Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar

A thesis submitted for the degree of Doctor of Philosophy

Elaine Uí Dhonnchadha

Dublin City University

Supervisor: Prof. Josef Van Genabith

December 2008

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, 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.:_____ Date: _____

Abstract

In this thesis, we present the development and evaluation of a suite of annotation tools for unrestricted Irish text, which go from tokenization, morphological analysis, part-of-speech tagging, right through to partial parsing. In order to develop such tools, a large body of texts is required for testing purposes. We, therefore, begin by describing our involvement in the creation of a 30 million word corpus of Irish texts (New Corpus for Ireland). From this corpus, we randomly extracted 3,000 sentences which we annotated and manually corrected in order to create a Gold Standard Corpus for evaluation purposes.

We then present the annotation tools. Firstly, we describe scaling a proof-of-concept implementation of finite-state tokenization and morphological analysis based on Xerox Finite State Tools (Uí Dhonnchadha, 2002, p146), to unrestricted text. After semi-automatic population of the finite-state morphology (FSM) lexical resources, the morphological analyser contains a lexicon of 30K lemmas, which together with a set of morphological guessers assign at least one morphological analysis to all tokens in unrestricted texts. Following this, we describe our POS tagger for Irish, implemented using Constraint Grammar Disambiguation Rules, and *vislcg2* software. The POS tagger currently achieves an f-score of 95% on development data and 94.35% on unseen test data. This tagger has been used to tag the 30 million word corpus of Irish.

Finally, we present our implementation of partial parsing, which is a combination of dependency analysis overlaid with finite-state chunking. As this is the first attempt at implementing a partial parser for Irish, (to our knowledge), there were no guidelines or precedents available. The dependency analysis uses Constraint Grammar Dependency Mapping Rules, and the chunker is implemented using regular expressions and Xerox Finite-State Tools. The dependency analysis currently achieves an f-score of 93.60% on development data and 94.28% on unseen test data. The chunker achieves an f-score of 97.20% on development data and 93.50% on unseen test data.

Acknowledgements

Sincere thanks to my supervisor Prof. Josef van Genabith for his advice and direction. Thank you to Dr. Carl Vogel for his valuable comments and suggestions. Buíochas go háirithe, do Ghearóid, grá mo chroí, gan do chuid tacaíochta ní bheadh sé críochnaithe go deo. Go raibh maith agat chomh maith as do chuid comhairle ó thaobh comhréire de. Buíochas speisialta do Bhéibhinn, Éanna agus Síomha as gliondar a chur orm i gcónaí.

Abbreviations

CG Constraint Grammar

CNG Corpas Náisiúnta na Gaeilge

FSM Finite State Morphology

MRD Machine-Readable Dictionary

MWE Multi-word expression

NCI New Corpus for Ireland (Irish & Irish English)

NCII New Corpus for Ireland - Irish Only

POS Part-of-Speech

Typographical Conventions

All Irish language examples in the text are in *italic* typeface followed by the translation in single quotation marks, e.g. Irish: *cos* 'foot'.

Single quotation marks are also used to highlight English words described in the text e.g. the plural of 'woman' is 'women'.

When a particular word is being discussed, it is highlighted using **bold** typeface e.g. Irish: ***cathair*** 'city'.

TABLE OF CONTENTS

ABSTRACT 3

ACKNOWLEDGEMENTS 4

ABBREVIATIONS 5

TYPOGRAPHICAL CONVENTIONS 5

LIST OF FIGURES 10

LIST OF TABLES 12

APPENDICES 13

OVERVIEW 14

AIMS OF THE THESIS 14

STRUCTURE OF THE THESIS 18

PART I CORPUS CREATION AND ANNOTATION METHODOLOGIES 23

1 DEVELOPMENT OF A CORPUS OF MODERN IRISH 24

1.1 INTRODUCTION 24

1.2 CORPUS DESIGN AND COLLECTION 24

1.3 MORPHOSYNTACTIC ANNOTATIONS 26

1.4 TEXT PRE-PROCESSING 27

1.5 HEADER INFORMATION DATABASE 30

1.6 XML CORPUS ENCODING STANDARD (XCES) MARK-UP 32

1.7 CORPUS EVALUATION 33

1.8 SUMMARY 33

2 LINGUISTIC ANNOTATION OF CORPORA 35

2.1 INTRODUCTION 35

2.2 LINGUISTIC ANNOTATION 36

2.3 TECHNIQUES FOR PART-OF-SPEECH TAGGING 36

2.4 PART-OF-SPEECH TAGGING OF IRISH 42

2.5 TECHNIQUES FOR SYNTACTIC PARSING 44

2.6 PARTIAL PARSING OF IRISH 52

2.7 RELATED RESEARCH 54

2.8 LINGUISTIC ANNOTATION: A WORKED EXAMPLE 55

2.9 SUMMARY 57

3 A GOLD STANDARD EVALUATION CORPUS 59

3.1 INTRODUCTION 59

3.2 TEXT SELECTION FOR GOLD STANDARD CORPUS 59

3.3 MANUAL DISAMBIGUATION 61

3.4 GOLD STANDARD DEPENDENCY CORPUS AND GOLD STANDARD CHUNKED CORPUS 63

3.5 EVALUATION MEASURES 64

3.6 SUMMARY 65

PART II AUTOMATIC PART-OF-SPEECH TAGGING FOR IRISH 67

4 FINITE-STATE TOKENIZATION 68

4.1 INTRODUCTION 68

4.2 TOKENIZATION ISSUES 68

4.3 IMPLEMENTATION OF THE FINITE-STATE TOKENIZER 72

4.4 EVALUATION OF THE TOKENIZER 77

4.5 SUMMARY 79

5 FINITE-STATE MORPHOLOGICAL ANALYSIS 80

5.1 INTRODUCTION 80

5.2 SEMI-AUTOMATIC EXTENSION OF FSM LEXICONS 81

5.3 EVALUATION OF RESULTS OF SEMI-AUTOMATIC POPULATION OF LEXICONS 86

5.4 ADDITION OF DERIVATIONAL MORPHOLOGY RULES 87

5.5 EVALUATION OF MORPHOLOGICAL ANALYSIS COVERAGE 89

5.6 COMPOUND RECOGNITION 90

5.7 MORPHOLOGICAL GUESSERS 98

5.8 EVALUATION OF GUESSERS 101

5.9 MORPHOLOGICAL ANALYSIS LOOKUP STRATEGY 104

5.10 SUMMARY OF TOKEN RECOGNITION RATES 105

5.11 SUMMARY 107

6 POS TAGGING USING MORPHOSYNTACTIC DISAMBIGUATION 108

6.1 INTRODUCTION 108

6.2 PRINCIPLES OF CONSTRAINT GRAMMAR 108

6.3 CG MORPHOSYNTACTIC DISAMBIGUATION RULES FOR IRISH 113

6.4 DISAMBIGUATION CHALLENGES 119

6.5 EVALUATION OF POS DISAMBIGUATION RATE 126

6.6 EVALUATION OF POS TAGGING 128

6.7 SUMMARY 131

PART III PARTIAL PARSING OF IRISH 132

7 DEPENDENCY ANALYSIS OF IRISH 133

7.1 INTRODUCTION 133

7.2 GRAMMATICAL FUNCTIONS AND DEPENDENCY RELATIONS FOR IRISH 134

7.3 ANNOTATION SCHEME 135

7.4 SENTENCE TEMPLATES FOR DEPENDENCY ANALYSIS 152

7.5 IMPLEMENTATION 181

7.6 EVALUATION 189

7.7 SUMMARY 194

8 CHUNKING 195

8.1 INTRODUCTION 195

8.2 ANNOTATION SCHEME FOR NESTED CHUNKING 195

8.3 IMPLEMENTATION OF THE FINITE-STATE CHUNKER 198

8.4 EVALUATION 203

8.5 SUMMARY 208

9 CONCLUSION 209

9.1 SUMMARY 209

9.2 MAIN CONTRIBUTIONS 211

9.3 NLP TOOLS FOR IRISH 212

9.4 LINGUISTIC RESOURCES FOR IRISH 212

9.5 FUTURE RESEARCH 213

GLOSSARY OF TERMS **214**

PUBLICATIONS RESULTING FROM RESEARCH REPORTED IN DISSERTATION **216**

REFERENCES **217**

List of Figures

FIGURE 1 CORPUS ANNOTATION PROCESSING ARCHITECTURE 19
FIGURE 2 XCES SAMPLE 26
FIGURE 3 DOCUMENT HEADER INTERRUPTING BODY TEXT 28
FIGURE 4 POEM MARK-UP 29
FIGURE 5 DRAMA MARK-UP 30
FIGURE 6 CONSTITUENCY HIERARCHY 47
FIGURE 7 FLAT CONSTITUENCY STRUCTURE 49
FIGURE 8 DEPENDENCY REPRESENTATION 50
FIGURE 9 DEVELOPMENT - EVALUATION CYCLE 65
FIGURE 10 TOKENIZER DEFINITIONS: CONTRACTIONS 74
FIGURE 11 TOKENIZER DEFINITIONS: ABBREVIATIONS 75
FIGURE 12 TOKENIZER DEFINITIONS: ENGLISH POSSESSIVE APOSTROPHE 75
FIGURE 13 TOKENIZER DEFINITIONS: XML TAGS 75
FIGURE 14 TOKENIZER DEFINITIONS: NUMERIC EXPRESSIONS AND LIST NUMBERING 75
FIGURE 15 TOKENIZER DEFINITIONS: URLS AND E-MAIL ADDRESSES 76
FIGURE 16 TOKENIZER DEFINITIONS: INITIAL MUTATION HYPHEN 76
FIGURE 17 TOKENIZER DEFINITIONS: MULTI-WORD EXPRESSIONS 77
FIGURE 18 MACHINE-READABLE DICTIONARY TEXT 82
FIGURE 19 SAMPLE OF LEXC COMPATIBLE INPUT AUTOMATICALLY DERIVED FROM MRD 84
FIGURE 20 SAMPLE OF SCANNED DATA 85
FIGURE 21 SAMPLE OF LEXC COMPATIBLE INPUT DERIVED FROM SCANNED DATA 85
FIGURE 22 EXTRACT FROM COMPOUNDING REGULAR EXPRESSION SCRIPT 91
FIGURE 23 EXTRACT 1 FROM VERB GUESSER REGULAR EXPRESSION SCRIPT 99
FIGURE 24 EXTRACT 2 FROM VERB GUESSER REGULAR EXPRESSION SCRIPT 100
FIGURE 25 EXTRACT FROM NOUN GUESSER TYPE 2 REGULAR EXPRESSION SCRIPT 101
FIGURE 26 CG COHORTS AND READINGS 109
FIGURE 27 EXAMPLE OF CG2 SYNTAX 113
FIGURE 28 TEMPLATE FOR SENTENCE WITH FINITE MAIN VERB (ANALYTIC) 152
FIGURE 29 TEMPLATE FOR SENTENCE WITH FINITE MAIN VERB (SYNTHETIC) 153
FIGURE 30 TEMPLATE FOR SENTENCE WITH FINITE MAIN VERB 154
FIGURE 31 TEMPLATE FOR SUBSTANTIVE VERB *BÍ* 'TO BE' 156

FIGURE 32 TEMPLATE FOR PROGRESSIVE ASPECT 158
FIGURE 33 TEMPLATE FOR IDENTITY COPULA 162
FIGURE 34 TEMPLATE FOR CLASSIFICATORY COPULA 163
FIGURE 35 TEMPLATE FOR OWNERSHIP COPULA 163
FIGURE 36 TEMPLATE FOR COMPARATIVE COPULA 164
FIGURE 37 TEMPLATE FOR FRONTING USING A COPULA 165
FIGURE 38 TEMPLATE FOR FRONTED COPULAR CONSTRUCTION 166
FIGURE 39 TEMPLATE FOR IDIOMATIC USE OF THE COPULA 167
FIGURE 40 TEMPLATE FOR FMV INTRODUCING COPULAR COMPLEMENTS 168
FIGURE 41 TEMPLATE FOR COPULA INTRODUCING COPULAR COMPLEMENTS 168
FIGURE 42 TEMPLATE FOR INFINITIVE WITH AUXILIARY VERB 168
FIGURE 43 TEMPLATE FOR DIRECT RELATIVE CLAUSES 170
FIGURE 44 TEMPLATE FOR INDIRECT RELATIVES 172
FIGURE 45 TEMPLATE FOR WH-QUESTIONS 173
FIGURE 46 TEMPLATE FOR PASSIVE USING AUTONOMOUS VERB FORM 174
FIGURE 47 TEMPLATE FOR PASSIVE USING VERBAL ADJECTIVE 175
FIGURE 48 TEMPLATE FOR SENTENCE WITH FINITE PHRASAL VERB 177
FIGURE 49 DEPENDENCY ANALYSIS FLOWCHART 181
FIGURE 50 DEPENDENCY ANNOTATION: CLAUSE BOUNDARIES 184
FIGURE 51 DEPENDENCY ANNOTATION: FINITE MAIN VERBS 185
FIGURE 52 DEPENDENCY ANNOTATION: PREPOSITIONAL PHRASES 185
FIGURE 53 DEPENDENCY ANNOTATION: DEPENDENT MODIFIERS 186
FIGURE 54 DEPENDENCY ANNOTATION: SUBJECTS 1 186
FIGURE 55 DEPENDENCY ANNOTATION: SUBJECTS 2 186
FIGURE 56 DEPENDENCY ANNOTATION: SUBJECTS 3 187
FIGURE 57 DEPENDENCY ANNOTATION: SUBJECTS 4 187
FIGURE 58 DEPENDENCY ANNOTATION: OBJECTS 1 187
FIGURE 59 DEPENDENCY ANNOTATION: OBJECTS 2 188
FIGURE 60 DEPENDENCY ANNOTATION: OBJECTS 3 188
FIGURE 61 DEPENDENCY ANNOTATION: PREDICATES 188
FIGURE 62 DEPENDENCY ANNOTATION: TEMPORAL ADVERBIALS 189
FIGURE 63 DEPENDENCY ANNOTATION: OTHER NOUNS 189
FIGURE 64 CHUNKER DEFINITIONS: GENERAL 201
FIGURE 65 CHUNKER DEFINITIONS: VERB CHUNKS 201
FIGURE 66 CHUNKER DEFINITIONS: PREPOSITIONAL CHUNKS 203
FIGURE 67 CHUNKER DEFINITIONS: ASPECTUAL CHUNKS 203

