Irish Dependency Treebanking and Parsing

Teresa Lynn

B.Sc. in Applied Computational Linguistics

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

Doctor of Philosophy (Ph.D.)

under a Double Award Agreement

at the

School of Computing
Dublin City University

Department of Computing
Macquarie University

Supervised by
Dr. Jennifer Foster
Associate Prof. Mark Dras

January 2016

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. 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.: 98300890

Date:

Contents

List of Figures

List of Tables

Irish Dependency Treebanking and Parsing

Teresa Lynn

Abstract

Despite enjoying the status of an official EU language, Irish is considered a minority language. As with most minority languages, it is a 'low-density' language, which means it lacks important linguistic and Natural Language Processing (NLP) resources. Relative to better-resourced languages such as English or French, for example, little research has been carried out on computational analysis or processing of Irish.

Parsing is the method of analysing the linguistic structure of text, and it is an invaluable processing step that is required for many different types of language technology applications. As a verb-initial language, Irish has several features that are uncharacteristic of many languages previously studied in parsing research. Our work broadens the application of NLP methods to less-studied language structures and provides a basis on which future work in Irish NLP is possible.

We report on the development of a dependency treebank that serves as training data for the first full Irish dependency parser. We discuss the linguistic structures of Irish, and the motivation behind the design of our annotation scheme. Our work also examines various methods of employing semi-automated approaches to treebank development. We overcome the relatively small pool of linguistic and technological resources available for the Irish language with these approaches, and show that even in early stages of development, parsing results for Irish are promising. What counts as a sufficient number of trees for training a parser varies according to languages. Through empirical methods, we explore the impact our treebank's size and content has on parsing accuracy for Irish. We also discuss our work in cross-lingual studies through converting our treebank to a universal annotation scheme.

Finally we extend our Irish NLP work to the unstructured user-generated text of Irish tweets. We report on the creation of a POS-tagged corpus of Irish tweets and the training of statistical pos-tagging models. We show how existing resources can be leveraged for this domain-adapted resource development.

Acknowledgments

What a journey – as much an emotional journey as an intellectual one! Firstly, thank you to my family (Mam, Captain Dermot and Stephen) for believing in me from the start, and in particular when times got tough. I am forever indebted.

I owe an immeasurable degree of gratitude to my two supervisors, both at opposite ends of the world, who managed to seamlessly put me, and keep me, on the right track over this eventful journey. To Mark Dras, in Macquarie University for seeing my PhD potential in the first place, and being so supportive after the sudden and unexpected changes in geographical location. Thank you for being such an unphased, steady and dependable mentor. You have been an inspiration in many ways. To Jennifer Foster, for being a friend, mentor and superb role-model. I couldn't have been in better hands after moving back to Dublin. I am especially grateful for how you supported my 'plan' to finish up the writing of this thesis. Thank you so much for always being so dependable and timely with your feedback.

A particular thank you to Josef van Genabith who supervised me in the early days of the PhD, you set me off on a very concrete path from the beginning. Thank you for seeing my potential. Thank you also to Andy Way for encouraging me to pursue the possibility of a cotutelle PhD with DCU, back on Bloomsday 2010. It was particularly significant to submit this thesis on Bloomsday 2015! And collectively, thank you both for inspiring me to pursue this field back in my undergraduate days.

I am grateful to Elaine Uí Dhonnchacha for sharing your own research so willingly and generously with me, and for partaking in the early stages of the treebank development. Your input was invaluable and I am honoured to be able to join you in our quest for improving Irish LT.

A special thank you to Kevin Scannell for so graciously giving me the opportunity to work in St. Louis University with you on a Fulbright Scholarship – accepting my proposal without batting an eyelid, allowing me to share your enthusiasm for Irish LT, and of course, getting my spoken Irish back up to scratch! Míle buíochas a chara, táim ag tnúth le níos mó taighde a dhéanamh leat sa todhchaí.

To Nancy Stenson, to whom I was serendipitously led as a fellow Fulbrighter, thank you for your extensive and easy to digest discussion of Irish syntax. Your work made life so much easier than you can imagine. Thanks for your friendship and support also.

A huge thank you to Djamé Seddah for his consistent encouragement with treebank development. It meant a lot to have that mutual enthusiasm to feed me when times were challenging! And of course to Corentin Ribeyre for his annotation tool and magic scripting skills to extract my treebank stats!

To the research group in Macquarie University who made the PhD student life that little bit more fun every day, in particular the LFG gang Francois, Jette, Yasaman and Benni. To the lab friends who showed such relentless support during my time in DCU – Maria, Lorraine, Jennifer, Zeliha and Lamia. To John Judge for jumping on board the Irish language technology bandwagon with me and proving to be a team to be reckoned with!

On a very personal note, I want to sincerely thank Adam McLean in Sydney and Susan Fitzpatrick in Dublin for keeping me on the road to sanity. A PhD is a difficult journey for anyone, the events you helped me through cannot be overlooked nor my gratitude summed up with a simple "thank you". There is no doubt that I would not have got this far without you both.

In the same vein, thank you to the friends and family, both in Australia and Ireland, who were my reinforced concrete during the near- 'pack it in' stages – Jane, Laura, Lasairfhíona, Anne-Marie, Ruth, Lisa, Clare, Sarah, Ciara and Sally. A special thanks to Nuala for being a fantastic housemate and for putting up with my flighty comings and goings over the last year!

Thanks to Eimear Maguire, one of our newest recruits in the Irish LT gang, for proof-reading, checking my Irish grammar and for your annotation work. Hope you are inspired to carry the torch...

Special thanks is also due to the Fulbright Commission of Ireland and Enterprise

Ireland for the wonderful opportunity to do research at St. Louis University, Missouri. What a memorable experience – made even more special through the support of Colleen Dube and Lisa Nic an Breitheamh of Fulbright, two wonderful female role-models!

Last, but certainly not the least, I am thankful to CNGL at DCU for funding my final year of research and for providing travel grants to allow me to present and discuss my work at wider range of venues.

Chapter 1

Introduction

The role of technology in our daily lives is evolving at an enormous speed. The way in which we access knowledge and information has taken a dramatic shift towards digital media in the last fifteen years. Most people now opt to inform themselves, not only through print newspapers, magazines or books, but also through alternative digital resources such as the Internet, accessing news sites, social media, and blogs for example, often favouring digital over print. Today, most people have access to these new media either through home or work computers, tablets or smart phone technology. Schools world-wide are engaging with technology as a means of providing educational content, while slowly setting aside the traditional books and paper-based approach to learning. The content of all of these media, of course, is communicated through human language.

Language technology is the meeting of information technology and human language. For a language to have a presence on these digital platforms, language content (text) needs to be both computationally accessible and processable. A recent EU report on the status of language technology for European languages (Judge et al., 2012) compares this digital revolution to Gutenberg's invention of the printing press. At that time, printing made written content more widely available, facilitating communication, standardisation of written forms of major languages, and teaching and translation across languages. However, the repercussions of this meant that minority languages that thrived mostly on oral tradition did not have literature printed to the same degree as major languages. They slowly became restricted in their use, particularly in contexts where a major language was a dominant alternative, thus ultimately leading to their decline.

With this analogy, it is easy to understand the concerns that minority languages have in this new information technology age with regards to the impact that technology may have on their survival. This concern is real in the case of Irish as a minority language, where all speakers of Irish in Ireland also speak English. This means that if users find that digital content or language technology tools are not accessible in Irish, they can easily default to using English resources instead. In schools, if language learning for other better-resourced languages such as French and German, for example, is more attractive and fun through the use of technology, then Irish could become a less appealing subject, and deemed out-of-date. Compared to other EU languages, Irish is regarded as a low-resourced language in terms of language technology (Judge et al., 2012). If the Irish language is not sufficiently digitally accessible, this digital revolution could contribute to a further decline in its use, thus threatening its survival.

In this thesis, we report on our contribution to addressing this lack of language technology resources by developing (i) a dependency treebank, (ii) a deep syntactic parsing model for Irish, (iii) a gold-standard POS-tagged corpus of Irish tweets and (iv) a domain-adapted part-of-speech tagging model for Irish tweets.

Our treebank is a resource that provides linguistic information on the Irish language in a digitally accessible format. This treebank will be valuable not only for linguistic research, but also for the development of Natural Language Processing (NLP) tools. The treebank has already proved to be invaluable for training the first Irish statistical dependency parser. This parser enables automatic extraction of syntactic (grammatical) information from previously unseen Irish text, contributing to potential future development of other NLP tools such as hybrid machine translation, grammatical error detection, question answering, text summarisation, sentiment analysis, information retrieval and information extraction systems.

As part of a Fulbright[1] research visit to St. Louis University, I worked with Professor Kevin Scannell on the development of a POS-tagged corpus of Irish tweets and statistical part-of-speech tagging models for Irish Twitter text. We also see this as a significant contribution to Irish language research in the context of social media. In the past few years, an increased online presence of Irish speakers has enabled new ways in which to communicate through Irish. The variation of Irish that is used on Twitter is representative of an evolving language, and will prove interesting to socio-linguistic studies of Irish language use today. This POS tagger is valuable for a digital analysis of this new use of Irish.

Research and development of NLP resources for Irish is not only valuable from a language conservation perspective, but it also contributes to work in the area of less-resourced languages (LRLs). It helps us to understand a language better, and by evaluating new NLP techniques the outcomes may be beneficial to work on other minority languages or texts of new domains.

This chapter presents an overview of the considerations required in the development of these new resources. In Section 1.1, we explain the current status of the Irish language in relation to language technology. Through this, we highlight the need for new resources such as our treebank and parser. From there, in Section 1.2, we outline the work involved in developing a treebank for Irish, reflecting the difficulties encountered when dealing with a low-resourced language. In Section 1.3, we focus on the development of our parsing models for Irish and introduce the various ways in which we carried out parsing experiments using the Irish treebank. We also highlight the steps we took to address the impact a lack of resources had on the performance of the parser. In Section 1.4 we discuss the expansion of our work to social media NLP, specifically a part-of-speech analysis of Irish Tweets. In Section 1.5, we outline our research questions, followed by an outline of the structure of the thesis in Section 1.6 and a list of our publications in Section 1.7.

[1] Fulbright Enterprise Ireland Award 2014-2015.

1.1 Irish Language Technology

Many of the world's languages are referred to as *low density* languages. In the field of Natural Language Processing (NLP), 'low-density' refers to a lack of resources, usually digital, for a language. Basic language text resources make invaluable contributions to NLP research and provide the basis of the development of advanced NLP tools. Such resources include, among others:

- electronic dictionaries: often known as lexicons. These lexicons can contain semantic information, translations, synonyms, hypernyms, part-of-speech information and morphological information.

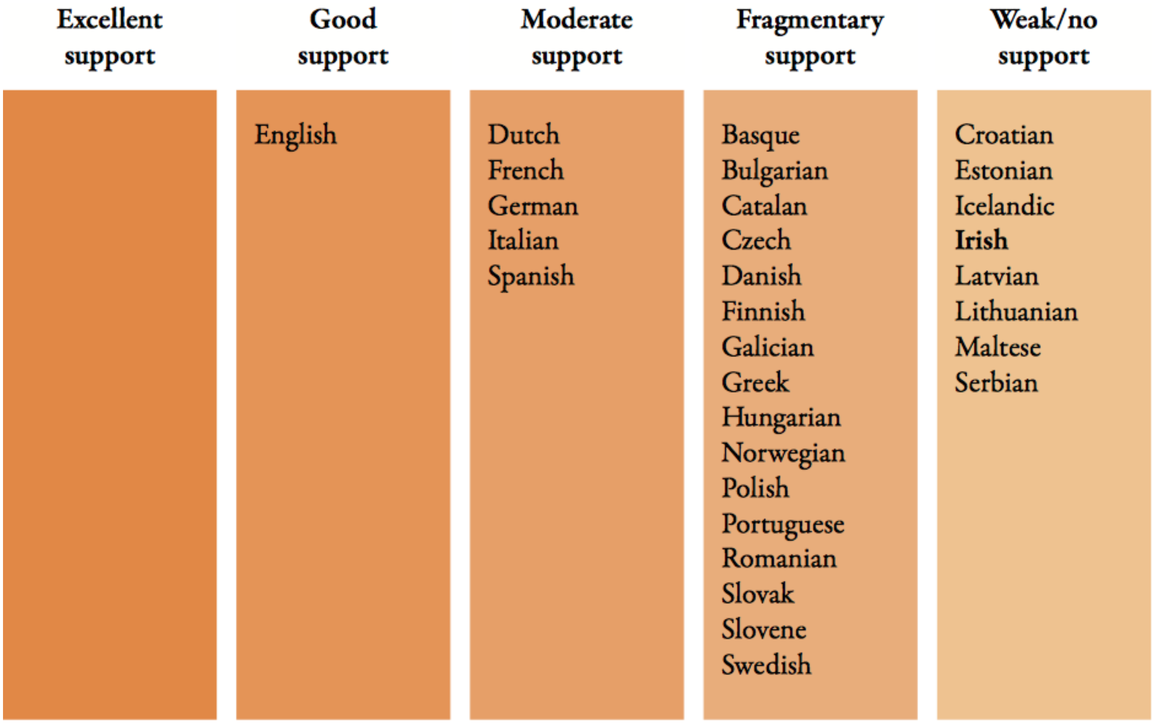
- bilingual corpora: aligned translations of text. Such resources are valuable for statistical machine translation.

- part-of-speech tagged corpora: text annotated with data which provides information on the part of speech of a word. This is particularly invaluable when confronted with ambiguous terms in a language. For example, in English, the word 'walk' may be a noun or a verb, and in Irish the word *glac* may be a noun ('handful') or a verb ('accept').

- treebanks: corpora that have been annotated with syntactic information. Parsing provides grammatical information regarding the structure of text, which is often crucial for successful computational analysis of a language.

- semantically tagged text: text which has been annotated with information that indicates the semantic category to which it belongs. For example, nouns may be animate or inanimate, countable or mass, common nouns or proper nouns, and so on.

To date, little research has been carried out on the Irish language with regards to computational analysis or processing, resulting in a lack of vital linguistic resources. From the list of basic resources above, only the first three items were available before the completion of our treebank. There are a number of electronic dictionaries and term bases available online, including *focloir.ie, potafocal.com, tearma.ie and teanglann.ie.*[2] Some of these resources are also available through smart phone applications. With regards to bilingual corpora, most of the parallel data that is freely available is legislative (e.g. *gaois.ie*, EU Joint Research Council[3]). A web-crawled parallel corpus is available for queries only (not for full download), and this text is more general domain.[4] Finally, a 3,000 sentence gold-standard POS-tagged corpus was made available by Uí Dhonnchadha (2009). There is no semantically annotated corpus available for Irish.



Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch	Basque	Croatian
		French	Bulgarian	Estonian
		German	Catalan	Icelandic
		Italian	Czech	**Irish**
		Spanish	Danish	Latvian
			Finnish	Lithuanian
			Galician	Maltese
			Greek	Serbian
			Hungarian	
			Norwegian	
			Polish	
			Portuguese	
			Romanian	
			Slovak	
			Slovene	
			Swedish	

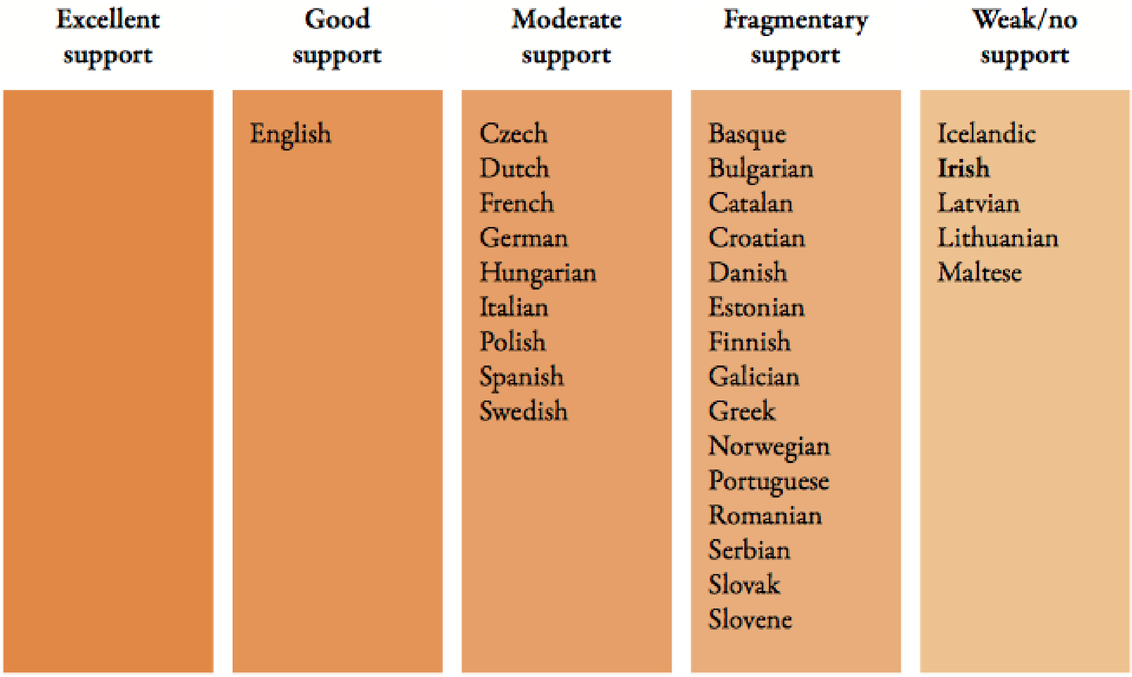
Figure 1.1: Text Analysis – state of language technology support for 30 EU languages (Judge et al., 2012)

Figures 1.1 and 1.2 from Judge et al. (2012)'s report show how Irish is positioned in relation to other European languages with regards to language technology. It is clear that Irish is significantly under-resourced when it comes to language technology. Since the publication of the report, a survey of Welsh language technology was also

[2] Fiontar, DCU is responsible for the development of the terminology database *tearma.ie*. They have also developed cultural online databases such as *duchais.ie, ainm.ie* and *logainm.ie.*

[3] The JRC provides a number of corpora at https://ec.europa.eu/jrc/

[4] Available to query at http://borel.slu.edu/corpas/index.html



Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech	Basque	Icelandic
		Dutch	Bulgarian	**Irish**
		French	Catalan	Latvian
		German	Croatian	Lithuanian
		Hungarian	Danish	Maltese
		Italian	Estonian	
		Polish	Finnish	
		Spanish	Galician	
		Swedish	Greek	
			Norwegian	
			Portuguese	
			Romanian	
			Serbian	
			Slovak	
			Slovene	

Figure 1.2: Speech and Text resources – state of language technology support for 30 EU languages (Judge et al., 2012)

made available (Evas, 2013). Welsh, another Celtic minority language, is just as under-resourced and falls into the same category as Irish (weak/ no support) in both instances.

In light of the status of Irish language technology, the significance of the research and development involved in this project is:

- Development of linguistic resources as a platform for future work in Irish NLP
- Contribution to methods for constructing linguistic resources for low-density languages
- Contribution to research in dependency parsing for verb-initial languages

1.2 Dependency Treebanking

A major part of our work in this thesis is the development of a dependency treebank. A treebank is a large collection of texts (a corpus) that has been parsed at the sentence level. In other words, its sentences have been annotated with information regarding its syntactic structure. Treebanks are invaluable resources for the development of NLP applications. They provide a rich representation of linguistic phenomena of a language and are a solid platform for linguistic analysis. Traditionally treebanks have played an important role in the discipline of corpus linguistics. More recently however, successful implementation of machine learning approaches to NLP have resulted in exploitation of parsed corpora for the development of parsers. To put their value into context, the Penn Treebank (Marcus et al., 1993) is one of the most influential pieces of work in NLP, with more than 5,000 citations to date.

Traditionally treebanks were created using Phrase Structure Grammars (PSG), often known as Context Free Grammars (CFG), that identified phrases in text at a sentence level to produce phrase structure trees or constituents. In more recent years however, there has been much interest in the application of 'dependency' grammars to the development of treebanks: Arabic Prague Dependency Treebank (Hajič and Zemánek, 2004), Czech Prague Dependency Treebank (Hajič, 1998), Turkish Dependency Treebank (Oflazer et al., 2003), Slovene Dependency Treebank (Dzeroski et al., 2006), Finnish (Haverinen et al., 2010), Greek (Prokopidis et al., 2005) and Danish Dependency Treebank (Kromann, 2003) for example.

Dependency grammars use links to define the relationship between words. The relationship exists only between two words and is identified by labelling one word as the head and the other as its dependent. The link between both words is referred to as the dependency relation. Dependency grammars vary depending on the relationships that are defined within the language. In this project, we have defined and implemented Lexical Functional Grammar (LFG) inspired dependencies for the Irish language upon which the treebank was built. As a verb initial language, with a Verb-Subject-Object word order, Irish diverges greatly from the structure of most languages of current research in the domain of treebank and parser development. Our work involved an extensive syntactic analysis of the Irish language, along with the design of a dependency annotation scheme suited to Irish syntax.

In the context of treebanking for other languages, we also report on work we carried out in a Universal Dependencies project, in which treebanks for several languages (including Irish) were mapped to a universal annotation scheme for the purposes of cross-lingual studies.

1.3 Dependency Parsing

Parsing is the process of analysing text in order to identify its grammatical structure. Language parsing has changed greatly over recent years, profiting from the robustness of statistical methods. Traditionally, parsing was achieved by linguists who defined grammar rules and implemented them computationally through what is termed a rule-based parser. While some rule-based systems provide high-quality output, they are both time and language resource intensive. Current research focuses on the use of statistical models in the development of parsers. Statistical parsing models are developed using training data. This data consists of large collections of previously parsed text for the language in question, i.e. a treebank. Statistical parsers use this previously parsed text to assign probabilities to parse candidates of a sentence. The most likely parse is then identified.

In recent years, some progress has been made in the collection and development of syntactic linguistic resources for Irish. A part-of-speech (POS) tagger was developed by Uí Dhonnchadha (2009) and used to create a gold-standard POS-tagged corpus, which proved to be a fundamental resource for our treebank. Uí Dhonnchadha also reports on some initial attempts at parsing, and while the design of this constraint grammar partial parser was useful as a starting point for our own linguistic analysis, the output of this shallow parser was not sufficient to either (i) build a treebank or (ii) to allow for extraction of deep syntactic information for other NLP tools, such as hybrid machine translation or grammatical error detection systems, for example. Therefore, the need for the development of a full, deep parser for Irish was clear.

This thesis reports on the training and testing of the first statistical dependency parsing models for Irish. We used the treebank highlighted in Section 1.2 as our training and test data. We have developed parsing models specifically for dependency parsing, which identify relationships between words such as 'subject of', 'modifier of', and so on.

There are several frameworks available within which dependency parsers can be trained. These parsing platforms are data driven and rely on large collections of training data in the form of dependency treebanks. Frameworks such as MaltParser (Nivre et al., 2006), which we use in most of our experiments reported in this thesis, combine parsing algorithms, machine learning algorithms and feature models to induce dependency parsers. Parsing algorithms are used to build dependency graphs from text. Feature models combine features (such as parts-of speech, dependency relations) from the input string with features of the treebank to assist in parse decisions. The data-driven nature of these parsers means that they are language independent. With sufficient resources, parsers can therefore be developed for any language.

In our work, we explored various approaches to parsing with the Irish Dependency Treebank (IDT). We showed how parsing models built in the early days of the treebank's development could help bootstrap further development of the corpus by semi-automating the annotation process with Passive and Active Learning. If the parser can be improved, ultimately it would result in a smaller annotation effort within a bootstrapping (semi-automated) annotation environment. This leads us to consider improving the parser's accuracy by leveraging existing un-annotated resources through means of semi-supervised parsing. In addition, we show how we explored the possibility of cross-lingual transfer parsing through mapping the treebank to a universal annotation scheme. We take all of these approaches as methods for overcoming the limited resources of a minority language. Yet, we show that the semi-supervised and cross-lingual transfer parsing approaches do not prove to be promising solutions for bootstrapping the Irish parser. In particular, we note from a cross-lingual perspective, that, similar to previous studies McDonald et al. (2013), transfer parsing results are relatively low. We see that, as Irish is so typographically different from the languages for which there are large treebanks, it is an unlikely solution to our parser development. A clear conclusion we can draw from these unremarkable bootstrapping parsing results, is that it is difficult to replace the efforts of a human annotator in the context of treebank development.

1.4 Expanding Irish NLP to Social Media

To date, all of the linguistic resources and NLP tools developed for Irish have been directed at processing standard and grammatical Irish text. However, in recent years there has been a notable increase in the Irish language content available online. Much of the content made available on social media such as Facebook and Twitter is user-generated, i.e. not edited or reviewed for grammatical accuracy. Previous POS tagging and parsing studies on user-generated unstructured English text (e.g. Foster et al. (2011)) has shown how this language variety affects the quality of standard NLP tools. For the same reasons, existing NLP tools are not equipped to deal with processing this variation of Irish language. Rule-based systems are tied to strict grammatical maxims and statistically-driven tools can only successfully process language similar to that seen in training data, which until now has been well-structured Irish text. We extend our contribution of Irish NLP resources to this social media domain by building a gold-standard POS-tagged corpus of Irish tweets. With this corpus, we successfully train statistical POS taggers, achieving state-of-the art results and show how they outperform a rule-based tagger that has been designed for grammatical text. Through this we demonstrate why domain adaptation is necessary for these kinds of tools when processing noisy user-generated Irish text.

1.5 Research Questions

The primary contribution of this thesis is the Irish Dependency Treebank. The work involved in the construction of such a resource related to many aspects of development. Firstly, it included extensive exploration of theoretical linguistic issues in order to establish a suitable syntactic analysis. From there further research was involved in deciding on the most suitable formalism to draw on, and how this formalism should be adapted for Irish. It was also important to choose an appropriate set of dependency relations for our annotation scheme, ensuring that the set was comprehensive enough to cover all linguistic constructions we had encountered in our Irish data. The manual annotation effort required also played a significant part in this project, particularly in the early stages where the annotation guide was being developed concurrently as new linguistic phenomena were encountered.

However, in addition to developing this resource, we also seek to answer the following research questions:

- What is an appropriate linguistic analysis of Irish for a dependency treebank, drawing on and synthesising traditional descriptive analyses and theoretical work?
- Can an approach such as Active Learning, that has been suggested to be applicable to bootstrapping the development of treebanks, prove to be useful when deployed in the actual construction of a treebank?
- Given the existence of proposed techniques for development of parsers for low-resource language or improving performance of such a parser – cross-lingual transfer parsing and the use of unlabelled data – can these help when combined with a small gold standard treebank used for training?
- In what way can we leverage existing Irish NLP tools for processing Irish tweets?

1.6 Thesis Structure

This thesis comprises eight chapters including the current introductory chapter. Chapter 2 provides a summary of the status of the Irish language and an overview of its distinctive linguistic features. We also provide an overview of dependency syntax, which is the grammatical analysis we used in the creation of our treebank, along with an explanation of dependency treebank data formats. In Chapter 3, we summarise the processes involved in developing the treebank. With regards to our choice of syntactic representation, we discuss why we chose dependency syntax over constituency syntax. We also discuss our choice of labelling scheme, in addition to the existing resources upon which we built the treebank. This chapter also reports our inter-annotator agreement study. In Chapter 4, we introduce the treebank's annotation scheme, and detail our choice of syntactic analysis for certain linguistic phenomenon that are specific to the Irish language. In Chapter 5, we present our work on mapping the Irish Dependency Treebank annotation scheme to a universal annotation scheme for the purposes of cross-lingual studies. All of our parsing experiments are presented in Chapter 6. The experiments include establishing baseline scores, post IAA-study experiments, Active Learning experiments, semi-supervised parsing experiments and cross-lingual transfer parsing experiments. Finally, in Chapter 7, we report on how we extend our development of NLP resources to the social media domain. We present a gold standard POS-tagged Irish Twitter corpus and report on building statistical POS tagger models. This work also gives an overview of the differences involved with processing user-generated Irish text when compared with standard grammatical Irish text, and the impact this has on the development of NLP tools.

The Annotation Guidelines for the Irish Dependency Treebank are provided in Appendix A and the Annotation Guidelines for Irish Twitter Part-of-Speech Tagging are provided in Appendix B. Finally, some statistics on the content of the IDT are presented in Appendix C.

1.7 Publications

During the course of the treebank and parser development, much of our work was published at NLP workshops and conferences. We list them in chronological order here:

- Lynn et al. (2012a) describes the methodology behind building a treebank and the steps we took to leverage existing resources. We also report our baseline parsing scores.

  Lynn, T., Çetinoğlu, Ö., Foster, J., Dhonnchadha, E. U., Dras, M., and van Genabith, J. (2012a). Irish treebanking and parsing: A preliminary evaluation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 1939–1946, Istanbul, Turkey.

- Lynn et al. (2012b) reports our Inter-annotator agreement study and how it was used to improve our labelling scheme. As the treebank had grown in size at this stage, we reported on our efforts to use Active Learning as a bootstrapping mechanism.

  Lynn, T., Foster, J., Dras, M., and Dhonnchadha, E. U. (2012b). Active learning and the Irish treebank. In Proceedings of the Australasian Language Technology Workshop (ALTA), pages 23–32, Dunedin, New Zealand

- Lynn et al. (2013) reports our semi-supervised experiments and the various ways we tried to improve parsing accuracy with unlabelled data.

  Lynn, T., Foster, J., and Dras, M. (2013). Working with a small dataset – semi-supervised dependency parsing for Irish. In Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 1–11, Seattle, Washington, USA.

- Lynn et al. (2014) reports our cross-lingual transfer parsing and universal treebank studies.

Lynn, T., Foster, J., Dras, M., and Tounsi, L. (2014). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In Proceedings of the First Celtic Language Technology Workshop, pages 41–49, Dublin, Ireland.

- Nivre et al. (2015) and Agić et al. (2015) report respectively on the first and second release of the Universal Dependency project. The Irish Dependency Treebank is mapped to this UD scheme, and this Irish data and documentation is included in these releases.

  Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., **Lynn, T.**, Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal Dependencies 1.0.

  Agić, Ž., Aranzabe, M. J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A. T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., **Lynn, T.**, Manning, C., Martínez, H. A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.1.

- Lynn et al. (2015) reports on our domain-adaptation approach to building a POS-tagged corpus of Irish tweets and training POS-tagging models for use with unstructured, user-generated Irish text.

  Lynn, T., Scannell, K., and Maguire, E. (2015). Minority Language Twitter: Part-of- Speech Tagging and Analysis of Irish Tweets. In Proceedings of the 1st Workshop on Noisy User-generated Text (W-NUT 2015), Beijing, China.

Chapter 2

Background

Our work deals with the development of resources for a low-resourced, minority language. In Section 2.1, we give an overview of the status of the Irish language and explain how, as a national language, it came to be a minority and low-resourced language. We also provide an overview of some linguistic features of Irish in Section 2.2. Firstly, we describe some of the intricacies of Irish syntax and how there is still much disagreement regarding Irish structures in discussions of theoretical syntax. We highlight these issues as factors that have consequently impacted our linguistic analysis of the language, thus influencing the design of our treebank's annotation scheme. We also provide an overview of Irish morphology and describe the various inflectional processes that occur.

As we will see in Chapter 3, our treebank is based on dependency syntax analyses. In order to fully explain dependency structures, in Section 2.3, we provide an overview of dependency treebanks and dependency syntax. We show how a dependency syntax analysis is represented firstly as dependency graphs and secondly in machine readable formats that are acceptable training and evaluation input for data-driven parsers.

2.1 The Irish Language

The history of the Irish language is covered in some depth in sources such as Mac Giolla Chríost (2013) and Doyle (2015). Here we give an overview leading up to the status of the language today.

Irish is one of the oldest written forms of language in Europe. A Celtic language, it evolved from Old Irish (roughly 600-900 AD) to Middle Irish (900AD - 1200AD) to Early Modern Irish (1200–1600) to our present form of Modern Irish. There are three distinct Modern Irish dialects across Ireland - Ulster, Connacht and Munster. In the 1950's and 1960's an Official Standard (An Caighdeán Oifigiúil) was introduced to standardise spelling (diverging from the old written script) and vocabulary in order to make Irish more accessible through the education system to all.

Historically, Irish has experienced centuries of both decline and revival. Much decline was seen during the English colonialism era as Ireland was slowly becoming more Anglicised. In the 14th century, under Anglo-Norman Rule, the English monarchy forbade the Anglo-Normans from speaking Irish, which saw more widespread use and influence of the English language. However, by the start of the 16th century, the Anglo-Normans began to speak Irish and the percentage of Irish speakers rose again, with the language even adopting some English vocabulary. In the 17th century, during the time of the plantations (when Irish land was taken from native Irish and given to English and Scottish planters), the English language influence began to spread once more. Another sharp decline occurred in the early and mid 19th century, which was brought about by several factors. The secondary school system began to be modelled on the British education system with the English language at the fore. National schools did not even have Irish on the syllabus. Many of these schools forbade students from speaking Irish. More significantly, Irish people needed to learn English to be equipped to deal with legal matters, as English had become the administrative language in Ireland. The Catholic Church began to use English instead of Irish, and the Great Hunger resulted in both the death and emigration of many of the country's Irish speakers. By this stage, Irish had become mostly a spoken language in rural regions. In light of this, by the end of the 19th century, a movement known as the Gaelic Revival took strides towards reversing this decline, particularly in terms of the written language and literature. In particular, it saw the foundation of the Gaelic League (Conradh na Gaeilge) as a promotor of the Irish language, which is still in existence today. Following the establishment of Ireland as a Free State in the 1920's, Irish was re-introduced as a compulsory subject at state-run schools.

These days, Irish, a minority EU language, is the first official and national language of the Republic of Ireland. English is the second official language. Yet, in the 2011 census, only 41% of the population are reported as Irish speakers. However, the figures for those who actually speak and use Irish on a daily basis as a first language are even lower. Despite Irish being a core subject in the Irish primary and secondary education curricula, following graduation, the majority of Irish people do not continue to speak Irish in their daily lives.

The use of Irish in daily activities as a first language is mostly restricted to the *Gaeltacht* areas, which are predominantly in the West of Ireland. There are three categories of Gaeltacht areas as defined by the percentage of daily speakers, the environment for use (family, community, schools, linguistic networks) and age-cohorts of Irish speakers (Giollagáin et al., 2007). These Gaeltacht areas are protected by the State, with the government Department for Arts, Heritage and the Gaeltacht (DAHG) taking responsibility for ensuring that sufficient services and provisions are provided in order to allow people to use Irish as their first language, and to ensure that Irish continues as their primary medium of communication. In recent years, concerns have grown for the continued survival of these Gaeltacht areas as reports and studies show that younger generations in these communities are using English more frequently as their medium of communication outside of family life (Giollagáin et al., 2007).

Outside of the main Gaeltacht areas, frequent use of Irish is restricted to small linguistic networks in urban areas, and also to those students studying Irish at school. Thus, Irish is listed as a UNESCO endangered language and an increased effort has been made by the Irish State to ensure its survival. There are four Irish language radio stations (Raidió na Gaeltachta, Raidió na Life, Raidió Rí-Rá and Raidió Fáilte) and a dedicated Irish language TV channel (TG4) to cater for those who use Irish as their first language, and also to help promote the use of Irish language for those who are keen to continue using it as a second language. A number of printed Irish language publications such as Foinse and Nós also contribute to Irish language media. There are also *gaelscoileanna* schools nationwide through which all subjects are taught in Irish.

In recent years, there has been a resurgence of Irish language use outside of Gaeltacht areas, particularly among a younger generation of speakers, and online. The internet has become a new haven for languages resources such as Irish language blogs, e-zines, news articles and forums. In particular, social media platforms such as Facebook and Twitter have allowed Irish-speaking networks to grow, connecting speakers from both across the country and across the world. The Indigenous Tweets project reports that there have been over one million Irish tweets since Twitter began.[1] These tweets have been sent by over 8,000 tweeters worldwide, including countries such as the U.S., Canada, Australia, Brazil, Spain, Norway and Finland. In fact, there is also a growing interest in the use of Irish outside of Ireland with the formation of overseas linguistic groups. DAHG provides funding to these overseas organisations and directly assists funding of the teaching of Irish language at U.S. universities through the Fulbright Commission of Ireland.

In light of reports in a recent EU white paper on the status of Irish language technology (Judge et al., 2012), there has been an increased awareness of the need for technology in efforts to ensure the survival of the language. The DAHG have recognised the role that Machine Translation can play in meeting the State requirement that official documents need to be provided in both English and Irish (Official

[1] `http://indigenoustweets.com/ga/` (accessed June 2015)

Languages Act 2003). The ABAIR project has seen the development of a speech synthesiser which can assist language learners through CALL applications.[2] The development of web-based termbases, electronic dictionaries, grammar checkers, a part-of-speech tagged corpus and morphological analyser have also contributed to this initiative. However, relatively speaking and at an EU level, Irish is regarded as a low-resourced language with much more work and investment needed the area of speech and language technology. It is worth noting that while Irish is a minority language, compared to many other languages in the world, it has a relatively larger speaking population which would support opportunities to build and use language technology resources.[3]

2.2 An Overview of Irish linguistics

The Irish language is a Celtic language of the Indo-European language family. There are six Celtic languages: Irish, Scots Gaelic, Manx, Welsh, Cornish and Breton. In linguistic terms, Irish, Scots Gaelic and Manx are most closely related. In fact, Scots Gaelic originated through migration from Ireland to Scotland before the 5th century AD. Cornish and Manx are spoken only by small communities of language revivalists. Irish shares distinctive features with other Celtic languages such as Verb-Subject-Object (VSO) word order and rich morphology (Stenson, 1981). Both of these distinct linguistic features strongly influenced our treebank and parser development.

The development of our treebank and parser, as linguistic NLP resources for Irish, relied heavily on existing work in theoretical linguistics (where available) and previously developed NLP resources. Much of the groundwork for the development of the treebank involved an analysis of the syntax of Irish. In terms of literature on this topic, very early discussions of Irish linguistic studies appeared during the Gaelic Revival (e.g. Craig (1897)), but the most concise description of Irish grammar was not published until the 1960's – originally in Irish (Christian-Brothers, 1960)

[2] Accessible at `www.abair.tcd.ie`

[3] `http://www.ethnologue.com/cloud/gle` (accessed June 2015)

and later in English (Christian-Brothers, 1962; The Christian Brothers, 1988). Since then there have been several syntax and grammar books published. Some examples that this thesis in particular draws on are McCloskey (1979), Stenson (1981), The Christian Brothers (1988) and Ó Siadhail (1989), articles such as Ahlqvist (1972), McCloskey (1983), and Carnie (2005), along with paper publications (e.g. Sulger (2009b) and Asudeh (2002)) and Sulger (2009a)'s Masters thesis. Of these resources, we found that The Christian Brothers (1988) and Stenson (1981) provided the most comprehensive description of Irish. All other materials focused on analyses or discussions of only specific linguistic features of the language.[4] We found however, that within the limited scope of previous work in this field, there is still significant disagreement as to how the structure of Irish should be analysed.

The following provides an overview of the type of syntactic structures and morphology present in the Irish language. It is not a comprehensive summary as the rules governing both syntax and morphological changes are too extensive and at times too complex to document here. Instead, our summary aims to outline (i) some of the aspects (and issues) of Irish syntax that affected the development of our treebank in Section 2.2.1 and (ii) the many inflectional processes in Irish that impacted the development of our parser in Section 2.2.2.

2.2.1 Irish Syntax

Comparative studies have been made across the Celtic languages, and there exists much disagreement in theoretical syntax as to the nature of these VSO languages. Here we summarise some of the distinctive features of Irish as a Celtic language. These features commonly occur in standard Irish use and therefore require discussion in the context of treebank development. As we have noted in Section 2.2, Irish theoretical syntax is relatively under-researched, yet this summary shows that even within the limited work carried out in this area thus far, there still remain many

[4] While Ó Siadhail (1989)'s coverage of the language is quite broad, it focusses mainly on the dialectal variations of Irish.

unresolved disagreements as we show here.

VSO clause structure Both main clauses and subordinate clauses follow a VSO structure in Irish. There are only a couple of exceptional circumstances under which an element can appear between the verb and the subject (see Example 1) and while various elements may occur between the subject and object, such as prepositional phrases and adverbs (see Example 2), the verb-subject-object order is strict (McCloskey, 1983, pp. 10-11).

(1) Tá ar ndóigh daoine a chreideann...
Are of-course people who believe...
'There are of course people who believe...'

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

VSO languages pose problems for theoretical grammar frameworks that assume a VP (verb phrase) constituent across all languages. VPs typically contain a verb followed by the object NP. However, in the case of Irish, the subject NP always directly follows the verb. Arising from this, much of the discussion and disagreement in Irish syntax literature revolves around the structure of an Irish sentence – whether it has a flat structure such as that in Figure 2.1 as argued by Stenson (1981, p.41) and employed in Lexical Functional Grammar (Bresnan, 2001, p.399), for example, or derived from an underlying SVO structure such as that in Figure 2.2, as argued by Bobaljik and Carnie (1996), for example. Even within the pro-SVO camp and those influenced by Chomskyan transformational grammar, there is further disagreement as to the transformational processes involved to realise the VSO surface structure.

Particles There are a number of different particles in Irish including verb particles, adverbial particles, numerical particles, quantifier particles, complementisers and relative particles, among others. The *a* particle that occurs in Example 3 is regarded

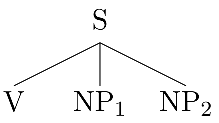


Figure 2.1: Flat VSO Constituent Structure

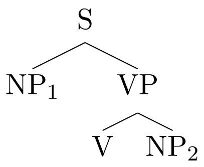


Figure 2.2: Hierarchical SVO Constituent Structure

as a relativiser by Stenson (1981, p.32). However it is contested by McCloskey (1979, pg.11) that it is in fact a complementiser as it has the same phonetic and morphological realisations on a following verb as the complementiser *go.* Sells (1984) argues further that this is a preverbal particle marking resumptive pronouns.

(3) an fear **a** thug an litir duit
the man REL gave the letter to-you
‘the man who gave the letter to you’

Copula There are two verbs ‘to be’ in Irish – the copula *is* and the substantive verb ‘*bí*’. Stenson (1981, p.92) and others consider the copula to be a verb form, although it does not inflect for mood, gender and number in the same way other verbs do. On the other hand, Ahlqvist (1972, p.271) argues that the copula is a particle that gives a verbal nature to its predicate, and Carnie (1997) claims it to be a complementizer particle which equates two noun phrases, without labelling either element as a subject or predicate.

Non-finite phrases There is no infinitive form in Irish (The Christian Brothers, 1988, p.128). Instead, a verbal noun is used to denote non-finite phrases (e.g. *léamh/ a léamh* ‘to read’). Verbal nouns in Irish are said to clearly originate from verbs, and are used in cases similar to the English gerund, yet they exhibit the same morphological inflection as nouns (Stenson, 1981, p.29). When an infinitive phrase expresses an object, that object precedes the verbal noun and is accompanied by the particle *a* (see Example 4).

(4) leabhar **a** léamh

book INF read

'to read a book'

The Christian Brothers (1988, p.128) refers to this *a* as a preposition and Stenson (1981, p.32) as a particle. The disagreement over the labelling of this element aside, there is much debate on the positioning of the object to the left of the verbal noun in non-finite clauses, and how it came to be there (e.g. Chung and McCloskey (1987) vs Bobaljik and Carnie (1996)).

Progressive Aspectual Phrases There is no present participle in Irish (The Christian Brothers, 1988, p.128). Instead, *ag* + verbal noun construction is used to indicate progressive aspectuals (e.g. *ag léamh* 'reading'). However, there is disagreement as to what *ag* really is. It resembles the preposition *ag* 'at' and is regarded as that by Stenson (1981, p.139), recognising the verbal noun as a prepositional phrase object. Yet, McCloskey (1983) rejects this in favour of a 'progressive particle' analysis.

In terms of developing the treebank, we – like Uí Dhonnchadha (2009) – have tended to adopt the syntactic analysis of Stenson in general. We explain this in more detail in Section 4, but in summary, Stenson's flat structure analysis allows us to make a minimum of theoretical assumptions as to underlying structure; this philosophy aligns with the framework we draw on for our analysis, discussed in Section 3. Our treebank annotation focuses mainly on the functional roles of words in the Irish clause, such as subject or object, rather than the elements of constituents. Thus, we are more concerned with identifying relationships between words and ensuring that we are consistent in whatever labelling we choose to apply.

2.2.2 Irish Morphology

Inflection in Irish mainly occurs through suffixation, but initial mutation through lenition and eclipsis is also common (The Christian Brothers, 1988). Lenition is a phonological change that softens or weakens the articulation of a consonant. The eclipsis process renders voiced segments as nasalised and voiceless segments as being voiced (Stenson, 1981, p.18). A prominent feature of Irish (also of Scottish Gaelic and Manx), which influences inflection, is the existence of two sets of consonants, referred to as 'broad' and 'slender' consonants (Ó Siadhail, 1989, p.9). Consonants can be slenderised by accompanying the consonant with a slender vowel, either *e* or *i*. Broadening occurs through the use of broad vowels; *a*, *o* or *u*. In general, there needs to be vowel harmony (slender or broad) between stem endings and the initial vowel in a suffix. See Example 5 for the effect of this process on two forms of the verb *buail*. The verbal noun ending for *buail* is -*adh*. The stem is therefore broadened to *bual* to ensure vowel harmony. In contrast, the impersonal form has a slender suffix -*eadh*, which is in harmony with the slender form of the stem.

(5) *buail* 'hit' → *ag bu**a**l**a**dh na liathróide* 'hitting the ball' (Verbal Noun)
buail 'hit' → *bua**ile**adh an liathróid* 'the ball was hit' (Impersonal Form)

A process known as syncopation also occurs when words with more than one syllable have a vowel-initial suffix added. See Example 6 where the verbs *imir* and *labhair* drop the vowel(s) in the second syllable before adding a suffix.

(6) *imir* 'to play' → ***imr**ím* 'I play'
labhair 'to speak' → ***labhr**aím* 'I speak'

Verbs Verbs inflect for number and person, as well as mood and tense. Verbs can incorporate their subject, inflecting for person and number through suffixation. Such forms are referred to as synthetic verb forms. Most verbs tend to incorporate a subject when it is first person singular or plural. These synthetic forms are generally restricted to the Present Tense, Imperfect Tense, Conditional Mood and Imperative

Mood. See Example 7. In addition, verb tense is often indicated through various combinations of initial mutation, syncopation and suffixation.

(7) *scríobh* 'write'

*scríobh**aim*** 'I write'

*scríobh**faimid*** 'we will write'

However, second person singular and plural subjects are incorporated in some verb tenses and moods as show in Example 8.

(8) *nigh* 'wash'

*ni**teá*** 'you used to wash'

*ní**gí**!* '(you pl.) wash!'

Tense is also marked by lenition on some verb forms, as per Example 9:

(9) *dún* 'close'

*d**h**ún mé* 'I closed'

*d**h**únfainn* 'I would close'

Lenition occurs after the negative particle *ní*, as per Example 10:

(10) *tugaim* 'I give'

*ní t**h**ugaim* 'I do not give'

tabharfaidh mé 'I will give'

*ní t**h**abharfaidh mé* 'I will not give'

Eclipsis (initial mutation) occurs following clitics such as interrogative particles (*an, nach*); complementisers (*go, nach*); and relativisers (*a, nach*) (Stenson, 1981, pp. 21-26). Following on from the usage of the verb *tabhair* in Example 10, we can see the effects of mutation in Example 11.

(11) *an **d**tugann sé?* 'does he give?'

*nach **d**tugann sé* 'that he does not give'.

*go **d**tabharfadh sé* 'that he would give'

Nouns While Old Irish employed several grammatical cases, Modern Irish uses only three: Nominative, Genitive and Vocative. The nominative form is sometimes regarded as the 'common form' as it is now also used for accusative and dative forms. Nouns in Irish are divided into five classes, or declensions, depending on the manner in which the genitive case is formed. In addition, there are two grammatical genders in Irish - masculine and feminine. Case, declension and gender are expressed through noun inflection. For example, *páipéar* 'paper' is a masculine noun in the first declension. Both lenition and slenderisation are used to form the genitive singular form: ***p**h**áipé**i**r***. Other examples of this kind of inflection are given in Example 12.

(12) *an dochtúir* 'the doctor'
*cóta an dochtúr**a*** 'the doctor's coat'

an fheoil 'the meat'
*boladh an feol**a*** 'the smell of the meat'

an coinín 'the rabbit'
*ainm an c**h**oinín* 'the rabbit's name'

an siopa 'the shop'
*cúl an **t**siopa* 'the back of the shop'

Máire 'Mary'
*a M**h**áire!* 'Mary!' (Vocative)

In addition, possessive determiners cause noun inflection through lenition, eclipsis and prefixation. See Example 13.

(13) *teach* 'house'
*mo t**h**each* 'my house'
*ár **d**teach* 'our house'

ainm 'name'
*a **h**ainm* 'her name'
*a **n**-ainm* 'their name'

Adjectives In general, adjectives follow nouns and agree in number, gender and case. Depending on the noun they modify, adjectives can also inflect. The Christian Brothers (1988, p.63) note eight main declensions of adjectives. They can decline for genitive singular masculine, genitive singular feminine and nominative plural as per Example 14.

(14) *bacach* 'lame'
*baca**igh*** (Gen.Sg.Masc)
*baca**í*** (Gen.Sg.Fem)
*bacach**a*** (Nom.PL).

Comparative adjectives are also formed through inflection as shown in Example 15.

(15) *láidir* 'strong', *níos láid**re*** 'stronger'
déanach 'late', *is déana**í*** 'latest'.

Prepositions Irish has simple prepositions (e.g. *ar* 'on') and compound prepositions (e.g. *in aghaidh* 'against'). Most of the simple prepositions can inflect for for a pronominal object that indicates person and number (known as prepositional pronouns or pronominal prepositions), thus including a nominal element. Compare *le* and *leis* in Example 16:

(16) *bhí sé ag labhairt **le** fear* 'he was speaking **with** a man'
*bhí sé ag labhairt **leis*** 'he was speaking **with him**'

These forms are used quite frequently, not only with regular prepositional attachment where pronominal prepositions operate as arguments of verbs or modifiers of nouns and verbs, but also in idiomatic use where they express emotions and states. See Example 17.

(17) *tá brón* ***orm*** 'I am sorry'

(lit. 'is sorrow on-me')

tá súil ***agam*** 'I hope'

(lit. 'is expectation with-me')

Irish has been described as a noun-centered language (Greene, 1966), and nouns are often used to convey the meaning that verbs would convey. Pronominal prepositions are often used in these types of structures, as shown in Example 18.

(18) *bhain mé geit* ***aisti*** 'I frightened her'

(lit. extracted I shock **from her**)

bhain mé mo chóta ***díom*** 'I took off my coat'

(lit. extracted I my coat **from me**)

bhain mé úsáid ***as*** 'I used it'

(lit. extracted I use **from it**)

bhain mé triail ***astu*** 'I tried them'

(lit. extracted I attempt **from them**)

Emphatics and diminutives An emphatic suffix *-sa/-se* (both broad and slender form) can attach to nouns or pronouns. It can also be attached to any verb that has been inflected for person and number, and also to pronominal prepositions. See Example 19.

(19) *mo thuairim* 'my opinion' → *mo thuairim****se*** '**my** opinion'

tú 'you'(sg) → *tu****sa*** '**you**'

cloisim 'I hear' → *cloisim****se*** '**I** hear'

liom 'with me' → *liom****sa*** 'with **me**'

In addition, the diminutive suffix *-ín* can attach to all nouns to form a derived diminutive form. The rules of slenderisation apply here also as shown in Example 20.

(20) *buachaill* 'boy' → *buacha**illín*** 'little boy'
tamall 'while' → *tama**illín*** 'short while'

The inflectional processes in Irish we describe here give rise to data sparsity in our treebank, which in turn affects the quality of the parser that is built on that treebank. This is a phenomenon that has been shown to occur for many highly inflected languages (Bohnet et al., 2013). In other words, due to the fact that there can be many forms of one word, if a specific form does not appear in the treebank (which is training data for the parser), then the parser is less likely to know how to process it. We take steps to overcome this through the inclusion of the lemma form (base form) of each word when training the parser. This is described in more detail in Section 6.

2.3 Dependency Treebanks

A treebank is a corpus of text that has been annotated with syntactic information describing the grammatical structure of each sentence. Primarily there are two kinds of approaches to syntactic analysis – constituency analysis, in which a sentence is divided up into hierarchical phrases or constituents, and dependency analysis, which is based on extracting sets of labelled relations between pairs of words in a sentence. As we will see in Chapter 3, the Irish Dependency Treebank's syntactic annotations are based on a dependency syntax annotation scheme.[5] Thus, in this section, we explain dependency syntax and show how this type of structural analysis can be represented in a treebank. We only attempt to give an overview here; for more detail we refer the reader to Kubler et al. (2009).

[5] We motivate this choice of representation in more detail in Section 3.1.1.

2.3.1 An Overview of Dependency Syntax

Dependency structure grammars grew from a theory of structural syntax which focused on connections and grammatical relations between words (Tesnière, 1959). The syntactic structure of a sentence is described through defining a set of binary relations between words. These binary relations are described as dependencies in which one word (*subordinate*) is a dependent of another (*governor*). Current work in this area refers to the pair as a *head-modifier* dependency relationship. Dependencies are regarded as syntactic representations that focus on the relationship between words in a sentence with regards to the functional role the words play. Dependency structures are often represented as *dependency graphs* as per the example given in Figure 2.3.

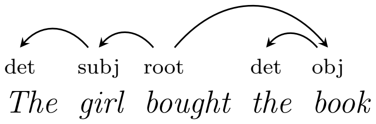


Figure 2.3: A typed dependency graph

We follow Kubler et al. (2009)'s definition of dependency graphs and dependency trees, and outline the main properties here:

Sentences : $S = w_0w_1.....w_n$

S denotes a sentence as a sequence of tokens (words). w_0 is an artificial ROOT word.

Nodes : $V \subseteq \{w_0, w_1, ...w_n\}$

V is a set of the nodes (tokens) in the sentence for use as vertices in the dependency graph to be defined.

Relations : $R = \{r_1, ..., r_m\}$ is a set of possible *dependency relation types* that can hold between two words in a sentence. $r \in R$ where r is a relation type (or arc label).

In the case of *typed dependencies*, the relationship is marked by a function label, or relation type that identifies the grammatical role of the dependent in relation to the head. For example, in the graph representation given in Figure 2.3, there are four head-modifier relationships: `det(girl,the)`, `subj(bought,girl)`, `det(book,the)`, `obj(bought,book)`. In other words, *the* is a determiner of *girl*, *girl* is the subject of *bought*, the second *the* is a determiner of *book* and *book* is the object of *bought*.

Arcs : $A \subseteq V \times R \times V$:

A is a set of arcs through which nodes (V) are connected, labelled with a relation from R. An arc $(w_i, r, w_j) \in A$ represents a dependency relation from head w_i to dependent w_j, labelled with the relation type r. The direction of the arcs is usually from the head to the dependent. Although, some representations choose the opposite direction.

Arc Restriction : If $(w_i, r, w_j) \in A$ then $(w_i, r\prime, w_j) \notin A$, where $w_i, w_j \in V$, $r, r' \in R$, and $r \neq r'$.

There can only be one dependency relation arc defined between two nodes.

Graphs : A *dependency graph* $G = (V, A)$

A dependency graph is a representation of a sentence, where each word in the sentence is a node (V), and directed edges, or arcs (A) link the nodes as head-modifier dependency relations.

Trees : A *dependency tree* is a well formed dependency graph that originates out of node w_0 and has a *spanning* node set $V = V_S$, which is the set that contains all the words of a sentence.

Dependency trees have an artificial ROOT word w_0 which is connected to the root node in the sentence (any one of the nodes in V). Note that the root node is labelled as `root` in some treebanks, and labelled as `top` in the IDT.

- The *spanning property* of a dependency tree requires each node to have some relevance to the dependency analysis of the entire sentence.

- A *dependency tree* $G = (V, A)$ satisfies the *single head property*, which states for all $w_i, w_j \in V$, if $(w_i, r, w_j) \in A$ then there does not exist $(w_i\prime, r\prime, w_j) \in A$

Dependency graphs are an instance of *directed acyclic graphs* (DAGs) as there is no way to follow from one token, through a sequence of edges, to loop back to that original token. DAGs can allow a word to have more than one head. However, this is not the case with *dependency trees*, as each word may have multiple dependents, but only one head.

The Irish Dependency Treebank contains dependency trees like the example given in Figure 2.4.

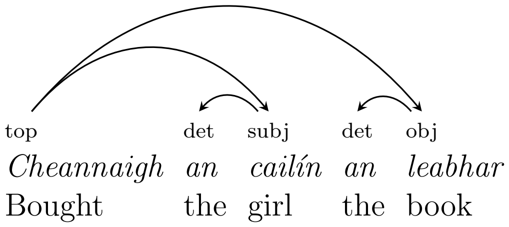


Figure 2.4: Dependency tree representation of *Cheannaigh an cailín an leabhar* 'the girl bought the book'

2.3.2 Projectivity

A dependency tree can be *projective* (which means that no arcs should cross) or *non-projective* (with crossing arcs). Non-projectivity is shown to affect many languages (Buchhloz and Marsi, 2006) and is required for capturing some languages, particularly those of freer word order. However, non-projectivity can cause problems for some parsing algorithms. Some systems such as MaltParser take a pre-processing step to use graph transformation techniques to produce non-projective structures before parsing (Nivre and Nilsson, 2005). Others (e.g. MSTParser (McDonald et al., 2005)) can easily handle non-projective trees.

In Irish, most sentences produce projective trees, yet non-projective structures can occur. For example, non-projectivity can arise due to the relative freedom of the positioning of oblique prepositional phrases. See Figure 2.5 for example. The oblique argument *acu* 'at-them' is placed within the noun phrase *suim sa Ghaeilge* 'interest in Irish'.

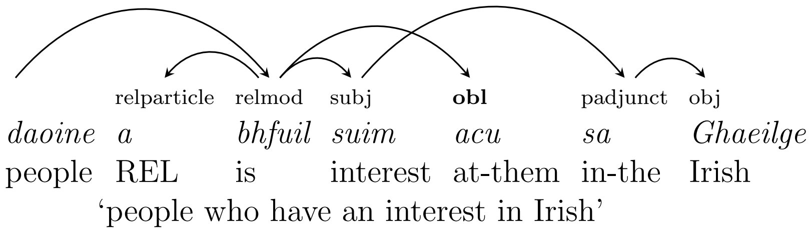

relparticle
relmod
subj
obl
padjunct
obj
daoine a bhfuil suim acu sa Ghaeilge
people REL is interest at-them in-the Irish
'people who have an interest in Irish'

Figure 2.5: Example of non-projective tree for an Irish sentence

2.3.3 Dependency Tree Data Formats

Parsing systems such as MaltParser (Nivre et al., 2006) and MSTParser (McDonald et al., 2005) are language-independent systems that allow users to build parsing models using their own choice of treebank. In order to train and test a parsing system such as MaltParser, for example, the dependency tree structure needs to be encoded in an easily-read format. The information available to the parser normally contains: each token, its index in the sentence, its lemma, its part-of-speech tag, (optional morphological data), the index of the head (where it attaches) and the description of that attachment (dependency label). A statistical data-driven parser can use this data as *features* when learning patterns in the training data, which we discuss in more detail in Section 6.1.1.

There is a range of different formats used for various parsers, including: XML as shown in Example 21, a space separated format where a sentence spans 3-4 lines of meta-data like the MSTParser input in Example 22, or a tab-delimited format such as the CoNLL-X format where each token in a sentence is on a new line, each line containing part-of-speech information, dependency label and attachment information, see Example 23.

(21)
```
<sentence id=“2” user=“malt” date=“”>
<word id=“1” form=“Cheannaigh” postag=”VERB” head=“0” deprel=“root”/ >
<word id=“2” form=“an” postag=“ART” head=“3” deprel=“det”/ >
<word id=“3” form=“cailín” postag=“NOUN” head=“1” deprel=“subj”/ >
<word id=“4” form=“an” postag=“ART” head=“5” deprel=“det”/ >
<word id=“5” form=“leabhar” postag=“NOUN” head=“1” deprel=“obj”/ >
</sentence>
```

(22)
```
Cheannaigh an  cailín   an leabhar
V          ART    NOUN  ART    NOUN
root       det   subject det   object
0            3      1    5       1
```

(23)
```
1 Cheannaigh Ceannaigh Verb VI _ 0 root _ _
2 an an Art Art _ 3 det _ _
3 cailín cailín Noun Noun _ 1 subj _ _
4 an an Art Art _ 5 det _ _
5 leabhar leabhar Noun Noun _ 1 obj _ _
```

The CoNLL-X format was introduced for a multilingual dependency parsing Shared Task in the 2006 Conference on Natural Language Learning. This standard format was implemented by all users in establishing a benchmark for evaluating their parsers across multiple languages. Most statistical dependency parsers now accept the CoNLL-X format as input, and also allow for easy conversion from their default input format to this more widely used format.

The parse format of Example 23 can be interpreted as follows: Each word (token) is represented on one line.[6] The token *Cheannaigh* ‘bought’ is a Verb and is the `root` of the sentence. The value of the head of the root is always 0. The second token is the determiner *an*. Its head is the noun *cailín*, which is token #3. The relationship

[6] Each sentence in a text is separated by a blank line.

between these two tokens is det (determiner). The third token, *cailín*, is the subject (subj) modifier of the first token (#1), and so on.

The dependency label names used in these examples are taken from the dependency label tagset for the Irish Dependency Treebank and is discussed in more detail in Section 4. Label tagsets vary across labelling schemes and treebanks, but are usually intuitive descriptions of the type of relationship between a head and its modifier. The IDT is based on the CoNLL format, as per Example 23.

2.4 Summary and Conclusion

In this chapter, we have provided an overview of both the Irish language and dependency treebanks – two subjects that are central to this thesis.

In the Irish language overview, we explained the status of Irish, an official and national language, and how it came to be a minority language in Ireland, secondary to English. We also showed that its minority status has led to a lack of research in Irish syntax, along with unresolved theoretical issues in the literature. Our analysis of these linguistic features of the language played an important role in our treebank's development. We also provide an overview of the morphological complexity of the Irish language, a feature that impacts data sparsity in our treebank.

In the dependency treebank overview, we have explained dependency syntax and how a sentence analysed with a dependency grammar can be represented through dependency graphs. As this thesis also deals with the practical implementation of dependency syntax, we also highlighted some of the various data formats of dependency trees that can be processed by data-driven parsers.

In the next chapter, we introduce the Irish Dependency Treebank and the approach we took to its development.

Chapter 3

Irish Dependency Treebank

Treebanks provide a rich representation of linguistic phenomena of a language and are a solid platform for linguistic analysis. In corpus linguistics, linguists use treebanks to test linguistic theories and study syntactic structures. Treebanks are also invaluable resources for the development of NLP applications, specifically data-driven statistical parsers. These parsers learn patterns of syntactic structure through machine learning processes, as we will see later in Chapter 6.

In this chapter, we report on the development of the first treebank for Irish. We refer to the treebank as the Irish Dependency Treebank (IDT). The current status of the treebank is at 1,018 dependency trees[1] and 23,684 tokens. The average sentence length is 23 tokens.

In Section 3.1, we begin by reporting on the varieties of treebanks currently available, and the various factors that influence their design and contribute to their fundamental differences. These factors include syntactic representations, labelling schemes, and the types of resources used. We then step through how each of these main factors influenced the development of the IDT.

As human annotators were involved with the development of this treebank, in Section 3.2, we also report on two inter-annotator agreement studies, along with workshops that took place between those studies to discuss and finalise our linguistic

[1] Since the final parsing experiments, two sentences with only one token have been removed as they are not trees.

analysis of Irish. We show how this conference-in-progress approach to treebank development helped to improve our annotation guide and labelling scheme.[2]

3.1 Building the Irish Dependency Treebank — the starting point

Treebanks exist for many languages (e.g. German TIGER Treebank (Brants et al., 2002), Finnish (Haverinen et al., 2010), French (Abeillé et al., 2003)). Some languages have multiple treebanks that are based on varying annotation representations, linguistic formalisms and content. For example, the Penn-II Treebank (Marcus et al., 1994) contains (among others) a collection of articles from the Wall Street Journal (WSJ) that has been annotated according to a constituency grammar scheme. The same corpus has been converted to various other annotation representations including an LFG-inspired dependency grammar (Parc 700 Dependency Treebank (King et al., 2003)) which is based on a subsection of section 23 of the WSJ, and Combinatory Categorial Grammar (CCG Bank (Hockenmaier and Steedman, 2007)) based on the entire Penn Treebank. In addition, the LinGO Redwoods treebank, which is parsed according to the Head-driven Phrase Structure Grammar (HPSG) framework (Oepen et al., 2002), is a collection of text from varying domains and sources such as Wikipedia, e-commerce, the tourist domain and a semantically annotated subset of the Brown Corpus (Miller et al., 1993). More recently, there has been a notable increase in research on less formal online text, motivating the development of a corpus for web-based text only (The English Web Treebank).[3] This corpus has been annotated with constituency structures and has also been converted to a dependency structure representation (Silveira et al., 2014).

All of these treebanks vary according to different considerations that are taken

[2] Later, in Section 6.3, we also show how these updates to the labelling scheme and treebank also resulted in an increase in parsing accuracy.

[3] Linguistic Data Consortium release LDC2012T13, `https://catalog.ldc.upenn.edu/LDC2012T13`

during development, including:

- Type of Syntactic Representation

  There are a number of syntactic representations or grammar formalisms to choose from when designing a treebank, which are based on varying linguistic theories and formalisms. For example, a phrase structure grammar representation hierarchically denotes constituents and phrases within sentences, while a dependency grammar labels connections between words within a sentence according to their functional roles.

- Labelling Scheme

  Labelling schemes define how linguistic structures are represented and labelled in a treebank. They are often closely linked to the chosen syntactic representation or chosen formalism. They are also influenced greatly by specific linguistic phenomena in the language in question.

- Resources Used

  There are various ways of assisting treebank development by leveraging from existing NLP resources such as POS taggers, morphological analysers and existing corpora.

The following is a summary of the way in which these factors shaped the development of the Irish Dependency Treebank.

3.1.1 Choice of syntactic representation

Many of the earlier treebanks contain phrase structure constituent trees. In constituency structure representations, nodes can be either terminals (lexical items) or non-terminals (syntactic categories) as part of a hierarchical structure that identifies phrases or constituents. The constituent tree in Figure 3.1 is an example of parsed text from the Wall Street Journal section of the Penn-II Treebank (Marcus et al., 1994). The sentence "Others have tried to spruce up frequent-flier programs" has been fully parsed to not only indicate that the sentence (S) consists of an noun phrase (NP) and a verb phrase (VP), but also to show the subconstituents which make up these constituents.

While modern dependency grammars date back to the late 1950's and were even adopted in the early days of computational linguistics (Hays, 1964), constituency grammars, however, continued to dominate the field of linguistics and parsing for a significant period of time. Mel'čuk (1988) proposed a number of reasons for this, in particular the fact that during that time, work on modern syntax and thus constituency grammars was led by English speaking linguists (e.g. Noam Chomsky) and that English was their main source of research data. Word order is central to constituency structures and is congruous to the strict word order of English. He also argued that there was a lack of interest in semantics, which constituency grammar looks to almost as an afterthought, generating semantic structures from syntactic structures.

This constituency analysis contrasts with dependency graphs, where each node is a terminal, representing a word in the sentence. Figure 3.2 shows how a dependency graph represents dependency relations between words, as per the Stanford dependency scheme (de Marneffe and Manning, 2008). For example, *others* is the subject of *tried*, and *spruce* is the head of an open complement which is dependent on the matrix clause.

Dependency grammars have become more popular in the last twenty years and dependency treebanks have been developed for various languages including Czech (Hajič, 1998; Böhmová et al., 2001), Arabic (Hajič and Zemánek, 2004), Danish (Kromann, 2003) and Turkish (Oflazer et al., 2003), to name a few.

