEs was included in a review of the three MWE labels used in the Irish UD Treebank (McGuinness et al., 2020) and contributed towards the development of a systematic approach for the treatment of such constructions in the Irish-specific UD guidelines. The approach was applied to version 2.6 of the Irish UD Treebank, as well as subsequent versions. Simultaneously, this review helped to inform our own approach to the categorisation of three MWE types: `Nominal Compounds`, `Named Entities`, and `Fixed Expressions`.

4. **A typology and categorisation scheme for Irish MWEs:** Our typology of MWEs in Irish aligns reasonably well with both linguistic analysis of MWEs in the Irish literature, as well as widely adopted cross-lingual frameworks incorporating MWE information. By aligning our typology with these sources, we allow for cross-lingual research and comparisons of Irish MWEs with MWEs in other languages while maintaining consistency with existing discussions of Irish MWEs. The decisions made in our typology are largely motivated by empirical analysis of our data. Furthermore, the categorisation scheme we developed draws heavily from the PARSEME multilingual annotation guidelines for verbal MWEs. Our categorisation scheme extends the scope of these guidelines to annotate non-verbal MWEs, and enables future annotation of Irish MWEs across 9 categories.

5. **Ilfhocail lexicon of Irish MWEs:** This work represents the first lexicon of Irish MWEs developed for NLP purposes. It serves as a source of data on the types of MWEs present in Irish, allowing for their analysis, which contributes towards our understanding of MWE forms in Irish. As demonstrated in Chapter 6, the lexicon can also be incorporated as MWE information in other datasets, allowing for integration with downstream NLP applications.

6. **PARSEME manually annotated corpus of Irish verbal MWEs:** This open-source corpus represents a language resource with gold-standard annotations of Irish verbal MWEs, which has been used by systems participating in the PARSEME shared task on the automatic identification of verbal MWEs. The annotations are consistent with the multilingual annotation scheme, allowing for cross-lingual comparisons and research. We apply this corpus as training data in our experiments in developing an automatic system for identifying Irish verbal MWEs.

7. **Automatically-tagged parallel GA-EN corpus:** This corpus represents a large (835,867 sentences) bilingual parallel corpus automatically tagged on both sides with fixed and semi-fixed MWEs from the Ilfhocail corpus, and a combination of lexical resources for English MWEs. It serves as a source of training data for experiments in developing MT systems. Additionally, this corpus allows for the analysis of MWEs in running texts, across four domains.

8. **Annotated test sets of Irish and English MWEs:** These two corpora form two parallel datasets with gold-standard MWE annotations on the source side of both, according to the categorisation schemes developed for each of these languages. The datasets can be used in the training of MWE-aware NLP applications. This data can also be used for evaluation purposes, both the evaluation of MT systems (as shown in Chapter 6) and the evaluation of related tasks in the automatic processing of MWEs, such as automatic tagging or identification techniques (as shown in Chapter 5).

9. **Experiments in incorporating MWEs in NMT systems:** These preliminary experiments represent the first attempts at enhancing GA↔EN MT systems with MWE information. While the results from these experiments are not conclusive, the results indicate that including MWEs as linguistic features does not improve either GA→EN or EN→GA NMT systems. However, our evaluation of the output of these systems reveal patterns that could prove

promising in future experiments, notably, investigating the relationship between the proportion and prevalence of MWE types and certain domains, and investigating the impact of including MWEs as linguistic features for NMT systems trained on the legal domain exclusively.

10. **SOTA system for the automatic identification of Irish verbal MWEs:** Our system developed in Chapter 8 represents the state-of-the-art in the task of identification of MWEs for Irish. Our optimised system uses the gaBERT pre-trained language model, the first such monolingual language model trained for Irish. Our experiments in the identification of Irish verbal MWEs serves as a test case for this model, proving the capabilities of this model over a multilingual model (i.e. mBERT). Our experiments also contribute towards addressing model instability that is particularly seen in Transformer-based neural architectures. We demonstrate how increasing the number of epochs trained for, as well as tuning the learning rate, can combat this instability.

9.2 Addressing Research Questions

***RQ1**: Is it possible to classify Irish MWEs under a multilingual classification framework?*

Research Question 1 is addressed in Chapter 4, where we describe a typology of Irish MWEs that aligns with two multilingual frameworks: PARSEME and Universal Dependencies. By drawing on the treatment of MWEs in both of these frameworks, we have proposed a classification of Irish MWEs that allows for cross-lingual applications (see Chapter 6). That said, there are some features of Irish that do not easily align with these multilingual frameworks. Where these discrepancies exist, we have proposed alternative measures for the annotation of Irish MWEs, allowing empirical evidence, and linguistically-motivated discussions of these constructions, to inform our decisions.

***RQ2**: How can existing Irish resources be leveraged to generate Irish MWE-specific resources?*

Research Question 2 is addressed through our description of three MWE-specific resources for Irish in Chapter 5, those resources being: (i) Ilfhocail, a lexicon of Irish MWEs, and including English translations and POS information; (ii) the PARSEME Annotated Corpus of Irish vMWEs, a treebank resource enhanced with manual annotations of Irish verbal MWEs; and (iii) the MWE-Annotated Parallel Corpus, an EN-GA parallel corpus automatically tagged with MWEs.[1] By leveraging existing resources for Irish, such as dictionaries, WordNets, terminology databases, corpora, both monolingual and bilingual, and treebanks, we demonstrate how MWE-specific resources can be constructed, using both manual and automatic approaches, and how such resources can then be applied in downstream NLP tasks.

***RQ3**: What challenges exist for developing a system of automatic identification of Irish MWEs?*

Research Question 3 is addressed through our exploration of the task of MWE identification in Chapters 7 and 8. In Chapter 7, we examine the task at large, and some of the challenges inherent to this task. Our exploration of the inclusion of Irish in Edition 1.2 of the PARSEME Shared Task highlights what aspects may be contributing towards the difficulty of this task for Irish in particular, comparing the results achieved on the Irish dataset with other participating languages. We identify several potential challenges, such as the small size of the Irish dataset, and the high rate of unseen vMWEs in the data (which corresponds to the high rate of variability for Irish MWEs and the balanced nature of the Irish data). In Chapter 8, we detail our approach in designing experiments for the optimisation of a model for this task. We highlight in particular the issues of model instability and labelling of

[1]This last resource also includes a pair of manually annotated test sets, with gold-standard annotations of Irish and English MWEs.

MWEs as two challenges to overcome, given our approach. Our results demonstrate the success of certain aspects of our approach, such as employing a monolingual pre-trained language model for this task, and careful selection of hyperparameters to tune. We also find some inconclusive results following from our experiments in optimisation of the data, including employing alternative labelling schemes for capturing MWEs in our data.

***RQ4**: Is knowledge of MWEs useful in an Irish-English and English-Irish MT system? (Case study)*

> ***RQ4a**: How do we represent knowledge of MWEs in an MT system?*
>
> ***RQ4b**: How can we evaluate the effect of adding MWE information to an Irish↔English MT system?*

Research Question 4 is addressed through our experiments in incorporating MWE information in Neural MT systems (see Chapter 6). While the results of these experiments are inconclusive due to our small test datasets and sparsity issues, we have laid groundwork for future experiments in this topic. We describe a method for incorporating MWE information in our datasets in Chapter 5, using automatic tagging to annotate the text with an MWE label, similar to approaches of including linguistic features to improve MT systems, addressing Research Question 4a. Research Question 4b is also addressed in Chapter 6, as we employ several metrics for our evaluation of MWE-aware MT systems: BLEU, CHRF, and `Score_mwe`. In addition, we perform a manual inspection of the data, employing tools to analyse different decisions made by the different models.

9.3 Future Work

9.3.1 Cleaning and Publication of Ilfhocail

The Ilfhocail lexicon represents a first step towards an open source lexical resource of Irish MWEs for NLP purposes. Our manual inspection and annotation of a sample of this lexicon revealed some quality issues that we plan to address for future versions of this lexicon. Issues to address include removing non-MWE entries, generating canonical forms for entries (and removing non-lexicalised elements), extracting informative headwords for each entry, and generating POS information and English translations where such information is missing.

9.3.2 Discovery of New MWE Types

In Appendix B, we describe some exploratory experiments in the task of automatic discovery of MWEs. The methods attempted include filtering a large monolingual corpus using automatic metrics to extract statistically idiosyncratic N-grams, exploiting misalignments between parallel corpora to discover MWE candidates, and employing discovery tools, such as the MWE-toolkit (Ramisch et al., 2010). We would like to continue to explore these methods, with the benefit of a better understanding of how MWEs appear in Irish. Additionally, we would like to extend the Ilfhocail lexicon, through application of these discovery methods.

9.3.3 Extending the PARSEME Annotated Corpus for Irish

The PARSEME annotated corpus for Irish, described in Chapter 5, forms a valuable resource of gold-standard MWE annotations in text. This resource allowed for the development of systems for the identification of Irish vMWEs in the PARSEME shared task, and the development of our own gaBERT-based model for the identification of Irish vMWEs. By extending this corpus, we hope to improve the SOTA for systems tackling this task.

9.3.4 Enhancement of Typology and Annotation Guidelines for Irish and English MWEs

Our typology described in Chapter 4 and the annotation guidelines provided in Appendix A represent the first of their kind for the purposes of annotating MWEs in Irish in running text. As we continue to analyse the kinds of MWE information that is useful to capture for NLP applications, this typology (and subsequently, the annotation scheme) is likely to evolve. We aim to address the difficult cases listed in Chapter 4, and find a suitable treatment of these constructions, that aligns with both with language-specific features of Irish, as well as multilingual frameworks. We also aim to concurrently enhance the annotation guidelines for English, particularly for those categories of MWEs that are not yet considered by the PARSEME annotation guidelines. We aim to align this treatment of English MWEs with our treatment of Irish MWEs, to allow for cross-lingual analysis and the creation of parallel and bilingual resources.