List of Tables

TABLE 1 NCII COLLECTION TARGETS 25
TABLE 2 NCII TEXT SOURCES 26
TABLE 3 USE OF <GAP> TAG 29
TABLE 4 HEADER INFORMATION DATABASE 31
TABLE 5 NCII: TARGETS VS. ACTUAL COLLECTION 33
TABLE 6 SAMPLE OF BROWN TAGS 38
TABLE 7 SAMPLE OF PAROLE TAGS FOR IRISH 38
TABLE 8 SAMPLE OF PAROLE SHORT TAGS FOR IRISH 38
TABLE 9 DEPENDENCY RELATIONS 52
TABLE 10 COMPOSITION OF GOLD STANDARD (3000) POS CORPUS 60
TABLE 11 COMPOSITION OF GOLD STANDARD (250) DEPENDENCY CORPUS 63
TABLE 12 TOKENIZATION EVALUATION 77
TABLE 13 DEVELOPMENT SET: ERROR ANALYSIS OF TOKENIZATION 78
TABLE 14 DEVELOPMENT SET: AFTER CORRECTION 79
TABLE 15 SUMMARY OF FOCLÓIR PÓCA DATA 82
TABLE 16 SAMPLE OF MRD DATA 83
TABLE 17 EXTENDED FSM LEXICONS 86
TABLE 18 COVERAGE OF MORPHOLOGICAL ANALYSERS 89
TABLE 19 COMPOUND RECOGNISER 1: ERROR ANALYSIS 92
TABLE 20 ANALYSIS OF POS AND FEATURE ASSIGNMENT TO COMPOUNDS 93
TABLE 21 COMPOUND RECOGNISER 1: ANALYSIS OF OMITTED COMPOUNDS 94
TABLE 22 COMPOUND RECOGNISER 2: ERROR ANALYSIS 97
TABLE 23 COMPOUND RECOGNISER 2: ANALYSIS OF OMITTED COMPOUNDS 98
TABLE 24 DEVELOPMENT SET: GUESSER PRECISION 102
TABLE 25 VERB GUESSER: ERROR ANALYSIS 102
TABLE 26 TEST SET: GUESSER PRECISION 103
TABLE 27 SUMMARY OF TOKEN RECOGNITION RATES 106
TABLE 28 DISAMBIGUATION: ERROR ANALYSIS OF TOKEN *A* 120
TABLE 29 CONFUSION MATRIX FOR PARTICLE *A* 120
TABLE 30 HOMONYMOUS NUMBER FORMS 124
TABLE 31 DEVELOPMENT SET: RATE OF DISAMBIGUATION 126
TABLE 32 DEVELOPMENT SET: DETAILED POS TAGGING RESULTS 130
TABLE 33 GRAMMATICAL FUNCTION AND HEAD/MODIFIER DEPENDENCY LABELS 137
TABLE 34 DEPENDENCY ANNOTATION: OVERALL EVALUATION RESULTS 190
TABLE 35 DEVELOPMENT SET (150): DEPENDENCY ANNOTATION RESULTS 192
TABLE 36 DEPENDENCY ANNOTATION CONFUSION MATRIX 193
TABLE 37 BRACKETED CHUNK LABELS 196
TABLE 38 CHUNK DEPENDENCY TAGS 197

TABLE 39 TEST SUITE (225): EVALB BRACKET SCORING SUMMARY 205
TABLE 40 DEVELOPMENT SET (150): EVALB BRACKET SCORING SUMMARY 206
TABLE 41 TEST SET (100): EVALB BRACKET SCORING SUMMARY 206
TABLE 42 CHUNKER: DEVELOPMENT SET ERROR ANALYSIS 208

Appendices

A PAROLE MORPHOSYNTACTIC DESCRIPTIONS

B FINITE-STATE MORPHOLOGICAL FEATURE TAGS

C GUIDELINES FOR MANUAL POS DISAMBIGUATION

D CONSTRAINT GRAMMAR POS DISAMBIGUATION RULES

E TEST SUITE SENTENCES

F CONSTRAINT GRAMMAR DEPENDENCY RULES

G FINITE-STATE CHUNKER REGULAR EXPRESSIONS

H FINITE-STATE TO PAROLE TAG MAPPINGS

Overview

Aims of the Thesis

This thesis sets as its central aim the design, implementation and evaluation of a suite of Natural Language Processing (NLP) tools for automatic linguistic annotation of Irish texts, as well as the creation of a Gold Standard Annotated Corpus.

Specifically, we aim to develop tools, methods, and linguistic guidelines for the automatic part-of-speech (POS) tagging and partial parsing of Irish. The primary goal of the current research is to develop a POS tagger and Lemmatizer for unrestricted Irish text, and to carry out exploratory research into partial parsing of Irish. In order to do this we have developed a tokenizer, morphological analyser, disambiguator, dependency tagger, and chunker.

In the modern communication age, the use of technology is pervasive in all aspects of life; in the home, and for leisure, as well as business, educational, and religious activities. The use of computers, mobile phones, internet, and electronic games is increasing all the time, and all of these technologies employ natural language interfaces. One measure which we can take to help maintain linguistic diversity, is to ensure that minority languages, such as Irish, can benefit from the technology available to the major languages. We can do so by taking advantage of the research into NLP of these more technologically advanced languages.

In order for software developers and businesses to provide language specific end-user applications and services (e.g. word processing, speech synthesis, automatic call answering etc.) the basic linguistic tools and resources need to be in place (Krauwer, 2003). These tools include morphological analysers, part-of-speech taggers, and parsers. To date, there has been little research in the area of Computational Linguistics for Irish, largely due to the dominance of English in Ireland. In this work, we hope to redress the balance in some small way.

This work seeks to create robust tools, which handle real-world data in a reliable, consistent, and efficient manner. Wherever possible, we use existing, tried and tested, language independent tools, such as Finite-State Morphology and Constraint Grammar, which allows us to concentrate on language specific issues. The Finite-State Morphology and Constraint Grammar methodologies which we use for tagging and partial parsing reflect the nature of Irish as they make extensive use of inflectional and derivational morphology and make use of the strict word order constraints of Irish in order to linguistically annotate strings.

A substantial part of the effort of POS tagging involves deciding on the most appropriate POS tag to assign to functional categories and particles. In all cases we strive for consistency and choose the most generally applicable POS categories for particular lexical items. For example, although prepositions are used in a number of different constructions in

Irish, (e.g. locative, temporal, aspectual etc.) we choose to make these functional distinctions at a higher (e.g. phrasal) level rather than at the POS level. All instances of prepositions are tagged as such and it is through looking at the wider context during dependency analysis that we attempt to distinguish function.

In order to deal with unrestricted corpus data, containing sentences which are grammatical, ungrammatical or something in between, we use a *reductionist* method of tagging (Karlsson et al., 1995, p13). Firstly, we generate the choice of possible morphological analyses for each word. We then remove impossible or unlikely options. The rules explicitly define what is not grammatical as well as defining grammatical structures. In the words of Karlsson (1995, p37) 'everything is licensed unless explicitly ruled-out'. The last remaining analysis is never removed, therefore, we are able to provide a POS and partial parse for every input.

After POS tagging, the next step is to identify larger syntactic units in the text. The first task is to decide what those syntactic units are and how they should be annotated. In parsing a language for the first time, this constitutes a major part of the work. We then investigate how automatic partial parsing/chunking can be implemented.

There are two main schools of thought regarding syntactic annotation among the existing parsed corpora (treebanks) for other languages. Some implement a constituency based analysis (Marcus et al., 1993), others have a dependency based analysis (Hajič, 1998) and a few combine elements of both (Brants et al., 2003). There is a substantial overlap between both types of analysis.

Our primary aim in this exploration of partial parsing of Irish is to account for as much of the linguistic phenomena as possible and to decide on an initial style guide for the partial syntactic annotation of the language. In order to be comprehensive, we have implemented both partial dependency analysis and partial constituency parsing (i.e chunking). We have annotated dependency relations and grammatical functions using Constraint Grammar and have overlaid this with chunk boundaries using a regular-expression grammar.

In our dependency analysis, we identify clause boundaries and head-modifier dependencies within clauses, as well as the grammatical functions of subject, object, predicate, and various types of prepositional phrase (Karlsson, 1995; Tapanainen, 1996; 1999). As is usual for dependency analysis, we annotate the tokens present in the input string without introducing any abstract categories (phrasal nodes or elipted or elided items). This results in a partial, rather than full parse of the texts.

Using the dependency tags we identify phrase-like structures known as 'chunks'. Identifying the relationships between chunks (i.e. PP attachment and co-ordination) is beyond the scope of the current work, as are issues relating to long-distance dependencies.

We have applied finite-state techniques (Beesley and Karttunen, 2003) to a new language, namely Irish, and we find that finite-state techniques successfully and efficiently handle all of the tokenization and morphological phenomena associated with Irish.

Our partial parsing is preliminary and tentative in nature, as there are several issues in the theorethical syntax of Irish which have yet to be resolved. Some issues are the result of a lack of research into VSO languages in general, i.e. the status of VP in Irish, and other theoretical issues such as the nature of periphrastic aspectual structures in Irish are unclear. We intend, therefore, that this current research will provide a useful basis for future work in the parsing of Irish.

Processing Overview

In the following section, we give a overview of the linguistic annotation applied to a simple sentence. For example, in the sentence in (1), the initial step of tokenization results in four tokens as shown in (2).

```
(1)   Chan an  cailín.
      Sang the girl
      'The girl sang'
```

```
(2)   Chan
      an
      cailín
      .
```

The morphological analysis of (2) is given in (3), where each analysis contains a word form followed by its lemma, its part-of-speech category, and its morphosyntactic features. Details of all the morphosyntactic feature tags used in this thesis may be found in Appendix B.

```
(3)   "<Chan>"
         "chan" CU Part Vb Neg
         "chan" CU Cop Pres Neg
         "can" Verb VTI PastInd Len
         "can" Verb VTI PastInd Q Len
         "can" Verb VTI PastInd NegQ Len
         "can" Verb VTI PastInd Neg Len

      "<an>"
         "is" Cop Pres Q
         "is" Cop Pres Dep Q
```

```
        "an" Art Sg Def
        "an" Part Vb Q

    "<cailín>"
        "cailín" Noun Masc Com Sg
        "cailín" Noun Masc Com Sg DefArt
        "cailín" Noun Masc Gen Sg

    "<.>"
        "." Punct Fin
```

The output of the morphological analyser is then disambiguated in order to arrive at (in most cases[1]) one unambiguous POS category (and morphosyntactic features) for each token in the input string (4).

```
(4)   "<Chan>"     "can" Verb VTI PastInd Len
      "<an>"       "an" Art Sg Def
      "<cailín>"   "cailín" Noun Masc Com Sg DefArt
      "<.>"        "." Punct Fin
```

At this point, we have POS tagged text (4) which we can either convert to XML Corpus Encoding Standard (XCES) (Ide et al., 2000) formatted corpus text (5), or which we can use as the basis for dependency analysis processing, as in (6).

```
(5)   <s>
      <w base="can" tag="Verb VTI PastInd Len">Chan</w>
      <w base="an" tag="Art Sg Def">an</w>
      <w base="cailín" tag="Noun Masc Com Sg DefArt">cailín</w>
      <w base="." tag="Punct Fin">.</w>
      </s>
```

In (6), after dependency analysis, the verb *chan* 'sing' and subject *cailín* 'girl' have received functional labels `@FMV` (finite main verb) and `@SUBJ` (subject). The token *an* 'the', which has been identified as an article, is annotated with `@>N` meaning that it is dependent on the noun to its right.

[1] Some ambiguities may remain unresolved.

```
(6)   "<Chan>"    "can" Verb VTI PastInd Len @FMV
      "<an>"      "an" Art Sg Def @>N
      "<cailín>"  "cailín" Noun Masc Com Sg DefArt @SUBJ
      "<.>"       "." Punct Fin
```

After dependency analysis, to facilitate further processing, the lemma and morphosyntactic features are concatenated and each sentence is converted into a string of token-tag pairs, as shown in (7).

```
(7)   "Chan" "can"+Verb+VTI+PastInd+Len+@FMV "an"
      "an"+Art+Sg+Def+@>N "cailín"
      "cailín"+Noun+Masc+Com+Sg+DefArt+@SUBJ "." "."+Punct+Fin
```

Finite-state regular expressions are applied to each sentence to identify syntactic chunks such as verb, noun, and adverbial chunks, as shown in (8).

```
(8)   [S [V "<Chan>" "can"+Verb+VTI+PastInd+Len+@FMV V]
      [NP "<an>" "an"+Art+Sg+Def+@>N "<cailín>"
      "cailín"+Noun+Masc+Com+Sg+DefArt+@SUBJ NP]"<.>" "."+Punct+Fin
      S]
```

Structure of the Thesis

This thesis consists of three parts:

- Chapters 1 - 3: Part I - Background
- Chapters 4 - 6: Part II - Part-of-Speech Tagging of Irish
- Chapters 7 - 8: Part III - Dependency Annotation and Partial Parsing of Irish
- Chapter 9: Conclusions

Part I provides a high-level and informal overview of the research reported in this thesis, including an overview of the creation of a 30-million word corpus of Irish texts. Part II details the design, implementation, and evaluation of Part-of-Speech Tagging of Irish, while Part III details the design, implementation, and evaluation of Dependency Annotation and Partial Parsing of Irish.

Figure 1 shows the processing architecture developed in this thesis.

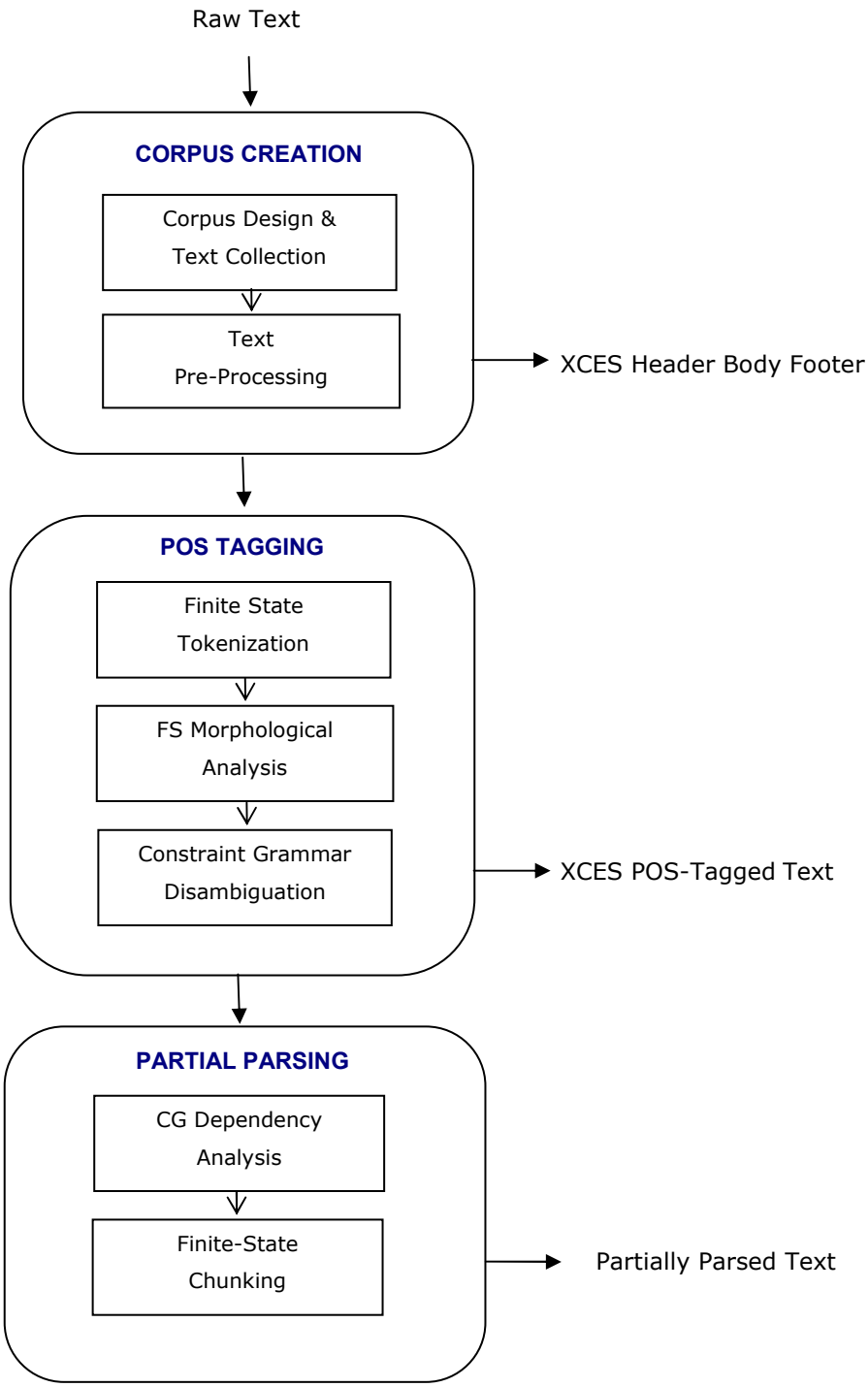


Figure 1 Corpus Annotation Processing Architecture

Part I - Corpus Creation and Annotation Methodologies

In Part I of the thesis, we provide a background discussion of corpus annotation, particularly the concepts of part-of-speech tagging and syntactic parsing. This is followed by an outline of the tools and methodologies we have chosen to use in the POS tagging and partial parsing of Irish texts. As linguistic annotation tools can only be developed and tested in conjunction with a corpus, we describe the creation of a 30-million corpus of Irish texts, and finally, we present our Gold Standard Annotated Corpus and our evaluation methodology.