We have chosen to build a dependency rather than constituency treebank for Irish. We outline our reasons for this here:

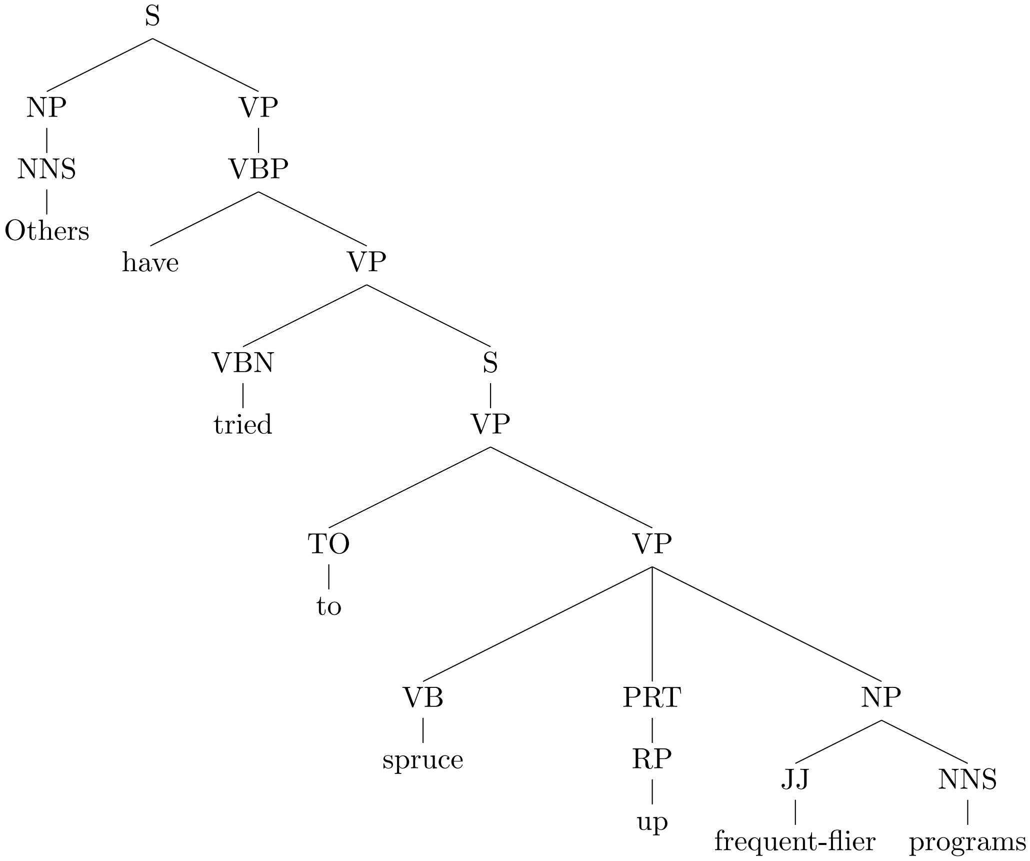


Figure 3.1: Constituency representation of parsed text

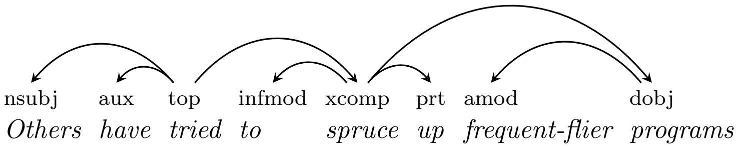


Figure 3.2: Stanford dependency representation of parsed text

Better handling of freer word order The increase in popularity of dependency grammars follows from the same arguments that linguists like Tesnière (1959) and Mel’čuk (1988) originally proposed, and were later reinforced in the early days of dependency parsing by Covington (1990). That is, that dependency grammars are less anglo-centric. They are language independent and especially more suited to languages with a freer word order. Dependency nodes represent each word in a sentence. These nodes are terminals and do not require the abstract representations such as NPs (noun phrases) or InflPs (inflection phrases) that make up additional non-terminal nodes required by constituency structures.

Node simplicty While Irish does not have a free word order, its under-researched VSO structure is more suited to a dependency analysis than a constituency analysis. As we have previously discussed in Section 2.2.1, an examination of the existing literature in Irish theoretical syntax (including, but not limited to McCloskey (1979); Stenson (1981) and Carnie (2005)), shows a lack of sufficient agreement on the syntactic representation of some fundamental linguistic phenomena. Their syntactic analyses differ even at a basic level, such as for example, the disagreement on whether Celtic languages have a flat VSO structure or an underlying SVO structure. In fact, among those who advocate an underlying SVO structure, McCloskey (1983), Doherty (1992) and Bobaljik and Carnie (1996), for example, disagree as to what type of movement or raising is involved to realise the VSO surface structure. Discussions like these on topics such as deep structure or movement (represented by traces), for example, would have been highly relevant for a constituency-based treebank constructed on Chomskyan principles. The question of the existence of a VP constituent would also strongly impact a constituency analysis. It is more feasible therefore, in our treebank development, to identify the functional relationships within sentences (dependencies) than to try to address all the unsolved complexities of Irish syntax, which would lie outside the scope of our work. As a dependency grammar is regarded as less restrictive than a phrase structure grammar, it is thus more conducive to parsing a language such as Irish with a such a divergent (VSO) word order.

In Section 2.3.2, we noted, for example, how it is possible in Irish to insert a prepositional phrase into a noun phrase, even though it is attached at a higher level. In constituency grammars, these type of instances are regarded as discontinuous constituents, and they pose difficulties for phrase structure parsing. In contrast, some dependency parsers, such as McDonald et al. (2005), for example, can easily handle these structures.

Clean mapping to semantic predicate-argument structure In addition, by using dependency grammars, it is easier to extract semantic information and details of functional roles within a grammatical structure. We are able to answer questions such as 'WHO did WHAT to WHOM?'. Parsers based on these representations play an important role in the development of applications such as Question-Answering systems (Shen, 2007; Verberne et al., 2008), Machine Translation (Och et al., 2004; Quirk et al., 2005; Xu et al., 2009; Cai et al., 2014), Quality Estimation for Machine Translation (Amigó et al., 2009; Kaljahi et al., 2013), Grammatical Error Detection (Tetreault et al., 2010), Educational Applications (Higgins et al., 2014), Discourse Analysis (Fisher and Roark, 2007), Text Summarisation (Bhaskar and Bandyopadhyay, 2010; Kikuchi et al., 2014), Information Retrieval (Carmel et al., 2014) and Sentiment Analysis systems (Johansson and Moschitti, 2010; Bakliwal et al., 2013). We therefore see the decision of building a dependency treebank for Irish as providing the option of making a more direct use of syntactic dependencies in future Irish NLP research. This is also aligned with the current momentum in the research community towards dependency parsing NLP application.

While this approach requires fewer theoretical assumptions about previously unseen structures, we were still left with a significantly challenging task, as we outline later in Section 4.

3.1.2 Choice of dependency labelling scheme

Dependency labelling schemes vary widely across treebanks. Nivre (2015) proposes that this large variation can be attributed to different theoretical preferences among treebank developers, or most particularly, as a result of descriptive grammatical traditions that have been established over time for specific languages. While the traditions can hold similarity across languages, there are often subtle differences in terminology and notation. In essence, there is no standard approach to designing a dependency labelling scheme for any one language.

We have based our dependency labelling scheme on that of Çetinoğlu et al.

(2010). This scheme was inspired by the functional relations defined within Lexical Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001), a theory that incorporates c(onstituent) and f(unctional) structures. We provide an example Irish sentence in Example 24. Figure 3.3 shows a c-structure representation of this. The sentence is broken into constituents in a hierarchical structure. Figure 3.4, on the other hand, shows an LFG f-structure of the same sentence that focuses on the grammatical functions within the sentence structure. This representation shows that the predicate (in this case the main verb) has two arguments – f_1 and f_2. These arguments co-reference the subject (`SUBJ`) and oblique (`OBL`) arguments, whose internal structures in turn describe their features and arguments.

(24) Chuir mé isteach ar an phost sin
Put me in on the job DEM[4]
'I applied for that job'

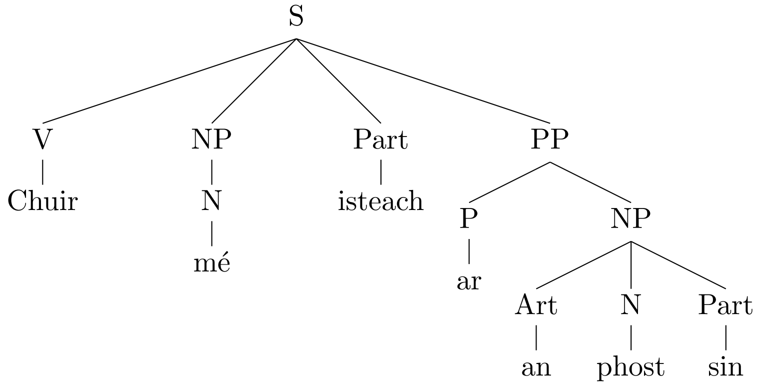


Figure 3.3: C-structure representation of *Chuir mé isteach ar an phost sin* 'I applied for that job'. (Bresnan, 2001)

Çetinoğlu et al. (2010) build upon an LFG f-structure Annotation Algorithm (AA) (Cahill et al., 2004, 2008) to create LFG-inspired dependencies. The AA essentially converts the output of a constituency parser (c-structures) into f-structures. For the purpose of comparative parsing experiments, Çetinoğlu et al. (2010) extend this work by creating an LFG dependency parsing pipeline. The pipeline takes the AA output (f-structures) and converts them into dependency trees in order to train a dependency parser. This conversion is not straightforward, however as there are

$$\left[\begin{array}{ll}
\text{PRED} & \text{`cur-isteach}\langle f_1,\, f_2\rangle\text{'} \\
\text{LOC} & + \\
\text{TENSE} & \text{PAST} \\
\text{SUBJ} & f_1\left[\begin{array}{ll}\text{PRED} & \text{`PRO'}\\ \text{NUM} & \text{SG}\\ \text{PERS} & 1\end{array}\right] \\
\text{OBL}_{AR} & f_2\left[\begin{array}{ll}\text{CASE} & \text{OBL}_{AR}\\ \text{OBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`post'}\\ \text{DEF} & +\\ \text{DEIXIS} & \text{DISTANT}\end{array}\right]\end{array}\right]
\end{array}\right]$$

Figure 3.4: LFG F-structure representation of *Chuir mé isteach ar an phost sin* 'I applied for that job' (adapted from Bresnan (2001))

fundamental differences between the nature of f-structures, which are represented by Directed Acyclic Graphs (DAGs) and bilexical labelled dependency trees.

These differences meant that the conversion involved multiple changes to the representation: (i) As we saw in Section 2.3.1, it is a requirement of dependency trees that each token is labelled. LFG f-structures can represent some words (e.g. verb particles) as features of predicates, and these needed to be extracted and represented at token level. (ii) The output of the AA can result in multiple heads for non-local dependencies (e.g. relative clauses). As dependency trees do not allow multiple-head dependencies, these are removed before being converted to trees. (iii) Similarly, dependency trees only allow a single root, whereas f-structures may have multiple roots. Pre-processing steps are also taken to address this.

The trees are created using a conversion dependency tagset of 25 labels. The tagset is based on functional labels from the LFG f-structures, along with newly introduced labels required for tokens that are not overtly represented in the abstract f-structures (e.g. `possmarker`: possessive marker 's). The conversion tagset is presented in Table 3.1. The labels marked with asterisks are dependencies that are not part of basic LFG theory but have been introduced for the generation of dependency trees.

LFG is relatively language-independent due to the abstract nature of the f-

adjunct	adjunct		poss	possessive
app	apposition		*possmarker	possessive marker 's
comp	complement		*prepositionhead	MWE's in LFG
coord	coordination item		*punctuation	punctuation
*dep	dependency (dummy)		quant	quantifier
det	determiner		relmod	relative modifier
focus	focus		subj	subject
obj	object		*toinfinitive	to infinitive
obj2	2nd object (obj-th)		*top	root of dependency tree
obl	oblique object		topic	topic
*obl2	2nd oblique object		topicrel	relative topic
obl-ag	oblique agent		xcomp	open complement
*particlehead	head of particle			

Table 3.1: LFG-inspired conversion tagset (Çetinoğlu et al., 2010). Asterisks indicate newly introduced dependencies that are not part of LFG theory.

structure component, which is the main motivation behind the multilingual LFG ParGram project (Butt et al., 2002). Thus, although the LFG-inspired dependency scheme was designed to describe English sentences, its roots in LFG theory make it a good starting point for developing resources for a language such as Irish with syntactic structures that are significantly different to English. We extend the LFG conversion tagset to create our own tagset of 47 labels. We discuss our Irish tagset in more detail in Section 4.

As discussed earlier in Section 2.2, there is relatively little research conducted on Irish syntax when compared to other better-resourced languages. However, a limited range of Irish linguistic phenomena has been covered to date in LFG research that are relevant to the design of our annotation scheme. For example, Asudeh (2002) focuses on an analysis of Irish preverbal particles and adjunction, Attia (2008) reports on an analysis of copula constructions taking Irish as an example and Sulger (2009b) carried out an analysis of Irish copula and cleft constructions. These previous studies also serve as valuable starting points for the design of our annotation scheme.

Annotation schemes can vary greatly not only in the dependency labelling (e.g. `infmod` vs `toinfinitive`) but also in their structural analysis. Compare Figures 3.2 (Stanford) and Figure 3.5 (LFG-inspired). Note that the verbs which are regarded as the root of the sentence differ across the two formalisms. The Stanford analysis regards the verb `tried` as the head of two dependents – `have` and `spruce`. LFG theory however, regards `have` as the head of (governor of) its dependent `tried`, which subsequently governs `spruce`.

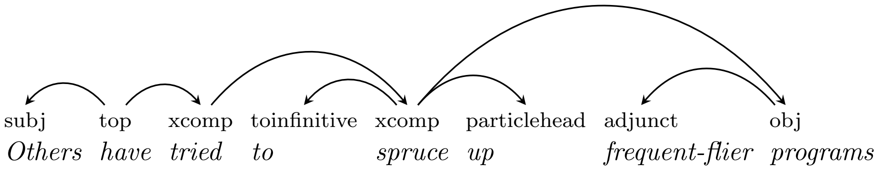


Figure 3.5: LFG-inspired dependency representation of parsed text

Other key linguistic phenomena, such as co-ordination and punctuation are treated differently across various formalisms. For example, by comparing Figure 3.6 and Figure 3.7, we can see the different ways in which coordination is handled across these two formalisms. Figure 3.6 shows a representation of Stanford coordination dependencies. This type of coordination is referred to as right-adjunction, where the first coordinate, `the cat`, is the head of the coordination and the rest of the phrase, `and the dog`, is adjoined to the right. This follows the argument that left/right branching conjunctions are often asymmetrical, with the dependencies between words on left-branching structures being shorter than right-branching.

In contrast, Figure 3.7 shows how LFG labels the coordinating conjunction `and` as the head, with the coordinates `the cat` and `the dog` as its dependents. This approach is often more favoured for retaining scope information.

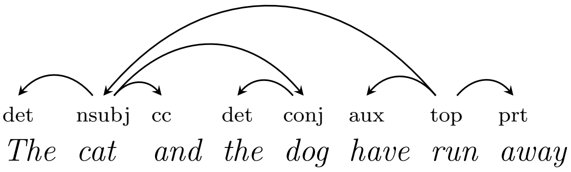


Figure 3.6: Stanford dependency representation of parsed text

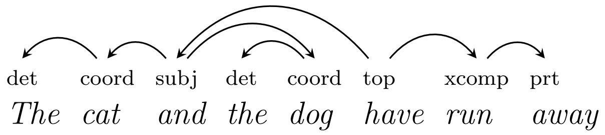


Figure 3.7: LFG-inspired dependency representation of parsed text

3.1.3 Developing upon existing NLP tools

When undertaking the development of a new NLP resource, particularly in the case of minority languages, it is important to build upon existing resources when possible and to leverage findings of previous studies.

In recent years, some progress has been made in the collection and development of linguistic resources for Irish. A 30 million word corpus of Modern Irish text (NCII)[5] was developed in 2004 for *Foras na Gaeilge*.[6] In addition, corpus annotation tools, namely a morphological analyser (Uí Dhonnchadha et al., 2003), a part-of-speech (POS) tagger (Uí Dhonnchadha and van Genabith, 2006) and a shallow parser (Uí Dhonnchadha and van Genabith, 2010) have been developed. A 3,000-sentence gold standard POS-annotated corpus was produced as a by-product of this work. These sentences were randomly selected from the NCII corpus and consist of text from books, newspapers, websites and other media, forming a solid representation of Modern Irish language data. We use this POS-tagged corpus as a basis for our treebank. The tags are based on the PAROLE Morphosyntactic Tagset (ITÉ, 2002).

Uí Dhonnchadha (2009) also made available a small corpus of 225 chunked Irish sentences. These sentences represented a test suite for a shallow parser (Uí Dhonnchadha and van Genabith, 2010) which is based on Constraint Grammar Dependency Mapping Rules (Karlsson, 1995) and implemented using Xerox Finite State Tools.[7] The dependency analysis for this parser was based on Constraint Grammar (CG), developed by Karlsson (1995). We present an example Irish sentence in Example 25. Figure 3.8 shows how CG rules annotate tokens of this sentence

[5] New Corpus for Ireland – Irish. See http://corpas.focloir.ie

[6] A government body in Ireland responsible for the promotion of the Irish language – http://www.forasnagaeilge.ie

[7] See http://xrce.xerox.com/ for more details on XFST

with grammatical functions such as @SUBJ (meaning the token is the subject) and dependency relations such as @>V (meaning the token is dependent on the verb to the right). Figure 3.9 shows the same chunked output without morphological tags. Finite State regular expressions are then applied (using Xerox XFST) to the annotated text to mark linguistic chunks. Chunks are groups of words that represent phrases, and are labelled as NP (noun phrase), V (verb), PP (prepositional phrase), for example. Dependency relation tags identify links between tokens within chunks, not between chunks.

(25) D' fhan siad ansin le fiche bliain
PAST stay they there with twenty years
'They stayed there for twenty years'

```
[S
[V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV]
[NP siad siad+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP]
[AD ansin ansin+Adv+Loc+@ADVL]
[PP le le+Prep+Simp+@PP_ADVL [NP fiche fiche+Num+Card+@>N
 bliain bliain+Noun+Fem+Com+Sg+@P<  NP]  PP]
. .+Punct+Fin+<<< S]
```

Figure 3.8: Example of chunked output (with morphological tags) for *D'fhan siad ansin le fiche bliain* 'They stayed there for twenty years'

```
[S
[V D' @>V fhan]
[NP siad]
[AD ansin]
[PP le  [NP fiche @>N bliain @P< ]]
. < S]
```

Figure 3.9: Example of chunked output for *D'fhan siad ansin le fiche bliain* 'They stayed there for twenty years'

The shallow nature of this chunking parser means that the dependency analysis does not extend to cover coordination, prepositional attachment, long-distance dependencies or clausal attachment. However, these 225 invented sentences cover the major syntactic phenomena of Irish and provided a valuable starting point for this treebank development. Many of our attachment rules within phrases are closely aligned with this particular grammar.

Our first step involved reviewing the dependency analysis for the shallow parser and adapting it to fit our chosen dependency scheme, as discussed in Section 3.1.2. We modified and extended the parses in this small chunked corpus to produce deep, full syntactic parses. Figure 3.8 is example output from the shallow parser for *D'fhan siad ansin le fiche bliain* 'They stayed there for twenty years'. The sentence is parsed into 4 chunks; V (Verb), NP (Noun Phrase), AD (Adverb) and PP (Prepositional Phrase). This output also indicates the kind of data available to us in the POS-tagged corpus that we use in our treebank. For example, the token *D'* is tagged with surface form (*D'*), lemma (*do*), coarse-grained POS-tag (*Part – particle*) and fine-grained POS-tag (*Vb – verbal*).

Using the chunked corpus as a starting point highlighted the type of linguistic analysis required for defining a new dependency annotation scheme. It also prompted discussions around the type of formalism our scheme should be based on, as discussed in more detail in Section 3.1.2. We then added these fully parsed sentences to the 3,000 gold standard POS-tagged corpus and subsequently randomised the data so that the relatively simple 225 sentences were dispersed throughout the corpus[8].

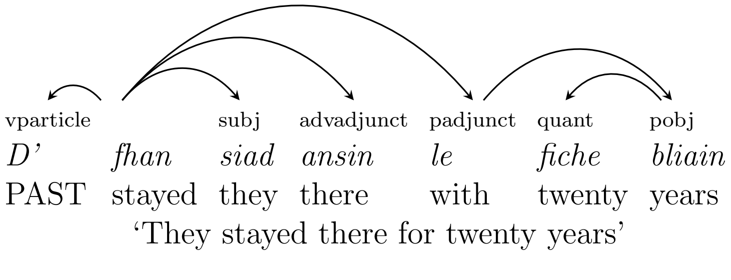


Figure 3.10: The fully parsed sentence of Figure 3.8

Figure 3.10 presents our extended parse analysis for the same sentence as Fig-

[8] The hand-crafted nature of the 225 sentences rendered them more simple structures than naturally occurring sentences in the corpus. This random dispersion prevented a chunk of the treebank being biased towards short, easy to parse sentences.

ure 3.8, showing in particular adverbial and prepositional attachment to the verb. A deeper syntactical analysis such as this provides a more informative linguistic description of Irish text. Consider, for example, the attachment of prepositional phrases. A shallow parse output does not identify the phrase that is being modified by the prepositional phrase. For example, an ambiguous sentence will have differing prepositional attachment – (i) to the verb as in Figure 3.11 or (ii) to the object of a prepositional phrase as in Figure 3.12.

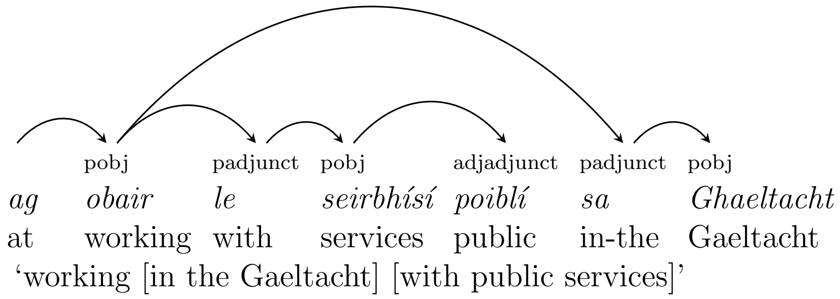

pobj
padjunct
pobj
adjadjunct
padjunct
pobj
ag *obair* *le* *seirbhísí* *poiblí* *sa* *Ghaeltacht*
at working with services public in-the Gaeltacht
'working [in the Gaeltacht] [with public services]'

Figure 3.11: Prepositional phrase ambiguity, where the head of the preposition *sa* is *obair*.

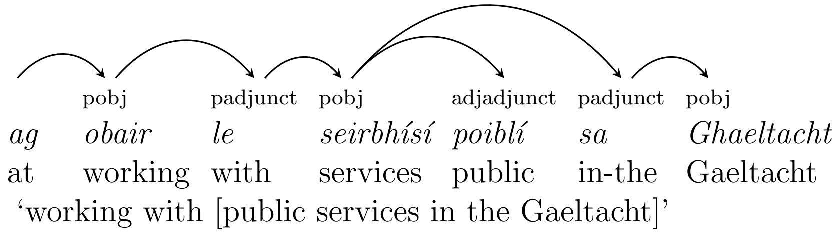

pobj
padjunct
pobj
adjadjunct
padjunct
pobj
ag *obair* *le* *seirbhísí* *poiblí* *sa* *Ghaeltacht*
at working with services public in-the Gaeltacht
'working with [public services in the Gaeltacht]'

Figure 3.12: Prepositional phrase ambiguity, where the head of the preposition *sa* is *seirbhísí*.

Figure 3.13 demonstrates how this type of ambiguity cannot be resolved with a shallow parser. We can see that there is no connection between the aspectual phrase *ag obair* and either of the prepositional phrases.

3.2 Inter-annotator Agreement

When working on a manual annotation or classification task, it is important to report inter-annotator agreement (IAA). This is a calculation of the normalised proportion

```
[PP-ASP ag  [NP obair @P<]]
[PP le  [NP seirbhísí poiblí @P< ]]
[PP sa [NP Ghaeltacht @P< ]]
```

Figure 3.13: Example of chunked output for *ag obair le seirbhísí poiblí* 'working with public services in the Gaeltacht'

of times annotators agree in their choice of labelling, which reveals the consistency of annotators. It is argued that this measurement can indicate usefulness or reliability of the data, and as we show below, identify gaps in a labelling scheme's annotation guide.

Carletta (1996)'s report on agreement assessment among NLP classificatory tasks found that simple calculations of agreement cited by researchers at that time were uninterpretable and misleading, as they did not account for the level of agreement expected by chance. The suggested solution was to adopt reliability measures from other fields, such as content analysis and medicine, that report an agreement measurement that corrected for chance. The Kappa coefficient of agreement, which normalises for chance, is now widely regarded as a standard for calculating IAA for corpus annotation tasks (Di Eugenio and Glass, 2004; Artstein and Poesio, 2008). This method of measurement has been adopted for assessing inter-annotator agreement in tasks such as discourse annotation (Poesio, 2004), word-sense annotation (Bruce and Wiebe, 1998) and POS annotation (Mieskes and Strube, 2006), for example.

At the time of our inter-annotator study, we observed that while agreement scores were reported in some of the dependency treebank literature, there did not appear to be a standard approach to measuring IAA for dependency parse annotation. Reports vary from "Labeled Attachment Kappa"[9] at a chunk level (Gupta et al., 2010), to multiple Kappa results (e.g. Uria et al. (2009)) and double-blind experiments (Voutilainen and Purtonen, 2011). In addition, Bhat and Sharma (2012) report a Kappa score for annotators of an Urdu dependency treebank. More re-

[9] Note: the authors do not state how this labeled attachment Kappa is calculated.

cently, Ragheb and Dickinson (2013) report IAA on the annotation of an English learner treebank, marked up with multiple layers of annotation: morphosyntactic, morphological, syntactic dependencies and subcategorisation information. They use MASI (Measuring Agreement on Set-valued Items (Passonneau et al., 2006)) as a metric to report on sets of unlabelled attachment agreement, labelled attachment agreement and label only agreement. They note that Kappa is not used in their calculations as the number of classes from which annotators need to choose is so large that chance agreement is unlikely. Since our study, Skjærholt (2014) has introduced a new metric that calculates agreement between dependency labels and attachments.

Below, we discuss the calculation of inter-annotator agreement on the Irish Dependency Treebank using labelled attachment scores, unlabelled attachment scores and the Kappa coefficient. We describe how subsequent workshops, where disagreements were analysed and discussed, lead to improvements of the annotation scheme and guidelines. We also report a second (improved) IAA score based on these updates.[10]

3.2.1 IAA and the Irish Dependency Treebank

Our two annotators are computational linguists with an advanced fluency of Irish and understanding of Irish syntax.[11] We calculated an inter-annotator agreement (IAA) measure on 50 sentences[12] of the Irish Dependency Treebank to assess consistency between both of our annotators (IAA-1). This task differs somewhat to other annotation tasks in that the agreement of the (head, label) pair of a dependency annotation cannot be measured in the same way as the agreement of single-value tags (e.g. POS-tags, discourse units, word-senses).

[10] This supports the findings of Hahn et al. (2012), who showed that iterative rounds of annotation and refinement of guidelines can be a successful approach to achieving accurate annotation of a corpus of named entity types.

[11] The primary annotator is the author, and the second annotator is Dr. Jennifer Foster, PhD supervisor at Dublin City University.

[12] After having manually annotated 300 sentences in our randomised corpus, we chose the next sequential 50 sentences for this study.

For that reason, we divided the assessment into two measurements:

(i) calculation of accuracy on (head, label) pair values through LAS/ UAS scores, taking the primary annotator's set as gold-standard data. LAS or Labelled Attachment Score is the percentage of words for which the two annotators have assigned the same head and label. UAS or Unlabelled Attachment Score is the percentage of words for which the two annotators have assigned the same head.

(ii) a kappa measurement of agreement on dependency tags (label values). There are two ways of calculating kappa – by assigning equal (Siegel and Castellan, 1988) or separate (Cohen, 1960) probability distributions among annotators. We report on Cohen's kappa coefficient measurement here for agreement between dependency label types. Our calculations do not take punctuation into account.

The Kappa statistic is defined as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of observed agreement among annotators, and $P(E)$ is the proportion of expected agreement. By correcting for $P(E)$, this measurement accounts for the fact that the annotators are expected to agree a proportion of times just by chance. Di Eugenio and Glass (2004) present the calculation of Cohen's $P(E)$ as:

$$P(E) = \sum\nolimits_j p_{j,1} \times p_{j,2}$$

where $p_{j,a}$ is the overall proportion of items assigned to a label $_j$ by annotator $_a$.

3.2.2 IAA Analysis and Treebank Update

The results of our two IAA studies are presented in Table 3.2. The Kappa score for IAA-1 is 0.7902 (where $P(A) = 0.8036$ and $P(E) = 0.0640$), and the score for IAA-2 is 0.8463 (where $P(A) = 0.8550$ and $P(E) = 0.0569$). We use Landis and Koch (1977)'s metric shown in Table 3.3 for interpretation of our Kappa results.

	Kappa (labels)	**LAS**	**UAS**
IAA-1	0.7902	74.37%	85.16%
IAA-2	0.8463	79.17%	87.75%

Table 3.2: Inter-annotator Agreement results.

Kappa value	**Strength of Agreement**
< 0.00	None
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Table 3.3: Landis and Koch's interpretation of Cohens's Kappa

We can see that our agreement scores increase from `Substantial` (0.79) in IAA-1 to `Almost Perfect` (0.85) in IAA-2. Here we discuss the relationship between the two inter-annotator agreement studies.

We held three workshops with our two annotators and one other fluent Irish speaker/linguist to analyse the results of IAA-1. We took both annotators' files from IAA-1 to assess the types of disagreements that were involved. The analysis highlighted many gaps in the annotation guide along with the requirement for additional labels or new analyses.

Following the analysis of the disagreements between annotators, we updated our annotation scheme and the annotation guide to address these issues. The updates are discussed in more detail in Section 4.

With our updated scheme, we then carried out a second IAA assessment (IAA-2) on a second set of 50 sentences from our randomised corpus. The results are shown in the second row of Table 3.2. A notable improvement in IAA-2 results, with Kappa increasing from 0.79 to 0.85, demonstrates that the post-IAA-1 analysis, the resulting workshop discussions and the subsequent updates to the annotation scheme and guidelines were highly beneficial steps towards improving the quality of the treebank.

3.2.3 Sources of annotator disagreements

The analysis of IAA results provided information valuable for the improvement of the annotation scheme. This analysis involved the comparison of both annotators' files of 50 sentences to see where they disagreed and the types of disagreements involved. Close examination of the disagreements allowed us to categorise them as: (i) Interpretation disagreements (ii) Errors (iii) Gaps in annotation guide (iv) Outstanding issues with the dependency scheme. These are discussed below.

3.2.3.1 Interpretation disagreements

The treebank data was extracted from the NCII which contains many examples of Irish legislative text. Some of these sentences are over 200 tokens in length and use obscure terminology or syntactic structures. Both annotators encountered difficulties in (i) interpreting these sentences and (ii) analysing their structures. Sources of disagreement included long distance dependencies and coordinated structures. The example provided in Example 26 is legal text from the official Irish translation of text from an EU Legal document:

(26) *Má dhéantar iarratas nó doiciméad eile nós imeachta a seoladh chuig an gCúirt Chéadchéime a thaisceadh de dhearmad*
'Where an application or other procedural document addressed to the Court of First Instance is lodged by mistake'.

Annotator 1 interpreted the verb *seoladh* 'address' as being in the infinitive form, and part of a complex autonomous verb construction, attached to the matrix verb *dhéantar*. Annotator 2 correctly annotated *a seoladh* as a relative modifier of *iarratas nó doiciméad* 'application or document', and *a thaisceadh* 'to lodge' as contributing to the autonomous verb construction. The confusion arose from the similar structure of infinitive verbs (*a seoladh*) and relative autonomous forms (*a thaisceadh*): 'a' + verbal noun, along with the unusual legalese use of an autonomous form *déan* 'to do' with an infinitive verb, resulting in a long distance dependency.

Similarly, autonomous relative forms caused confusion for the interpretation of the sentence in Example 27:

(27) *Iriseoir agus craoltóir í Seosaimhín Ní Bheaglaoich a tógadh i mBaile na bPoc, Contae Chiarraí.*

A1: 'Seosaimhín Ní Bheaglaoich, who grew up in Ballynabuk, Co. Kerry, is a journalist and broadcaster'

A2: Seosaimhín Ní Bheaglaoich, is a journalist and broadcaster who grew up in Ballynabuk, Co. Kerry'

Annotator 1 interpreted this sentence as 'Seosaimhín Ní Bheaglaoich, who grew up in Ballynabuk, Co. Kerry, is a journalist and broadcaster'. Annotator 2 interpreted it as 'Seosaimhín Ní Bheaglaoich, is a journalist and broadcaster who grew up in Ballynabuk, Co. Kerry'. The contrasting interpretations resulted in different head attachments for the relative modifier *tógadh* 'grew up'.

3.2.3.2 Errors

Human error played a relatively small role as both annotators carried out careful reviews of their annotations. Nevertheless, some discrepancies were due to an annotator applying the wrong label even though they had been aware of the correct one in previous annotations.

3.2.3.3 Gaps in the annotation guide

Gaps relate to a lack of sufficient examples in the annotation guide or lack of coverage for certain structures. For example, our analysis of IAA-1 confusions revealed that differences between the labels `padjunct` (prepositional modifier) and `obl` (oblique) had not been described clearly enough. The annotation guide was updated accordingly.

3.2.3.4 Outstanding issues in the dependency scheme

We also noted during the workshops that there were still some issues we had yet to fully resolve. Other outstanding issues involved linguistic phenomena that had not arisen in the data during earlier annotations and thus required discussion at this stage. The final decisions surrounding these analyses are presented in detail in Section 4.

3.3 Summary and Conclusion

In this chapter, we have described the steps involved with the development of the Irish Dependency Treebank. We started by providing a brief overview of the various types of treebanks available. We noted the various factors that influence the development of each of these treebanks and how these factors also influenced decisions related to the development of the IDT.

Firstly, we discussed two main syntactic representations, constituency and dependency trees. We explained the background of dependency syntax and how it has become a more popular approach to linguistic analysis and NLP in the past two decades. We reported the reasons that motivated us to choose to use a dependency syntax analysis in the IDT. We showed how the language-independent nature of a syntactic dependency analysis lends itself to our work on the Irish language, especially in light of the lack of consensus over constituency analyses of Irish in theoretical syntax.

We also explained our choice of labelling scheme, one that is heavily influenced by Çetinoğlu et al. (2010)'s LFG-inspired labelling scheme for English. The f-structure component of LFG, like dependency grammars, abstracts from phrase structures and focuses more on functional roles. We therefore easily adapted this scheme to our dependency analysis of Irish. We showed how the structure of our dependency analysis is influenced by LFG in the cases of coordination and clausal heads.

Like most NLP resources, a treebank relies heavily on the availability of other

NLP and language resources. We showed how existing resources such as a wide-domain corpus of Irish, a morphological analyser, a POS tagger, a POS-tagged corpus and shallow chunker proved to be an invaluable basis on which we could carry out our own research and build our treebank and parser.

Finally, we reported on two inter-annotator (IAA) studies carried out to assess the agreement level between our two treebank annotators. We showed how the metrics for reporting IAA results for dependency parsing vary in the literature. Our results are reported through LAS, UAS and Cohen (1960)'s kappa score, showing an increase in agreement from 0.79 (`Substantial`) to 0.85 (`Almost Perfect`) between the first and second study. We explained this increase through updates we made to the treebank and our annotation guide following a series of workshops where both annotators and one other linguist analysed the disagreements found in the first study. Our work showed that the benefits of conducting IAA studies not only includes assessment of accuracy and objectivity of the annotators, but also a medium by which a labelling scheme can be assessed. We conclude that this is a valuable stage in treebank development.

More information on the contents of the treebank, (including label and tag statistics, projectivity etc.), are presented in Appendix C.

In the next chapter we will discuss the specific linguistic considerations that were required to develop an Irish dependency treebank. We will show how many of these decisions, which were outcomes of the IAA study described here, have influenced the design of our labelling scheme.

Chapter 4

Defining the IDT Annotation scheme

In the previous chapter, we noted that a dependency structure analysis is applied to the Irish Dependency Treebank (IDT). We also discussed how our dependency annotation scheme is based on the LFG-inspired labelling scheme of Çetinoğlu et al. (2010) (refer to Section 3.1.2). One of our main challenges involves deciding on an appropriate linguistic analysis of Irish for a treebank, a question that has not been previously explored. In order to design our annotation scheme, we drew on existing analyses of Irish syntax (as outlined in Section 2.2.1) and examined the ways in which these analyses applied to our chosen formalism. Through synthesising both theoretical work and computational representations of syntactical structures, we have designed a broad-coverage syntactic analysis and labelling scheme that (i) covers at least all linguistic phenomenon that we have encountered thus far in the treebank data, and (ii) covers a number of syntactic characteristics in the Irish language that grammar writers have noted as unusual or problematic.

This chapter presents some aspects of the Irish Dependency Annotation Scheme in detail. Firstly, in Section 4.1, we present our labelling scheme. Secondly, in Section 4.2 we explain in detail the linguistic and annotation analysis that required specific consideration for Irish while designing the annotation scheme and adapting it from that of Çetinoğlu et al. (2010). We also adopt some labels (e.g. `csubj` and `pobj`) from the Stanford Typed Dependencies Scheme (de Marneffe and Manning, 2008). For the complete IDT annotation scheme, see Appendix A.

It should be noted here that there is an overlap between syntax, semantics and morphology as criterion for deciding on the labelled relations in our tagset. This is partly a result of the annotation schemes which influence our scheme. It also arises from the varying types of extractable information we want to identify, and expect will be most valuable for understanding, processing and interpreting Irish text.

4.1 Irish Dependency Labelling Scheme

The final labelling scheme for the IDT has 47 labels and is presented in Table 4.1. Our scheme contains a hierarchical structure with 7 parent labels and 17 sub-labels. The hierarchical structure provides an option for mapping to a less fine-grained scheme if required. Only the `pred` label is not used in the standard version of the treebank, but is an optional label to which its sub-labels can be mapped.

4.2 Language-specific choices

The following is a summary of the language-specific decisions we made regarding the annotation scheme that were driven by linguistic nuances of the Irish language. We highlight the motivation behind these analyses within our annotation scheme. Some of these choices were made during three workshops that took place as part of an inter-annotator agreement study, as discussed in Section 3.2. The following points should be noted regarding the discussion below:

- Compound prepositions are tokenised as single tokens: (e.g. *in_aghaidh* 'against', *os_cionn* 'above'). All other words are tokenised on white space. This means that contractions such as *d'ith* 'ate' are split (e.g. *d'*   *ith*).

dependency label	**function**
top	root
addr	addressee
adjunct	adjunct
adjadjunct	adjectival modifier
advadjunct	adverbial modifier
nadjunct	nominal modifier
padjunct	prepositional modifier
subadjunct	subordinate conjunction
advadjunct_q	adverbial adjunct (question)
app	noun in apposition
aug	augment pronoun
comp	closed complement
coord	coordinate
dem	demonstrative pronoun
det	determiner
det2	post or pre-determiner
for	foreign (non-Irish) word
obj	object
pobj	object of preposition
vnobj	object of verbal noun
obj_q	object (question)
obl	oblique object
obl2	second oblique object
obl_ag	oblique agent
particle	particle
relparticle	relative particle
cleftparticle	cleft particle
advparticle	adverbial particle
nparticle	noun particle
vparticle	verb particle
particlehead	particle head
qparticle	quantifier particle
vocparticle	vocative particle
poss	possessive pronoun
pred	predicate
ppred	prepositional predicate
npred	nominal predicate
adjpred	adjectival predicate
advpred	adverbial predicate
punctuation	internal and final punctuation
quant	quantifier
relmod	relative modifier
subj	subject
csubj	clausal subject
subj_q	subject (question)
toinfinitive	infinitive verb marker
xcomp	open complement

Table 4.1: The Irish Dependency Treebank labels: sublabels are indicated in bold, indented below the parent label.

- All of the examples we provide here as illustrations are taken or adapted from the IDT.

- The label `top` is used to indicate the root of a sentence. If the example provided is only a fragment or a phrase, no label is used.

4.2.1 Labelling of predicates

In the early stages of designing the annotation scheme, we labelled all predicates of both the copula *is* and the substantive verb *bí* as `xcomp`. The predicate complement function is referred to as `xcomp` (open complement) in LFG (Bresnan, 2001, p. 267). The complement type can be X = V(erb), N(oun), A(dverb) or P(reposition). Both Dalrymple et al. (2004) and Sulger (2009b) discuss the use of `xcomp` in copular constructions in particular. For example, they note how `xcomp` can be used in an LFG f-structure for the French sentence *elle est petite* 'she is small'.

However, in Irish, open complement verbs (non-finite clauses and progressive verb phrases) are also labelled as `xcomp` in our scheme.[1] In order to differentiate these different kinds of functions, we adopted a new `pred` (predicate) label hierarchy of `npred` (nominal), `ppred` (prepositional), `adjpred` (adjectival) and `advpred` (adverbial). While a more fine-grained labelling scheme can result in more data sparsity, it also results in a more precise description of Irish syntax. The hierarchical structure allows for mapping back to a more coarse-grained tag if necessary. Figure 4.1 gives an example of a predicate in a copular construction: *An leatsa an teach?* can be literally translated as 'Is it yours the house?'. Possession can be denoted through idiomatic phrases such as this, using prepositional phrases or inflected prepositions such as *leat* 'with you'. In this example, *leatsa* contains the emphatic suffix *-sa*. Figure 4.2 gives an example of a predicate with a substantive verb, where *déanta* 'made' is a verbal adjective or 'past particle'.

4.2.2 Irish copula

A copula is a word which links a subject and a predicate. In some languages the copula is regarded as a verb (copular verb), even though it might not always present

[1] LFG uses the grammatical function `xcomp` to represent non-finite complements (Carnie, 2007).

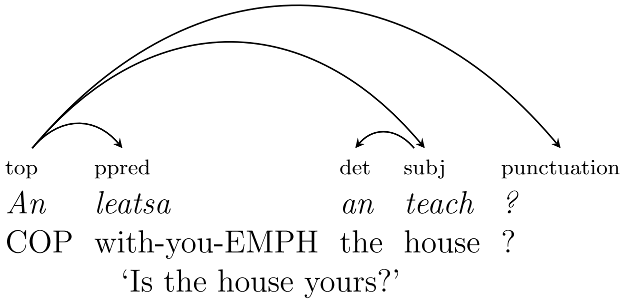


Figure 4.1: Copula construction with prepositional predicate.

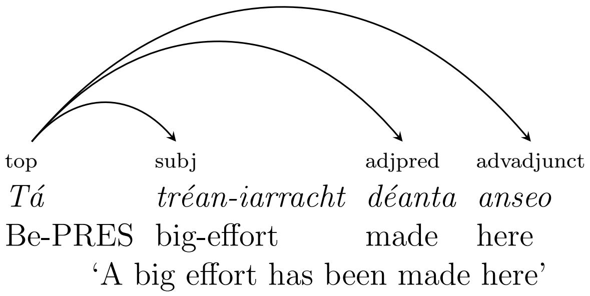


Figure 4.2: Substantive verb construction with adjectival predicate.

the same behaviour as regular verbs. In English, the copula is the verb 'to be'. In Irish, however, there is a distinction between the substantive verb *bí* 'to be', which inflects for tense, mood and person as per all Irish verbs, and the copula *is*, which only has two tensed forms – present/future and past/conditional.

Bí, as a verb, uses separate particles in negative and interrogative constructions with all tense and mood forms. Some examples are given in Example 28 (where the particles are bolded).

(28) *tá tú* 'you are'

níl *tú* 'you are not'

an *bhfuil tú* 'are you?'

bhí tú 'you were'

ní *raibh tú* 'you were not'

nach *raibh tú* 'were you not?'

beidh tú 'you will be'

an *mbeidh tú* 'will you be?'

nach *mbeidh tú?* 'will you not be?'

bheifeá 'you would be'

an *mbeifeá* 'would you be?'

nach *mbeifeá?* 'would you not be?'

The copula *is* uses its own forms in these constructions as is shown in Example 29. The main forms are: *is* (positive – present/ future), *ní* (negative – present/ future), *ba* (positive – conditional/ past), *níor* (negative – conditional/ past), *an* (interrogative/ positive – present/ future), *ar* (interrogative/ positive – conditional/past), *nach* (interrogative/ negative – present/ future), *nár* (interrogative/ negative – conditional/past).

(29) ***is*** *maith leat* 'you like'

ní *maith leat* 'you don't like'

ba *mhaith leat* 'you would like'

níor *mhaith leat* 'you would not like'

an *maith leat?* 'do you like?'

ar *mhaith leat?* 'would you like?'

nach *mhaith leat?* 'do you not like?'

nár *mhaith leat?* 'would you not like?'

Irish copula constructions present interesting questions when being defined by dependency relations. The order of elements is in general: copula, predicate (new or focussed information), and subject (The Christian Brothers, 1988, p.123). The equative (identification) copula example in Figure 4.3 translates to English as 'You are the teacher'. However non-intuitive to an English speaker, following The Christian Brothers (1988) we identify *tusa* 'you' as the predicate and *múinteoir* 'teacher' as the subject. The Christian Brothers (1988, p.124) explain this role-labelling by the fact that it answers the question 'Who is the teacher?'. The answer in Irish reads literally as 'The teacher is you'.

In addition, it is worth considering that the interrogative form of this sentence is *An tusa an múinteoir?* 'Are you the teacher' (lit. 'Is the teacher you?'). It is not possible to swap the elements in this instance to **An an múinteoir thú?*. It is only possible to swap the elements in a predicative construction with 'you' as the subject, if the predicate 'teacher' is indefinite (new information): *An múinteoir thú?* 'Are you **a** teacher?'.

It may be worth noting here that in some analyses (e.g. Carnie (1997)), the Irish copula in this construction is regarded as a complementizer particle which links two noun phrases, but not labelling either element as a subject or predicate, and therefore not identifying either as a focussed or unfocussed dependent. We do not consider this unlabelled analysis, since, according to the dependency scheme we have adopted, all relations must be labelled, and we also want to differentiate the roles of each of the noun phrases.

Our analysis defines the copula as the head of the construction, with the subject and predicate as its dependents.

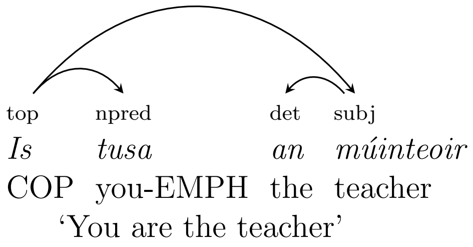


Figure 4.3: Dependency tree for Irish copula equative construction.

We follow this copula-predicate-subject analysis as per Uí Dhonnchadha (2009), which applies to other constructions that we list here and describe in detail below:

- classificatory constructions: *Is lá deas é* 'It is a nice day'

- ownership constructions: *Is liomsa é* 'It is mine'

- fronting constructions: *Is ise a chonaic sé* 'It is she whom he saw' (see Section 4.2.4)

- identification constructions: *Is iad na buaiteoirí* 'They are the winners'

- idiomatic use: *Is maith liom tae* 'I like tea'

Classificatory constructions The Christian Brothers (1988) note that classification sentences are used to tell us what a person or thing is. The predicate is always indefinite. See Figure 4.4 for example.

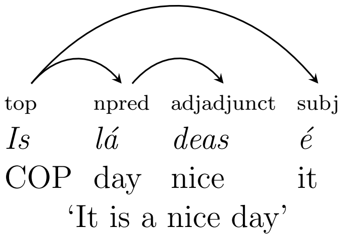


Figure 4.4: Dependency tree for classificatory constructions.

Ownership constructions Ownership can be denoted through the use of both the substantive verb *bí* and the copula. In copula ownership constructions, the preposition *le* is used, and can incorporate an inflected object pronoun. See Figure 4.5 for an example of use with the inflected preposition *liomsa*.[2]

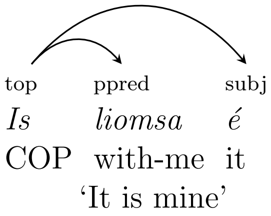


Figure 4.5: Dependency tree for ownership constructions.

Identification constructions In identification sentences, the subject cannot be indefinite and the predicate is always a definite noun or pronoun (The Christian

[2] See Section 4.2.8 for more detail on these prepositions.

Brothers, 1988). It is important to note the presence of an agreement pronoun in the case of identification sentences, where the copula precedes a definite NP. Stenson (1981, p.96) refers to this pronoun as a 'subpredicate' and Doherty (1997) refers to it as an 'augment pronoun' – a term which we adopt here. The augment pronoun agrees in both number and gender with the NP. We make the dependency attachment between the two, with the pronoun modifying the NP, using the label `aug`, as shown in Figure 4.6.

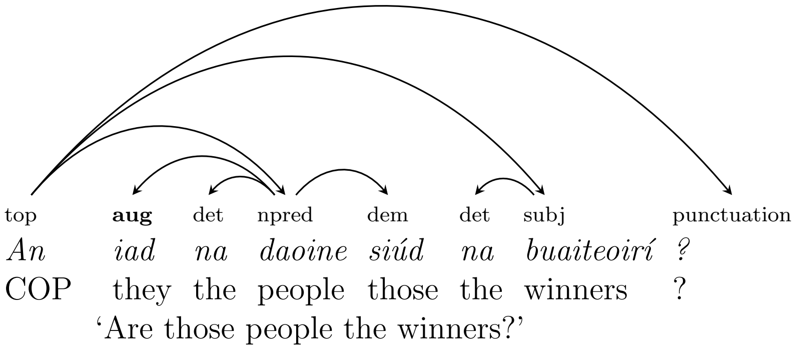

top
aug
det
npred
dem
det
subj
punctuation
An *iad* *na* *daoine* *siúd* *na* *buaiteoirí* *?*
COP they the people those the winners ?
'Are those people the winners?'

Figure 4.6: Dependency tree for identification constructions with an augment pronoun.

Idiomatic constructions The copula can also be used in idiomatic constructions to express feelings or desires (Mac Congáil, 2002). See Figure 4.7 for example.

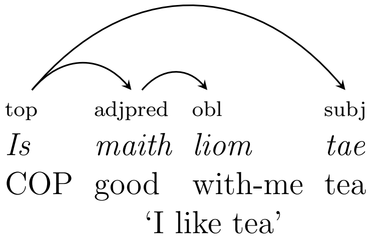


Figure 4.7: Dependency tree for an Irish copula predicate with oblique argument.

These constructions follow the order of copula – predicate – subject, where the predicate head is normally an adjective or noun with a prepositional dependent. We attach this prepositional phrase (often in the form of prepositional pronoun) to the predicate as an `obl` argument. Note that this differs somewhat from Sulger

(2009b)'s LFG analysis which regards this attachment as *adjunct*. We however, do not regard the preposition as an optional modifier in these idiomatic clauses.

4.2.3 Copular subject complements

In copular constructions, the grammatical subject may take the form of a finite verb clause. We initially labelled the head of the clause (the verb) as a subject (`subj`), in alignment with the analysis discussed in Section 4.2.2. However, we now choose instead to highlight the clausal nature of these finite verb subjects with a more specific label, i.e. subject complement (`csubj`[3] – a subtype of `subj`). See Figure 4.8 for example.

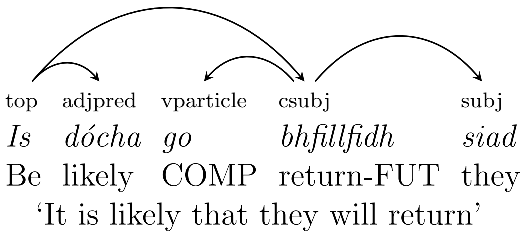

top adjpred vparticle csubj subj
Is dócha go bhfillfidh siad
Be likely COMP return-FUT they
'It is likely that they will return'

Figure 4.8: Dependency structure with subject complement labelling.

4.2.4 Cleft constructions – cleftparticle

Clefting or fronting is a commonly used structure in the Irish language. Elements are fronted to predicate position to create emphasis or focus. Irish clefts differ from English clefts in that there is more freedom with regards to the type of sentence element that can be fronted (Stenson, 1981, p.99). In Irish, the structure is as follows: Copula, followed by the fronted element (Predicate), followed by the rest of the sentence (Relative Clause). The predicate can take the form of a noun phrase (headed by pronoun, noun, verbal noun), or adjectival, prepositional or adverbial phrases. For example:

[3] This label is also used in the English Stanford Dependency Scheme (de Marneffe and Manning, 2008).

- Adverbial Fronting:

  Is ***laistigh de bhliain*** *a déanfar é*: 'It's **within a year** that it will be done'

- Pronoun Fronting:

  Is ***ise*** *a chonaic siad inné*: 'It is **she** whom they saw yesterday'

Stenson (1981, p.111) describes the cleft construction as being similar to copular identity structures with the order of elements as Copula, Predicate, Subject. This is the basis for the cleft analysis provided by Sulger (2009b) in the Irish LFG literature. We follow this analysis but with a slight difference in the way we handle the *a*. According to Stenson, the *a* is a relative particle which forms part of the relative clause. However, there is no surface head noun in the relative clause – it is missing an NP. Stenson refers to these structures as having an 'understood' nominal head such as *an rud* 'the thing' or *an té* 'the person/the one', e.g. *Is ise* ***[an té]*** *a chonaic siad inné.* When the nominal head is present, it becomes a copular identity construction: *She is the one who they saw yesterday.* In the absence of a head noun and because we do not want to introduce empty elements in the dependency tree, we label the verb as the head of the clause. To distinguish the *a* in these cleft sentences from those that occur in relative clauses with surface head nouns, we introduce a new dependency label `cleftparticle` and we attach *a* to the verb *chonaic* using this relation. This is shown in Figures 4.9 and 4.10.

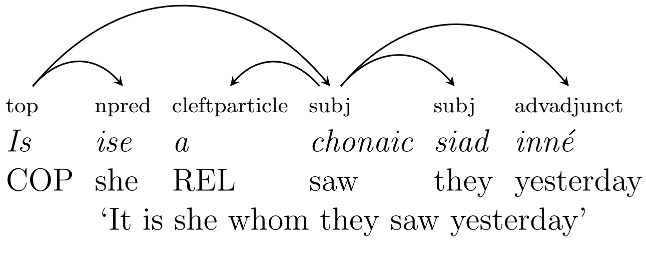


Figure 4.9: Dependency structure for cleft construction (nominal fronting).

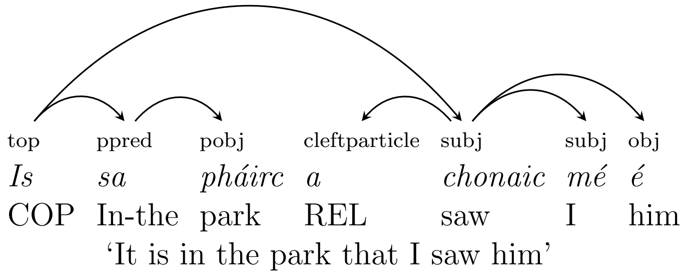

top ppred pobj cleftparticle subj subj obj
Is sa pháirc a chonaic mé é
COP In-the park REL saw I him
'It is in the park that I saw him'

Figure 4.10: Dependency structure for cleft construction (prepositional fronting).

4.2.5 Copula drop

It is possible to drop the copula from most sentences when it is not "marked by features of tense or mood, negation or interrogation" (Stenson, 1981, p.94) or when there is an idiomatic predicate (Stenson, 1981, p.125). Despite deletion, the copula is understood. As the copula is normally the head of a clause in our analysis, we propose that the predicate can raise to head position when the copula is dropped. See Figure 4.11 for an example of how the drop is analysed.

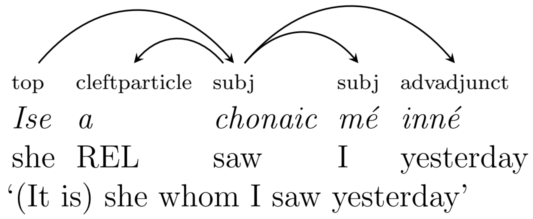

top cleftparticle subj subj advadjunct
Ise a chonaic mé inné
she REL saw I yesterday
'(It is) she whom I saw yesterday'

Figure 4.11: Dependency structure for cleft construction with copula drop.

4.2.6 Copula – Complementiser form

If a copula is the head of a complement phrase, by merging with the complementiser *go* it adopts a complementiser form (Stenson, 1981, p.93): *gur* in the present-future form (*gurb* before vowels) and *gur(bh)* in the past-conditional form. As the verb is the governor of a complementiser in general, the contracted form is attached to the matrix clause as `comp`.

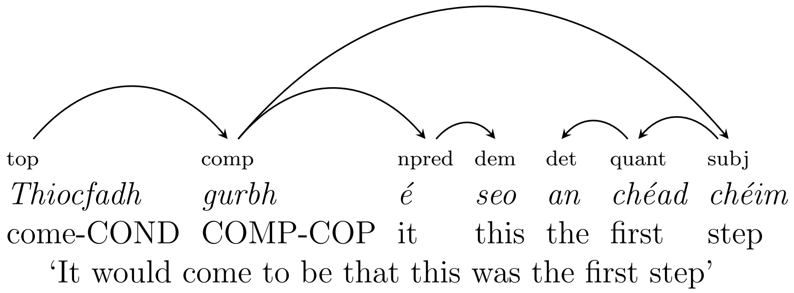


Figure 4.12: Dependency structure for copula complementiser form.

4.2.7 Copular subordinator

In some cases, the copula head of a subordinate clause can combine with a subordinate conjunction (e.g. *má* 'if'), resulting in one contracted word form (e.g. *más*). In these cases, we use the label `subadjunct`, treating the unit as an inflected form of the conjunction. The would-be dependents of the copula thus become dependents of the copular subordinator. For example, see the attachments of `npred` (nominal predicate) and `xcomp` (open complement) in Figure 4.13.

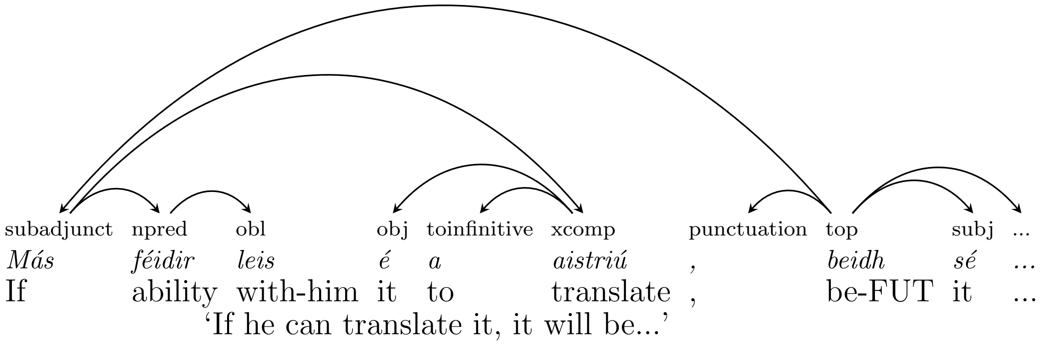


Figure 4.13: Dependency structure for copular-subordinator.

4.2.8 Prepositional attachment

While prepositional phrase attachment has already been extensively studied in other languages, Irish possesses some unusual prepositional behaviour that requires some discussion here.

Irish has simple prepositions (e.g. *le* ‘with’, *ar* ‘on’, *ag* ‘at’), and compound prepositions (e.g. *in aghaidh* ‘against’, *os cionn* ‘above’). Most of the simple prepositions can inflect for person and number to indicate a personal pronoun, thus including a nominal element (prepositional pronouns/ pronominal prepositions). When comparing *leat* ‘with you’ (Figure 4.14) and *leis an bhfear* ‘with the man’ (Figure 4.15) it is clear that such inflection creates data sparsity within the treebank.

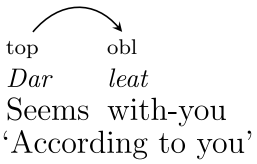


Figure 4.14: Attachment of prepositional pronoun

The PAROLE prepositional pronoun part-of-speech coarse and fine-grained tags are Pron (pronoun) Prep (preposition). We choose to see the relationship between prepositional pronouns and their heads (normally a verb) as a prepositional attachment as opposed to a nominal attachment. Therefore we use either `padjunct` (prepositional adjunct – when the attachment is optional) or `obl` (oblique – when the attachment is not optional and the prepositional pronoun is closely tied to the verb).

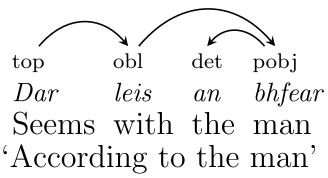


Figure 4.15: Attachment of oblique preposition.

Prepositional phrases are also used in idiomatic expressions for denoting psychological states and ownership (Stenson, 1981, pp 57,98). These usually follow the pattern of *Tá NP ag/ar NP*. ‘NP is at/on NP’. For example *Tá bron orm* ‘I am sorry’ (lit. sorrow is on me) and *Tá peann agam* ‘I have a pen’ (lit. A pen is at me). See Figure 4.16 for this analysis. See also Section 4.3 for other examples of copular idiomatic use with prepositions.

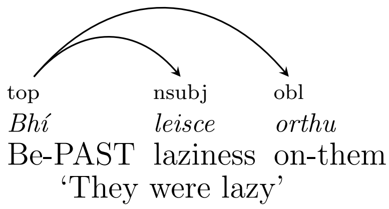


Figure 4.16: Attachment of idiomatic oblique preposition.

Progressive aspectual phrases represent another interesting preposition function. As presented by Uí Dhonnchadha (2009), these types of phrases, such as *Tá sé ag iascaireacht* 'He is fishing' are constructed using the substantive verb *Tá* 'is' as an auxiliary, along with a non-finite complement (a prepositional phrase consisting of a preposition *ag* 'at' and a verbal noun *iascaireacht* 'fishing'). The verbal nature of these types of prepositional phrases means that they cannot be labelled as adjuncts, which is often the case for prepositional attachment. Instead, we regard them as predicates. As discussed in Section 4.2.1, non-verbal predicates such as prepositional phrases can be labelled as open complements (xcomp) in LFG. While examples of prepositional phrase predicates can be found in English (e.g. 'She kept out of the argument'), they are used in limited circumstances. In contrast, periphrastic constructions like this involving prepositional phrase predicates occur frequently in Irish. See in Figure 4.17 for progressive aspectual dependency representation.

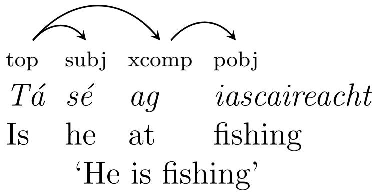


Figure 4.17: Dependency tree for Irish progressive aspectual phrase.

4.2.9 Prepositions and complement phrases

Verbal noun complements are negated by the preposition *gan.* In these cases where infinitive phrases follow a preposition, we attach the verbal noun (infinitive verb) to the preposition. There were two options available to us for labelling this attachment.

Either (i) as per all other infinitive phrase cases, attach as `xcomp`, or (ii) treat the infinite verb as an object of the preposition `pobj`. Given its verbal noun part of speech, we choose `pobj` as is shown in Figure 4.18.

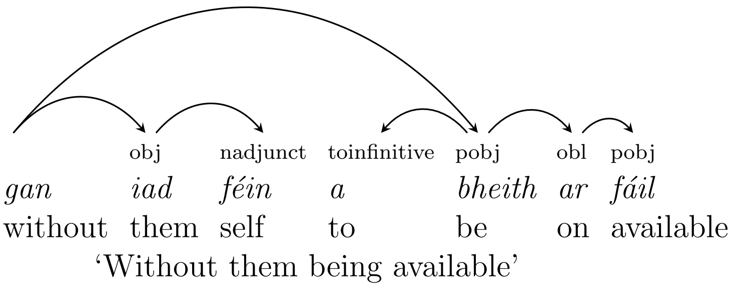

obj
nadjunct
toinfinitive
pobj
obl
pobj
gan *iad* *féin* *a* *bheith* *ar* *fáil*
without them self to be on available
'Without them being available'

Figure 4.18: Dependency structure for attachment of complement phrases to prepositions.

4.2.10 Gerund object attachment

The attachment of objects of verbal nouns in progressive aspectual phrases also required consideration. Take the sentence *Bhí sé ag lorg tacaíochta* 'He was seeking support' for example. Two attachment options appeared possible at first: (i) attaching the object *tacaíochta* to the preposition *ag* or (ii) attaching it to the verbal noun *lorg*. Given that the object noun is in the genitive case, we choose to regard them as modifiers of the verbal noun and attach them accordingly. Thus, *Bhí sé ag lorg tacaíochta* could be interpreted literally as 'He was at support-seeking'. We label the relation as `vnobj`. See Figure 4.19 for a progressive phrase example.

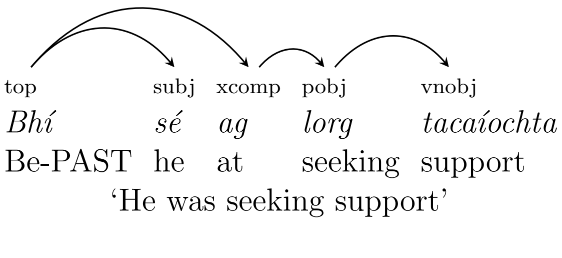

top
subj
xcomp
pobj
vnobj
Bhí *sé* *ag* *lorg* *tacaíochta*
Be-PAST he at seeking support
'He was seeking support'

Figure 4.19: Dependency structure for gerund object attachment.

4.2.11 Gerund adverb attachment

We also considered two options for adverbial attachment to verbal nouns in progressive aspectual phrases. Take the phrase *ag teacht ar ais* 'coming back' for example. Two attachment options appeared possible: (i) attaching the adverbial *ar ais* to the preposition *ag* or (ii) to the verbal noun *teacht.* Consideration was given to the fact that verbal nouns can also play a nominal role, i.e. *teacht ar ais* can occur independently of *ag* as a subject or object. See Examples 30 and 31.

(30) Bhí [teacht ar_ais] ar intinn aici
Was [coming back] on mind at-her
'She had returning in mind'

(31) Bhí a lán deacrachtaí ag baint leis an [teacht ar_ais]
Was PART many difficulties at relating with the [coming back]
'There were a lot of difficulties associated with the return'

For this reason, we make the attachment between the verbal noun and the adverb. See Figure 4.20 for example.

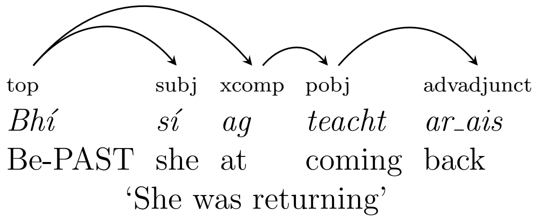


Figure 4.20: Dependency structure for gerund adverb attachment.

4.2.12 Complement phrase attachment to verbal nouns.

In keeping with other attachments to verbal nouns in progressive aspectual phrases, we choose to attach the head of a dependent complement phrase to the verbal noun instead of *ag*. Figure 4.21 shows how an infinitive complement phrase *a mbealach a dhéanamh* 'to make their way' is attached to the verbal noun of a progressive aspectual phrase (*iarraidh*).

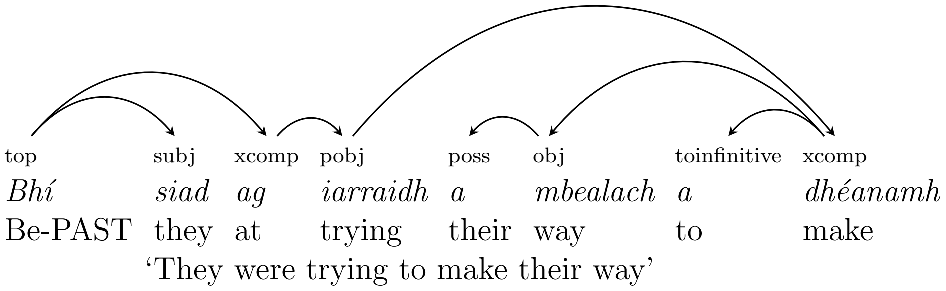


Figure 4.21: Dependency structure for complement phrase attachment to verbal nouns

4.2.13 Objects of infinitive phrases

We regard the infinitive verb form (verbal noun) as the head of an infinitival phrase. An object of the infinitive verb is labelled *obj*, but note that there are structural differences when compared to English; the object precedes the verb: *a mbealach a dhéanamh* 'to make their way' (lit. 'their way to make'). Figure 4.21 also shows this.

4.2.14 Objects of autonomous verbs

Irish does not have an equivalent to the English passive construction (The Christian Brothers (1988, p.120) and Stenson (1981, p.145)). Stenson identifies autonomous verbs and stative passives (see Section 4.2.15) as constructions that are used instead. While only transitive verbs can be used in the English passive voice, all Irish verbs (apart from the copula) have an autonomous form. Additionally, English passives can specify agents (e.g. 'the door was closed by him'), yet Irish autonomous verbs cannot express agents (e.g. **dúnadh an doras aige*).