9.3.5 Incorporating Discontiguous MWEs in NMT

In our experiments in incorporating MWEs in NMT, we limited our experiments to syntactically fixed and semi-fixed MWEs only. These MWEs were necessarily contiguous constructions, and contained proportionally few verbal MWEs. This extension was similarly posed in Zaninello and Birch (2020). Furthermore, Barreiro et al. (2013) noted many translation errors arose from non-contiguous constructions, particularly LVCs and IAVs.

By applying a hybrid tagging approach using both the lexicon to look up contiguous MWEs, as well as our system for the automatic annotation of verbal MWEs, it will be possible to capture both of these types of MWEs in the data. Further experiments in incorporating such MWEs in NMT models may reveal more conclusively whether annotating such information improves the performance of GA↔EN NMT. Such results may also require human evaluation studies to fully investigate

the impact of including MWE information in our training data.

9.3.6 Improved Models for the Identification of Irish Verbal MWEs

A SOTA system for the identification of Irish verbal MWEs, described in Chapter 8, was created by following the example of successful models in Edition 1.2 of the PARSEME shared task. By incorporating other methodologies explored in this task, we aim to improve this model's performance even further. For instance, the best performing model overall (Taslimipoor et al., 2020) employed a multi-task learning approach, by jointly learning MWEs and parsing for this task. The results of this methodology appear promising, and may result in a better performing system when combined with our gaBERT language model.

Bibliography

Acosta, O., Villavicencio, A., and Moreira, V. (2011). Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA. Association for Computational Linguistics.

Ahrenberg, L. (2007). LinES: An English-Swedish parallel treebank. In *Proceedings of NODALIDA*, pages 270–273, Tartu, Estonia.

Al-Haj, H., Itai, A., and Wintner, S. (2013). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27:130–170.

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and treatment of multiword expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55, Barcelona, Spain. Association for Computational Linguistics.

Anastasiou, D. (2010). *Idiom treatment experiments in machine translation*. PhD thesis, Universität des Saarlandes.

Antunes, S. and Mendes, A. (2013). MWE in Portuguese: Proposal for a typology for annotation in running text. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 87–92, Atlanta, Georgia, USA. Association for Computational Linguistics.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. MIT Press, Cambridge, Massachusetts and London, England.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Aulamo, M., Sulubacak, U., Virpioja, S., and Tiedemann, J. (2020). OpusTools and parallel corpus diagnostics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Baldwin, T. (2004). Multiword expressions. In *Advanced course at the Australasian Language Technology Summer School (ALTSS 2004)*, Sydney, Australia.

Baldwin, T. (2008). A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pages 1–2, Marrakech, Morocco.

Baldwin, T. and Kim, S. N. (2010). Multiword Expressions. *Handbook of Natural Language Processing, Second Edition*, pages 267–285.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Mitrofan, M., Dimitrova, T., and Todorova, M. (2019). Hear about verbal multiword expressions in the Bulgarian and the Romanian Wordnets straight from the horse's mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 2–12, Florence, Italy. Association for Computational Linguistics.

Barreiro, A., Monti, J., Orliac, B., and Batista, F. (2013). When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice, France.

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Ó Meachair, M. J., and Foster, J. (2022). gaBERT – an Irish Language Model. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France.

Bayda, V. (2015). Irish constructions with bain. *Yn llawen iawn, yn llawn iaith: Proceedings of the 6th International Colloquium of Societas Celto-Slavica. Vol. 7 of Studia Celto-Slavica. Johnston, D., Parina, E. and Fomin, M. (eds)*, 7:213–228.

Bayda, V. (2016). Progress on Irish idiomatic constructions. The Eighth International Colloquium of Societas Celto-Slavica.

Bejček, E. and Straňák, P. (2010). Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44:7–21.

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. (2000). SIMPLE: A general framework for the development of multilingual lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82:233 – 278.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Berk, G., Erden, B., and Güngör, T. (2019). Representing overlaps in sequence labeling tasks with a novel tagging scheme: bigappy-unicrossy. Computational Linguistics and Intelligent Text Processing, Springer International Publishing.

Berzins, A., Choukri, K., Giagkou, M., Lösch, A., Mazo, H., Piperidis, S., Rigault, M., Schnur, E., Small, L., van Genabith, J., Vasiljevs, A., Adamson, A., Anastasiou, D., Avraamides-Haratsi, N., Bel, N., Bódi, Z., Branco, A., Budin, G., Dadurkevicius, V., de Smeytere, S., Dobreva, H., Domeij, R., Dunne, J., Eide, K., Foti, C., Gavriilidou, M., Grouas, T., Gruzitis, N., Hajic, J., Heinisch, B., Hoste, V., Jönsson, A., Kakoyianni-Doa, F., Kirchmeier, S., Koeva, S., Konturová, L., Kotzian, J., Krek, S., Kristmannsson, G., Kuhmonen, K., Lindén, K., Lynn, T., Magone, A., Mazo, H., Melero, M., Mihailescu, L., Montemagni, S., Ó Conaire, M., Odijk, J., Ogrodniczuk, M., Pecina, P., Arild Olsen, J., Sandford Pedersen, B., Perez, D., Repar, A., Rigouts Terryn, A., Rögnvaldsson, E., Rosner, M., Routzouni, N., Soria, C., Soska, A., Spiteri, D., Tadic, M., Tiberius, C., Tufis, D., Utka, A., Vale, P., van den Berg, P., Váradi, T., Vare, K., Witt, A., Yvon, F., Ziedins,

J., and Zumrik, M. (2019). *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe-Why Language Data Matters: ELRC White Paper*. European Language Resource Coordination Consortium, Germany.

Bhattasali, S., Fabre, M., and Hale, J. (2018). Processing MWEs: Neurocognitive bases of verbal MWEs and lexical cohesiveness within MWEs. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 6–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bhreathnach, Úna, C. F. and Nic Pháidín, C. (2013). *Téarmaíocht don Aontas Eorpach. Taithí na hÉireann: Tionscadal GA IATE/ Terminology for the European Union. The Irish Experience: The GA IATE Project*. Galway: Cló Iar-Chonnacht.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Hirst, G. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English Web Treebank. Linguistic Data Consortium. `https://catalog.ldc.upenn.edu/LDC2012T13`.

Bird, S. and Simons, G. (2002). Seven dimensions of portability for language documentation and description. *CoRR*, cs.CL/0204020.

Bird, S. and Simons, G. (2003). Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities*, 37:375–388.

Bloch-Trojnar, M. (2009). On the Nominal Status of VNs in Light Verb Constructions in Modern Irish. In *PASE Papers 2008. Vol. 1: Studies in Language and Methodology of Teaching Foreign Languages*, page 25–33, Wrocław: Oficyna Wydawnicza ATUT.

Bloch-Trojnar, M. (2010). Semantic constraints on light verb constructions in modern Irish. In *Proceedings of the Fourth International Colloquium of Societas Celto-Slavica*, Poland.

Bobrow, S. and Bell, S. (1973). On catching on to idiomatic expressions. *Memory Cognition*, (1):343–346.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M.,

Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bolinger, D. (1971). *The Phrasal Verb in English*. Cambridge: Harvard University Press.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012a). Automatic construction of a MultiWord expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 95–108, Mumbai, India. The COLING 2012 Organizing Committee.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012b). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 674–679, Istanbul, Turkey. European Language Resources Association (ELRA).

Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 109–114, Uppsala, Sweden. Association for Computational Linguistics.

Bräsicke, L. (2003). Gramadach na Gaeilge. `http://nualeargais.ie/gnag/gram.htm`. Accessed: 2022/01/15.

Buljan, M. and Šnajder, J. (2017). Combining linguistic features for the detection of Croatian multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 194–199, Valencia, Spain. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE map. harmonising community descriptions of resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089, Istanbul, Turkey. European Language Resources Association (ELRA).

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, R., Macleod, C., and Zampolli, A. (2002a). Towards best practice for multiword expressions in computational lexicons. *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 1934–1940.

Calzolari, N., Zampolli, A., and Lenci, A. (2002b). Towards a standard for a multilingual lexical entry: The EAGLES/ISLE initiative. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 264–279, Berlin, Heidelberg. Springer-Verlag.

Candito, M. and Constant, M. (2014). Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Baltimore, Maryland. Association for Computational Linguistics.

Carpuat, M. and Diab, M. (2010). Task-based evaluation of Multiword Expressions: A pilot study in Statistical Machine Translation. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June):242–245.

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, Dublin, Ireland. Association for Computational Linguistics.

Chang, B., Danielsson, P., and Teubert, W. (2002). Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of COLING-02: The First SIGHAN Workshop on Chinese Language Processing*, pages 1–5.

Cholakov, K. and Kordoni, V. (2014). Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 196–201, Doha, Qatar. Association for Computational Linguistics.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, 50 edition.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Fluhr, C. and Walker, D. E., editors, *Computer-Assisted Information Retrieval (Recherche d'Information et ses*

Applications) - RIAO 1988, 2nd International Conference, Massachusetts Institute of Technology, Cambridge, MA, USA, March 21-25, 1988. Proceedings, pages 609–623. CID.

Christian Brothers, T. (1960). *Graiméar Gaeilge na mBráithre Críostaí*. M.H. Mac an Ghoill agus a Mhac, Tta.

Christian Brothers, T. (1999). *New Irish Grammar*. Dublin: Fallon.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 76–83.

Ciaramita, M. and Johnson, M. (2003). Supersense tagging of unknown nouns in Wordnet. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 168–175, Sapporo, Japan. Association for Computational Linguistics.