In Chapter 1, we give an overview of our involvement in the development of the Irish part of the New Corpus for Ireland (NCII) (Kilgarriff, Rundell and Uí Dhonnchadha, 2007). We begin with a brief description of corpus design and the decisions relating to the type and level of linguistic annotation required, as well as a summary of text collection results. All texts are normalised into a standard character encoding (Irish has accented characters outside of the basic ASCII range) and format. XCES (XML Corpus Encoding Standard) was chosen as the text format and the ISO 8859-1 character encoding standard was used initially. Texts were subsequently converted to Unicode (UTF8). We also describe text pre-processing and validation in detail. This important task must be carried out before texts are ready for structural and linguistic mark-up. The initial quality of the text has implications for the quality of the annotation process. In this chapter, we also describe the implementation of the XML Corpus Encoding Standard (XCES) including header creation and document structure mark-up.

In Chapter 2, we introduce the main concepts in the linguistic annotation of corpora, focusing in particular on POS tagging and partial parsing. We give an overview of current methodologies for automatic annotation of corpora. We also describe the tools and methodologies chosen for the POS tagging and partial parsing of Irish.

In Chapter 3, we describe evaluation methods and the development of a Gold Standard Annotated Corpus. The linguistic annotation of text is carried out in a series of stages, with each stage providing input to the subsequent stage. Since the quality of each stage depends on the quality of the output of the previous stage, systematic and early evaluation of results is vitally important in order to ensure a good overall result. In order to evaluate the automatic annotation, we created a Gold Standard Annotated Corpus. The Gold Standard Corpus was created by randomly selecting approximately 3,000 sentences from the 30 million word NCII corpus. These sentences were randomly distributed into a Development Set (2,000 sentences approx.) and a Test Set (1,000 sentences approx.). The sentences were then automatically annotated and manually corrected. Each tool is developed incrementally by comparing its output with the manually corrected development data. Error analysis is carried out on the results, improvements are made to the tools, and they are re-tested against the

Development Set. Finally, evaluation of each tool is carried out by comparing the automatically annotated data with the manually corrected Gold Standard Test data set.

We use three evaluation measures: precision, recall and f-score. Precision is the percentage of tags automatically assigned which are correct compared to the Gold Standard tags. Recall is the percentage of Gold Standard tags which were correctly identified in the automatic tagging. For example, a tagger might correctly assign a noun tag to a small number of nouns, giving high precision. However, if there were many more tokens in the Gold Standard which were nouns then recall would be low. F-score is the harmonic mean of the two measures and this is the figure we will cite in the summaries.

Part II - Automatic Part-of-Speech Tagging of Irish

In Part II, we give a detailed account of the development and evaluation of tools for POS tagging of Irish. The prerequisite for POS tagging of corpus texts is tokenization. POS tagging itself, is carried out in two stages: firstly each token is analysed in order to assign all of its possible POS tags based on the finite-state morphological analyser, and in the second stage, we disambiguate in order to choose the appropriate tag for the token, given the particular context in which it is used.

In Chapter 4, we describe the tokenization of corpus texts. This entails segmenting the text input stream into separate tokens which will be passed on to the morphological analyser. In the tokenizer the default token is any item bounded by white-space. Multi-word expressions which we wish to keep together (e.g. idioms, place names etc.) and contractions, tokens which we wish to divide (e.g. *d'fhéach* 'looked', *m'aghaidh* 'my face' etc.), are specified and dealt with as exceptions. By default all punctuation is separated, and any exceptions to this rule (e.g. abbreviations, titles, mathematical formulae etc.) are specified in the tokenizer. The tokenizer is implemented using Xerox Finite-State Tools.[2]

In Chapter 5, we describe the scaling for use on unrestricted text of a prototype finite-state morphological analyser (Uí Dhonnchadha, 2002). This involves semi-automatically extending the basic lexicon, the addition of named entities (names, places, organisations etc.), and the addition of derivational morphology rules. The effect is to increase coverage by more than 10% resulting in over 95% of tokens receiving at least one analysis. 60% of tokens on average receive more than one analysis.

In Chapter 5, we also describe the development of a series of morphological guessers in order to further extend morphological analysis. The morphological guessers handle the

[2] See http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html (Accessed 10/05/2008)

remaining 5% of tokens, which were not recognised by the morphological analyser. These guesser transducers concatenate stems in the lexicon to identify possible compounds, as well as concatenating stems with prefixes or suffixes to identify possible derived words. The remaining tokens, which do not appear to be related to any stems in the lexicon, are analysed according to any distinguishing characteristics which they may have. For example, they may contain syllables which are indicative of a part-of-speech category (e.g. inflectional suffixes on verbs) or other morphological features (e.g. gender of nouns).

In Chapter 6, we describe morphological disambiguation. In this architecture, part-of-speech tagging consists of choosing the correct analysis for each token having more than one analysis. This disambiguation task is carried out by writing Constraint Grammar (Karlsson et al., 1995) rules which look at the local context of each token (within the scope of the sentence) in order to select the right analysis. Based on comparison with the Gold Standard Corpus, the tagger chooses the correct POS tag with an f-score of 95.01% on the Development Set and an f-score of 94.35% on the Test Set.

Part III - Automatic Dependency Annotation and Partial Parsing of Irish

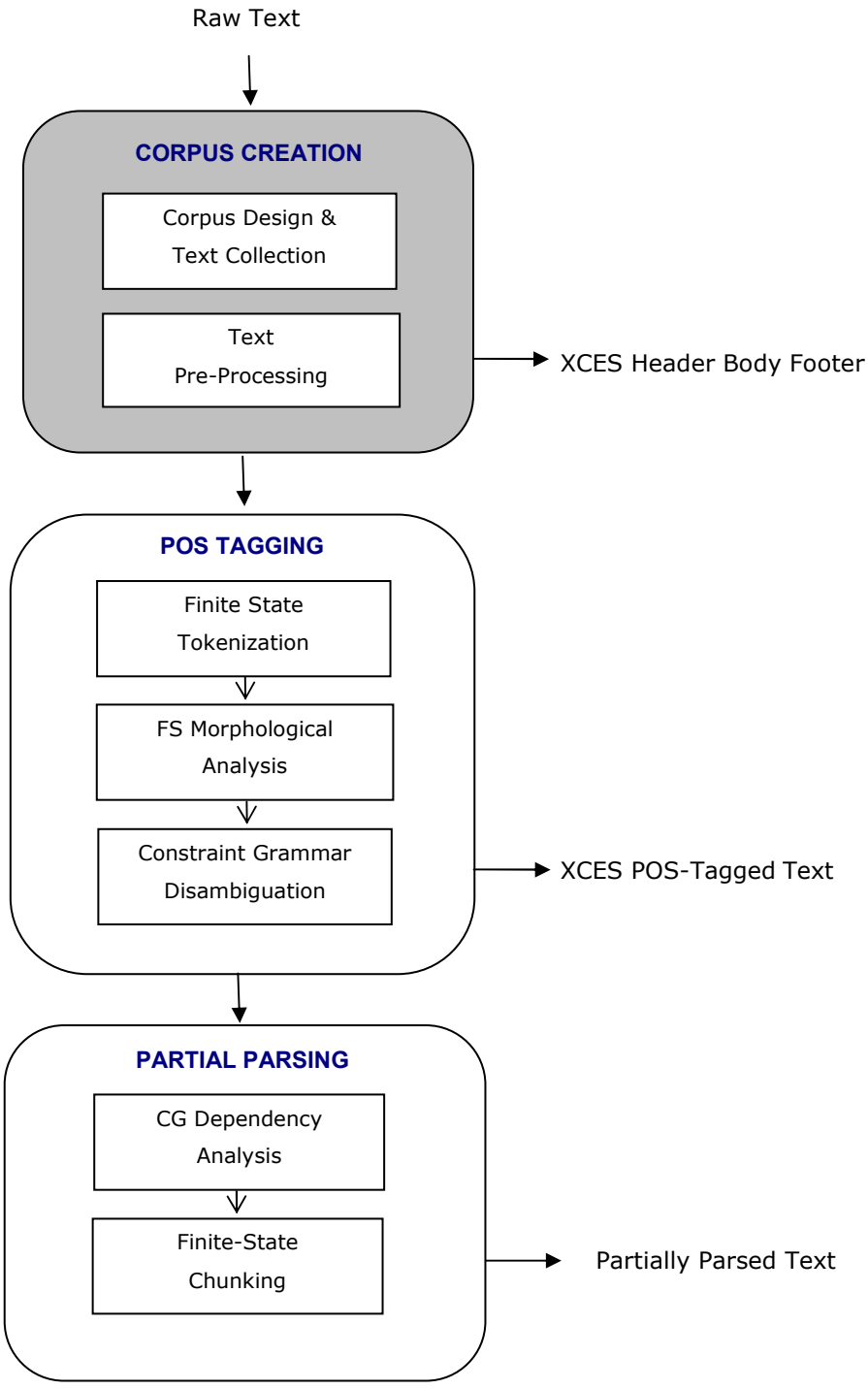
The remaining chapters are concerned with the Partial Parsing of Irish using dependency annotation and chunking.

We begin in Chapter 7 with a basic survey of Clause Structure in Irish. In this chapter, we aim to cover all of the basic syntactic structures. The sample sentences used form part of a test suite for the initial development of the Constraint Grammar dependency analysis rules. As this manually constructed test suite may not cover all the basic linguistic phenomena and does not take into account frequency of usage, we also test using a subset of the Gold Standard Corpus. We evaluate the Dependency Analysis by comparing the output of automatic annotation with the Gold Standard sentences. The Dependency Analysis Tagger currently achieves an f-score of 93.60% on the Development Set and an f-score of 94.28% on the Test Set.

Chapter 8 is concerned with the development and testing of a Finite-State Chunker. This is applied to the dependency annotated sentences. In this chapter, we describe our chunk annotation as well as the implementation of the Chunker using Xerox Finite-State Tools. The Finite-State Chunker currently achieves an f-score of 97.20% on the Development Set and an f-score of 93.50% on the Test Set.

In Chapter 9, we summarise the research reported in this thesis and suggest future directions.

Part I Corpus Creation and Annotation Methodologies



1 Development of a Corpus of Modern Irish

1.1 Introduction

A corpus is a body of texts collected for a specific purpose. The New Corpus for Ireland (NCI), was created in 2004 for *Foras na Gaeilge*, the government body responsible for the promotion of Irish in Ireland. The corpus was initially designed to support dictionary development, but it was also envisaged as a general purpose linguistic resource. To facilitate effective searching of the corpus, the texts need to be annotated at the word level, i.e. each word must be annotated with its lemma, part-of-speech category and also have its morphosyntactic features annotated as fully as possible.

The focus of this thesis is the development of corpus annotation tools, i.e. a part-of-speech (POS) tagger and partial parser for Irish. These tools were developed and tested using a subset of the 30 million word NCI-Irish (NCII) corpus of texts (see Chapter 3). In the current chapter we describe the creation of the NCII corpus, including the author's involvement in text collection and in the supervision of text pre-processing and structural mark-up of texts.

The remainder of this Chapter is laid out as follows. In Section 1.2, we describe corpus design and text collection, followed in Section 1.3 by a description of the text encoding format and the POS tagset chosen for morphosyntactic annotation. In Section 1.4 we describe text pre-processing. In Section 1.5, we describe the header information database used for XML headers, and in Section 1.6, we introduce the XML Corpus Encoding Standard (XCES). Finally, in Section 1.7, we describe XML file validation, and compare actual text collection figures to the design targets.

1.2 Corpus Design and Collection

In a corpus development project there are a number of design decisions which must be made at the outset:

- composition: what types of text are needed and in what quantity, as well as the ratio of spoken to written language to be included.
- text format: character encoding and formatting.
- text annotation: what level of structural and linguistic annotation is required.
- annotation tools: are the appropriate tools available or must they be developed.

The first task in corpus design is to decide what types of text the corpus should contain and in what proportions. This is achieved by deciding what will optimally support the intended use

of the corpus, in this case dictionary development, as well as the provision of a valuable data resource for Irish in the areas of linguistic research, NLP applications and pedagogy.

The corpus design for NCI, which has an Irish and an Irish English (Hiberno English) component, was carried out by Lexicography MasterClass Ltd. (LMC, 2004). The Irish part of NCI (NCII) contains 30 million words[3] of written text, fifty percent of which comes from books and the remainder from a variety of other written media. Table 1 gives an overview of the categories and quantities of the text types required at the outset of the project (see Table 5 p33, for actual collection figures).

Table 1 NCII collection targets

Text Category	***Words (in millions)***
Books-informative	(6 mill.)
Books-imaginative	(9 mill.)
Books Total	15.0
Newspapers	4.5
Periodicals	2.5
Official/Govt	1.5
Broadcast	1.0
Websites	5.5
TOTAL	**30.0**

NCII is composed of written rather than spoken texts, apart from some scripted dialogue and one transcribed radio interview. This is a result of practical considerations, as to our knowledge, there is no transcribed spoken material available and there was neither the time nor resources to carry out transcription in the one-year timeframe within which the NCII Corpus was created.

Texts were obtained by approaching publishers and typesetters of Irish language books and newspapers, and asking them to supply us with texts which were already in electronic format. Scanning of texts and transcription of spoken material were not part of this project but may be carried out in the future in order to improve representation in certain areas.

NCII builds on previous Irish corpus initiatives, in which the author was also involved. The Parole Corpus of Irish (ITÉ, 2001; Ó Cróinín and Uí Dhonnchadha, 1998) was created and

[3] The Irish English (Hiberno English) part of NCI contains 25 million words.

developed while the author worked as a research assistant in Institiúid Teangeolaíochta Éireann (ITÉ), during its participation in the EU funded LE-PAROLE project (1996-1999). When this project came to an end ITÉ continued to collect texts and in 2003 an enhanced version, Corpus Náisiúnta na Gaeilge (CNG) (ITÉ, 2003), was issued. It consisted of the Parole Corpus plus a small number of additional texts which had been processed at that time. The NCII incorporates Corpus Náisiúnta na Gaeilge and a large portion of the 20+ million words of unprocessed texts collected by ITÉ, together with other texts collected by Lexicography MasterClass Ltd., and texts harvested from the Internet by Infogistics Ltd. (Table 2).