The subject is not specified (overt) in these impersonal verb forms, but understood. It is argued that the third person subject is marked as an inflection in the autonomous form (Stenson, 1981, p.147). For example *scaoileadh iad* translates roughly as 'they were released', but literally translates as 'somebody released them'. The accusative case is marked in Irish pronouns, and in this example *iad* 'them' (the accusative pronoun form of *siad* 'they') confirms an object role. The object therefore does not assume subject position as is the case in English passives, and thus we use the `obj` label. See Figure 4.22 for example.

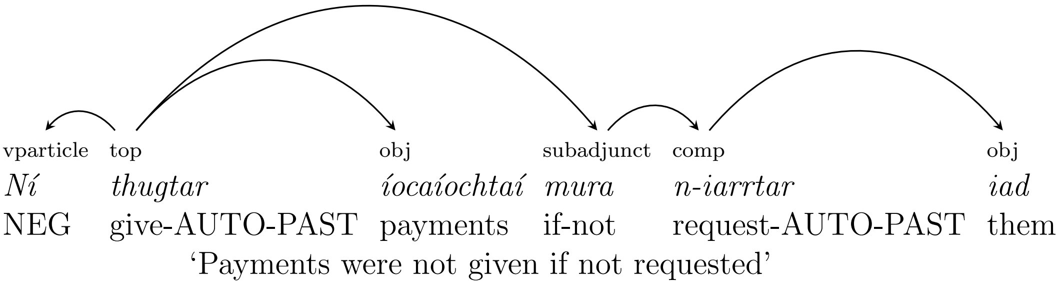

vparticle
top
obj
subadjunct
comp
obj
Ní *thugtar* *íocaíochtaí* *mura* *n-iarrtar* *iad*
NEG give-AUTO-PAST payments if-not request-AUTO-PAST them
'Payments were not given if not requested'

Figure 4.22: Dependency structure for autonomous verbs

4.2.15 Oblique agents of stative passives

Stative passive constructions consist of a substantive verb *bí* and a verbal adjective. Unlike autonomous verbs, these constructions may specify an external agent, that is, the person or thing that caused the action or state. We introduce a fine-grained oblique label (`obl_ag`) to mark the agent of the verbal adjective. The agent is essentially a prepositional phrase using the preposition *ag* 'at' or its various forms inflected for person and number. For example, in Figure 4.23, we can see that *an leanbh* 'the child' is the agent, and is the object of the preposition *ag* 'at'.

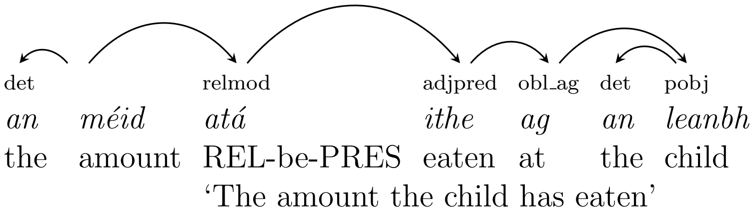

det
relmod
adjpred
obl_ag
det
pobj
an *méid* *atá* *ithe* *ag* *an* *leanbh*
the amount REL-be-PRES eaten at the child
'The amount the child has eaten'

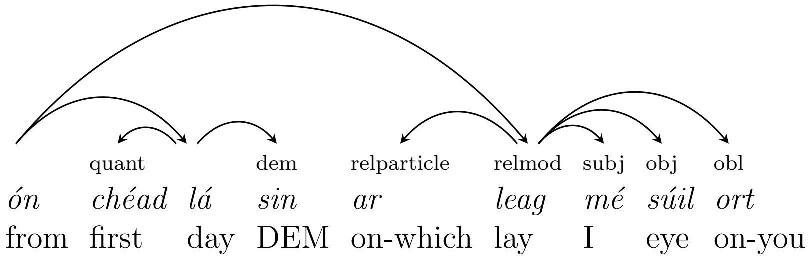
Figure 4.23: Dependency structure for internal comma and quotation marks

4.2.16 Relative particle

Asudeh (2002) provides an overview of some of the disagreements in Irish syntax literature regarding the role of the relative particle *a*. McCloskey (1979, p.12) regards it as a complementiser, putting it into the same category as *go*, and its variants *nach, nár, gur* as referred to in Section 4.2.6. Sells (1984) argues that this is a preverbal particle marking resumptive pronouns, and Stenson (1981, p.34) and The Christian Brothers (1988, pp. 143–146) refer to them as relativisers. We view their role as relativisers (indirect and relative particles) but do not mark the resumptive pronoun role in our analysis, i.e. we do not label them according to the functional role of the noun to which they refer (e.g. subject, object).[4] Instead we mark them as `relparticle` indicating their particle features, yet differentiating them from other preverbal particles such as interrogative particles (e.g. *an*) and negative particles (e.g. *ní*).

This analysis extends to both the direct and indirect particle *a* and their negative variants *nach, nár* (see Figure 4.24), along with relatives that occur after prepositions (e.g. *inar*, *as a* – see Figure 4.25). When relative particles are merged with a verb form (e.g. *atá*), the `relmod` label is used.

Note that this relative particle labelling does not apply to particles in wh-questions such as *Cad a d'ith sé* 'What did he eat?' (see Section 4.2.21 for further explanation).


quant
dem
relparticle
relmod
subj
obj
obl
ón chéad lá sin ar leag mé súil ort
from first day DEM on-which lay I eye on-you

'From the first day I laid eyes on you'

Figure 4.24: Dependency structure for indirect relative particles

[4] This differs from the Universal Dependencies analysis we follow in Section 5.2.2.2

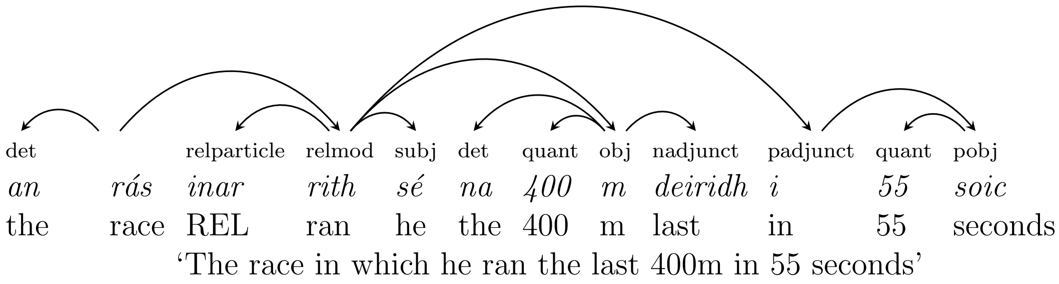

det
relparticle
relmod
subj
det
quant
obj
nadjunct
padjunct
quant
pobj
an *rás* *inar* *rith* *sé* *na* *400* *m* *deiridh* *i* *55* *soic*
the race REL ran he the 400 m last in 55 seconds
'The race in which he ran the last 400m in 55 seconds'

Figure 4.25: Dependency structure for relative particles following prepositions

4.2.17 English chunks

If an English word occurs alone within Irish text, and it has been correctly pos-tagged as a syntactic component of the sentence (i.e. not tagged as `Foreign`), then the dependency label used for the attachment of this word reflects its syntactic role within the sentence.

However, for a string of two or more English words, the first word of the string is parsed as though it is an Irish word, and the remaining English words attach to it using the label `for`. See Figure 4.26. Note that there are relatively few cases of English in the IDT, and they are mainly cases of quotations or untranslated titles.

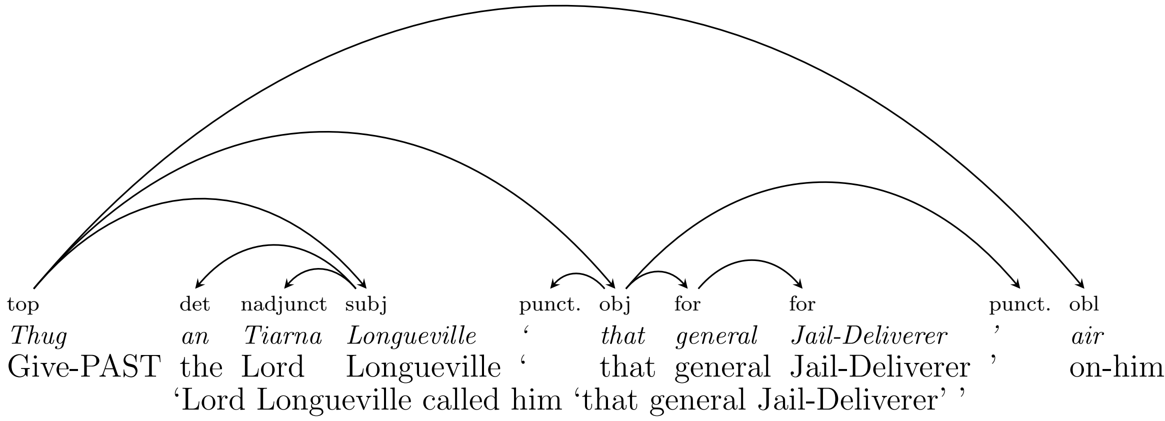

top
det
nadjunct
subj
punct.
obj
for
for
punct.
obl
Thug *an* *Tiarna* *Longueville* *'* *that* *general* *Jail-Deliverer* *'* *air*
Give-PAST the Lord Longueville ' that general Jail-Deliverer ' on-him
'Lord Longueville called him 'that general Jail-Deliverer' '

Figure 4.26: Dependency structure English chunks

4.2.18 Internal punctuation

Punctuation can vary across annotation schemes. In general, final punctuation is usually a dependent of the root of the sentence. However, internal punctuation rules can vary. For example, some dependency analyses, such as the conversion[5] of the TIGER treebank (Brants et al., 2002) to dependency structures attaches each punctuation token to the token to the left. Others, such as the Universal Dependency scheme (Nivre et al., 2015) determines punctuation attachment according to its type (e.g. separating coordinated units, preceding or following subordinated units, or paired punctuation).

Our analysis of internal punctuation is as follows:

- Internal commas, colons (:), semi-colons (;) and forward slashes (/) are attached to the head of the immediately following phrase (Figure 4.27).
- Quotation marks of direct speech are attached to the prompting verb (e.g. *arsa/ dúirt* 'said') (Figure 4.27). All other quotation marks are treated as paired punctuation.
- Paired punctuation (parenthesis): the head of the phrase within parentheses is the head of both the opening and closing brackets (Figure 4.28).
- Coordinator: Sometimes, punctuation (e.g. commas, forward slashes) are used as coordinators. These are discussed in more detail in Section 4.2.20.

4.2.19 Multiple coordination

As discussed in Section 3.1.2, we follow an LFG-inspired analysis of coordination, the coordinating conjunction is the head, and the dependents are the coordinate phrases. Problems arise when there are multiple coordinates and it is not clear how to combine them. This multiple use of 'and' in a single sentence is more common

[5] Tiger2Dep conversion tool, available at: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/Tiger2Dep.en.html`

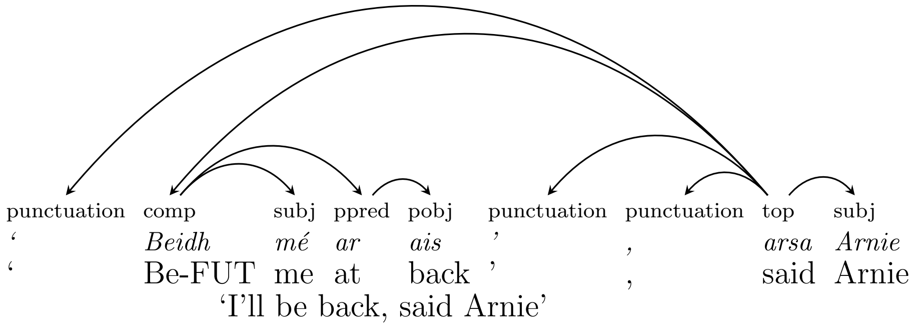

punctuation comp subj ppred pobj punctuation punctuation top subj
‘ Beidh mé ar ais ’ , arsa Arnie
‘ Be-FUT me at back ’ , said Arnie
‘I’ll be back, said Arnie’

Figure 4.27: Dependency structure for internal comma and quotation marks

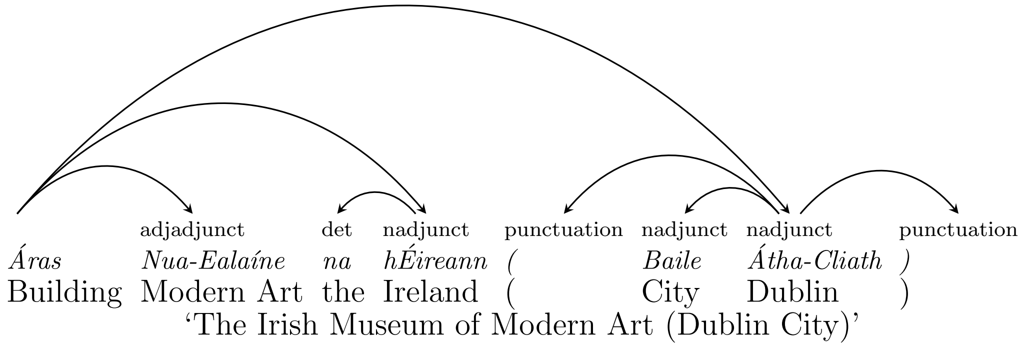

adjadjunct det nadjunct punctuation nadjunct nadjunct punctuation
Áras Nua-Ealaíne na hÉireann (Baile Átha-Cliath)
Building Modern Art the Ireland (City Dublin)
‘The Irish Museum of Modern Art (Dublin City)’

Figure 4.28: Dependency structure for parenthesis

and acceptable in Irish than in English. In cases of clustered coordination that are unequal in number, we treat the second group as a cluster – see Figure 4.29.

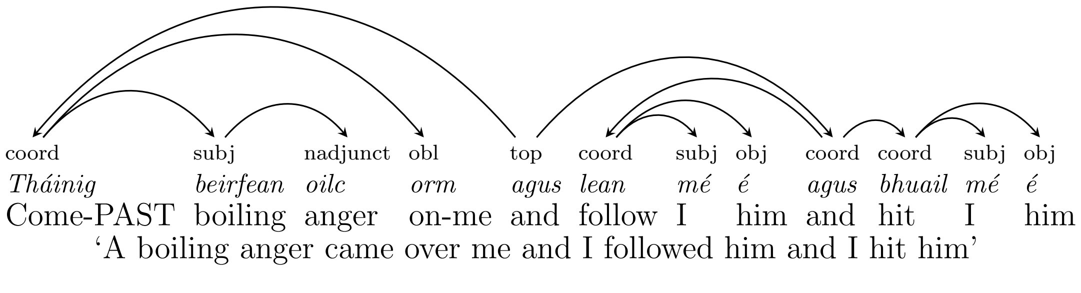

coord subj nadjunct obl top coord subj obj coord coord subj obj
Tháinig beirfean oilc orm agus lean mé é agus bhuail mé é
Come-PAST boiling anger on-me and follow I him and hit I him
‘A boiling anger came over me and I followed him and I hit him’

Figure 4.29: Dependency structure for clustered coordination

4.2.20 Punctuation as a coordinator

It is common in Irish to use a comma instead of a coordinating conjunction. In these cases, we are faced with the question of how to analyse these types of coordination as the coordinator is usually labelled with the role of its coordinates. We allow punctuation (including '/' as 'or') to be treated as a coordinating conjunction. See Figure 4.30 for example.

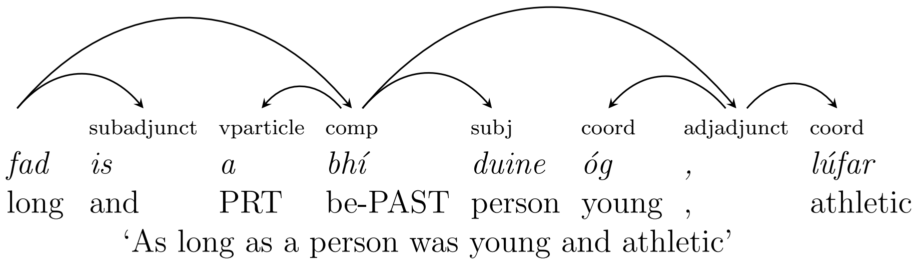

subadjunct
vparticle
comp
subj
coord
adjadjunct
coord
fad *is* *a* *bhí* *duine* *óg* *,* *lúfar*
long and PRT be-PAST person young , athletic
'As long as a person was young and athletic'

Figure 4.30: Dependency structure for coordinate punctuation

4.2.21 Wh-questions

Notwithstanding the observation that WH-questions are syntactically similar to cleft sentences (Stenson, 1981, p.107), we choose to treat them differently so that their predicate-argument structure is obvious and easily recoverable. Instead of regarding the WH-word as the head (just as the copula is the head in a cleft sentence), we instead regard the verb as the sentential head and mark the WH-element as a dependent of that verb, labelled as `subj_q`, `obj_q` or `advadjunct_q`. An example of `obj_q` is in Figure 4.31. Also note that the rest of the structure is similar to a relative clause. However, we label the *a* particle as `vparticle` to differentiate them from relative particles in regular relative clauses.

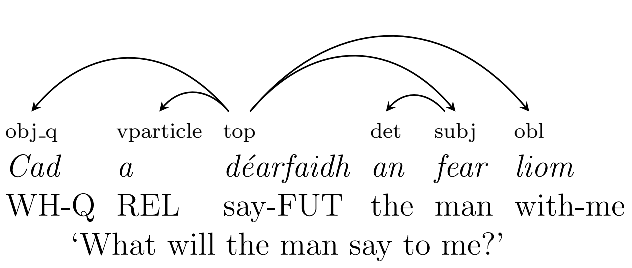

obj_q
vparticle
top
det
subj
obl
Cad *a* *déarfaidh* *an* *fear* *liom*
WH-Q REL say-FUT the man with-me
'What will the man say to me?'

Figure 4.31: Dependency structure for question construction

When there is no verb present, however, WH-elements such as *cad*, for example, should be treated as an interrogative copula. See Figure 4.32 for example.

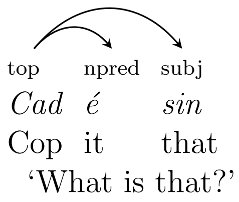


Figure 4.32: Dependency structure for question construction without a verb

4.3 Summary and Conclusion

In this chapter we have presented the final label tagset for the Irish Dependency Annotation Scheme. We show how, while the tagset is fine-grained and contains 47 labels, it is also possible to map it, if required, to a more coarse-grained set (21 labels) as a result of the hierarchical nature of some of our label types.

We have also highlighted and discussed in detail many of the linguistic phenomena in the Irish language which required significant consideration when designing the annotation scheme. We have also discussed the various studies in the literature that have helped with the analyses of many of these linguistic features.

In the next chapter, we discuss how we map this annotation scheme to two universal annotation schemes (developed subsequently to this work) for purposes of cross-lingual studies.

Chapter 5

Universal Dependencies

The hypothesis of a Universal Grammar (UG) (Chomsky, 1986) is an ongoing topic of controversy in the field of linguistics. The underlying belief of UG is that humans are born with an innate ability to understand the structure of languages and that grammar rules are hard-wired into the brain. A recent discussion by Dabrowska (2015) presents an overview of the UG linguistic approach and highlights arguments from both sides.

Those arguments put forward in favour of UG include the fact that all languages share fundamental similarities, often referred to as linguistic universals (Chomsky, 1965). In addition, children produce sentences that they have never heard before (the poverty of stimulus argument (Chomsky, 1980)), children worldwide acquire language at similar stages and in the same order, and constructions are understood by children to be ungrammatical even though rules governing this are not explicitly expressed.

However, it is also argued that while a Universal Grammar identifies a list of linguistic universals, they do not need to be used by all languages, and that newly discovered language properties are merely added to the list (Evans and Levinson, 2009). In addition, it can be noted that the worldwide age brackets in which language is primarily acquired (e.g. 9-15 months for referential words) are significantly wide in terms of child development and that some language constructions are acquired at varying stages depending on the language in question (Berman, 1985). Also, studies show that requests from adults for clarification are interpreted by children as negative feedback on ungrammatical structures (e.g. Chouinard and Clark (2003)).

The UG approach contradicts some traditional beliefs of language acquisition that are linked to behaviourism theories (e.g. Skinner (1957)), which views a learning environment of linguistic input from parents, relatives, etc., combined with responses and reinforcement as the fundamental factors influencing language learning.

It is possible to see the benefits for a universal grammar approach, without necessarily subscribing to a full alignment with the theory, however. From a cross-lingual syntactic study perspective, the notion of a universal grammar is appealing. Take research in the area of treebanking, for example. As we discussed in Chapter 3, dependency treebanks exist for many languages (e.g. Turkish (Oflazer et al., 2003), Czech (Hajič, 1998), Danish (Kromann, 2003), Slovene (Džeroski et al., 2006) and Finnish (Haverinen et al., 2010)). However, these schemes vary significantly and are independently tailored and developed according to the language in question. The labelling notations and linguistic analyses are specific to that language, and often influenced by linguistic theories that the developers subscribe to. Thus, cross-lingual research is often hampered by variations that exist across the annotation schemes of treebanks. Comparison of treebanks is more effective if a universal grammar is used to overcome these variations.

More specifically from a parsing point of view, if the labelled training data for both languages is based on different annotation schemes, parser output in one language cannot be easily compared to another. It is also impossible to accurately compare the difficulty between parsing different languages for a number of reasons. Parsing results can be influenced by annotation scheme discrepancies such as the number of labels in a scheme that a parser has to choose from, or varying linguistic analyses across languages, for example. McDonald et al. (2011) demonstrated these difficulties through a cross-lingual parsing study. Cross-lingual transfer parsing involves training a parser on one language, and parsing data of another language.

Their work describes two types of cross-lingual parsing: (i) direct transfer parsing in which a delexicalised version of the source language treebank is used to train a parsing model, which is then used to parse the target language, and (ii) a more complicated projected transfer approach in which the direct transfer approach is used to seed a parsing model which is then trained to obey source-target constraints learned from a parallel corpus.

These experiments evaluated the use of delexicalised parsers (using non-lexical features only) trained on a number of source languages and tested on a set of target languages. Both direct transfer and projected transfer (using parallel treebanks) parsing experiments were carried out. The results revealed that languages that were typologically similar were not necessarily the best source-target pairs, sometimes due to variations between their language-specific annotation schemes.

There has been a movement towards identifying linguistic universals in an attempt to facilitate cross-lingual studies. Petrov et al. (2012) designed what is referred to as the Google Universal POS-tagset, which identifies 12 main part-of-speech tags, aimed at covering the most frequent categories that exist in any language. By mapping 25 different language datasets to their universal set, they were able to accurately compare POS-tagger performance across languages.

In parsing, McDonald et al. (2013) built upon this work to develop a set of Universal Dependency Treebanks, which were based on their design of a universal standard for annotating treebanks. They defined a set of 41 dependency labels and a universal annotation scheme which they use to convert 10 language treebanks for the purpose of a cross-lingual parsing study. They reported improved results on cross-lingual direct transfer parsing using their uniformly annotated treebanks. While their results confirm that parsers trained on data from languages in the same language group (e.g. Romance and Germanic) show the most accurate results, they also show that training data taken across language-groups also produces promising results, confirming that some language features are certainly common across language groups.

More recently however, a new Universal Dependency project (Nivre et al., 2015; Agić et al., 2015) released data for 10 treebanks that have been converted according to a new universal scheme. This new annotation scheme is based on (universal) Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008).

For the purposes of cross-lingual studies, and to compare Irish linguistic and NLP resources to that of other languages, we mapped the Irish Dependency Treebank (IDT) to both universal schemes described above. For the rest of this thesis we will refer to the universal dependency scheme of McDonald et al. (2013) as UD13, and to the project of Nivre et al. (2015) as UD15. Due to the fact that Irish is officially part of the UD15 project, and it now supersedes UD13, we provide more detail in the discussion of this conversion. This chapter explains our mapping from the IDT to both of these schemes. The associated experiments are presented later in Section 6.7.

5.1 A Universal Dependency Scheme for the Irish Dependency Treebank

In this section, we describe how we created a "universal" version of the Irish Dependency Treebank. We do this by first mapping the Irish POS tagset to the Google Universal POS tags, and then mapping the original dependency scheme to McDonald et al. (2013)'s universal dependency scheme (UD13).

5.1.1 Mapping the Irish POS tagset to the Universal POS tagset

The Google Universal POS tagset (Petrov et al., 2012) was designed to facilitate unsupervised and cross-lingual part-of-speech tagging and parsing research, by simplifying POS tagsets and unifying them across languages. The Irish Dependency Treebank was built upon a POS-tagged corpus developed by Uí Dhonnchadha and van Genabith (2006), and is based on the PAROLE Morphosyntactic Tagset (ITÉ, 2002). The treebank's tagset contains both coarse- and fine-grained POS tags which we map to the Universal POS tags (e.g. Prop Noun → NOUN). Table 5.1 shows the mappings.

Part-of-speech (POS) mappings			
Univ.	**Irish**	**Univ.**	**Irish**
NOUN	Noun Noun, Pron Ref, Subst Subst, Verbal Noun, Prop Noun	ADP	Prep Deg, Prep Det, Prep Pron, Prep Simp, Prep Poss, Prep CmpdNoGen, Prep Cmpd, Prep Art, Pron Prep
PRON	Pron Pers, Pron Idf, Pron Q, Pron Dem	ADV	Adv Temp, Adv Loc, Adv Dir, Adv Q, Adv Its, Adv Gn
VERB	Cop Cop, Verb PastInd, Verb PresInd, Verb PresImp, Verb VI, Verb VT, Verb VTI, Verb PastImp, Verb Cond, Verb FutInd, Verb VD, Verb Imper	PRT	Part Vb, Part Sup, Part Inf, Part Pat, Part Voc, Part Ad, Part Deg, Part Comp, Part Rel, Part Num, Part Cp
DET	Art Art, Det Det	NUM	Num Num
ADJ	Prop Adj, Verbal Adj, Adj Adj	X	Item Item, Abr Abr, CM CM, CU CU, CC CC, Unknown Unknown, Guess Abr, Itj Itj, Foreign Foreign
CONJ	Conj Coord, Conj Subord	.	. . .. ? ? ! ! : : ? . Punct Punct

Table 5.1: Mapping of Irish Coarse and Fine-grained POS pairs (coarse fine) to Universal POS tagset.

Most of the POS mappings made from the Irish POS tagset to the universal tagset are intuitive. However, some decisions require explanation.

Cop → VERB As we have discussed in Section 4.2.2, there are two verbs 'to be' in Irish: the substantive verb *bí* and the copula *is*. For that reason, the Irish POS tagset differentiates the copula by using the POS tag `Cop`. In the literature on Irish syntax, there is some discussion over its syntactic role, whether it is a verb or a linking particle (Carnie, 1997). The role normally played is that of a linking element between a subject and a predicate. However, our syntactic analysis of the copula is in line with that of Stenson (1981), regarding it as a verb. In addition, because the copula is often labelled in the Irish annotation scheme as the syntactic head of the matrix clause, we have chosen `VERB` as the most suitable mapping for this part of speech.

Pron Prep → ADP *Pron Prep* is the Irish POS tag for pronominal prepositions, which are also referred to as prepositional pronouns (see Section 4.2.8). Characteristic of Celtic languages, they are prepositions inflected with their pronominal objects – compare, for example, ***le*** *mo chara* 'with my friend' with ***le****is* 'with him'. While the Irish POS labelling scheme labels them as pronouns in the first instance (i.e. their coarse grained tag), our dependency labelling scheme treats the relationship between them and their syntactic heads as `obl` (obliques) or `padjunct` (prepositional adjuncts). Therefore, we map them to `ADP` (adpositions).

It should be noted that when a fine-grained POS tagset is mapped to a coarse-grained POS tagset, information is often lost. This would generally be the case for most (if not all) languages mapped to this Universal POS tagset. Some fine-grained POS tags carry more important information than others, and when they are subsumed by coarse-grained tags, the meaning and relationship between tokens and structure is often lost. The following are some examples of this type of loss:

(i) `Cop` → `VERB`: the linguistic structure of copula constructions differ significantly from regular verbal constructions. By mapping all copula forms to 'Verb' this distinction is not easily recoverable and is also likely to generate noise for verbal patterns within the training data.
(ii) `Prop Noun` → `NOUN`: our treebank does not mark semantic properties of tokens.

Proper nouns often have different semantic properties (e.g. person/ organisation) than regular common nouns. By subsuming proper nouns by a tag that also covers common nouns, the inherent semantic restrictions associated with certain verbal use are lost.

(iii) `Verbal Adj` → `ADJ`: as discussed in Section 4.2.15, Irish passive constructions are expressed as stative passives using verbal adjectives. When the verbal adjective POS tag is subsumed by the general adjective tag, information regarding the 'passive' nature of these constructions is lost. This is further compounded by the loss of the `obl_ag` (oblique agent) dependency label, as shown in Table 5.2.

(iv) `Verbal Noun` → `NOUN`: As discussed in Section 2.2.1, the Irish language uses a verbal noun to denote an infinitive verb form and progressive aspectual phrases. By mapping verbal nouns to the general noun POS tag, this verbal nature is lost.

5.1.2 Mapping the Irish Dependency Scheme to the Universal Dependency Scheme (UD13)

The departure point for the design of the Universal Dependency (UD13) Treebanks (McDonald et al., 2013) was the Stanford typed dependency scheme (de Marneffe and Manning, 2008), which was adapted based on a cross-lingual analysis of six languages: English, French, German, Korean, Spanish and Swedish. As a result of this study, universal dependency treebanks were developed initially for these six languages, followed by subsequent development of UD treebanks for five languages (Brazilian Portuguese, Finnish[1], Indonesian, Italian and Japanese).[2] Approaches to development of these treebanks varied. Existing English and Swedish treebanks were automatically mapped to the new universal scheme. The rest of the treebanks were developed manually to ensure consistency in annotation. The study also reports some structural changes (e.g. Swedish treebank coordination structures).[3]

[1] The Finnish data was not available at the time of our experiments.

[2] Version 2 data sets downloaded from `https://code.google.com/p/uni-dep-tb/`

[3] There are two versions of the annotation scheme: the *standard* version (where copulas and adpositions are syntactic heads), and the *content-head* version which treats content words as syntactic

There are 41 dependency relation labels to choose from in the universal annotation scheme.[4] McDonald et al. (2013) use all labels in the annotation of the German and English treebanks. The remaining languages use varying subsets of the label set. In our study we map the Irish dependency annotation scheme to 30 of the universal labels. The mappings are given in Table 5.2.

UD13 Dependency Label Mappings			
Universal	**Irish**	**Universal**	**Irish Label**
root	top	*csubj*	csubj
acomp	adjpred, advpred, ppred	*dep*	for
adpcomp	comp	*det*	det, det2, dem
adpmod	padjunct, obl, obl2, obl_ag	*dobj*	obj, vnobj, obj_q
adpobj	pobj	*mark*	subadjunct
advcl	comp	*nmod*	addr, nadjunct
advmod	adjunct, advadjunct, quant, advadjunct_q	*nsubj*	subj, subj_q
amod	adjadjunct	*num*	quant
appos	app	*p*	punctuation
attr	npred	*parataxis*	comp
aux	toinfinitive	*poss*	poss
cc	NEW	*prt*	particle, vparticle, nparticle, advparticle, vocparticle, particlehead, cleftparticle, qparticle, aug
ccomp	comp	*rcmod*	relmod
compmod	nadjunct	*rel*	relparticle
conj	coord	*xcomp*	xcomp

Table 5.2: Mapping of Irish Dependency Annotation Scheme to UD13 Annotation Scheme

As with the POS mapping discussed in Section 5.1.1, mapping the Irish dependency scheme to the universal scheme was relatively straightforward, due in part, perhaps, to a similar level of granularity suggested by the similar label set sizes (Irish 47; standard universal 41). That said, there were significant considerations made in the mapping process, which involved some structural change in the treebank and the introduction of more specific analyses in the labelling scheme. These are discussed below.

heads. We are using the *standard* version.

[4] The vmod label is used only in the content-head version.

5.1.2.1 Structural Differences

The following structural changes were made manually before the dependency labels were mapped to the universal scheme.

coordination The most significant structural change made to the Irish treebank was an adjustment to the analysis of coordination. The original Irish Dependency Treebank subscribes to the LFG coordination analysis, where the coordinating conjunction (e.g. *agus* 'and') is the head, with the coordinates as its dependents, labelled `coord` (see Figure 5.1 and refer to Section 3.1.2 for further discussion). The Universal Dependency Annotation scheme, on the other hand, uses right-adjunction, where the first coordinate is the head of the coordination, and the rest of the phrase is adjoined to the right, labelling coordinating conjunctions as `cc` and the following coordinates as `conj` (Figure 5.2).

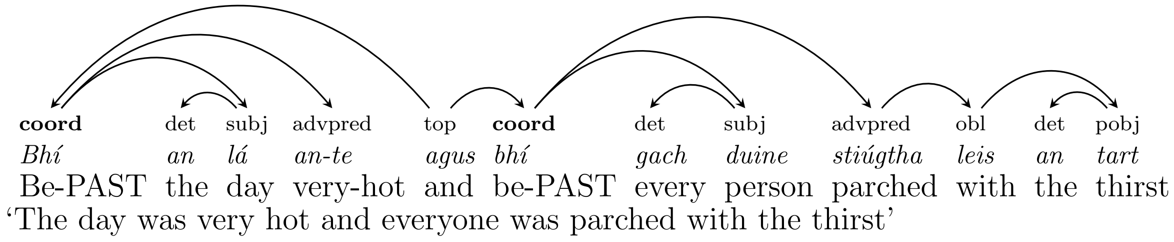

coord det subj advpred top **coord** det subj advpred obl det pobj
Bhí an lá an-te agus bhí gach duine stiúgtha leis an tart
Be-PAST the day very-hot and be-PAST every person parched with the thirst
'The day was very hot and everyone was parched with the thirst'

Figure 5.1: LFG-style coordination of original Irish Dependency Treebank.

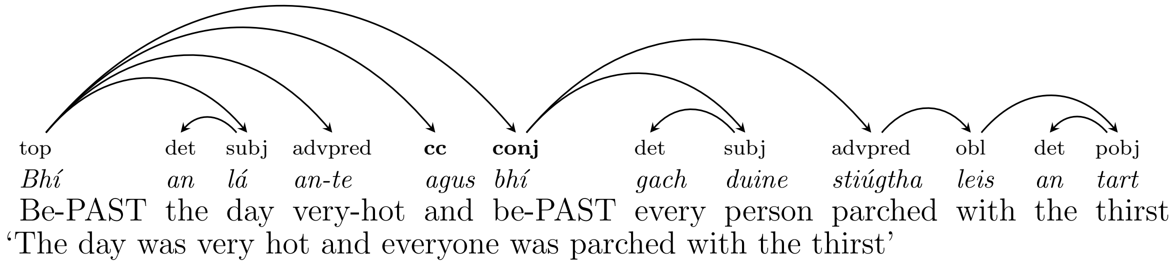

top det subj advpred **cc** **conj** det subj advpred obl det pobj
Bhí an lá an-te agus bhí gach duine stiúgtha leis an tart
Be-PAST the day very-hot and be-PAST every person parched with the thirst
'The day was very hot and everyone was parched with the thirst'

Figure 5.2: Stanford-style coordination changes to original Irish Dependency Treebank.

subordinate clauses In the Irish Dependency Treebank, the link between a matrix clause and its subordinate clause is similar to that of LFG: the subordinating conjunction (e.g. *mar* 'because', *nuair* 'when') is a `subadjunct` dependent of the matrix verb, and the head of the subordinate clause is a `comp` dependent of the subordinating conjunction (Figure 5.3). In contrast, the universal scheme is in line with the Stanford analysis of subordinate clauses, where the head of the clause is dependent on the matrix verb, and the subordinating conjunction is a dependent of the clause head (Figure 5.4).

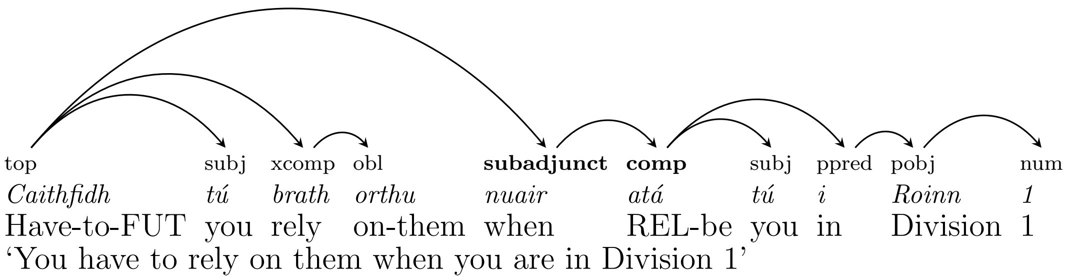


Figure 5.3: LFG-style subordinate clause analysis (with IDT labels)

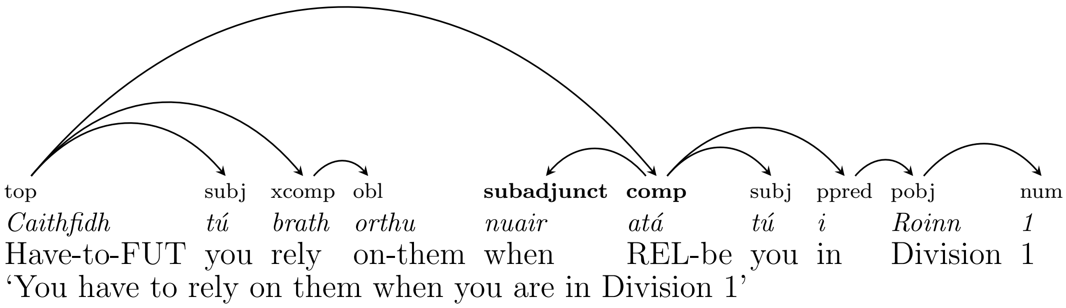


Figure 5.4: Stanford-style subordinate clause analysis (with IDT labels)

5.1.2.2 Differences between dependency types

We found that the original Irish scheme makes distinctions that the universal scheme does not – this finer-grained information takes the form of the following Irish-specific dependency types: `advpred`, `ppred`, `subj_q`, `obj_q`, `advadjunct_q`, `obl`, `obl2`. In producing the universal version of the treebank, these Irish-specific dependency types are mapped to less informative universal ones (see Table 5.2). Conversely, we found that the universal scheme makes distinctions that the Irish scheme does not. Some of these dependency types are not needed for Irish. For example, there is no indirect object `iobj` in Irish, nor is there a passive construction that would require the labels `nsubjpass`, `csubjpass` or `auxpass`. Also, in the Irish Dependency Treebank, the copula is usually the root (`top`) or the head of a subordinate clause (e.g. `comp`) which renders the universal type `cop` redundant. Others that are not used are `adp`, `expl`, `infmod`, `mwe`, `neg`, `partmod`. However, we did identify some dependency relationships in the universal scheme that we introduce to the UD13 Irish Dependency Treebank (`adpcomp, adposition, advcl, num, parataxis`). These are explained below.

comp → adpcomp, advcl, parataxis, ccomp The following new mappings were previously subsumed by the IDT label `comp` (complement clause). The mapping for *comp* has thus been split between *adpcomp*, *advcl*, *parataxis* and *ccomp*.

- `adpcomp` is a clausal complement of an adposition. An example from the English data is 'some understanding of what the company's long-term horizon should **begin** to look like', where 'begin', as the head of the clause, is a dependent of the preposition 'of'. An example of how we use this label in Irish is in Figure 5.5.

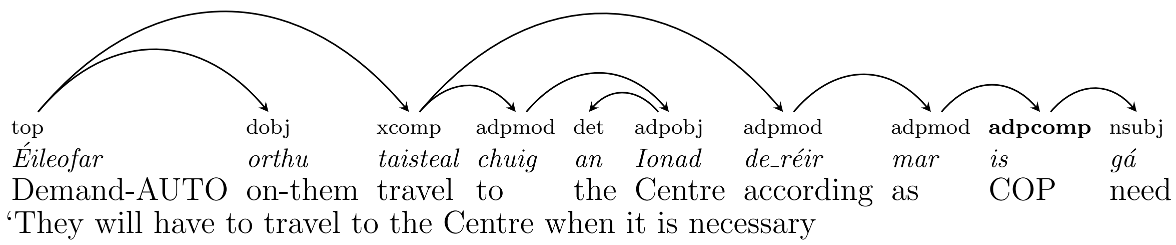
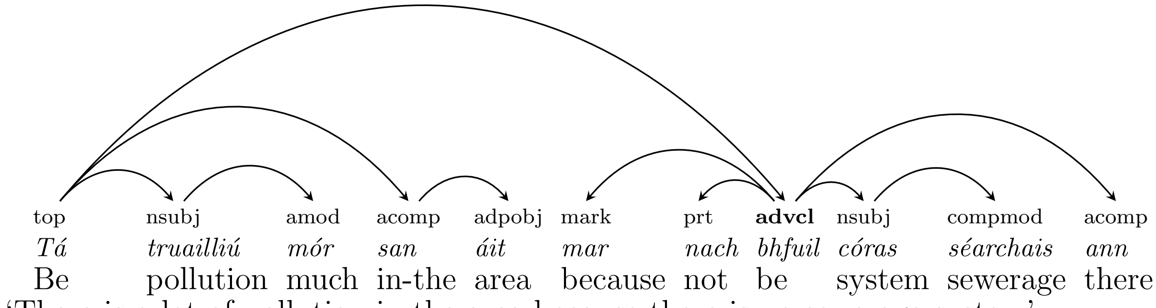


Figure 5.5: UD13 `adpcomp` complement clause

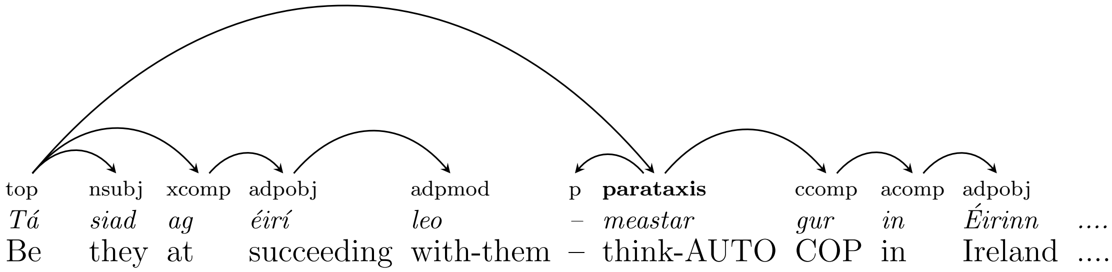
- `advcl` is used to identify adverbial clause modifiers. In the English data, they are often introduced by subordinating conjunctions such as 'when', 'because', 'although', 'after', 'however', etc. An example is 'However, because the guaranteed circulation base is being **lowered**, ad rates will be higher'. Here, 'lowered' is an advcl dependent of 'will'. Equivalent subordinating conjunctions in Irish are *mar* 'because', *nuair* 'when', *cé* 'although', for example. An example of Irish usage is given in Figure 5.6.


top nsubj amod acomp adpobj mark prt **advcl** nsubj compmod acomp
Tá truailliú mór san áit mar nach bhfuil córas séarchais ann
Be pollution much in-the area because not be system sewerage there

'There is a lot of pollution in the area because there is no sewerage system'

Figure 5.6: UD13 advcl adverbial clause modifier

- parataxis labels clausal structures that are separated from the previous clause with punctuation such as – ... : () ; and so on. See Figure 5.7 for example.


top nsubj xcomp adpobj adpmod p **parataxis** ccomp acomp adpobj
Tá siad ag éirí leo – meastar gur in Éirinn ....
Be they at succeeding with-them – think-AUTO COP in Ireland ....

'They are succeeding - it is believed that in Ireland...'

Figure 5.7: UD13 parataxis clauses

- ccomp covers all other types of clausal complements. For example, in English, 'Mr. Amos says the Show-Crier team will probably **do** two live interviews a day'. The head of the complement clause here is 'do', which is a comp dependent of the matrix verb 'says'. An Irish example is given in Figure 5.8.

quant → num, advmod The IDT Scheme uses one dependency label (quant) to cover all types of numerals and quantifiers. We now use two labels from the

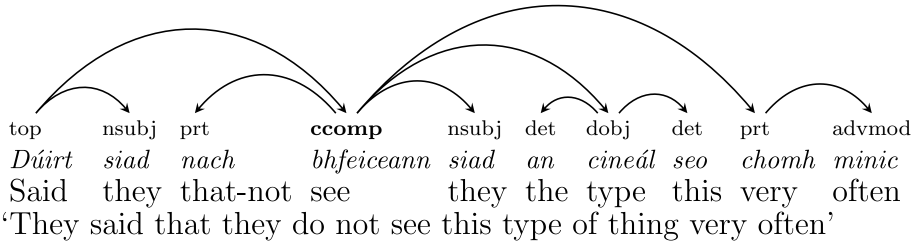

top nsubj prt **ccomp** nsubj det dobj det prt advmod
Dúirt siad nach bhfeiceann siad an cineál seo chomh minic
Said they that-not see they the type this very often
'They said that they do not see this type of thing very often'

Figure 5.8: UD13 `ccomp` clausal complements

universal scheme to differentiate between quantifiers such as *mórán* 'many' (`advmod`) and numerals such as *fiche* 'twenty' (`num`).

nadjunct → nmod, compmod The IDT label `nadjunct` accounts for all nominal modifiers. However, in order to map to the universal scheme, we discriminate two kinds: (i) nouns that modify clauses are mapped to `nmod` (e.g. ***bliain*** *ó shin* 'a **year** ago') and (ii) nouns that modify nouns (usually genitive case in Irish) are mapped to `compmod` (e.g. *plean* ***margaíochta*** '**marketing** plan').

As with the POS tag mapping, information is also lost through the dependency label mapping process. This is because some UD labels are too general to describe fully the nature of the dependency relation between tokens. This is clear from the numerous 'many-to-one' mappings shown in Table 5.2. Some of the lost information is explained in more detail here:

(i) (`adjpred, advpred, ppred`) → `comp`; `npred` → `attr`: all of these predicate arguments are used in a similar pattern. By separating the nominal predicate mapping to a separate label, the common behaviour and linguistic patterns across all predicates are lost.

(ii) `obl_ag` → `adpmod`: by subsuming the oblique agent label under a general modifier label, the 'passive' function of these oblique modifiers and the nature of the structures they describe are lost.

(iii) `prt`: the universal 'particle' label subsumes all of the fine-grained Irish particles (apart from relative particles). Particles are a significant feature of the Irish language, carrying out many functions, and their distinction can help semantic disambiguation of some forms (e.g. *a* can be a vocative, quantifier, cleft and time particle).

It is worth noting, however, that in the more recent Universal Dependency Scheme (see Section 5.2.2), it is possible to account for the loss of information in these (and other mappings) through the use of language-specific sub-labels.

5.2 A new Universal Dependency Scheme (UD15)

The following is a summary of the conversion and mapping of the IDT to a new universal scheme, as part of the Universal Dependency Project (Nivre et al., 2015; Agić et al., 2015).

This scheme aims to give greater consideration to the varying linguistic differences across languages, and provides the option of defining language-specific label sub-types where necessary. In October 2014, guidelines for this new scheme were released to allow for mappings and conversions of existing treebanks. Ten converted treebanks were released in January 2015. The languages included in this release were Czech, English, Finnish, French, German, Hungarian, Irish, Italian, Spanish and Swedish.

Below we report on the work we carried out on the Irish dataset for this release in converting the IDT to an updated universal POS tagset and a new universal annotation scheme (UD15).

5.2.1 New Universal POS tagset

There are 17 tags in the UD15 Universal POS tagset. We provide a mapping from the Irish PAROLE tagset to the UD15 tagset in Table 5.3. Tags that were not in the Google tagset are bolded, and tag name changes (compared to the Google tagset) are marked with †. We discuss the changes below.

The UD15 Universal POS tagset has 17 POS-tags (compared to 12 tags in the

Part-of-speech (POS) mappings			
New Univ.	**Irish**	**New Univ.**	**Irish**
NOUN	Noun Noun, Pron Ref, Subst Subst, Verbal Noun,	ADP	Prep Deg, Prep Det, Prep Pron, Prep Simp, Prep Poss, Prep CmpdNoGen, Prep Cmpd, Prep Art, Pron Prep
PROPN	Prop Noun	ADV	Adv Temp, Adv Loc, Adv Dir, Adv Q, Adv Its, Adv Gn
PRON	Pron Pers, Pron Idf, Pron Q, Pron Dem	PART†	Part Vb, Part Sup, Part Inf, Part Pat, Part Voc, Part Ad, Part Deg, Part Comp, Part Rel, Part Num, Part Cp,
VERB	Cop Cop, Verb PastInd, Verb PresInd, Verb PresImp, Verb VI, Verb VT, Verb VTI, Verb PastImp, Verb Cond, Verb FutInd, Verb VD, Verb Imper	NUM	Num Num
DET	Art Art, Det Det	X	Item Item, Abr Abr, CM CM, CU CU, CC CC, Unknown Unknown, Guess Abr, Foreign Foreign
ADJ	Prop Adj, Verbal Adj, Adj Adj	PUNCT†	. . .. ? ? ! ! : : ? . Punct Punct
CONJ	Conj Coord	**INTJ**	Itj Itj
SCONJ	Conj Subord	**SYM**	(Abr)

Table 5.3: Mapping of Irish Coarse and Fine-grained POS pairs (coarse fine) to 2015 Universal Dependency POS tagset.

Google Universal POS tagset). However, we only map to 16 of these tags as there are no auxiliary verbs in Irish to require the inclusion of `AUX`. Some naming differences between the universal tagsets include `PRT` (Google) → `PART` (UD15) as a particle tag, and `.` (Google) → `PUNCT` (UD15) as a punctuation tag. The following is a summary of new POS tags introduced in the UD15 set:[5]

`PROPN`: proper nouns Proper nouns were subsumed by the `NOUN` tag in the Google POS tagset. A proper noun is a noun that is the name of an individual, place, object

[5] POS descriptions given here are adaptations of the UD annotation guidelines. `http://universaldependencies.github.io/docs/u/pos/all.html`

or organisation. In Irish, proper nouns always have initial capitalisation. Days of the week and months of the year, while capitalised, are marked as common nouns. Personal names are treated as a sequence of proper nouns. Examples of proper nouns include:

- *Lá 'le Pádraig* 'St. Patrick's Day'
- *Eoraip* 'Europe'

INTJ: **interjections** Interjections were subsumed by the catch-all X tag in the Google POS tagset. An interjection is a word that is used most often as an exclamation or part of an exclamation.

- *Och* 'but'/ 'aw'
- *á* 'aw'
- *Ó* 'Oh'

SCONJ: **subordinating conjunctions** This new tagset differentiates coordinating conjunctions CONJ and subordinating conjunctions SCONJ, rather than using one tag (CONJ) for both. In Irish, some subordinate conjunctions, such as *nuair* 'when' in Example 32 normally precede a subordinate clause marker (*a*). Subordinate conjunctions link clauses by making one of them a constituent of the other. There is also a special case of using *agus* 'and' (normally POS-tagged as CONJ) as a subordinate conjunction, where the subordinate clause is missing a surface verb 'to be' (see Example 33).

(32) Tháinig sí ar_ais **nuair** a chuala sí an nuacht
Came she back **when** PART heard she the news
'She came back when she heard the news'

(33) Seo pictúir a tógadh dhó **agus** é briste
This picture REL taken-AUTO to-it **and** it broken
'Here's a picture taken of it broken'

`SYM`: symbols A symbol is a word-like entity that differs from ordinary words by form, function, or both. Due to the domain type of our treebank data, there is currently only one example of a token tagged as `SYM`: post@clubsult.com (email address). This is tagged as `Abr` (abbreviation) according to the PAROLE tagset, however it does not hold that all `Abr` instances should be mapped to `SYM`. These cases need to be identified on an individual basis. According to the UD guidelines, other examples would include:

- $, %, §, ©,
- $+, -, \times, \div, =, <, >$,
- ☺, post@clubsult.com

5.2.2 2015 Universal Dependency Scheme

As we have shown in Table 5.3 (through the use of bolding), there were minimal differences between the UD13 and UD15 Universal POS tagsets. This resulted in a relatively easy mapping from the Irish PAROLE tagset to the new UD15 Universal POS tagset.

However, there are significant differences between the UD13 and the UD15 annotation schemes. Thus, the IDT to UD15 treebank conversion required extensive additional work on dependency relation renaming, mapping and new structural changes. Here, we focus on a mapping from the Irish Dependency scheme to the UD15 dependency labels. We provide a mapping in Table 5.4 and describe the changes below.

UD15 Dependency Label Mappings			
2015 Universal	**Irish**	**2015 Universal**	**Irish**
root	top	*foreign*	for
acl:relcl	relmod	*list*	quant
advcl	comp †	*mark*	subadjunct, toinfinitive
advmod	adjunct †, advadjunct, advadjunct_q, quant †	*mark:prt*	advparticle, cleftparticle, particle, qparticle, vparticle
amod	adjadjunct	*name* ±	nparticle, nadjunct †
appos	app	*neg*	vparticle
case ±	padjunct, obl_ag	*nmod*	aug, pobj †±
case:voc	vocparticle	*nmod:poss*	poss
cc ±	–	*nmod:prep*±	obl, obl2
ccomp	comp †	*nmod:tmod*	advadjunct, padjunct †, pobj †, relparticle †
compound	nadjunct	*nsubj*	relparticle †, subj, subj_q
compound:prt	particlehead	*nummod*	quant
conj ±	coord	*parataxis*	comp †
cop ±	NEW	*punct*	punctuation
csubj:cop	csubj	*vocative*	addr
det	det, det2, dem	*xcomp*	xcomp
discourse	adjunct †	*xcomp:pred*	adjpred, advpred, npred, ppred ±
dobj	obj, vnobj, obj_q		

Table 5.4: Mapping of Irish Dependency Annotation Scheme to UD15 Annotation Scheme. † marks one-to-many mappings, and ± marks structural changes

5.2.2.1 UD labels not used in the Irish UD Treebank

The following is a list of labels in the UD15 annotation scheme that do not apply to the Irish language.

- `aux`: This label is used for non-main verbs in a clause, i.e. auxiliary verbs. Examples in English are '*has* opened', '*will* be', '*should* say'. There are no equivalent auxiliary verbs in Irish.[6]

- `auxpass, nsubjpass, csubjpass`: These labels are used in passive constructions, respectively as: passive auxiliary verbs, passive nominal subjects and clausal passive subjects. There is no passive form in Irish.[7]

[6] Stenson (1981, p.86) notes that modal verbs such as *caithfidh* inflect as per regular verbs and are considered the main verb.

[7] See Section 4.2.14 for further discussion.

- `iobj`: In English, an example is ‘Mary gave *John* the book’. There are no indirect objects in Irish, and constructions like these must follow the normal ditransitive verb structure using a preposition, as per Example 34.

  (34) Thug Máire an leabhar do Sheán
  Gave Mary the book to John
  ‘Mary gave the book to John’

There are also some labels that are not used in our treebank, due to lack of instances observed in the data. The reason for this may be related to the well-structured, grammatically correct nature of the text in our corpus (newswire, legal documents, literature).

- `reparandum`: This label is used to indicate disfluencies in text. The Irish data does not contain any disfluencies.
- `goeswith`: This label links to parts of a word that has been split, due to poor editing. There are no instances of this in the Irish data.
- `dep`: This catch-all label is used for unknown relations. We do not require this in the Irish data.

In addition, there are some Universal labels that we have not included in the first release version of this treebank, but which we expect will be included in future releases.

- `expl`: There is no existential ‘there’ in Irish. However, we have not yet fully researched uses of other types of expletives in our data.
- `mwe`: Multiword expressions are not marked in the IDT. There is not sufficient linguistic literature on this topic for Irish on which we could base a complete analysis of idioms or multiword units in the treebank. This analysis therefore remains as a possible future enhancement to the treebanks when such resources are available.

- `remnant`: This label is used for remnants in ellipsis, where a predicate or verb is dropped (e.g. 'Marie went to Paris and Miriam [] to Prague'). Instances of remnants in Irish are not easily identified. Further study is required to identify cases, if any, including a possible analysis of crossing dependencies.
- `dislocated`: This label is used for fronted or postposed elements that are not core grammatical elements of a sentence. Example, 'he must not eat it, *the playdough*'. We have not yet identified such cases in the Irish data.

5.2.2.2 Manual label updates

Some of the treebank conversion was automated with straightforward mappings. However, there were a number of labels that needed to be manually mapped because they were one-to-many label mappings. These instances are marked with † in Table 5.4.

relative particles In the IDT, the relative particle *a* was attached to a relative modifier verb with the label `relparticle`. In the UD15 scheme, this particle is labelled with the syntactic role it plays in the relative clause – a type of annotation that cannot be automated in the absence additional data on the semantic properties of the element the relativised refers to. The *a* can therefore fulfil the role of `nsubj`, `dobj`, `nmod` or `nmod:tmod`. For example *an rud deireanach* **a** *chonaic sé* 'the last thing that he saw' is shown in Figure 5.9. In this case *a* refers to *rud* 'thing', and therefore is labelled as a `dobj` of *chonaic* 'saw'. See Figure 5.10 for an example of nominal relativiser.

quant → nummod, list, advmod Similar to the UD13 mapping, numerals and quantifiers are given more fine-grained descriptions in UD15 than the single IDT `quant` label. In addition, list numbering is now represented by its own label `list`.

comp → advcl, ccomp, parataxis The tokens labelled in the IDT with the closed complement label `comp` have been divided among three new labels. Note

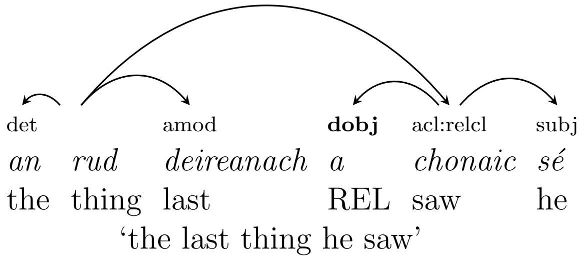

det
amod
dobj
acl:relcl
subj
an *rud* *deireanach* *a* *chonaic* *sé*
the thing last REL saw he
'the last thing he saw'

Figure 5.9: UD15 `dobj` relative particle analysis

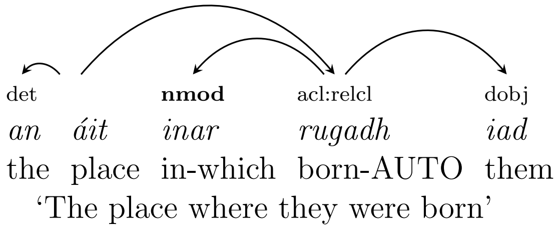

det
nmod
acl:relcl
dobj
an *áit* *inar* *rugadh* *iad*
the place in-which born-AUTO them
'The place where they were born'

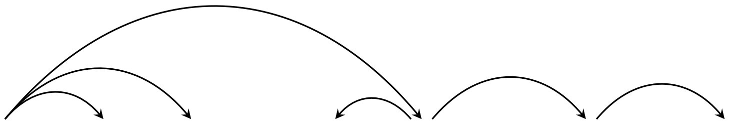
Figure 5.10: UD15 `nmod` relative particle analysis

that unlike UD13, there is no `adpcomp` label in this scheme. The UD15 labels are: `advcl` adverbial clause (normally connected with a subordinator such as *nuair* 'when', *má* 'if' etc); `ccomp` complement clauses that are normally introduced by the complementiser *go*, *nach*, *gur*, or quoting direct speech; `parataxis` labels two phrases or sentences set side-by-side without explicit linking through coordination or subordination, for example. Sometimes punctuation such as colons or semicolons connects the pairs. *Sa tseanam, bhí an cál an-ghann*; *b'fheidir nach mbeadh i ngach baile ach aon gharraí amháin.* 'In olden times, kale was very scarce**;** maybe there would only be one garden in every town'.

nadjunct → compound, name The compound label is used for nominal modifiers. In Irish this could take the form of compounding (one noun modifying another) such as *deireadh seachtaine* 'weekend' (Example 35), or ownership *teach Mhicil* 'Michael's house' (Example 36). Compounding can occur with a string of nouns as per the example in Figure 5.11.

(35) deireadh seachtaine
end week
'weekend'

(36) teach Mhichil
house Michael
'Michael's house'



root	nsubj	advmod	nummod	dobj	**compound**	**compound**
Chaill	*sí*	*beagnach*	*ocht*	*mbliana*	*riaráistí*	*pinsin*
Lost	she	almost	eight	years	arrears	pension

'She lost almost eight years of pension arrears'

Figure 5.11: UD15 compounding analysis

The new label `name` is explained below in more detail in Section 5.2.2.3.

5.2.2.3 Structural Changes

Other labels required a manual annotation because they related to structural changes required in the treebank. Along with the structural changes required for coordination and subordination during the UD13 conversion, as described in Section 5.1.2.1, additional structural changes were required in the UD15 scheme mapping. These changes are described below:

cop In the IDT, the copula is treated similarly to a verb, and can function as the root of a sentence, or as the head of a dependency clause. However, the UD15 scheme analyses copula constructions differently. Instead, the predicate is regarded as the head of the phrase, and the copula is its dependent, as indicated by the `cop` label. This also applies to copula use in fronting or cleft structures. See Figure 5.12 and Figure 5.13 for comparison.[8]

[8] The labels have also been mapped between examples, but the structural change is of interest here.

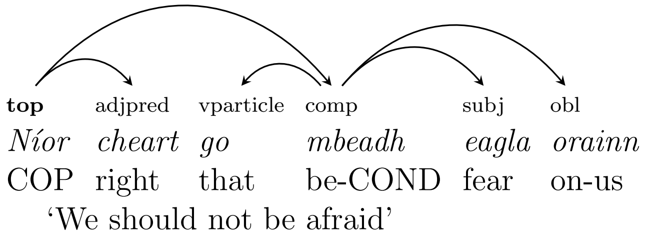

top adjpred vparticle comp subj obl
Níor cheart go mbeadh eagla orainn
COP right that be-COND fear on-us
'We should not be afraid'

Figure 5.12: IDT copula analysis

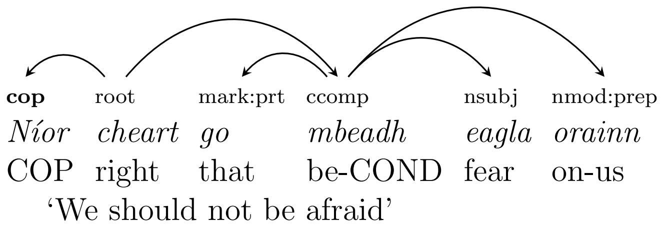

cop root mark:prt ccomp nsubj nmod:prep
Níor cheart go mbeadh eagla orainn
COP right that be-COND fear on-us
'We should not be afraid'

Figure 5.13: UD15 copula analysis

name The UD relation `name` is used with compounding proper nouns, typically for names of people, places, organisations and so on. In Irish, this not only includes surnames, but also surname particles such as *Mac, Mc, Ó, de, Uí* and *Ní*. In the Irish Dependency Treebank, we identify the surname as the head noun, and its dependents can either be first names (`nadjunct`) or nominal particles (`nparticle`). See Figures 5.14 for example. However in the UD15 analysis, all words in the name phrase modify the first one as `name`. See Figure 5.15 for comparison.

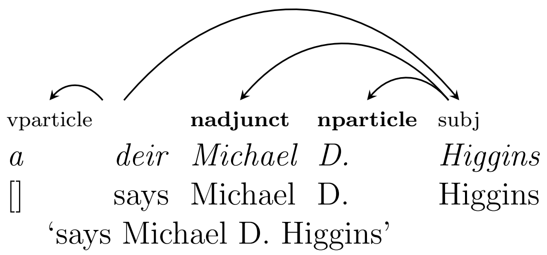

vparticle **nadjunct** **nparticle** subj
a deir Michael D. Higgins
[] says Michael D. Higgins
'says Michael D. Higgins'

Figure 5.14: IDT name analysis

nmod, case, xcomp:pred The head of a prepositional phrase has changed from the preposition to the head noun of the object noun phrase. This affects the Irish treebank in a number of ways:

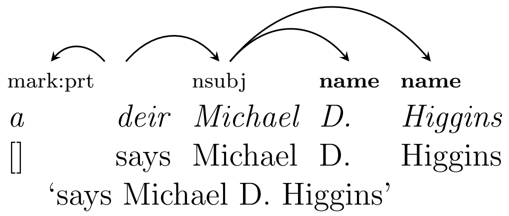


Figure 5.15: UD15 name analysis

In the UD15 analysis, the head of regular preposition phrases (object of the preposition) is attached to the verb as `nmod` (formerly `pobj` in IDT). The preposition is a dependent of the object, and this relation is labelled as `case`. Compare Figures 5.16 and 5.17 to observe the difference in analyses.

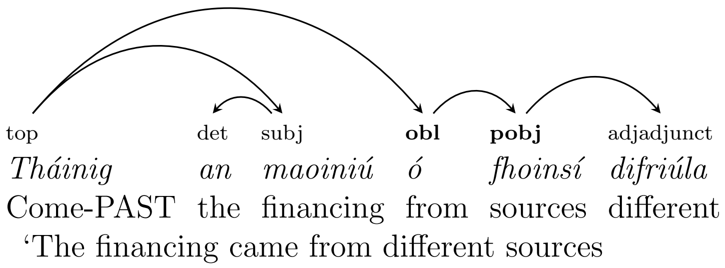


Figure 5.16: IDT prepositional phrase analysis

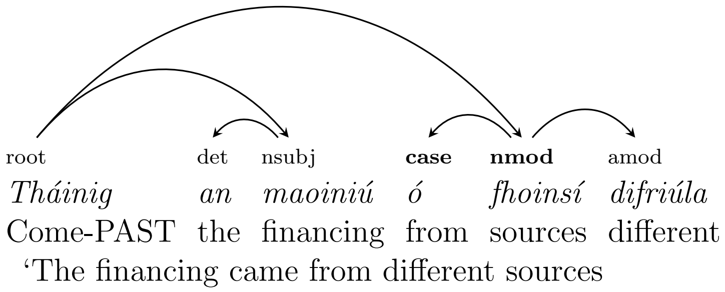


Figure 5.17: UD15 prepositional phrase analysis

Progressive aspectual phrases are constructed with the preposition *ag* followed by a verbal noun. The IDT regards *ag* as the head of the prepositional phrase, and thus the `xcomp` relation is identified between the matrix verb and the preposition. In the UD15 scheme however, the verbal noun is regarded as the head of the prepositional phrase. Compare Figures 5.18 and 5.19.

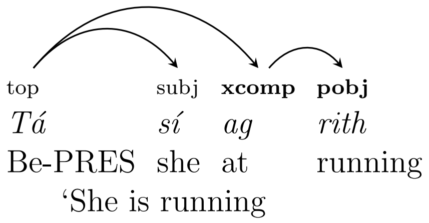


Figure 5.18: IDT progressive aspectual phrase analysis

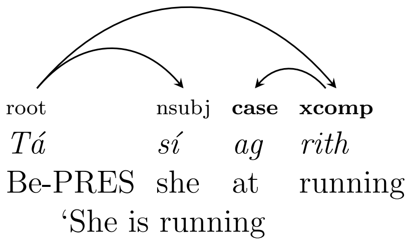


Figure 5.19: UD15 progressive aspectual phrase analysis

Prepositional predicates are labelled as `ppred` in the Irish Dependency Treebank. In keeping with the other PP analyses, the preposition is the head of the prepositional phrase. The label `ppred` maps to `xcomp:pred` in the UD15 scheme.[9] In addition, the object of the preposition is now regarded as the head of the phrase. See Figures 5.20 and 5.21 for comparison of prepositional predicate analyses.

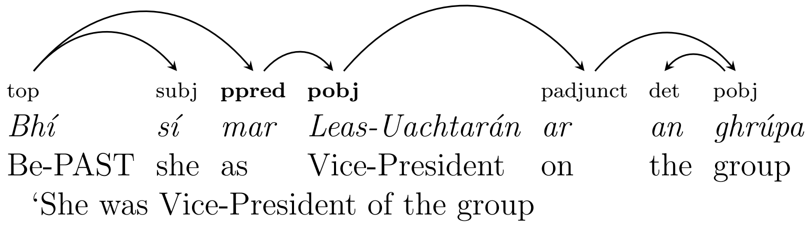


Figure 5.20: IDT prepositional predicate analysis

5.2.2.4 Irish-specific relations

One advantage of the UD15 scheme is that it provides scope to include language-specific subtype labels. The label name format is *universal:extension*, which ensures that the core UD15 relation is identified, making it possible to revert to this coarse

[9] The label `xcomp:pred` is an Irish-specific label, these language specific labels are discussed in Section 5.2.2.4.