Colson, J.-P. (2020). HMSid and HMSid2 at PARSEME shared task 2020: Computational corpus linguistics and unseen-in-training MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 119–123, online. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087. `http://arxiv.org/abs/1710.04087`.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Constant, M. and Nivre, J. (2016). A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 161–171, Berlin, Germany. Association for Computational Linguistics.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Copestake, A. and Flickinger, D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., and Flickinger, D. (2002). Multiword expressions: linguistic precision and reusability. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Cowie, A. P. (1981). The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics*, II(3):223–235.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

CSO (2016). Census of Population 2016 – profile 10: Education, Skills and the Irish language. Central Statistics Office, Ireland. `https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/`.

de Bhaldraithe, T. (1959). *English-Irish Dictionary*. An Gúm, 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'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

de Marneffe, M.-C., Padó, S., and Manning, C. D. (2009). Multi-word expressions in textual inference: Much ado about nothing? In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer)*, pages 1–9, Suntec, Singapore. Association for Computational Linguistics.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). Bertje: A Dutch BERT model. arXiv:1912.09582.

Declerck, T., Siegel, M., and Racioppa, S. (2019). Using OntoLex-Lemon for representing and interlinking German multiword expressions in OdeNet and MMORPH. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 22–29, Florence, Italy. Association for Computational Linguistics.

Defauw, A., Szoc, S., Vanallemeersch, T., Bardadym, A., Brabers, J., Everaert, F., Scholte, K., Winckel, K. V., and den Bogaert, J. V. (2019). Developing a Neural Machine Translation system for Irish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 32–38, Dublin, Ireland. European Association for Machine Translation.

Deksne, D., Skadiņš, R., and Skadiņa, I. (2008). Dictionary of multiword expressions for translation into highly inflected languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Di Sciullo, A. M. and Williams, E. (1987). On the definition of word. *Linguistic Inquiry monographs*, (14).

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. A. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.

Dowdall, J., Rinaldi, F., Ibekwe-SanJuan, F., and SanJuan, E. (2003). Complex structuring of term variants for question answering. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 1–8, Sapporo, Japan. Association for Computational Linguistics.

Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivastava, A., and Judge, J. (2015). Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish. In *Proceedings of the 4th LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*, Poznan, Poland.

Dowling, M., Lynn, T., Poncelas, A., and Way, A. (2018). SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop*

on Technologies for MT of Low Resource Languages (LoResMT 2018), pages 12–20, Boston, MA. Association for Machine Translation in the Americas.

Dowling, M., Lynn, T., and Way, A. (2019). Leveraging backtranslation to improve machine translation for Gaelic languages. In *Proceedings of the Celtic Language Technology Workshop*, pages 58–62, Dublin, Ireland. European Association for Machine Translation.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Eryiğit, G., Adali, K., Torunoğlu-Selamet, D., Sulubacak, U., and Pamay, T. (2015). Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 70–76, Denver, Colorado. Association for Computational Linguistics.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Stuttgart.

Evert, S. and Kermes, H. (2003). Experiments on candidate data for collocation extraction. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–86, Budapest, Hungary. Association for Computational Linguistics.

Farahmand, M. and Martins, R. (2014). A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 10–16, Gothenburg, Sweden. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Fellbaum, C., editor (2009). *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. Corpus and Discourse. Bloomsbury Publishing.

Fernando, C. and Flavell, R. (1981). *On Idiom: Critical Views and Perspectives*. Exeter linguistic studies. University of Exeter.

Fillmore, Charles, B. C. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, Pittsburgh.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Findlay, J. (2019). *Multiword Expressions and the Lexicon*. PhD thesis.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Forcada, M. L. (2017). Making sense of neural machine translation. In *Translation Spaces*, volume 6, page 291–309, Universitat d'Alacant, Spain.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, 6(1):22–42.

Ghoneim, M. and Diab, M. (2013). Multiword expressions in the context of statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1181–1187, Nagoya, Japan. Asian Federation of Natural Language Processing.

Gläser, R. (1986). *Phraseologie der englischen Sprache*. Tübingen: Walter de Gruyter.

Gombert, S. and Bartsch, S. (2020). MultiVitaminBooster at PARSEME shared task 2020: Combining window- and dependency-based features with multilingual contextualised word embeddings for VMWE detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 149–155, online. Association for Computational Linguistics.

Green, S., de Marneffe, M.-C., and Manning, C. D. (2013). Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, 39(1):195–227.

Greene, D. (1966). *The Irish Language*. Dublin, The Three Candles.

Gurrutxaga, A. and Alegria, I. (2013). Combining different features of idiomaticity for the automatic classification of Noun+Verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).

Hartmann, S., Szarvas, G., and Gurevych, I. (2012). Mining multiword terms from Wikipedia. *Semi-Automatic Ontology Development: Processes and Resources*, pages 226–258.

Hashimoto, C. and Kawahara, D. (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001, Honolulu, Hawaii. Association for Computational Linguistics.

Hawwari, A., Attia, M., and Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 48–56, Doha, Qatar. Association for Computational Linguistics.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):1–19.

Jackendoff, R. (1997a). *The Architecture of the Language Faculty*. Linguistic inquiry monographs. MIT Press.

Jackendoff, R. (1997b). Twistin' the night away. *Language*, 73(3):534–559.

Jackendoff, R. (2002). English particle constructions, the lexicon, and the autonomy of syntax. *Dihé, Jackendoff, McIntyre, Urban*, pages 67–94.

Jacquemin, C. (1996). A symbolic and surgical acquisition of terms through variation. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, page 425–438, Berlin, Heidelberg. Springer-Verlag.

Jaquette, D., Cieri, C., and DiPersio, D. (2020). Related works in the linguistic data consortium catalog. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC2020)*, pages 3438–3442, Marseille, France. European Language Resources Association (ELRA).

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. P., and Uí Dhonnchadha, E. (2012). *The Irish Language in the Digital Age*. Springer Publishing Company, Incorporated.

Junczys-Dowmunt, M., Dwojak, T., and Hoang, H. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Keysar, B. and Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34:89–109.

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(2):127–152.

Kim, J. D., Brown, R., and Carbonell, J. (2010). Chunk-based EBMT. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. 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.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kuiper, K., McCann, H., Quinn, H., Aitchison, T., and van der Veer, K. (2003). SAID. Linguistic Data Consortium. `https://catalog.ldc.upenn.edu/LDC2003T10`.

Kurfalı, M. (2020). TRAVIS at PARSEME shared task 2020: How good is (m)BERT at seeing the unseen? In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 136–141, online. Association for Computational Linguistics.

Kyed, L. (2020). The longest word in the Greenlandic language. `https://guidetogreenland.com/travel-blogs/lasse-kyed/the-longest-word-in-the-greenlandic-language/`. Accessed: 2022-04-17.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Lankford, S., Afli, H., and Way, A. (2021). Machine translation in the covid domain: an English-Irish case study for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 144–150, Virtual. Association for Machine Translation in the Americas.

Laporte, E. (2018). Choosing features for classifying multiword expressions. In Sailer, M. and Markantonatou, S., editors, *Multiword expressions. Insights from a multi-lingual perspective*, chapter 6, pages 143–186.

Laskova, L., Osenova, P., Simov, K., Radev, I., and Kancheva, Z. (2019). Modeling MWEs in BTB-WN. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 70–78, Florence, Italy. Association for Computational Linguistics.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Lipka, L. (1992). *An outline of English lexicology: lexical structure, word semantics, and word-formation*. Tübingen: Niemeyer.

Logar, N., Gantar, P., and Kosem, I. (2014). Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 2(1):41–61.

Lohar, P., Popovic, M., Alfi, H., and Way, A. (2019). A systematic comparison between SMT and NMT on translating user-generated content. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, La Rochelle, France.

Losnegaard, G. S., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S., and Monti, J. (2016). PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).

Lynn, T. (2016). *Irish Dependency Treebanking and Parsing*. PhD thesis, Dublin City University and Macquarie University, Sydney.

Lynn, T. (2022). Report on the Irish language. https://european-language-equality.eu/deliverables/. Technical Report D1.20, European Language Equality Project.

Lynn, T. and Foster, J. (2016). Universal Dependencies for Irish. In *Proceedings of the Second Celtic Language Technology Workshop*, pages 79–92, Paris, France.

Lynn, T., Foster, J., and Dras, M. (2017). Morphological features of the Irish Universal Dependency Treebank. In *TLT 2017 : Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, volume 1779, pages 111–122, Bloomington, U.S.

Lyse, G. I. and Andersen, G. (2012). Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text. *Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 79–110.

Manning, C. D. and Schütze, H. (1999a). *Foundations in Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, London, England.

Manning, C. D. and Schütze, H. (1999b). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Mapelli, V., Popescu, V., Liu, L., and Choukri, K. (2016). Language resource citation: the ISLRN dissemination and further developments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1610–1613, Portorož, Slovenia. European Language Resources Association (ELRA).

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Matthews, P. (1997). *The Concise Oxford Dictionary of Linguistics*. Oxford paperback reference. Oxford University Press.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking WordNet using Lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

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). Association for Computational Linguistics.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. *Idioms: structural and psychological perspective*, pages 167–232.

Mel'čuk, I. A., Clas, A., and Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot, Montréal: AUPELF-UREF.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Monti, J., Barreiro, A., Elia, A., Marano, F., and Napoli, A. (2011). Taking on new challenges in multi-word unit processing for machine translation. *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 11–19.

Monti, J. and Todirascu, A. (2015). Multiword units translation evaluation in machine translation: another pain in the neck? In *Proceedings of Multiword Units in Machine Translation and Translation Technology (MUMTTT15)*, pages 25–30, Malaga, Spain.

Moon, R. (1998). *Fixed expressions and idioms in English : a corpus-based approach*. Oxford : Clarendon Press.

Morishita, M., Suzuki, J., and Nagata, M. (2018). Improving neural machine translation by incorporating hierarchical subword features. In *Proceedings of the 27th*

International Conference on Computational Linguistics, pages 618–629, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines.

Moszczyński, R. (2007). A practical classification of multiword expressions. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 19–24, Prague, Czech Republic. Association for Computational Linguistics.

Měchura, M. B. (2017). Pota focal gluais tí. `http://www.potafocal.com/gt/`. Accessed: 2022-06-10.