Table 2 NCII Text Sources

Text Source	Words (millions)
ITÉ Corpus Náisiúnta na Gaeilge	8.7
ITÉ Other Texts	15.3
Lexicography MasterClass Ltd.	1.0
Infogistics Ltd.	5.0
TOTAL	**30.0**

Most publishers and typesetters who were contacted were willing to provide electronic versions of texts. The copyright owners (publisher or author) were contacted in order to obtain permission to use their texts in the new corpus.

The main Irish language book publishers - An Gúm, Cló Iar-Chonnachta and Coiscéim - were major contributors to the corpus. Many electronic texts were provided by the two main Irish typesetting companies Peanntrónaic Teo. and Evertype Teo. All of the Irish medium newspapers, *Foinse*, *LÁ* and *Anois* are well represented, as are two popular periodicals, *Feasta* and *Comhar*.

1.3 Morphosyntactic Annotations

The Parole Irish Morphosyntactic Description tagset was chosen (see Section 2.3.1 and Appendix A for further details) together with XML Corpus Encoding Standard (XCES) (Ide et al., 2000; Ide and Suderman, 2002) as the final delivery format for the NCI corpus. The Parole tags can be used in attribute-value pairs in XCES mark-up as shown in Figure 2. The attribute `base` is used to encode the lemma. The Parole tags can be truncated to bare POS tags, by using only the first two characters.

```
<w tag = "Ncfsg" base = "cathair">cathrach</w>
```

Figure 2 XCES Sample

A number of adjustments were made to the Parole tagset specified for Irish during the LE-PAROLE project (1996-1999), to facilitate consistent POS tagging. A new class was created for the copula *is* 'is' which was previously included under verbs, and also for verbal particles which were previously included in the Unique Membership Class. The verbal noun and verbal adjective, which were previously categorised under the verb category, are now classified under nouns and adjectives respectively, as they share features with these categories and appear in the same syntactic constructions. Further details of modifications can be found in Appendix A.

1.4 Text Pre-processing

The NCI texts (books, newspaper, magazines), both Irish and English, which were received from publishers and typesetters came in a variety of desktop publishing and word processor formats for both Mac and PC systems. Some books were received in several files (e.g. a file per chapter) and these were concatenated into one file in order to have one distinct header per text.

1.4.1 Conversion to Plain Text

All texts were transformed from proprietary formats into a uniform plain text format. This meant acquiring a copy of the software used to produce the original documents e.g. MS Word for Mac and PC, Quark for Mac and PC, PageMaker for Mac etc. (or the co-operation of someone who had the necessary software and hardware), in order to save the documents in a plain text format.

As Irish texts can have an acute accent on the five vowels, in both lowercase and uppercase, ISO-8859-1 character encoding was used for all texts. (The texts were later converted to UTF-8).

Where there were less than 20 of a particular type of text file, e.g. PDF, (mainly books), the conversion to plain text was carried out on each document individually using cut and paste commands. In other cases it was possible to convert batches of text files automatically using MS Word Visual Basic macros and also using the text extraction plug-in, TeXtractor, for Quark Xpress (Mac).

Before converting to plain text, a number of other text preparation tasks were carried out, such as removing tables of contents and indexes etc. This is a time-consuming task and three interns[4] worked full-time for six months during the text pre-processing phase of the

[4] Lisa Nic Sheáin, Dan Xu, Eamon Keegan, (CA3, 2004) School of Computing, Dublin City University.

project, under the supervision of the author, preparing the various texts either manually, or whenever possible by writing programs to carry out tasks automatically. As noted by Manning & Schütze (1999, p117) this is an important and often underestimated stage of corpus creation. The quality of the text at this stage has implications for all subsequent stages of tool development and will of course impact on the utility of the corpus to end users.

1.4.2 Removal of Front and End Matter in Books

In the corpus we only wish to include the chapter content, but the text files in general contained at least some ancillary text. Therefore, all "front matter" except for the first occurrence of the title, i.e. everything prior to the start of the first chapter/introduction/foreword etc. was removed. This included author, illustrator, designer, publication, copyright details and table of contents etc. All of these details were recorded in a header-details database (described in Section 1.5) before deletion from the text.

"End matter" i.e. everything after the main text, was also removed. This usually consisted of lists such as indexes, glossaries, word-lists, bibliographies and references.

1.4.3 Removal of Header and Footer Text

Page header and footer details were removed where they interrupt the flow of the body text. This is shown in Figure 3 for one English text file originally received in PDF format.

```
"Instinctively I started walking in the direction of the docks.
Ten minutes
THERE IS A TIME
10
later, I was sitting on a mooring bollard in the centre of the
dockyard overlooking the Shannon river. I tried to put my
confused thoughts in order but …" (Brandon/Duhan There Is A
Time)
```

Figure 3 Document Header Interrupting Body Text

1.4.4 Deletions in Body Text

All of the previous deletions of front and end matter are made without explicitly recording the fact in the text as they do not affect the body of the text. However, deletion of items within the body of the text such as tables, illustrations and formulas, long quotations in another language etc., whose deletion interrupt the flow of the text are explicitly recorded using the `<gap>` tag. Examples of the use of this tag are given in Table 3.

Table 3 Use of <gap> Tag

Tag and attribute/value pair	Type of material removed from text
<gap desc="table"/>	table of data
<gap desc="note"/>	footnotes and endnotes
<gap desc="bibl"/>	lists of authors and titles
<gap desc="formula"/>	mathematical formulas (textbooks mainly)
<gap desc="english"/>	sentences of text not in target language
<gap desc="glossary"/>	lists of words, etc., not in sentence form
<gap desc="contact_info"/>	contact details incl. name, addr., e-mail, phone & fax numbers etc.

Occasional foreign words embedded in a target language sentence are acceptable. However, entire paragraphs or sentences of text in a language other than the corpus language are removed and replaced with a <gap> tag. The primary goal is that texts should comprise of *complete sentences* in the target language. Only fragments such as titles or list items are allowed remain.

1.4.5 Poetry and Drama Mark-up

Poems, verses and songs are tagged semi-automatically using the poem <poem>, line group <lg> and line <l> mark-up as shown in Figure 4. The beginning and end of poems etc. were manually marked-up, and line and line-group tags were later inserted automatically.

```
<p>
<s> text </s>
</p>
<poem>
    <lg>
          <l>line 1</l>
          <l>line 2</l>
          <l>line 3</l>
    </lg>
</poem>
<p>
```

Figure 4 Poem Mark-Up

Dramas and plays are marked-up, as shown in Figure 5, using the spoken paragraph tag <sp> and speaker tag <speaker> along with the usual paragraph <p> and sentence <s> tags. This was also carried out semi-automatically in a manner similar to that used for poems.

```
<sp>
<speaker>SEÁN</speaker>
<p>
<s> speech here </s>
</p>
<stage> stage instructions here </stage>
</sp>
```

Figure 5 Drama Mark-Up

1.4.6 Clean-Up of Newspapers/Periodicals

Newspapers and periodicals contain many items such as crosswords, TV listings, names, addresses, dates, forms, advertisements, racing results, lists of team members etc. which are not suitable for inclusion in the corpus and which are removed. The <gap> tag was not used in these cases, as these were separate items which were not embedded in another unit of text.

Newspaper texts contain many hyphenated words due to the columnar format. These alter the word and impede linguistic analysis, e.g. if the word 'competition' appears as 'com-petition' or 'com-' 'petition' in the text, it will not be found in the lexicon. This problem was alleviated by generating a large list of words without hyphens from the corpus texts. Each time a hyphenated word is encountered in the text, the list was searched to see if the same word exists without the hyphen. If so, we assume that it was OK to remove the hyphen - otherwise the hyphenated form remains.

1.4.7 Clean-Up of Web Text

In order to eliminate recurring text in web pages, a list of frequently occurring button texts etc., was compiled and automatically removed from the web pages. Recurring headers and footers, and advertisements in newspapers and magazines/journals can also be dealt with in the same way.

1.5 Header Information Database

Information relating to each text was recorded in a web-based Php/MySQL database application when the text was first processed. In the case of books this information was entered manually, mainly from the front matter included with the text. At the end of automatic cleanup the number of words in the text was recorded in the database. Header information for issues of newspapers, journals and web pages was generated automatically from their filenames, and loaded into the database.

The web interface meant that the data could be easily maintained and added to, by different personnel at different locations. The data is exported from the database to generate XML headers (which are stored separately from the body text), and these headers can be quickly regenerated anytime the database has been updated.

Table 4 below, adapted from Table 4 in "Efficient corpus development for lexicography: building the New Corpus for Ireland" (Kilgarriff, Rundell and Uí Dhonnchadha, 2007), shows the type of header information stored in the database.

Table 4 Header Information Database

Feature	**Values**	**Note**
Docid	unique 8-character document ID	
Title	free text	
Author	free text	
Publisher	free text	
Pubplace	free text	Publication place
Pubdate	free text	Publication date
Author-birthplace	free text	Author place of birth
Author-DOB	free text	Author date of birth
Author-residence	free text	
Language	ga, en	ISO 639 language codes
Langvariety	ie br am	Hiberno/british/american: applies to english only
NativeSp	y, n, u	Native speaker; yes/no/unknown
NativeSpDialect	connacht, munster, ulster, u	Dialect area or unknown
Translation	y, n	
Time	1883-1959, 1960-1999, 2000-on, u	Publication year/unknown
Biographical	yes no auto	Applies to irish only; default is ‘no’
Mode	written, spoken	

Feature	Values	Note
Medium	book, newspaper, magazine, periodical, acad-journal, website-news, website-other, email-webchat, dissertation, official-govt, unpublished, ephemera, broadcast-radio, broadcast-tv, conversation, interview, lecture, meeting, unknown	Used in defining target proportions; several values (e.g. Email-webchat, dissertation) were unused.
Genre	inf, imag	Informative/imaginative
Genre2	fiction, poetry, drama, non-fiction, information, instruction, official, unknown	A more fine-grained genre classification is recorded where known.
Topic	hard-applied-science, social-science, govt, politics, history, religion-philosophy, business-finance, arts-culture, leisure, geography, health, news, legislation, unknown	
Target-readers	general, schools, academic, teenagers, children, adult-learner, unknown	

1.6 XML Corpus Encoding Standard (XCES) Mark-Up

After clean-up, texts were converted into XML files according to the XML Corpus Encoding Standard (XCES) and validated against the XCES DTD (Ide et al., 2000; Ide and Suderman, 2002). XCES is a member of the SGML family of mark-up standards.

These files contain standard XML header and body mark-up, including `<p>` paragraph tags and a reference to an external header file as well as the tags inserted as part of pre-processing.

Some symbols such as `&`, `<` and `>` have special meaning in XML and where they occur naturally in the text, they must be converted to entity references, i.e. `&amp;`, `&lt;` and `&gt;`. We also convert quotation marks to ".

1.7 Corpus Evaluation

Two aspects of the corpus were evaluated at this point. Firstly, all XML texts were checked to ensure that they were well formed and valid with respect to the XCES DTD. The XMLWriter program was used for the batch validation of XML files.

Secondly we evaluated how closely the quantities and categories of text collected and pre-processed matched the Corpus Design Targets. This was carried out quite straightforwardly by summarising the data in the header database. Targets were met in most cases, except for fiction (imaginative books) which proved to be the most difficult target to achieve (see Table 5 adapted from (Kilgarriff, Rundell and Uí Dhonnchadha, 2007)). This is due to a general lack of Irish fiction, as well as the fact that copyright clearance is usually more difficult to obtain for literary fiction than other categories of text.

Table 5 NCII: Targets vs. Actual Collection

Irish Text Category		*Words: target*		*Words: actual*	*Diff*
Books-informative	6,000,000		8,400,000		+1.4
Books-imaginative	9,000,000		7,600,000		- 1.5
Books total		15,000,000		16.000,000	
Newspapers	4,500,000		4,500,000		0
Periodicals	2,500,000		2,600,000		+1.0
News+Per. total		7,000,000		7,100.000	+1.0
Official/Govt		1,500,000		1,200,000	- 0.8
Broadcast		1,000,000		400,000	- 0.4
Websites		5,500,000		5,500,000	0
TOTAL		**30,000,000**		**30,200,000**	**+1.0**

1.8 Summary

In this chapter, we gave a brief overview of the NCII corpus in terms of types, quantities and sources of texts, as well as an overview of the chosen POS tagset (Parole) and text encoding format (XCES). We also described the essential but laborious task of text pre-processing, and the creation of a text header information database.

At this point in the development of the NCII Corpus we have a corpus of clean texts with headers and body structure marked-up but no linguistic annotation as yet.

In the next chapter, we present background information on linguistic annotation of corpora. In particular, we present techniques for POS tagging and partial syntactic parsing. In addition, we introduce the methodology used for POS tagging and partial syntactic parsing for Irish.

2 Linguistic Annotation of Corpora

2.1 Introduction

A corpus of raw texts is a very useful repository of information about a language. We can automatically extract lists of words, collocations etc., and we can compute information about the relative frequencies of words. We can also identify regular patterns, suggesting possible prefixes and suffixes. But, raw texts do not explicitly encode information about the function of individual words, or how they are related to each other, either in morphological paradigms or in syntactic phrases.

Quite often, if we wish to know how a word is used, we are interested not just in one particular form of the word, but in all inflected forms of the word. In order to be able to do this we must associate each word with the canonical form (lemma) representing its paradigm. This is known as lemmatisation. Alternatively, we may be interested in a word when it is functioning as a noun but not as a verb, or vice versa. In order to do this, we must associate the appropriate part-of-speech (POS) category with each word form. In order to study certain linguistic phenomena, or to automatically extract a grammar, we need more detailed information about phrases, constituents and the hierarchical structure of a sentence. For translation purposes, or for information extraction, it is important to know the grammatical functions of words. Identifying the phrases and constituents in a sentence is known as partial parsing (or, shallow parsing or chunking), whereas deep parsing requires the full hierarchical structure of the sentence to be specified. In this thesis we describe partial parsing (deep parsing is beyond the scope of the current work).

As well as the types of linguistic annotation mentioned above, corpus annotation usually involves mark-up of the structure of the texts. A variety of tags are used to indicate section, paragraph, and sentence boundaries as well as to identify text fragments such as titles, captions, formulae etc. It is customary in a corpus to include header details, giving information about the provenance and type of each of the texts.

All of this mark-up entails inserting extra information into the texts. In order that this can be searched efficiently, and interpreted automatically by specialised software, it must follow a particular formatting standard. The current work uses XCES (XML Corpus Encoding Standard), to encode the following types of information:

Text Header:

- Metadata about the text, e.g. title, author, date etc. (see Table 4, p31)

Text Body:

- Structural mark-up
    - paragraph, sentence boundaries
    - title, chapter, section etc.
    - drama and poetry mark-up
- Linguistic annotation at word-level
    - POS tags and morphosyntactic features
    - lemmas
- Linguistic annotation at sentence-level
    - clause boundaries
    - grammatical functions
    - phrase/chunk boundaries

In Section 2.2, we discuss the motivation for linguistic annotation in corpora. In Section 2.3, we describe the main techniques for POS tagging, followed in Section 2.4 by our method of POS tagging for Irish. In Section 2.5, we describe the main techniques for parsing, followed in Section 2.6 by our method of partial parsing for Irish. In Section 2.7, we highlight some recent research in the area of Irish Natural Language Processing. Finally in Section 2.8 we present a worked example of the linguistic annotations produced as a result of the research presented in this dissertation.