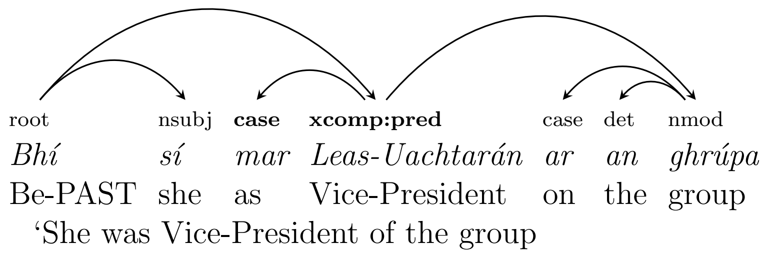

root
nsubj
case
xcomp:pred
case
det
nmod
Bhí *sí* *mar* *Leas-Uachtarán* *ar* *an* *ghrúpa*
Be-PAST she as Vice-President on the group
'She was Vice-President of the group

Figure 5.21: UD15 prepositional predicate analysis

label for cross-lingual analysis. During the conversion of the IDT, we defined some labels required to represent Irish syntax more concisely. These labels are discussed below.

acl:relcl This label is used for relative clause modifiers. This subtype label is more specific than its supertype counterpart `acl`, which covers examples in English such as 'This case to **follow**' and 'He entered the room **sad**'. These types of constructions have not been observed in the Irish data. This subtype label `acl:relcl` is used in cases where the head of the relative clause is a predicate (usually a verb), and is dependent on a noun in a preceding clause. It is also used in the English, Finnish and Swedish schemes. An example of this subtype used in the converted IDT is in Figure 5.22.

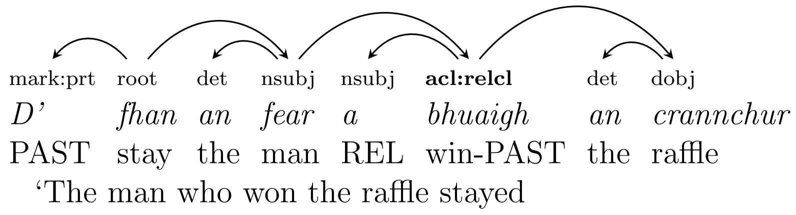

mark:prt
root
det
nsubj
nsubj
acl:relcl
det
dobj
D' *fhan* *an* *fear* *a* *bhuaigh* *an* *crannchur*
PAST stay the man REL win-PAST the raffle
'The man who won the raffle stayed

Figure 5.22: UD15 relative clause analysis

case:voc The vocative particle *a* is a case marker in Irish and precedes an addressee. From observing how the UD15 scheme recognises prepositions as case markers through the `case` label, we chose to retain the `case` information for vocatives also. For example, *Slán* ***a*** *chara* 'Goodbye, friend'.

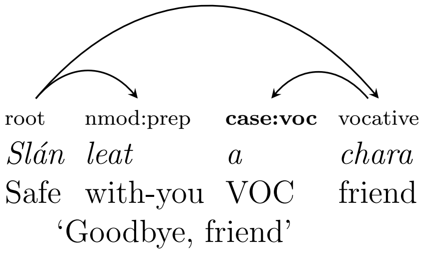


Figure 5.23: UD15 vocative particle analysis

csubj:cop The supertype label `csubj` indicates a clausal subject (a clause that acts as the subject of another). In English '[what she `said`] makes sense'. However, Finnish uses an additional specific subtype label `csubj:cop` to indicate clausal subjects that act as a subject of a copular clause. We observed in our data that clausal subjects in Irish are only ever subjects of copula clauses. For this reason we use only the subtype label `csubj:cop` for clausal subjects. See Figure 5.24 for an example of use.

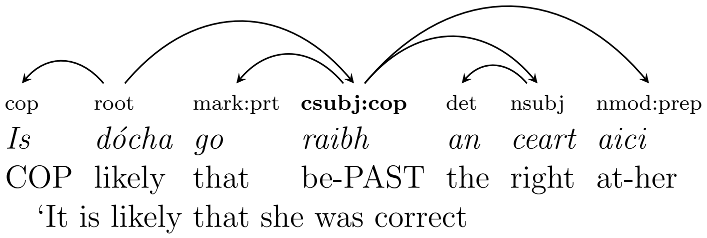

cop root mark:prt **csubj:cop** det nsubj nmod:prep
Is dócha go raibh an ceart aici
COP likely that be-PAST the right at-her
'It is likely that she was correct

Figure 5.24: UD15 copular clausal subject analysis

mark:prt The UD15 scheme does not use the UD13 `prt` label to which we previously mapped Irish particles. Therefore we introduce the new subtype label `mark:prt` for: adverbial particles, cleft particles, quantifier particles, comparative/ superlative particles, verb particles and days of the week particles.[10]

nmod:poss In Irish, possession is denoted by possessive pronouns (*mo, do, a, ár, bhur*). English, Finnish and Swedish, use the subtype label `nmod:poss` to indicate possession, and we also adopt it for Irish. An example is given in Figure 5.25.

[10] The infinitive marker, previously subsumed by the UD13 `prt`, is now mapped to the supertype `mark` as per other languages such as English and French.

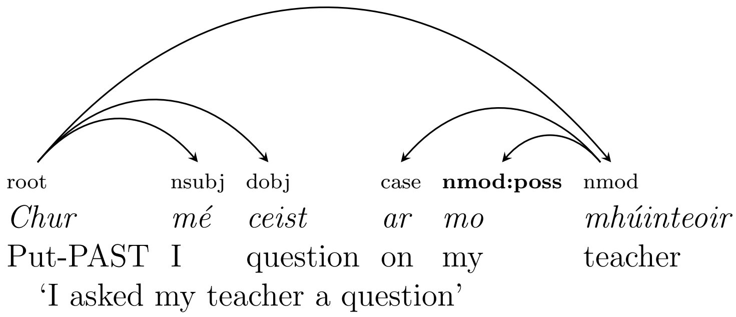


Figure 5.25: UD15 possessive analysis

nmod:prep 16 of the most common Irish simple prepositions can be inflected to mark pronominal objects (see Section 5.1.1 for a more detailed description). These are referred to as pronominal prepositions or prepositional pronouns, and were most frequently marked as `obl` or `obl2` in the IDT. In the UD15 scheme, where the object is the head of a preposition phrase, we regard these as playing nominal roles instead of prepositional roles (yet their POS-tag remains `ADP`). We introduce the language-specific label `nmod:prep` so as not to lose information regarding the presence of the preposition within this synthetic form. An example is given in Figure 5.26. Note that in some cases, prepositional pronouns behave like nominal modifiers of noun phrases. E.g. *an bheirt* ***acu*** 'the two **of them**'. These cases take the label `compound`.

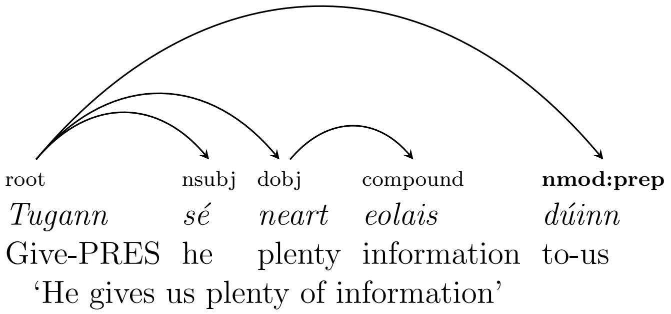


Figure 5.26: UD15 prepositional pronoun analysis

nmod:tmod Temporal modifiers specifying time, in nominal form, are labelled as `nmod`. English also uses this subtype label. An example is given in Figure 5.27.

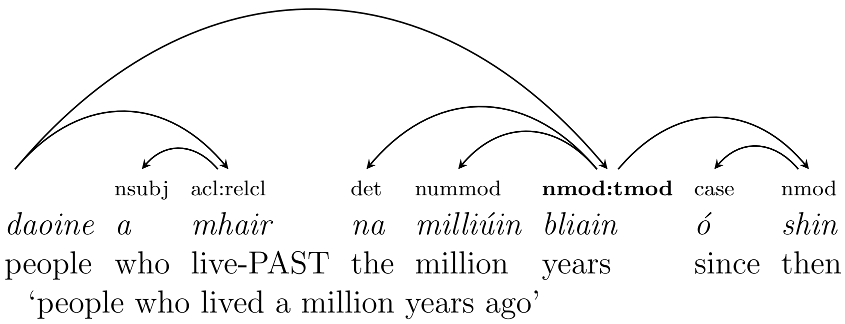

nsubj
acl:relcl
det
nummod
nmod:tmod
case
nmod
daoine a mhair na milliúin bliain ó shin
people who live-PAST the million years since then
'people who lived a million years ago'

Figure 5.27: UD15 temporal modifier analysis

xcomp:pred The IDT uses the following fine-grained labels for predicates: `npred` (nominal), `adjpred` (adjectival), `advpred` (adverbial) and `ppred` (prepositional). These were typically used in copular constructions and are now no longer relevant in the UD15 where the predicate heads the copular phrase. However, adjective, adverbial and prepositional predicates can also be arguments of the substantive verb *bí*. In LFG, the xcomp (open complement) label is used to represent predicates. Therefore, we extend the open complement label to include the subtype `xcomp:pred`. See Figure 5.28 for an example of an adjectival predicate.

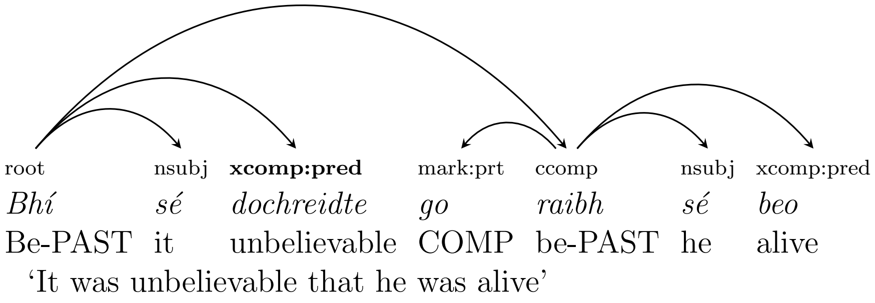

root
nsubj
xcomp:pred
mark:prt
ccomp
nsubj
xcomp:pred
Bhí sé dochreidte go raibh sé beo
Be-PAST it unbelievable COMP be-PAST he alive
'It was unbelievable that he was alive'

Figure 5.28: UD15 adjectival predicate analysis

5.3 Summary and Conclusion

In this chapter, we have discussed the role of a universal grammar in recent NLP research and the importance of a universal annotation scheme when attempting to compare and cross-evaluate different treebanks and parsers. We have described the mapping of the Irish Dependency annotation scheme to two universal annotation schemes – namely UD13 (McDonald et al., 2013) and UD15 (Nivre et al., 2015; Agić et al., 2015). We have described in detail the mapping and conversion process, including structural changes, for the release of the Irish UD15 treebank as part of the Universal Dependencies project.[11] We have also discussed linguistic analyses and motivation for choice of language-specific label types for Irish.

In the next chapter we discuss using both the Irish Dependency Treebank and the Irish Universal Dependency Treebank as training data for parsing. We report on various parsing experiments for both mono-lingual and cross-lingual studies.

[11] The Irish UD treebank is available to download from The Universal Dependency project repository: v1.0 `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1464` and v1.1 `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/LRT-1478`

Chapter 6

Parsing with the Irish Dependency Treebank

Syntactic parsing is a process by which grammatical structures are identified and assigned to sentences within a text. Typically, parsing is an automated process and plays a significant role in the development of NLP tools. Treebanks are not only valuable for linguistic research and corpus analysis, but they also provide training data on which statistical parsing models can be built. In this chapter, we show how the Irish Dependency Treebank is used as training data for building the first full Irish dependency parser.

Statistical data-driven parsers learn how to syntactically analyse sentences from a set of examples. This data set, in the form of a treebank, is referred to as training data. What counts as a sufficient number of trees for training an Irish statistical dependency parser remains an open question. However, what is clear is that the parser needs to have encountered a linguistic phenomenon in training in order to learn how to accurately analyse it. Creating a treebank is a resource-intensive process which requires extensive linguistic research in order to design an appropriate labelling scheme, as well as considerable manual annotation (parsing). In general, manual annotation is desired to ensure high quality treebank data. Yet, as is often encountered when working with language, the task of manually annotating text can become repetitive, involving frequent encounters with similar linguistic structures.

In this chapter, we provide an overview of dependency parsing in Section 6.1 and report a baseline parsing score for our Irish dependency parsing experiments in Section 6.2. These baseline scores increased following our inter-annotator study, as we show in Section 6.3. Some research has been carried out on experimenting with automating, or at least, semi-automating the process of treebank annotation and improving parser accuracy. In general, this previous work has been experimental and simulated, but not actually tried out in the context of building resources for a low-resourced language. We extend on these approaches to address the actual problem of building a treebank in the early stages of its development. In Section 6.4, we therefore discuss how, by examining the accuracy of our parsing models, we bootstrap the development of our treebank through Active Learning methods. From an annotation perspective, if a parser can be used to pre-parse data for manual correction, then a high accuracy parser would result in a smaller annotation (correction) effort. For that reason, in Section 6.5, Section 6.6 and Section 6.7, we show how we explore bootstrapping parser development for a low-resourced language by leveraging existing minimal resources and those of other better-resourced languages.

We conclude however, that despite our various attempts at semi-automating the development of our parser (which, in effect, could help the effort of treebank development), we find that human input cannot easily be replaced by technology in these type of tasks. In essence, we show that there is no reasonable substitute for the manual work required in building a treebank.

6.1 An Overview of Dependency Parsing

Studies in dependency grammars have become more popular in the last two decades, and towards the end of the 1990's there was a shift towards dependency parsing. From a computational perspective, this shift has been attributed to an increased efficiency in learning and parsing through dependency parsing. One of the main factors contributing to this is the smaller search space required when dealing with less complex bilexical labelled dependency structures, when compared to processing or predicting constituency structures that have additional hierarchical non-terminal nodes to consider (Nivre and Scholz, 2004; McDonald et al., 2005).

The earliest computational implementation of dependency grammars was by Covington (1990, 1994). Not long afterwards, Finite State approaches (e.g. Karlsson (1995)) gave rise to rule-based transducers being developed to parse text based on dependency relations.[1] This era also saw new approaches to parsing such as the extraction from corpora of argument structures such as verbal subcategorisation classes, in order to improve statistical constituency parsing (e.g. Briscoe and Carroll (1997)). Ideally, statistical dependency parsers need data on which to both evaluate and train a system. In these early days, resources were few. Despite seminal work on the earliest development of a number of probabilistic models for dependency parsing (Eisner, 1996a,b), evaluation took place on the constituency-parsed Penn Treebank. In addition, the SUSANNE corpus (Sampson, 1993), a constituency-parsed corpus that also labels functional roles, proved to be a valuable resource that was used for the evaluation of the Minipar parser (Lin, 2003), a parser which enabled extraction of dependency relationships from its parsed output. Extracting dependencies from phrase structures was also the approach taken by Bouma et al. (2000) to develop a dependency treebank for Dutch. However, despite the trend of converting from constituents to dependencies to create data resources, during this time, development also began of pure dependency resources, such as the Prague Dependency Treebank for Czech (Hajič, 1998).

In more recent advancements in statistical data-driven parsing, an increase in the availability of resources (dependency treebanks are now available for a large number of languages) has resulted in the development of dependency parsing platforms such as MSTParser (McDonald et al., 2005), MaltParser (Nivre et al., 2006) and

[1] Section 3 shows how Karlsson (1995)'s Constraint Grammar was used to develop a shallow/ partial parser for Irish.

Mate Parser (Bohnet, 2010), to name a few. Probabilistic (statistical data-driven) dependency parsers predict dependency trees after having learned patterns within the training data (treebank), which help to assess the probability of two words being part of a head-modifier relationship (see Section 6.1.1 for more detail on how parsers work). Converting dependency resources such as the Prague Dependency Treebank to constituent trees in order to train and test probabilistic parsers (e.g. Collins et al. (1999)) was no longer necessary following the development of these dependency parsing platforms.

From a parsing efficiency perspective, dependency syntax lends itself to much more efficient parsing algorithms than constituency parsing. This is as a result of dependency parsers only needing to process and predict labels for the tokens in the sentence. Constituency parsers need to consider both tokens (terminals) and phrasal categories (non-terminals) during processing time. We now provide an overview of how data-driven dependency parsers process their input data and predict parse trees through statistical methods.

6.1.1 How do Transition-Based Dependency Parsers work?

In the last decade, data-driven dependency parsing has come to fore, with two main approaches dominating – graph-based and transition-based. Graph-based dependency parsing involves the construction of a parse tree by predicting the Maximum Spanning Tree (MST) in the digraph for the input sentence (McDonald et al., 2005). In this digraph, each word corresponds to a vertex, and these vertices are all connected by directed edges (arcs). Based on frequency counts in the training data (treebank), each arc in the graph is assigned a score at the learning or training stage. Making a common assumption of arc factorisation, the score of the graph is the sum of all the arc scores (weights). The challenge of the parser is to find the highest-scoring tree, that is, a subgraph including all vertices and only the minimum number of arcs to be connected (see Section 2.3.1) – the MST, when choosing from amongst the proposed candidates. MSTParser (McDonald et al., 2005) is a graph-based parser. Mate Parser (Bohnet, 2010) also has a graph-based component.

On the other hand, in classic transition-based dependency parsing, the training phase consists of learning the correct parser action to take, given the input string and the parse history, and the parsing phase consists of the greedy application of parser actions as dictated by the learned model. These actions are based on that of a shift-reduce parser, which involves progressing through the input string and moving tokens onto a stack from a buffer (shift) or removing them once they have been fully processed (reduce). Depending on the type of algorithm used, there are other actions involved, which we describe in more detail in Figure 6.1. MaltParser (Nivre et al., 2006) is an example of a transition-based parser and we describe how it works here in more detail than a graph-based parser as it is the main parser we used in most of our parsing experiments in Section 6.

In transition-based parsing, the parser moves from left to right through a sentence, making decisions as to which words will make up dependency pairs with the help of a classifier. The transition-based parsing algorithms use a *buffer* containing the sentence tokens in linear order, a *stack* onto which each token is pushed as part of the processing step and an *arc list* that contains the proposed head-modifier relations (Nivre, 2003; Nivre and Nilsson, 2005; Kubler et al., 2009). In this approach, the parser looks to see what is on the top of the stack and appearing next in the buffer. Due to the fact that it does not look beyond the next item in the buffer nor does it undo any decisions it has already made, it is referred to as a greedy algorithm. Jurafsky and Manning (2012) identify the four main contributors (listed below) that help to calculate the probabilities of a dependency pair. In the training stage, the parser uses a classifier when looking at features in order to a predict parsing action, given a particular configuration. These features also help it to decide what dependency label to apply to the relation pair.

- **Lexical** information (based on data previously observed in a treebank) can tell the parser if two words are likely to be a dependency pair (e.g. 'small child').

```
  Initialise:
  $\sigma = [\mathrm{ROOT}]$, $\beta = \mathrm{w}_1, ..., \mathrm{w}_n$, $\mathrm{A} = \phi$
  repeat
    Choose one of the transition operations 1-4
  until $\beta = \phi$
Transitions
1. Left-Arc$_r$ $\sigma| \mathrm{w}_i, \mathrm{w}_j|\beta, \mathrm{A} \rightarrow \sigma, \mathrm{w}_j|\beta, \mathrm{A} \cup \{\mathrm{r}(\mathrm{w}_j, \mathrm{w}_i)\}$
precondition: $r\prime\ (\mathrm{w}_k, \mathrm{w}_i) \notin \mathrm{A}, \mathrm{w}_i \neq \mathrm{ROOT}$
2. Right-Arc$_r$ $\sigma| \mathrm{w}_i, \mathrm{w}_j|\beta, \mathrm{A} \rightarrow \sigma| \mathrm{w}_i| \mathrm{w}_j, \beta, \mathrm{A} \cup \{\mathrm{r}(\mathrm{w}_i, \mathrm{w}_j)\}$
3. Reduce $\sigma| \mathrm{w}_i, \beta, \mathrm{A} \rightarrow \sigma, \beta, \mathrm{A}$
precondition: $r\prime\ (\mathrm{w}_k, \mathrm{w}_i) \in \mathrm{A}$
4. Shift $\sigma, \mathrm{w}_i\ |\beta, \mathrm{A} \rightarrow \sigma| \mathrm{w}_i, \beta, \mathrm{A}$
```

Figure 6.1: The arc-eager shift-reduce algorithm. σ represents the stack, β is the buffer and A is the set of dependencies.

- The **distance** between two words can also indicate the probability that these two words will be connected. Dependency distance tends to be short, although long-distance dependencies do exist. For example, determiners and their nouns are usually in close proximity. In Irish, the subject and verb are usually close in proximity, whereas inserted elements such as adverbs or subject modifiers can increase the distance between a verb and its object.

- **Intervening words:** the parser can use the words occurring between the two words likely to attach to determine whether the attachment is possible. For example, punctuation and verbs are unlikely to occur between dependencies.

- The **valency** of a word is the number of arguments it is likely to have. Depending on its part of speech (e.g. noun, verb), the valency of words can differ, and vary as to whether the arguments will be to the left or right. For example, the verb *tabhair* 'give' has a valency of 3: subject, object and an oblique argument.

The *arc-eager* algorithm is outlined in Figure 6.1.[2] At each transition step, the parser consults the classifier to determine a word's dependencies.

[2] Taken and adapted from Jurafsky and Manning (2012).

The Start configurations are such that σ indicates the **stack**, onto which tokens will be pushed as they are encountered, for which the notation is $\sigma|w_i$ for a token w_i. At initialisation there is just one element on the stack, that is the `ROOT`. β represents a **buffer** with the list of tokens from the input (the sentence) as $word_1, word_2$, and so on. **A** represents a set of dependency arcs $r(\mathrm{w}_i, \mathrm{w}_j)$, where r is a dependency label. At initialisation, A is empty (ϕ). There are four possible operations as the parser iterates through each word on the buffer: **Left-Arc**, **Right-Arc**, **Reduce**, **Shift**. The precondition of Left-Arc is that the token on the top of the stack cannot already be a dependent of another word (thus preventing multiple heads), and it cannot be the root of the sentence. The precondition of the Reduce operation is that a token can only be removed from the stack if it has been made a dependent of another word. The process is finished when the buffer is empty.

Let us take the Irish sentence in Example 37 and parse it with the arc-eager algorithm.

(37) Cheannaigh an cailín an leabhar
Bought the girl the book
'The girl bought the book'

The Start configurations are:

σ = [ROOT]; β = Cheannaigh, an, cailín, an, leabhar; A = ϕ

1. *Cheannaigh*, a verb, is likely to be the root of the sentence, therefore apply a **Right-Arc**$_{root}$ operation, which takes the word on the top of the stack (*ROOT*) and makes it the head of the next word in the buffer *Cheannaigh*. Add to A:

   A = {root(ROOT, Cheannaigh)}

   Part of Right-Arc operation is to push *Cheannaigh* onto the stack:

   σ = [ROOT | Cheannaigh]; β = an, cailín, an, leabhar;

2. **Shift** *an* onto the stack (an unlikely dependent of *Cheannaigh*)

   σ = [ROOT, Cheannaigh | an]; β = cailín, an, leabhar

3. The next word in the buffer (*cailín*) is likely to be the head of *an* – apply the **Left-Arc**$_{det}$ operation, which takes the word on the top of the stack (*an*) and makes it the dependent of the next word in the buffer *cailín*. Add to A:

   A = {root(ROOT, Cheannaigh), **det(cailín, an)**}

4. **Reduce** – Take *an* off the stack (it is already a dependent)

   σ = [ROOT | Cheannaigh]; β = cailín, an, leabhar

5. Apply the **Right-Arc**$_{subj}$ operation. *cailín* is a good candidate as a subject dependent of *Cheannaigh*. Add to A.

   A = {root(ROOT, Cheannaigh), det(cailín, an), **subj(Cheannaigh, cailín)**}

   We add *cailín* to the stack (σ| w$_i$| w$_j$):

   σ = [ROOT, Cheannaigh | cailín]; β = an, leabhar

6. **Reduce**: As *cailín* does not have any right dependents, and it is already identified as a dependent of another token, it can be taken off the stack:

   σ = [ROOT | Cheannaigh]; β = an, leabhar

7. **Shift** *an* onto the stack (*an* is not likely to be the dependent of *Cheannaigh*)

   σ = [ROOT, Cheannaigh | an]; β = leabhar

8. The next word in the buffer (*leabhar*) is likely to be the head of *an* – apply the **Left-Arc**$_{det}$ operation and add to A:

   A = {root(ROOT, Cheannaigh), det(cailín, an), subj(Cheannaigh, cailín), **det(leabhar, an)**)}

9. **Reduce** – Take *an* off the stack (it is already a dependent)

   σ = [ROOT | Cheannaigh]; β = leabhar

10. Apply the **Right-Arc**$_{obj}$ operation. *leabhar* is a good candidate as an object dependent of *Cheannaigh*. Add to A.

    A = {root(ROOT, Cheannaigh), det(cailín, an), subj(Cheannaigh, cailín), det(leabhar, an), **obj(Cheannaigh, leabhar)**}

    We add *leabhar* to the stack (σ| w$_i$| w$_j$)

    σ = [ROOT, Cheannaigh | leabhar]; $\beta = \phi$

11. **Finish** – as the buffer is empty.

It is clear in this example that we are assuming that all the decisions made by the classifier are correct. Of course, the classifier is not always reliable and the parser cannot get all choices correct each time. Yet, we can see how the various features it uses assists it in making the correct decisions. For example, we can see in the steps above that at one stage *Cheannaigh* is on the top of the stack, and *an* is the next token on the buffer. We note that *an* is unlikely to be a dependent of *Cheannaigh* at this stage. Yet, if we were parsing the (present tense) interrogative form of the sentence (*An gceannaíonn an cailín an leabhar?* 'Does the girl buy the book?'), the first word An_1 is a present tense interrogative particle, and should in fact be attached to the verb *gceannaíonn* as a dependent. How then, would the classifier know the difference between what is possible for An_1 and an_2 (of *an cailín*)? In this case, the length of the proposed dependency arc for both cases is equal, so that feature will not bear any weighting. The part-of-speech tags will differ however. The POS for An_1 is `Part` (particle) and the POS for an_2 is `Art` (Article). In addition, An_1, in sequence occurs to the left of the verb, requiring a Left-Arc operation, therefore is more likely to be a dependent than an_2 which occurs to the right of the verb (requiring a Right-Arc operation).

6.1.2 Evaluation Metrics

Parsers are evaluated by comparing the parsed output of test data to a gold standard version of the test data. In dependency parsing, by comparing the trees, it is possible to calculate the accuracy of each parse. The score is based on the number of correct dependencies, i.e. the percentage of dependency relations that the parser got right:

$$accuracy = \frac{\text{\# correct dependencies}}{\text{total \# dependencies}}$$

In general, the accuracy of a dependency parser is reported with two scores (i) Unlabelled Attachment Score (UAS) is the accuracy of the parser's prediction of attachments (assigned head) only, and (ii) Labelled Attachment Score (LAS) reports the accuracy based on predictions of both the correct attachments and correctly assigned dependency labels to each head-modifier relation.

6.2 Establishing a Baseline Parsing Score

In the early stages of the treebank's development, the sentences in our corpus were manually annotated. As our corpus sentences had been randomised, we sequentially chose blocks of sentences to annotate. Once we reached the first **300** fully-annotated and reviewed (gold standard) trees, we were able to train a statistical model and establish a parsing baseline score against which all future parsing models could be compared. We trained MaltParser (Nivre et al., 2006) on our 300 gold trees and tested for accuracy using 10-fold cross-validation.[3]

We apply the *stacklazy* and *LIBLINEAR learner* algorithms. The *stacklazy* algorithm differs slightly to the algorithm described in Section 6.1.1, in that the arcs are created between the top two nodes on the stack instead of between the top nodes of both the stack and the next node on the buffer. The stacklazy algorithm can directly handle non-projective structures that result in crossing dependencies, which is important for our parsing experiments as our analyses show that the Irish data contains some non-projective structures.[4] The *liblinear* algorithm has been proven to reduce processing time, compared to other classifiers such as LibSVM (Cassel, 2009).

We test a variety of feature models which make use of various combinations of the following information extracted from the 3,000-sentence POS-tagged corpus: surface form, lemma, fine-grained and coarse-grained POS tags (see Section 3.1.3). Section 6.1.1 also discusses the use of features in inducing parsing models. The

[3] As a held out test set would be too small of a representation of the linguistic variances in Irish, k-fold cross validation is a more favoured approach to testing models on small data sets. The entire data set is split into k number of sets, with each set iterated over as a test set, using the remaining data as training set. The scores are then averaged across all sets. (Jurafsky and Martin, 2000, pp. 154-155).

[4] See Section 2.3.2 for a discussion of projectivity.

Model	**LAS**	**UAS**
Form+POS:	60.6	70.3
Lemma+POS:	61.3	70.8
Form+Lemma+POS:	61.5	70.8
Form+CPOS:	62.1	72.5
Form+Lemma+CPOS:	62.9	72.6
Form+CPOS+POS:	63.0	72.9
Lemma+CPOS+POS:	63.1	72.4
Lemma+CPOS:	63.3	72.7
Form+Lemma+CPOS+POS:	**63.3**	**73.1**

Table 6.1: Preliminary parsing results with MaltParser. LAS is Labelled Attachment Score – the proportion of labels and attachments the parser correctly predicts. UAS is Unlabelled Attachment Score – the proportion of attachments the parser correctly predicts.

results are shown in Table 6.1.

The small size of our seed set meant that the differences between the various models were not statistically significant.[5] Nevertheless, we chose the best-performing model as our baseline model in the bootstrapping process with a UAS score of 73.1% and LAS score of 63.3%. This model uses information from the word form, lemma and both POS tags, and these are the features we use in all our MaltParser parsing experiments discussed in this chapter.

To put these baseline scores into perspective, the CoNLL 2007 Shared Task on Dependency Parsing (Nivre et al., 2007) categorises the parsing LAS results for ten languages as follows: low scores (76.31-76.94), medium (79.19-80.21) and high (84.4-89.61). The size of the training data sets for these languages vary widely (as shown in Table 6.2), but the general trend is that the larger training sets gained the higher parsing scores. Most of these dependency treebanks have pure dependency annotations or both dependency and constituency annotations, others were converted from constituency trees (as indicated with †).

[5] We used Dan Bikel's Randomized Parsing Evaluation Comparator, (`http://www.cis.upenn.edu/~dbikel/software.html#comparator`, downloaded 28 March 2011), to calculate statistical significance between the output of the parsing models. We use this same tool for all statistical significant testing reported in this thesis, and base significance on the p-value of 0.05.

Training Data	**Ar**	**Ba**	**Ca †**	**Ch †**	**Cz**	**En †**	**Gr**	**Hu †**	**It**	**Tu**
# Tokens (k)	112	51	431	337	432	447	65	132	71	65
# Trees (k)	2.9	3.2	15	57	25.4	18.6	2.7	6.0	3.1	5.6
Best LAS	76.5	76.9	88.7	84.7	80.2	89.6	76.3	80.3	84.4	79.8
Best UAS	86.1	82.8	93.4	88.9	86.3	90.6	84.1	83.6	87.9	86.2

Table 6.2: CoNLL 2007 Shared Task on Dependency Parsing: Treebank sizes and best LAS/UAS results, rounded to one decimal point. (Note: † denotes treebanks converted from constituency trees to dependency format.)
Ar:Arabic, Ba:Basque, Ca:Catalan, Ch:Chinese, Cz:Czech, En:English, Gr:Greek, Hu:Hungarian, It:Italian, Tu:Turkish.

Model	**LAS-1**	**UAS-1**	**LAS-2**	**UAS-2**
Form+POS:	60.6	70.3	64.4	74.2
Lemma+POS:	61.3	70.8	64.6	74.3
Form+Lemma+POS:	61.5	70.8	64.6	74.5
Form+CPOS:	62.1	72.5	65.0	76.1
Form+Lemma+CPOS:	62.9	72.6	66.1	76.2
Form+CPOS+POS:	63.0	72.9	66.0	76.0
Lemma+CPOS+POS:	63.1	72.4	66.0	76.2
Lemma+CPOS:	63.3	72.7	65.1	75.7
Form+Lemma+CPOS+POS:	63.3	73.1	**66.5**	**76.3**

Table 6.3: Preliminary MaltParser experiments with the Irish Dependency Treebank: pre-IAA results (as per Table 6.1) and post-IAA results.

6.3 Improvements over Baseline following IAA Study

In Section 3.2, we reported on our inter-annotator agreement study. We discussed how subsequent workshops, where the annotators and a third linguistic expert analysed the disagreements found in this study, highlighted improvements required for the annotation guide and scheme. Following the changes we made to the labelling scheme as a result of the our first IAA study, we re-ran the same parsing experiments on the newly updated seed set of 300 sentences. We used 10-fold cross-validation on the same feature sets (various combinations of form, lemma, fine-grained POS and coarse-grained POS). The parser's accuracy scores increased to 66.5% (LAS) and 76.3% (UAS). The improved results, as shown in the final two columns of Table 6.3, reflect the value of undertaking an analysis of IAA-1 results, with a subsequent improvement to the labelling scheme and annotation guide. The improvements also are a direct result of identifying the sources of annotator disagreements outlined in

Section 3.2.3 and updating the treebank accordingly.

6.4 Bootstrapping Treebank Development – Active Learning

Total # Gold Trees	**Tree batch size**	**Source**
300	300	Manual annotation

Table 6.4: Treebank Status

Human input is fundamental in treebank development for the purpose of ensuring accuracy and quality. However, there are steps that have been proposed to semi-automate the annotation task so as to reduce the manual effort and time required. One *bootstrapping approach* involves training a parser to pre-parse the trees, allowing the annotator to correct the parser's output instead of annotating from scratch. The newly parsed data is added to the training data and the cycle continues until all data is parsed. We refer to this as a basic bootstrapping method as it differs considerably to the various other bootstrapping approaches that we will discuss in later sections. Note that the term bootstrapping can sometimes imply that there is no manual annotation effort involved. We clarify here that our basic bootstrapping method involves a manual correction stage to ensure that gold trees are added to the treebank on each iteration.

In order to therefore speed up our treebank development process, we applied the following bootstrapping approach to annotating the subsequent 150 sentences in our corpus (similar to that of Judge et al. (2006) and more recently, Seraji et al. (2012)):

1. Manually annotate a seed set of n sentences.
2. Train a baseline parsing model on the seed set.
3. Parse m sentences with the baseline model and manually correct the output.
4. Add the m automatically parsed and manually corrected trees to the training set and train a new model.

5. Parse another m sentences with the new model and manually correct the output.
6. Repeat steps 4 and 5 until all sentences have been parsed and used as training data.

This kind of iterative approach to parsing a corpus allowed us to take advantage of the presence of repetition in the data. The parser learns more at each iteration as a result of exposure to repetitive syntactic structures. Despite starting from such a small training set, frequent structures, such as determiner or adjective attachment to nouns, are easily learned. Thus, with this type of approach, the parser is expected to correctly annotate the frequently encountered and learned structures, leaving the annotator with the manual correction of only the infrequent, previously unseen or difficult parses each time. Through the addition of newly parsed data to the training material at each iteration, the learning process becomes quicker. Using this approach, we increased the treebank size to **450 trees**.[6]

However, one can argue that it may be futile to continue to add examples of structures already learned – those which the parser already finds easy to parse – and instead, it would be a more productive use of resources and time to annotate more difficult constructions. An enhancement to this standard bootstrapping approach is *Active Learning*, which is a form of selective sampling whereby only 'informative' sentences are chosen to be annotated and added to the parser's training data on each iteration. Sentences are regarded as informative if their inclusion in the training data is expected to fill gaps in the parser's knowledge.

In Section 6.4.1, we discuss previous work on active learning in NLP. In Section 6.4.2, we explain how our active learning experiments were carried out and we discuss those results in Section 6.4.3. We then provide a discussion of our findings in Section 6.4.4, followed by a summary of experiments using unlabelled data in Section 6.4.5. Finally, we provide some conclusions on active learning for treebank development in Section 6.4.6.

[6] 100 of these trees were used in the IAA study.

6.4.1 Related Work

Active learning is a general technique applicable to many tasks involving machine learning. Two broad approaches are Query By Uncertainty (QBU) (Cohn et al., 1994), where examples about which the learner is least confident are selected for manual annotation; and Query By Committee (QBC) (Seung et al., 1992), where disagreement among a committee of learners is the criterion for selecting examples for annotation. Active learning has been used in a number of areas of NLP such as information extraction (Scheffer et al., 2001), text categorisation (Lewis and Gale, 1994; Hoi et al., 2006) and word sense disambiguation (Chen et al., 2006). Olsson (2009) provides a survey of various approaches to active learning in NLP.

For our own work, and from a parsing point of view, the following studies are the most relevant application of active learning. Thompson et al. (1999) apply active learning sampling methods to semantic parsing. Hwa et al. (2003) use selective sampling for co-training to reduce the annotation effort and time required to annotate labelled training data for syntactic constituency parsing. Reichart and Rappoport (2007) use sample selection methods to assess parse quality. More closely in line with our own studies, Osborne and Baldridge (2004) use active learning in HPSG (Head-driven Phrase-Structure Grammar) parsing. The goal of their work was to improve parse selection for HPSG: for all the analyses licensed by the HPSG English Resource Grammar (Baldwin et al., 2004) for a particular sentence, the task is to choose the best one using a log-linear model with features derived from the HPSG structure. The supervised framework requires sentences annotated with parses, which is where active learning can play a role. Osborne and Baldridge (2004) apply both QBU with an ensemble of models, and QBC, and show that this decreases annotation cost, measured both in number of sentences to achieve a particular level of parse selection accuracy, and in a measure of sentence complexity, with respect to random selection.

In corpus development, Hughes et al. (2005) use active learning to assist them in selecting the most useful sentences for annotation in their development of a wide-coverage corpus of CCG (Combinatory Category Grammar) derivations. Similarly, Atserias et al. (2010) took an active learning approach to developing a dependency treebank of questions. Since our own study, Ghayoomi and Kuhn (2013) have explored (through simulation) various active learning approaches for the future expansion of a Persian HPSG treebank. Their aim is similar to our work on Irish, as Persian is also a low-resourced language.

In addition, most active learning work in NLP has used variants of QBU and QBC where instances with the *most* uncertainty or disagreement (respectively) are selected for annotation. Some work by Sokolovska (2011) in the context of phonetisation and named entity recognition has suggested that a distribution over degrees of uncertainty or disagreement may work better: the idea is that examples on which the learners are more certain or in greater agreement might be more straightforwardly added to the training set. We also consider this approach in the context of treebank construction, so that examples selected by active learning for annotation are a mix of easier and more complex.

6.4.2 Active Learning Experiments

Total # Gold Trees	**Tree batch**	**Source**
	300	Manual annotation
450	150	Automated & manual correction (incl. IAA trees)

Table 6.5: Treebank Status

We carried out experiments to assess the extent to which active learning can play a role in treebank and parser development. We assessed whether such an approach could be considered useful by comparing a QBC active learning bootstrapping approach to a passive one in which sentences are chosen at random for manual revision. The following describes the steps involved in setting up this experiment.

Our treebank/parser bootstrapping algorithm is given in Algorithm 1. In an initialisation step, a parsing model is trained on a seed set of gold standard trees. In

Algorithm 1 Our bootstrapping algorithm

```
$A$ is a parser.
$M^i_A$ is a model of $A$ at step $i$.
$P^i_A$ is a set of X trees produced using $M^i_A$.
$U$ is a set of sentences.
$U^i$ is a subset of $U$ at step $i$.
$L$ is the manually labelled seed training set.
$L^i_A$ is labelled training data for $A$ at step $i$.
**Initialise:**
$L^0_A \leftarrow L$.
$M^0_A \leftarrow Train(A,L^0_A)$
**for** $i = 1 \rightarrow N$ **do**
    $U^i \leftarrow$ Add set of unlabelled sentences from $U$
    $P^i_A \leftarrow Parse(U^i\ ,\ M^i_A)$
    $Pr^i_A \leftarrow$ Select a subset of $Y$ parsed trees from $P^i_A$
    $Pr^i_{Agold} \leftarrow$ Hand-correct $Pr^i_A$
    $L^{i+1}_A \leftarrow L^i_A + Pr^i_{Agold}$
    $M^{i+1}_A \leftarrow Train(A,L^{i+1}_A)$
**end for**
```

each iterative step, a new batch of unseen sentences is retrieved, the parsing model is used to parse these sentences, a subset of these automatically parsed sentences is selected (in a manner described below), the parse trees for the sentences in this subset are manually corrected, the corrected trees are added to the training set and a new parsing model is trained. This process is repeated, ideally until parsing accuracy converges.

We experiment with two versions of this bootstrapping algorithm. In the *passive learning* variant, the Y trees that are added to the training data on each iteration are chosen at random from the batch of X unseen sentences. In the *active learning* variant, we select these trees based on a notion of how informative they are, i.e. how much the parser might be improved if it knew how to parse them correctly. We approximate informativeness based on QBC, specifically, disagreement between a committee of two parsers – the less accurate we suspect a parse to be, the more likely its correction is to be informative. Thus, we rank the set of X trees (P^i_A) based on their disagreement with a second reference parser. The top Y trees (Pr^i_A) from this ordered set are manually revised and added to the training set for the next iteration.

This assessment of disagreement between two trees is based on the number of dependency relations they disagree on, which is similar to the idea of the F-complement used by Ngai and Yarowsky (2000) in their work on the comparison of rule-based and data-driven approaches to noun phrase chunking. Disagreement between two trees, t_1 and t_2 is defined as $1 - LAS(t_1, t_2)$.

We use MaltParser as the only parser in the passive learning setup and the main parser in the active learning setup. We use another dependency parser Mate (Bohnet, 2010) as our second parser in the active learning setup. Since we have 450 gold trees by this stage of development, we split them into a seed training set of 150 trees, a development set of 150 and a test set of 150. We run the two versions of the algorithm for four iterations, and on each iteration 50 (Y) parse trees were hand-corrected from a set of 200 (X).[7] This means that the final training set size for both setups is 350 trees (150 + (4*50)). However, the 4*50 training trees added to the seed training set of 150 are not the same for both setups. The set of 200 unseen sentences in each iteration is the same but, crucially, the subsets of 50 chosen for manual correction and added to the training set on each iteration are different — in the active learning setup, QBC is used to choose the subset and in the passive learning setup, the subset is chosen at random. The primary annotator carried out all the manual correction. At the final iteration, the treebank size is **650 trees**.

6.4.3 Results

The results of our bootstrapping experiments are shown in Figures 6.2 and 6.3. Figure 6.2 graphs the labelled attachment accuracy for both the passive and active setups over the four training iterations. Figure 6.3 depicts the unlabelled attachment accuracy. The highest scoring models occur on the third iteration, achieving UAS 78.49% and LAS 68.81%. All results are on our development set.

[7] The next 4 sequential blocks of 200 sentences were chosen from our randomised corpus for this experiment.

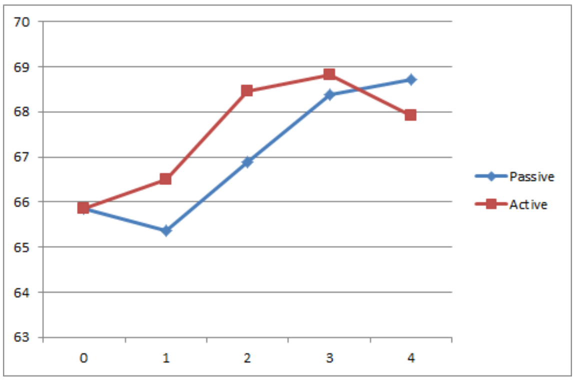


Figure 6.2: Passive versus Active Learning results on development set: **Labelled Attachment Accuracy**. The x-axis represents the number of training iterations and the y-axis the labelled attachment score.

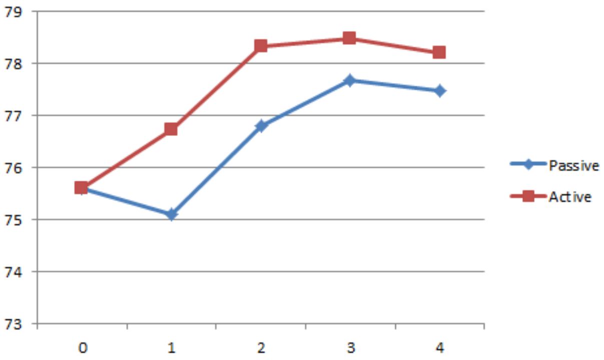


Figure 6.3: Passive versus Active Learning results on development set: **Unlabelled Attachment Accuracy**. The x-axis represents the number of training iterations and the y-axis the unlabelled attachment score.

6.4.4 Analysis

On the whole, the results in Figures 6.2 and 6.3 confirm that adding training data to our baseline model is useful and that the active learning results are superior to the passive learning results (particularly for unlabelled attachment accuracy). However, the drop in labelled attachment accuracy from the penultimate to the final iteration in the active learning setup is curious. We measured the sentence length of both the passive and active learning training sentences and found that we find one sentence in the active learning set contained 308 tokens, a feature which clearly contributed in a drop in parsing accuracy, due to an increased number of long distance dependencies that would require processing.

Since the training sentences were manually corrected before adding them to the training sets, this meant that we could measure how much correction was involved by measuring the level of disagreement between the automatic parses and their gold-standard corrected versions. This represents an approximation of annotation difficulty.

Compare the following examples to see how the automatically generated parse (Figures 6.4) can differ from the hand-corrected parse (Figure 6.5). The hand-corrected parse is considered the gold-standard parse. As we are calculating LAS (labelled attachment score) we are counting how many times the parser correctly predicts both the labels and attachment values in one sentence.

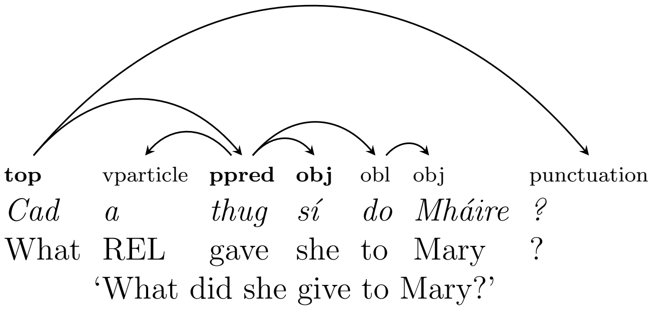

top vparticle **ppred** **obj** obl obj punctuation
Cad *a* *thug* *sí* *do* *Mháire* *?*
What REL gave she to Mary ?
'What did she give to Mary?'

Figure 6.4: Automatically generated annotation

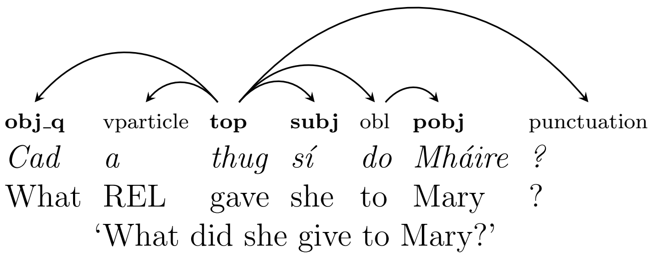

obj_q vparticle **top** **subj** obl **pobj** punctuation
Cad *a* *thug* *sí* *do* *Mháire* *?*
What REL gave she to Mary ?
'What did she give to Mary?'

Figure 6.5: Hand-corrected (gold) annotation

From comparison, we can see that the parser incorrectly predicted three labels – (i) `top` instead of `obj_q`, (ii) `ppred` instead of `top` and (iii) `obj` instead of `pobj` – and three attachments – (i) the head of *Cad* should be *thug*, (ii) the head of *thug* should be the `ROOT` and (iii) the head of *?* should be *thug*. This means that the parser only predicted both the correct label and attachment values for two tokens

	It. 1	It.2	It.3	It.4
	Correction Effort			
Passive	23.8	30.2	27.0	23.8
Active	36.7	37.6	32.4	32.8
	Average Sentence Length			
Passive	18.6	28.6	23.9	24.5
Active	18.8	25.5	24.8	35.9

Table 6.6: Differences between active and passive training sentences. Correction effort is the level of disagreement between the automatic parse and its correction (1-LAS).

in the sentence. As there are seven tokens in the sentence in total, the correction effort is 0.72, and calculated as:

$$1 - LAS = 1 - \frac{\#correctTokens}{\#tokenCount} = 1 - \frac{2}{7} = 0.72$$

The results for the correction effort involved in manually correcting all of the active learning and passive learning sentences at each iteration are shown in Table 6.6. We can see that the correction effort figures confirm that the active learning sentences required more correction than the passive learning sentences. This demonstrates that the QBC metric is successful in predicting whether a sentence is hard to parse but it also calls into doubt the benefits of active learning over passive learning, especially when resources are limited. Do the modest gains in parsing accuracy warrant the extra annotation effort involved?

We should note here that Baldridge and Palmer (2009)'s active learning study for POS tagging involved a time-based evaluation. We did not have similar technology available to us to carry out this type of evaluation, but the primary annotator's experience of this annotation task was that it took longer to correct the batch of active learning selected trees when compared to correcting the batch of passive learning selected trees.

The average sentence lengths of the active learning and passive learning sentences at each iteration are also shown in Table 6.6. We can see that there is no notable difference in average sentence length between the active and passive learning sets (apart from the final iteration).[8] We can also see that an increase in average sentence length does not necessarily correlate with an increase in annotation effort.

It is interesting that the biggest difference in sentence length is in iteration 4 where there is also a drop in active learning performance on the development set when adding them to the parser. If we examine the 50 trees that are corrected, we find one that contains 308 tokens. If this was omitted from the training data, labelled attachment accuracy rises from 67.92% to 69.13% and unlabelled attachment accuracy rises from 78.20% to 78.49%. It is risky to conclude too much from just one example but this appears to suggest that if sentences above a certain length are selected by the QBC measure, they should not be revised and added to the training set since the correction process is more likely to be lengthy and error-prone.

The test set showed similar trends to the development set. The baseline model obtains a LAS of 63.4%, the final passive model a LAS of 67.2% and the final active model a LAS of 68.0%, (increasing to 68.1% when the 308-token sentence is removed from the training set). Metric gain of one parser over another does not always represent a statistically significant gain. We calculate statistical significance values in order to assess whether the difference between the two parsers is really a reflection of improved models or just random. The difference between the active and passive learning results in this case is not statistically significant.

6.4.5 Making Use of Unlabelled Data

One criticism of the active learning approach to parser/treebank bootstrapping is that it can result in a set of trees which is an unrepresentative sample of the language since it is skewed in favour of the type of sentences chosen by the active learning informative measure.[9] One possible way to mitigate this is to add automatically (unchecked) labelled data in addition to hand-corrected data. Taking the third

[8] The notable increase in average sentence length of the active learning set in iteration 4 is due to one sentence which contains 308 tokens.

[9] See a discussion on this concern by Hal Daumé at `http://nlpers.blogspot.ie/2011/10/active-learning-far-from-solved.html`.

active learning iteration with a training set of 300 sentences as our starting point, we added automatic parses from the remaining sentences in the unlabelled set for that iteration. The unlabelled set is ordered by disagreement with the reference parser and so we incrementally added from the bottom of this set until we reached the subset of 50 trees which were manually corrected, i.e. we prioritised those parses that show the highest agreement with the reference parser first because we assumed these to be more accurate. The results, shown in Figure 6.6, demonstrate that the addition of the automatic parses makes little difference to the parsing accuracy. This is not necessarily a negative result since it demonstrates that the training sentence bias (concerns of which were highlighted above) can be adjusted without additional annotation effort and without adversely affecting parsing accuracy (at least with this limited training set size).

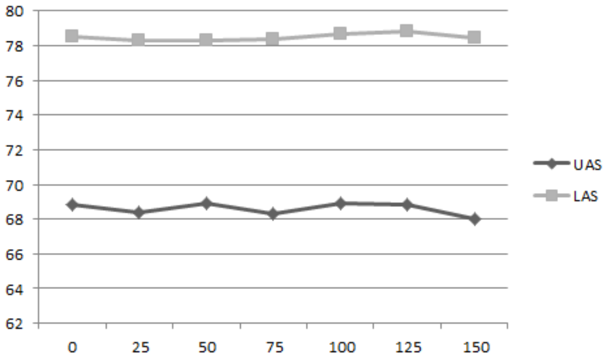


Figure 6.6: Adding Automatically Parsed Data to the Training set: the x-axis shows the number of automatically parsed trees that are added to the training set and the y-axis shows the unlabelled and labelled attachment accuracy on the development set.

6.4.6 Active Learning Conclusion

While previous studies showed positive benefits from applying active-learning methods to parsing, this differs from the task of constructing a resource that is intended to be reused in a number of ways such as a treebank. Other studies related to resource development observed potential trade-offs that arise from applying selective sampling. First, as Baldridge and Osborne (2004) show, when "creating labelled training material (specifically, for them, for HPSG parse selection) and later reusing it with other models, gains from active learning may be negligible or even negative": the simulation of active learning on an existing treebank under a particular model, with the goal of improving parser accuracy, may not correspond to a useful approach to constructing a treebank. Second, in the actual task of constructing a resource — interlinearized glossed text — Baldridge and Palmer (2009) show that the usefulness of particular example selection techniques in active learning varies with factors such as annotation expertise. They also note the importance of measures that are sensitive to the cost of annotation: the sentences that active learning methods select are often difficult to annotate as well, and may result in no effective savings in time or other measures.

In keeping with previous findings and suggestions, we have shown that we can reach a certain level of parsing accuracy with a smaller training set using active learning, but that the advantage over passive learning is relatively modest and may not be enough to warrant the extra annotation effort involved. While it is interesting to see how passive and active approaches differ in effectiveness of treebank development, the active approach cannot be the sole basis of our treebank's development. Active learning looks at only adding informative sentences to the treebank. Our resources for developing this treebank are currently limited to a 3,000 sentence POS-tagged corpus and it is expected that all of these sentences will eventually be part of the IDT.

6.5 Bootstrapping Parser Development – Semi-supervised Learning

As we have seen, developing a data-driven statistical parser relies on the availability of a parsed corpus for the language in question. Once our treebank reached over

Total # Gold Trees	Tree batch size	Source
	300	Manual annotation
	150	Automated & manual correction (incl. IAA trees)
	200	Active Learning experiments
803	153	Automated & manual correction

Table 6.7: Treebank Status

800 gold trees in size,[10] and after considering how labour- and time-intensive the annotation process is, we opted to evaluate other approaches to bootstrapping the parser.[11] The active learning approach we described earlier is regarded as a supervised learning approach. That is, the trees that are parsed are manually checked and corrected before being added to the training data. In this section we report on the evaluation of various semi-supervised approaches to bootstrapping an Irish statistical parser with a set of unlabelled sentences. This way we could ascertain how accurate parsing output could be at that stage of treebank development without solely relying on the limited gold-labelled data. Our previous preliminary attempts at adding unchecked labelled data to the training set, as we outlined in Section 6.4.5, show that while a small data set of this kind does not increase parsing accuracy, neither does it have a considerable negative impact. We therefore try semi-supervised approaches with much larger data sets here to see how they can help improve parser performance. We carried out a number of different semi-supervised bootstrapping experiments using self-training, co-training and sample-selection-based co-training:

- *Self training* involves adding a parser's output to its training data, without manual correction.

- *Co-training* involves adding the parsed output of one parser to another parser's training data, without manual correction.

[10] An additional 153 trees were added through the basic bootstrapping annotation/correction method.

[11] Note that the treebank does not grow in size through these semi-supervised experiments, as we are only interested in adding manually reviewed trees to the IDT.

- *Sample-selection-based co-training* involves adding to a parser's training data in a co-training manner, but only those trees meeting an accuracy threshold.

Our studies differ from previous similar experiments as our data was taken from a work-in-progress treebank. Thus, aside from the small treebank (803 trees) which was used for training the initial seed model and for testing, there was no additional gold-labelled data available to us at that stage to directly compare supervised and semi-supervised approaches using training sets of comparable sizes. This type of comparison could be interesting at a later date when the treebank grows in size. Nevertheless, these experiments represent the first of their kind for the Irish language.

In Section 6.5.1, we discuss how we apply a self-training approach to bootstrapping the parser. Section 6.5.2 explains how we use two parsers to apply a co-training approach to bootstrapping. In Section 6.5.3, we complement the co-training approach by using sample-selection methods to choose the parse trees that are added to the training data. Finally, in Section 6.6, we show how we explore the incorporation of morphological features in our parsing model in an attempt to overcome data-sparsity.

6.5.1 Self-Training

6.5.1.1 Related Work

Self-training, the process of training a system on its own output, has a long and chequered history in parsing. Early experiments by Charniak (1997) concluded that self-training is ineffective because mistakes made by the parser are magnified rather than smoothed during the self-training process. The self-training experiments of Steedman et al. (2003) also yielded disappointing results. Reichart and Rappaport (2007) found, on the other hand, that self-training could be effective if the seed training set was very small. McClosky et al. (2006) also report positive results from self-training, but the self-training protocol that they use cannot be considered to be pure self-training as the first-stage Charniak parser (Charniak, 2000) is retrained on the output of the two-stage parser (Charniak and Johnson, 2005). They later show that the extra information brought by the discriminative reranking phase is a factor in the success of their procedure (McClosky et al., 2008). Sagae (2010) reports positive self-training results even without the reranking phase in a domain adaptation scenario, as do Huang and Harper (2009) who employ self-training with a PCFG-LA parser.

6.5.1.2 Experimental Setup

We carried out our experiments using a small seed set of manually parsed trees and a larger, yet still relatively small, set of unlabelled sentences. In our study, we employed Malt (Nivre et al., 2006), a transition-based dependency parsing system, and Mate (Bohnet, 2010), a graph-based parser – and compared results for both.

The labelled data available to us at this stage of treebank development is 803 gold standard trees (following our basic bootstrapping and active learning annotations). This small treebank includes the 150-tree development set and 150-tree test set used in the active learning experiments discussed in Section 6.4. We use the same development and test sets for this study. As for the remaining 503 trees, we remove any trees that have more than 200 tokens. The motivation for this is two-fold: (i) we had difficulties training Mate parser with long sentences due to memory resource issues, and (ii) in keeping with the findings of the active learning experiments, the large trees were sentences from legislative text that were difficult to analyse for automatic parsers and human annotators. This leaves us with 500 gold-standard trees as our seed training data set.

For our unlabelled data, we take the next 1945 sentences from the gold standard 3,000-sentence POS-tagged corpus referred to in Section 3.1.3. When we remove sentences with more than 200 tokens, we are left with 1938 sentences in our unlabelled set.

The main algorithm for self-training is given in Algorithm 2. We carry out two separate experiments using this algorithm. In the first experiment we use Malt. In the second experiment, we substitute Mate for Malt.[12]

The steps are as follows: Initialisation involves training the parser on a labelled seed set of 500 gold standard trees (L_A^0), resulting in a baseline parsing model: M_A^i. We divide the set of gold POS-tagged sentences (U) into 6 sets, each containing 323 sentences U^i. For each of the six iterations in this experiment ($i = 1 \ldots 6$), we parse U^i. Each time, the set of newly parsed sentences (P_A) is added to the training set L_A^i to make a larger training set of L_A^{i+1}. A new parsing model (M_A^{i+1}) is then induced by training with the new training set.

Algorithm 2 Self-training algorithm

```
$A$ is a parser.
$M_A^i$ is a model of $A$ at step $i$.
$P_A^i$ is a set of trees produced using $M_A^i$.
$U$ is a set of sentences.
$U^i$ is a subset of $U$ at step $i$.
$L$ is the manually labelled seed training set.
$L_A^i$ is labelled training data for $A$ at step $i$.
Initialise:
$L_A^0 \leftarrow L$.
$M_A^0 \leftarrow Train(A, L_A^0)$
for $i = 1 \rightarrow N$ do
    $U^i \leftarrow$ Add set of unlabelled sentences from $U$
    $P_A^i \leftarrow Parse(U^i, M_A^i)$
    $L_A^{i+1} \leftarrow L_A^i + P_A^i$
    $M_A^{i+1} \leftarrow Train(A, L_A^{i+1})$
end for
```

6.5.1.3 Results

The results of our self-training experiments are presented in Figure 6.7. The best Malt model was trained on 2115 trees, at the 5th iteration (70.2% LAS). UAS scores did not increase over the baseline (79.1%). The improvement in LAS over the baseline is not statistically significant. The best Mate model was trained on 1792 trees, at the 4th iteration (71.2% LAS, 79.2% UAS). The improvement over the baseline is not statistically significant.

[12] Versions used: Maltparser v1.7 (stacklazy parsing algorithm); Mate tools v3.3 (graph-based parser).

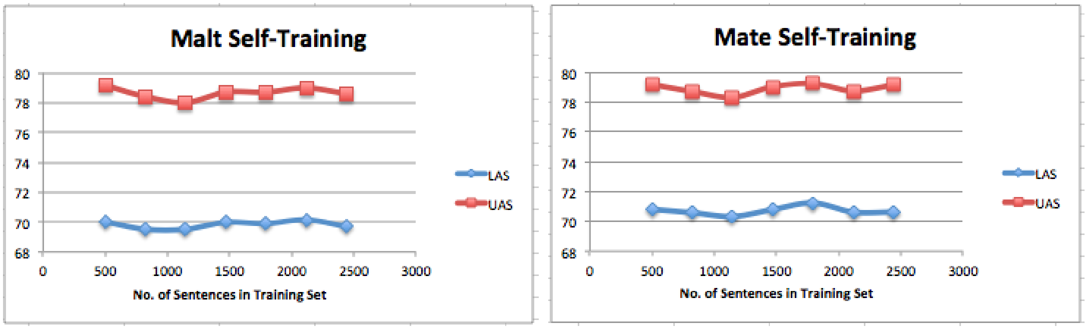


Figure 6.7: Self-Training Results on the Development Set

6.5.2 Co-Training

6.5.2.1 Related Work

Co-training involves training a system on the output of a different system. Co-training has found more success in parsing than self-training, and it is not difficult to see why this might be the case as it can be viewed as a method for combining the benefits of individual parsing systems. Steedman et al. (2003) directly compare co-training and self-training of constituency parsers and find that co-training outperforms self-training. Sagae and Tsujii (2007) successfully employ co-training in the domain adaption track of the CoNLL 2007 shared task on dependency parsing. Their methods involved (i) using two out-of-domain trained models to parse domain-specific unlabelled data, (ii) using sample selection to identify output that was identical for both models (iii) adding these (unchecked) parsed trees to the original out-of-domain labelled training set.

6.5.2.2 Experimental Setup

In this and all subsequent experiments, we use both the same training data and unlabelled data that we refer to in Section 6.5.1.2.

Our co-training algorithm is given in Algorithm 3 and it is the same as the algorithm provided by Steedman et al. (2003). Again, our experiments are carried out using Malt and Mate. This time, the experiments are run concurrently as each parser is bootstrapped from the other parser's output.

The steps are as follows: Initialisation involves training both parsers on a labelled seed set of 500 gold standard trees (L_A^0 and L_B^0), resulting in two separate baseline parsing models: M_A^i (Malt) and M_B^i (Mate). We divide the set of gold POS-tagged sentences (U) into 6 sets, each containing 323 sentences U^i. For each of the six iterations in this experiment ($i = 1 \ldots 6$), we used Malt and Mate to parse U^i. This time, the set of newly parsed sentences P_B^i (Mate output) is added to the training set L_A^i to make a larger training set of L_A^{i+1} (Malt training set). Conversely, the set of newly parsed sentences P_A^i (Malt output) is added to the training set L_B^i to make a larger training set of L_B^{i+1} (Mate training set). Two new parsing models (M_A^{i+1} and M_B^{i+1}) are then induced by training Malt and Mate respectively with their new training sets.

Algorithm 3 Co-training algorithm

```
$A$ and $B$ are two different parsers.
$M_A^i$ and $M_B^i$ are models of $A$ and $B$ at step $i$.
$P_A^i$ and $P_B^i$ are a sets of trees produced using $M_A^i$ and $M_B^i$.
$U$ is a set of sentences.
$U^i$ is a subset of $U$ at step $i$.
$L$ is the manually labelled seed training set.
$L_A^i$ and $L_B^i$ are labelled training data for $A$ and $B$ at step $i$.
Initialise:
$L_A^0 \leftarrow L_B^0 \leftarrow L$.
$M_A^0 \leftarrow Train(A, L_A^0)$
$M_B^0 \leftarrow Train(B, L_B^0)$
for $i = 1 \rightarrow N$ do
    $U^i \leftarrow$ Add set of unlabelled sentences from $U$
    $P_A^i \leftarrow Parse(U^i , M_A^i)$
    $P_B^i \leftarrow Parse(U^i , M_B^i)$
    $L_A^{i+1} \leftarrow L_A^i + P_B^i$
    $L_B^{i+1} \leftarrow L_B^i + P_A^i$
    $M_A^{i+1} \leftarrow Train(A, L_A^{i+1})$
    $M_B^{i+1} \leftarrow Train(B, L_B^{i+1})$
end for
```

6.5.2.3 Results

The results of our co-training experiment are presented in Figure 6.8. The best Malt model was trained on 2438 trees, at the final iteration (71.0% LAS and 79.8% UAS). The improvement in UAS over the baseline is statistically significant. Mate's best model was trained on 823 trees on the second iteration (71.4% LAS and 79.9%

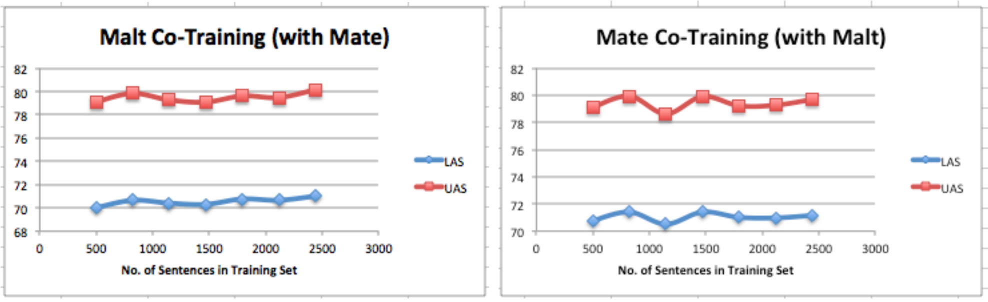


Figure 6.8: Co-Training Results on the Development Set

UAS). The improvement over the baseline is not statistically significant.

6.5.3 Sample-Selection-Based Co-Training

6.5.3.1 Related Work

Sample selection involves choosing training items for use in a particular task based on some criteria which approximates their accuracy in the absence of a label or reference. In the context of parsing, Rehbein (2011) chooses additional sentences to add to the parser's training set based on their similarity to the existing training set – the idea here is that sentences that are similar to training data are likely to have been parsed properly and so are "safe" to add to the training set. In their parser co-training experiments, Steedman et al. (2003) sample training items based on the confidence of the individual parsers (as approximated by parse probability).

In active learning research (see Section 6.4), the Query By Committee selection method (Seung et al., 1992) is used to choose items for annotation – if a committee of two or more systems disagrees on an item, this is evidence that the item needs to be prioritised for manual correction. Steedman et al. (2003) discuss a sample selection approach based on differences between parsers – if parser A and parser B disagree on an analysis, parser A can be improved by being retrained on parser B's analysis, and vice versa. In contrast, Ravi et al. (2008) show that parser *agreement* is a strong indicator of parse quality, and in parser domain adaptation, Sagae and Tsujii (2007) and Le Roux et al. (2012) use agreement between parsers to choose which automatically parsed target domain items to add to the training set.

Sample selection can be used with both self-training and co-training. We restrict our attention to co-training since our previous experiments have demonstrated that it has more potential than self-training. In the following set of experiments, we explored the role of both parser agreement and parser disagreement in sample selection in co-training.

6.5.3.2 Agreement-Based Co-Training

Experimental Setup The main algorithm for agreement-based co-training is given in Algorithm 4. Again, Malt and Mate are used. However, this algorithm differs from the co-training algorithm in Figure 3 in that rather than adding the full set of 323 newly parsed trees (P_A^i and P_B^i) to the training set at each iteration, selected subsets of these trees ($P_A^i\prime$ and $P_B^i\prime$) are added instead. To define these subsets, we identify the trees that have 85% or higher **agreement** between the two parser output sets.[13] As a result, the number of trees in the subsets differ at each iteration. For iteration 1, 89 trees reach the agreement threshold; iteration 2, 93 trees; iteration 3, 117 trees; iteration 4, 122 trees; iteration 5, 131 trees; iteration 6, 114 trees. The number of trees in the training sets is much smaller compared with those in the experiments of Section 6.5.2.