Měchura, M. and Ó Raghallaigh, B. (2010). The Focal.ie national terminology database for Irish. In *Proceedings of the XIV Euralex International Congress*, pages 937–948, Leewarden.

Měchura, M. and Ó Raghallaigh, B. (2012). The logainm.ie Placenames Database of Ireland: software demonstration. In *Proceedings of Placenames Workshop: Management and dissemination of toponymic data online*, Dublin: Fiontar, DCU.

Nakayama, H. (2018). seqeval: A Python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Neubig, G., Dou, Z., Hu, J., Michel, P., Pruthi, D., Wang, X., and Wieting, J. (2019). Compare-MT: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

Nic Niallais, A. (2020). Ag dul chun spairne leo - Sracshúil ar bhriathra frásacha na Gaeilge. *Léann Teanga: An Reiviú.*

Nic Pháidín, C., Ó Cleircín, G., and Bhreathnach, (2010). Building on a terminology resource - the Irish experience. In *Proceedings of the XIV Euralex International Congress*, pages 954–965, Leewarden.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nivre, J. and Nilsson, J. (2004). Multiword units in syntactic parsing. In *Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications, LREC*, Lisbon, Portugal.

Ní Chasaide, A., Barnes, E., Ní Chiaráin, N., McGuirk, R., Morrin, O., Nic Corcráin, M., and Cummins, J. (2022). Challenges in assistive technology development for an endangered language: an Irish (Gaelic) perspective. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 80–87, Dublin, Ireland. Association for Computational Linguistics.

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 2017*, pages 2113–2117.

Ní Chiaráin, N., Nolan, O., Comtois, M., Robinson Gunning, N., Berthelsen, H., and Ni Chasaide, A. (2022). Using speech and NLP resources to build an iCALL platform for a minority language, the story of An scéalaí, the Irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 109–118, Dublin, Ireland. Association for Computational Linguistics.

Ní Loingsigh, K. (2016). *Tiomsú agus Rangú i mBunachar Sonraí ar Chnuasach Nathanna Gaeilge as Saothar Pheadair Uí Laoghaire*. PhD thesis, Dublin City University.

Ní Loingsigh, K. (2019). "Tusa an t-oide, mise an mac léighinn": comhairle teanga ón Athair Peadar Ó Laoghaire. In *Irishleabhar Mhá Nuad*, pages 46–63.

Ní Loingsigh, K. (2021). Irish-language idioms: an overview and analysis of idioms in the published work of peadar Ó laoghaire. In *Intercontinental Dialogue 4 Project: Reproducible language units in an interdisciplinary perspective*, pages 201–216. University of Bialystok Publishing House.

Ní Loingsigh, K. and Ó Raghallaigh, B. (2016). Starting from scratch – the creation of an Irish-language idiom database. In Tinatin Margalitadze, G. M., editor, *Proceedings of the 17th EURALEX International Congress*, pages 726–734, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.

Ní Loingsigh, K. and Ó Raghallaigh, B. (2018). Harping on a new string: the search for Irish-language idioms. In *Phraséologie et discours*, pages 393–401. Paris: Honoré Champion.

Ó Corráin, A. (1989). *A concordance of idiomatic expressions in the writings of Séamus Ó Grianna*. The Queen's University of Belfast, Belfast.

Ó Domhnalláin, T. and Ó Baoill, D. (1975). *Réamhfhocail le briathra na Gaeilge*. Tuarascáil taighde. Institiúid Teangeolaíochta Éireann.

Ó Dónaill, N. (1977). *Foclóir Gailge-Béarla*. An Gúm, An Roinn Oideachas.

Ó Dónaill, N. and Ua Maoileoin, P. (1991). *An Foclóir Beag*. An Gúm, Baile Átha Cliath.

Ó Huallacháin, C. and Department of Education, I. (1966). *Buntús Gaeilge: réamhthuarascáil ar thaighde teangeolaíochta a rinneadh sa Teanglann, Rinn Mhic Gormáin*. Oifig an tSoláthair.

Ó Mianáin, P. (2020). *Concise English-Irish Dictionary*. An Gúm, Ireland.

Ó Siadhail, M. (1991). *Modern Irish: Grammatical Structure and Dialectal Variation*. Cambridge University Press, United Kingdom.

Olah, C. (2015). Understanding LSTM networks. `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`. Accessed: 2022-05-04.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Parra Escartín, C., Nevado, A., and Martínez, E. (2018). Spanish multiword expressions: Looking for a taxonomy. In Sailer, M. and Markantonatou, S., editors, *Multiword expressions. Insights from a multi-lingual perspective*, chapter 10, pages 271–323.

Pasikowska-Schnass, M. (2020). European Day of Languages: Digital survival of lesser-used languages. `https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2020)652086`. Accessed: 2022-06-03.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020a). Seen2Unseen at PARSEME shared task 2020: All roads do not lead to unseen verb-noun VMWEs.

In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online. Association for Computational Linguistics.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020b). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Pecina, P. (2009). Lexical association measures: Collocation extraction. Studies in Computational and Theoretical Linguistics, pages 137–158. Ústav formální a aplikované lingvistiky.

Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. 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, Istanbul, Turkey. European Language Resources Association (ELRA).

Phi, M. (2018). Illustrated Guide to Recurrent Neural Networks. Accessed: 2022-05-04 `https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9`.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.*, 20(1):1934–1965.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI*, pages 1–10.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing series XIV. Springer.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008). An Evaluation of Methods for the Extraction of Multiword Expressions. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, (June):50–53.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International*

Conference on Language Resources and Evaluation (LREC'10), pages 662–669, Valletta, Malta. European Language Resources Association (ELRA).

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. V. (2010). Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1/2):1–5.

Rikters, M. and Bojar, O. (2017). Paying attention to multi-word expressions in neural machine translation. In *Proceedings of The 16th Machine Translation Summit*, Nagoya, Japan.

Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L., and Mitkov, R. (2019). Bridging the gap: Attending to discontinuity in identification of multiword expressions. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:2692–2698.

Rondon, A., Caseli, H., and Ramisch, C. (2015). Never-ending multiword expressions learning. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 45–53, Denver, Colorado. Association for Computational Linguistics.

Rosén, V., De Smedt, K., Losnegaard, G. S., Bejček, E., Savary, A., and Osenova, P. (2016). MWEs in treebanks: From survey to guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2323–2330, Portorož, Slovenia. European Language Resources Association (ELRA).

Rosén, V., Losnegaard, G. S., De Smedt, K., Bejcek, E., Savary, A., Przepiórkowski, A., Osenova, P., and Mititelu, V. B. (2015). A survey of multiword expressions in treebanks. *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 179.

Ruder, S. (2018). A review of the neural history of natural language processing. `https://ruder.io/a-review-of-the-recent-history-of-nlp/`.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing*, pages 1–15.

Salehi, B., Cook, P., and Baldwin, T. (2016). Determining the Multiword Expression Inventory of a Surprise Language. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, (2015):471–481.

Savary, A., Cordeiro, S., Lichte, T., Ramisch, C., Iñurrieta, U., and Giouli, V. (2019a). Literal occurrences of multiword expressions: Rare birds that cause a stir. *Prague Bulletin of Mathematical Linguistics*, 112:5–54.

Savary, A., Cordeiro, S., and Ramisch, C. (2019b). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Savary, A., Escartín, C. P., Bond, F., Mitrović, J., and Mititelu, V. B., editors (2019c). *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, Florence, Italy. Association for Computational Linguistics.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G. S., Parra Escartín, C., Waszczuk, J., Constant, M., Osenova, P., and Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Scannell, K. P. (2003). Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-Mer, France.

Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014a). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Schneider, N., Hovy, D., Johannsen, A., and Carpuat, M. (2016a). SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Schneider, N., Hwang, J., Srikumar, V., Green, M., Conger, K., O'Gorman, T., and Palmer, M. (2016b). A corpus of preposition supersenses in English web reviews. In *Proceedings of the 10th Linguistic Annotation Workshop*, Berlin, Germany.

Schneider, N., Hwang, J. D., Srikumar, V., Prange, J., Blodgett, A., Moeller, S. R., Stern, A., Bitan, A., and Abend, O. (2018). Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014b). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Schneider, N. and Smith, N. A. (2015). A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Senellart, J., Dienes, P., and Váradi, T. (2001). New generation systran translation system. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer Netherlands.

Silveira, N., Dozat, T., Marneffe, M. D., Bowman, S. R., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *In Proceedings of LREC*, pages 2897–2904, Reykjavík, Iceland.

Simova, I. and Kordoni, V. (2013). Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 62–71, Nice, France.

Sinha, K., Parthasarathi, P., Pineau, J., and Williams, A. (2021). UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Skadina, I. (2016). Multi-word Expressions in English-Latvian machine translation. *Baltic Journal of Modern Computing*, 4:811–825.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Stenson, N. (1981). *Studies in Irish syntax*. Ars linguistica. Tübingen: Gunter Narr Verlag.

Stenson, N. (1997). Language contact and the development of Irish directional phrase idioms. *Dán do oide: essays in memory of Conn R. Ó Cléirigh*, pages 559–577.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tan, L. and Pal, S. (2014). Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tanabe, T., Takahashi, M., and Shudo, K. (2014). A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. *Comput. Speech Lang.*, 28(6):1317–1339.

Taslimipoor, S., Bahaadini, S., and Kochmar, E. (2020). MTLB-STRUCT @Parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Tayyar Madabushi, H., Gow-Smith, E., Scarton, C., and Villavicencio, A. (2021). AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tesnière, L. (1959). *Eléments de Syntaxe Structural*. Paris: Librairie C. Klincksieck.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Tóth, K., Farkas, R., and Kocsor, A. (2008). Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm. *Acta Cybern.*, 18:463–478.

Tredinnick-Rowe, J. (2017). The semiotics of minority language branding: A study of the Celtic languages. *International Journal of Marketing Semiotics*, 5:2–16.

Tsvetkov, Y. and Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(4):549–573.

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.