2.2 Linguistic Annotation

There are a great variety of linguistic phenomena that can be annotated in corpora to aid empirical linguistic analysis and NLP development, e.g. phonetics, prosody, part-of-speech, syntactic structure, semantics, anaphora, appositions, discourse markers etc. The methodologies involved range from mainly automatic (e.g. POS tagging), to mainly manual (e.g. anaphora resolution). In this thesis we focus on two types of automated linguistic annotation: POS tagging and partial parsing.

2.3 Techniques for Part-of-Speech Tagging

Part-of-speech (POS) tagging consists of assigning the appropriate part-of-speech category to each token in a corpus of text (which can be written text or transcribed spoken language). The major part-of-speech categories are noun, verb, adjective, pronoun, adverb, conjunction, preposition, determiner and article, as well as other functional items such as particles, numerals and punctuation. The exact set used will vary from language to language. In addition to the basic POS category, other morphological information such as number, gender, case, tense, aspect etc. is usually encoded.

POS tagged text makes some of the inherent structure in language available to us without needing to understand or encode the full syntactic hierarchy or semantic content (Manning and Schütze, 1999, p341). POS tagged text can be used in both practical applications and theoretical research and is an intermediate step towards full parsing.

- POS tagged corpora are widely used in dictionary compilation (lexicography) and the development of reference grammars.
- Some Machine Translation systems incorporate POS tagging in their analysis of source and target languages.
- In speech processing, knowing the underlying POS category of tokens is an aid to prosodic modelling in speech synthesis; POS tagging is also an aid in automatic speech recognition (ASR).
- Many branches of linguistics make use of POS tagged text (especially where fully parsed text is not available), e.g. in the areas of syntactic analysis, discourse analysis and child language acquisition etc.
- Clinical studies of language can make use of tagged data to compare normal and abnormal language acquisition and production.
- In language pedagogy, POS tagged text can be used for error analysis and correction purposes.
- In literature studies, tagged texts can be used to find and analyse stylistic/cultural/dialectal differences etc. in the texts under consideration.

2.3.1 Annotation Schemes

In order to make explicit the linguistic structure of a text, a standard set of annotations must be devised. This is commonly known as a tag set. A tag set can be described in terms of granularity; the more detail encoded, the finer the granularity; and conversely, less detail means a coarser granularity. The type of tag set used will depend on both the morphology of the language in question and the intended application of the tagged data.

The first part-of-speech tagged corpus for English was the Brown Corpus which was created in the 1960's (Kuĉera and Francis, 1967), and which used a tag set of 87 tags of the type shown in Table 6. This provided the pattern for many later tagsets for English, e.g. the Penn Treebank tag set, (45 tags) and Claws C5 (62 tags) used on the BNC (Leech et al., 1994), Susanne (Sampson, 1993) and LOB corpora (Johansson, 1986).

Table 6 Sample of Brown Tags

Tag	Description
NN	noun singular
NNS	noun plural
NPS	noun proper plural
JJ	adjective
VB	verb - present
VBD	verb - past

These first tagsets were designed specifically for English, a language with limited morphological forms; therefore, they do not include gender, case or many of the verb forms found in other (European) languages. A more complete set of morphological descriptions was developed for the European funded LE-PAROLE project (1996-1999), which covered 14 European languages including Irish. This tag set incorporated Multext (1996) and EAGLES (1996) recommendations. Some examples of the 350 (approx.) Parole tags applied to Irish are given in Table 7 below. These tags are used as an output format only, the fuller morphologcal descriptions (Appendix B) are used during processing. The mapping between the full morphologcal descriptions and the output Parole tags is given in Appendix H.

Table 7 Sample of Parole Tags for Irish

Tag	Description
Ncfsg	noun common feminine singular genitive
Ncmpc-e	noun common masculine plural common case emphatic
Pp1-s-e	pronoun personal 1st person singular emphatic
Vmc-2p-d	verb main conditional 2nd person plural dependent
Aqafsc	adjective qualifying attributive feminine singular common case

In Table 8 we have a sample of the truncated Parole tags used in the evaluation of POS tagging (Section 6.6). A description of the Parole Tagset, as well as the complete shortened tagset (39 tags) may be found in Appendix A.

Table 8 Sample of Parole Short Tags for Irish

Tag	Description
Nc	common noun
Pp	personal pronoun
Vm	main verb
Aq	adjective

2.3.2 POS Tagging Methodologies

In natural languages, it is common for a word to have a number of possible part-of-speech categories depending on its context. For example, in (9) the word 'chair' is functioning as a verb, whereas in (10) it is functioning as a noun. The challenge in POS tagging is to choose the correct POS tag for the context in which the word is being used, wherever possible. Given a corpus of millions of words, manual labelling of text would be very time-consuming and error-prone.

(9) I will chair the meeting.

(10) Where is the chair?

In order to develop an automatic tagger, we must provide the system with information about the language. This can either be provided explicitly in the form of rules, or implicitly in the form of manually tagged text from which rules can be automatically derived, or some combination of both. Automatic tagging methodologies fall into three broad categories:

- Rule based taggers
- Statistical taggers
- Transformation taggers

2.3.3 Rule-Based Taggers

Rule-based POS tagging is a two-stage process. In the first stage text is tokenized, i.e. segmented into units for analysis, and each token is tagged with all possible POS tags using a lexicon or a morphological analyser. A wide-coverage morphological analyser (or lexicon) is required to provide analyses for all of the tokens, in the first stage. In the second stage, hand-crafted linguistically motivated rules are developed which seek to select the most appropriate tag or to eliminate inappropriate tags, ideally leaving the one correct POS tag for the token appropriate to the context. The tagger applies the rules to texts and the results are evaluated, usually, by comparison with a gold standard. Problem areas can easily be targeted and the rules can be amended and added to.

Rule-based systems can be developed in an incremental fashion and this approach is often used where there is a lack of pre-existing linguistic resources for a language (e.g. a reliable POS-tagged training corpus) and where there are limited financial and human resources. Examples of rule-based taggers include Taggit (Kučera and Francis, 1967) which was used to tag the Brown Corpus, and EngCG using Constraint Grammar (Karlsson, 1995) which was

used to tag the Bank of English[5] (COBUILD) corpus. The Brill Transformation Tagger (Brill, 1995a) uses rules and the Xerox POS Tagger (Cutting et al., 1992) uses rules to a lesser extent for lexical and transition biases. Both EngCG and the Xerox Tagger achieve accuracy levels of over 97%.

The main drawback with rule-based tagging is the difficulty of manually producing all the rules necessary to describe a natural language. A human rule-generator although benefiting from linguistic knowledge and intuition, cannot operate on the scale or with the speed and consistency of an automatic rule-generator.

2.3.4 Statistical Taggers

Supervised, machine-learning-based statistical taggers require a substantial amount of accurately tagged training data as raw material. They estimate the probabilities of tags in new texts, based on frequency data observed in a manually tagged training corpus. This is possible because although a word may have more than one possible POS category, they are not all equally likely to occur, particularly when local context is taken into account. In general texts, *chair* occurs more frequently as a noun than as a verb, (however this may not be the case in some domain specific texts, e.g. minutes of meetings). Word/tag frequency counts are known as unigrams. It is estimated that by simply always assigning a word its most frequent tag (for English), the overall result will be 90%[6] correct (Manning and Schütze, 1999, p344). This may sound impressive, but taking an average of twenty words per sentence, it means that on average every sentence in the corpus could contain two errors.

As well as unigrams (probabilities of individual word/tag combinations), bigrams and trigrams are often used. With bigrams or trigrams, the probability of a word/tag pair occurring is conditioned on the previous one or two tags (or surrounding tags) in the text.

The main drawbacks of statistical taggers are that:

- A large amount of manually-tagged training data is required to train a tagger in a new language (anecdotally, 50K words minimum).
- Results can be inconsistent; the tagger will perform well on text similar to the training text but could perform quite poorly in a different domain. Rule-based taggers by their nature tend to be more corpus independent.

[5] See http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html for details (last accessed 30 June 2008).

[6] In comparison the early rule-based Taggit Tagger achieved 70% accuracy (before manual correction).

- Statistical taggers are prone to over-training, as it is not clear when iterative training should stop – the aim is to extract the generalities of the language without incorporating specific characteristics of the texts which happen to be in the training corpus – and there is no clear way of knowing when this point is reached.
- Dealing with unknown words in a text (i.e. items not encountered in training data) is a problem. Various smoothing techniques are employed to avoid items which were not seen in the training data from being assigned zero probability.
- When there is an obvious problem in the tagged text, in a purely statistical tagger there is no direct way of intervening to correct it other than by adding more manually tagged text and re-training the tagger.
- When using bigrams, trigrams and higher n-grams it can be difficult to calculate reliable probabilities due to sparseness of data (but see Brants and Franz (2006) for web-extracted n-grams).

However, most of these difficulties have been addressed in the most successful taggers, by using sophisticated statistical and stochastic techniques in conjunction with large amounts of varied training data, resulting in over 97% accuracy being achieved.[7]

2.3.5 Transformation Based Taggers

A third type of tagger, a transformation based tagger, i.e. the Brill Tagger (Brill, 1995a), combines elements of the two previous approaches. The training phase uses both a training corpus, and a lexicon which specifies the most frequent tag for each item (usually generated from another tagged corpus), in order to iteratively generate rules rather than probabilities. Although statistics are used in the training, the resulting tagger is rule based.

A small quantity of training data is sufficient to begin training. The tagger takes the raw text from a manually tagged or corrected corpus and automatically tags it with the most frequent tag in the lexicon. Unknown words are tagged with the most likely tag; the default setting being singular common noun. The tagger then compares the results with the manually tagged data and builds up a set of transformation rules which will result in the automatically tagged text being as close as possible to the manually tagged text. The tagger generates both contextual and lexical rules. The contextual rules arise from examining the context surrounding the tokens that were incorrectly tagged and looking for statistical patterns in the errors. The lexical rules arise from looking at individual tokens and discerning prefixes and/or suffixes which appear to correlate to specific tags. The lexical transformation rules are

[7] See (CLAWS) (Garside, 1987; Garside, 1995), Xerox POS tagger (Cutting et al., 1992) Maximum Entropy tagger (Ratnaparkhi, 1996) etc. for details of some popular statistical taggers.

particularly useful for dealing with unknown lexical items. Both types of rule can be manually inspected and corrected or augmented. Training is carried out incrementally, by manually correcting the automatically tagged output at each stage and then adding this new text to the training corpus.

This tagging methodology is quite successful, and has all the advantages of using probabilities while at the same time incorporating linguistic (and human readable) rules. Its main drawbacks are that a) it requires a training corpus and a lexicon which includes tag frequency information, and b) the format of the rules is not very flexible or user-friendly. Unless there is a large body of accurately tagged data for training, this tagger is, therefore, more amenable for use with a simple tag set (e.g. Penn, Claws etc.) than with the type of detailed tag set which is necessary for languages with richer inflectional morphologies.

2.3.6 Unsupervised POS Tagging

Tagger training which requires a tagged corpus is known as supervised training. Alternatively, training can be carried out using an untagged corpus and a lexicon which specifies the alternative POS options for each word. In this method, known as unsupervised training, the quality of the lexicon is of great importance (Banko and Moore, 2004; Brill, 1995b).

2.4 Part-of-Speech Tagging of Irish

Tagging methodologies, as already mentioned, vary from statistical systems to linguistic theory-driven rule-based systems, as well as systems which use various combinations of the two. All three types of tagger (described in Section 2.3) are capable of performing equally well, given the appropriate circumstances and resources.

For the research reported in this thesis, statistical tagging was not a realistic possibility as no tagged text was available on which to train the tagger. Some experiments were carried out using the Brill transformational tagger. Progress in training the tagger was slow. A precision of 85% was easily achieved through use of the lexicon. Through training this gradually increased to 89%, but further progress was difficult due to limited training data, resulting in a data sparseness problem given the size of the tag set (350+ Parole tags).

Improvements were made to the tagger by manually editing the automatically generated rules, to add useful rules and remove inappropriate rules. However, as there was no way of generalising rules over sets of tags, this solution was cumbersome and inelegant. For example, a rule relating to all possible nouns would have to be repeated to account for all possible configurations of noun tag, i.e. to accommodate all number, gender and case

combinations. A further disadvantage was that our lexicon contained no information about relative frequency of tags for each lexeme, which is a requirement of the Brill tagger.

For these reasons, a two stage approach to tagging, consisting of Finite-State Morphological Analysis followed by Constraint Grammar Disambiguation, was examined and adopted. This rule based approach exploits the output of the morphological analyser, and provides a framework within which progress is incremental and can be easily measured.

There are several advantages to using this tagging methodology. Firstly, the computational efficiency of finite-state processing is used in the morphological analysis stage. Also, part-of-speech tags and lemmas are assigned in one integrated step using the two-level morphological analyser.

Furthermore, full morphological descriptions are used during all processing stages, and only as a final step are these transformed into (condensed) POS tags of choice, in this case Parole tags. Similarly, other mappings to BNC tags, Penn tags, Childes tags etc. could be created if desired.

In addition, during disambiguation, Constraint Grammar allows one to leave some ambiguities unresolved if it is not possible to make a safe choice, i.e. the system does not force one to make a final choice. Multiple tags (and lemmas) can be accommodated in XCES (Ide et al., 2000; Ide and Suderman, 2002). For example, in (11) there are three possible POS analyses for the token *an* 'the', i.e. article, verb particle or copula, and two possible lemmas *an* or *is*. This ambiguity can be encoded in an XCES word tag `<w>` tag as shown in (12). `Art Sg Def` is mapped to `Td-s`, `Part Vb Q Pres` is mapped to `Qq`, and `Cop Pres Q` is mapped to `Wp-q` in the Parole tag set. There is, however, a loss of information in this representation, as we have not specified which lemma (base) is associated with which POS tag. Full details of both tagsets may be found in Appendices A and B).

```
(11)  "<an>"
            "an" Art Sg Def
            "an" Part Vb Q Pres
            "is" Cop Pres Q
```

```
(12)  <w tag = "Td-s|Qq|Wp-q" base = "an|is">an</w>
```

Successful Constraint Grammar implementations of POS tagging exist for English, Finnish, Danish, Portuguese, German, French, Spanish and others are in development. In this thesis the application of the Constraint Grammar tagging methodology (Karlsson, 1995;

Tapanainen, 1996) to Irish will be described. The advantages of this methodology for POS tagging are:

- It does not require a manually tagged training corpus (such a corpus was not available for Irish).
- As CG relies heavily on the lexical and morphological features, it builds on existing work for Irish, i.e. finite-state tokenization and two-level finite-state morphological analysis (Uí Dhonnchadha et al., 2005).
- There is a freely available source code implementation for Constraint Grammar.[8]
- It is capable of producing results comparable to both statistical (Chanod and Tapanainen, 1995a) and transformation-based taggers

2.5 Techniques for Syntactic Parsing

POS tagging is concerned with words. The next level of corpus annotation, partial parsing, is concerned with larger syntactic units such as phrases, clauses and sentences. Parsing involves assigning a syntactic analysis to a sentence according to some grammar. A computer program which carries out parsing is known as a parser. Automatic parsing of natural language is a more complex task than POS tagging, as it must deal with a greater number of structures, i.e. not just words, but also phrases and clauses (Meyer, 2002, p93).