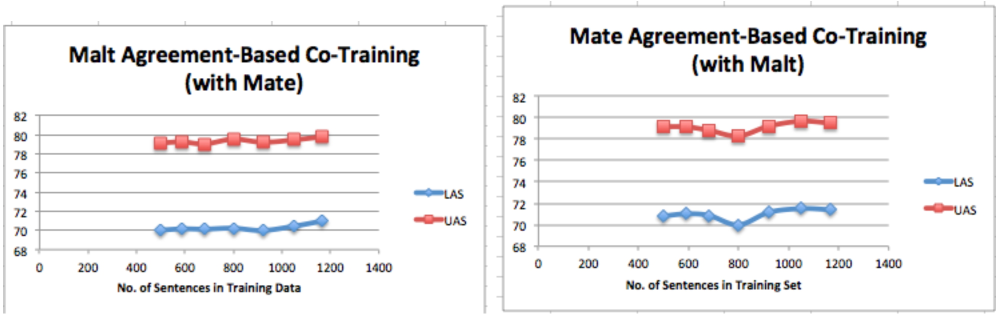


Figure 6.9: Agreement-based Co-Training Results on the Development Set

[13] We chose 85% as our cut-off as it was more relaxed than 100% agreement, yet seemed a respectable threshold for quality trees when we regarded the proportion of the agreement between trees in the development set.

Algorithm 4 Sample selection Co-training algorithm

```
$A$ and $B$ are two different parsers.
$M^i_A$ and $M^i_B$ are models of $A$ and $B$ at step $i$.
$P^i_A$ and $P^i_B$ are a sets of trees produced using $M^i_A$ and $M^i_B$.
$U$ is a set of sentences.
$U^i$ is a subset of $U$ at step $i$.
$L$ is the manually labelled seed training set.
$L^i_A$ and $L^i_B$ are labelled training data for $A$ and $B$ at step $i$.
Initialise:
$L^0_A \leftarrow L^0_B \leftarrow L$.
$M^0_A \leftarrow Train(A,L^0_A)$
$M^0_B \leftarrow Train(B,L^0_B)$
for $i = 1 \rightarrow N$ do
    $U^i \leftarrow$ Add set of unlabelled sentences from $U$
    $P^i_A \leftarrow Parse(U^i , M^i_A)$
    $P^i_B \leftarrow Parse(U^i , M^i_B)$
    $P^i_{A'} \leftarrow$ a subset of X trees from $P^i_A$
    $P^i_{B'} \leftarrow$ a subset of X trees from $P^i_B$
    $L^{i+1}_A \leftarrow L^i_A + P^i_{B'}$
    $L^{i+1}_B \leftarrow L^i_B + P^i_{A'}$
    $M^{i+1}_A \leftarrow Train(A,L^{i+1}_A)$
    $M^{i+1}_B \leftarrow Train(B,L^{i+1}_B)$
end for
```

Results The results for agreement-based co-training are presented in Figure 6.9. Malt's best model was trained on 1166 trees at the final iteration (71.0% LAS and 79.8% UAS). Mate's best model was trained on 1052 trees at the 5th iteration (71.5% LAS and 79.7% UAS). Neither result represents a statistically significant improvement over the baseline.

6.5.3.3 Disagreement-based Co-Training

Experimental Setup This experiment uses the same sample selection algorithm we used for agreement-based co-training (Figure 4). For this experiment, however, the way in which the subsets of trees ($P^i_{A'}$ and $P^i_{B'}$) are selected differs. This time we choose the trees that have 70% or higher **disagreement** between the two parser output sets. Again, the number of trees in the subsets differ at each iteration. For iteration 1, 91 trees reach the disagreement threshold; iteration 2, 93 trees; iteration 3, 73 trees; iteration 4, 74 trees; iteration 5, 68 trees; iteration 6, 71 trees.

Results The results for our disagreement-based co-training experiment are shown in Figure 6.10. The best Malt model was trained with 831 trees at the 4th iteration

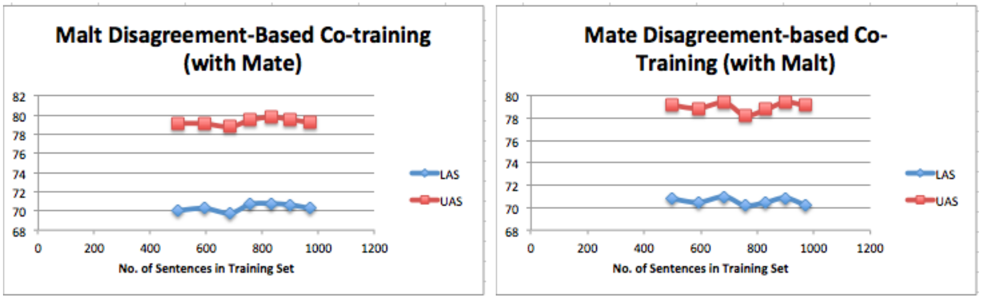


Figure 6.10: Disagreement-based Co-Training Results on the Development Set

(70.8% LAS and 79.8% UAS). Mate's best models were trained on (i) 684 trees on the 2nd iteration (71.0% LAS) and (ii) 899 trees on the 5th iteration (79.4% UAS). Neither improvement over the baseline is statistically significant.

6.5.3.4 Non-Iterative Agreement-based Co-Training

In this section, we explore what happens when we add the additional training data at once rather than over several iterations. Rather than testing this idea with all our previous setups, we choose sample-selection-based co-training where agreement between parsers is the criterion for selecting additional training data.

Experimental Setup Again, we also follow the algorithm for agreement-based co-training as presented in Figure 4. However, two different approaches are taken this time, involving only one iteration in each. For the first experiment (ACT1a), the subsets of trees ($P_A^i\prime$ and $P_B^i\prime$) that are added to the training data are chosen based on an agreement threshold of 85% between parsers, and are taken from the *full* set of unlabelled data (where $U^i = U$), comprising 1938 trees. In this instance, the subset consisted of 603 trees, making a final training set of 1103 trees.

For the second experiment (ACT1b), only trees meeting a parser agreement threshold of 100% are added to the training data. 253 trees ($P_A^i\prime$ and $P_B^i\prime$) out of 1938 trees ($U^i = U$) meet this threshold. The final training set consisted of 753 trees.

Results ACT1a proved to be the most accurate parsing model for Mate overall. The addition of 603 trees that met the agreement threshold of 85% increased the LAS and UAS scores over the baseline by 1.0% and 1.3% to 71.8 and 80.4 respectively. This improvement is statistically significant. Malt showed a LAS improvement of 0.93% and a UAS improvement of 0.42% (71.0% LAS and 79.6% UAS). The LAS improvement over the baseline is statistically significant.

The increases for ACT1b, where 100% agreement trees are added, are less pronounced and are not statistically significant. Results showed a 0.5% LAS and 0.2% UAS increase over the baseline with Malt, based on the 100% agreement threshold (adding 235 trees). Mate performs at 0.5% above the LAS baseline and 0.1% above the UAS baseline.

6.5.4 Analysis

We performed an error analysis for the Malt and Mate baseline, self-trained and co-trained models on the development set. We observed the following trends:

- All Malt and Mate parsing models confuse the `subj` and `obj` labels. A few possible reasons for this stand out: (i) It is difficult for the parser to discriminate between analytic verb forms and synthetic verb forms. For example, in the phrase *phósfainn thusa* 'I would marry you', *phósfainn* is a synthetic form of the verb *pós* 'marry' that has been inflected with the incorporated pronoun 'I'. Not recognising this, the parser decided that it is an intransitive verb, taking *thusa*, the emphatic form of the pronoun *tú* 'you', as its subject instead of object. (ii) Possibly due to a VSO word order, when the parser is dealing with relative clauses, it can be difficult to ascertain whether the following noun is the subject or object.

(38)	an	cailín	a	chonaic	mé	inné
	the	girl	REL	saw	me/I	yesterday

'the girl who saw me/ whom I saw yesterday'

Example 38 shows an ambiguous relative clause.[14] (iii) There is no passive verb form in Irish. The autonomous form is most closely linked with passive use and is used when the agent is not known or mentioned. A 'hidden' or understood subject is incorporated into the verbform. *Casadh eochair i nglas* 'a key was turned in a lock' (lit. somebody turned a key in a lock). In this sentence, *eochair* 'key' is the object.

- For both parsers, there is some confusion between the labelling of `obl` and `padjunct`, both of which mark the attachment between verbs and prepositions. Overall, Malt's confusion decreases over the 6 iterations of self-training, but Mate begins to incorrectly choose `padjunct` over `obl` instead. Mixed results are obtained using the various variants of co-training.
- Mate handles coordination better than Malt.[15] It is not surprising then that co-training Malt using Mate parses improves Malt's coordination handling whereas the opposite is the case when co-training Mate on Malt parses, demonstrating that co-training can both eliminate and introduce errors.
- Other examples of how Mate helps Malt during co-training is in the distinction between `top` and `comp` relations, between `vparticle` and `relparticle`, and in the analysis of `xcomps`.
- Distinguishing between relative and cleft particles is a frequent error for Mate, and therefore Malt also begins to make this kind of error when co-trained using Mate. Mate improves using sample-selection-based co-training with Malt.
- The sample-selection-based co-training variants show broadly similar trends to the basic co-training.

6.5.5 Test Set Results

The best performing parsing model for Malt on the development set is in the final iteration of the basic co-training approach in Section 6.5.2. The best performing

[14] Naturally ambiguous Irish sentences like this require context for disambiguation.

[15] Nivre and McDonald (2007) make a similar observation when they compare the errors made by graph and transition based dependency parsers.

Parsing Models	**LAS**	**UAS**
Development Set		
Malt Baseline:	70.0	79.1
Malt Best (co-train) :	71.0	80.2
Mate Baseline:	70.8	79.1
Mate Best (85% threshold ACT1a):	**71.8**	**80.4**
Test Set		
Malt Baseline:	70.2	79.5
Malt Best (co-train) :	70.8	79.8
Mate Baseline:	71.9	80.1
Mate Best (85% threshold ACT1a):	**73.1**	**81.5**

Table 6.8: Results for best performing models

parsing model for Mate on the development set is the non-iterative 85% threshold agreement-based co-training approach described in Section 6.5.3.4. The test set results for these optimal development set configurations are also shown in Table 6.8. The baseline model for Malt obtains a LAS of 70.2%, the final co-training iteration a LAS of 70.8%. This increase is not statistically significant. The baseline model for Mate obtains a LAS of 71.9%, and the non-iterative 85% agreement-based co-trained model obtains a LAS of 73.1%. This increase is statistically significant, with a p-value of 0.029.

6.5.6 Semi-supervised Learning Conclusion

In this chapter, we have a set of experiments where our aim was to improve dependency parsing performance for Irish, based on a small treebank seed training set size of 500 trees. In these experiments, we tried to overcome the limited treebank size by increasing the parsers' training sets using automatically parsed sentences. While we did manage to achieve statistically significant improvements in some settings, it is clear from the results that the gains in parser accuracy through semi-supervised bootstrapping methods were fairly modest. Yet, in the absence of more gold labelled data, it is difficult to know now whether we would achieve similar or improved results by adding the same amount of gold training data. This type of analysis will be interesting at a later date when the unlabelled trees used in these experiments are eventually annotated and corrected manually.

In line with similar experiments carried out on English (Steedman et al., 2003), we found that co-training is more effective than self-training. Co-training Malt on the output of Mate proved to be the most effective method for improving Malt's performance on the limited data available for Irish. Yet, the improvement is relatively small (0.6% over the baseline for LAS, 0.3% for UAS) for the best co-trained model. The best Mate results were achieved through a non-iterative agreement-based co-training approach, in which Mate is trained on trees produced by Malt which exhibit a minimum agreement of 85% with Mate (LAS increase of 1.2% and UAS of 1.4%).

6.6 Parsing with Morphological Features

Small data sets can lack sufficient examples of some linguistic phenomena in a treebank leading to data sparsity. In addition, the morphologically rich nature of the Irish language also confounds data sparsity in the treebank. Lexical data sparsity arises in treebanks of morphologically rich languages, where inflection reduces the likelihood of a word form variant being seen by the parser in training. As discussed in Section 2.2.2, the Irish language is highly inflected with both inflectional and derivational morphology. This results in our data containing a number of possible inflected forms for a given root form, making it more difficult for a parser to recognise and learn frequent patterns. For example, the verb *tabhair* 'give' is inflected to create the following verb forms in the current treebank (803 trees): *thug, dtugtar, thugtar, tugtar, tugadh, dtugadh, thabharfadh, dtabharfadh, thugann, thugaidís, dtabharfaidh.*

With this in mind, and following on from the discussion in Section 6.5.4, we carry out further parsing experiments in an attempt to make better use of morphological information during parsing. We addressed this in two ways: by reducing certain words to their lemmas and by including morphological information in the optional FEATS (features) field. The reasoning behind reducing certain word forms to lemmas is to further reduce the differences between inflected forms of the same word, and the reasoning behind including morphological information is to make more explicit the similarity between two different word forms inflected in the same way. All experiments are carried out with MaltParser and our seed training set of 500 gold trees. We focus on two phenomena: prepositional pronouns or pronominal prepositions and verbs with incorporated subjects (see Section 2.2.2).

In the first experiment, we included extra morphological information for pronominal prepositions. We ran three parsing experiments: (i) replacing the value of the surface form (FORM) of pronominal prepositions with their lemma form (LEMMA), for example *agam*→*ag*, (ii) including morphological information for pronominal prepositions in the FEATS column. For example, in the case of *agam* 'at me', we include `Per=1P|Num=Sg`, (iii) we combine both approaches of reverting to lemma form and also including the morphological features. The results are given in Table 6.9.

In the second experiment, we included morphological features for verbs with incorporated subjects: imperative verb forms, synthetic verb forms and autonomous verb forms such as those outlined in Section 6.5.4. For each instance of these verb types, we included `incorpSubj=true` in the FEATS column. The results are also given in Table 6.9.

6.6.1 Results and Conclusion

The aim of these experiments was to mitigate some of the data sparsity issues in the Irish treebank by exploiting morphological characteristics of the language, thus reducing word forms to lemmas and introducing morphological features in certain cases. These changes, however, did not bring about an increase in parsing accuracy.

The experiments on the pronominal prepositions show a drop in parsing accuracy while the experiments carried out using verb morphological information showed no change in parsing accuracy. Although the total number of correct attachments are the same, the parser output is different. In the case of inflected prepositions, it is possible we did not see any improvement because we did not focus on a linguistic phenomenon which is critical for parsing. More experimentation is necessary to establish the exact morphological characteristics of Irish that have the largest impact on parsing.

Parsing Models (Malt)	**LAS**	**UAS**
Baseline:	70.0	79.1
Lemma (Pron_Prep):	69.7	78.9
Lemma + Pron_Prep Morph Features:	69.6	78.9
Form + Pron_Prep Morph Features:	69.8	79.1
Verb Morph Features:	**70.0**	**79.1**

Table 6.9: Results with morphological features on the development set

6.7 Cross-lingual Transfer Parsing

Total # Gold Trees	**Tree batch size**	**Source**
	300	Manual annotation
	150	Automated & manual correction (incl. IAA trees)
	200	Active Learning experiments
	153	Automated & manual correction
1020	217	Automated & manual correction

Table 6.10: Treebank Status

As previously discussed, annotating additional trees to increase the parser's training set is a labour-intensive task. The small size of the treebank affects the accuracy of any statistical parsing models learned from this treebank. Therefore, we investigate whether training data from other languages could be successfully utilised to improve Irish parsing. Leveraging data from other languages in this way is known as cross-lingual transfer parsing.

Cross-lingual transfer parsing involves training a parser on one language, and parsing data of another language. McDonald et al. (2011) describe two types of cross-lingual parsing, direct transfer parsing in which a delexicalised version of the source language treebank is used to train a parsing model which is then used to parse the target language, and a more sophisticated projected transfer approach in which the direct transfer approach is used to seed a parsing model which is then trained to obey source-target constraints learned from a parallel corpus. These experiments revealed that languages that were typologically similar were not necessarily the best source-target pairs, sometimes due to variations between their language-specific annotation schemes. In more recent work, however, McDonald et al. (2013) reported improved results on cross-lingual direct transfer parsing using the UD13 universal annotation scheme, to which six chosen treebanks are mapped for uniformity purposes.[16] Underlying the experiments with this new annotation scheme is the universal part-of-speech (POS) tagset designed by Petrov et al. (2012). While their results confirm that parsers trained on data from languages in the same language group (e.g. Romance and Germanic) show the most accurate results, they also show that training data taken across language groups also produces promising results. In this section, we present our experiments with direct transfer cross-lingual parsing, using Irish as the target language.

Since the Irish language belongs to the Celtic branch of the Indo-European language family, the natural first step in cross-lingual parsing for Irish would be to look to those languages of the Celtic language group, i.e. Welsh, Scots Gaelic, Manx, Breton and Cornish, as a source of training data. However, these languages are just as, if not further, under-resourced. Thus, we explore the possibility of leveraging from the languages of the UD13 universal dependency treebanks (McDonald et al., 2013) that are discussed in Section 5.

In Section 6.7.1, we describe the datasets used in our experiments and explain the experimental design. In Section 6.7.2, we present the results, which we then discuss in Section 6.7.3.

6.7.1 Data and Experimental Setup

Firstly, we present the datasets used in our experiments and explain how they are used. Irish is the target language for all our parsing experiments.

[16] See Section 5 for more detail on UD13.

Universal Irish Dependency Treebank UD13 is the universal version of the Irish Dependency Treebank (now containing **1020 gold-standard trees**),[17] which have been mapped to Petrov et al. (2012)'s Universal POS tagset and McDonald et al. (2013)'s Universal Dependency Annotation Scheme (see Section 5). In order to establish a monolingual baseline against which to compare our cross-lingual results, we performed a five-fold cross-validation by dividing the full data set into five non-overlapping training/test sets. We also tested our cross-lingual models on a *delexicalised* version of this treebank, in which we replaced all tokens and lemmas (language-specific data) with XX, and fine- and coarse-grained POS tags with the universal POS tags.

Transfer source training data For our direct transfer cross-lingual parsing experiments, we used 10 of the UD13 standard version harmonised training data sets[18] made available by McDonald et al. (2013): Brazilian Portuguese (PT-BR), English (EN), French (FR), German (DE), Indonesian (ID), Italian (IT), Japanese (JA), Korean (KO), Spanish (ES) and Swedish (SV). For the purposes of uniformity, we selected the first 4447 trees from each treebank – to match the number of trees in the smallest data set (Swedish). We delexicalised all treebanks and used the universal POS tags as both the coarse- and fine-grained values.[19] We trained a parser on all 10 source data sets outlined and use each induced parsing model to parse and test on a *delexicalised* version of the Universal Irish Dependency Treebank.

Largest transfer source training data – Universal English Dependency Treebank English has the largest source training data set (sections 2-21 of the Wall Street Journal data in the Penn Treebank Marcus et al. (1993) contains 39,832 trees). As with the smaller transfer datasets, we delexicalised this dataset and use the universal POS tag values only. We experimented with this larger training set in

[17] An additional 217 trees were added through the basic bootstrapping annotation/correction method.

[18] Version 2 data sets downloaded from `https://code.google.com/p/uni-dep-tb/`

[19] Note that the downloaded treebanks had some fine-grained POS tags that were not used across all languages: e.g. VERBVPRT (Spanish), CD (English).

order to establish whether more training data helps in a cross-lingual setting.

Parser and Evaluation Metrics We used MaltParser (a transition-based dependency parsing system) for all of our experiments. In each case we report Labelled Attachment Score (LAS) and Unlabelled Attachment Score (UAS).[20]

6.7.2 Results

All cross-lingual results are presented in Table 6.11. Note that when we trained and tested on Irish (our monolingual baseline), we achieved an average accuracy of 78.54% (UAS) and 71.59% (LAS) over the five cross-validation runs. The cross-lingual results were substantially lower than this baseline. The LAS results range from 0.84% (JA) to 43.88% (ID) and the UAS from 16.74% (JA) to 61.69% (ID).

Experiment	**Baseline**
Training	GA
UAS	78.54%
LAS	71.59%

Experiment	***SingleT***									
Training	EN	FR	DE	ID	IT	JA	KO	PT-BR	ES	SV
UAS	51.72	56.84	49.21	**61.69**	50.98	16.74	18.02	57.31	57.00	49.95
LAS	35.03	37.91	33.04	**43.88**	37.98	0.84	9.35	42.13	41.94	34.02

Experiment	***MultiT***	***LargestT***
Training	All	EN
UAS	57.69	51.59
LAS	41.38	33.97

Table 6.11: Multi-lingual transfer parsing results

A closer look at the single-source transfer parsing evaluation results (*SingleT*) shows that some language sources are particularly strong for parsing accuracy of certain labels. For example, `ROOT` (for Indonesian), `adpobj` (for French) and `amod` (for Spanish). In response to these varied results, we explored the possibility of combining the strengths of all the source languages (*multi-source direct transfer* (*MultiT*) – also implemented by McDonald et al. (2011)). A parser is trained on a concatenation of all the delexicalised source data described in Section 7.4.1 and tested on the full delexicalised Universal Irish Dependency Treebank. Combining

[20] All scores are micro-averaged meaning that they are calculated using weighting, which is based on the number of tokens in a sentence, instead of just calculating the scores at a sentence level.

all source data produced parsing results of 57.69% (UAS) and 41.38% (LAS), which is outperformed by the best individual source language model (Indonesian).

Parsing with the large English training set (*LargestT*) yielded results of 51.59% (UAS) and 33.97% (LAS) compared to a UAS/LAS of 51.72/35.05 for the smaller English training set. We investigated more closely why the larger training set did not improve performance by incrementally adding training sentences to the smaller set – none of these increments reveal any higher scores, suggesting that English is not a suitable source training language for Irish.

Although the best cross-lingual model failed to outperform the monolingual model, we looked at combining the strengths of the Indonesian and Irish treebanks instead. We performed 5-fold cross-validation on the combined Indonesian and Irish data sets. The results, 73.6% (UAS) and 65.03% (LAS), did not improve over the Irish model. We then analysed the extent of their complementarity by counting the number of sentences where the Indonesian model outperformed the Irish model. This happened in only 20 cases, suggesting that there is no benefit in using the Indonesian data over the Irish data nor in combining them at the sentence-level.

6.7.3 Discussion

McDonald et al. (2013)'s single-source transfer parsing results show that languages within the same language groups make good source-target pairs. They also show reasonable accuracy of source-target pairing across language groups. For instance, the baseline when parsing French is 81.44% (UAS) and 73.37% (LAS), while the transfer results obtained using an English treebank are 70.14% (UAS) and 58.20% (LAS). Our baseline parser for Irish yields results of 78.54% (UAS) and 71.59% (LAS), while Indonesian-Irish transfer results are 61.69% (UAS) and 43.88% (LAS).

The lowest scoring source language is Japanese. This parsing model's output shows less than 3% accuracy when identifying the `ROOT` label. This suggests the effect that the divergent word orders have on this type of cross-lingual parsing – VSO (Irish) vs SOV (Japanese). Another factor that is likely to be playing a role is the length of the Japanese sentences. The average sentence length in the Japanese training data is only 9 words, which means that this dataset is comparatively smaller than the others. It is also worth noting that the UD13 Japanese treebank uses only 15 of the 41 universal labels (the universal Irish treebank uses 30 of these labels).

As our best performing model (Indonesian) is an Austronesian language, we investigate why this language does better when compared to Indo-European languages. We compare the results obtained by the Indonesian parser with those of the English parser (*SingleT*). Firstly, we note that the Indonesian parser captures nominal modification much better than English, resulting in an increased precision-recall score of 60/67 on `compmod`. This highlights that the similarities in noun-noun modification between Irish and Indonesian helps cross-lingual parsing. In both languages the modifying noun directly follows the head noun, e.g. 'the statue of the hero' translates in Irish as *dealbh an laoich* (lit. statue the hero); in Indonesian as *patung palawan* (lit. statue hero). Secondly, our analysis shows that the English parser does not capture long-distance dependencies as well as the Indonesian parser. For example, we have observed an increased difference in precision-recall of 44%-44% on `mark`, 12%-17.88% on `cc` and 4%-23.17% on `rcmod` when training on Indonesian. Similar differences have also been observed when we compare with the French and English (*LargestT*) parsers.

6.7.4 Cross-lingual Transfer Parsing Conclusion

In this study, we had hoped that we would be able to identify a way to bootstrap the development of the Irish Dependency Parser through the use of delexicalised treebanks of other languages that were annotated with the UD13 Annotation Scheme. Uniformity as a result of the UD13 Annotation scheme and the delexicalised nature of the treebanks allows for a more conclusive comparison and sharing of resources across languages. While the current treebank data might capture certain linguistic phenomena well, we expected that some cross-linguistic regularities could be taken advantage of through cross-lingual transfer parsing.

While none of the cross-lingual parsing models outperformed our baseline monolingual model, our analysis of our best performing parsing model (Indonesian) and our Irish parsing model output revealed similar linguistic traits that explained their compatibility. We could also see that while some of the other language source data contributed to accurately predicting certain labels, overall, their disparities with Irish in general resulted in much lower scores.

Our relatively low cross-lingual parsing results suggest that this is not a suitable bootstrapping approach to our parser development. Yet, our study has provided an interesting insight into the difficulties of parsing Irish, both from the perspective of limited training data and in comparison to the linguistic features of other languages.

6.8 Summary

We started this chapter with an overview of dependency parsing and a description of how dependency parsers work. We gave an overview of the history of dependency parsing and discuss the shift from traditional constituency parsing over time through the development of resources. We then stepped through the parsing process of an example Irish sentence to fully illustrate how transition-based dependency parsing works.

We then moved on to discuss the link between the Irish Dependency Treebank and statistically-driven parsers. When training a parser with a treebank for a new language, it is difficult to accurately guess how large that treebank should be to induce a sufficiently accuracy parsing model. There are many factors that can affect parser accuracy, ranging from data set size, to morphological features resulting in further data sparsity, to the type of parsing systems used for a particular language. In this chapter, we explored the effects these factors have on Irish parsing throughout the course of the treebank development. In addition, we also explored various semi-automated approaches to bootstrapping the parser's development, through (i) bootstrapping the treebank's development with a parser (ii) leveraging the Irish

Stage	**Treebank size**	**Training size**	**Test type/size**	**LAS**	**UAS**
Baseline	300	270	Cross-val. (10-fold)	63.3	73.1
Baseline (post IAA)	300	270	Cross-val.(10-fold)	66.5	76.3
Passive Learning	650	350	Dev set (150)	68.4	77.7
Active Learning	650	350	Dev set (150)	68.8	78.5
Co-training	803	1103†	Dev set (150)	71.8	80.4
Cross-lingual (UD – ga)	1020	816	Cross-val. (5-fold)	71.6	78.5
Final Treebank	1020	816	Cross-val. (5-fold)	71.4	80.1

Table 6.12: Summary of treebank parsing results throughout development. (Note: † indicates unchecked trees.)

parser's own uncorrected output as training data and (iii) leveraging treebanks of other languages as training data. Table 6.12 highlights the parsing trends accruing to the treebank's development stages.

We began by establishing a baseline score of UAS 73.1% and LAS 63.3% on a small seed set of 300 manually labelled trees. Our baseline increased to UAS 76.3% and LAS 66.5% following an improvement of our labelling scheme, annotation guide and treebank annotation. These updates were made following our IAA study discussed in Section 3.2.

From there we experimented with various bootstrapping approaches in order to speed up the annotation process and reduce the manual effort required. Firstly, we showed how we increased the treebank from 300 to 450 trees by following a basic bootstrapping algorithm which allows a parser, trained on the seed set to pre-parse the next set of trees for manual correction.

We then expanded on this approach through active learning where sample-selection allowed us to choose the most informative trees for correction, thus increasing the accuracy of the treebank more quickly. With this method, our treebank increased in size from 450 to 650 trees. Further parsing experiments show an increase in accuracy to UAS 78.49% and LAS 68.81%. Yet, we concluded that the extra effort involved in this approach outweighed the parser accuracy gains.

Following from these supervised learning approaches, we then discussed some semi-supervised learning approaches where we attempted to improve the parsers performance by incrementing the gold training data with automatically generated parse trees that are not corrected. At this stage, we had increased the gold standard treebank to 803 trees through the basic bootstrapping annotation/correction method. The highest parsing accuracy we achieved through adding uncorrected trees was UAS 80.4% and LAS 71.8%.

We also showed how we attempted to address the data sparsity issues that arise from the morphologically rich nature of Irish. While no improvements were gained from these experiments, we note that we would like to explore the role of morphology in Irish parsing through future versions of the treebank.

We described how we used treebanks from other languages to bootstrap our parser's training data, which is now at 1020 trees (again, through further basic bootstrapping). We showed how a uniform Universal Dependency annotation scheme (McDonald et al., 2013) made it possible to perform multiple cross-lingual transfer parsing experiments in an attempt to improve our parser results through leveraging the treebank data of other languages. Despite the availability of extra training data from other treebanks, none of the cross-lingual models outperformed our Irish parser's current accuracy of UAS 78.54% and LAS 71.59%.

Finally, experiments carried out on the final version of the 1020-tree Irish Dependency Treebank (using 5-fold cross-validation) show scores of UAS 71.4% and LAS 80.1%.

A clear conclusion we can make from the results of all of our attempts to semi-automate or bootstrap the Irish parser, is that using a solid gold-annotated treebank as training data is a fundamental requirement in achieving decent parsing accuracy. We have therefore shown that human annotation effort cannot be easily replaced by technology for this type of task.

In the next chapter, we move from processing standard, well-structured Irish text to user-generated text found on Twitter, and explain the ways in which this different variation of text influences NLP development; but to do this, we must return to the task of POS tagging, as other work (Gimpel et al., 2011) suggests that POS taggers cannot be straightforwardly applied to such text.

Chapter 7

Irish Twitter POS-tagging

The Irish language is listed by UNESCO as an endangered language. By law, Irish is a compulsory subject in primary and secondary schools but the majority of school leavers outside the Gaeltacht areas (Irish speaking regions) will not continue to use it once they have graduated. However, everyday use outside of academic environments has seen a recent resurgence on social media platforms such as Facebook and Twitter. This evolution of a new generation of Irish language online users shows promising signs in terms of the future direction of the language.

We have thus been inspired to develop new language resources tailored to assisting research on the growth of the Irish language in social media. The linguistic variations that have developed in the language through social media use, without the constraints of academic assessment, are of particular interest not only to NLP research groups but also to sociolinguists.

The language style variation used on social media platforms, such as Twitter for example, is often referred to as noisy user-generated text. Tweets can contain typographical errors and ungrammatical structures that pose challenges for processing tools that have been designed for and tailored to high quality, well-edited text such as that found in newswire, literature and official documents. Previous studies, Foster et al. (2011) and Petrov and McDonald (2012) for example, have explored the effect that the style of language used in user-generated content has on the performance of standard NLP tools. Other studies by Gimpel et al. (2011), Owoputi et al. (2013), Avontuur et al. (2012), Rehbein (2013) and Derczynski et al. (2013) (POS-tagging), Ritter et al. (2011) (named entity recognition), Kong et al. (2014) and Seddah et al. (2012) (parsing) have shown that NLP tools and resources need to be adapted to cater for the linguistic differences present in such text.

When considering data-driven NLP tasks, a lack of resources can also produce additional challenges. Therefore our study on Irish (a low-resourced language) in this context proves interesting on many levels. Through our analysis, we examine the impact of noisy user-generated text on the existing resources for the language. We find that the linguistic variation in Irish Twitter data differs greatly from standard written Irish, which has a negative result on the performance of existing NLP tools. From a lesser-resourced language point of view, we explore options for leveraging from existing resources to produce a new domain-adapted POS-tagger for processing Irish Twitter data. Our steps involve:

- defining a new POS tagset for Irish tweets
- providing a mapping from the PAROLE Irish POS-tagset to this new one
- manually annotating a corpus of 1537 Irish tweets
- training three statistical taggers on our data and reporting results

Our work is inspired by Gimpel et al. (2011) and Owoputi et al. (2013)'s earlier work on English tweet POS tagging. Our study is similar to Gimpel et al. (2011) in that they designed a Twitter-specific POS tagset, pre-tagged $1,827$ tweets with an out-of-domain tagger (the WSJ-trained Stanford POS tagger (Toutanova et al., 2003)), manually corrected by 17 annotators over two months, and trained POS tagging models on this data. They, however, designed and built a tagging system tailored to English tweets that included features of frequently capitalised English tokens, distributional similarity obtained from a large set of unlabelled English tweets and English phonetic normalisation. These are all resources we do not have for Irish.

Owoputi et al. (2013) extended their work by improving the tagging system with unsupervised clustering methods, using 56 million English tweets (an approach also taken by Rehbein (2013) for POS tagging German tweets). Again, this is an option not available to our Irish study due to the limited number of Irish tweets. They also had a number of various test sets available to evaluate their work. Our testing was limited to a subset of our single Twitter data set.

On the other hand, our work differs from theirs in a couple of ways. We do not build a tagging system, instead we re-train existing versions of the Stanford Tagger (Toutanova et al., 2003) and Owoputi et al. (2013)'s Tweet NLP ARK Tagger with our Irish Twitter data. The other ways in which our work differs is mainly due to linguistic characterisations in Irish tweets. Our work extends to evaluating Morfette (Chrupała et al., 2008), a tagger that uses lemma information as features in its tag predictions. In contrast to English, Irish is a highly inflected language. Inflection can cause data sparsity in a data set and thus impact tagging accuracy. Another issue for Irish tweet processing is code-switching, a phenomenon that did not impact the experiments on English tweet processing. In addition, the tweet tokeniser used for English tweets (and which we used for Irish tweets) is tailored to the English language, which meant our data required an additional post-processing step before annotation. Our annotation was carried out over three months, by two annotators.

This chapter is structured as follows: Section 7.1 gives a summary of Twitter and issues specific to the Irish Twitter data. Section 7.2 discusses the new part-of-speech tagged corpus of Irish tweets. In Section 7.3 we describe our inter-annotator agreement study and the observations we note from annotator disagreements. In Section 7.4, we report our tagging accuracy results on three state-of-the-art statistical taggers.

7.1 Irish Tweets

Twitter is a micro-blogging platform which allows users (*tweeters*) to create a social network through sharing or commenting on items of social interest such as ideas, opinions, events and news. Tweeters can post short messages called *tweets*, of up to 140 characters in length, that can typically be seen by the general public, including the user's *followers*. Tweets can be classified by topic by using *hashtags* (e.g. #categoryname) and linked to other tweeters through the use of *at-mentions* (e.g. @username).

The first tweets in Irish appeared not long after the launch of Twitter in 2006, and there have been more than a million tweets in Irish since then, by over 8000 tweeters worldwide.[1]

The social nature of tweets can result in the use of informal text, unstructured or ungrammatical phrases, and a variety of typographical errors. The 140 character limit can also lead to truncated ungrammatical sentences, innovative spellings, and word play, such as those discussed by Eisenstein (2013) for English. From our analysis, this phenomenon appears to extend also to Irish tweets.

In Figure 7.1, we provide an example of an Irish tweet that contains some of these NLP challenges:

Freezing i dTra Li,Ciarrai chun cinn le cuilin.
Freezing i dTr**á** Lí, **tá** Ciarra**í** chun cinn le c**ú**il**í**n.
'Freezing in Tralee, Kerry (is) ahead by a point.'

Figure 7.1: Example of noisy Irish tweet

Diacritics Irish, in its standard orthography, marks long vowels with diacritics (á,é,í,ó,ú). Our analysis of Irish tweets revealed that these diacritics are often replaced with non-accented vowels (cúilín => cuilin). There are a number of word pairs that are differentiated only by the presence or absence of these diacritics (for example, *cead* 'permission' : *céad* 'hundred'). There are many possible reasons for

[1] `http://indigenoustweets.com/ga/` (accessed June 2015)

omitting diacritics, including shortening the time required to tweet (the example tweet in Figure 7.1 is from a spectator at a Gaelic Football match), a lack of knowledge on how to find diacritics on a device's keyboard, carelessness, or uncertainty about the correct spelling.

Code-switching Alternating between English and Irish is common in our dataset. This is unsurprising as virtually all Irish speakers are fluent English speakers, and many use English as their first language in their daily lives. In the example given, there is no obvious reason why "Freezing" was used in place of various suitable Irish words (e.g. *Préachta*), other than perhaps seeking a more dramatic effect. Sometimes, however, English is understandably used when there is no suitable Irish term in wide use, for example 'hoodie' or 'rodeo-clown'. Aside from occurring at an intra-sentential level, code-switching at an inter-sentential level is also common in Irish: *an t-am seo an t7ain seo chugainn bei 2 ag partyáil le muintir Ráth Daingin! Hope youre not too scared #upthevillage.* In total, of the 1537 tweets in our gold-standard corpus, 326 (21.2%) contain at least one English word with the tag G.[2]

Verb drop We can see in this example that the verb *tá* 'is' has been dropped. This is a common phenomenon in user-generated content for many languages. The verb is usually understood and can be interpreted through the context of the tweet.

Spacing Spacing after punctuation is often overlooked (i) in an attempt to shorten messages or (ii) through carelessness. In certain instances, this can cause problems when tokenizing tweets; Li,Ciarrai => Li, Ciarrai.

Phonetic spelling Linguistic innovations often result from tweeters trying to fit their message into the 140 character limit. Our dataset contains some interesting examples of this phenomenon occurring in Irish. For example *t7ain* is a shortened version of *tseachtain* 'week'. Here the word *seacht* 'seven' is shortened to its numeral

[2] The tag G is used for foreign words, abbreviations, items and unknowns, as shown in Table 7.1

form and the initial mutation *t* remains attached. Other examples are *gowil (go bhfuil), beidir (b'fhéidir), v (bhí).*

Abbreviations Irish user-generated text has its own set of frequently used phrase abbreviations – referred to sometimes as text-speak. Forms such as *mgl:maith go leor*, 'fair enough' and *grma:go raibh maith agat* 'thank you' have been widely adopted by the Irish language community.

The linguistic variation of Irish that is used in social media is relatively unexplored, at least not in any scientific manner. We expect therefore that the part-of-speech tagged corpus and taggers that we have developed for Irish language tweets will contribute to further research in this area.

7.2 Building a corpus of annotated Irish tweets

Unlike rule-based systems, statistical data-driven POS-taggers require annotated data on which they can be trained. Therefore, we build a gold-standard corpus of 1537 Irish tweets annotated with a newly defined Twitter POS tagset. The following describes this development process.

7.2.1 New Irish Twitter POS tagset

As discussed in Chapter 5, the rule-based Irish POS-tagger (Uí Dhonnchadha and van Genabith, 2006) for standard Irish text is based on the PAROLE Morphosyntactic Tagset (ITÉ, 2002). We used this as the basis for our Irish Twitter POS tagset. We were also inspired by the English-tweet POS tagset defined by Gimpel et al. (2011), and have aimed to stay closely aligned to it in order to facilitate any future work on cross-lingual studies.

We started by selecting a random sample of 500 Irish tweets to carry out an initial analysis and define our tagset. We choose to keep our tagset at a comparable level of granularity to the English tagset for comparison purposes to Gimpel et al. (2011) and Owoputi et al. (2013)'s results. While our tagset is also closely aligned with the English-tweet POS tagset, we introduce the following tags that the English set does not use:

- `VN`: **Verbal Noun** As we discussed earlier in Section 2.2.1, while Irish verbal nouns have clear verbal origins, they inflect morphologically as nouns. Verbal nouns have a distinct role from common or proper nouns. Firstly, verbal nouns are used to denote non-finite phrases. Sometimes this is indicated by a preceding infinitive marker *a* (Example 39), and sometimes not (Example 40). These cases need to be differentiated from regular nouns.

  (39) an locht a **chur** orthu
  the blame INF put on-them
  'to put the blame on them'

  (40) beidh ort **teacht** ar_ais
  will-be on-you come back
  'you will have to come back'

  Also, progressive aspectual phrases in Irish are denoted by the preposition *ag* followed by a verbal noun, as shown in Example 41.

  (41) bhí @aodhanodea ag **labhairt**
  was @aodhanodea at speaking
  '@aodhanodea was speaking'

  We therefore choose to differentiate between N and VN to avoid losing this verbal information in what would otherwise be a regular prepositional phrase.

- `#MWE`: **Multiword hashtag** These are hashtags containing strings of words used to categorise a text (e.g. #WinterIsComing). We retain information on the multi-word nature of these hashtags in order to facilitate future syntactic analysis efforts. Multiword tagging will also assist future work on named entity recognition. For example *#LáNollag* is a hashtag containing two nouns *Lá* 'Day', *Nollaig* 'Christmas' (i.e. Christmas Day).

We also adapt the `T` particle from the English tagset to suit Irish linguistic features.

- `T`: **Particle** We extend the `T` tag to not only cover verb particles, but all other Irish particles: relative particles, surname particles, infinitive particles, numeric particles, comparative particles, the vocative particle, and adverbial particles.

We do not use the following tags from the English set: `S, Z, L, M, X, Y,` as the linguistic cases they apply to do not occur in either standard or non-standard Irish.

- S: nominal + possessive (e.g. someone's)
- Z: proper noun + possessive (e.g. America's)
- L: nominal + verbal (he's)
- M: proper noun + verbal (Mark'll)
- X: existential there, predeterminers
- Y: X + verbal (there's)

The final set of 21 POS-tags is presented in Table 7.1.

Most of the tags in the tagset are intuitive to an Irish language speaker, based on a knowledge of Irish grammar. However, some tags require specific explanation in the guidelines. Hashtags and at-mentions can be a syntactic part of a sentence or phrase within a tweet. When this is the case, we apply the relevant syntactic POS tag. For example, *Beidh*$_V$ *mé*$_O$ *ar*$_P$ *chlár*$_N$ *@SplancNewstalk*$_\wedge$ *anocht*$_R$ *ag*$_P$ *labhairt*$_{VN}$ *leis*$_P$ *@AnRonanEile*$_\wedge$ *faoi*$_P$ *#neknomination*$_N$ 'I will be on @SplancNewstalk tonight speaking to @AnRonanEile about #neknomination'.[3] Otherwise if they are not

[3] $\wedge$ is the tag used for proper nouns.

Tag	**Description (PAROLE TAGS)**
N	common noun (Noun, Pron Ref, Subst)
^	proper noun (Prop Noun)
O	pronoun (Pron Pers, Pron Idf, Pron Q, Pron Dem)
VN	verbal noun (Verbal Noun)
V	verb (Cop, Verb†)
A	adjective (Adj, Verbal Adj, Prop Adj)
R	adverb (Adv†)
D	determiner (Art, Det)
P	preposition, prep. pronoun (Prep†, Pron Prep)
T	particle (Part†)
,	punctuation (Punct)
&	conjunction (Conj Coord, Conj Subord)
$	numeral, quantifier (Num)
!	interjection (Itj)
G	foreign words, abbreviations, item (Foreign, Abr, Item, Unknown)
~	discourse marker
#	hashtag
#MWE	multi-word hashtag
@	at-mention
E	emoticon
U	URL/email address/XML (Web)

Table 7.1: Mapping of Irish Twitter tagset to PAROLE tagset. († indicates the entire fine-grained set for that coarse tag is relevant.)

part of the syntactic structure of the tweet (typically appended or prepended to the main tweet text), they are tagged as @ and # (or #MWE). In our gold standard corpus, 554 out of 693 hashtags (79.9%), and 1604 out of 1946 at-mentions (82.4%) are of this non-syntactic type.

With some Twitter clients, if a tweet exceeds the 140 character limit, the tweet is truncated and an ellipsis is used to indicate that some text is missing. We leave this appended to the final (usually partial) token (which was often a URL). We marked these cases as `G`. For example *http://t.co/2nvQsxaIa7...*.

Some strings of proper nouns contain other POS elements, such as determiners and common nouns. Despite being a proper noun phrase syntactically, we tag each token as per its POS. For example, *Cú*$_{\wedge}$ *na*$_{D}$ *mBaskerville*$_{\wedge}$ 'The Hound of the Baskervilles'.

7.2.2 Tweet pre-processing pipeline

We took a random sample of 1550 Irish tweets from the approximately 950,000 Irish tweets that had been sent between Twitter's launch in 2006 and September 2014, and processed them as follows:

(1) We tokenised the set with Owoputi et al. (2013)'s version of `twokenise`,[4] which works well on web content features such as emoticons and URLs.

(2) Using a list of multiword units from Uí Dhonnchadha (2009)'s rule-based Xerox FST tokeniser,[5] we rejoined multiword tokens that had been split by the language-independent tokenizer (e.g. the compound preposition *go_dtí*).

(3) Using regular expressions, we then split tokens with the contractions *b' (ba), d' (do), m' (mo)* prefixes. For example *b'fhéidir* 'maybe'; *d'ith* 'ate'; *m'aigne* 'my mind'.

(4) We took a bootstrapping approach by pre-tagging and lemmatising the data with the rule-based Irish POS-tagger first, and then automatically mapped the tags to our new Twitter-specific tagset.

(5) In cases where the rule-based tagger failed to produce a unique tag, we used a simple bigram tag model (trained on the gold-standard POS-tagged corpus from Uí Dhonnchadha (2009) – see Section 7.4.1) to choose the most likely tag from among those output by the rule-based tagger.

[4] Available to download from `http://www.ark.cs.cmu.edu/TweetNLP/#pos`

[5] Available to download from `https://github.com/stesh/apertium-gle/tree/master/dev/irishfst`

(6) Finally, we manually corrected both the tags and lemmas to create a gold-standard corpus.

7.2.3 Annotation

The annotation task was shared between two annotators.[6] Correction of the first 500 tweets formed a basis for assessing both the intuitiveness of our tagset and the usability of our annotation guide. Several discussions and revisions were involved at this stage before finalising the tagset. The next 1000 tweets were annotated in accordance with the guidelines, while using the first 500 as a reference. At this stage, we removed a small number of tweets that contained 100% English text (errors in the language identifier). All other tweets containing non-Irish text represented valid instances of code-switching.

The annotators were also asked to verify and correct the lemma form if an incorrect form was suggested by the morphological analyser. All other tokeniser issues, often involving Irish contractions, were also addressed at this stage. For example *Tá'n* $->$ *Tá an*.

7.3 Inter-Annotator Agreement

As we previously showed in Section 3.2, Inter-Annotator agreement (IAA) studies are carried out during annotation tasks to assess consistency, levels of bias, and reliability of the annotated data. We carry out a similar study here on our POS tagging task. For our study, we chose 50 random Irish tweets, which both annotators tagged from scratch. This differed from the rest of the annotation process, which was semi-automated.[7] However, elimination of possible bias towards the pre-annotation output allowed for a more disciplined assessment of agreement level between the

[6] The author was the primary annotator and Eimear Maguire, a recent computer science/ linguistics graduate was the second annotator.

[7] This also differed from the parsing IAA study in which the annotators corrected pre-parsed text.

annotators. We achieved an agreement rate of 90% and a κ score (Cohen, 1960) of 0.89.

Smaller tagsets make an annotation task easier due to the constraint on choices available to the annotator, and is certainly one reason for our high IAA score. This result also suggests that the tagging guidelines were clear and easy to understand. A closer comparison analysis of the IAA data explains some disagreements. The inconsistency of conflicts suggests that the disagreements arose from human error. Some examples are given below.

Noun vs Proper Noun The word *Gaeilge* 'Irish' was tagged on occasion as `N` (common noun) instead of $\wedge$ (proper noun). This also applied to some proper noun strings such as *Áras an Uachtaráin* (the official name of the President of Ireland's residence).

Syntactic at-mentions A small number of at-mentions that were syntactically part of a tweet (e.g. *mar chuid de @SnaGaeilge* 'as a part of @SnaGaeilge') were incorrectly tagged as regular at-mentions (`@`).

Retweet colons One annotator marked ':' as punctuation at random stages rather than using the discourse marker tag `~`.

7.4 Experiments

7.4.1 Data

We took the finalised set of Irish POS-tagged tweets and divided them into a test set (148 tweets), development set (147 tweets) and training set (1242 tweets). Variations of this data are used in our experiments where we normalise certain tokens (described further in Section 7.4.2.)

We also automatically converted Uí Dhonnchadha (2009)'s 3198 sentence (74,705 token) gold-standard POS-tagged corpus using our mapping scheme. This text is from the New Corpus for Ireland – Irish (NCII)[8], which is a collection of text from books, newswire, government documents and websites. The text is well-structured, well-edited, and grammatical, and of course lacks Twitter-specific features like hashtags, at-mentions, and emoticons, thus differing greatly from our Twitter data. The average sentence length in this corpus is 27 tokens, diverging significantly from the average tweet length of 17.2 tokens. Despite this, and despite the fact the converted tags were not reviewed for accuracy, we were still interested in exploring the extent to which this additional training data could improve the accuracy of our best-performing model. We refer to this set as `NCII_3198`.

7.4.2 Part-of-Speech Taggers

We trained and evaluated three state-of-the-art POS-taggers with our data. All three taggers are open-source tools.

Morfette As Irish is an inflected language, inclusion of the lemma as a training feature is desirable in an effort to overcome data sparsity. Therefore we trained Morfette (Chrupała et al., 2008), a lemmatization tool that also predicts POS tags and uses the lemma as a training feature. We report on experiments both with and without optional dictionary (`Dict`) information. We used the dictionary from Scannell (2003), which contains $350,418$ surface forms, lemmas and coarse-grained POS tags. Our baseline Morfette data (`BaseMorf`) contains the token, lemma and POS-tag. The lemmas of URLs and non-syntactic hashtags have been normalised as $< URL >$ and $< \# >$, respectively.

We then evaluated the tagger with (non-syntactic) $< \# >$, $< @ >$ and $< URL >$ normalisation of both token form and lemma (`NormMorf`). Both experiments are rerun with the inclusion of our dictionary (`BaseMorf+Dict,NormMorf+Dict`).

[8] New Corpus for Ireland - Irish. See `http://corpas.focloir.ie`

ARK We also trained the CMU Twitter POS-tagger (Owoputi et al., 2013), which in addition to providing pre-trained models, allows for re-training with new languages. The current release does not allow for the inclusion of the lemma as a feature in training, however. Instead, for comparison purposes, we report on two separate experiments, one using the surface tokens as features, and the other using only the lemmas as features (`ArkForm, ArkLemma`). We also tested versions of our data with normalised at-mentions, hashtags and URLs, as above.

Stanford tagger We re-trained the Stanford tagger (Toutanova et al., 2003) with our Irish data. This tagger is not typically used for tagging tweets, but we report these experiments here for comparison with Gimpel et al. (2011)'s Stanford results with English tweets. We experimented by training models using both the surface form only (`BestStanForm`) and the lemma only (`BestStanLemma`). The best performing model was based on the feature set `left3words, suffix(4), prefix(3), wordshapes(-3,3), biwords(-1,1)`, using the `owlqn2` search option.[9]

Baseline Finally, to establish a baseline (`Baseline`), and more specifically to evaluate the importance of domain-adaptation in this context, we evaluated a slightly-enhanced version of the rule-based Irish tagger on the Twitter dataset. When the rule-based tagger produced more than one possible tag for a given token, we applied a bigram tag model to choose the most likely tag, as we did in creating the first draft of the gold-standard corpus. In addition, we automatically assigned the tag `U` to all URLs, # to all hashtags, and @ to all at-mentions.

7.4.3 Results

The results for all taggers and variations of data-setup are presented in Table 7.2. Firstly, our best performing single model (`ArkLemma#URL@`) on the test set achieves a score of 91.46%, which is 8 points above our rule-based baseline score of

[9] All other default settings were used.

Training Data	**Dev**	**Test**
Baseline		
Rule-Based Tagger	85.07	83.51
Morfette		
BaseMorf	86.77	88.67
NormMorf	87.94	88.74
BaseMorf+Dict	87.50	89.27
NormMorf+Dict	88.47	90.22
ARK		
BaseArkForm	88.39	89.92
ArkForm#@	89.36	90.94
ArkForm#URL@	89.32	91.02
BaseArkLemma#URL	90.74	91.62
ArkLemma#URL@	**91.46**	**91.89**
Stanford		
BestStanForm	82.36	84.08
BestStanLemma	87.34	88.36
Bootstrapping Best Model		
ArkLemma#URL@+NCII	**92.60**	**93.02**

Table 7.2: Results of evaluation of POS-taggers on new Irish Twitter corpus

83.51%. This confirms that tailoring training data for statistically-driven tools is a key element in processing noisy user-generated content, even in the case of minority languages. It is worth noting that the best-performing model learns from the lemma information instead of the surface form. In particular, it is the case that all results based on lemma inclusion are better than corresponding results that exclude the lemma information. This clearly demonstrates the effect that the inflectional nature of Irish has on data sparsity. The Twitter-specific tokens such as URLs, hashtags and at-mentions have been normalised which demonstrates the impact the relative uniqueness of these tokens has on the learner.

All of our results are comparable with state-of-the-art results produced by Gimpel et al. (2011) and Owoputi et al. (2013). This is interesting, given that in contrast to their work, we have not optimised our system with unsupervised word clusters due to the lack of sufficient Irish tweet data. Nor have we included a tag dictionary, distribution similarity or phonetic normalisation – also due to a lack of resources.

We carried out a closer textual comparison of Owoputi et al. (2013)'s English tweet dataset (daily547) and our new Irish tweet dataset. After running each dataset through a language-specific spell-checker, we could see that the list of highly ranked OOV (out of vocabulary) tokens in English are forms of text-speak, such as *lol* 'laugh out loud', *lmao* 'laugh my ass off' and *ur* 'your', for example. Whereas the most common OOVs in Irish are English words such as 'to', 'on', 'for', 'me', and words misspelled without diacritics. This observation shows the differences between textual challenges of processing these two languages. It may also suggest that Irish Twitter text may follow a more standard orthography than English Twitter text, and will make for an interesting future cross-lingual study of Twitter data.

Finally, we explored the possibility of leveraging from existing POS-tagged data by adding NCII_3198 to our best performing model ArkLemma#URL@. We also duplicated the tweet training set to bring the weighting for both domains into balance. This brings our training set size to 5682 (117, 273 tokens). However, we find that a significant increase in the training set size only results in just over a 1 point increase in POS-tagging accuracy. At a glance, we can see some obvious errors the combined model makes. For example, there is confusion when tagging the word *an*. This word functions as both a determiner and an interrogative verb particle. The lack of direct questions in the NCII corpus results in a bias towards the D (determiner) tag. In addition, many internal capitalised words (e.g. the beginning of a second part of a tweet) are mislabelled as proper nouns. This is a result of the differing structure of the two data sets – each tweet may contain one or more phrases or sentences, while the NCII is split into single sentences.

7.5 Summary and Conclusion

We have expanded our NLP research of the Irish language to include processing of Irish social media text. In this chapter, we have reported how we developed the first dataset of gold-standard POS-tagged Irish language tweets and produced training models for a selection of POS-taggers. We show how, with these resources, we have been able to carry out some preliminary linguistic analysis of the language variation used in Twitter Irish.

We have also shown how we have leveraged existing work to build these resources for a low-resourced language, to achieve state-of-the-art results. In addition, we confirm through empirical methods that the NLP challenges arising from noisy user-generated text can also apply to a minority language.

Our data and models are available to download from `https://github.com/tlynn747/IrishTwitterPOS`. The annotation guide is presented in Appendix B.

Chapter 8

Conclusion

In this chapter, we summarise the contributions our work has made to NLP research for the Irish language, namely the development of the first syntactic treebank, the first statistical parsing models, a POS-tagged corpus of Irish tweets and the first statistical POS tagging models for Irish Twitter text. We also revisit the research questions we proposed in Chapter 1 and provide answers to them. Finally, we discuss the various possibilities for further research that have arisen through our work.

8.1 Summary and Contributions

In this digital age, there is a demand for language processing tools that will make digital content and text-based technology available to linguistic groups in their own language. If not available, there is a real risk that these users will opt to use another language that facilitates these technical options instead. Minority and low-resourced languages are at risk in the context of this increased shift towards online or computer-based language use. In particular, the Irish language is at risk, due to the fact that all Irish speakers in Ireland are fluent English speakers. As English is an easy second option for technology users, this can result in less engagement with Irish, particularly among young people. The current status of Irish in NLP, as discussed in Chapter 1, and the language's minority status, as discussed in Chapter 2, reflect the need for more resources if we want to computationally process and interpret Irish language text.

Our research has taken a step towards addressing this need by providing basic text resources and processing tools upon which further NLP research will be possible. We summarise our contribution here:

- The dependency treebank we have developed for Irish (IDT), and discussed in Chapter 3, is a text resource that will not only provide a basis for a linguistic analysis of Irish, but is also a digitally readable corpus from which linguistic information can be automatically extracted and serve as input to a range of NLP applications.[1]

- Two additional elements of this treebank are the Dependency Labelling Scheme outlined in Chapter 4 and the Annotation Guidelines for the Irish Dependency Treebank in Appendix A, which together provide a detailed linguistic analysis and description of Irish. This is a valuable addition to the limited collection of Irish syntax reference resources we highlighted in Chapter 2. They are both necessary resources for continued development of the treebank by other annotators.

- In Chapter 5 we report on the mapping of the IDT to two different universal dependency (UD) annotation schemes. The Irish UD treebank based on the 2013 scheme allowed us to carry out cross-lingual parsing experiments (also Chapter 6) in order to examine how treebanks from other languages could be leveraged for our own work. The Irish UD treebank based on the 2015 scheme has seen Irish become part of The Universal Dependencies project which seeks to develop an annotation scheme that is cross-linguistically uniform and will aid further research in multi-lingual parser development. Our involvement in this project has helped tie the Irish language closely to the international effort for developing resources, and these links should prove beneficial in the future.

[1] This treebank is available to download under an open source licence from: `https://github.com/tlynn747/IrishDependencyTreebank`

- The statistical parsing models we have trained with the IDT achieve an accuracy of UAS 78.54% and LAS 71.59%, as per our reports in Chapter 6. In comparison to better-resourced languages, the scores reflect the need for additional treebank data on which we can train the systems. Yet based trends shown in Table 6.12, it is not possible to predict how many additional trees would result in state-of-the art parsing accuracy.

- In terms of building a treebank for a low-resourced language, we have shown how Active Learning could be used to assist this development and help to overcome resource limitations.

- Despite the acknowledged contribution of Active Learning to bootstrapping the treebank's development, we have also shown (and in partial answer to the question raised by Hal Daumé),[2] through various parsing experiments that did not yield impressive results, the value of human annotation in treebank development, and how this cannot be easily replaced by technological means.

- The 1537-tweet POS-tagged corpus we provide will be a useful resource for both linguistic and socio-linguistic research on Irish language. This contribution highlights the evolving use of the Irish language online and how the style variation of Irish in social media differs considerably from standard well-structured and grammatical Irish text. Our corpus may also be useful to those studying code-switching and computational approaches to dealing with code-switching. We provide an annotation guide for POS tagging of Irish tweets in Appendix B.

- In the context of building statistical tagger models for Irish Twitter data, we have shown how this variation of Irish cannot be easily processed using standard NLP tools. However, we have also demonstrated how existing Irish language resources can be leveraged to produce domain-adapted tools that are

[2] Refer to Section 6.4.5 and to the following blog post: `http://nlpers.blogspot.ie/2011/10/active-learning-far-from-solved.html`

tailored to user-generated content such as tweets.

8.2 Research Questions Answered

- *What is an appropriate linguistic analysis of Irish for a dependency treebank, drawing on and synthesising traditional descriptive analyses and theoretical work?*

  We have designed a broad-coverage annotation scheme for the Irish language, which is based on LFG-inspired dependencies and extended to contain analyses and dependency labels that are specific to the Irish language. Based on the broadly representative sentences contained in the treebank, we have shown that this type of dependency analysis sufficiently addresses the main syntactic structures that occur within the language, despite some unresolved syntactic theoretical issues that are present in the literature.

- *Can an approach such as Active Learning, that has been suggested to be applicable to bootstrapping the development of treebanks, prove to be useful when deployed in the actual construction of a treebank?*

  If the purpose of a treebank's development is to serve as training data for a parser, then Active Learning is a worthwhile approach. We show how the Query By Committee approach to Active Learning (AL) can ensure that the trees added to the treebank for annotation will add to the linguistic richness of the data at an early stage. In other words, it is possible to ensure that a wide range of linguistic constructions are present in the data that sufficiently represent the various nuances of a language even when the treebank is relatively small in size. However, the purpose of the AL approach is to identify sentences that prove informative or difficult to parse automatically. Through calculations of correction effort, we show the extra degree of correction required by a human annotator in the AL setup, when compared to the Passive Learning (PL) setup.

The annotator also noted that the AL corrections took longer and required more thought than the PL corrections. Based on the small size of seed data we worked with in this study, it should be noted that the increase in the parsing accuracy scores resulting from the new data did not justify the extra annotation effort involved in this approach. In the context of Irish, considering that there is a limit to the number of gold-standard POS-tagged sentences available to us, all of these sentences are likely to ultimately end up in the treebank. The AL approach would only serve to set the order in which all of these trees would eventually be added.

- *Given the existence of proposed techniques for development of parsers for low-resource language or improving performance of such a parser – the use of unlabelled data and cross-lingual transfer parsing – can these help when combined with a small gold-standard treebank used for training?*

  Some of our semi-supervised built parsing models using unlabelled data achieved a statistically significant increase in scores over the baseline. Yet these improvements were still fairly modest. At this stage, and because the treebank is a work-in-progress, it is not yet possible to establish how the increase we did achieve would compare to adding the same trees in gold-standard form.

  In our cross-lingual transfer parsing study, none of the 10 languages involved revealed to be a language that would prove suitable for bootstrapping the Irish parser. While the use of Indonesian data as training data for parsing Irish showed an increase of accuracy over all other languages in the set, the results did not outperform our Irish baseline scores. However, this is not to say that cross-lingual bootstrapping is not a method for Celtic languages in general, as our analysis of the Indonesian parsing results revealed that similar linguistic characteristics across source and target languages are valuable for cross-lingual transfer parsing. In this context, we would expect that the Irish Dependency Treebank data would be a useful additional training data set for other Celtic languages.

- *In what way can we leverage existing Irish NLP tools for processing Irish tweets?*

  We drew on previous studies for other languages to explore the adaption of current resources to process user-generated text. While we did not adapt the rule-based Irish POS tagger, we showed that it was easy to leverage it in our creation of a gold-standard POS-tagged corpus of Irish tweets. The rule-based tagger, while achieving an accuracy of just 83.51% on the final data set, provided us with a pre-tagged corpus which only required manual correction, rather than manual annotation from scratch. This allowed us to create a corpus of 1537 tweets with just two annotators in a short amount of time.

8.3 Proposed Future Work

8.3.1 Further Treebank Development

In comparison to treebanks of better-resourced languages, the Irish treebank of 1018 trees is relatively small. This of course is due to lack of significant resources. Financial support and human annotators can prove difficult to source for low-resourced languages. As we have seen, human annotation cannot be easily replaced by automated methods. Financial support for better-resourced languages is often driven by the size of demand (e.g. the number of speakers/application users) and the opportunity for financial gain (e.g. benefits of machine translation global industries). Finding speakers of a minority language who also have a suitable skill set for NLP research is also a challenge.

Therefore, if future opportunities for treebank expansion arises, it is important that more bootstrapping methods such as the following are explored. Recent work by Mirroshandel and Nasr (2011) demonstrates how relevant substrings within identified problematic sentences can be isolated for correction. This is similar to selective-sampling methods we use in Chapter 6, only it is at the substring level instead of the sentence level. The theory behind this approach is that a parser will often successfully parse easy structures, such as determiner-noun attachments, for example, and only require correction or human input on more difficult substrings, such as multiple coordination. Other work focuses on developing methods for automatically detecting errors in dependency parses. For example, Dickinson (2010) and Dickinson and Smith (2011) extract grammar rules from a gold-annotated corpus and compare these to rules extracted from a corpus of predicted annotations. If the rules from the predicted annotations do not fit well with the gold grammar, they are flagged. In fact, the latter expanded approach (which also looks at the value of a small gold grammar) is noted as being particularly beneficial to low-density languages.

If there is an option to expand on our Active Learning experiments, our interesting dip in results in our final experiment (due to sentence length) suggests that it may be worth considering setting an upper limit on sentence length before presenting new trees for correction or perhaps, more generally, adding length constraints to the Active Learning process.

Bootstrapping lesser-resourced languages through the use of parallel texts is also possible. This involves exploiting tools of the more highly-resourced language of the language pair (e.g. Hwa et al. (2005); Wróblewska and Frank (2009)). This approach may be a possibility for our treebank development as there is a large number of English-Irish parallel official documents available from both Irish and European Parliament proceedings. However, given the difficulties that the length of some legal text poses for Irish parsing, the use of a parallel corpus collected by Scannell (2005) could prove a more valuable resource if we were to consider this approach.

8.3.2 Exploiting Morphological Features

Previous studies on parsing morphologically rich languages have shown that the inclusion of morphological features can improve parsing accuracy (Bohnet et al., 2013). The Irish Dependency Treebank does not currently contain morphological features. In Section 6.6, our preliminary experiments using morphological features did not yield interesting results. However, the morphological features included were minimal and our MaltParser feature models were manually optimised.

One possible enhancement to the treebank in the future involves retrieving morphological information from the 3,000 sentence gold-standard corpus (Uí Dhonnchadha, 2009) on which we based our treebank.[3] In future experiments, it would be interesting to experiment with augmenting parsing models with MaltOptimizer (Ballesteros, 2012), an open-source tool that facilitates MaltParser optimisation.

8.3.3 Exploiting the Hierarchical Dependency Scheme

In terms of further parsing experiments with the treebanks, an additional avenue of research would be to exploit the hierarchical nature of the dependency scheme as outlined in Chapter 4. It would be interesting to see if a more coarse-grained version of the tagset would lend itself to higher parsing accuracy. It would also be interesting to see if this could help us to arrive at more flexible way of measuring agreement or disagreement in sample selection, such as the experiments we report on in Section 6.5.

8.3.4 Semi-supervised Parsing

Further to the semi-supervised experiments we describe in Section 6.5, a lack of gold annotated data, on which we can compare our results to, makes it difficult to fully assess the significance of this type of bootstrapping. This type of analysis would be interesting at a later date when the unlabelled trees used in these experiments

[3] From a data management perspective, omitting this morphological information from the first version of the treebank made the development task more manageable for a single developer.

are eventually annotated and corrected manually, serving as a comparison gold data set.

While the semi-supervised experiments did not produce promising results from the perspective of parser accuracy improvement, there are many directions this parsing research could take us in the future. Our gold POS-tagged unlabelled data set contained 1938 trees annotated with gold POS tags. While the trends over the stages of the experiments do not suggest that an increase in this unlabelled set size would make much difference to the parsing scores, it may be worth considering taking advantage of the fully unlabelled, untagged data in the New Corpus for Ireland – Irish, which consists of 30 million words. We would also like to experiment with a fully unsupervised parser using this dataset.

8.3.5 Cross Lingual Studies

In our studies in cross-lingual studies, we discovered that other languages that are linguistically and typologically different to Irish did not prove useful from a bootstrapping perspective. However, Irish, as a Celtic language, is unique with regards to the groupings of the other languages involved in the study. We believe that a cross-lingual approach such as this would be more beneficial if applied to languages in the same language family as Irish. In comparison to other Celtic languages such as Scottish Gaelic for example, Irish is in fact better resourced. It would be interesting to see how the Irish Dependency Treebank could assist with future treebank or parser development for Scottish Gaelic. In light of the recent development of a statistical Scottish Gaelic ↔ Irish machine translation system[4] (Scannell, 2014), a Scottish Gaelic POS tagger (Lamb and Danso, 2014) and preliminary work on a Scottish Gaelic dependency treebank (Batchelor, 2014), this type of bootstrapping looks very promising.

[4] `http://www.intergaelic.com/gd-ga/` (accessed June 2015)

8.3.6 NLP for Social Media

Limited resources and time prevented exploration of some options for improving our POS-tagging results. One of these options is to modify the CMU (English) Twitter POS-tagger to allow for inclusion of lemma information as a feature. Another option, when there is more unlabelled data available (i.e. more Irish tweets online), would be to include Irish word cluster features in the training model.

We expect that this new data resource (the POS-tagged Twitter corpus) will provide a solid basis for linguistic and sociolinguistic study of Irish on a social media platform. This new domain of Irish language use can be analysed in an empirical and scientific manner through corpus analysis by means of our data. This type of research could contribute to future strategy planning for the language.

From a tool-development perspective, we expect our Twitter corpus and the derived POS-tagging models could be used in a domain-adaptation approach to parsing Irish tweets, similar to the work of Kong et al. (2014). This would involve adapting our Irish statistical dependency parser for use with social media text. Our corpus could provide the basis of a treebank for this work. It is expected that a treebank of tweets could be easier to develop than the IDT, as each tweet is limited to 140 characters. This means that there are likely to be fewer sub-clauses and long-distance dependencies to analyse. It should also be considered, however, that the ungrammatical structures present in some tweets may equally introduce additional parsing difficulties and would lead to an interesting study.