Ultach (2020). I've discovered that almost every single article on the Scots version of Wikipedia is written by the same person - an American teenager who can't speak Scots. `https://www.reddit.com/r/Scotland/comments/ig9jia/ive_discovered_that_almost_every_single_article/`. Accessed: 2022-03-30.

van der Goot, R. (2021). We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vanmassenhove, E. and Way, A. (2018). SuperNMT: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73, Melbourne, Australia. Association for Computational Linguistics.

Varile, G. B. and Lau, P. (1988). Eurotra practical experience with a multilingual machine translation system under development. In *Second Conference on Applied Natural Language Processing*, pages 160–167, Austin, Texas, USA. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Veselinović, E. (2006). How to put up with *cur suas le rud* and the Bidirectionality of Contact. *The Celtic Englishes*, 4:173–190.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding, MUC6 '95*, pages 45–52.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. (2005). Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comput. Speech Lang.*, 19(4):365–377.

Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004). Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain. Association for Computational Linguistics.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic. Association for Computational Linguistics.

Vincze, V. (2012). Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2381–2388, Istanbul, Turkey. European Language Resources Association (ELRA).

Vincze, V., Nagy T., I., and Berend, G. (2011). Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria. Association for Computational Linguistics.

Vincze, V., Nagy T., I., and Farkas, R. (2013). Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Sofia, Bulgaria. Association for Computational Linguistics.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Walsh, A., Bonial, C., Geeraert, K., McCrae, J. P., Schneider, N., and Somers, C. (2018). Constructing an annotated corpus of verbal MWEs for English. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193–200, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Walsh, A., Lynn, T., and Foster, J. (2020). Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, L. and Yu, S. (2010). Construction of Chinese idiom knowledge-base and its applications. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 11–18, Beijing, China. Coling 2010.

Waszczuk, J., Savary, A., and Parmentier, Y. (2017). Multiword expression-aware A* TAG parsing revisited. In *Proceedings of the 13th International Workshop on Tree Adjoining Grammars and Related Formalisms*, pages 84–93, Umeå, Sweden. Association for Computational Linguistics.

Way, A. and Gaspari, F. (2019). PRINCIPLE: Providing resources in Irish, Norwegian, Croatian and Icelandic for the purposes of language engineering. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 112–113, Dublin, Ireland.

Wigger, A. (2008). Advances in the lexicography of Modern Irish verbs. In *Bloch-Rozmej, A. (ed.): Issues in Celtic Linguistics. Lublin Studies in Celtic Languages 5*, pages 233–250.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., and Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X.,

Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J. R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G. S., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Yirmibeşoğlu, Z. and Güngör, T. (2020). ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135, online. Association for Computational Linguistics.

Yu, L. and Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

Zaninello, A. and Birch, A. (2020). Multiword Expression aware Neural Machine Translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökirmak, M., Nedoluzhko, A., Cinková, S., Jan Hajič Jr., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Alonso, H. M., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada.

Appendix A

Guidelines for Annotating Irish MWEs

These guidelines have been partially adapted from PARSEME Annotation Guidelines.

As a note, embedded MWEs are not annotated in this pass. For example, the `VPC` *cuir suas* in the `IAV` *cuir suas le* would be annotated as `IAV`, to match the outermost MWE label.

A.1 Nominal Compound (NC)

Nominal compounds (NCs) are compound noun phrases that consist of a head noun and a dependent noun or adjective, where the compound can be said to be semantically semi-compositional or non-compositional.

Also included in this category are technical or specialised language terms, such as those included in the Tearma corpus.

Category	**Example**	**Tag**
Nominal Compounds	*mac tíre*	NC
Named Entity	*Baile Átha Cliath*	NE
Institutionalised Phrase	*domhan uile*	IP
Copular Construction	*is maith le*	CC
Light Verb Constructions	*déan obair*	LVC
Inherently Adpositional Verb	*bain le*	IAV
Verbal Idiom	*an lá atá inniu*	VID
Verb Particle Construction	*leag amach*	VPC
Fixed Expression	*in aghaidh*	FE

Table A.1: Table of categories and codes

Test NC.1: Noun phrase is non-compositional

Does the combination of noun and noun, or adjective and noun, lend a non-compositional or semi-compositional meaning to the noun?

If yes, annotate as `NC`.

(165) *mac tíre*

(166) *Mí na meala*

If not, continue to test `NC.2`.

(167) *nósanna imeachta*

(168) *Cothrom an lae*

Test NC.2: Specialised term

Can the construction can be considered a technical term or does it have a specific meaning within a certain domain, even if the meaning is compositional?

If yes, annotate as `NC`.

(169) *Uiscebhealaí intíre*

(170) *Aigéad sulfarach*

If not, do not annotate.

(171) *Grúpa cannon*

(172) *Taifead ábhartha*

A.2 Named Entities (NE)

Named entities are all proper noun phrases that make up a recognisable entity; these include but are not limited to proper names, organisations or agencies, place names, compound months and foreign titles.

The category is similar to the MWE flat relation in the UD guidelines for the IDT, however there are some differences. While the flat relation is always used for syntactically flat constructions (i.e. there is no internal syntax or hierarchy), named entities may include syntactically structured constructions such as official committee titles (e.g. "*An Ghníomhaireacht Eorpach chun Comhar Oibríochtúil a Bhainistiú ag Teorainneacha Seachtracha Bhallstáit an Aontais Eorpaighle*"). Also, syntactically flat constructions that do not form a recognisable entity, such as dates, are annotated with the flat relation, but are not considered here as named entities, due to their productivity.

Test NE.1: Noun phrase forms an entity

Does the construction takes the form of a noun phrase that represents some recognisable entity?

If yes, annotate as `NE`.

(173) *Deireadh Fómhair*

(174) *Na Náisiún Aontaithe*

(175) *Comhairle Cathrach Bhaile Átha Cliath*

A.3 Fixed Expressions (FE)

Fixed expressions (FE) are those that have no internal modification permitted, and do not inflect for any grammatical change.

This category closely aligns with the fixed label in the UD annotation guidelines, but is expanded slightly to include constructions that are semi-fixed.

Test FE.1: Unit of language

Does the construction form a linguistic or semantic unit, i.e. all the words in the construction contribute towards a non-productive constituent whose words cannot be replaced by others of the semantic class, and in which the entire construction modifies the sentence in some particular way.

If yes, or unsure, continue to test `FE.2`.

(176) *Ceart go leor*

(177) *Chomh maith*

(178) *Go dtí*

If no, do not annotate.

(179) *Ar an*

(180) *Bhain sé*

Test FE.2: Syntactically fixed

Can the construction undergo inflection or internal modification?

If yes or unsure, continue to test `FE.3`.

(181) *ina dhiaidh*

If no, annotate as `FE`.

(182) *Tar éis*

(183) *Le linn*

(184) *Ar chor ar bith*

Test FE.3: Syntactically semi-fixed

Is the modification or inflection that can occur only to show grammatical information such as number or person? If yes, annotate as fixed.

(185) *Ina dhiaidh* → *i mo/do dhiaidh*

Otherwise, do not annotate.

A.4 Institutionalised Phrases (IP)

Institutionalised Phrases (IP) are expressions whose meaning may not be entirely idiomatic, but the lexical items become fixed through conventions of language use.

It is similar to the category of fixed expression, but while fixed expressions are typically non-productive, shorter constructions that modify the sentence in some way, institutionalised phrases can take the form of longer, more complex constructions.

Test IP.1: Fixed lexical usage

Does conventional language use tend to favour a certain selection of lexical items, rather than a semantically equivalent lexical item?

If yes, annotate as `IP`.

(186) *Gruth agus meadhg*

(187) *Seomra is cistin*

(188) *Scuaine lachan*

(189) *Chomh críonna le sionnach*

If no, do not annotate.

(190) *Greim docht*

A.5 Light verb constructions (LVC)

Light verb constructions (LVC) are formed by a verb v and a (single or compound) noun n, which either directly depends on v (and possibly contains a case marker or a postposition), or is introduced by a preposition.

(191) *Bain triail as*

(192) *Cuir lúchair ar*

(193) *Déan iarracht ar*

(194) *Cuir fuil-shrón le*

(195) *Déan dearmad ar*

(196) *Tabhair faoi deara*

(197) *Déan dreas cainte*

The (single or compound) noun n is predicative and refers to an event (e.g. decision, visit) or a state (e.g. fear, courage). Predicative nouns are nouns that have semantic arguments, that is, they express predicates whose meaning is only fully specified by their semantic arguments.

Test LVC.1: Noun is abstract

Is the noun abstract (i.e. does it denote a quality, state, or idea)?

If yes, continue to test `LVC.2`.

(198) *dearmad*

(199) *trial*

(200) *fearg*

If no (i.e. denotes a concrete object), do not annotate.

(201) *cathaoir*

(202) *lámh*

(203) *leabhar*

Test LVC.2: Noun is predicative

Does the noun n have at least one semantic argument, implying that it is a predicative noun?

If yes, continue to test `LVC.3`.

(204) *Tabhair cuairt ar* → event with two arguments: the visitor and the visitee

If no, do not annotate.

(205) *Cuir tuí*

(206) *Tabhair peann ar*

Test LVC.3: Verb's subject is noun's semantic argument

Is the subject of the verb a semantic argument of the noun? In other words, is the verb linking the predicative noun to one of its semantic arguments that occurs as the subject of the verb?

If yes, continue to test `LVC.4`.

(207) *Chaith Seán vóta* → Seán is the subject of the verb and a semantic argument (the voter) of the noun

If no, do not annotate.

(208) *Chomhair Séan na vótaí* → Vote does not have a semantic argument of the counter

Test LVC.4: Verb with light semantics

Is v semantically light, that is, is the semantics that v adds to n restricted to: (i) what stems from its morphological features (e.g. future, plural, perfective aspect, etc.), (ii) pointing at the semantic role of n played by v's subject?

If yes, continue to test `LVC.5`.