Parsers, like POS taggers, can be categorised as *rule-based* e.g. EngCG (Voutilainen et al., 1992), FDG (Tapanainen and Järvinen, 1997) or *probability based* parsers e.g. Fidditch (Hindle, 1993) or a combination of both. Parsers can also be categorised according to whether they generate *partial* parses or *full* parses, i.e. whether they generate a full hierarchical syntactic structure or not. Systems which produce unattached phrases (or chunks) (Abney, 1996b) or annotated tokens only (Karlsson et al., 1995), are types of partial parsers, whereas parsers such as Fidditch (Hindle, 1993), or FDG (Tapanainen and Järvinen, 1997) produce a full syntactic parse. In addition, parsers can be described as *shallow* or *deep* depending on how detailed their syntactic annotation is. The Penn I Treebank (Marcus et al., 1993) is an example of shallow syntactic analysis as it contains only 'skeletal' constituency markup. This markup was later supplemented with functional categories giving a deeper analysis, in the Penn II Treebank (Marcus et al., 1994).

Nivre (2006, Ch.2) makes an important distinction between *grammar parsing* and *text parsing*. According to Nivre "grammar parsing is an abstract problem, which can be studied

[8] See http://sourceforge.net (last accessed October 2006) a publicly available version of Constraint Grammar developed by the VISL project at Syddansk Universitet, Denmark.

using formal methods and internal evaluation criteria, while text parsing is an empirical problem, where formal methods need to be combined with experimental methods and external evaluation criteria". In grammar parsing, there is a formal grammar of natural language syntax. Formal grammars range from early transformational grammars to more recent frameworks such as LFG or HPSG. With formal grammars, only sentences which are part of the language defined by the grammar receive an analysis. In text parsing, there are no assumptions about the syntactic completeness of a sentence, therefore there is no formal grammar defining the language. The text parsing problem requires the mapping from input language to syntactic representation, where a well-defined abstract problem is used as an approximation for the real text parsing problem. While research into grammar parsing has been carried out for Irish (Carnie and Guilfoyle, 2000; Duffield, 1995; McCloskey, 1979; Stenson, 1981) to our knowledge no such research exists in the area of text parsing.

Text parsing may be either grammar-driven or data-driven. In a grammar-driven approach, sentences are analysed by "constructing" a syntactic representation in accordance with the rules of the grammar. Alternatively, in an "eliminative" parser, a syntactic analysis which violates any of a set of constraints, is rejected, e.g. Constraint Grammar. In a data-driven approach, the mapping is induced from a body of pre-analysed texts (e.g. a Treebank) which are used to propose analyses for new sentences.

There are two main aspects to syntactic analysis: constituent structure and relational dependency structure (Van Valin, 2001, p4). This has led to two traditions in syntactic analysis: constituency analysis (or phrase-structure analysis) and dependency analysis. Constituency analysis defines the groups of words (or word types) in a sentence which form a single unit or phrase, while relational structure looks at the dependencies between pairs of words in a sentence. The main difference between constituent structure and dependency structure is that dependency structure has no phrasal nodes (Nivre, 2007, p2), i.e. apart from a root node all nodes are terminal nodes.

In a constituency-based phrase-structure analysis of language, the focus is on the syntactic structure of language, which according to generative theories can be studied independently of meaning, as the "Colorless green ideas sleep furiously" example (Chomsky, 1957, p15), seeks to demonstrate. We can judge this sentence to be syntactically well-formed, but semantically meaningless. Mel'čuk (1988), a proponent of dependency analysis, characterises this methodology as "generate structures first, and ask questions about meaning later".

In a dependency-based analysis, there is a closer relationship between syntax and semantics. Relationships between pairs of words in a sentence are represented in terms of predicate-argument relationships, or head-modifier relationships. This use of lexical

dependencies is an important aid to parsing (Jurafsky and Martin, 2000, p463). Dependency analysis is independent of word-order, unlike constituency-based analysis which is more heavily reliant on word-order.

There are a number of syntactic theories based on dependency analysis, including Meaning Text Model (MTM) (Mel'čuk, 1988), Relational Grammar (Perlmutter and Rosen, 1984), Word Grammar (Hudson, 2007), and Lexicase (Starosta, 1988).

Many phrase-structure syntactic oriented theories also include dependency and/or functional relations. In LFG (Bresnan, 2001), grammatical roles are encoded in f-structures (functional structures), in HPSG (Pollard and Sag, 1994) subcategorisation frames are used, in Government-Binding Theory (Chomsky, 1988) theta roles are used, and in Case Grammar (Cook, 1989; Fillmore, 1968) semantic dependencies are used.

A fully syntactically parsed corpus is known as a treebank. There are several treebanks in existence, some of which have a constituency based mark-up, e.g. Penn I Treebank (Marcus et al., 1993), others a dependency relation based mark-up e.g. Prague Dependency Treebank (Hajič, 1998), and others which use a hybrid approach, e.g. NEGRA Treebank (Brants et al., 2003) and Penn II Treebank (Marcus et al., 1994). Treebanks can be used to train parsers, to automatically extract grammars and to test linguistic hypotheses.

In this dissertation, we describe the work carried out on partial parsing of Irish using rule-based dependency analysis. This includes the annotation of grammatical functions and unlabelled dependency relations. We follow this by bracketing together heads and their dependants into phrase-like units known as chunks. The generation of a full hierarchical structure (a parse tree) for a sentence is beyond the scope of the current work.

In the following sections we take a closer look at constituent structure, chunking, dependency relations, grammatical functions, predicate-argument dependencies and head-modifier dependencies.

2.5.1 Constituent Structure

All languages have some scope for varying the word-order in a sentence, often to emphasise or focus in on some part of the statement. For instance, we could focus 'the table' in (13)a by saying 'On the table , the man put a book', as shown in (13)b. In (13)c. (13)a has been relativized as an NP.

(13) a. The man put a book on the table.
b. On the table, the man put a book.
c. The table, the man put a book on.

However, words do not always move independently of one another. There are groups of words which must be moved (or replaced) as a unit. These units are known as constituents. Constituent structure is, usually, determined by evidence from movement, replacement and omission tests, together with grammaticality judgements.

In (13), we cannot simply move 'table' on its own to the front as in (14), as 'table' cannot be separated from its determiner 'the'. In general, we must also move the preposition 'on' in order to preserve the meaning. This is because the noun phrase (NP) 'the table' is part of a larger prepositional phrase (PP) as shown in Figure 6.

There are some instances where the locative preposition can be left behind, as in (13), but only in limited contexts, e.g. in answer to a question or as part of a larger relative clause structure.

(14) *Table the man put a book on the.

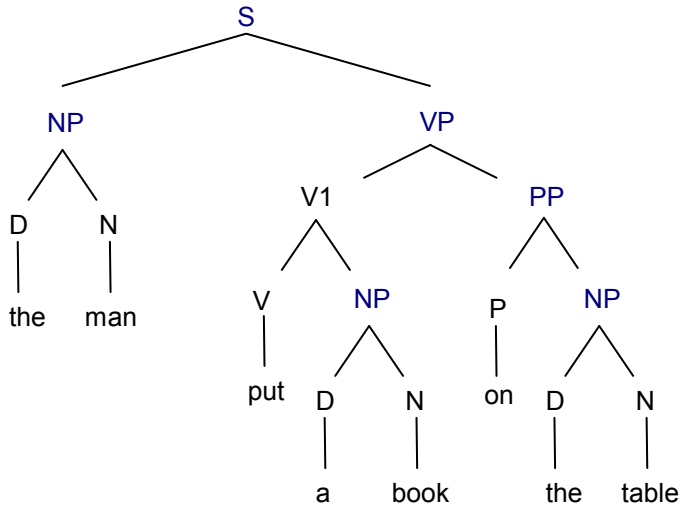


Figure 6 Constituency Hierarchy

The most common method of modelling constituent structure is the Context Free Grammar (CFG) or Phrase Structure Grammar, which consists of production rules (Jurafsky and Martin, 2000, p236). These rules encode immediate dominance and linear precedence. The sentence in Figure 6 can be derived using rules of the following type:

(15) S -> NP VP
PP -> P NP
NP -> D N

In a constituency based mark-up, the words which are grouped together to form constituents (or phrases) and the hierarchical relationship between these phrases in the sentence are represented as a tree. This tree can be linearized by bracketing. The brackets usually

include phrase labels which apply to non-terminals, as in (16), and the nesting of brackets shows the hierarchical structure of the sentence.

(16) [S[NP The man] [VP put [NP a book] [PP on [NP the table]]]]

2.5.2 Chunking

Some of the most challenging aspects of full parsing include prepositional phrase attachment (17) and coordination (18). In automatically parsed text, these items usually require manual checking and correction. This manual intervention can take place, either before parsing as in the ICE-GB corpus (Wallis, 2003) or after as in the case of the Penn Treebank (Taylor et al., 2003).

(17) [John] [killed] [the man with a gun] OR
[John] [killed] [the man] [with a gun]

(18) [John] [ate] [with a [small [fork and spoon]]] OR
[John] [ate] [with a [small fork] and [spoon]]

The following example, (19), given by Meyer (2002, p95) illustrates the difficulties a parser faces with some types of co-ordination. In this example, it is difficult for a parser to decide whether 'wrist' should be co-ordinated with 'arm' or 'mother'.

(19) The boy broke his arm and his wrist and his mother called the doctor.

Other problems which parsers commonly experience in practice, include failing to find a global parse of the sentence or finding too many parses.

Because of these difficulties, Abney (1991) proposed a method known as chunking, in which he splits parsing into two distinct phases. In the first phase, the sentence is divided into chunks which are similar to phrases, but which do not have recursion or embedding (except for NPs embedded in PPs) (20). Typically, they consist of a “single content word surrounded by a constellation of function words, matching a fixed template” (Abney, 1991). Adjectives are included in NP chunks.

(20) [NP The man] [V put] [NP a book] [PP on [NP the table]]

In the second phase, an “attacher” module links the chunks by inserting the nodes required to create the syntactic hierarchy. It also deals with any items that were not included in a chunk in the first phase.

When the hierarchical relationships between the chunks are not fully specified i.e. brackets are not nested recursively to give the full phrase-structure tree, we get a flat structure, as in Figure 7 (as opposed to the more hierarchical diagram in Figure 6).

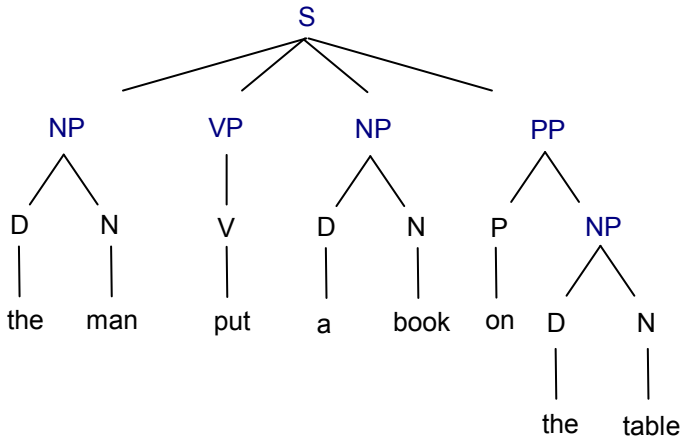


Figure 7 Flat Constituency Structure

As constituency based mark-up deals in general with contiguous items, handling discontinuous constituents presents a challenge. In the next section we look at dependency analysis where non-contiguous elements are not an issue.

2.5.3 Dependency Relations

Rather than bracketing consecutive elements of a sentence as constituents, we can annotate dependency relations between pairs of words, which need not be adjacent. This methodology is attributed originally to Tesnière (1959), and although applicable to any language, it is particularly useful for dealing with free word order languages and for discontinuous constituents. A constituency analysis can be inferred from dependency relations using head-modifier information, whereas, the opposite is not always the case, e.g. if the constituency analysis contains no grammatical function information.

To date there exists a rich variety of dependency-based linguistic formalisms and despite general similarities, there is no general agreement as to the analyses and terminology used in the various approaches. Below we chart some of the main approaches.

In a dependency analysis, two or more elements within a sentence may be related in such a way that one element is dependent on the other. The dominant element is known as the "head" and the other element or elements will be "dependants" of the head (Van Valin, 2001, p87). In Figure 8 we give a dependency representation of example (20) with an additional optional adverbial adjunct 'yesterday'. In this representation, we show dependent words pointing to their heads (the opposite notation can also be used).

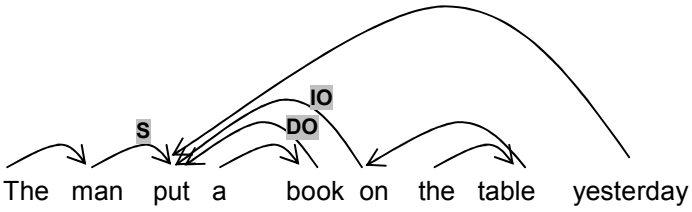


Figure 8 Dependency Representation

Figure 8 shows both labelled and unlabelled dependencies, with the arguments of the verb being labelled with their grammatical functions (i.e. subject (S), direct object (DO) and indirect object (IO)). These labelled dependencies can also be termed predicate-argument dependencies. The remaining dependencies are unlabelled.

2.5.4 Grammatical Functions

Grammatical functions, which include subject, direct object and indirect object, label the relationships which exist between the predicate and the various noun phrases in the clause. In Figure 8, 'man' is the subject and 'book' is the direct object of 'put'. The indirect object (preceded by the preposition 'on') is 'table'.

In addition to bracketing and phrase labels, a constituency based annotation may also include functional information. Example (21) shows how (16) (page 48) can be represented using Penn II Treebank notation (Bies et al., 1995).

```
(21)  (S
        (NP-SBJ The man  )
        (VP  put
            (NP a book)
            (PP-LOC on
                 (NP the table ) ) )
```

2.5.5 Predicate-Argument Dependencies

The term "predicate" is used to refer to the element of a clause which names an action, event or state (Kroeger, 2004, p7). This is usually a verb, but can also be a noun or adjective. The argument structure describes the number and type of roles, (i.e. valency) required by the predicate. Additional *optional* pieces of information in a clause are known as adjuncts.

A typical dependency is that which holds between a verb and its arguments. A verb can be described as being a 1-place, 2-place or 3-place predicate, i.e. requiring one, two or three arguments respectively. Some predicates such as 'bet' and 'insure' can take four and five arguments. These arguments are often described in terms of their grammatical (or semantic) function, i.e. subject, direct object and indirect object, and the verb is described as being intransitive (requires subject only), transitive (requires subject and direct object) or ditransitive (requires subject, direct object and indirect object). The predicate, whether verbal, nominal or adjectival, is considered to be the head, as it determines the number of arguments and often the characteristics of those arguments (e.g. whether animate or inanimate etc.).

The subject and object are referred to as "direct arguments" or "terms", while any other arguments are considered to be "indirect" or "oblique" arguments. In English, the direct arguments are usually noun phrases whereas the indirect or oblique arguments are often prepositional phrases (Kroeger, 2004, p15).

In (22), the predicate of the clause (in this case a simple sentence) is the verb 'put'. This requires three arguments; the instigator of the action, the object in question and the location i.e. where the object was put. In (22)b we have an optional extra piece of information, 'yesterday', i.e. the time at which the action was carried out . This is an adverbial adjunct of time.

(22) a. The man put a book on the table.
b. The man put a book on the table yesterday

2.5.6 Head-Modifier Dependencies

Another type of dependency is that between modifier and modified. The modifier is dependent on the element which it modifies, e.g. a determiner is dependent on the noun which it modifies. In a prepositional phrase the object NP is dependent on the preposition.