Following our discovery of the extent that code-switching is present our Irish Twitter data, we feel future studies on this phenomenon would be of interest to research groups working on computational approaches to code-switching (e.g Solorio et al. (2014)). In order to do that, we suggest updating the corpus with a separate tag for English tokens (that is, a tag other than `G`, which is also used for abbreviations, items and unknowns) before carrying out further experiments in this area.

Finally, as part of speech tags have proven to be useful for improving machine translation (Hoang, 2007), it would be interesting to see how our work can assist in the automated translation of Irish tweets.

8.3.7 Universal Dependencies Project

The universal dependencies project (UD15) we refer to in Chapter 5 is an ongoing project. In parallel to the IDT development, we also expect to expand the universal version of our data. We hope to increase its size and also to include the morphological features we discuss in Section 8.3.2. We also intend to review some of the UD labels we did not employ in the first release (e.g. `remnant` – remnant in ellipsis, or `expl` – expletives) and fully review the Irish data to see if they are in fact applicable.

8.4 Concluding Remarks

The major contribution to this thesis is the development of the first syntactic treebank for Irish. We believe that, with this resource, we have put in place a foundation for future treebank development and parser development. The treebank will also provide a solid linguistic reference corpus for the Irish linguistic research community. We have also shown, through our various attempts at bootstrapping the treebank's development, the importance of human annotation in this type of task and how it is not easy to overcome the lack of these costly resources through the use of automated methods.

We have also reported on the inclusion of Irish as a language in the Universal Dependencies project. This project is an important milestone in multilingual parsing, and from the Irish NLP research community perspective, it is significant that Irish, a minority language, is playing a part in this.

We hope that our work will also benefit the wider community, by providing a basis upon which tools that will assist Irish-speaking groups in their daily lives. Irish schools are in much need of CALL (Computer Assisted Language Learning) systems. For example, from a language learning perspective, the treebank provides syntactic information in machine-readable and processable format that can assist with tasks such as error detection and grammar checking. POS tagging and morphological analysis has already proven useful in some preliminary work in this area (Keogh et al., 2004; Ward, 2014).

There has been a recent uptake of using parser output for improving information retrieval tasks (e.g. Gillenwater et al. (2013)). Currently, there are no information retrieval systems available that are tuned to searching for Irish content documents. We hope that the treebank or parser can contribute to research in this area to some extent.

In addition, we hope that the treebank data will be used to improve current developments in Irish↔English Statistical Machine Translation (SMT). The author has been instrumental in securing funding to develop an SMT system for an Irish government department (DAHG), to assist with their in-house English↔Irish translation demands. The project has just moved from a pilot phase to a stage where a hybrid approach, involving word-reordering, is being explored. The divergent word order between the two languages can affect translation quality, and previous work involving word-reordering on divergent languages (e.g. Xu et al. (2009)) has proven successful in this respect.

Also, from a language shift perspective, our work on Irish tweets presents a new perspective for both Irish NLP and sociolinguistic research. We see this as an important step towards recognising an evolving language that now has a wider use in terms of its increased online presence, and is being influenced by a new generation of Irish speakers. We hope that our preliminary work in this area will open up future efforts to cultivating the use of Irish in social media.

Finally, Figure 8.1 and Figure 8.2 give an overview of the type of future applications that are possible following the development of the Irish NLP resources in this thesis.

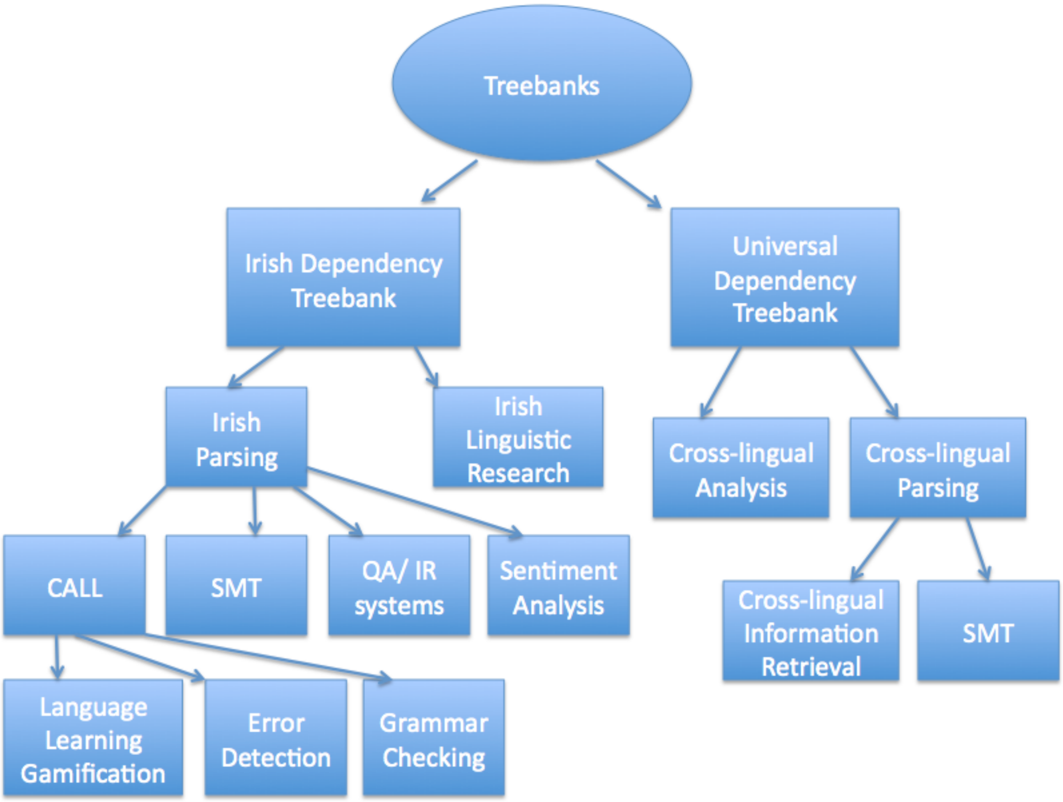

Treebanks
Irish Dependency Treebank
Universal Dependency Treebank
Irish Parsing
Irish Linguistic Research
Cross-lingual Analysis
Cross-lingual Parsing
CALL
SMT
QA/ IR systems
Sentiment Analysis
Cross-lingual Information Retrieval
SMT
Language Learning Gamification
Error Detection
Grammar Checking

Figure 8.1: Applications of the treebanks developed in this thesis.

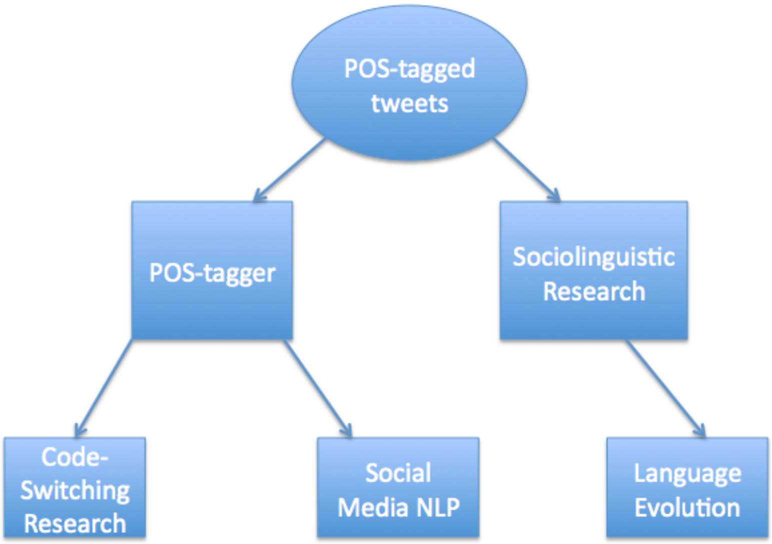


Figure 8.2: Applications of the Twitter POS-tagged corpus developed in this thesis.

Bibliography

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a Treebank for French. In *Treebanks : Building and Using Parsed Corpora*, pages 165–188. Springer.

Agić, Ž., Aranzabe, M. J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A. T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., Lynn, T., Manning, C., Martínez, H. A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.1.

Ahlqvist, A. (1972). Some aspects of the copula in Irish. *Éigse 14/4*, pages 269–274.

Amigó, E., Giménez, J., Gonzalo, J., and Verdejo, F. (2009). The contribution of linguistic features to automatic machine translation evaluation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 306–314, Stroudsburg, PA, USA.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asudeh, A. (2002). The syntax of preverbal particles and adjunction in Irish. In *Proceedings of the 2002 Lexical Functional Grammar Conference*, pages 1–18, Athens, Greece.

Atserias, J., Attardi, G., Simi, M., and Zaragoza, H. (2010). Active learning for building a corpus of questions for parsing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Attia, M. (2008). A Unified Analysis of Copula Constructions in LFG. In *Proceedings of the Lexical Functional Grammar '08 Conference*, pages 89–108, Sydney, Australia.

Avontuur, T., Balemans, I., Elshof, L., van Noord, N., and van Zaanen, M. (2012). Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal*, 2:34–51.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia.

Baldridge, J. and Osborne, M. (2004). Active Learning and the Total Cost of Annotation . In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain.

Baldridge, J. and Palmer, A. (2009). How well does Active Learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore.

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., and Oepen, S. (2004). Road testing the English Resource Grammar over the British National Corpus. In

Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 2047–2050, Lisbon, Portugal.

Ballesteros, M. (2012). Maltoptimizer: A System for MaltParser Optimization. In *Proceedings of the Eighth International Conference on Linguistic Resources and Evaluation (LREC)*, pages 2757–2763, Istanbul, Turkey.

Batchelor, C. (2014). gdbank: The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 60–65, Dublin, Ireland.

Berman, R. A. (1985). The Acquisition of Hebrew. *The Crosslinguistic Study of Language Acquisition*, 1.

Bhaskar, P. and Bandyopadhyay, S. (2010). A query focused multi document automatic summarization. In *The 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24)*, pages 545–554, Sendai, Japan.

Bhat, R. A. and Sharma, D. M. (2012). A dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop*, LAW VI '12, pages 157–165, Stroudsburg, PA, USA.

Bobaljik, J. D. and Carnie, A. (1996). A miniminalist approach to some problems of Irish word order. In Borsley, B. and Roberts, I., editors, *The Syntax of the Celtic Languages*, pages 223–240. Cambridge University Press.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The Prague Dependency Treebank: Three-level annotation scenario. In Abeillé, A., editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 89–97, Beijing, China.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajic, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics (TACL)*, 1:415–428.

Bouma, G., Noord, G. V., Malouf, R., and Groningen, R. (2000). Alpino: Wide-coverage computational analysis of Dutch. *Computational Linguistics in the Netherlands Journal*, pages 45–59.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol, Bulgaria.

Bresnan, J. (2001). *Lexical Functional Syntax.* Oxford: Blackwell.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 356–363, Washington DC, USA.

Bruce, R. and Wiebe, J. (1998). Word sense distinguishability and inter-coder agreement. In *Proceedings of 3rd Empirical Methods in Natural Language Processing (EMNLP-98)*, pages 53–60, Granada, Spain.

Buchhloz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, pages 149–164, New York City, USA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan.

Cahill, A., Burke, M., O'Donovan, R., Riezler, S., van Genabith, J., and Way, A. (2008). Wide-coverage deep statistical parsing using automatic dependency structure annotation. *Computational Linguistics, Volume 34, Number 1, March 2008.*

Cahill, A., Burke, M., O'Donovan, R., Van Genabith, J., and Way, A. (2004). Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 319–326, Barcelona, Spain.

Cai, J., Utiyama, M., Sumita, E., and Zhang, Y. (2014). Dependency-based pre-ordering for Chinese-English machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 155–160, Baltimore, Maryland.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carmel, D., Mejer, A., Pinter, Y., and Szpektor, I. (2014). Improving term weighting for community question answering search using syntactic analysis. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 351–360, New York, USA.

Carnie, A. (1997). Two types of non-verbal predication in Modern Irish. *Canadian Journal of Linguistics*, 42/1-2:57–73.

Carnie, A. (2005). Flat Structure, Phrasal Variability and VSO. *Journal of Celtic Linguistics.*

Carnie, A. (2007). *Syntax: A Generative Introduction.* Blackwell Publishing, Blackwell Publishing, Malden, MA, USA, second edition.

Cassel, S. (2009). Maltparser and liblinear – transition-based dependency parsing with linear classification for feature model optimization. Master's Thesis, Uppsala University.

Çetinoğlu, Ö., Foster, J., Nivre, J., Hogan, D., Cahill, A., and van Genabith, J. (2010). LFG without C-structures. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories*, Tartu, Estonia.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th national conference on artificial intelligence and 9th conference on Innovative applications of artificial intelligence*, pages 598–603, Providence, Rhode Island.

Charniak, E. (2000). A maximum entropy inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, pages 132–139, Seattle, Washington, USA.

Charniak, E. and Johnson, M. (2005). Course-to-fine n-best-parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180, Ann Arbor, Michigan.

Chen, J., Schein, A., Ungar, L., and Palmer, M. (2006). An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 120–127, New York City, USA.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA.

Chomsky, N. (1980). *Rules and Representations.* Columbia Univeristy Press, New York.

Chomsky, N. (1986). *Knowledge of Language. Its Nature, Origin, and Use.* Convergence. Praeger, New York/Westport/London.

Chouinard, M. M. and Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(03):637–669.

Christian-Brothers (1960). *Graiméar Gaeilge na mBráithre Críostaí.* Dublin: M.H. Mac an Ghoill agus a Mhac, Tta.

Christian-Brothers (1962). *New Irish Grammar.* Dublin: C J Fallon.

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

Chung, S. and McCloskey, J. (1987). Government, barriers and small clauses in Modern Irish. In *Linguistic Inquiry*, volume 18, pages 173–237.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

Collins, M., Hajic, J., Ramshaw, L., and Tillmann, C. (1999). A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, Maryland, USA.

Covington, M. A. (1990). A Dependency Parser for Variable-Word-Order Languages. Technical Report AI-1990-01, University of Georgia, Athens, GA.

Covington, M. A. (1994). Discontinuous dependency parsing of free and fixed word order: Work in progress. Technical Report AI-1994-02, University of Georgia, Athens, GA.

Craig, J. (1897). The Irish verbal noun and present particple. *Gaelic Journal.*

Dabrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, 6(852).

Dalrymple, M. (2001). *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics.* Academic Press, San Diego, London, Boston.

Dalrymple, M., Dyvik, H., and King, T. H. (2004). Copular complements: Closed or open? In *The Proceedings of the LFG '04 Conference*, pages 188–198, Christchurch, New Zealand.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and D. Manning, C. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4585–4592, Reykjavik, Iceland.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure trees. In *Proccedings of the 5th international conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454, Genoa, Italy.

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Workshop on Crossframework and Cross-domain Parser Evaluation (COLING2008)*, Manchester, U.K.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Recent Advances in Natural Language Processing*, pages 198–206, Hissar, Bulgaria.

Di Eugenio, B. and Glass, M. (2004). The Kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Dickinson, M. (2010). Detecting errors in automatically-parsed dependency relations. In *Proceedings of the 48th Annual Meeting on Association for Computational Linguistics (ACL2010)*, pages 729–738, Uppsala, Sweden.

Dickinson, M. and Smith, A. (2011). Detecting dependency parse errors with minimal resources. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*, Dublin, Ireland.

Doherty, C. (1992). Clause structure and the Modern Irish copula. *Syntax at Santa Cruz*, 1.

Doherty, C. (1997). *The Pronominal Augment in Irish Identificational Sentences.* Dublin: Institiúid Teangeolaíochta Éireann.

Doyle, A. (2015). *A History of the Irish Language: From the Norman Invasion to Independence*. Oxford University Press.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., and Žele, A. (2006). Towards a Slovene dependency treebank. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC2006)*, pages 1388–1391, Genoa, Italy.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia.

Eisner, J. M. (1996a). An empirical comparison of probability models for dependency grammar. Technical Report IRCS-96-11, Institute for Research in Cognitive Science, University of Pennysylvania.

Eisner, J. M. (1996b). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 340–345, Copenhagen, Denmark.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences*, 32(5):429–48; discussion 448–494.

Evas, J. (2013). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age.* META-NET White Paper Series: Europe's Languages in the Digital Age. Springer.

Fisher, S. and Roark, B. (2007). The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 488–495, Prague, Czech Republic.

Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., Van Genabith, J., et al. (2011). #hardtoparse: POS Tagging and Parsing the Twitterverse. In *Proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25, San Francisco, CA.

Ghayoomi, M. and Kuhn, J. (2013). Sampling methods in Active Learning for treebanking. In *Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT'12)*, pages 49–60, Heidelberg, Germany.

Gillenwater, J., He, X., Gao, J., and Deng, L. (2013). End-to-End Learning of Parsing Models for Information Retrieval. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA.

Giollagáin, C. Ó., Donnacha, S. M., Chualáin, F. N., Shéaghdha, A. N., and O'Brien, M. (2007). Comprehensive linguistic study of the use of Irish in the Gaeltacht. Technical report, National University of Ireland.

Greene, D. (1966). *The Irish Language.* Dublin: Cultural Relations Committee of Ireland.

Gupta, M., Yadav, V., Husain, S., and Sharma, D. M. (2010). Partial Parsing as a Method to Expedite Dependency Annotation of a Hindi Treebank. In *Proceed-*

ings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), pages 1930–1935, Valletta, Malta.

Hahn, U., Beisswanger, E., Buyko, E., Faessler, E., Traumüller, J., Schröder, S., and Hornbostel, K. (2012). Iterative refinement and quality checking of annotation guidelines - how to deal effectively with semantically sloppy named entity types, such as pathological phenomena. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 3881–3885.

Hajič, J. and Zemánek, P. (2004). Prague Arabic Dependency Treebank: Development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117, Cairo, Egypt.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In Hajičová, E., editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

Haverinen, K., Viljanen, T., Laippala, V., Kohonen, S., Ginter, F., and Salakoski, T. (2010). Treebanking Finnish. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT-9)*, pages 79–90, Tartu, Estonia.

Hays, D. (1964). Dependency theory: A formalism and some observations. *Language*, 40:511–525.

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., Flor, M., Madnani, N., Tetreault, J. R., Blanchard, D., Napolitano, D., Lee, C. M., and Blackmore, J. (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *arXiv preprint*, arXiv:1403.0801.

Hoang, H. (2007). Factored translation models. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL*, pages 868–876.

Hockenmaier, J. and Steedman, M. (2007). CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Hoi, S. C. H., Jin, R., and Lyu, M. (2006). Large-scale text categorization by batch mode Active Learning. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)*, pages 633–642, Edinburgh, UK.

Huang, Z. and Harper, M. (2009). Self-training PCFG grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 832–841, Suntec, Singapore.

Hughes, B., Haggerty, J., Nothman, J., Manickam, S., and Curran, R. J. (2005). A distributed architecture for interactive parse annotation. In *Proceedings of the Australasian Language Technology Workshop 2005*, volume 3, pages 207–214, Sydney, Australia.

Hwa, R., Osborne, M., Sarkar, A., and Steedman, M. (2003). Corrected co-training for statistical parsers. In *Proceedings of the Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington DC, US.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*.

ITÉ (2002). PAROLE Morphosyntactic Tagset for Irish. Institiúid Teangeolaíochta Éireann.

Johansson, R. and Moschitti, A. (2010). Syntactic and semantic structure for opinion expression detection. In *Proceedings of the 14th Conference on Computational Natural Language Learning*, pages 67–76, Uppsala, Sweden.

Judge, J., Cahill, A., and van Genabith, J. (2006). Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 497—504, Sydney, Australia.

Judge, J., Chasaide, A. N., Dhubhda, R. N., Scannell, K. P., and Uí Dhonnchadha, E. (2012). *The Irish Language in the Digital Age.* META-NET White Paper Series: Europe's Languages in the Digital Age. Springer.

Jurafsky, D. and Manning, C. D. (2012). Natural Language Processing: Stanford online course.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Kaljahi, R., Foster, J., Rubino, R., Roturier, J., and Hollowood, F. (2013). Parser accuracy in quality estimation of machine translation: A tree kernel approach. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1092–1096, Nagoya, Japan.

Kaplan, R. and Bresnan, J. (1982). Lexical Functional Grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge MA.

Karlsson, F. (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, volume 4. Mouton de Gruyter.

Keogh, K., Koller, T., Uí Dhonnchadha, E., van Genabith, J., and Ward, M. (2004). CL for call in the primary school. In *eLearning for Computational Linguistics and*

Computational Linguistics for eLearning. International Workshop in Association with COLING 2004., pages 79–85.

Kikuchi, Y., Hirao, T., Takamura, H., Okumura, M., and Nagata, M. (2014). Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland.

King, T. H., Crouch, R., Riezler, S., Dalrymple, M., and Kaplan, R. M. (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, pages 1–8, Budapest, Hungary.

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.

Kromann, M. (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*, pages 217–220, Växjö, Sweden.

Kubler, S., McDonald, R., Nivre, J., and Hirst, G. (2009). *Dependency Parsing.* Morgan and Claypool Publishers.

Lamb, W. and Danso, S. (2014). Developing an automatic part-of-speech tagger for scottish gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5, Dublin, Ireland.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. In *Biometrics*, volume 33, pages 159–174. International Biometric Society.

Le Roux, J., Foster, J., Wagner, J., Kaljahi, R. S. Z., and Bryl, A. (2012). DCU-Paris13 systems for the SANCL 2012 shared task. In *Working Notes of First*

Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), pages 1–4, Montreal, Canada.

Lewis, D. and Gale, W. (1994). A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACL-SIGIR Conference on Research and Development of Information Retrieval*, pages 3–12, Dublin, Ireland.

Lin, D. (2003). Dependency-based evaluation of Minipar. In Abeillé, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 317–329. Springer Netherlands.

Lynn, T., Çetinoğlu, Ö., Foster, J., Uí Dhonnchadha, E., Dras, M., and van Genabith, J. (2012a). Irish treebanking and parsing: A preliminary evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1939–1946, Istanbul, Turkey.

Lynn, T., Foster, J., and Dras, M. (2013). Working with a small dataset – semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA.

Lynn, T., Foster, J., Dras, M., and Tounsi, L. (2014). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, Dublin, Ireland.

Lynn, T., Foster, J., Dras, M., and Uí Dhonnchadha, E. (2012b). Active learning and the Irish treebank. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 23–32, Dunedin, New Zealand.

Lynn, T., Scannell, K., and Maguire, E. (2015). Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the 1st Workshop on Noisy User-generated Text (W-NUT 2015)*, Beijing, China.

Mac Congáil, N. (2002). *Leabhair Gramadaí Gaeilge.* Cló Iar-Chonnacht.

Mac Giolla Chríost, D. (2013). *The Irish language in Ireland : from Goídel to globalisation.* Number 3 in Routledge Studies in Linguistics. Routledge.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology (HLT'94)*, pages 114–119, Plainsboro, NJ, USA.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

McCloskey, J. (1979). *Transformational syntax and model theoretic semantics: a case in Modern Irish.* Dordrecht: Reidel.

McCloskey, J. (1983). A VP in a VSO language? In Gazdar, G., Klein, E., and Pullum, G. K., editors, *Order, Concord and Constituency.* Foris Publications, Dordrecht.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA.

McClosky, D., Charniak, E., and Johnson, M. (2008). When is self-training effective for parsing? In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 152–159, Manchester, UK.

McDonald, R., Nivre, J., Quirmbach-brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu, N., and Lee, C. J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL '13*, pages 92–97, Sofia, Bulgaria.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language*

Technology Conference on Empirical Methods in Natural Language Processing, pages 523–530, Vancouver, British Columbia, Canada.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice.* State University of New York Press.

Mieskes, M. and Strube, M. (2006). Part-of-speech tagging of transcribed speech. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 935–938, Genoa, Italy.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Princeton, New Jersey.

Mirroshandel, S. A. and Nasr, A. (2011). Active Learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*, pages 140–149, Dublin, Ireland.

Ngai, G. and Yarowsky, D. (2000). Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. In *Proceedings of 38th Annual Meeting on Association for Computational Linguistics (ACL '00)*, pages 117–125, Hong Kong.

Nivre, J. (2003). An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, pages 149–160, Nancy, France.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Process-*

ing, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal Dependencies 1.0.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

Nivre, J. and McDonald, R. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131, Prague, Czech Republic.

Nivre, J. and Nilsson, J. (2005). Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 99–106, Stroudsburg, PA, USA.

Nivre, J. and Scholz, M. (2004). Deterministic dependency parsing of english text. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA.

Ó Siadhail, M. (1989). *Modern Irish: Grammatical structure and dialectal variation.* Cambridge: Cambridge University Press.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, MA, USA.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2002). LinGO Redwoods - a rich and dynamic treebank for HPSG. In *Beyond PARSEVAL Workshop at the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 575–596, Las Palmas, Spain.

Oflazer, K., Say, B., Hakkani-Tür, D. Z., and Tür, G. (2003). Building a Turkish treebank. In Abeille, A., editor, *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. Technical report, Swedish Institute of Computer Science.

Osborne, M. and Baldridge, J. (2004). Ensemble-based Active Learning for Parse Selection. In *HLT-NAACL 2004: Main Proceedings*, pages 89–96, Boston, USA.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia.

Passonneau, R., Habash, N., and Rambow, O. (2006). Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC2006)*, pages 1951–1956, Genoa, Italy.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096.

Petrov, S. and McDonald, R. (2012). Overview of the 2012 Shared Task on Parsing the Web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, Montreal Canada.

Poesio, M. (2004). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain.

Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., and Piperidis, S. (2005). Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 149–160, Barcelona, Spain.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, Michigan.

Ragheb, M. and Dickinson, M. (2013). Inter-annotator agreement for dependency annotation of learner language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Atlanta, Georgia.

Ravi, S., Knight, K., and Soricut, R. (2008). Automatic prediction of parser accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Honolulu, Hawaii.

Rehbein, I. (2011). Data point selection for self-training. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2011)*, Dublin, Ireland.

Rehbein, I. (2013). Fine-grained pos tagging of German tweets. In Gurevych, I., Biemann, C., and Zesch, T., editors, *GSCL*, volume 8105 of *Lecture Notes in Computer Science*, pages 162–175. Springer.

Reichart, R. and Rappaport, A. (2007). Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623, Prague, Czech Republic.

Reichart, R. and Rappoport, A. (2007). An Ensemble Method for Selection of High Quality Parses. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 408–415, Prague, Czech Republic.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA.

Sagae, K. (2010). Self-training without Reranking for Parser Domain Adapation and its Impact on Semantic Role Labeling. In *Proceedings of the ACL Workshop on Domain Adaptation for NLP*, pages 37–44, Uppsala, Sweden.

Sagae, K. and Tsujii, J. (2007). Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 1044–1050, Prague, Czech Republic.

Sampson, G. (1993). The SUSANNE Corpus.

Scannell, K. (2005). Applications of parallel corpora to the development of monolingual language technologies.

Scannell, K. (2014). Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland.

Scannell, K. P. (2003). Automatic thesaurus generation for minority languages: an Irish example. *Actes de la 10e conférence TALNa Batz-sur-Mer*, 2:203–212.

Scheffer, T., Decomain, C., and Wrobel, S. (2001). Active hidden Markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA-2001)*, pages 309–318, Cascais, Portugal.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India.

Sells, P. (1984). *Syntax and Semantics of Resumptive Pronouns*. PhD thesis, University of Massachusetts, Amherst.

Seraji, M., Megyesi, B., and Nivre, J. (2012). Bootstrapping a Persian Dependency Treebank. *Linguistic Issues in Language Technology*, 7.

Seung, S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 287–294, Pittsburgh, USA.

Shen, D. (2007). Using semantic role to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Siegel, S. and Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw–Hill, Inc., second edition.

Silveira, N., Dozat, T., Marneffe, M.-C. D., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Skinner, B. (1957). *Verbal Behaviour*. Copley Publishing Group.

Skjærholt, A. (2014). A chance-corrected measure of inter-annotator agreement for syntax. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–944, Baltimore, USA.

Sokolovska, N. (2011). Aspects of Semi-Supervised and Active Learning in Conditional Random Fields. In *Proceedings of the European Conference on Machine Learning (ECML PKDD) 2011*, pages 273–288, Athens, Greece.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Proceedings of the first workshop on computational approaches to code switching. pages 62–72, Doha, Qatar.

Steedman, M., Osborne, M., Sarkar, A., Clark, S., Hwa, R., Hockenmaier, J., Ruhlen, P., Baker, S., and Crim, J. (2003). Bootstrapping statistical parsers from small datasets. In *Proceedings of the Tenth Conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 331–338, Stroudsburg, PA, USA.

Stenson, N. (1981). *Studies in Irish Syntax*. Tübingen: Gunter Narr Verlag.

Sulger, S. (2009a). Irish Clefting – The LFG Perspective. Master's thesis, Universität Konstanz.

Sulger, S. (2009b). Irish clefting and information-structure. In *Proceedings of the LFG09 Conference*, Cambridge, UK. CSLI Publications.

Tesnière, L. (1959). *Eléments de Syntaxe Structurale*. Klincksieck, Paris.

Tetreault, J., Foster, J., and Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 353–358, Stroudsburg, PA, USA.

The Christian Brothers (1988). *New Irish Grammar.* Dublin: C J Fallon.

Thompson, C., Califf, M. E., and Mooney, R. (1999). Active Learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Machine Learning Conference (ICML-99)*, pages 406–414, Bled, Slovenia.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA.

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.

Uí Dhonnchadha, E., Pháidín, C. N., and van Genabith, J. (2003). Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, 18:173–193.

Uí Dhonnchadha, E. and van Genabith, J. (2006). A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Uí Dhonnchadha, E. and van Genabith, J. (2010). Partial dependency parsing for Irish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 19–21, Valletta, Malta.

Uria, L., Estarrona, A., Aldezabal, I., Aranzabe, M. J., Díaz De Ilarraza, A., and Iruskieta, M. (2009). Evaluation of the syntactic annotation in EPEC, the reference corpus for the processing of Basque. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, (CICLing '09), pages 72–85, Berlin, Heidelberg. Springer-Verlag.

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P.-A. (2008). Using syntactic information for improving why-question answering. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 953–960, Manchester, UK.

Voutilainen, A. and Purtonen, T. (2011). A double-blind experiment on interannotator agreement: the case of dependency syntax and Finnish. In Bolette Sandford Pedersen, G. N., Pedersen, I. S. B. S., Nešpore, G., and Skadiņa, I., editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, pages 319–322, Riga, Latvia.

Ward, M. (2014). Using Irish NLP resources in primary school education. In *Proceedings of the First Celtic Language Technology Workshop*, pages 6–17, Dublin, Ireland.

Wróblewska, A. and Frank, A. (2009). Cross-lingual Projection of LFG F-Structures: Building an F-Structure Bank for Polish. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, Milan, Italy.

Xu, P., Kang, J., Ringgaard, M., and Och, F. (2009). Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08).*, pages 213–218, Marrakech, Morocco.

Appendix A

Annotation Guidelines for the Irish Dependency Treebank

In this appendix, we provide the set of annotation guidelines that were used to manually annotate the Irish Dependency Treebank. Our tagset has 47 labels (see Section 4 for the full list) to choose from. Some points to note when using this manual are:

- The bilexical labelled dependency notation reads as follows:
  deplabel(Head, Dependent)
- All examples are taken or adapted from the Irish dependency treebank.
- The guide is organised according to the possible dependents of each part of speech.
- The `top` label, which denotes the root label, is only used in examples using full sentences (not fragments or phrases).

A.1 Verb Dependents

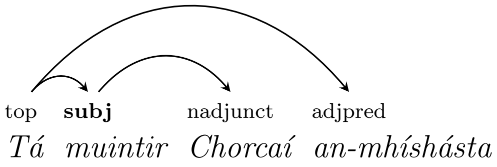
A.1.1 subj : subject – (Verb)

- Nominative subject of a verb

Example

Tá ***muintir*** *Chorcaí an-mhíshasta* : 'Cork **people** are very unhappy'

subj(Tá, muintir)

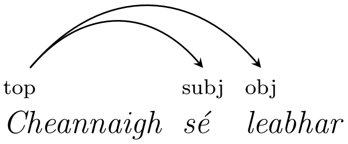


'Cork people are very unhappy'

Example

Cheannaigh sé ***leabhar*** : '**He** bought a book'

subj(Cheannaigh, sé)



'He bought a book'

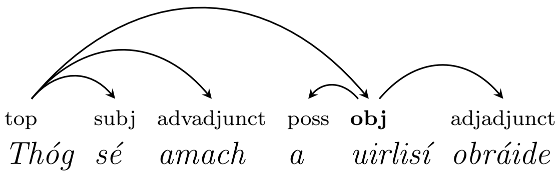
A.1.2 obj : object – (Verb)

- Direct object

Example

Thóg sé amach a ***uirlisí*** *obráide* : 'He took out his surgical **instruments**'

obj(thóg, uirlisí)



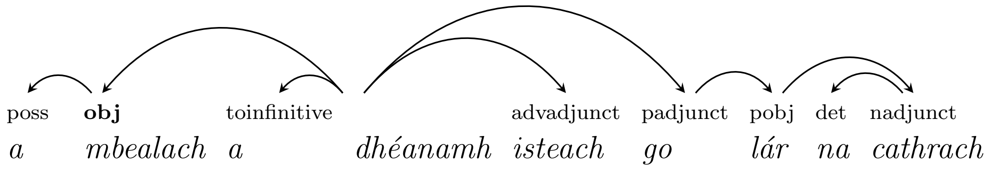
‘He took out his surgical instruments ’

- <u>Object of infinitive verb</u>

Example

*a **mbealach** a dhéanamh go lár na cathrach* : ‘to make their **way** into the city centre’

obj(dhéanamh, mbealach)



‘to make their way into the city centre’

- <u>Object of autonomous verb form</u>

Note: In Irish there is an understood (hidden) subject in the ‘briathar saor’ form of a verb

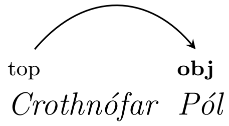
Example

*Crothnófar **Pól*** : ‘Paul will be remembered’ (lit. someone will remember Paul)

obj(Crothnófar, Pól)

Example

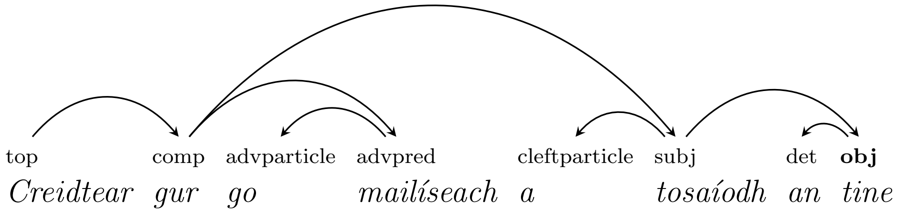
*Creditear gur go mailíseach a tosaíodh an **tine*** : ‘It is believed that the **fire**



‘Paul will be remembered’

was started maliciously’

obj(tosaíodh, tine)



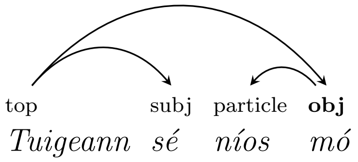
‘It is believed that the fire was started maliciously’

- Quantative object

Example

Tuigeann sé níos ***mó*** *anois* : ‘He understands **more** now’

obj(tuigeann, mó) [1]



‘He understands more’

A.1.3 obl/ obl2 : oblique nouns – (Verb)

Used for prepositions that are closely attached to a verb i.e. arguments/ complements as opposed to adjuncts. These include inflected prepositions.

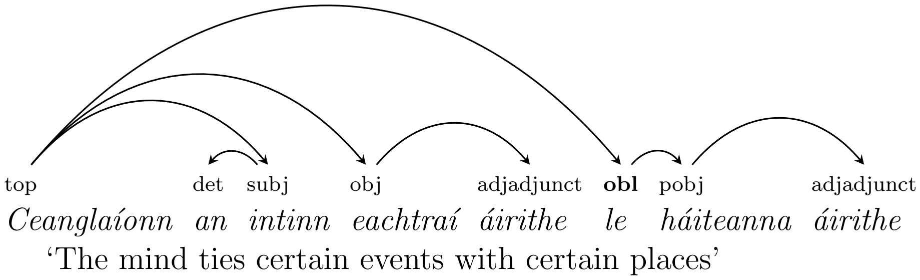
[1] Where *mó* is not modifiying an object to mean ‘more of’. (It is an object in itself)

- The verb selects for a transitive preposition

Example

Ceanglaíonn an intinn eachtraí áirithe ***le*** *háiteanna áirithe* : ‘The mind ties certain events **with** certain places’

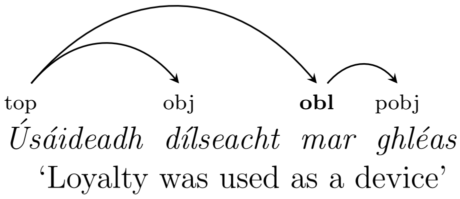
obl(gceanglaíonn, le)


top
det
subj
obj
adjadjunct
obl
pobj
adjadjunct
Ceanglaíonn an intinn eachtraí áirithe le háiteanna áirithe
‘The mind ties certain events with certain places’

Example

Úsáideadh dílseacht ***mar*** *ghléas le leatrom a dhéanamh..*: ‘loyalty was used as a device to oppress..’

obl(Úsáideadh, mar)


top
obj
obl
pobj
Úsáideadh dílseacht mar ghléas
‘Loyalty was used as a device’

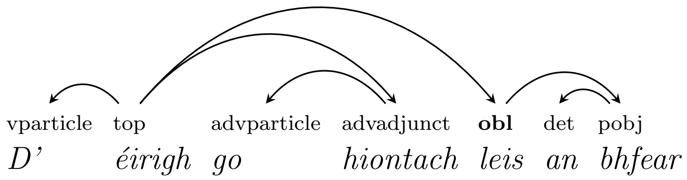
- These prepositions contribute to the meaning of the verb, they are not optional (arguments rather than modifiers) and take object complements.

Example

D’éirigh go hiontach ***leis*** *an bhfear*: ‘The man succeeded well’

obl(éirigh, leis) [2]

[2] These objects can take the form of prepositional pronouns (e.g. *D’éirigh go hiontach liom* ‘I succeeded well’)



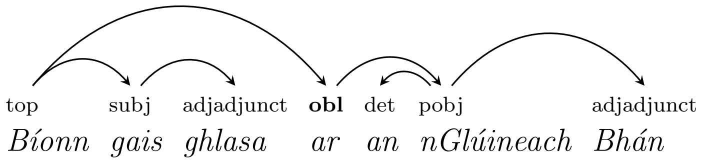
‘ The man succeeded well’

- Prepositions that are used with the verb *bí* to describe a state of a person/ thing in an idiomatic manner.

  Example

  Bíonn gais ghlasa ***ar*** *an nGlúineach Bhán* : ‘The Persicaria has green stalks’ (lit. Green stalks are [**on** the Persicaria])’.

  obl(Bíonn, ar)

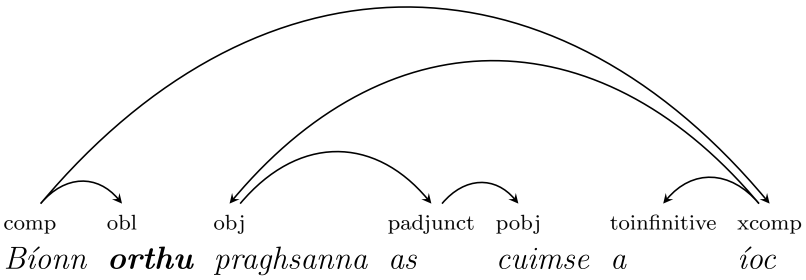


‘The Persicaria has green stalks’

Example

Bíonn ***orthu*** *praghsanna as cuimse a íoc* : ‘They have to pay incredible prices’

obl(Bíonn, orthu)

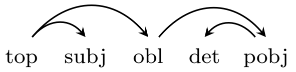


‘They have to pay incredible prices’

Example

Tá cáil ***ar*** *an leabhar* : ‘The book is famous’

obl(Tá, ar)



Tá cáil ***ar*** *an leabhar*

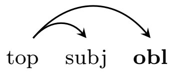
'The book is famous'

- Prepositions used with the verb *bí* 'be' to describe ownership of something.

  Example

  Tá súil ***agam***: 'I hope' (lit. Hope is at me)

  obl(tá, agam)



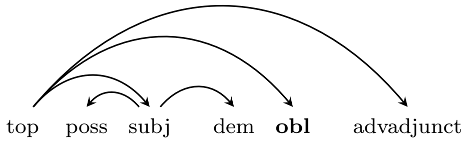
Tá súil agam

' I hope'

  Example

  Tá a fhios sin ***agat*** *anois*: 'You know that now'

  obl(tá, agat)



Tá a fhios sin agat anois

' You know that now'

- Oblique prepositions can precede the verb in indirect relative clause constructions

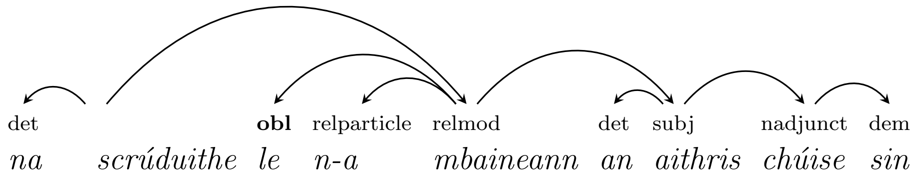
  (lena = le *(obl)* + n + a *(relparticle)*)

  Example

  na scrúduithe ***le*** *n-a mbaineann an aithris chúise sin*: 'the exams **to** which such representation relates'

  obl(mbaineann, le)

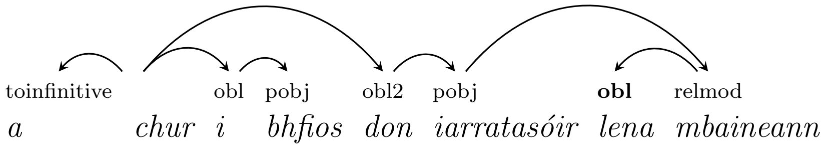
  Note: When the preposition and relative particle are combined (e.g. *lena*), the `relparticle` label is dropped in favour of `obl`.



‘the exams to which such representation relates ’

Example

a chur i bhfios don iarratasóir ***lena*** *mbaineann..* : ‘to communicate to the applicant **to whom** it concerns..’



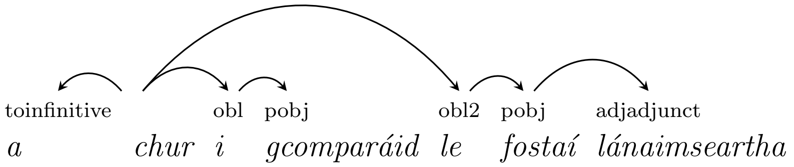
‘to communicate to the applicant concerned’

- When there are two oblique attachments to one verb, mark the second one as `obl2`.

Example

a chur ***i*** *gcomparáid* ***le***.. : ‘to compare with..’ (lit. ‘to put **in** comparison **with**)

obl(chur, i), *obl2(chur, le)*

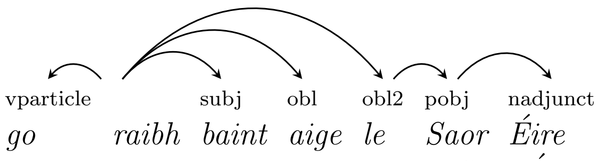


‘to compare with full-time employees’

Example

go raibh baint ***aige le*** *Saor Éire* : ‘that **he** had some association **with** Saor Éire’

obl(raibh, aige), obl2(raibh, le)

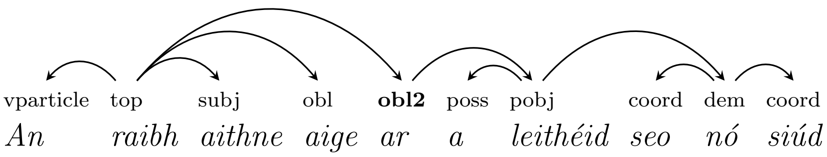


‘that he was associated with Saor Éire’

Example

An raibh aithne ***aige ar*** *a leithéid seo nó siúd?* : ‘Did he know this one or that one’

obl(raibh, aige), obl2(raibh, ar)



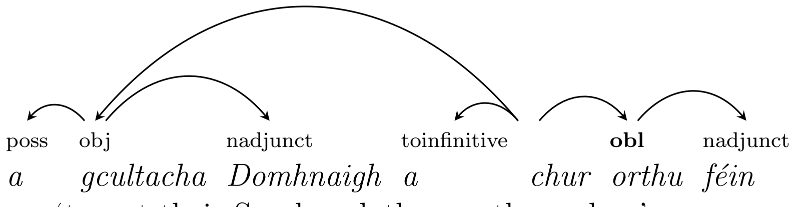
‘Did he know this one or that one?’

- Prepositional pronouns. These prepositions are inflected for an oblique object i.e. there is no overt *pobj*.

Example

a gcultacha Domhnaigh a chur ***orthu*** *féin* : ‘to put their Sunday clothes **on them**selves’

obl(chur, orthu)



‘to put their Sunday clothes on themselves’

A.1.4 particlehead – (Verb)

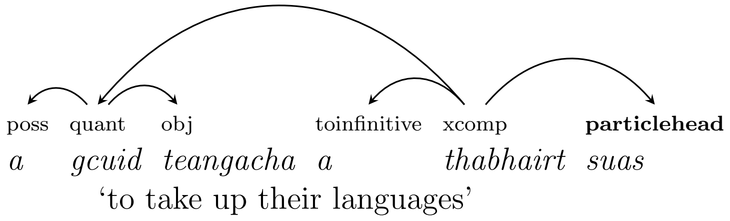
- A verb particle that is an adverb or a preposition.

Note: These verb particles cannot inflect for person or gender or be followed by a noun. These verbs are sometimes referred to as phrasal compound verbs.

Example

*a gcuid teangacha a thabhairt **suas*** : 'to take **up** their languages'

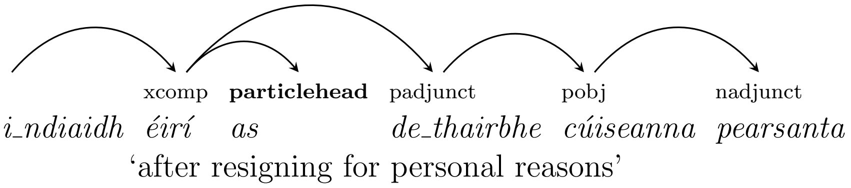
particlehead(thabhairt, suas)


poss
quant
obj
toinfinitive
xcomp
particlehead
a *gcuid* *teangacha* *a* *thabhairt* *suas*
'to take up their languages'

Example

*..i ndiaidh éirí **as** de thairbhe cúiseanna pearsanta* : '..after resigning (lit. rising **out**) for personal reasons'

particlehead(éirí, as)


xcomp
particlehead
padjunct
pobj
nadjunct
i_ndiaidh *éirí* *as* *de_thairbhe* *cúiseanna* *pearsanta*
'after resigning for personal reasons'

A.1.5 padjunct : prepositional adjunct – (Verb)

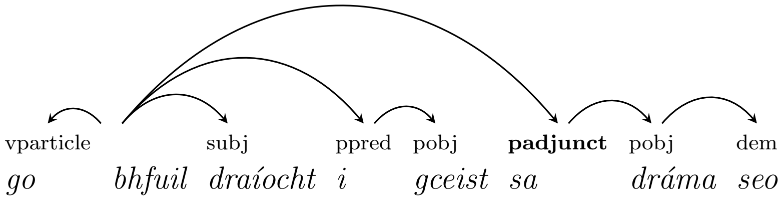
These prepositions are optional modifiers and tell us more about where or when something was done:

- Prepositions denoting time/ place

Example

*go bhfuil draíocht i gceist **sa** dráma seo* : 'that there is magic **in** this play'

padjunct(bhfuil, sa)

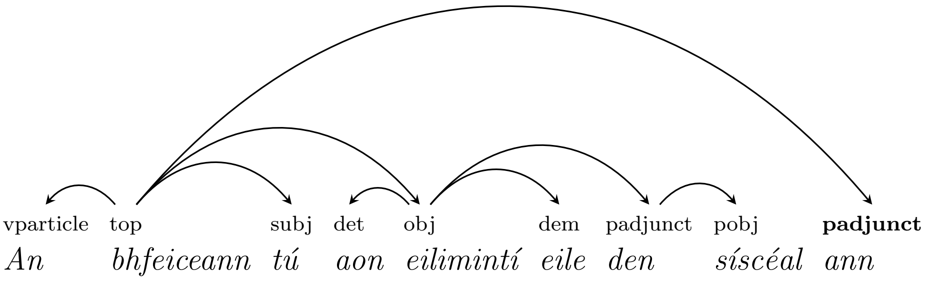


‘that there is magic in this play’

Example

An bhfeiceann tú aon eilimintí den síscéal eile ***ann****?* : ‘Do you seen any other elements of the fairytale **in it**?’

padjunct(bhfeiceann, ann)



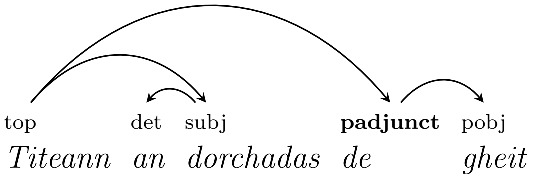
‘Do you see any other elements of the fairytale in it?’

- Adverbial prepositional phrases

Example

Titeann an dorchadas ***de*** *gheit* : ‘Darkness falls **with** a jolt’

padjunct(titeann, de)

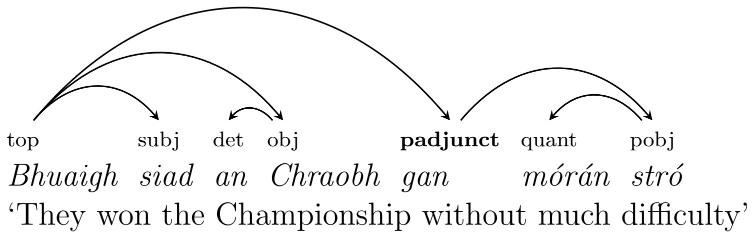


‘Darkness falls suddenly (with a jolt)’

Example

Bhuaigh siad an Chraobh gan mórán stró : 'They won the Championship without much difficulty'

padjunct(bhuaigh, gan)

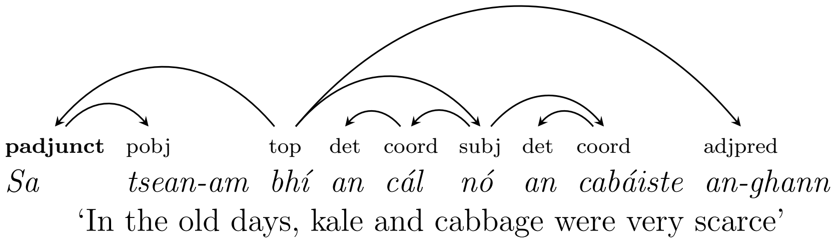

top
subj
det
obj
padjunct
quant
pobj
Bhuaigh siad an Chraobh gan mórán stró
'They won the Championship without much difficulty'

- Fronted prepositional phrases [3]

Example

Sa *tsean-am bhí an cál nó an cabáiste an-ghann* : '**In** the old days, kale and cabbage were very scarce'

padjunct(bhí, Sa)


padjunct
pobj
top
det
coord
subj
det
coord
adjpred
Sa tsean-am bhí an cál nó an cabáiste an-ghann
'In the old days, kale and cabbage were very scarce'

- Prepositional phrase cluster[4]

Note that the second preposition is dependent on the first

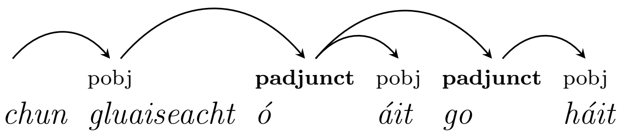
Example

chun gluaiseacht ***ó*** *áit* ***go*** *háit* : 'to move from place to place'

padjunct(gluaiseacht, ó), padjunct(ó, go)

[3] Note - the comma is not always present in Irish.

[4] Contrast with `obl2` above



‘to move from place to place’

A.1.6 advadjunct : adverbial adjunct – (Verb)

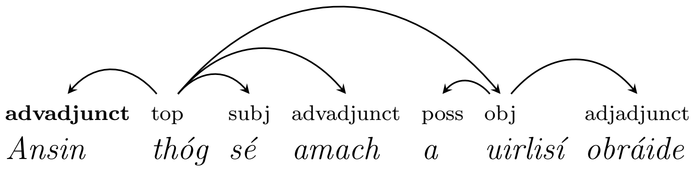
These are adverbs of manner/time/place (optional modifiers) that attach to the matrix verb.

- Adverbs of time

  Example

  Ansin *thóg sé amach a uirlisí obráide* : ‘**Then** he took out his surgical instruments’

  advadjunct(thóg, Ansin)



‘He took out his surgical instruments’

- Adverbs of manner

  Example

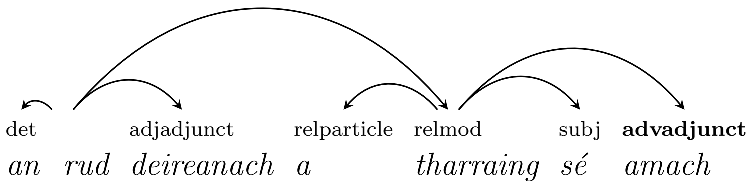
  : *an rud deireanach a tharraing sé* ***amach***..: ‘the last thing he pulled **out**..’

  advadjunct(tharraing, amach)

- Adverbs of place

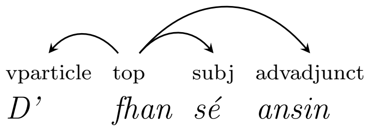
  Example

  D’fhan sé ***ansin***: ‘He stayed **there**’



‘the last thing he pulled out’

advadjunct(fhan, ansin)



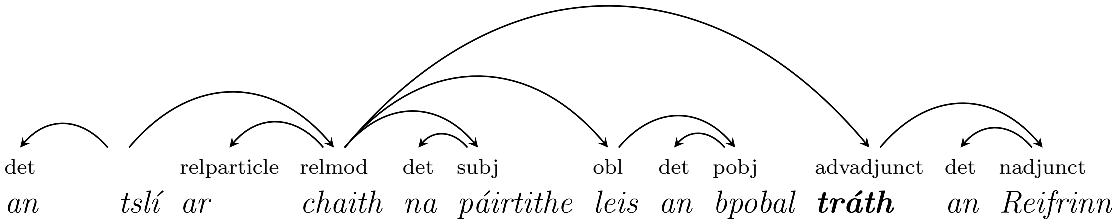
‘He stayed there’

- Nouns acting as adverbs

Example

..an tslí ar chaith na páirtithe leis an bpobal ***tráth*** *an Reifrinn* : ‘the way the parties treated the public **at the time of** the Referendum’

advadjunct(chaith, tráth)



‘the way the parties treated the public at the time of the Referendum’

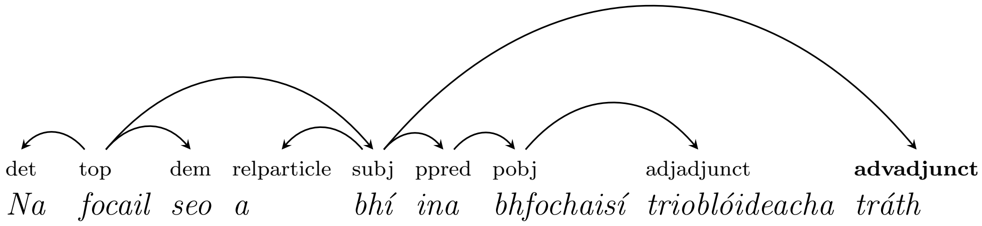
Example

Na focail seo a bhí ina bhfochaisí trioblóideacha ***tráth*** : ‘These words which were **once** troublesome’

advadjunct(bhí, tráth)

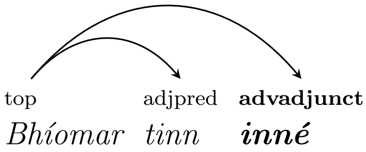
Example

Bhíomar tinn ***inné****.* : ‘We were sick **yesterday**’



‘These words which were once troublesome’

advadjunct(Bhíomar, inné)

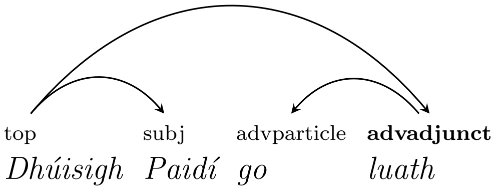


‘We were sick yesterday’

Example

Dhúisigh Paidí go ***luath*** : ‘Paidí woke **early**’

advadjunct(Dhúisigh, luath)



‘Paidí woke early’

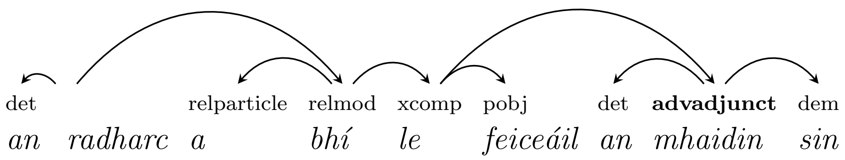
Example

an radharc a bhí le feiceáil an mhaidin sin: ‘the sight to behold that **morning**’

advadjunct(thugainn, mhaidin)

A.1.7 subadjunct : subordinate adjunct– (Verb)

Subordinate Clauses



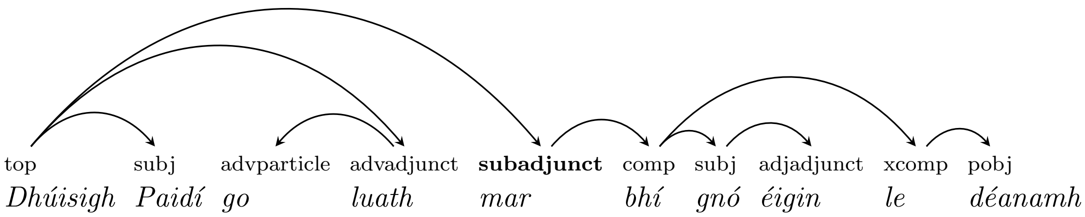
‘the sight to behold that morning’

- Subordinating conjunctions: *because, since-clauses*

Example

Dhúisigh Paidí go luath ***mar*** *bhí gnó éigin le déanamh* : ‘Paidí woke early **because** there was some work to do’

subadjunct(Dhúisigh, mar), *comp(mar, bhí)*

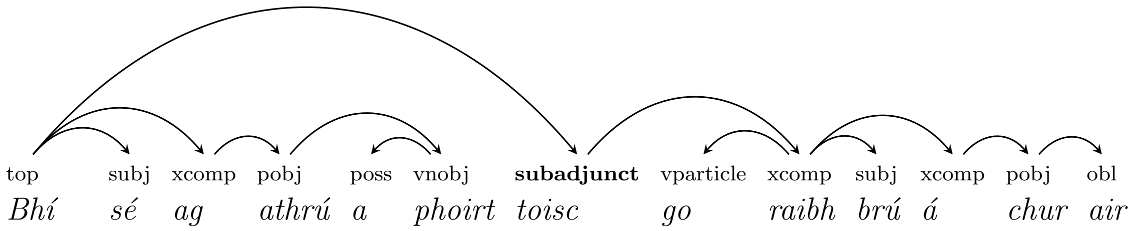


‘Paidí woke early because there was some work to do’

Example

Bhí sé ag athrú a phoirt toisc go raibh brú á chur air : ‘He was changing his tune since there was pressure on him’

subadjunct(Bhí, toisc), *comp(toisc, raibh)*



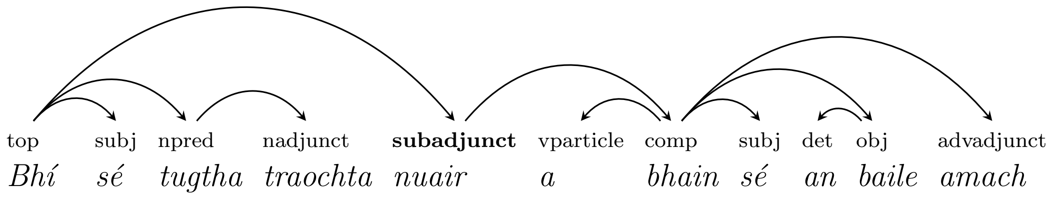
‘He was changing his tune since there was pressure on him’

- Subordinating conjunctions: *when-clauses*

Example

Bhí sé tugtha traochta ***nuair*** *a bhain sé an baile amach*: ‘He was exhausted **when** (by the time) he reached home’

subadjunct(bhí, nuair) , *comp(nuair, bhain)*

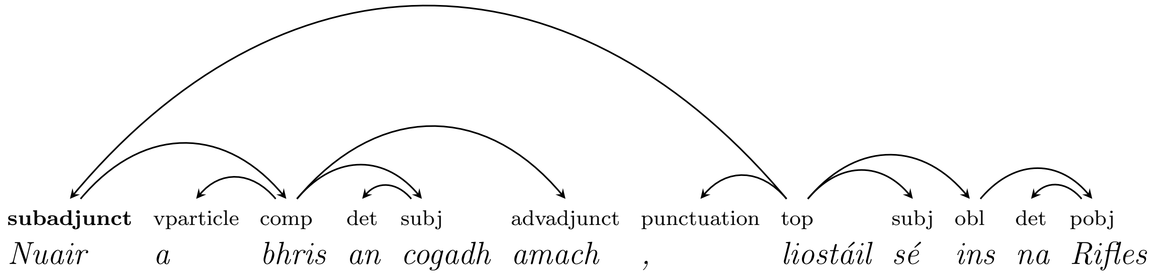


‘He was exhausted by the time he reached home’

Example

Nuair *a bhris an Cogadh Domhanda amach, liostáil sé ins na London Irish Rifles* : ‘**When** WWI broke out, he enlisted in the London Irish Rifles’

subadjunct(liostáil, Nuair), *comp(Nuair, bhris)*



‘When the war broke out, he enlisted with the Rifles’

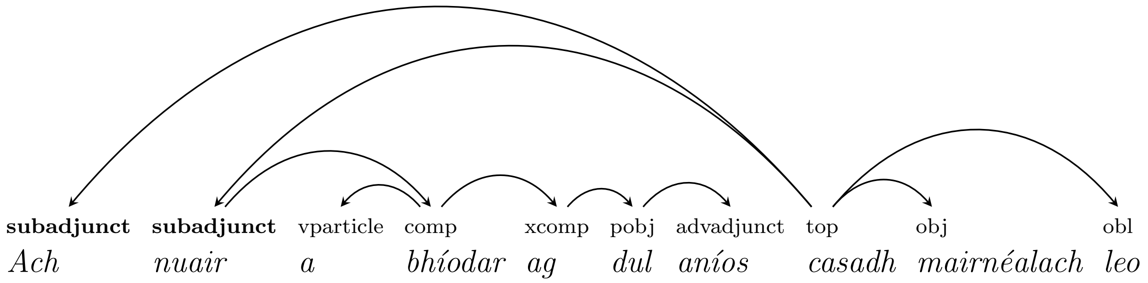
- Subordinating conjunctions: *but-clauses*

Note: If there are two subordinating conjunctions (e.g. But, when ..), the subordinate clause (`comp`) is attached to the one closest to it. Both are attached to the matrix clause as `subadjunct`.

Example

Ach *nuair a bhíodar ag dul aníos casadh mairnéalach leo* : ‘**But** when they were going down, they met sailors'

subadjunct(casadh, Ach), *subadjunct(casadh, nuair)*, *comp(nuair, bhíodar)*

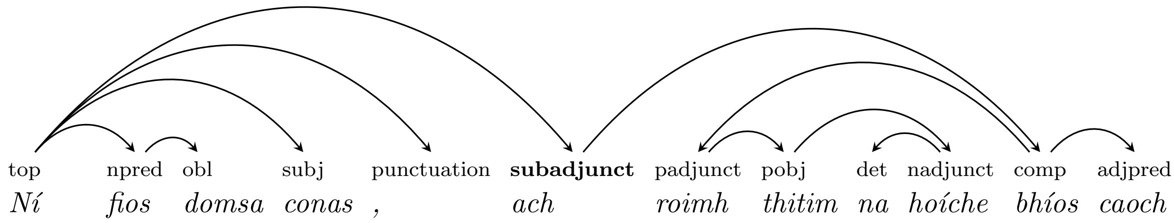


'But when they were going down, they met sailors'

Example

Ní fios domsa, ***ach*** *roimh thitim na hoíche bhíos caoch* : 'I don't know how, **but** before nightfall, I was blind-drunk'

subadjunct(Ní, ach), comp(ach, bhíos)



'I don't know how, **but** before nightfall, I was blind-drunk'

- Subordinating conjunctions. *if-clauses*: *má* (present), *dá* (conditional)

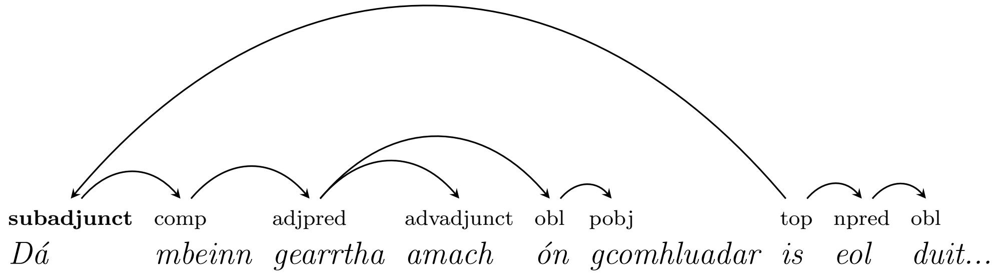
Example

Dá *mbeinn gearrtha amach ón gomhluadar is eol duit..* : '**If** I was cut off from the company, you know...'

subadjunct(is, Dá), *comp(Dá, mbeinn)*

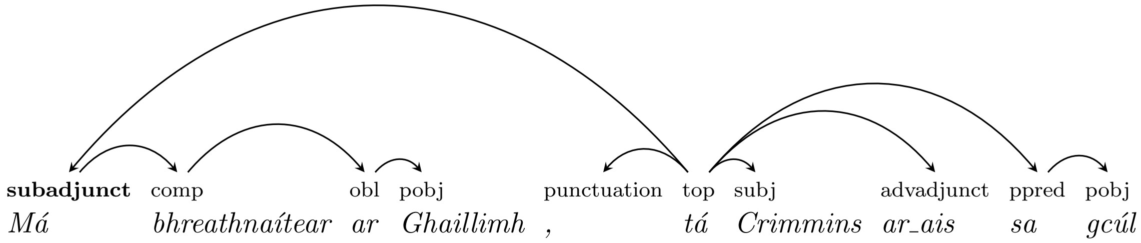
Example

Má bhreathnaítear ar Ghaillimh i dtosach, tá Michael Crimmins ar ais sa gcúl : 'If we look at Galway first, Michael Crimmins is back in goals'



‘If I was cut off from the company, you know...’

subadjunct(tá, má), *comp(má, bhreathnaítear)* [5]



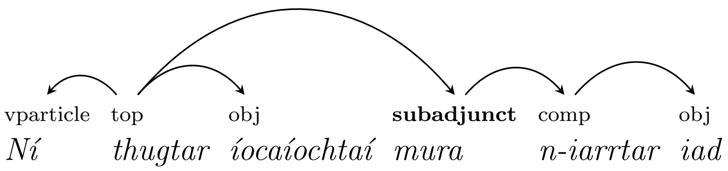
‘If we look at Galway, Crimmins is back in goals’

- Subordinating conjunctions. *unless-clauses*:

EXAMPLE

*Ní thugtar íocaíochtaí **mura** n-iarrtar iad* ‘Payments won’t be given **unless** they are requested’

subadjunct(thugtar, mura), *comp(mura, n-iarrtar)*



‘Payments won’t be given **unless** they are requested’

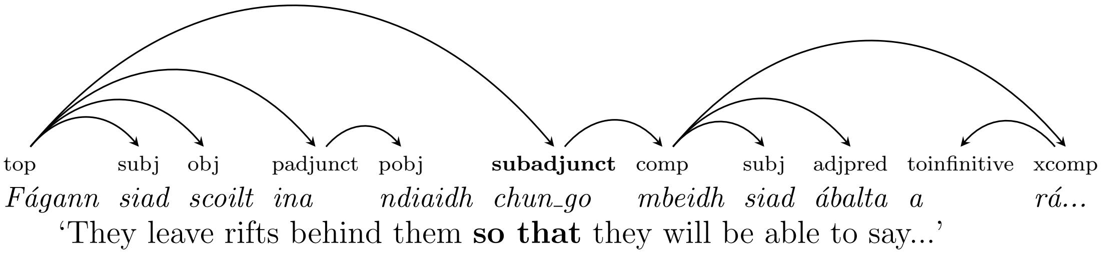
- Subordinating conjunctions. *so that, until-clauses*:

[5] For treatment of Más - see Section A.1.10

Example

Fágann siad scoilt ina ndiaidh ***chun go*** *mbeidh siad ábalta a rá....* 'They leave rifts behind them **so that** they will be able to say....'

subadjunct(Fágann, chun_go), *comp(chun_go, mbeidh)*

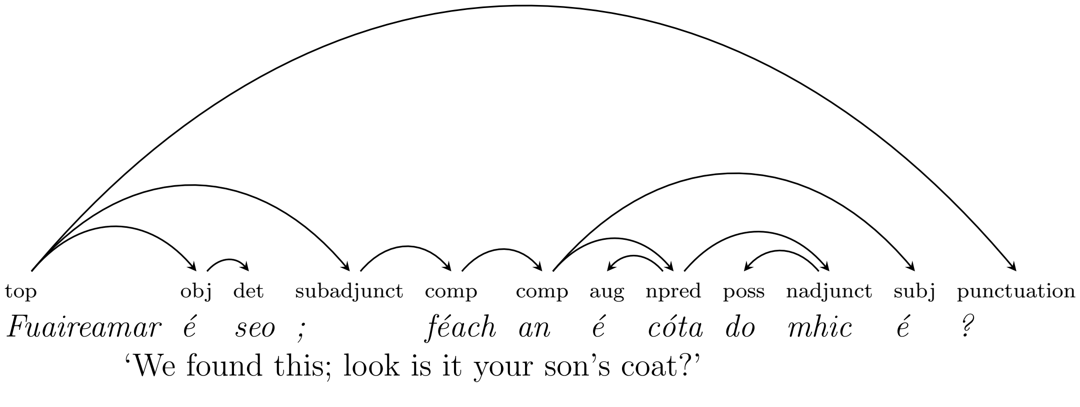


- Subordinating conjunctions. *Semi-colons*

Example

*Fuaireamar é seo****;*** *féach an é cóta do mhic é?* : 'We found this; look is it your son's coat?'

subadjunct(Fuaireamar, ;) comp(;, féach)



A.1.8 adjunct – (Verb)

These are verbal adjuncts that do not come under these categories: *advadjunct*, *padjunct*, or *subadjunct*.