(209) *Rinne mé iarracht ar m'obair bhaile* → *rinne* adds no meaning to *iarracht* except performing an activity

If not, go to test LVC.6.

(210) *Thosaigh mé iarracht ar m'obair bhaile* → *thosaigh* has an added aspectual meaning

Test LVC.5: Verb reduction

Try to build an NP without the verb, in which v's subject s becomes n's dependent. You might need to test several prepositions, possessives, case markers, etc, as long as you use no verb. Can this verbless NP refer to the same event or state as the candidate v+n construction does? (This is a test using an ownership construction).

If yes, annotate as `LVC.full`.

(211) *Déanann Mícheál cur síos ar* → *an* ***cur síos*** *do* ***Mhícheál***

(212) *Thug mé tacaíocht do Mháire* → ***mo thacaíocht féin***

If no, do not annotate.

(213) *Fuair Máire tacaíocht ó Sheán* → an tacaíocht a fuair Máire

Test LVC.6: Verb's subject is noun's cause

Is the subject of the verb expressing the cause of the predicate expressed by the noun? In other words, does the verb bring an additional participant to the scene, representing the source or cause of the event or state referred to by the noun?

If yes, annotate as `LVC.cause`.

(214) *Chuir Aoife áthas orm* → The happiness was inspired by Aoife

If not, do not annotate.

(215) *Chuir Aoife airgead sa bhanc* → The money was not caused by Aoife

A.6 Verb-particle constructions (VPC)

Verb-particle constructions (VPCs) are formed by a lexicalized head verb v and a lexicalized particle p dependent on v. Examples include constructions such as *tarraing anuas*, *cas as*, *tabhair amach*, etc.

The meaning of the VPC can be fully or partly non-compositional. In fully non-compositional `VPC` (`VPC.full`) the change in the meaning of v goes significantly beyond adding the meaning of p.

(216) *Cas as*

(217) *Tabhair amach*

In semi-non-compositional `VPCs` (`VPC.semi`), p adds a partly predictable but non-spatial meaning to v.

(218) *Tabhair suas*

(219) *Glan suas*

Test VPC.1: Verb without the particle refers to the same event/state

Can a sentence without the particle refer to the same event/state as the sentence with the particle? Special care must be taken when the same construction might or might not be a valid VPC depending on its context.

If no, annotate as `VPC.full`.

(220) *Cas as* → 'put out' does not imply to turn

(221) *Tarraing anuas* → draw down (i.e. bring up) does not imply to draw in a literal sense

(222) *Seas amach* → stand out does not imply to stand

If yes, go to test `VPC.2`.

(223) *Scríobh síos* → *scríobh* implies to write

(224) *Glan suas* → *glan* implies to clean

(225) *Féach amach* → *féach* implies to look

(226) *Éirigh amach* → *éirigh* against implies to rise

Test VPC.2: Spatial particle

Is the particle spatial in the context of the verb, i.e. does it express direction or position?

If no, annotate as `VPC`.

(227) *Glan suas* → *suas* is not directional here, but rather implies completely

(228) *Éirigh amach* → *amach* implies 'out' in a rebellious way

If yes, do not annotate.

(229) *Seas le chéile*

(230) *Féach amach*

A.7 Inherently adpositional verbs (IAVs)

Inherently adpositional verb (IAV) is considered a special and experimental category. It consists of a verb or VMWE and an idiomatic selected preposition that is either always required or, if absent, changes the meaning of the verb of VMWE significantly. `IAV` constructions should be annotated only after annotating `LVC` or `VPC` constructions, since this category can overlap with these two. However, we do not consider this category as overlapping with either `CC` or `VID` categories, so they must be annotated after this category.

(231) *Éirigh as*

(232) *Buail le*

Test IAV.1: Circumstantial question with no adposition

Note: This is an adaptation of STREUSLE's guideline on prepositional verbs by Nathan Schneider and Meredith Green. In response to a declarative sentence with the verb+adposition combination, is there a natural way to query the circumstances of the verbal event using the verb, but not the adposition?

If no, annotate as `IAV`.

(233) *Cuireann sé sin orm → #Cén sort rud a gcuireann tú? → Cuir ar* is annotated as IAV

If yes, do not annotate.

(234) *Sheas mé ar an mbord → Cén fáth ar sheas tú ann? → Seas ar* is not annotated as IAV

A.8 Idiomatic Copular Constructions (CC)

Idiomatic copular constructions are constructions formed with the copula and one or more arguments, where the construction has a meaning that is non-compositional from its component words.

Test CC.1: Non-compositional meaning

Is the meaning of the construction the same as the sum of its parts (i.e. do the components of the construction add meaning beyond their individual meanings to the construction?).

If yes, annotate as `CC`.

(235) *Is le* → idiomatic construction indicating possession

(236) *Is maith le* → idiomatic construction indicating enjoyment

If no or unsure, continue to test `CC.2`.

Test CC.2: Lexical inflexibility

Are each of the components of the construction lexically inflexible, so that replacing one token with another from the same semantic class would be incorrect?

If yes, annotate as `CC`.

(237) *Is chóir* → **is ceartas* ungrammatical despite the similar meaning of *ceartas* and *cóir*

A.9 Verbal Idioms (VID)

Verbal idioms constitute a universal category. A verbal idiom (VID) has at least two lexicalized components including a head verb and at least one of its dependents.

(238) *Is buí le bocht an beagán*

(239) *Ag cur madraí i bhfuinneoga*

(240) *Tá dhá thaobh ar an mbád*

(241) *Déan cat is dhá eireaball ar*

(242) *Gléasta go barr na méar*

(243) *Moll an óige agus tiocfaidh sí*

Idiomatic constructions with the copula are currently not annotated as VID, given that the syntactic head of the construction is not the verb in these cases:

Máire is ainm dom → head is noun *ainm*

Test VID.1 Fails other MWE tests

Does this construction fail as a different type of MWE (`LVC`, `VPC`, `IAV`, or `CC`)?

If yes, continue to test `VID.2`.

(244) *Caith an phingin i ndiaidh an phuint*

(245) *Bulla dall a dhéanamh de*

If no, annotate as required.

Test VID.2: Non-compositional meaning

Is the construction semantically non-compositional, i.e. the meaning cannot be derived entirely from the lexicalised components?

If yes, annotate as `VID`.

(246) *Bheith ar an bpláta beag* → sense of being in jail is not evident from components

(247) *Cat a scaoileadh as an mála* → sense of telling a secret not evident from components

If no, do not annotate.

(248) *Bheith ag obair*

(249) *Madra a fheiceáil ar an mbóthar*

Appendix B

Experiments in MWE Processing

The content in this Appendix details some exploratory experiments in the task of MWE discovery. This task is described in Chapter 2, and is concerned with the extraction of MWE types from text corpora. The output of this task is often a lexical list or similar resource, and as such, this task encompasses an important step for the building of lexical resources. Furthermore, as this task frequently employs unsupervised methods, it presents fewer obstacles for low-resource languages.

B.1 Statistical Measures

It is a common practice in linguistics and NLP to define a word not just based on the semantic meaning of the word itself, but on its relationship with other words. The oft-quoted phrase by Firth (1957) sums up this philosophy quite succinctly: "You shall know a word by the company that it keeps."

This idea can be applied to identifying multiword expressions, as the component words of these expressions are frequently linked with their companion words with unusual closeness—this is what is meant by statistical idiosyncrasy. Determining these close word links can be undertaken by employing association measures. Church and Hanks (1989) proposes a method for finding the association measure of two words x and y (termed *association ratio* by Church), through calculating the Mutual Information (MI) of x and y. MI is essentially the probability, for each element, of observing x and y together, considering the probability of observing each element separately. The formula for calculating MI is thus given as:

$$I(X;Y) = \Sigma_{x \in X} \Sigma_{y \in Y} log_2 \frac{P(x,y)}{P(x)P(y)}$$

The logarithm (base 2) is used to convert to a value that can be expressed in bits. If x and y are independent, $P(x,y)$ is equal to $P(x)P(y)$, and so MI will be 0.

To calculate the word probabilities $P(x)$ and $P(y)$, it is sufficient to count the occurrences of these words in a corpus, and divide by the size of the corpus. The co-occurrence counts of these words can be given an optional window size, wherein the words may co-occur. A smaller window size will reveal tightly fixed words such as those in a fixed expression, while a wider window may reveal a semantic or other relationship between the words.

Variations of this formula for calculating word association have been employed for MWE identification and automatic filtering of extracted candidate expressions:

- Ramisch et al. (2008) compare the results of three language- and type-independent statistical measures of association: mutual information (MI) mentioned above, χ^2 which is a standard measure of association, and permutation entropy (PE) which is measure of order association; and one measure which is language- and type-dependent, the entropy of permutation and insertion (EPI), which incorporates linguistic information about the MWE type.

- Pecina (2009) lists an inventory of 82 word association measures for two-word (bigram) collocations, which fall into the categories of statistical association (i.e. type and language independent) and context-based association measures (i.e. type and language dependent). Pecina compares the effectiveness of these association measures (AMs) on four datasets; *PDT-Dep* and *PDT-Surf* (manually annotated MWE candidates taken from the *Prague Dependency Treebank*), *CNC-Surf* (extracted from the *Czech National Corpus*), and *PAR-Dist* (consisting of Swedish verb-noun combinations extracted from the Swedish *PAROLE* corpus). By comparing the precision-recall curves, and the mean average precision scores, Pecina found that different AMs performed best on different datasets, indicating that there is no objectively superior measure of lexical association, and the performance of AMs will depend on the data and the task. Pecina also found that combining association measures can achieve a substantial improvement in performance.

Some initial experiments are discussed below.

B.2 Extracting MWEs from corpus using PMI scores

Pointwise-Mutual Information (PMI) is a variation of the Mutual Information (MI) measure discussed in Section B.1. While MI measures the expectation of correlation across all elements, PMI is the individual correlation measure for those two events,

and is used frequently for filtering statistically idiosyncratic MWEs (Chang et al., 2002; Villavicencio et al., 2007; Ramisch et al., 2010; Salehi et al., 2016). When we extracted bigrams from the New Corpus of Ireland (Kilgarriff et al., 2006)[1] and filtered candidates with a PMI score above a certain threshold, several MWE types emerged.