Dependencies occur within particular syntactic constructs such as clauses or phrases. A sentence consists of at least one clause, where a clause is defined as containing a finite verb or copula. Table 9 based on (Van Valin, 2001, p87) gives a summary of the main dependency relations and the syntactic constructs in which they are found.

Table 9 Dependency Relations

Head	Dependant	Syntactic Construct
Verb	Terms	Clause
Preposition	Object NP	Prepositional Phrase
Noun	Modifier(s)	Noun Phrase
Possessed Noun	Possessor NP	Noun Phrase

2.6 Partial Parsing of Irish

As initial steps towards parsing of Irish, we extend both processing paradigms already in use for POS tagging, i.e. Finite-State Methods and Constraint Grammar, in the following ways:

a) We automatically annotate clause boundaries, grammatical functions and dependency relations using Constraint Grammar.
b) We use finite-state transducers to introduce chunk boundaries.

There are a number of advantages of this approach. Firstly, we provide continuity with existing paradigms, i.e. Finite-State Methods and Constraint Grammar. Secondly, and more importantly, using Dependency Analysis allows us to bypass the theoretical questions concerning the VP and other syntactic structures in Irish, and concentrate on function rather than form.

2.6.1 Dependency Relations and Functional Mark-Up

As shown in (23), the Constraint Grammar formalism facilitates the annotation of grammatical functions to individual tokens (e.g. `@SUBJ` meaning that this item is the subject of the sentence) and unlabelled dependency relations (e.g. `@>N` meaning that the item is dependent on the following noun).

```
(23)  The          Det    @>N          Dependent on noun to the right
      man          N      @SUBJ        Subject of the clause
      put          V      @FMV         Finite main verb
      a            Det    @>N          Dependent on a noun to the right
      book         N      @OBJ         Object of the clause
      on           P      @PP_ADVL     Prepositional phrase: adverbial
      the          Det    @>N          Dependent on noun to the right
      table N      @P<                 Dependent on a prep. to the left
```

Constraint Grammar dependency mark-up has been criticised (Järvinen, 2003) in that it does not specify explicitly where the head of a dependency relation is, only whether it is to the right or to the left. For example `"the" DET @>N` specifies that the determiner 'the' is dependent on a noun to the right, but it does not tell us explicitly which noun. (In our implementation for Irish it will always be the first available head.) This issue is addressed in the Functional Dependency Grammar (FDG) parser developed by Tapanainen and Järvinen (1997). Bick (2006) addresses this issue by using an attacher module with modified Constraint Grammar output. Numbered tokens is also a feature of CG3, the latest implementation of VISLCG. We have not addressed this issue in the current work, but prefer to deal with it at a later stage by moving from the CG2 (currently used) to CG3,[9] in conjunction with the development of subcategorisation frames for Irish, which are necessary for PP-attachment.

2.6.2 Chunking

Abney (1991) defines chunks in terms of *major* heads, e.g. in 'the bald man', 'man' is the head even though 'bald' is also a content word. It is, in fact, a semantic head (s-head) rather than a syntactic head, according to Abney (1991), who also suggests (in the same paper) that chunks coincide to a large degree with prosodic phrases.

In the implementation presented in this dissertation, chunks consist of content words and their associated functional items. We have nested chunks but no recursion in the sense that a chunk never embeds a chunk of the same type.

The chunks are labelled as NP, V or VS, COP, PP, AD, PRED etc. We follow Abney (1991) in including adjectives and determiners in our definition of NP. We have not implemented a DP (determiner phrase) analysis of NPs, as there is no surface representation of the indefinite determiner in Irish, and to date we have not posited any abstract categories in our annotations. There is, however, no reason why this work could not be extended in the future to insert this and other abstract categories as required.

We have avoided using the chunk label VP as this is assumed to mean a verb and its object, which is inappropriate for Irish being a VSO language, (for a different view see McCloskey (1983)). As Stenson (1981) states, there is a greater association between verb and subject in Irish than between verb and object. Instead we use V (or VS where the verb has an incorporated subject), which consists of the verb and any dependent particles, but does not have an embedded object NP due to the VSO surface word order configuration of Irish.

[9] See http://beta.visl.sdu.dk/cg3.html for details (last accessed 30 June 2008).

2.7 Related Research

In this section we refer to related research on Natural Language Processing of VSO languages other than Irish, followed by some related research for Irish NLP.

Finite-state techniques have been applied to morphological analysis and generation for a number of VSO languages in addition to Irish, including Arabic (Beesley, 1998);(Attia, 2000), (Habash and Rambow, 2006) Hebrew (Wintner and Yona, 2003) and Welsh (Mittendorf and Sadler, 2006). Wintner states that finite-state techniques are currently accepted best practice for many NLP applications, but cautions that maintainability could be an issue as systems grow larger (Wintner, 2008).

Parsers are available for both Arabic and Hebrew, while parser development for Welsh is in progress (PARGRAM). For Arabic, a chunker was developed through machine learning using Support Vector Machines and the manually disambiguated Arabic Penn Treebank (Diab et al., 2005). In the case of Hebrew, a rule-based script automatically applied Morpho-Syntactic Dependencies (Guthmann et al., 2009) to the manually annotated Modern Hebrew Treebank (Sima'an et al., 2001). In both cases, unlike Irish or Welsh, parser development benefited from the use of a pre-existing Treebank.

The following are a number of recent publications relating to Irish and Natural Language Processing.

- Prof. Kevin Scannell's Natural Language Processing Website (Scannell, 2007).

Prof. Scannell has developed many utilities for automatically processing Irish and other languages, these include a web-crawler and search engine for Irish as well as a grammar checker.

- Advances in the lexicography of Modern Irish verbs (Wigger, 2007).

This paper, presented at the 38th Poznan Linguistic Meeting Poland, describes a project whose aims are "the empirical analysis of usage and the description of recurrent syntactical patterns and semantic differentiation for the majority of Irish verbs". Its expected outcome will be "a dictionary of Irish verbs and verbal locutions". The results of this research could provide valuable information on the subcategorisation of Irish verbs, which is necessary for effective parsing of the language.

- Towards a Machine-Learning Architecture for Lexical Functional Grammar Parsing (Chrupala, 2008).

The research hypothesis of this PhD thesis is "that by exploiting machine-learning algorithms to learn morphological features, lemmatization classes and grammatical functions from treebanks the amount of manual specification can be reduced and robustness, accuracy and domain- and language -independence for LFG parsing systems can be improved". As part of this research, experiments were carried out on several languages, including Irish, whereby morphological features and lemmatization classes are induced from Gold Standard Corpus training data.

The following PhD thesis examines valency of Irish verbs which is of relevance to automatic parsing.

- A Study of Valency in Modern Irish (Nolan, 2001)

This thesis is concerned with characterising the factors that underpin the syntactic and semantic valency of Irish verbs using a functional approach. In particular it aims "to define the relationship between the semantic representation of a verbal predicate in the context of a clause and its syntactic expression through the argument structure of the verb".

2.8 Linguistic Annotation: A Worked Example

In this section we show the stages involved in transforming raw text into linguistically annotated text, as a result of the processing pipeline presented in this dissertation. The steps include pre-processing, tokenization, morphological analysis and lemmatization, POS tagging through disambiguating morphological analyses in context, grammatical function and dependency annotation, and finally chunking.

Stage 1: Example Sentence

```
Tháinig an  bháisteach ar an  tríú  hoíche.
Came    the rain       on the third night
'The rain came on the third night'
```

Stage 2: Tokenized Text

```
Tháinig
an
bháisteach
ar
an
tríú
hoíche
.
```

Stage 3: Morphologically Analysed and Lemmatized Text

```
"<Tháinig>"
    "tar" Verb PastInd Neg Len
    "tar" Verb PastInd Len
"<an>"
    "an" Art Sg Def
    "an" Part Vb Q Cond
    "an" Part Vb Q Fut
    "an" Part Vb Q Past
    "an" Part Vb Q Pres
    "is" Cop Pres Q
    "is" Cop Pres Dep Q
"<bháisteach>"
    "báisteach" Noun Fem Voc Sg Len
    "báisteach" Noun Fem Com Sg Def
    "báisteach" Noun Fem Com Sg Len
    "báisteach" Verbal Noun Rel Len
"<ar>"
    "ar" Prep Simp
    "ar" Verb PastInd
    "ar" Part Vb Q Past
    "ar" Part Vb Rel
    "is" Cop Pres RelInd
    "is" Cop Past Q
    "is" Cop Past RelInd
"<an>"
    "an" Art Sg Def
    "an" Part Vb Q Cond
    "an" Part Vb Q Fut
    "an" Part Vb Q Past
    "an" Part Vb Q Pres
    "is" Cop Pres Q
    "is" Cop Pres Dep Q
"<tríú>"
    "trí" Num Ord
    "tríú" Noun Masc Gen Sg
    "tríú" Noun Masc Com Sg
    "tríú" Noun Masc Com Sg Def

"<hoíche>"
    "oíche" Noun Fem Com Sg
    "oíche" Noun Fem Gen Sg Def
<.>
    "." Punct Fin
```

Stage 4: POS Tagged (Morphosyntactically Disambiguated) Text

```
"<Tháinig>"     "tar" Verb PastInd Len
"<an>"          "an" Art Sg Def
"<bháisteach>"  "báisteach" Noun Fem Com Sg DefArt
"<ar>"          "ar" Prep Simp
"<an>"          "an" Art Sg Def
"<tríú>"        "trí" Num Ord
"<hoíche>"      "oíche" Noun Fem Com Sg
"<.>"           "." Punct Fin
```

Stage 5: Grammatical Function and Dependency Annotated Text

```
"<Tháinig>"     "tar" Verb PastInd Len @FMV
"<an>"          "an" Art Sg Def @>N
"<bháisteach>"  "báisteach" Noun Fem Com Sg DefArt @SUBJ
"<ar>"          "ar" Prep Simp @PP_ADVL
"<an>"          "an" Art Sg Def @>N
"<tríú>"        "trí" Num Ord @>N
"<hoíche>"      "oíche" Noun Fem Com Sg @P<
"<.>"           "." Punct Fin
```

Stage 5: Linguistically Annotated and Chunked Text.

```
[S
 [V Tháinig tar+Verb+PastInd+Len+@FMV]
 [NP an an+Art+Sg+Def+@>N bháisteach
     báisteach+Noun+Fem+Com+Sg+DefArt+@SUBJ]
 [PP ar ar+Prep+Simp+@PP_ADVL
    [NP an an+Art+Sg+Def+@>N tríú trí+ Num+Ord+@>N hoíche
       oíche+Noun+Fem+Com+Sg+@P< ]] . .+Punct+Fin
S]
```

2.9 Summary

In this chapter, we introduced the notion of different levels of corpus annotation. We described some current linguistic annotation schemes, and gave an overview of part-of-speech tagging and partial syntactic parsing. We described three POS tagging methodologies; statistical, rule-based and transformational. We introduced some of the main concepts in syntactic analysis and parsing, i.e. predicate-argument structure, grammatical relations, constituent structure, chunking, and dependency relations.

We outlined the stages and tools involved in POS tagging and partial parsing of Irish, as developed in the present dissertation. A two stage approach to the implementation of POS tagging for Irish was adopted:

- Tokenized text is first morphologically analysed using finite state transducers developed using the Xerox Finite State Tools, and
- The appropriate morphological analysis, given the context in which the token is used, is determined using Constraint Grammar disambiguation.

The work on partial parsing is exploratory in nature and the handling of some aspects of Irish syntax may require revision as more facts come to light. Nevertheless, we have implemented a framework for partial parsing which can be extended and modified as required. To accomplish this we use the two paradigms already in use for POS tagging.

- Dependency relations, grammatical functions and clause boundaries are annotated using Constraint Grammar
- Chunk boundaries are inserted using Finite-State regular expressions.

In the final chapter in Part I, we present the creation of a Gold Standard Annotated Corpus. This is followed in Parts II and II by a detailed description the tools developed for POS tagging and partial parsing of Irish.

3 A Gold Standard Evaluation Corpus

3.1 Introduction

The development of each of the annotation tools is an iterative process. In the early stages of development, the tools were tested on samples of text chosen from the various genres in the CNG corpus (ITÉ, 2003). The output was inspected, errors were noted, and the tools were revised and retested on these texts and additional texts. However, using this informal method, we could not be sure whether the results obtained in this manner were representative of the corpus as a whole.

In order to test and evaluate the quality of the various stages of automatic corpus annotation more formally, a randomly selected evaluation sub-corpus was created. This entailed selecting a random sample of sentences from the overall NCII corpus (Kilgarriff, Rundell, and Uí Dhonnchadha, 2007), automatically annotating it, and manually correcting the annotations. This is our Gold Standard POS Annotated Corpus. It is against this Gold Standard Corpus that we measure the quality of our POS annotation tools.

During manual disambiguation of the Development Set part of the Gold Standard Corpus, (which was automatically tagged using the POS tools under development), many shortcomings in the tools were noted (see Section 3.3 for further details). The process of manual disambiguation greatly enhanced the development process, as the problems observed which related to tokenization, morphological analysis, and automatic disambiguation were later systematically addressed and tools re-tested against the Gold Standard POS Corpus.

We repeated this process to create a Gold Standard Dependency Annotated Corpus and a Gold Standard Chunked Corpus.

In Section 3.2, we describe in detail how the Gold Standard Corpus was created. In Section 3.3, we describe the manual disambiguation of this Corpus and in Section 3.4, we describe the manual partial parsing of a subset of the Gold Standard Corpus. In Section 3.5, we describe the measures used to evaluate the output of the tools against the Gold Standard Corpus.

3.2 Text Selection for Gold Standard Corpus

In order to create the Gold Standard Corpus we extracted 3,000 sentences at random from the 30 million word NCII corpus (Kilgarriff, Rundell and Uí Dhonnchadha, 2007). As the NCII

Corpus is not in sentence-per-line format, therefore 3,000 random numbers between 1 and 30 million (i.e. number of words in corpus) were generated and stored in an array.

Each file of raw text making up the corpus was read consecutively, keeping a cumulative total of words. Whenever the word count matched a number in the array the next sentence was written out to a file. Punctuation was used to find the end of the sentence (i.e. ! or . or ?).

This file of random sentences was manually verified. Upon inspection, a number of sentences were found to be unsuitable for testing purposes as they consisted of fragments such as dates, names, numerical references, list items or phone numbers only. This was due to shortcomings in the NCII corpus. We, therefore, produced 3,020 new random numbers and removed unsuitable material as well as two sentences which were selected twice. Lines of poetry were removed if they were not representative of normal syntactic or punctuational conventions. In addition, some sentences were inadvertently truncated after a full stop in an abbreviation such as '*Dr.*' or '*Co.*' The end portion of these sentences was restored.

After checking the data 3,001 sentences of raw data remained. The average sentence length is 23 words, with a minimum sentence length of 1 word and a maximum sentence length of 312 words (from a legal text). These sentences were randomly distributed into two parts: a Development Set and a Test Set, in a ratio of approximately 2:1. Detailed figures are given in Table 10.

Table 10 Composition of Gold Standard (3000) POS Corpus

POS Tagged Data	Dev. Set	Test Set	Total
Sentences	2,036	965	**3,001**
Words	45,460	22,139	**67,599**
Tokens**	50,151	24,588	**74,739**
MWE*	422	261	**683**

* MWE = multi-word expression

** Tokens = words, punctuation and MWEs

3.3 Manual Disambiguation

Both Development and Test Sets were automatically tokenized and morphologically analysed, and then manually disambiguated. Ideally each text should be manually disambiguated by at least two human disambiguators independently, the results compared, and a single version agreed upon. The next best option would be for one person to disambiguate and for another person to check the work. However, due to lack of human resources neither option was available to us and as a consequence each text was only manually disambiguated once.