Note that we label *mar* 'because', *nuair* 'when' and *ach* 'but' as `adjuncts` if they do not introduce a subordinate clause.

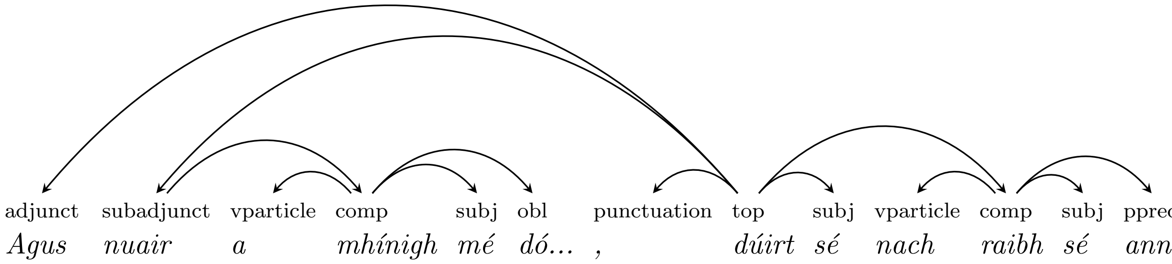
- Connectives

  agus is normally a coordinate conjunction, but it can also introduce a new sentence and is sometimes used alongside another subordinate conjunction, such as *nuair* 'when' for example.
  It is quite common and acceptable to start a sentence with 'Agus' in Irish.

  Example

  Agus *nuair a mhínigh mé dó..., dúirt sé..* : 'And when I explained to him.., he said..'

  adjunct(dúirt, Agus), *subadjunct(dúirt, nuair)*, *comp(nuair, mhínigh)*

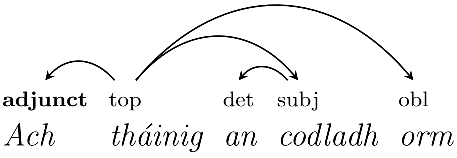


'And when I explained to him.., he said it wasn't there'

  Example

  Ach *tháinig an codladh orm* : '**But** I fell asleep' (lit. But sleep came on me)

  adjunct(tháinig, Ach)



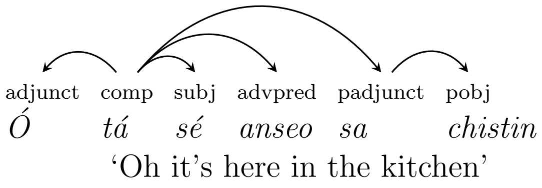
'But I fell asleep'

- Interjection

Example

'**Ó** tá sé anseo sa chistin' : '**Oh** it's here in the kitchen'

adjunct(tá, Ó)

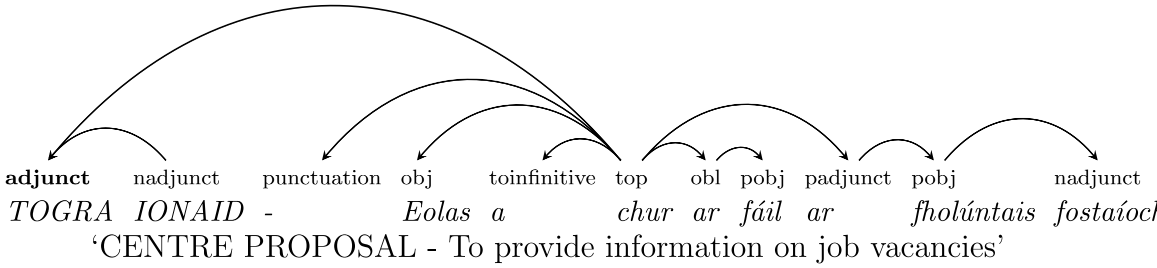


- Headings: the root of the heading (RH) is dependent on the root of the main sentence, and all other parts of the heading are dependent on RH.

  Example

  TOGRA IONAID *- Eolas a chur ar fáil ar fholúntais fostaíochta* : 'CENTRE PROPOSAL - To provide information on job vacancies'

  adjunct(Eolas, TOGRA), nadjunct(TOGRA, IONAID)



A.1.9 xcomp : open complement – (Verb)

A.1.9.1 xcomp : open complements

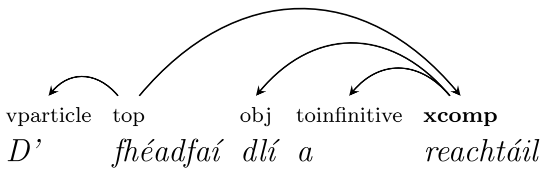
- infinitival phrases

  - share a subject with the matrix verb, i.e. the subject is not present in the infinitival phrase
  - preceded in Irish with *a* or *le*

Example

D'fhéadfaí dlí a ***reachtáil***; 'A law could be legislated (lit. someone could **legislate** a law)'

xcomp(fheadfaí, reachtáil)

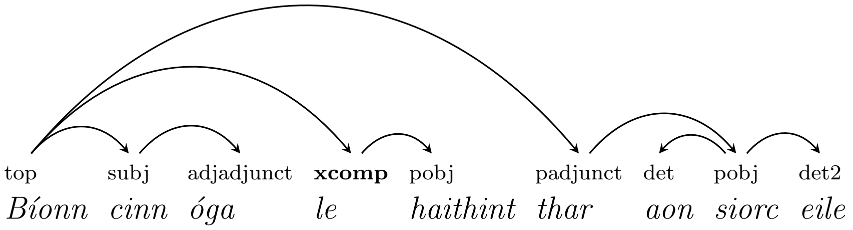


'A law could be legislated'

Example

Bíonn cinn óga le ***haithint*** *thar aon siorc eile* : 'Young ones are seen above any other shark'

xcomp(haithint, le)

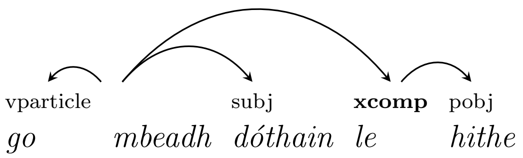


'Young ones are seen above any other shark'

Example

go mbeadh dóthain **le** *hithe* : 'that there would be enough to eat'

xcomp(mbeadh, le)

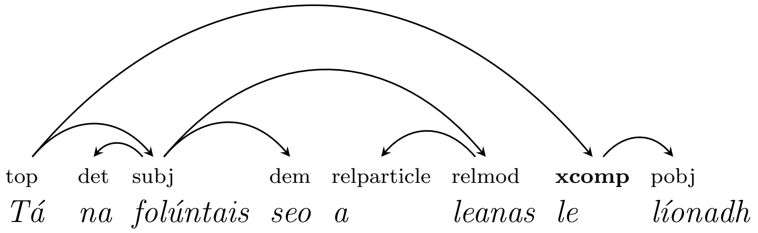


'that there would be enough to eat'

Example

Tá na folúntais seo a leanas **le** *líonadh* : 'The following vacancies are to be filled'

xcomp(tá, le)



'The following vacancies are to be filled'

A.1.9.2 xcomp : progressives

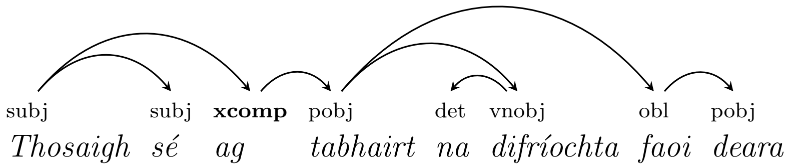
- <u>'ag' progressive aspectual phrases</u>

  – denoted through the use of the substantive verb 'bí' followed by the preposition 'ag' and a verbal noun. No equivalent in English.

Example

*Thosaigh sé [**ag** tabhairt na difríochta faoi deara]* : 'He started noticing the differences'

xcomp(Thosaigh, ag)



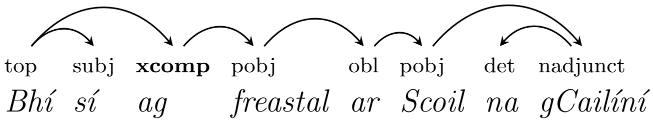
'He started noticing the differences'

Example

*Bhí sí [**ag** freastal ar scoil]* : 'She was attending the Girl's School'

xcomp(Bhí, ag)

- <u>'ar' progressive aspectual phrases</u>

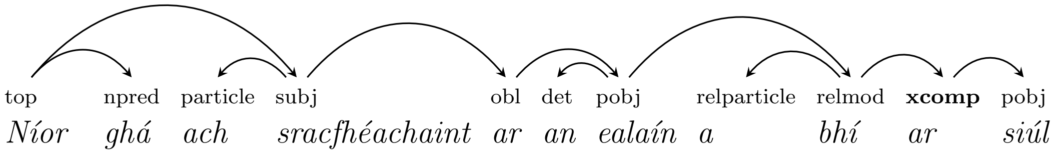


'She was attending the Girl's School'

Example

*Níor ghá ach sracfhéachaint ar an ealaín a bhí [**ar** siúl]* : 'You only have to glance at the art that was in progress'

xcomp(bhí, ar)

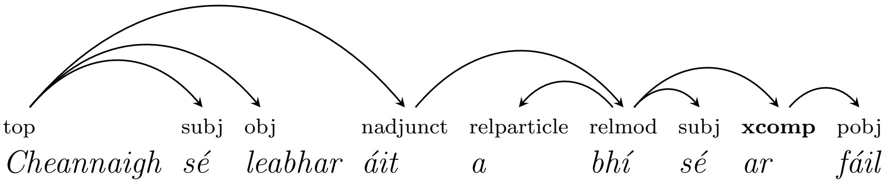


'You only have to glance at the art that was in progress'

Example

*Cheannaigh sé leabhar áit a bhí sé **ar** fáil*: 'He bought books anywhere they were available'

xcomp(bhí, ar)



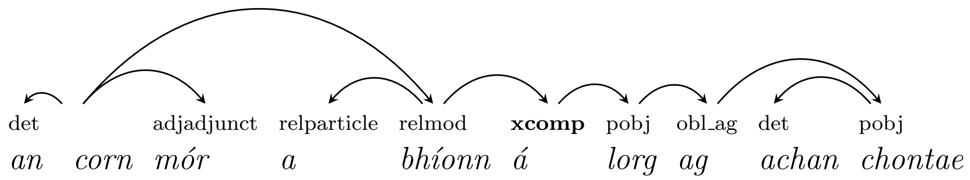
'He bought books anywhere they were available'

- *á* progressive aspectual phrases

Example

*an corn mór a bhíonn [**á** lorg]* : 'the big cup that every county seeks (lit. **at its** seeking)'

xcomp(bhíonn, á)



‘the big cup that every county seeks’

A.1.10 comp : closed complement – (Verb)

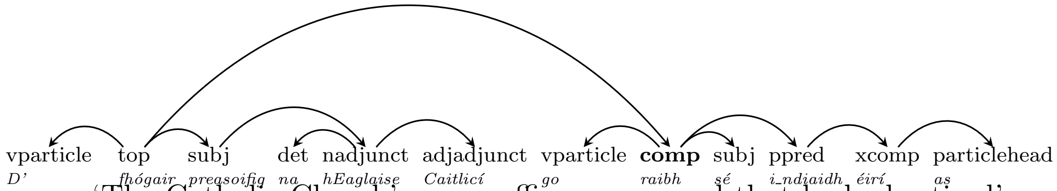
Closed complement clauses are clauses that contain a subject. The head is usually a verb. The link is usually between the matrix verb and the complement verb. Complements introduced by a copula (*gur*, *gurb* etc.) are referred to as copular complements.

- Clauses introduced by complementiser *go/gur/nach/nár* ‘that’, ‘that-not’

  Example

  D’fhógair preasoifig na hEaglaise Caitlicí [go ***raibh*** *sé i ndiaidh éirí as]* : ‘The Catholic Church press office announced that he had retired’

  comp(fhógair, raibh)



‘The Catholic Church’s press office announced that he had retired’

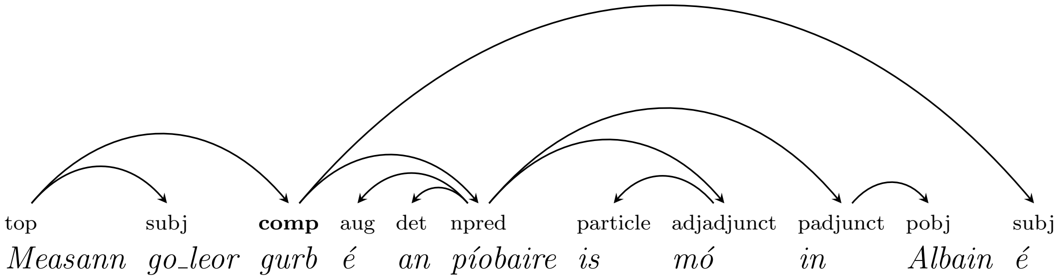
- Copular complements *gurb, narbh.*

  - both *gurb* and *narbh* are contraction of complementiser and main clausal copular verb
  - gur + is (*gurbh*), nar + is (*narbh*)

Example

*measann go leor [**gurb** é an píobaire is mó in Albain é]* : 'many believe **that** he **is** the biggest piper in Scotland'

comp(measann, gurb)



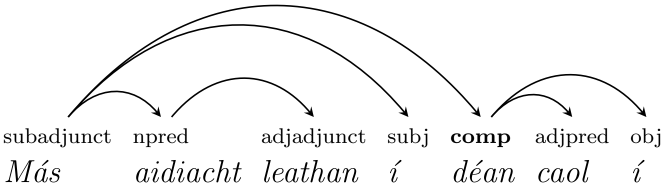
'Many believe that he is the biggest piper in Scotland'

- Más → Má + Is

Example

***Más** aidiacht leathan í déan caol í* : '**If it is** a broad adjective, make it slender'

comp(déan, Más)



'If it is a broad adjective, make it slender'

A.1.11 pred : predicate – (Verb)

Non-verbal predicates are labelled *pred.* This label is further divided into categories of *npred, adjpred, ppred, advpred* when used with the substantive verb *bí.*

A.1.11.1 adjpred - Adjectival Predicate

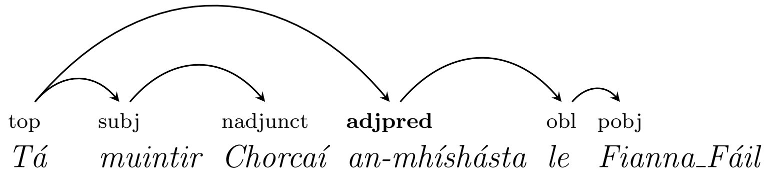
- Describing states

  – Adjectives used with the verb ‘to be’ (substantive verb *bí*) to describe a state

EXAMPLE

Tá muintir Chorcaí ***an-mhíshásta*** *le Fianna Fáil* : ‘Cork people are **very dissatisfied** with Fianna Fáil’

adjpred(Tá, an-mhíshásta)



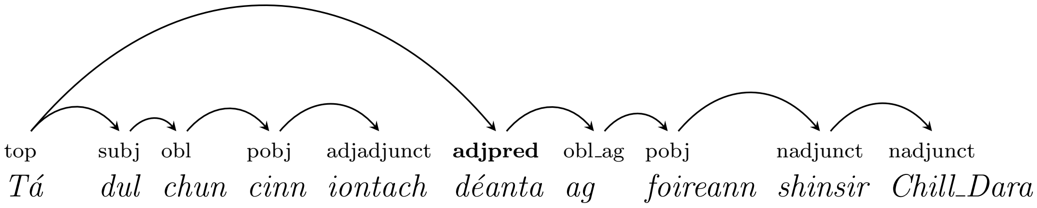
‘The people of Cork are very dissatisfied with Fianna Fáil’

- Verbal Adjectives

EXAMPLE

Tá dul chun cinn iontach ***déanta*** *ag foireann shinsir Chill Dara* : ‘Great progress has been **made** by Kildare Senior Team’

adjpred(tá, déanta)



‘Great progress has been made by Kildare Senior Team’

A.1.11.2 ppred - Prepositional Predicate

PP predicates are more common in Irish than in English. For example, they are used for describing a state, a profession, membership or denoting ownership. Note that the preposition is not normally realised in the English translation.

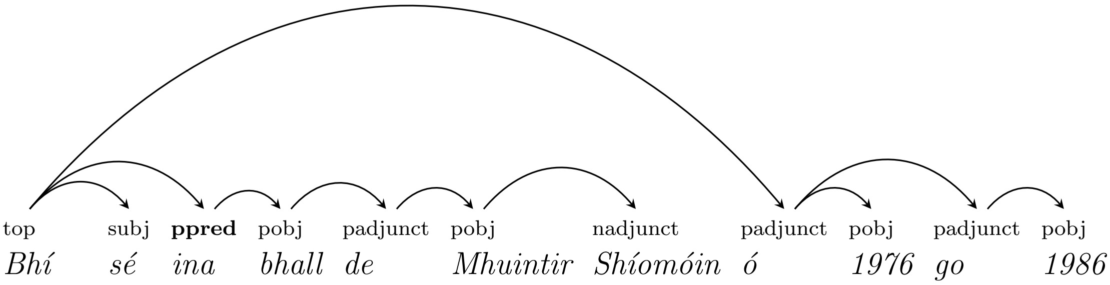
Also note, square brackets mark where `ppred` denotes a predicate phrase and the preposition heads that phrase.

- Prepositional predicates

Example

*Bhí sé [**ina** bhall de Mhuintir Shíomóin] ó 1976 go 1986* : 'He was [[6] a member of the Simon Community] from 1976 to 1986'

ppred(Bhí, ina)



'He was a member of the Simon Community from 1976 to 1986'

Example

*Tá sé [**i** gceist] an scéal a leathadh* : 'It is planned to spread the story'

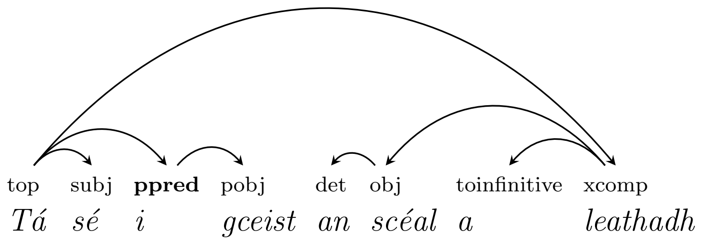
ppred(Tá, i)

Example

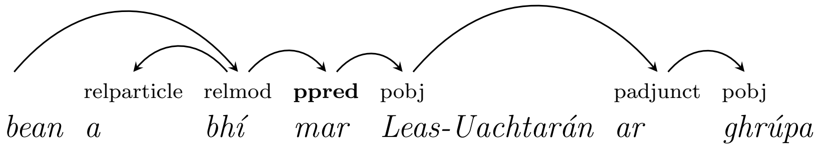
*..bean a bhí [**mar** Leas-Uachtarán] ar ghrúpa*: '..a woman who was [**(as)** Vice-President of a group]'

ppred(bhí, mar)

[6] literally '**in his** membership'



‘It is planned to spread the story’



‘a woman who was Vice-President of a group’

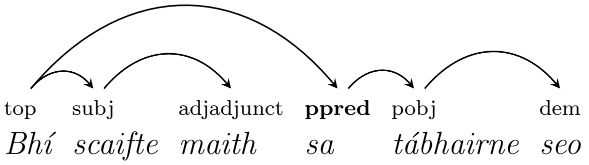
- Locative prepositions as predicates

  - These structures are equivalent to English ‘existential there’ constructions

Example

Bhí scaifte maith ***sa*** *tabhairne* : ‘There was a good crowd **in** the pub’

ppred(Bhí, sa)



‘There was a good crowd **in** the pub’

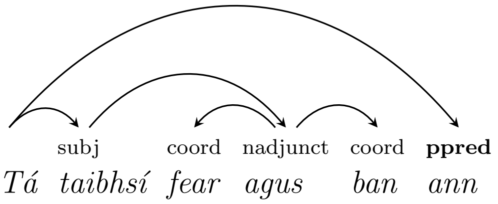
Example

Tá taibhsí fear agus ban ***ann*** : ‘There are male and female ghosts **(in it)**’

ppred(Tá, ann)

A.1.11.3 advpred - Adverbial Predicate

- Adverbs in predicate position

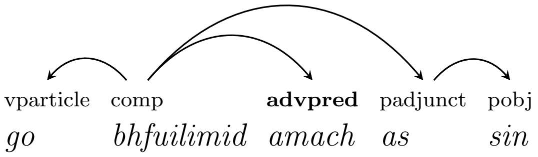


‘There are male and female ghosts’

Example

*go bhfuilimid **amach** as sin*: ‘that we are **out** of that’

advpred(bhfuilimid, amach)

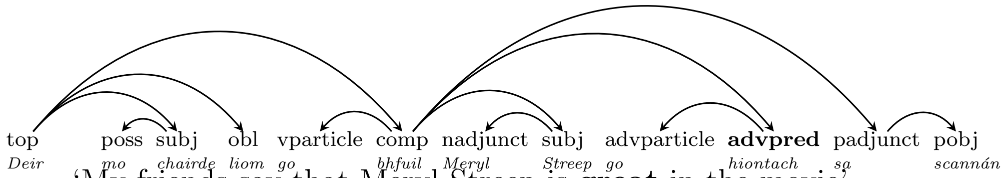


‘that we are **out** of that’

Example

*Deir mo chairde go bhfuil Meryl Streep go **hiontach** sa scannán* : ‘My friends say that Meryl Streep is **great** in the movie’

advpred(bhfuil, hiontach)

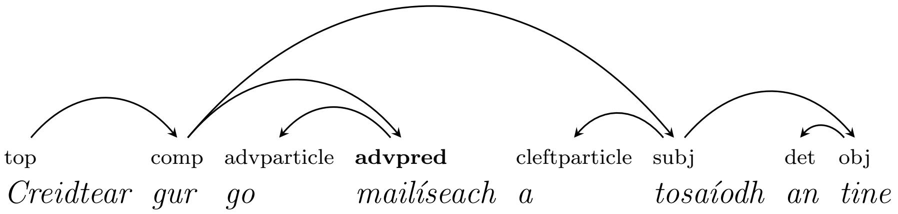


‘My friends say that Meryl Streep is **great** in the movie’

Example

*Creidtear gur go **mailíseach** a tosaíodh an tine* : ‘It’s believed that the fire was started **maliciously**’

advpred(gur, mailíseach)



‘It’s believed that the fire was started **maliciously**’

A.1.12 relparticle : relative particle – (Verb)

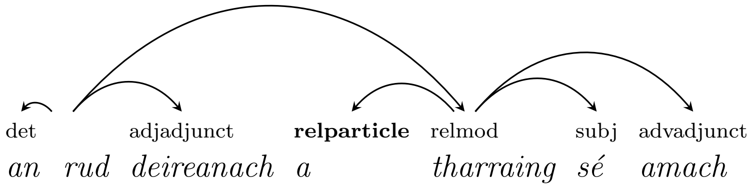
- *a* and *ar* relative particles

  - *a* is a direct relative, *ar* is an indirect relative
  - these particles precede, and are dependents of, relative modifier verbs (*relmod*).

EXAMPLE

..an rud deireanach ***a*** *tharraing*(relmod) *sé amach* : ‘..the last thing [**that**] he pulled out’

relparticle(tharraing, a)



‘the last thing he pulled out’

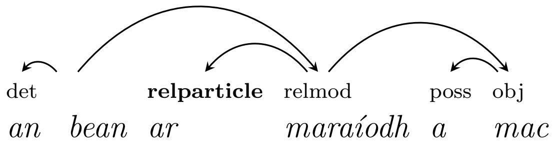
EXAMPLE

an bean ***ar*** *maraíodh a mac* : ‘the woman **whose** son was killed’

relparticle(maraíodh, ar)

- *dá* - relative particle

  - Similar in meaning to ‘that which is’

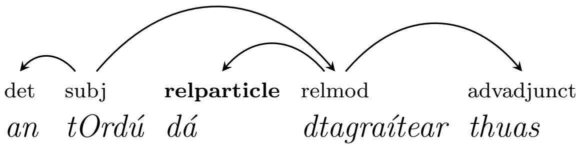


‘the woman **whose** son was killed’

Example

*an tOrdú **dá** dtagraítear thuas* : ‘the Order **that is** referred to above’

relparticle(dtagraítear, dá)



‘the Order **that is** referred to above’

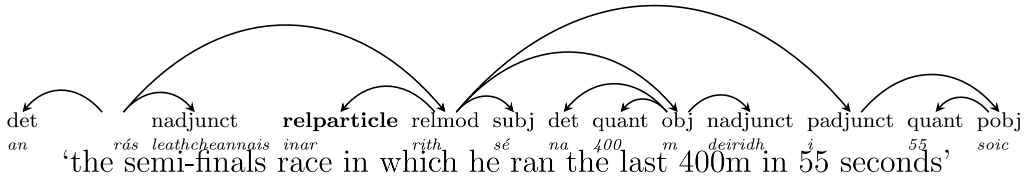
- *ina, inar* indirect relative particles

  - meaning ‘in which’

Example

*an rás leathcheannais **inar** rith sé na 400m deiridh i 55 soic* : ‘the semi-finals race **in which** he ran the last 400m in 55 seconds’

relparticle(rith, inar)



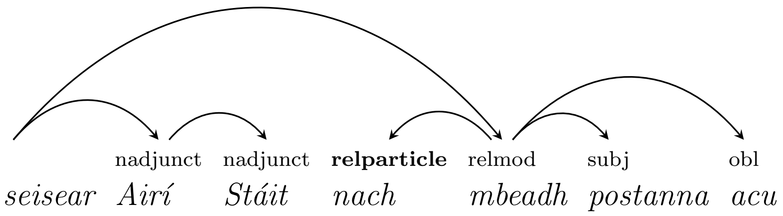
‘the semi-finals race in which he ran the last 400m in 55 seconds’

- *nár, nach* – negative relative particles

Example

*seisear Airí Stáit **nach** mbeadh postanna acu* : ‘six Ministers of State that wouldn't have jobs'

relparticle(mbeadh, nach), relmod(seisear, mbeadh)



'six Ministers of State who wouldn't have jobs'

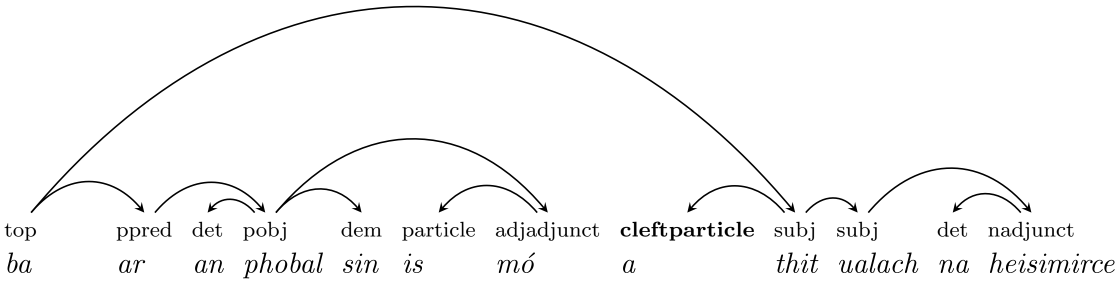
A.1.13 cleftparticle : cleft particle – (Verb)

- <u>Cleft particles</u>

  - We use this label to differentiate from *relparticle*, which is used only when the nominal head is present.

EXAMPLE

*Ba ar an phobal sin is mó **a** thit ualach na heisimirce* : 'It was mostly on that community that the burden of emigration hit'
cleftparticle(thit, a)

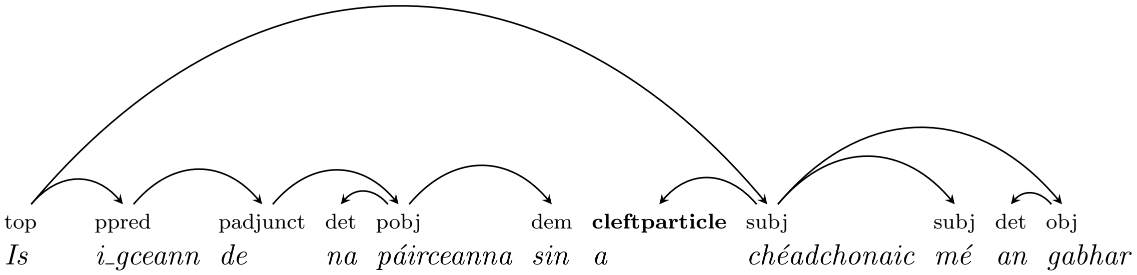


'It was mostly on that community that the burden of emigration hit'

EXAMPLE

*Is i gceann de na páirceanna sin **a** chéadchonaic mé an gabhar* : 'It was in one of those fields that I first saw the goat'

cleftparticle(chéadchonaic, a)



'It was in one of those fields that I first saw the goat'

A.1.14 vparticle : verb particle – (Verb)

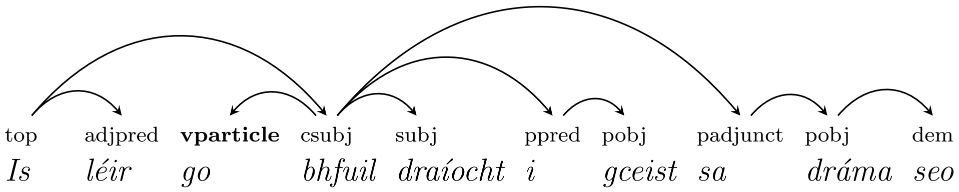
- Verb particles introducing complement clauses *go/gur/nach/nár*

  – similar to 'that' complementiser in English.

Example

Is léir ón teideal ***go*** *bhfuil draíocht i gceist sa dráma seo* : 'It's clear from the title **that** there is magic in this play'

vparticle(bhfuil, go)

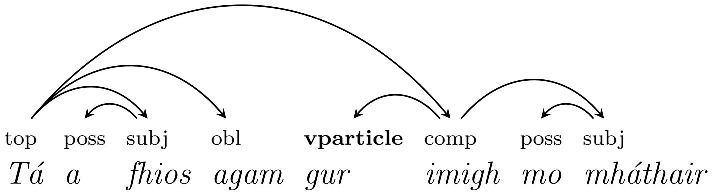


'It is clear that there is magic in this play'

Example

Tá a fhios agam ***gur*** *imigh mo mháthair* : 'I know **that** my mother left'

vparticle(imigh, gur)


top poss subj obl **vparticle** comp poss subj
Tá a fhios agam gur imigh mo mháthair

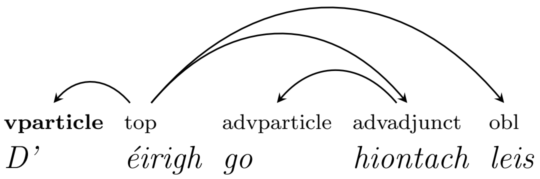
‘I know that my mother left’

- <u>Tensed verb particles</u>

Example

D’*eirigh go hiontach leis* : ‘He succeed**ed** greatly’

vparticle(éirigh, d’)

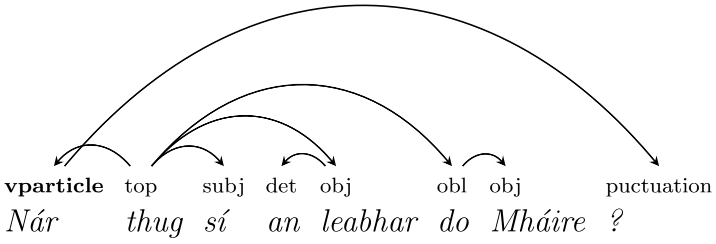

vparticle top advparticle advadjunct obl
D’ éirigh go hiontach leis

‘He succeeded well’

Example

Nár *thug sí an leabhar do Mháire?* : ‘**Did** she **not** give the book to Maire?’

vparticle(thug, nár)


vparticle top subj det obj obl obj puctuation
Nár thug sí an leabhar do Mháire ?

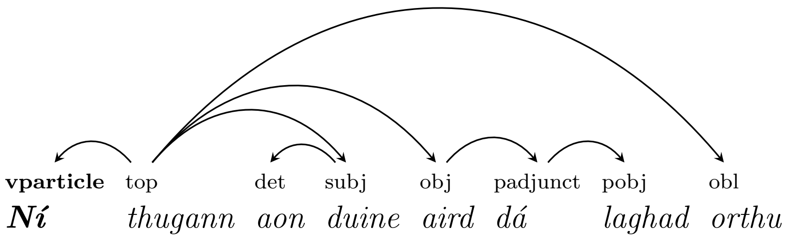
‘Did she not give the book to Maire?’

- <u>Negative verb particles</u>

Example

Ní *thugann aon duine aird dá laghad orthu* : ‘Nobody has any respect for them’

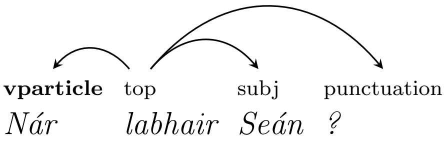
vparticle(thugann, Ní)



‘Nobody has any respect for them’

Example

Nár *labhair Seán?* : ‘**Didn’t** John speak?’



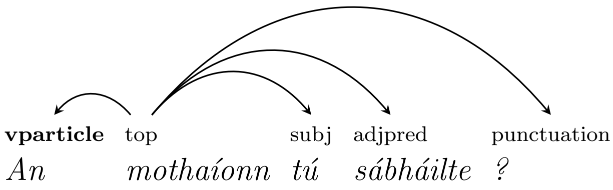
‘Didn’t John speak?’

- Interrogative verb particles

Example

An *mothaíonn tú sábháilte?* : ‘Do you feel safe?’

vparticle(mothaíonn, An)

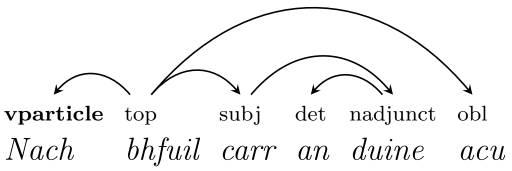


‘Do you feel safe?’

Example

Nach *bhfuil carr an duine acu?* : ‘Don’t they have a car each?’

vparticle(bhfuil, Nach)

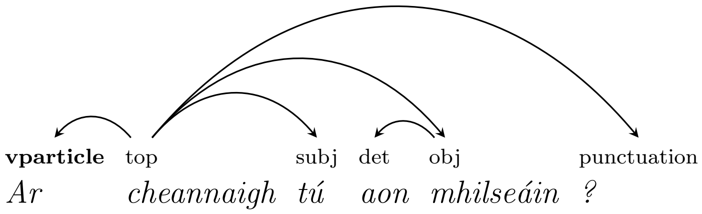


‘Don’t they have a car each?’

Example

Ar *cheannaigh tú aon mhilseáin?* : ‘Did you buy any sweets?’

vparticle(cheannaigh, Ar)



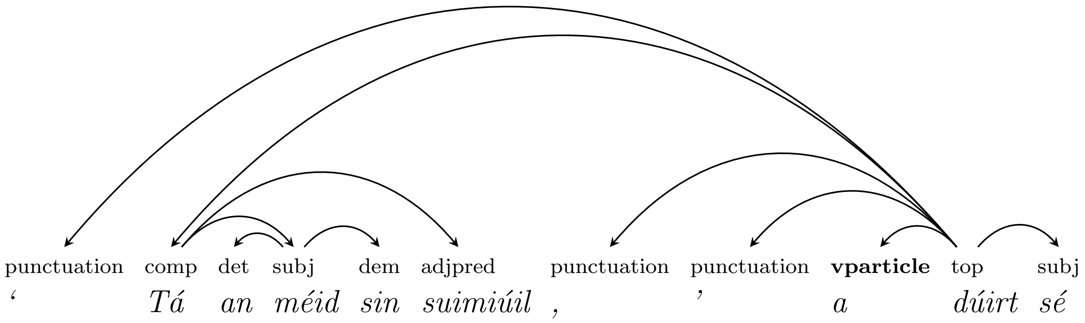
‘Did you buy any sweets?’

- Direct speech particles

Example

Tá an méid sin suimiúil a d’inis tú dom, ***a*** *dúirt sé* : “‘What you have told me is interesting”, [] he said.’

vparticle(dúirt, a)



“‘What you have told me is interesting”, [] he said.”

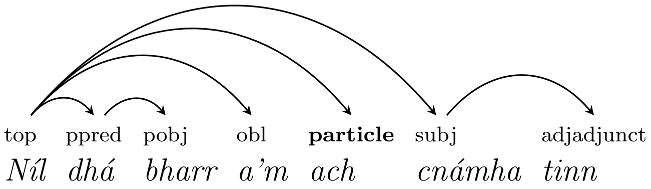
A.1.15 particle : verb particle — (Verb)

- 'only' particle – (níl ... ach)

Example

Níl dhá bharr a'm ***ach*** *cnámha tinn* : 'I have only sick hands' ('lit. there is not a result at-me **but** sick hands')

particle(níl, ach)

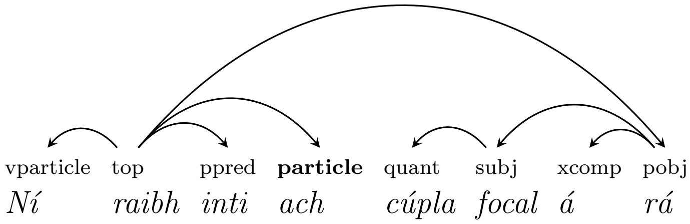


'I have only sick hands'

Example

Ní raibh inti ***ach*** *cúpla focal á rá*: 'she didn't have in her **but** a couple of words to say'/ 'she only had a couple of words to say'

particle(raibh, ach)



'She only had a couple of words to say'

A.1.16 addr : addressee – (Verb)

- Vocatives

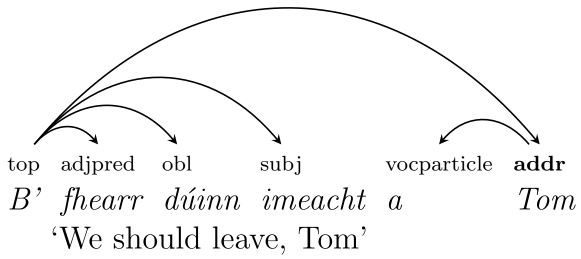
Addressees are a dialogue participant which is usually dependent on the main verb. Sometimes a vocative particle '*a*' is used. (There is no equivalent in

English.)

Example

B'fhearr dúinn imeacht a Tom: 'We should leave, Tom'

addr(B', Tom), vocparticle(Tom, a)



A.2 Noun Dependents

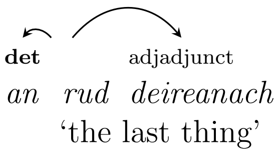
A.2.1 det : determiner – (Noun)

- Definite articles: *an* (singular), *na* (plural)

Example

an *rud deireanach*: 'the last thing'

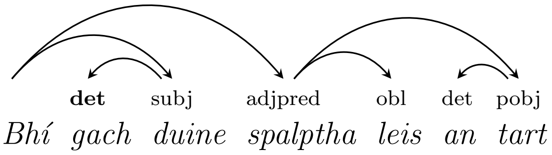
det(rud, an)



Example

Bhí ***gach*** *duine spalptha leis an tart* : '**Every**one was parched with thirst'

det(duine, gach)



‘Everyone was parched with thirst’

A.2.2 det2 : second determiner – (Noun)

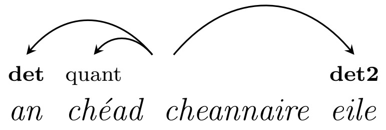
There are instances where two determiners are required. Sometimes the combination of both have just one interpreted meaning.

- Pre- and post-determiner combination

  Example

  an *chéad cheannaire* ***eile***: ‘the **next** leader’

  det(cheannaire, an), *det2(cheannaire, eile)*



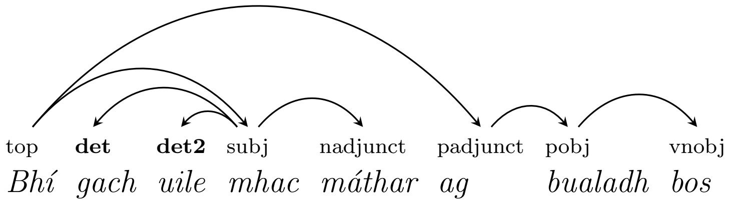
‘the next leader’

- Two pre-determiner combination (used for emphasis purposes).

  Example

  Bhí ***gach uile*** *mhac máthar ag bualadh bos*: ‘(**Each** and) **Every** mother’s son was clapping’

  det(mhac, gach), *det2(mhac, uile)*



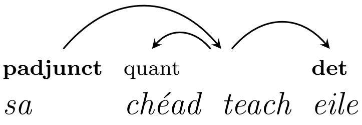
‘Every single mother’s son was clapping’

- NOTE in cases where a preposition incorporates the determiner, there is no need to label the following determiner as det2.

Example

Sa chead teach ***eile*** : ‘In the **next** house’

det(teach, eile)



‘in the next house’

A.2.3 dem : demonstrative – (Noun)

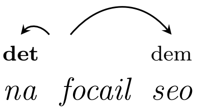
These demonstratives are used with definite articles.

- Demonstratives – (*seo, sin, úd, siúd*)

Example

Na *focail* ***seo***: ‘**These** words’

det(focail, Na), *dem(focail, seo)*



‘these words’

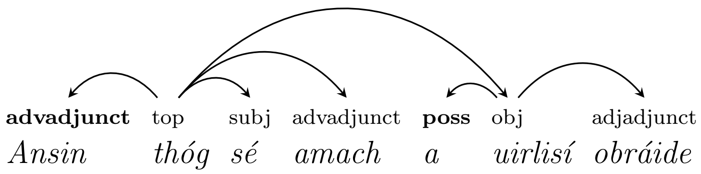
A.2.4 poss : possessive – (Noun)

- Possessive pronouns.

Example

Thóg sé amach ***a*** *uirlisí*: ‘He took out **his** tools’

poss(uirlisí, a)



‘He took out his surgical instruments’

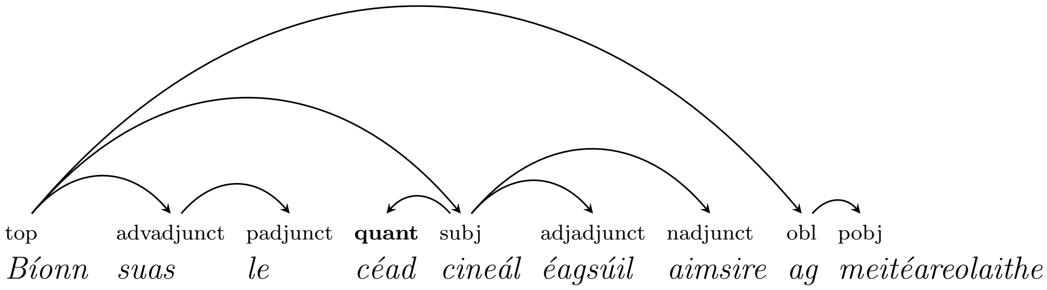
A.2.5 quant: quantifer – (Noun)

- Numbers

Example

Bíonn suas le ***céad*** *cineál éagsúil aimsire ag meitéareolaithe* : ‘Meteorologists have up to **one hundred** different kinds of weather’.

quant(cineál, céad)



‘Meteorologists have up to **one hundred** different kinds of weather’

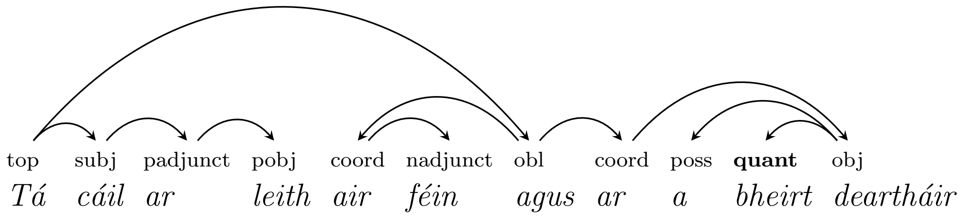
- Numerals for counting people

Example

Tá cáil ar leith air féin agus a ***bheirt*** *deartháir* : ‘He and his **two** brothers are famous’.

quant(deartháir, bheirt)

- Numbers - digits

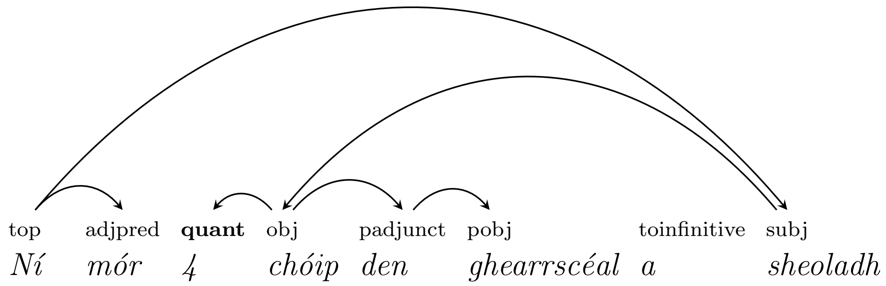


‘He and his two brothers are famous’

Example

Ní mór ***4*** *choip den ghearrscéal a sheoladh* : ‘**4** copies of the short-story have to be sent’.

quant(choip, 4)



‘4 copies of the short-story need to be sent’

- Non-numeric quantifiers:

Also *cuid* - ‘some of’, ‘part of’, ‘own’/ *neart* ‘plenty’

Example

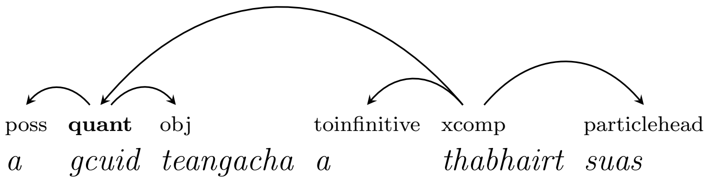
a ***gcuid*** *teangacha a thabhairt suss*: ‘to take up their own languages’

quant(teangacha, cuid)

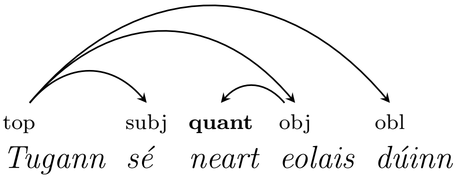
Example

Tugann sé ***neart*** *eolais dúinn*: ‘It gives us plenty of information’

quant(eolais, neart)



‘to take up their own languages’

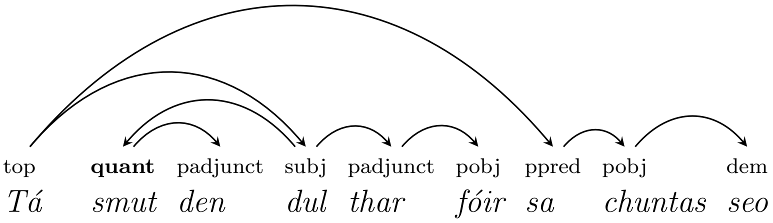


‘It gives us plenty of information’

Example

Tá ***smut*** *den dul thar fóir sa chuntas seo* : ‘There’s a little bit of exaggeration in this account’

quant(dul, smut)

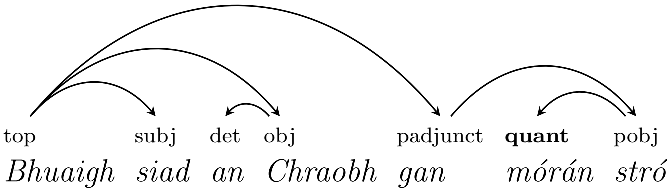


‘There’s a little bit of exaggeration in this account’

Example

Bhuaigh siad an Chraobh gan ***mórán*** *stró* : ‘They won the Cup without **much** effort’

quant(stró, mórán)

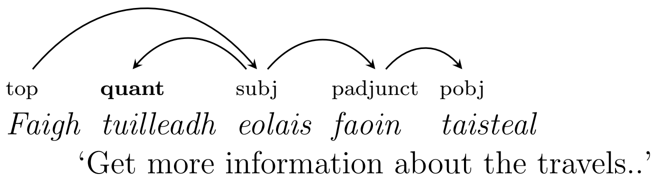


‘They won the Cup without much effort’

Example

Faigh ***tuilleadh*** *eolais faoin taisteal..* : 'Get **more** information about the travels..'

quant(eolais, tuilleadh)

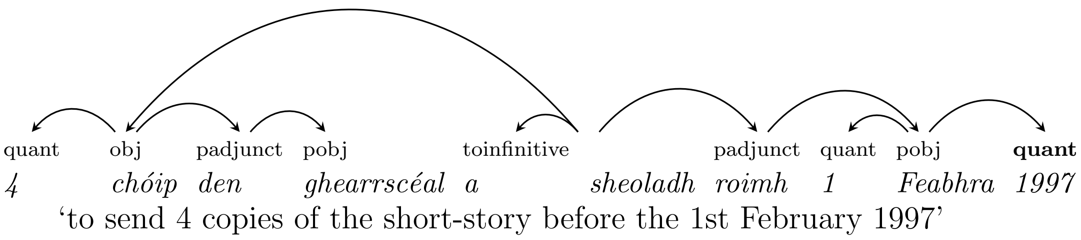


- Numbers - years

Example

4 choip den ghearrscéal a sheoladh roimh 1 Feabhra ***1997*** : 'to send 4 copies of the short-story before 1st February **1997**'.

quant(Feabhra, 1997)



- Numbers – adverbial use

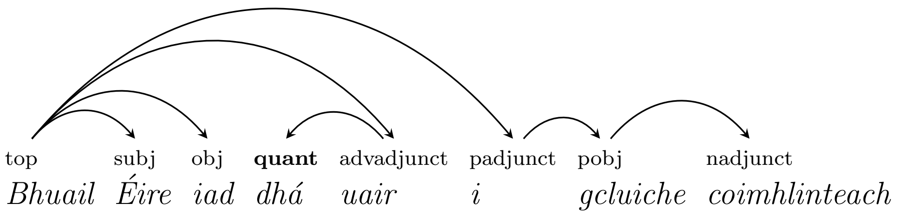
Example

Bhuail Éire iad dhá uair i gcluiche coimhlinteach : 'Ireland met them twice in competitive games'

quant(uair, dhá)

A.2.6 adjadjunct – (Noun)

- Adjectives



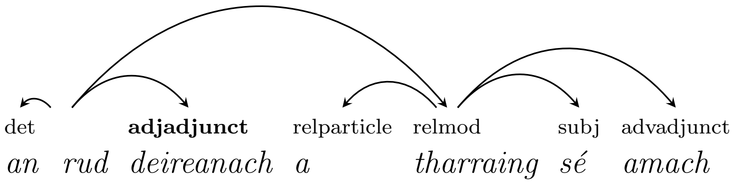
‘Ireland met them twice in competitive games’

Adjectives normally appear after the noun in Irish.

Example

Ba é an rud ***deireanach***: ‘It was the **last** thing’

adjadjunct(rud, deireanach)



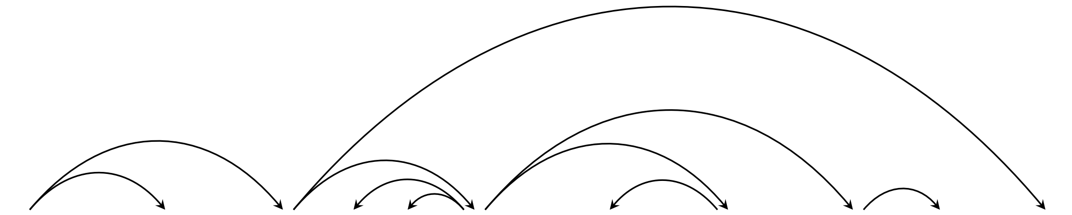
‘the last thing he pulled out’

- The same label (`adjadjunct`) is used for comparative/ superlative forms

Example

Measann go leor gurb é an píobaire is ***mó*** *in Albain é* ; ‘Many believe that he is the **biggest** piper in Scotland’

adjadjunct(píobaire, mó)



‘Many believe that he is the biggest piper in Scotland’

A.2.7 nadjunct : nominal adjunct – (Noun)

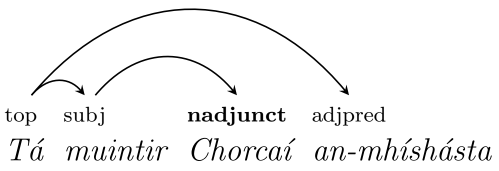
- Noun compounds

In Irish the second noun is in the genitive case, and thus treated as a noun modifying a noun.

Example

Tá muintir ***Chorcaí*** *an-mhíshásta*: '**Cork** people (people of Cork) are very dissatisfied'

nadjunct(muintir, Chorcaí)

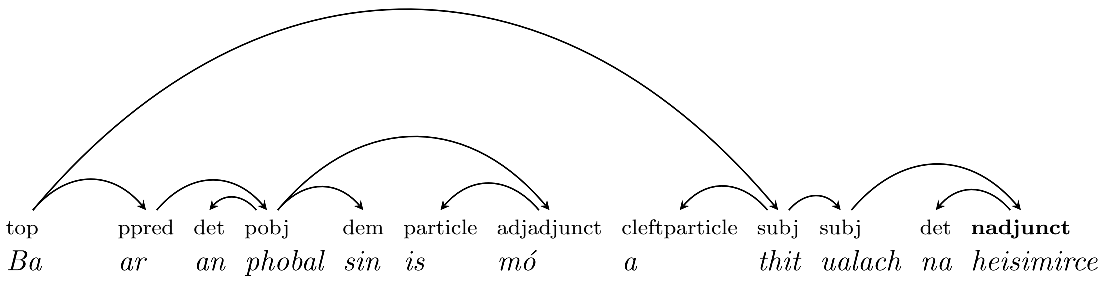


'Cork people are very unhappy'

Example

Ba ar an phobal sin is mó a thit ualach na ***heisimirce***: ' The burden of **emigration** hit that community the most'

nadjunct(ualach, heisimirce)



'It was mostly on that community that the burden of emigration hit'

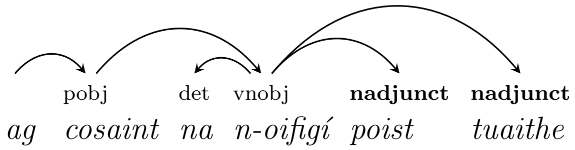
- String of nouns modifying one noun

Example

ag cosaint na n-oifigí ***poist tuaithe*** : ‘protecting the rural post offices’

(lit. the offices of the post of the country)

nadjunct(n-oifigí, poist), nadjunct(n-oifigí, tuaithe)



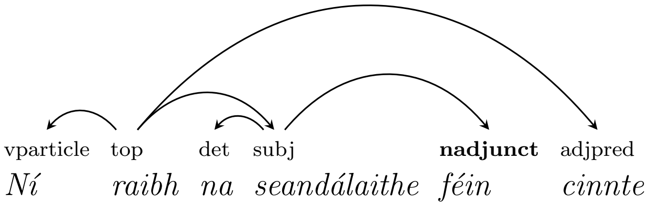
‘protecting the rural post offices’

- Reflexives

Example

Ní raibh na seandálaithe ***féin*** *cinnte* : ‘The archaeologists **themselves** were not sure’

nadjunct(seandálaithe, féin)



‘The archaeologists themselves were not sure’

- Names/ Titles

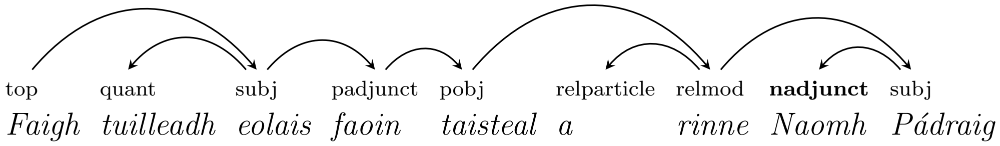
Example

Faigh tuilleadh eolais faoin taisteal a rinne ***Naomh*** *Pádraig* : ‘Get more information about the travels of **St** Patrick’

nadjunct(Pádraig, Naomh)

Example

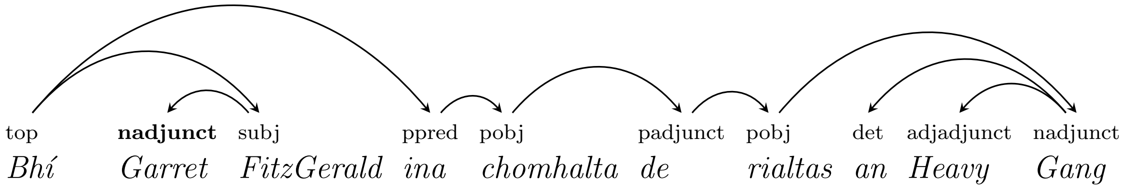
Bhí ***Garret*** *Fitzgerald ina chomhalta de rialtas an Heavy Gang* : ‘**Garret**



‘Get more information about the travels of St Patrick’

Fitzgerald was a member of the Heavy Gang administration’

nadjunct(Fitzgerald, Garret)



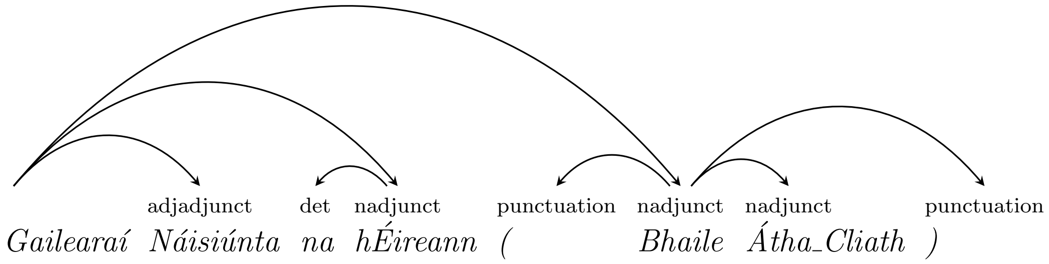
‘Garret Fitzgerald was a member of the Heavy Gang administration’

- Lines of addresses.

Example

*Gailearaí Náisiúnta na hÉireann (**Bhaile** Átha Cliath)* : ‘The National Gallery of Ireland (**Dubin**)’

nadjunct(Gailearaí, Bhaile)



‘The National Gallery of Ireland (Dubin)’

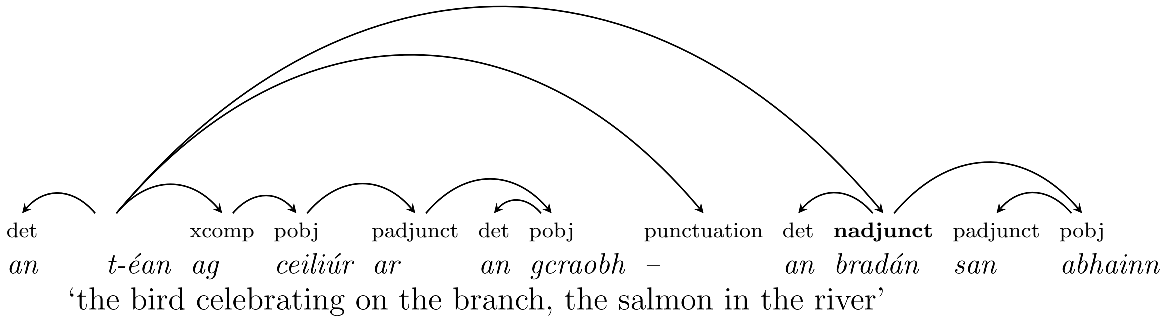
- Linking stanzas of poetry.

`nadjunct` can be used to link stanza clauses if they are noun phrases without clear coordination.

Example

an t-éan ag ceiliúradh ar an gcraobh – an bradán san abhainn ‘the bird celebrating on the branch, the **salmon** in the river’

adjunct(t-éan, bradán)



A.2.8 particle – (Noun)

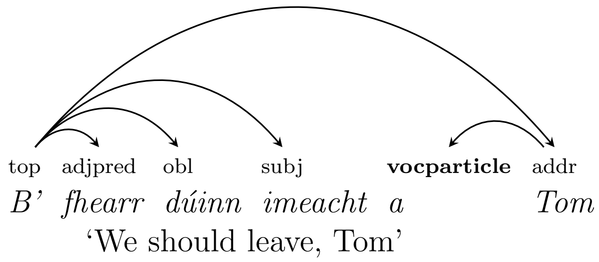
- Vocative particle ‘a’

  – used when addressing someone.

Example

B’fhearr dúinn imeacht a Tom: ‘We should leave, Tom’

vocparticle(Tom, a)



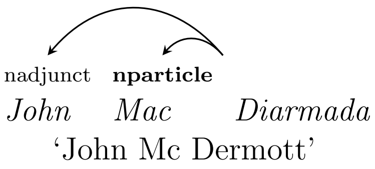
A.2.9 nparticle – (Noun)

- Surnames particles - *Mac, Ó, De,* etc.

Example

John Mac Diarmada

nparticle(Diarmada, Mac)



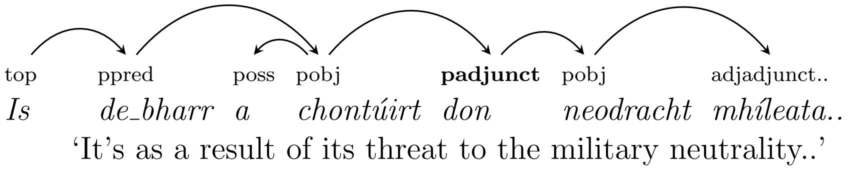
A.2.10 padjunct : prepositional adjunct – (Noun)

- Prepositions attached to Nouns

Example

Is de bharr a chontúirt ***don*** *neodracht mhíleata..* : 'It's as a result of its threat **to the** military neutrality..'

padjunct(chontúirt, don)



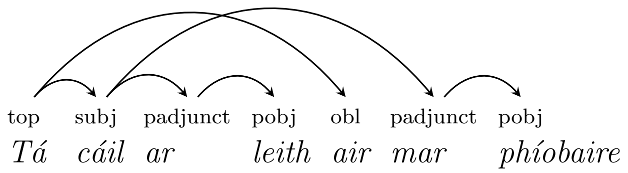
- 'mar' meaning 'as'.[7]

Example

Tá cáil ar leith air féin ***mar*** *phíobaire* : 'He is particularly known **as** a piper'

padjunct(cáil, mar)

[7] Note this creates crossing dependencies


top subj padjunct pobj obl padjunct pobj
Tá cáil ar leith air mar phíobaire

‘He is particularly known as a piper’

A.2.11 aug : augment pronoun – (Noun)

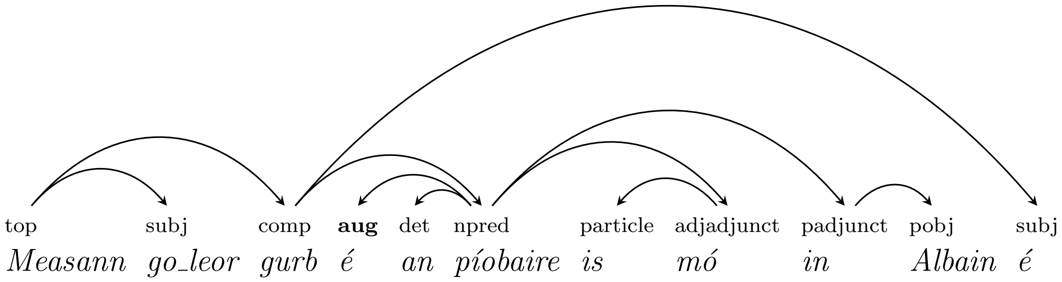
- Augment pronoun

Augment pronouns are typically used in copular constructions, always attached to (augmenting) and agreeing in person and number with the following noun.

Example

gurb ***é*** *an píobaire is mó in Albain é* : ‘that he was the biggest piper in Scotland’

aug(píobaire, é)

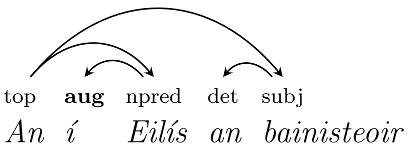

top subj comp **aug** det npred particle adjadjunct padjunct pobj subj
Measann go_leor gurb é an píobaire is mó in Albain é

‘Many believe that he is the biggest piper in Scotland’

Example

An ***í*** *Eilís an bainisteoir?* ; ‘Is the manager Eilís?’

aug(Eilís, í)



‘Is the manager Eilís?’

A.2.12 relmod : relative modifier – (Noun)

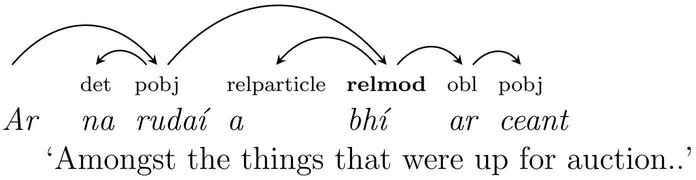
A verb that modifies a noun (head of a relative phrase)

- Subject (direct) relative modifier

  Example

  *Ar na rudaí a **bhí** ar ceant..* : 'Amongst the things that were up for auction..'

  relmod(rudaí, bhí)

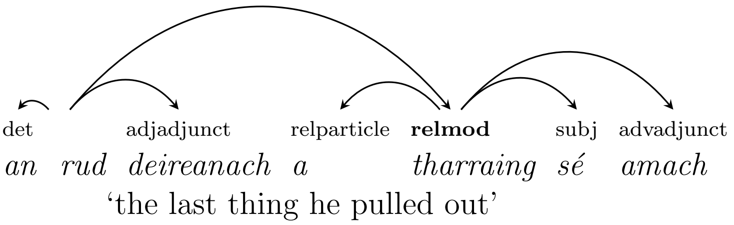


- Object (direct) relative modifier

  Example

  *..an rud deireanach a **tharraing** sé amach* : '..the last thing he **pulled** out'

  relmod(rud, tharraing)

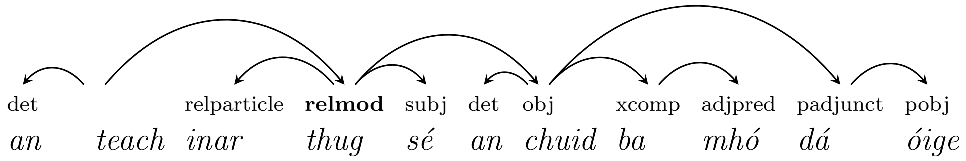


- Indirect relative modifier

  Example

  *an teach inar **thug** sé an chuid ba mhó dá óige* : 'The house in which he **spent** most of his youth'

  relmod(teach, thug)



‘The house in which he spent most of his youth’

A.2.13 app : nouns in apposition – (Noun)

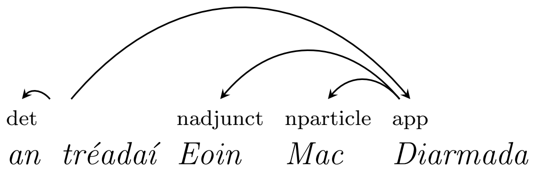
- Noun in apposition.