Setting the PMI score threshold to > 10 and the bigram frequency to > 100, a list of 349 unique terms was returned, mostly consisting of named entities (e.g. *Aontas Eorpach* 'European Union'). There were also examples of compound nouns (*dhuine uasail* 'gentleman/sir'), foreign terms ('Irish Times'), fixed expressions (*os comhair* 'in front of'), light verb constructions (*baint úsáid* 'make use') and potentially statistically idiosyncratic but otherwise non-idiomatic bigrams (*haois déag* '-teen years (old)', *cupán tae* 'cup of tea'). Also discovered were several instances of reduplication expressions such as *fite fuaite* 'firmly interwoven' and *frois frais* 'mess'.

B.3 Extracting MWEs from misalignments in parallel texts

One of the challenges associated with identifying MWEs in Irish (RQ3) is the lack of MWE-specific resources available in Irish, particularly for training supervised systems to automatically recognise and extract MWEs. To overcome this, some languages have attempted language-independent and linguistically-naive methods of identifying MWEs. One such experiment was conducted by Tsvetkov and Wintner (2012) on Hebrew text, in which they were able to extract MWEs using PMI scores in conjunction with word alignment scores in parallel text.

PMI measures are frequently criticised as a means of extracting MWEs due to the susceptibility of bias towards statistically frequent words. For instance, we can reasonably expect the combination 'of the' to occur frequently in any English text, but this bigram should not be treated as an MWE. As such, Tsvetkov and Wintner proposed only calculating PMI scores for bigram candidates that had already been selected as potential MWEs, by automatically word-aligning text from parallel corpora and extracting the one-to-many word alignments generated.

Applying this method to the task of classification of Irish MWEs (RQ1), we proposed an experiment to extract new MWE candidates from misaligned bigrams and filter these candidates using PMI scores. The resources required for this experiment are a monolingual dictionary, a monolingual corpus, a parallel corpus, all of which are available for the Irish language, and a language-independent word-alignment

[1]Approximately 30 million words

tool. Irish-English parallel data taken from the bilingual Gaois corpus[2] was used. The texts were aligned using FastAlign (Dyer et al., 2013). The results of this alignment attempt were not useful in this task, as word-alignments were over-generated, with every word in the Irish corpus aligned to nearly every word in the English corpus. This experiment will be revisited in the future, with a different alignment method, and a larger dataset.

B.4 MWEtoolkit and extracting from the UD treebank

Syntactic patterns are frequently used in automatic identification of MWEs. MWEtoolkit (Ramisch, 2015) is an open source, generic framework for extracting and processing MWEs, which makes use of regex-like queries to find MWE candidates in a language-independent context. The toolkit also includes tools for filtering and ranking these candidates, and extracting them to form a lexicon. The toolkit requires data that has been first preprocessed externally. The text must be tokenised, lemmatised, POS-tagged, and dependency parsed. Using regex, syntactic patterns can then be defined to identify MWE candidates of different types. By defining syntactic patterns, we extracted MWE candidates from the UD Irish treebank for some MWE types, including compound nouns, prepositional phrases, light verb constructions and inherently adpositional verbs (many of these candidates are annotated in the treebank using the `fixed`, `compound` and `compound:prt` labels). These candidates can then be ranked using a selection of Association Measures for better results.

[2] `https://www.gaois.ie/crp/en/`

Appendix C

Creating an Annotated Corpus of Verbal MWES for English

C.1 Data

There were several considerations when selecting appropriate text for inclusion in this corpus. This section describes the selection criteria, followed by a description of the annotation tool used. The suggestions for selecting an appropriate source of data were provided by PARSEME in the language leader guidelines, and were informed by version 1.0 of the shared task. Of those suggestions, the following criteria were deemed to be of the highest priority:

1. The corpus should be available under an open licence
2. The text must be originally written in English
3. The text should be annotated for morphosyntactic information
4. The size of the corpus should allow for at least 3,500 MWE annotations
5. The language must be of sufficiently high quality

There were several corpora considered for selection, including the DiMSUM corpus (Schneider et al., 2016b), the UP/TAP corpus,[1] Wikidata parallel text (Vrandečić and Krötzsch, 2014) and the Universal Dependencies (UD) treebanks.[2] Three corpora from the UD treebanks for English were ultimately selected as a source of data, as they alone fulfilled the criteria mentioned above: text was selected from the English-EWT corpus (Silveira et al., 2014),[3] the LinES parallel corpus (Ahrenberg, 2007) and the Parallel Universal Dependencies (PUD) treebank (Zeman et al.,

[1]Documentation for UP/TAP: `https://www.l2f.inesc-id.pt/~thomas/metashare/report-UP-TAP.pdf`

[2]Documentation for UD: `http://universaldependencies.org`

[3]Originally sourced from the English Web Treebank (Bies et al., 2012)

2017).[4] The files were extracted in CoNLL-U format and converted to FoLiA XML format (see section C.2) for annotating. The training, development and testing datasets for each treebank were concatenated, and then split into files of 201 sentences for annotation.

C.2 Annotation

During the data preparation period, annotators were trained in the use of the FoLiA Linguistic Annotation Tool (FLAT). FLAT is an open-source web-based environment,[5] using the XML-based FoLiA format. In order to aid annotators in annotating only verbal MWEs, FLAT highlights verbs using POS information taken from the CoNLL-U file. Figure C.1 shows a screenshot of the FLAT platform, demonstrating how the selecting and annotating of lexicalised components works. Annotators were also trained to recognize and categorise VMWEs of different types, detailed in the sections to follow.

The annotation team was comprised of volunteers who had experience with or interest in annotating multiword expressions, and were all native speakers of English. Four dialects of English were represented: Irish English, British English, American English and Canadian English.

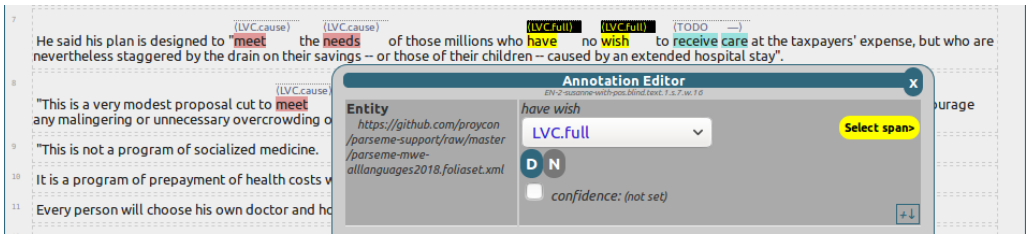


Figure C.1: Screenshot of the FLAT Platform

C.2.1 Categories of VMWE

Seven categories of VMWE were used in the English annotation task: Verbal Idioms (VID), Verb-Particle Constructions (VPC.full and VPC.semi),[6] Light-Verb Constructions (LVC.full and LVC.cause),[6] Multi-Verb Constructions (MVC) and Inherently Adpositional Verbs (IAV). The categories are divided into universal categories (valid for all participating languages), quasi-universal categories (valid for

[4]Though not a part of the task dataset, we have also fully annotated the Reviews portion of the UD English-EWT corpus by adding VMWE types to the existing VMWEs in STREUSLE (Schneider et al., 2014b; Schneider and Smith, 2015, `https://github.com/nert-gu/streusle/`); they were previously uncategorized. STREUSLE as of version 4.1 comprises 3812 sentences and 871 VMWE instances (121 IAV, 12 LVC.cause, 123 LVC.full, 310 VID, 206 VPC.full, 99 VPC.semi).

[5]`http://flat.readthedocs.io/en/latest/`

some language groups or languages), and an experimental category (which may be optionally considered for some languages).

Verbal idioms (VIDs) and the **Light-Verb Constructions** (LVCs) constitute universal categories. VIDs have at least two lexicalised components, including a head and at least one dependent. Dependents can be of different grammatical roles and parts of speech, meaning VIDs may be confused with other categories of VMWEs, such as LVCs. VIDs also include sentential expressions with no open slots, such as proverbs.

VID: *to* ***take*** *something* ***with a pinch of salt***: VMWE with an adverbial complement

LVCs are formed by a verb and a single or compound dependent noun. The noun must be abstract and predicative. The verb can be of two types: a 'light' verb, which arguably contributes no extra semantics to the expression beyond the semantics denoted by the predicative noun (annotated as LVC.full), and a 'causative' verb, which contributes only the semantics of causation, as the subject of the verb is the cause or source of the event or state expressed by the dependent noun (annotated as LVC.cause).

LVC.full: *to* ***make*** *a* ***decision***: verb adds nothing substantive to the semantics of 'decision'

LVC.cause: *to* ***give*** *a* ***headache***: the subject of 'give' is the cause of the headache

Verb-Particle Constructions (VPCs) and **Multi-Verb Constructions** (MVCs) are quasi-universal categories that are applicable to English. VPCs are formed by a verb and a dependent particle. The verb can be either fully non-compositional, where the addition of the particle changes the meaning of the verb significantly (annotated as VPC.full), or semi-non-compositional where the particle adds a partially predictable but non-spatial meaning to the verb (annotated as VPC.semi).

VPC.full: *to* ***check in*** *upon arrival*: omitting 'in' leads to very different meaning

VPC.semi: *to* ***eat*** *the cookies* ***up***: 'up' adds a sense of completion, but not a spatial meaning

MVCs are composed of two adjacent verbs, one of which is a governing verb and the other a dependent verb; together they function as a single predicate. The test for this category in English involves replacing the dependent verb with another verb from the same semantic class. If this leads to ungrammaticality or an unexpected change in meaning, the expression is categorised as MVC.

[6]New categories added to edition 1.1 of the shared task

MVC: *to* ***let go*** : replacing 'go' with 'depart', 'move', etc. changes the meaning significantly

Inherently Adpositional Verbs (IAVs) constitute an experimental category that has been included in the English annotation. IAVs consist of a verb and an adposition that is integral to the meaning of the expression. The guidelines include a test to differentiate between adpositions and particles, the former of which are exclusively used in IAVs.[7]

IAV: *to* ***come across*** *something*: omitting adposition 'across' leads to very different meaning

IAVs may also contain particles: e.g. *to* ***put up with*** *something* (verb+particle+preposition) means to endure it, and cannot have this meaning absent *up* or *with*.

C.2.2 Pilot Annotation Tasks

Three pilot annotation tasks were held to allow annotators to familiarize themselves with FLAT and the guidelines, as well as raise any potential issues and disagreements concerning the categorisation of VMWEs in English. Two small corpora were used for the first two pilot annotations, consisting of 200 sentences taken from the Brown corpus.[8]

Pilot annotation 1 was held in the beginning of June 2017, using version 1.0 of the guidelines. Following a discussion of this task, many disagreements seemed to stem from LVC tests that were difficult to apply and did not cover all cases. For example:

The grand jury ***took*** *a* ***swipe*** *at the State Welfare Department...*: While annotators felt this should be categorised as LVC, the original tests for LVC state that the noun must be used in one of its original senses, i.e. non-idiomatic use of the word, which would cause this expression to fail as an LVC. In response, the noun requirements within LVCs were generalized such that the noun must only be predicative, but need not retain one of its senses used outside of LVCs.

Annotators came across cases of LVCs that were fairly straightforward because the verb quite clearly adds little semantics beyond that of the predicative noun (e.g., *She has a terrible headache*). However, variant expressions with a different light verb were not clearly LVCs, given a minimal amount of causative semantics contributed by the verb (e.g., *The buzzing radio gave him a headache*). Such cases

[7]Adpositions are fixed in occurring exclusively before a noun phrase, unlike particles, which either modify an intransitive verb (***check in***) or are mobile with respect to full noun phrase complements (***eat*** *the cookies* ***up***/***eat up*** *the cookies*).

[8]Access the Brown Corpus Manual here: `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM`

were another source of disagreement in LVC annotations. To accommodate both types of LVCs while maintaining an acknowledgment of the causative semantics, it was decided after discussion to provide a distinction in the guideline tests between a fully light verb (LVC.full) and a causative light verb construction (LVC.cause). Causative light verbs, unlike fully light verbs, contribute the semantics of causation to the expression by licensing an outside causer or agent semantic role assigned to the verb's subject.

Other disagreements centered around unclear tests for particles, particularly particles which contribute aspectual or other subtle information, but do not significantly alter the meaning of the verb, leading to inconsistencies in VPC annotation. For example:

*...the Senate **passed** the bill **on** to the House*: Here the verb keeps its meaning but the particle contributes non-compositionally. Like LVCs, it was decided to subdivide the VPC category into the VPC.full and VPC.semi categories described above; thus improving agreement on borderline VPC.semi examples like this instance.

Pilot annotation 2 took place towards the end of November 2017, following the rewriting of the guidelines into version 1.1. In discussing the new annotation guidelines, some issues were raised. Many of these again centered around LVCs, including the productive nature of candidates in the new LVC.cause category (such productivity runs somewhat counter to the expectations of idiosyncrasy and lexicalisation for all VMWEs), as well as disagreements surrounding nouns categorized as either concrete or abstract (nouns within LVCs must be abstract and predicative). For example:

*A certain **vagueness** may also be **caused** by tactical appreciation of the fact...*: Here, and in all cases of *cause* in combination with an abstract/predicative noun, it was debated as to whether these should be considered LVC.cause. *Cause* expressions seemed to defy the normative expectations of idiosyncracy and lexicalization put forth for all VMWEs given that, unlike other light verbs, *cause* seems to combine productively with any predicative noun and the resulting expression is felicitous while maintaining its purely compositional semantics. After discussion, it was decided that such cases should be included as LVC.cause; however, a note that these cases do not exhibit some of the hallmarks of other LVCs and MWEs was added to the guidelines.

*The scholarship plan would **provide** federal **contributions** to each medical and dental school equal to $1500...*: It is unclear here if the noun *contributions* should be understood as abstract and predicative, or if it refers to the concrete contribution of the specific sum of money mentioned later in the sentence. If the noun

is understood as abstract and predicative, then the expression could be considered a case of LVC.cause, given that presumably the *scholarship plan* is an outsider causer of the contribution, while *federal* likely refers to the actual contributer. Additional guidance on distinguishing abstract and concrete nouns was added in response.

Related to distinguishing IAVs from VPCs, annotators also expressed confusion regarding the difference between particles and adpositions (and the recently added test to differentiate). For example:

...to ***set aside*** *the privilege resolution*: The categorisation of this expression was controversial because of uncertainty as to whether 'aside' could be considered a particle, and thus, belonging to a VPC. Following this confusion, tests for differentiating between adpositions and particles were featured more prominently as part of the decision tree for categorising VPCs.

After clarifying some of the intended interpretations and tests in the guidelines, it was decided to hold a third round of pilot annotations for English, reusing the corpus from the second pilot task, during the month of December 2017. **Pilot annotation 3** led to a more informed, robust discussion of the previous issues, and concluded with amendments to the guidelines, including notes regarding the productive characteristic of many LVC.cause VMWEs and additional pointers for distinguishing IAVs with adpositions from VPCs with particles.

Table C.1 in Section C.3 shows the number of VMWEs that were annotated during each pilot task, and the breakdown of categories that were annotated. Note that the categories VPC.full and LVC.full represent VPC and LVC respectively for Pilot 1, as the fine-grained labels did not exist in version 1.0 of the guidelines. Similarly, the optional category IAV was not considered for the first pilot task.

After iteration throughout piloting, the 1.1 edition of the guidelines were finalized for all languages. Several of the changes to the guidelines came about due to challenges with annotation of English VMWEs during the pilot annotation task, namely the subdivision of the LVC and VPC categories.

C.3 Corpus Annotation and Results

The annotation of the final corpus took place between the start of January 2018 and the end of February 2018. During this period, a total of 7437 sentences (124,202 tokens) were annotated. 4221 of these sentences were from the English Web Treebank, 3015 were from the LinES parallel corpus, and the remaining 201 sentences were from the PUD treebank. Out of a total of 14,121 verbs, 832 were annotated as VMWEs. Table C.1 displays the categories of VMWE that were annotated. The most commonly annotated category of English VMWE is full Verb-Particle Con-

structions, followed by full Light-Verb Constructions.

Following the end of the annotation period, the corpus was prepared for release. The annotated files were downloaded from FLAT in FoLiA XML format and aligned with the original CoNLL-U files. The annotated data from each annotator was consolidated, and a consistency check was performed to ensure that VMWEs were consistently annotated across all the data. Following this stage, the FoLiA files were then merged with the aligned CoNLL-U files to be converted into PARSEME TSV format, which is the format of the released data.[9]

Following the release of the annotated corpus, a portion of the corpus (804 sentences) was selected for annotation by all four annotators, in order to measure the quality of the corpus. The categorisation of VMWE types is shown in table C.2. The table shows the greatest level of disagreement in the categorisation of LVCs, particularly the LVC.cause category. Despite having provided additional guidance on the subject in the guidelines, the general VMWE definitional requirement of idiosyncrasy may have affected the categorisation of LVC.cause, as many instances of LVC.cause appear regular, and thus annotators may find it counter-intuitive to label these candidates as VMWEs.

Category	**Pilot 1**	**Pilot 2**	**Pilot 3**	**Final**
VPC.full	40	33	49	297
VPC.semi	0	25	25	45
LVC.full	37	43	82	244
LVC.cause	0	21	44	43
VID	38	19	30	139
MVC	0	2	1	4
IAV	0	15	34	60
Total	**115**	**158**	**265**	**832**

Table C.1: Number of annotations per category.

The IAA scores between all the pairs of annotators are given in Table C.3. The agreement between annotators is fair, showing moderate agreement when calculating the span of annotation (*F-score* and *Kappa*), and substantial agreement when calculating the agreement of categorisation only (*Kappa-cat*). We see from the table that the agreement between the two annotators who completed all three pilot tasks (A3 and A4) is higher than the agreement between the two annotators who did not participate in the three pilot tasks (A1 and A2), in as far as annotating the span of the VMWEs (*F-score* and *Kappa*). This is not the case when only the category of the VMWE is considered (*Kappa-cat*).

[9]The full PARSEME shared task data can be found at: `https://gitlab.com/parseme/sharedtask-data/tree/master/1.1`

Category	**A1**	**A2**	**A3**	**A4**
VPC.full	27	41	62	41
VPC.semi	17	3	9	23
LVC.full	77	32	43	42
LVC.cause	28	2	5	11
VID	13	14	25	41
MVC	4	0	0	1
IAV	22	9	9	17
Total	**188**	**101**	**153**	**176**

Table C.2: VMWEs in doubly annotated corpus.

Pair	**#X**	**#Y**	***F-score***	***Kappa***	***Kappa-cat***
1x2	188	101	0.436	0.396	0.661
1x3	188	153	0.452	0.402	0.647
1x4	188	176	0.478	0.427	0.635
2x3	101	153	0.480	0.446	0.773
2x4	101	176	0.513	0.479	0.636
3x4	153	176	0.529	0.487	0.625

Table C.3: IAA scores between annotator pairs (X and Y) for a subset (804 sentences) of the corpus. *F-score* is the F-measure between annotators, and is an optimistic measure that ignores agreement due to chance. The kappa scores used for *Kappa* and *Kappa-cat* are variants of 2-raters Cohen's kappa. *Kappa* is a calculation of the rate of agreement of annotation for all verbs in the corpus, while *Kappa-cat* takes into account only those VMWEs where both annotators agreed on the span, and measures the agreement of categorisation for these VMWEs.