  – – Linking of two noun phrases that relate to each other. They may or may not be separated by commas.

  Example

  an tréadaí Eoin Mac ***Diarmada*** : ‘the pastor Eoin Mac Diarmada’

  app(tréadaí, Diarmada)



‘the pastor Eoin Mac Dermott’

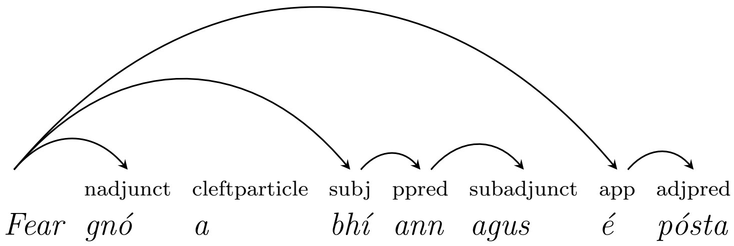
- Referential pronouns

  These are pronouns that do not modify a verb, but instead refer to a previous noun.

  Example

  Fear gnó a bhí ann, agus ***é*** *pósta* : ‘He was a business man, and **he** married’

  app(fear, é)



‘He was a business man, and he married”

A.2.14 obl: oblique – (Noun)

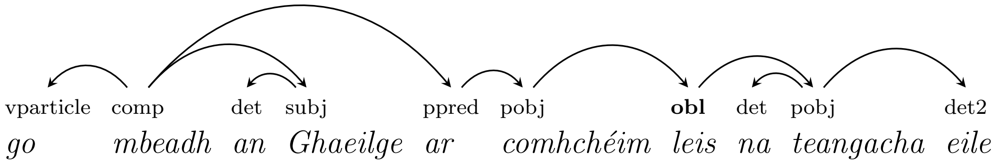
- Prepositions and prepositional pronouns

  – Prepositions that are closely linked to the noun (not the verb)

Example

*go mbeadh an Ghaeilge ar comhchéim **leis** na teangacha eile*: ‘that Irish would be on the same level **with** other languages’

obl(comhchéim, leis)

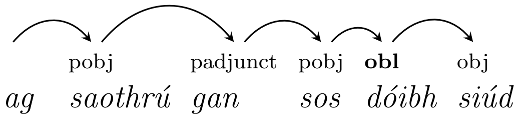


‘that Irish would be on the same level as other languages’

Example

*ag saothrú gan sos **dóibh** siúd* : ‘working without a break **for themselves**’

obl(sos, dóibh)



‘working without a break for themselves’

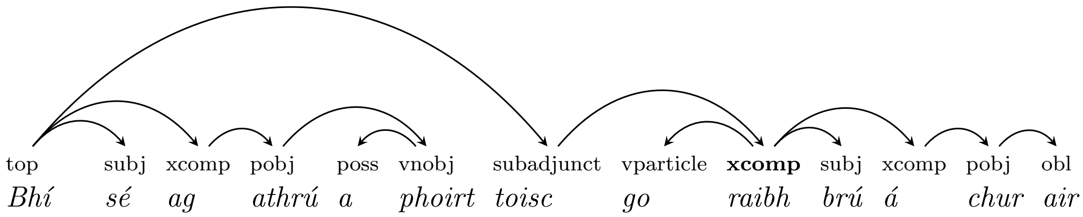
A.2.15 xcomp: open complement – (Noun)

- Nouns that play the role of subordinate conjunctions

Example

Bhí sé ag athrú a phoirt *toisc go* ***raibh*** *brú á chur air*: 'He was changing his tune since pressure was being put on him'

xcomp(toisc, raibh)



'He was changing his tune since there was pressure on him'

A.3 Preposition Dependents

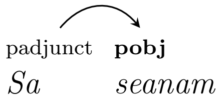
A.3.1 pobj: object of a preposition (Preposition)

- Head noun (object) of a prepositional phrases

Example

Sa ***seanam*** : 'In the **olden days**'

pobj(sa, seanam)

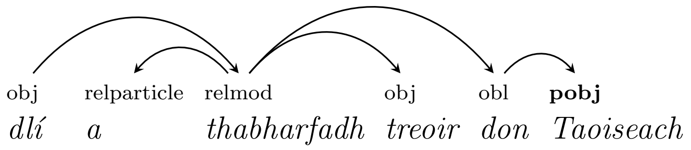


'In the olden days'

Example

dlí a thabharfadh treoir don ***Taoiseach*** : 'a law that would give guidance to the **Taoiseach**".

pobj(don, Taoiseach)



‘a law that would give guidance to the Taoiseach’

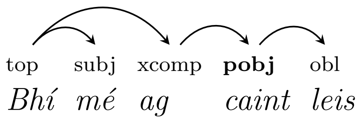
- Object in a progressive aspectual phrase

  - Progressives are formed in Irish using the preposition ‘ag’ followed by a verbal noun. The verbal noun is the object (`pobj`) of the preposition *ag.*

EXAMPLE

Bhí mé ag caint leis : ‘I was talking to him’ (lit. (at) **talking** with him)

pobj(ag, caint)



‘I was talking to him’

- Infinitival phrases[8]

  - The head of the infinitival phrase is an infinitive verb/verbal noun, which attaches to the preposition of the matrix clause.

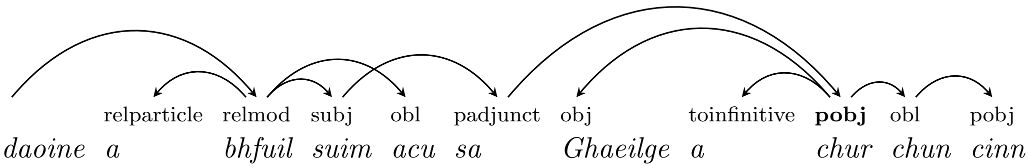
EXAMPLE

daoine a bhfuil suim acu sa [Ghaeilge a ***chur*** *chun cinn]*: ‘ people who have an interest in **promoting** Irish’

pobj(sa, chur) [9]

[8] Note that this can cause crossing dependencies

[9] see section A.1.7 for details on how to use chun_go



‘People who have an interest in promoting Irish’

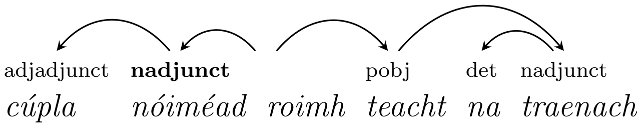
A.3.2 nadjunct: noun modifying a preposition (Preposition)

- nominal modifier (preceding the preposition)

Example

*cúpla **nóiméad** roimh teacht na traenach* : ‘a couple of **minutes** before the train’s arrival’

nadjunct(roimh, nóiméad)



‘a couple of minutes before the train’s arrival’

A.3.3 padjunct (Preposition)

- Prepositional cluster

  - A cluster of prepositions where the second preposition is attached to the first.

Example

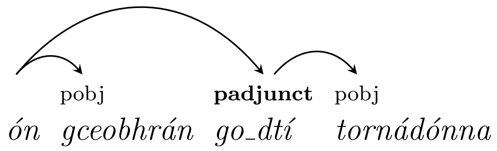
: *ón gceobhrán **go_dtí** tornádónna* : ‘from fog **to** tornadoes’

padjunct(ón, go_dtí)

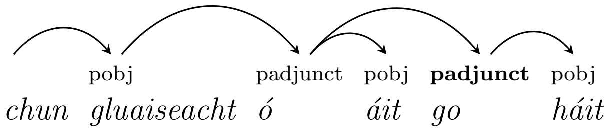
Example

*chun gluaiseacht ó áit **go** háit* : ‘to move from place **to** place’

padjunct(ó, go)



‘from fog to tornadoes’



‘to move from place to place’

A.4 Verbal Noun Dependents

A.4.1 vnobj : objects of verbal noun – (Verbal Noun)

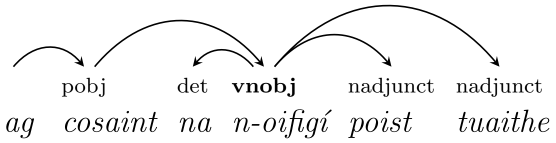
- Objects of verbal nouns.

  – These are objects of progressive verbs. They differ slightly from regular verbal objects because they are in the genitive case and seem to modify the verbal noun.

Example

ag cosaint na ***n-oifigí*** : ‘protecting the offices’ (lit. ‘at the protection of the offices’)

vnobj(cosaint, n-oifigí)



‘protecting the rural post offices’

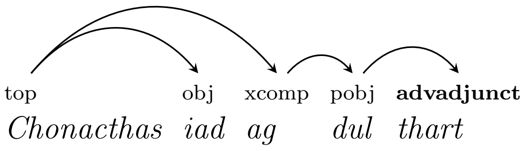
A.4.2 advadjunct : adverbial adjunct – (Verbal Noun)

- Adverbs[10]

Example

chonacthas iad ag dul ***thart*** : 'I saw them going **by**'

advadjunct(dul, thart)



'I saw them going by'

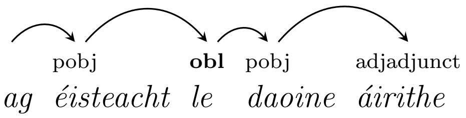
A.4.3 obl : oblique preposition – (Verbal Noun)

- Oblique
  - More closely attached to the verbal noun than regular preposition attachment
  - Can appear like collocations (suing for, dumped on, filled with, building on, set in)

Example

ag éisteacht ***le*** *daoine áirithe* : 'listening **to** certain people'

obl(éisteacht, le)



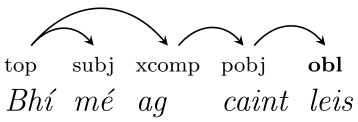
'listening to certain people'

[10] While the POS for these types of adverbs can sometimes be 'Adj', they should still be labelled as advadjunct.

Example

Bhí mé ag caint ***leis*** : 'I was speaking **to him**'

obl(caint, leis)



'I was talking to him'

A.4.4 xcomp : open complement – (Verbal Noun)

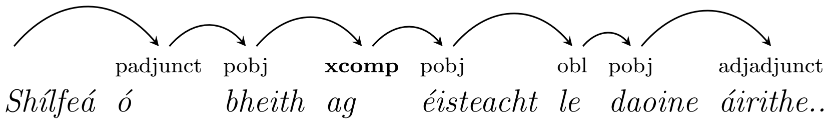
- Progressive aspectuals

`xcomp` is used to denote progressive aspectual phrases, with the preposition *ag* as the head.

Example

Shílfeá ó bheith ***ag*** *éisteacht le daoine áirithe..* : 'You would think from (to be) listening to certain people..'

xcomp(bheith, ag)



'You would think from listening to certain people..'

A.4.5 toinfinitive : infinitive marker – (Verbal Noun)

- Infinitive verb marker - *a*

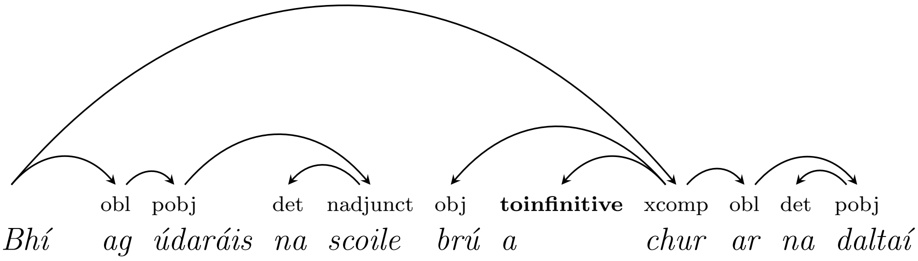
  – marks the verbal noun that immediately follows as infinitive.

Example

Bhí ag údaráis na scoile brú ***a*** *chur ar na daltaí* : 'The school authorities had

to put pressure on the students’

toinfinitive(chur, a)

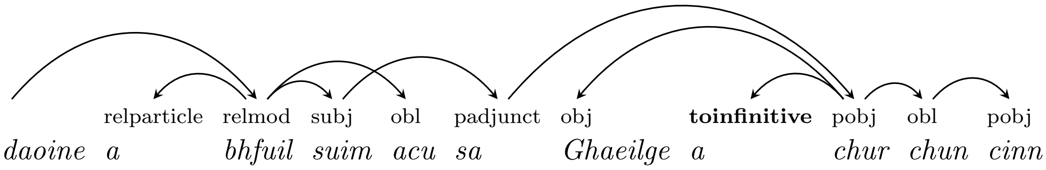


‘The school authorities had to put pressure on the students’

Example

daoine a bhfuil suim acu sa Ghaeilge ***a*** *chur chun cinn*: ‘people who have an interest in promoting Irish’

toinfinitive(chur, a)



‘People who have an interest in promoting Irish’

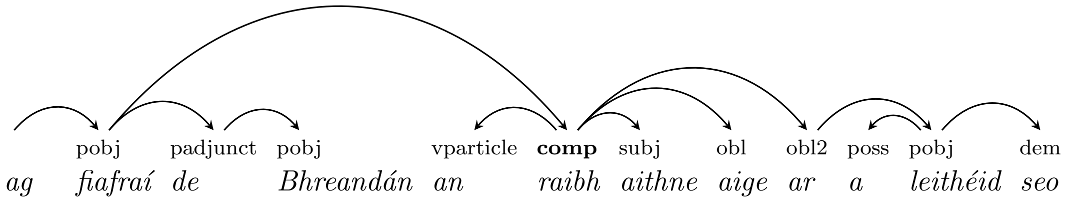
A.4.6 comp : closed complement – (Verbal Noun)

- complement phrases that have subjects

Example

ag fiafraí de Bhreandán an ***raibh*** *aithne aige ar a leithéid seo nó siúd* : ‘asking Brendan if he **knew** this or that’

comp(fiafraí, raibh)


pobj padjunct pobj vparticle **comp** subj obl obl2 poss pobj dem
ag fiafraí de Bhreandán an raibh aithne aige ar a leithéid seo

‘asking Brendan if he knew this or that’

A.5 Adjective Dependents

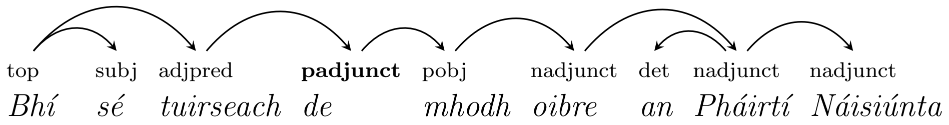
A.5.1 padjunct : prepositional adjunct – (Adjective)

- Dependents of adjectival predicates

Example

go raibh sé tuirseach ***de*** *mhodh oibre an Pháirtí Náisiúnta*: ‘that he was tired **of** the National Party’s line of action’

padjunct(tuirseach, de)

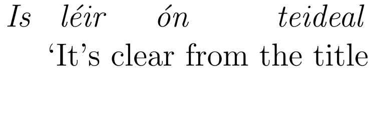

top subj adjpred **padjunct** pobj nadjunct det nadjunct nadjunct
Bhí sé tuirseach de mhodh oibre an Pháirtí Náisiúnta

‘He was tired of the National Party’s line of action ’

Example

Is léir ***ón*** *teideal go bhfuil draíocht i gceist* : ‘It’s clear **from** the title that there is magic involved’

padjunct(léir, ón)



‘It’s clear from the title’

A.5.2 particle – (Adjective)

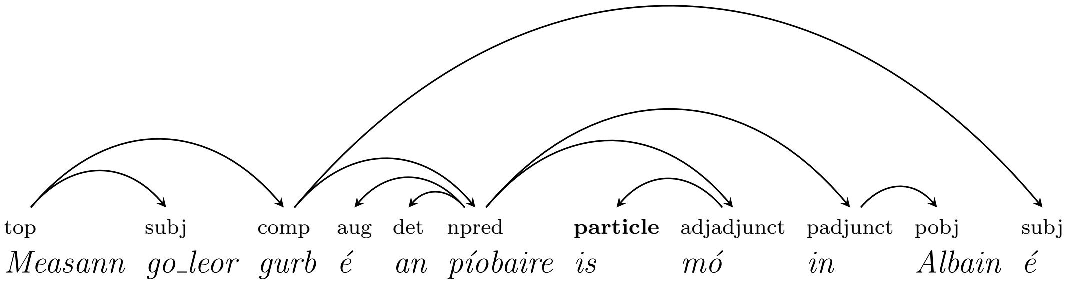
- Comparatives and superlatives

Particles used to indicate comparative *níos* or superlative *is* adjective forms.

Example

an píobaire ***is*** *mó in Albain* : 'the big**gest** piper in Scotland'

particle(mó, is)

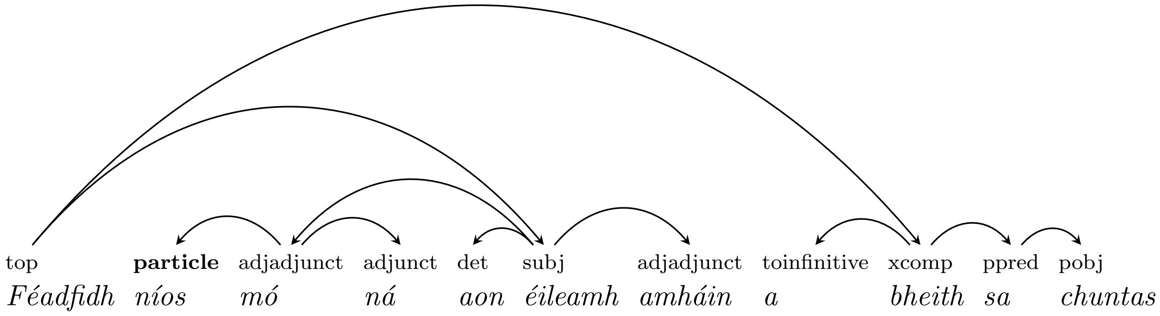


'Many believe that he is the biggest piper in Scotland'

Example

Féadfaidh ***níos mó*** *ná aon éileamh a bheith sa chuntas* : '**More** than one demand can be in the account'

particle(mó, níos)



'More than one demand can be in the account'

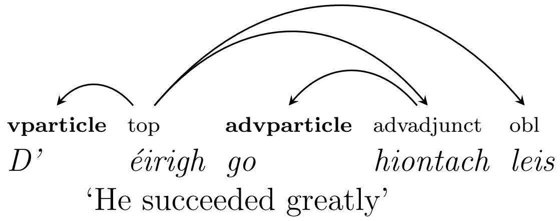
- Adverbial particle (no English equivalent)

- Despite having a `Adj` POS, when used with the particle *go*, an adjective takes the role of an adverb.

Example

D'eirigh go hiontach leis : 'he succeeded **greatly**'

advparticle(hiontach, go)



A.5.3 advadjunct: adverbial intensifier – (Adjective)

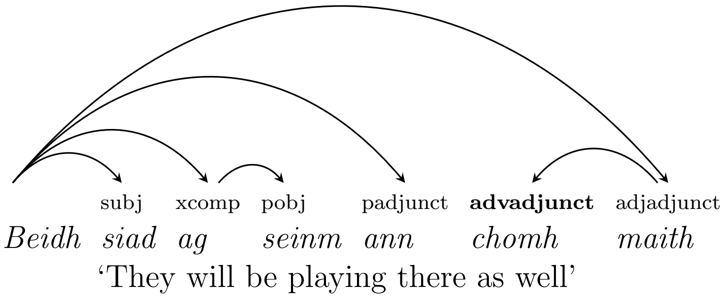
- Adjective Intensifier

  - Modifies an adjective to take on role of adverb.

Example

beidh siad ag seinm ann ***chomh*** *maith* : 'They will be playing there as well'

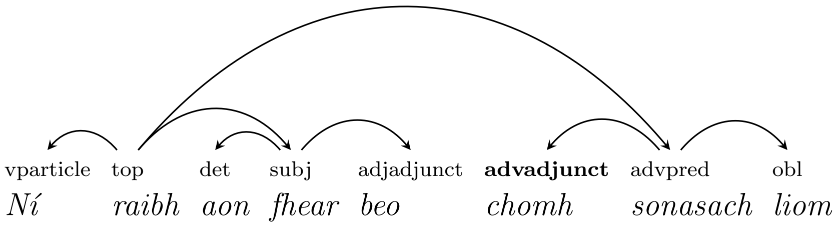
advadjunct(maith, chomh)



Example

Ní raibh aon fhear ***chomh*** *sonasach leis*: 'There was no man **as** happy as him’

advadjunct(sonasach, chomh)


vparticle top det subj adjadjunct **advadjunct** advpred obl
Ní raibh aon fhear beo chomh sonasach liom

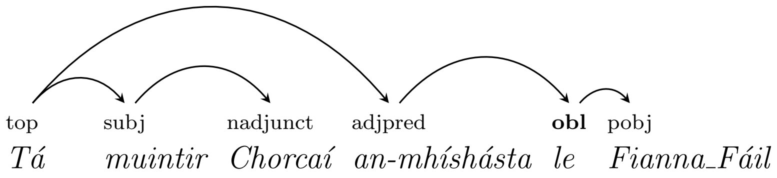
‘There was no man as happy as him’

A.5.4 obl: obliques – (Adjective)

Example

Tá muintir Chorcaí an-mhíshásta ***le*** *Fianna Fáil*: ‘Cork people are unhappy **with** Fianna Fáil’

obl(mhísháta, le)

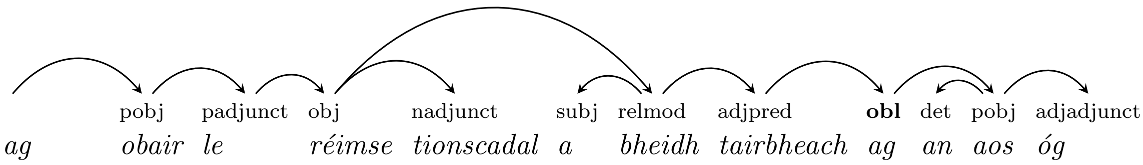

top subj nadjunct adjpred **obl** pobj
Tá muintir Chorcaí an-mhísháta le Fianna_Fáil

‘The people of Cork are very dissatisfied with Fianna Fáil’

Example

ag obair le réimse tionscadal a bheidh tairbheach ***ag*** *an aos óg* : ‘working with a range of projects that will be beneficial **for** the youth’

obl(tairbheach, ag)


pobj padjunct obj nadjunct subj relmod adjpred **obl** det pobj adjadjunct
ag obair le réimse tionscadal a bheidh tairbheach ag an aos óg

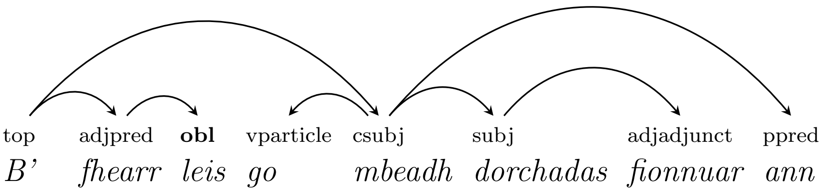
‘working with a range of projects that will be beneficial for the youth’

- equivalent to subject in English.

Example

B'fhearr ***leis*** *go mbeadh dorchadas fionnuar ann*: 'He would prefer if there was a cool darkness [11]'

obl(fhearr, leis)

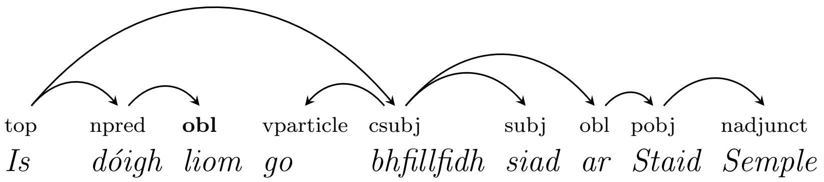


'He would prefer if there was a cool darkness'

Example

Is dóigh ***liom*** *go bhfillfidh siad ar Staid Semple* : 'I think that they will return to Semple Stadium' [12]

obl(dóigh, liom)



'I think that they will return to Semple Stadium'

A.6 Verbal Adjective Dependencies

A.6.1 obl_ag : oblique agent – (Verbal Adjective)

- agent of stative passives.

[11] lit. 'that there would be a cool darkness would be good **with him**'

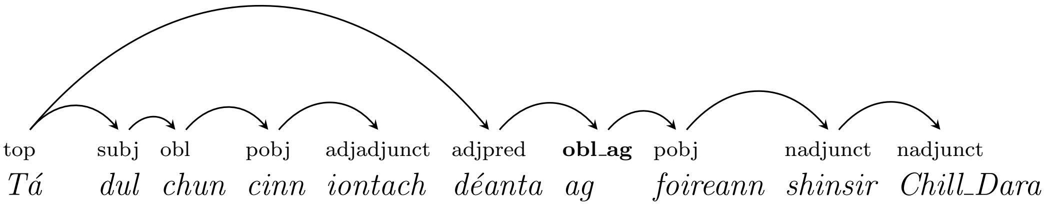
[12] lit. 'that they will return to Semple Stadium is likely **to me**'

– These translate to passive structures in English.

Example

Tá dul chun cinn iontach déanta ***ag*** *foireann shinsir Chill Dara*: 'Great progress has been made **by** Kildare Senior Team

obl_ag(déanta, ag)

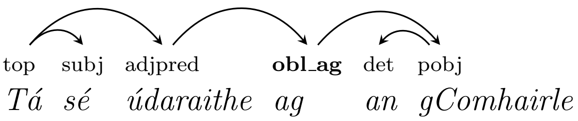


'Great progress has been made by Kildare Senior Team'

Example

Tá sé údaraithe ***ag*** *an gComhairle* : 'It is authorised **by** the Council'

obl_ag(údaraithe, ag)



'It is authorised by the Council'

A.6.2 obl : oblique argument – (Verbal Adjective)

- oblique prepositions or pronominal prepositions[13]

Example

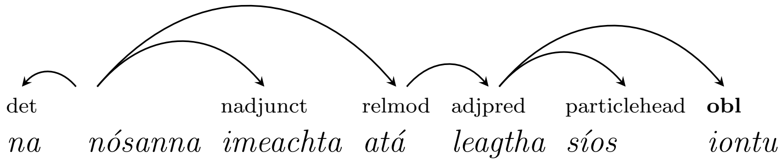
na nósanna imeachta atá leagtha síos ***iontu*** : 'the types of activities that are laid out **in them**'

obl(leagtha, iontu)

Example

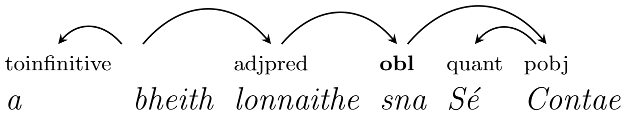
a bheith lonnaithe ***sna*** *Sé Contae* : 'to be based in the 6 Counties'

[13] Differs from obl_ag. See Section A.6.1



'the types of activities that are laid out in them'

obl(lonnaithe, sna)



'to be based in the 6 Counties'

A.7 Adverb Dependents

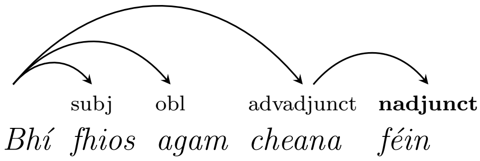
A.7.1 nadjunct : nominal adjunct – (Adverb)

- Reflexives

Example

Bhí fhios agam ***cheana*** *féin* : 'I knew already'

nadjunct(cheana, féin)



'I knew already'

A.8 Subordinate Conjunction Dependencies

A.8.1 comp : closed complement – (Subordinate Conjunction)

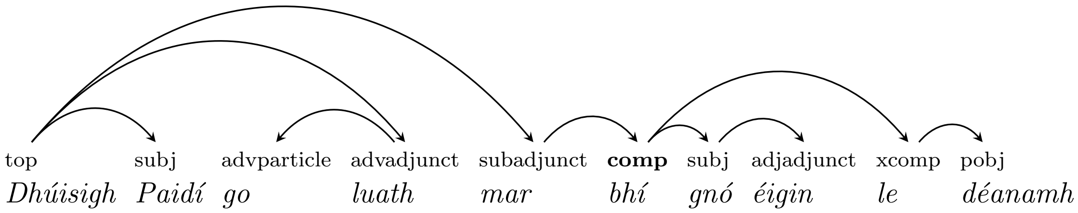
- complement phrases

- Normally subordinate conjunctions have complement phrases as dependents. These are usually full sentences. See section A.1.7 for more examples of the types of complement phrases that are dependents of subordinates.

Example

*Dhúisigh Paidí go luath mar **bhí** gnó éigin le déanamh*: 'Paidí woke early because there **was** some work to do'

comp(mar, bhí), *subadjunct(Dhúisigh, mar)*



'Paidí woke early because there was some work to do'

A.9 Copula Dependents

The copula is used in many constructions - identity, classificatory, ownership, comparative and idiomatic. Usually the structure is COP PRED SUBJ.

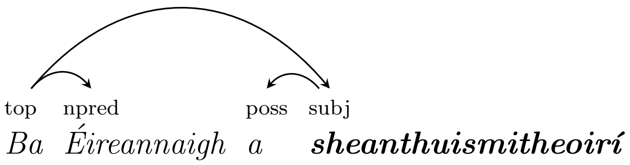
A.9.1 subj : subject – (Copula)

- Noun phrase subject

Example

*Ba Éireannaigh a **sheanthuismitheoirí***: 'His grandparents were Irish (people)'

subj(Ba, sheanthuismitheoirí)



‘His grandparents were Irish’

- Clefts

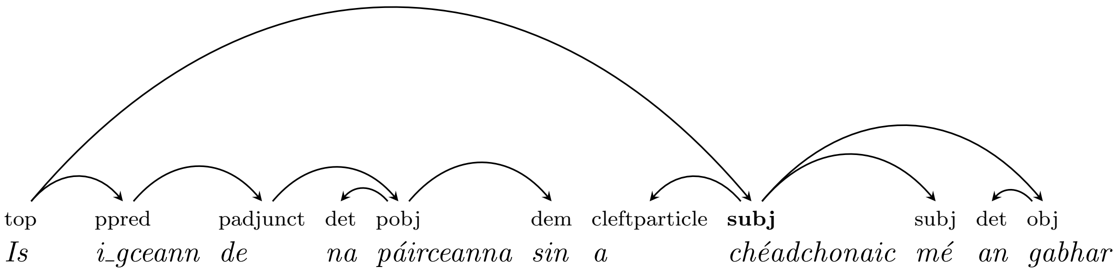
  The copula is also used for clefting (fronting). With clefting, the subject is the entire relative clause (The `subj` label is used to attach the head (the verb) to the copula.

  Example

  *Is [i_gceann de na páirceanna sin] [a **chéadchonaic** mé ghabhar]*[14]:

  lit. ‘It’s in one of those fields that I first saw the goat’ (‘I first saw the goat in one of those fields’)

  subj(Is, chéadchonaic)



‘It was in one of those fields that I first saw the goat’

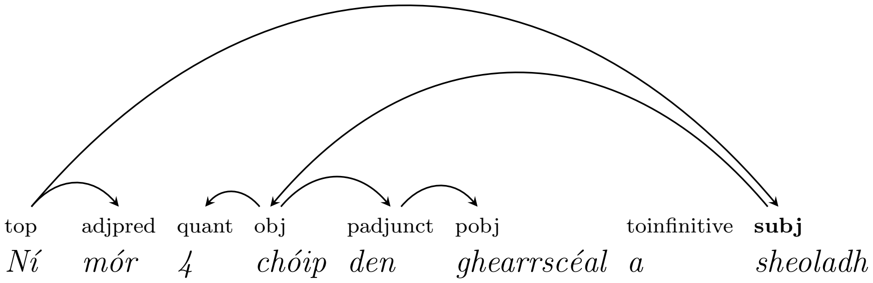
- In the following example, the infinitival phrase ‘to send 4 copies of the short-story’ is the subject.

  Example

  *Ní [mór] [4 choip den ghearrscéal a **sheoladh**]*: ‘4 copies of the short-story need to be sent’

  subj(Ní, sheoladh)

[14] The square brackets [] delimit the scope of the subject and predicate phrases.


top adjpred quant obj padjunct pobj toinfinitive **subj**
Ní mór 4 chóip den ghearrscéal a sheoladh

‘4 copies of the short-story need to be sent’

A.9.2 csubj : clausal subject – (Copula)

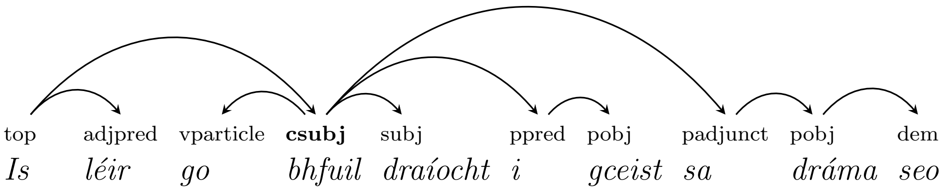
- Full clausal subject

  - The clause is in subject position of a copula construction

Example

*Is léir ón teideal [go **bhfuil** draíocht i gceist]* : ‘It’s clear **from** the title that there is magic involved’

csubj(Is, bhfuil)

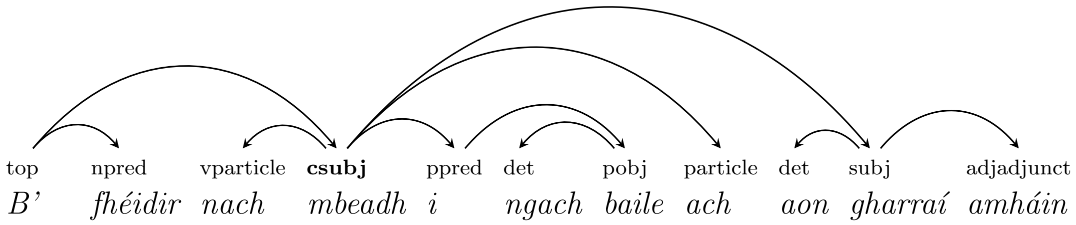

top adjpred vparticle **csubj** subj ppred pobj padjunct pobj dem
Is léir go bhfuil draíocht i gceist sa dráma seo

‘It is clear that there is magic in this play’

Example

*B’fhéidir nach **mbeadh** i ngach baile ach aon gharraí amháin* : ‘Maybe there would only be one garden in every town’

csubj(B’, mbeadh)



‘Maybe there would only be one garden in every town’

A.9.3 pred : predicates – (Copula)

In copular constructions, pred is usually used to label the predicate. There is a set of sublabels for `pred`: nominal predicate `npred`, adjectival predicate `adjpred`, prepositional predicate `ppred`, adverbial predicate `advpred`.

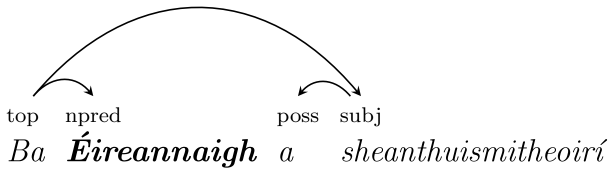
The order of copular constructions is usually: COP, PRED, SUBJ. This also applies to cleft constructions.

- npred - nominal predicates

Example

Ba ***Éireannaigh*** *a sheanthuismitheoirí*: ‘His grandparents were Irish (people)’

npred(Ba, Éireannaigh)



‘His grandparents were Irish’

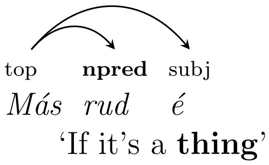
Example

Más ***rud*** *é* : ‘If it’s a **thing**’

npred(Más, rud)

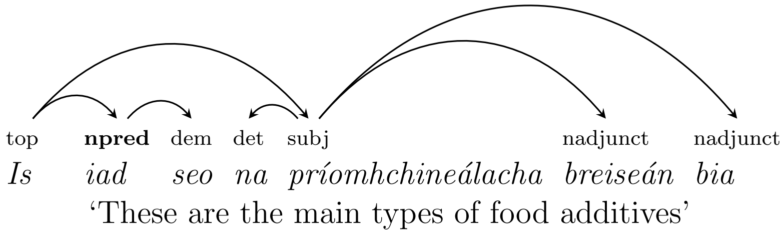
Example

Is ***iad*** *seo na príomhchineálacha breiseán bia* : ‘**These** are the main types of



food additives'

npred(Is, iad)

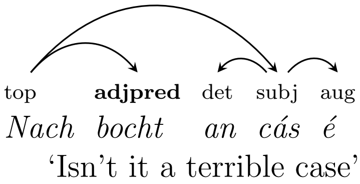

top
npred
dem
det
subj
nadjunct
nadjunct
Is iad seo na príomhchineálacha breiseán bia
'These are the main types of food additives'

- adjpred - adjectival predicates

Example

Nach bocht an cás é: 'Isn't it a terrible case'

adjpred(nach, bocht))



Example

Is ***dóigh*** *liom go bhfuil..* : 'I think that..' [15]

adjpred(Is, dóigh), obl(dóigh, liom)

- advpred - adverbial predicates

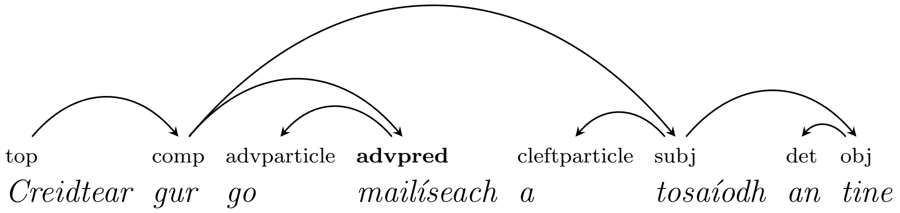
Example

Creidtear gur go ***mailíseach*** *a tosaíodh an tine* : 'It's believed that it was

[15] lit. It is **likely** to me that..

maliciously that the fire was started’

advpred(gur, mailíseach), *particle(mailíseach, go)*



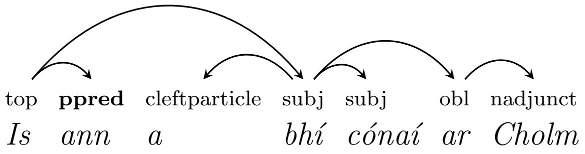
‘It is believed that the fire was started maliciously’

- ppred - prepositional predicates

Example

Is ***ann*** *a bhí cónaí ar Cholm* : ‘It’s **there** that Colm lived’

ppred(is, ann)

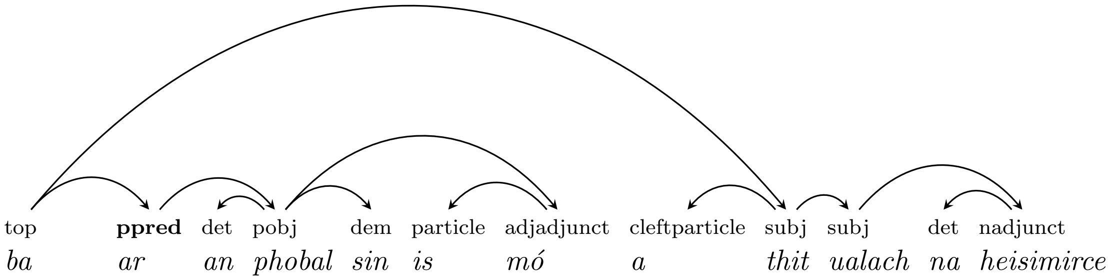


‘It’s there that Colm lived’

Example

*Ba [****ar*** *an phobal sin is mó] a thit ualach na heisimirce* : ‘It was [mostly **on** that community] that the burden of emigration has fallen”

ppred(is, ar)



‘It was mostly on that community that the burden of emigration hit’

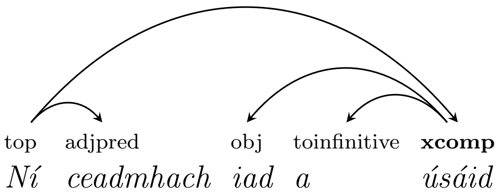
A.9.4 xcomp : open complement – (Copula)

- Infinitival phrases

Example

Ní ceadmhach iad a ***úsáid*** : ‘It is not permissable **to use** them’

xcomp(nach, úsáid)



‘It is not permissable to use them’

A.10 Quant Dependents

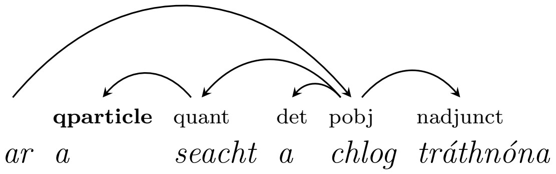
A.10.1 qparticle: quantifier particle (Quant)

- number particles (time)

Example

ar ***a*** *seacht a chlog tráthnóna* : ‘at 7 o’clock in the afternoon’

qparticle(dó, a)



‘at 7 o’clock in the afternoon’

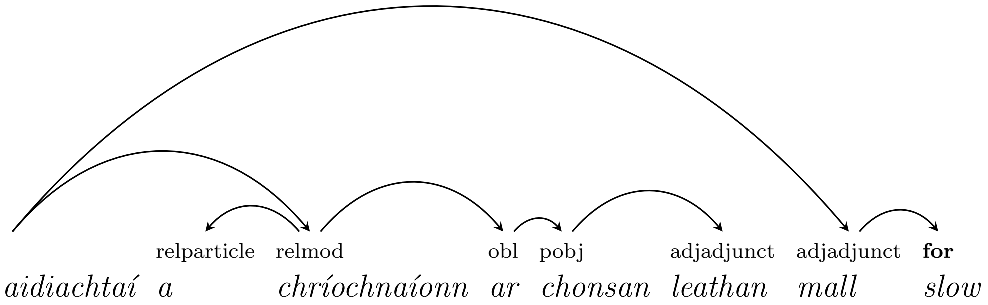
A.11 Foreign Words

- If there is only one foreign word within an Irish sentence, and it is simply a translation of the previous word or string, it should be labelled as `for`.

Example

aidiachtaí a chríochnaíonn ar chonsan leathan mall ***slow*** : 'adjectives that end in a broad consonant mall **slow**'

for(mall, slow)



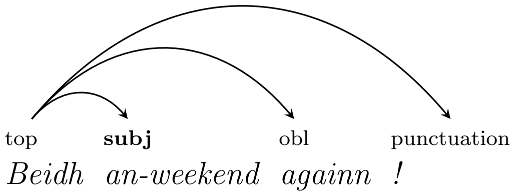
'Adjectives that end in a broad consonant mall slow'

- If there is only one foreign word within an Irish sentence, and it fits into the syntactic structure of the sentence, parse it as normal.

Example

Beidh ***an-weekend*** *againn!* : 'We will have a **great weekend**!'

for(Beidh, an-weekend)



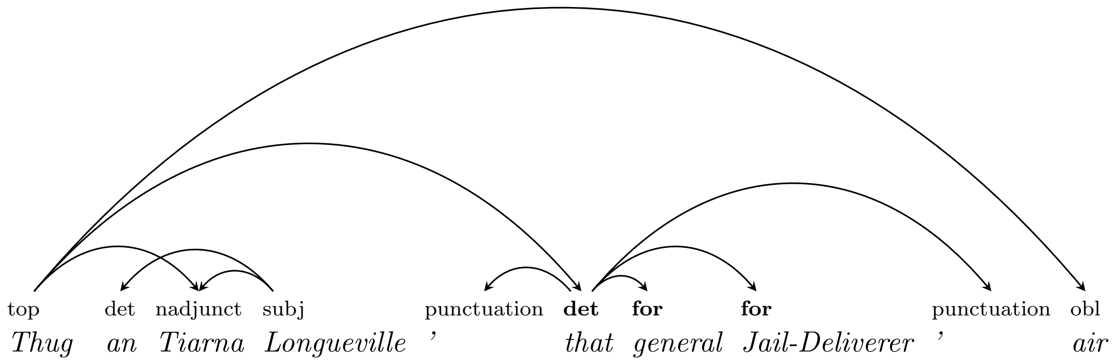
'We will have a great weekend!'

- If there is a string of foreign words, parse the first item as normal, but label the rest of the words as `for` dependents of that first item.

Example

Thug an Tiarna Longueville "that general Jail-Deliverer" air : 'Lord Longueville called him “that general Jail-Deliverer” ’

det(Thug, that), for(that, general), for(that, Jail-Deliverer)



’Lord Longueville called him “that general Jail-Deliverer”’

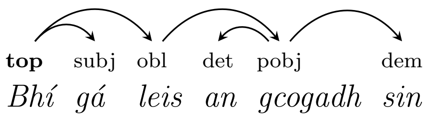
A.12 Sentence Root/ Head

- Normally the head of the sentence is the main verb.

  Example

  Bhí gá leis an gcogadh sin : ‘That war was necessary’

  top(root, Bhí)



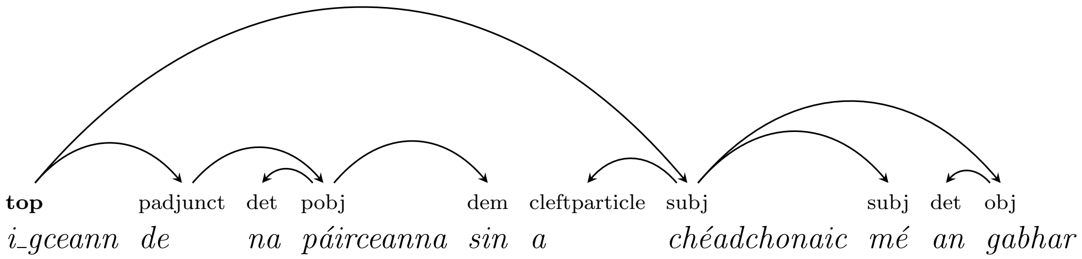
‘That war was necessary’

- But in the case of pseudo clefts, where the copula is dropped, the head of the clefted (fronted) part is promoted to head position. This can happen for nominal, prepositional, adjectival and adverbial clefting.

  Example

  [Is] ***I_gceann*** *de na páirceanna a chéadchonaic mé an gabhar* : ‘ **In** the park [is] where I first saw the goat’

  top(ROOT, I_gceann)



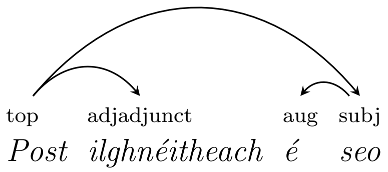
‘It was in one of those fields that I first saw the goat’

- The same applies to regular copular constructions (e.g. identity construction below) where the copula is dropped.

  Example

  : *[Is] Post ilghnéitheach é seo* : ‘[This is] a varied job’

  top(ROOT, Post)



‘This is a varied job’

A.13 Coordination

Coordinates are *agus* ‘and’, *nó* ‘or’, *ná*, *&*

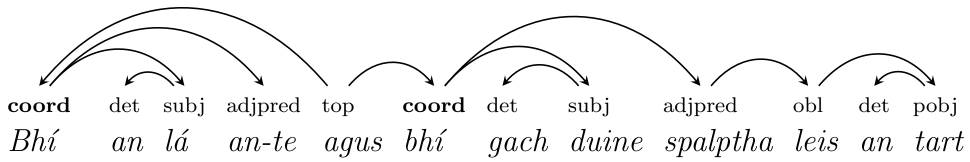
- We use LFG-inspired coordination where the coordinating conjunction is the head, and its coordinated phrases are dependents.

  Example

  Bhí an lá an-te agus bhí gach duine spalptha leis an tart ‘It was a hot day and everybody was parched with the thirst’:

  coord(agus, Bhí), coord(agus, bhí)

- This multiple use of ‘and’ is more common and acceptable in Irish than in English. In cases such as the following, where there is an uneven number of


coord det subj adjpred top coord det subj adjpred obl det pobj
Bhí an lá an-te agus bhí gach duine spalptha leis an tart

‘It was a hot day and everybody was parched with the thirst’

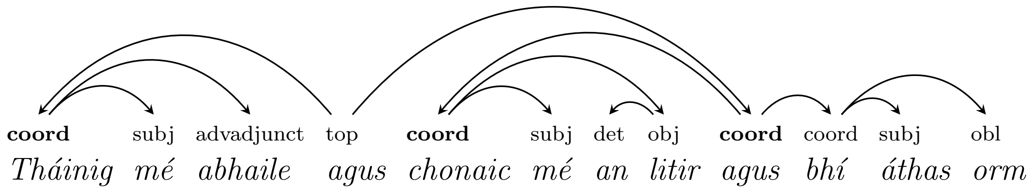
coordinates, we mark the final two as a cluster.

coord(agus$_1$, Tháinig), coord(agus$_1$, agus$_2$), coord(agus$_2$, lean), coord(agus$_2$, chuir)

Example

[Tháinig mé abhaile] agus$_1$ [chonaic mé an litir] agus$_2$ [bhí áthas orm] : ‘I came home and I saw the letter and I was happy’

coord(agus$_1$, Tháinig), coord(agus$_1$, agus$_2$), coord(agus$_2$, chonaic), coord(agus$_2$, bhí)


coord subj advadjunct top coord subj det obj coord coord subj obl
Tháinig mé abhaile agus chonaic mé an litir agus bhí áthas orm

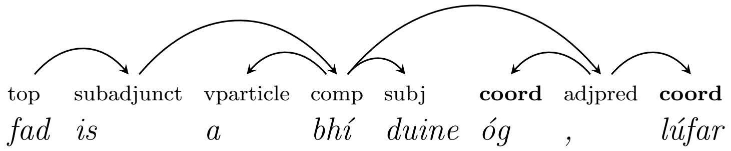
‘I came home and I saw the letter and I was happy’

- Punctuation can be the head of the coordination.

  Example

  fad is a bhí daoine óg, lúfar : ‘while people are young (and) agile’

  adjpred(bhí, ,), coord(, , óg), coord(, ,lúfar)

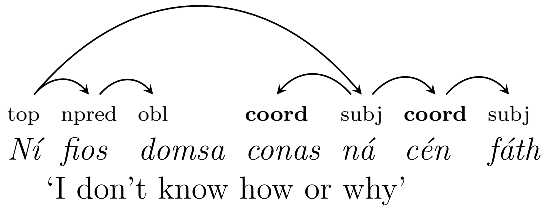

top subadjunct vparticle comp subj coord adjpred coord
fad is a bhí duine óg , lúfar

‘while people are young (and) agile’

- The particle *ná*, which has many other functions such as subordinating conjunction, can also mean ‘nor’.

Example

Ní fios domsa [conas] ***ná*** *[cén fáth]* : 'I don't know [how] **nor** [why]'


top npred obl **coord** subj **coord** subj
Ní fios domsa conas ná cén fáth
'I don't know how or why'

A.14 Punctuation

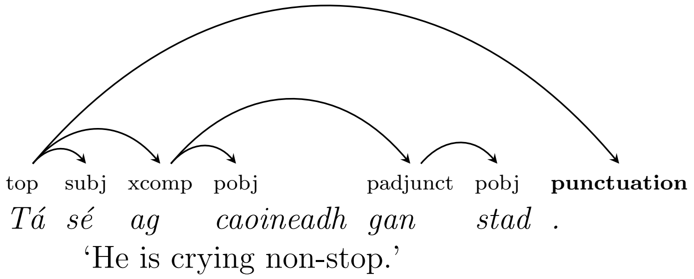
All punctuation is labelled `punctuation`.

- Initial and Final punctuation always attach to the root (top)

Example

Tá sé ag caoineadh gan stad ***.*** : 'He is crying non-stop.'

punctuation(Tá, .)

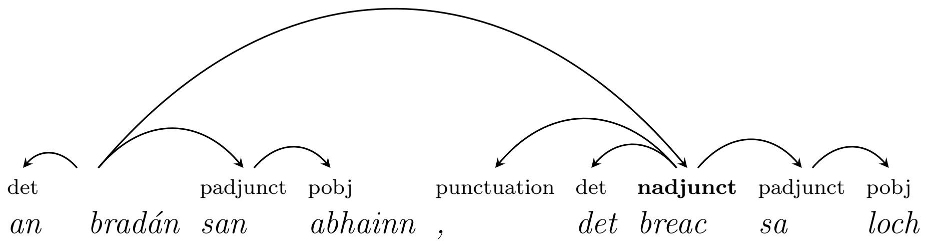

top subj xcomp pobj padjunct pobj **punctuation**
Tá sé ag caoineadh gan stad .
'He is crying non-stop.'

- Internal punctuation always attaches to the following head word.

Example

*an bradán san abhainn****,*** *an breac sa loch* : 'the salmon in the river, the trout in the lake'

punctuation(breac, ,)



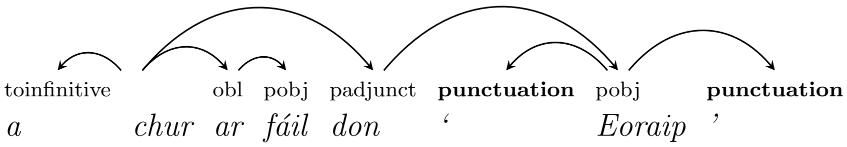
‘the salmon in the river, the trout in the lake’

- If a word or phrase is within quotes or brackets, both quotes or brackets are dependent on that word, or the head of the phrase.

Example

*a chur ar fáil don **‘Eoraip’*** : ‘to make available to **‘Europe**’’

punctuation(Eoraip, ‘), punctuation(Eoraip, ’)



‘to make available to **‘Europe**’’

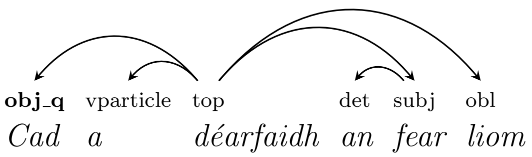
A.15 Questions

We regard the verb as the sentential head and mark the WH-element as a dependent of that verb, labelled as `subj_q`, `obj_q` or `advadjunct_q`.

Example

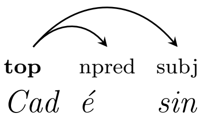
***Cad** a déarfaidh an fear liom* : ‘**What** will the man say to me?’

obj_q(déarfaidh, Cad)



‘What will the man say to me?’

- When there is no verb present, however, WH-elements such as *cad*, for example, should be treated as an interrogative copula.



‘What is that?’

A.16 MWEs - Multiword Expressions

We do not label multi-word-expressions in this release of the treebank as there has not been enough research carried out yet on idioms and multi-word expressions in Irish.[16]

[16] Katie Ní Loinsigh in Fiontar, DCU is doing a PhD in this area at present. Expected publication date is 2016.

Appendix B

Annotation Guidelines for Irish Twitter Part-of-Speech Tagging

In this appendix, we present the annotation guidelines for POS tagging Irish tweets. The style of language used in tweets is noisy and varies from standard canonical Irish text in a number of ways:

- phrases and sentences are limited to 140 characters
- tweets can contain typographical errors and ungrammatical structures
- diacritics are often omitted
- text can switch from Irish to English (code-switching)
- new 'text-speak' words are formed based on phonetic spelling
  (e.g. *7tain* (*seachtain*) 'week')
- words and phrases are often abbreviated
  (e.g. *mgl* (*maith go leor* 'fair enough'))
- twitter specific symbols are used
  (e.g. hashtags, at-mentions, retweet indicators)

For these reasons, we have developed a tweet-specific POS tagset (presented in Table B.1). We discuss tokenisation and POS annotation of Irish tweets below. In general, the tagging rules are similar to the assignment of the PAROLE Morphosyntactic Tagset (ITÉ, 2002) by the rule-based Irish POS-tagger (Uí Dhonnchadha and van Genabith, 2006). We discuss the cases that require specific explanation here.

Tag	**Description (PAROLE TAGS)**
N	common noun
^	proper noun
O	pronoun
VN	verbal noun
V	verb (incl. copula)
A	adjective
R	adverb
D	determiner
P	preposition, prepositional pronoun
T	particle
,	punctuation
&	conjunction
$	numeral, quantifier
!	interjection
G	foreign words, abbreviations, item
~	discourse marker
#	hashtag
#MWE	multi-word hashtag
@	at-mention
E	emoticon
U	URL/ email address/ XML

Table B.1: Irish Twitter part-of-speech tagset.

B.1 Tokenisation of Irish tweets

In general, the tweets are tokenised on white spaces. However, the following strings are split into separate tokens, and tagged individually:

- final punctuation (e.g. *oíche mhaith.* → *oíche mhaith .*)
- time (e.g. 15:15pm → 15:15 pm)

Punctuation is not split in the following contexts:

- the punctuation is part of a contraction (e.g. *Tá'n* (lit. *Tá an*))
- part of an email address (e.g. XYZ@gmail.com)
- an abbreviation (e.g. i.n. *iarnóin* 'afternoon')
- emoticon (e.g. *:0*)
- inflection such as an urú (e.g. *Lá na n-oibrithe* 'workers' day')
- compounds (e.g. *ró-ard*)

Multiword units (e.g. compound prepositions) are conjoined with an underscore and should be tagged with `P`.

- *go_dtí* 'to'
- *a_lán* 'a lot of/ many'
- *Ard_Mhaca* 'Armagh'
- *ar_ais* 'back'

If the tokeniser incorrectly splits a unit, it will be tagged as an unknown (`G`).

- *rphost/Twitter* (should be: *rphost / Twitter* 'email / Twitter')

B.2 POS tagging Irish Tweets

B.2.1 Prepostional pronouns/ pronominal prepositions

Most of the simple prepositions in Irish can inflect for pronominal objects. For example, *liom* 'with-me', *leo* 'with-them'. Despite these forms containing a nominal element, they are labelled as prepositions (`P`).

B.2.2 Hashtags

Hashtags are used to categorise a tweet. If hashtags are syntactically part of the sentence, mark them as per their correct syntactic POS.

- *Ní raibh an toghairm churtha amach roimh ré as #Gaeilge*$_{\wedge}$
  ‘The summons wasn’t sent out beforehand in #Irish’

Otherwise, if they are appended or prepended to a tweet, they are tagged as `#`.

- #offline$_{\#}$ oíche mhaith. tá mé tuirseach
  #offline good night. I am tired

B.2.3 Multi-word hashtags

If a hashtag contains a string of words, they are tagged as `#MWE`.

- *#BígíAnn* (*Bígí Ann*)
  ‘Be there with us’

- *#snaSAMrófhada* (*sna SAM ró-fhada*)
  ‘too long in the U.S.A’

B.2.4 At-mentions

At-mentions are used to refer to, or link a tweet to, another user. Similarly, if at-mentions are syntactically part of the sentence, mark them as per their correct syntactic POS.

- *Rugbaí Beo ar @TG4dotTV*$_{\wedge}$ *inniu ag 15:15pm*
  ‘Live Rugby on @TG4dotTV today at 15:15pm’

Otherwise, if they are appended or prepended to a tweet, they are tagged as `@`.

- *@EIREHUB*$_{@}$ *Tá fáilte romhat*
  ‘@EIREHUB You are welcome’

B.2.5 Retweets

Tweets can be “retweeted” by another user. They are introduced by the indicator *RT*, followed by the original tweet’s username and a colon. Both *RT* and the colon are tagged with the discourse marker tag ~. The following is a retweet of a tweet from @TG4TV.

- *RT*$_{\sim}$ *@TG4TV :*$_{\sim}$ *Go n-éirí linn anocht ag @IFTA*

  ‘RT @TG4TV Good luck to us tonight at @IFTA’

B.2.6 Verbal Nouns

Irish Verbal nouns have distinct roles from common or proper nouns and thus require a separate tag.

Verbal nouns are used to denote non-finite phrases. Sometimes this is optionally indicated by a preceding infinitive marker *a*:

- *Níl sé éasca bheith*$_{VN}$ *ag ithe yo-yos*

  ‘It’s not easy to eat (to be eating) yo-yos’

- *Áthas orm é sin a chloisteáil*$_{VN}$

  ‘I’m happy to hear that’

Progressive aspectual phrases in Irish are denoted by the preposition *ag* followed by a verbal noun:

- Bhí mé ag *éisteacht*$_{VN}$ *sa charr*

  ‘I was listening in the car’

B.2.7 Names

Names of people, organisations, places etc. are tagged as proper nouns. If the name is a string, all nouns are tagged as proper nouns.

- *Daft*$_\wedge$ *Punk*$_\wedge$ *as*$_P$ *Gaeilge*$_\wedge$

  'Daft Punk in Irish'

If there are other function words in the proper noun phrase they are tagged as per their normal POS.

- *Cú*$_\wedge$ *na*$_D$ *mBaskerville*$_\wedge$

  'The Hound of the Baskervilles'

The days of the week in Irish use the particle *Dé*. This particle is also labelled as a proper noun.

- *Dé*$_\wedge$ *Máirt*

  'Tuesday'

B.2.8 Foreign

Irish tweets contain code-switching from Irish to English. This can occur at an intra-sentential level or inter-sentential level. Non-Irish words are tagged as `G`.

- *Beagnach 500 likes*$_G$ *ag @NaGaeilOga ar facebook*

  '@NaGaeilOga have nearly 500 likes on Facebook'

- *an t-am seo an t7ain seo chugainn bei 2 ag partyáil le muintir Ráth Daingin! Hope*$_G$ *youre*$_G$ *not*$_G$ *too*$_G$ *scared*$_G$ *#upthevillage*

  this time next week we will be partying with the Ráth Daingin people! Hope youre not too scared #upthevillage

B.2.9 Typographical errors and non-standard spelling

The user-generated nature of the text found in tweets results in non-standard spelling and typos.

Irish tweeters sometimes omit diacritics, and instead use un-accented vowels. While these are technically incorrectly spelt, they should be tagged as per their intended form.

- ***Ce*** *acu is* ***mo*** *a* ***dheanann*** *dochar don teanga* → ***Cé*** *acu is* ***mó*** *a* ***dhéanann*** *dochar don teanga*
  ‘Which one of them harms Irish the most’

The text speak and online variation of Irish has seen the creation of new forms of spelling, the purpose of which is often to shorten the length of the string. These new forms are tagged as per their original POS.

- *Fuair mé seo an t****7****ain* → *Fuair mé seo an t****seacht****ain seo*
  ‘I got this this week’

- *v* → *bhí*
  ‘was’

Typos and mistakes can also result in misspellings. These should be tagged as per their intended form.

- *bhí sibh* ***at*** *fheabhas inniu* → *bhí sibh* ***ar*** *fheabhas inniu*
  ‘You were great today’

B.2.10 Abbreviations

There are regular and standard abbreviations in Irish include *srl* ‘etc’. Other types of abbreviations arise from ‘text speak’ where words or phrases are shortened. All abbreviations are tagged as `G`.

- *GRMA* → *Go Raibh Maith Agat*
  ‘Thank You’

- *mgl* → *maith go leor*
  ‘fair enough’

- *srl*
  ‘etc.’

B.2.11 Particles

There are numerous particles in Irish, all of which are tagged as `T`.

- Adjectival: An leabhar *is*$_T$ *deise*

  'The nicest book'

- Complementiser: *Ceapaim go*$_T$ *mbeidh mé críochnaithe*

  'I think I will be finished'

- Time: *Ag snámh anseo ag a*$_T$ *deich a chlog*

  'Swimming here at 10 o'clock'

- Interrogative: *An*$_T$ *mbualfaidh mé leat ansin?*

  'Will I meet you then?'

- Verbal: *d'*$_T$ *iarr mé orthu*

  'I asked them'

- Infinitive Marker: *Ní mór dúinn a*$_T$ *bheith airdeallach*

  'We have to be alert'

- Adverbial: *@labhaoisen beidh sé sin ag teacht go*$_T$ *luath :)*

  '@labhaoisen that will be coming soon :)'

B.2.12 Interjections

Interjections typically include exclamations like *Bhuel* 'Well' or *Ó* 'Oh'. User-generated text also contains many more informal interjections. All of these are tagged as `!`. Often English interjections are used where an Irish version is not available, and often when they are onomatopoeic. These are tagged also as `!` (and not as `G`).

- LOL (laugh out loud)

- Ehm

- Ouch

- *Féar_plé* ‘Fair play’

- Hmmm

- *Yó* ‘Yo’

- Haha

- AAAAHHHHH

- *Úúúú* ‘Oooh’

Appendix C

Irish Dependency Treebank Statistics

In this appendix, we present some statistics on the content of the Irish Dependency Treebank.

Firstly, Table C.1 provides an overall summary of the basic statistics associated with the treebank, much of which is discussed in the main body of this thesis (Chapter 3 and Chapter 4). We then provide an overview of the dependency labels distribution throughout the treebank in Table C.2, giving the frequency and relative frequency of the top 25 most used labels. Table C.3 gives a summary of the edge lengths throughout the treebank. This is the number of tokens an edge spans between a head token and its dependent. We also provide a summary of the sentence length distribution in Table C.4, within length groupings of 5. Table C.5 shows the top 25 dependency paths of length 2. Finally, we give a list of the frequencies for the top 25 coarse-grained POS dependency label pairs in Table C.6 (i.e. the combination of coarse-grained POS of a head token and the dependency label used to attach a dependent).

# trees	1020
# coarse-grained POS tags	31
# fine-grained POS tags	62
# coarse-grained dependency labels	21
# fine-grained dependency labels	47
# projective trees	924
# non-projective trees	96

Table C.1: Basic statistics for the Irish Dependency Treebank (as of January 2016)

	Label	**frequency**	**% relative frequency**
1	pobj	2932	12.4
2	punctuation	2407	10.2
3	nadjunct	1808	7.6
4	padjunct	1785	7.5
5	det	1660	7.0
6	subj	1567	6.6
7	coord	1466	6.2
8	obl	1031	4.4
9	top	1019	4.3
10	obj	707	3.0
11	adjadjunct	703	3.0
12	comp	697	2.9
13	advadjunct	535	2.3
14	vparticle	532	2.2
15	xcomp	474	2.0
16	subadjunct	446	1.9
17	relmod	434	1.8
18	relparticle	386	1.6
19	quant	381	1.6
20	adjpred	314	1.3
21	ppred	310	1.3
22	dem	298	1.3
23	toinfinitive	290	1.2
24	npred	234	1.0
25	poss	234	1.0

Table C.2: Top 25 dependency labels used (% to the nearest one decimal point)

edge length	**frequency**	**% relative frequency**
1-10	22110	93.35
11-20	974	4.11
21-30	319	1.35
31-40	143	0.60
41-50	53	0.22
51-60	26	0.11
61-70	12	0.05
71-80	10	0.04
81-90	4	0.02
91-100	5	0.02
100+	28	0.12

Table C.3: Summary of edge lengths (distance between head and dependent)

Sentence Length (#tokens) Group	**frequency**	**% relative frequency**
1-5	49	4.8
6-10	200	19.6
11-15	139	13.6
16-20	147	14.4
21-25	123	12.1
26-30	106	10.4
31-35	65	6.4
36-40	69	6.8
41-45	35	3.4
46-50	17	1.7
51-55	15	1.5
56-60	11	1.1
61-65	8	0.8
66-70	10	1.0
71-75	1	0.1
76-80	3	0.3
81-85	6	0.6
86-90	3	0.3
91-95	1	0.1
96-100	1	0.1
100+	10	0.1

Table C.4: Summary of sentence lengths

	Head-Dependent	**frequency**	**% relative frequency**
1	padjunct-pobj	1635	8.9
2	pobj-nadjunct	732	4.0
3	obl-pobj	583	3.2
4	pobj-det	442	2.4
5	nadjunct-det	433	2.4
6	comp-subj	425	2.3
7	pobj-padjunct	416	2.3
8	subj-det	357	1.9
9	relmod-relparticle	344	1.9
10	subadjunct-comp	321	1.7
11	coord-subj	265	1.4
12	comp-vparticle	252	1.4
13	nadjunct-nadjunct	251	1.4
14	xcomp-pobj	238	1.3
15	subj-nadjunct	232	1.3
16	ppred-pobj	223	1.2
17	pobj-adjadjunct	219	1.2
18	coord-coord	204	1.1
19	coord-padjunct	191	1.0
20	coord-pobj	181	1.0
21	obj-det	178	1.0
22	pobj-coord	174	0.9
23	comp-punctuation	166	0.9
24	comp-padjunct	151	0.8
25	coord-nadjunct	150	0.8

Table C.5: Top 25 dependency paths of length 2 (% to the nearest one decimal point)

POS/ label	**frequency**
Prep / pobj	2888
Noun / det	1464
Noun / nadjunct	1293
Conj / coord	1283
Verb / subj	1171
_ / top	1019
Verb / punctuation	874
Verb / padjunct	639
Verb / obl	598
Noun / adjadjunct	596
Noun / padjunct	587
Verb / vparticle	520
Verb / obj	429
Noun / punctuation	390
Conj / punctuation	386
Noun / relmod	373
Verb / relparticle	371
Verb / xcomp	323
Conj / comp	311
Noun / quant	309
Verbal / toinfinitive	284
Verbal / padjunct	284
Cop / subj	270
Verb / subadjunct	265
Verb / advadjunct	256
Prop / nadjunct	249
Noun / dem	247

Table C.6: Top 25 Coarse-grained POS/ dependency label pairs