During manual disambiguation some typographical errors (e.g. separate words joined or single words split in two) were removed from the Development Set raw text where it was felt that they would impede the evaluation process.

3.3.1 Manual Disambiguation Guidelines

In order to achieve consistency of POS tagging, we developed guidelines to aid manual disambiguation, particularly in cases where the choice of POS was not straightforward. In cases, where the surface form is systematically the same, e.g. Verbal Adjective and Verbal Noun (genitive case), or Demonstrative Pronoun and Demonstrative Determiner, it can be difficult to disambiguate consistently unless there are guidelines and examples. These guidelines may be found in Appendix C.

3.3.2 Issues Arising from Manual Disambiguation

During manual disambiguation of the Gold Standard Corpus Development Set a number of issues came to light. These problems were recorded and later examined and categorised according to the stage in the processing pipeline at which they occurred. The problems identified were as follows:

- Tokenizer
    - Multi-word expressions missing from the lexicon
        - Idioms, e.g. *go leor* 'plenty
        - Place names *Uíbh Fháillí* 'Offaly', *Baile Átha Cliath* 'Dublin'
        - Organisations e.g. political parties, *Fianna Fáil* 'Soldiers of Destiny', *Lucht Oibre* 'Labour Party'
    - Punctuation issues
        - It was apparent that some bracketed items should be kept together, e.g. '(6)', '(iii)' etc.

        - Where a sentence ending in a number, the full stop should separate from the number, e.g. '1996.'
        - In a number of special cases, i.e. contractions, punctuation should not be separated from the word, e.g. *im'* 'in my', *a's* 'and'
        - A number of common abbreviations were missing, e.g. *Uimh.* 'number'
        - Some inflected abbreviations were missing, e.g. *gCo.* 'county'
        - Possessive proper names needed special treatment, 'Madigan's', 'Pete's Pizzas'
    - E-mail and web addresses
        - 'panceltic@eircom.net', 'www.oneworld.com', 'http://10steps.ie'
    - Typographical errors in the text
        - Occurrences of single lexemes split in two parts, e.g. *fan faidh* (*fanfaidh*) 'will come'
        - Occurrences of two separate lexemes being joined, *agusan* (*agus an*) 'and the'
- Finite-state Lexicon
    - Missing dialectal variants of common function words
    - Missing lemmas
    - Lemma missing an additional part-of-speech option
    - Proper name inflectional morphology needed to be improved
    - Misspellings in the text
- Compound Recogniser
    - Too liberal - needed to be restricted
    - Should include Adj-Noun compounds as well as Noun-Noun compounds
- Guessers
    - Guessers needed to be reordered, and in some instances divided and reordered
    - A single guessed analysis was not sufficient in the case of some types of suffixes, e.g. *-adh*

This exercise clearly showed that problems with automatic POS tagging were distributed throughout all of the processing stages. The list of problems with tokenization, almost all of which are exceptions to the general rule of separating punctuation from text and splitting text on white space, caused misalignment between the manually disambiguated text and the automatically disambiguated text, but their impact on the overall analysis of a sentence was not serious and they were the easiest type of problem to rectify.

On the other hand, missing dialectal variants of common function words had a far more serious impact on the overall POS analysis of a sentence as these tokens were

predominantly guessed as noun or verb which had detrimental knock-on effects for neighbouring tokens.

It was found that the compound recogniser (Section 5.6) was over-generating, resulting in many non-existent compounds being suggested. For example the adjectival suffix *-each* was interpreted as the noun *each* (an old word for 'horse'), resulting a large number of unusual horsey compounds. In the case of guessers, the order in which they were used was found to be less than optimal (further details can be found in Section 5.9). Better results were achieved by re-ordering and splitting some of the guessers (Section 5.7).

3.4 Gold Standard Dependency Corpus and Gold Standard Chunked Corpus

In order to assess the quality of the tools for partial parsing of Irish, we also need a gold standard with which to compare the output of the tools. The gold-standard POS tagged text was used as a basis for the Gold Standard Dependency Corpus. The gold standard dependency data was in turn used as the basis for the Gold Standard Chunked Corpus.

Over 150 sentences were chosen at random from the Development Set and another 100 were randomly chosen from the Test Set (see Table 10). These sentences were automatically tagged with functional and dependency tags and were then manually corrected. Details are given in Table 11 .

Table 11 Composition of Gold Standard (250) Dependency Corpus

Dependency Tagged Data	Dev Set	Test Set	Total
Sentences	150	100	**250**
Words	4,036	2,314	**6,350**
Tokens	4,476	2,580	**7,056**

The 250 sentences of the Gold Standard Dependency Corpus where then automatically chunked and also manually corrected, to create the Gold Standard (250) Chunked Corpus.

3.5 Evaluation Measures

The development and evaluation of the tools is an iterative process on the Development Sets, with final evaluation carried out on the Test Sets, as shown in Figure 9. The tools are evaluated against the Gold Standards using precision, recall and F-score measures (Manning and Schütze, 1999, p268-269)). These are the standard measures used to evaluate annotation quality in Computational Linguistics.

To use POS-tagging as an example: a conservative tagger might tag a token with a particular tag only when absolutely sure, and, therefore, have a high precision rate. This could result in a low recall rate if there were many other tokens which the tagger should have been tagged with this particular tag. The reverse is also true; a tagger could have a high recall rate by tagging all possible tokens with a particular tag, thereby ensuring that all actual instances were tagged correctly. However, the number of incorrect tags would adversely affect the precision rate. The ideal situation is to maximise both precision and recall at the same time. The F-score is a combination of both measures.

Precision is calculated as: $$\frac{CorrectAutoTags}{AllAutoTags} \times \frac{100}{1}$$

Recall is calculated as : $$\frac{CorrectAutoTags}{GoldTags} \times \frac{100}{1}$$

F-score is calculated as : $$\frac{Recall \times Precision \times 2}{Recall + Precision}$$

To evaluate the POS and Dependency Annotation, a Perl program was written which counts the number of matching annotations and calculates precision, recall and F-scores for the automatic tagging. As well as overall results, in the case of the Development Set data we produce more fine-grained calculations. We output details of all tag mismatches as well as an analysis of each part-of-speech or dependency tag.

These measures are used in Sections 5.6.4, 5.8, 6.6, 7.6 and 8.4.

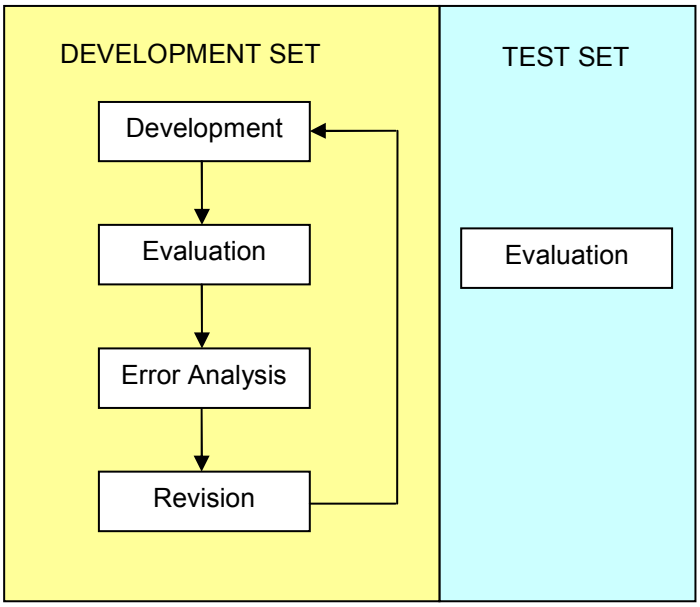


Figure 9 Development - Evaluation Cycle

3.6 Summary

In this chapter we outlined the need for an evaluation benchmark for the automatic annotation of texts, in the form of a Gold Standard Annotated Corpus. We described our method of randomly selecting 3,000 sentences from the 30 million word NCII Corpus in order that the Gold Standard Annotated Corpus be representative of the larger corpus. For POS tagging evaluation, the 3,000 sentences were random distributed into a Development Set and a Test Set of approximately 2,000 and 1,000 sentences, respectively.

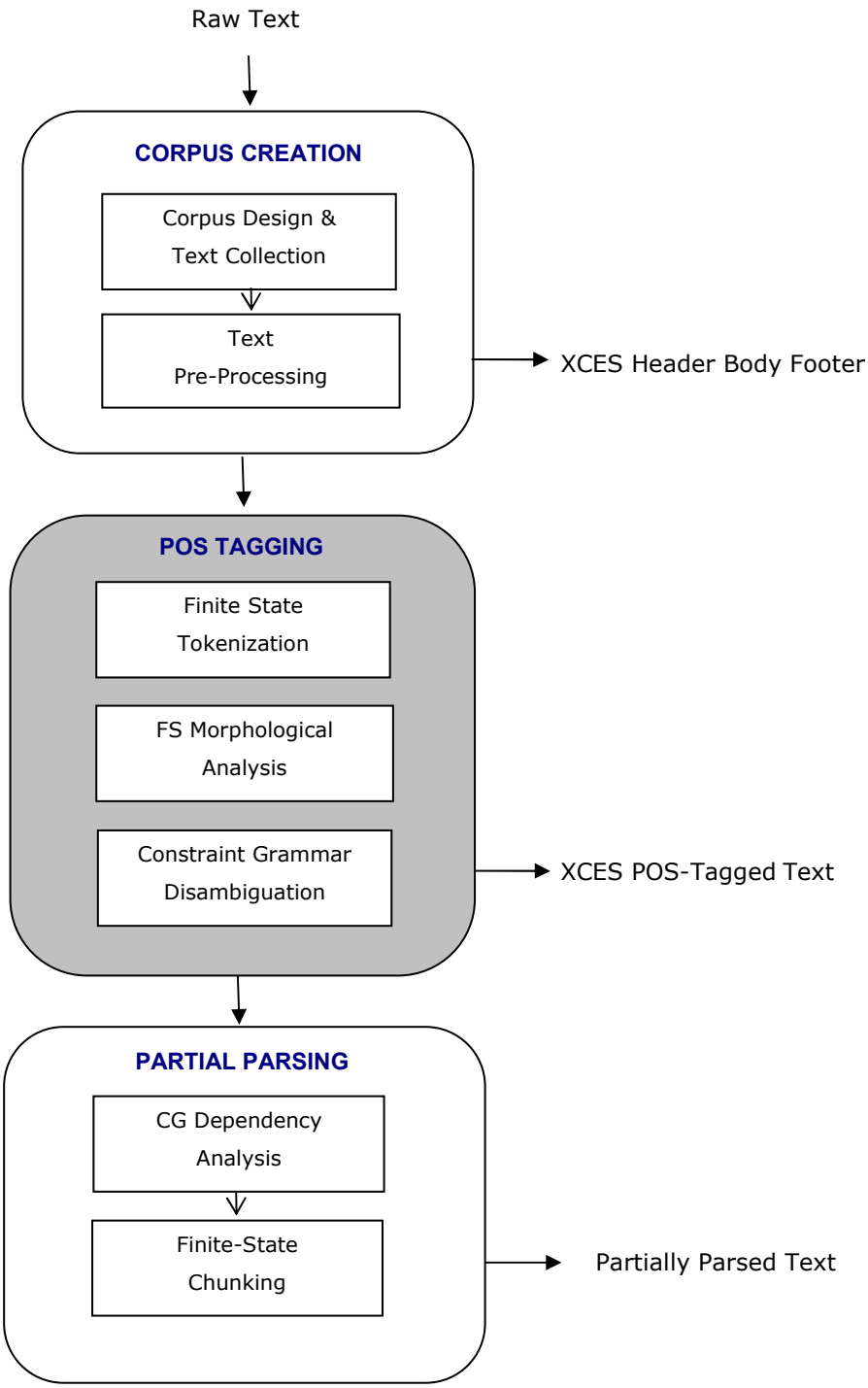
All of the sentences in the Gold Standard were morphologically analysed (automatically) and manually disambiguated. Based on our manual disambiguation of the Development Set, a number of shortcomings were identified in the tokenization and morphological analysis tools. We also developed a set of guidelines for disambiguation, during the manual disambiguation of texts.

Using the Gold Standard POS Tagged Corpus, we randomly selected 150 sentences from the Development Set and 100 sentences from the Test Set to create a Gold Standard for automatic dependency analysis. These sentences were automatically annotated with dependency and grammatical function tags, which were then manually corrected to create a Gold Standard Dependency Corpus. The Gold Standard Dependency Analysis sentences were then automatically bracketed into chunks and manually corrected to create a Gold

Standard Chunked Corpus. The process of inferring chunks from the dependency annotations is described in Section 8.3.

In Part II of the dissertation, the annotation tools for POS Tagging are described in detail, with chapters dealing in turn with the development and evaluation of the tokenizer, morphological analyser and guessers, and disambiguation rules.

Part II Automatic Part-of-Speech Tagging for Irish



4 Finite-State Tokenization

4.1 Introduction

The initial stage of POS tagging a text is to separate the text into a stream of individual tokens. These tokens may be words, punctuation markers, abbreviations, numbers or multi-word expressions (MWE). A MWE consists of a series of words separated by white space, which we prefer to keep together and treat as one token, e.g. we treat the preposition *tar éis* meaning 'after' as one token. In general, when the meaning is non-compositional (i.e. the meaning of whole cannot be inferred from the parts) we keep the lexical items together as a unit (i.e. a multi-word token).

On the other hand, there are words[10] such as contractions which consist of more than one element, which we would like to split into their constituent parts. For example, the possessive determiners *mo* 'my' and *do* 'your' always combine with a following vowel-initial noun, as in *m'athair* 'my father' or *d'athair* 'your father'. In this case it makes sense to treat this word as two separate tokens.

The stream of tokens resulting from tokenization forms the input to morphological analysis and disambiguation, and it is ultimately these tokens which are assigned part-of-speech tags and lemmas. Tokenization is therefore a very important step in the overall process, as decisions made at this early stage will effect all subsequent processing.

In Section 4.2 we discuss the issues which a general purpose tokenizer must address. In Section 4.3 we describe the implementation of our finite-state tokenizer. In Section 4.4 we provide details of our evaluation of the tokenizer.

4.2 Tokenization Issues

The are several issues which must be addressed by a tokenizer, and most of these are well documented in the literature (Grefenstette and Tapanainen, 1994; Habert et al., 1998; He and Kayaalp, 2006). (There are some domain-specific issues, such as the tokenization of chemical/mathematical formulae and biomedical terms (He and Kayaalp, 2006) which we do not address in this general purpose tokenizer.) In this section we will discuss tokenization issues as they relate to Irish.

Broadly speaking, the issues relating to tokenization can be summarised as follows:

[10] In this context, by 'word' we mean a sequence of characters bounded by white space or other delimiters.

1. Sentence Internal Punctuation
    a. commas, colons, quotation marks
2. Sentence Boundaries
3. Word Internal Punctuation
    a. Abbreviations
    b. Numerical Expressions
    c. XML/SGML tags
    d. E-mail addresses and URLs
    e. Enumerated lists
    f. Hyphenation
    g. Contractions
    h. English possessive marker
4. Multi-Word Expressions and Named Entities

4.2.1 Sentence Internal Punctuation

Irish texts, as with many other European languages, can to a large extent be segmented according to white space between words, as shown in (24). Sentence internal punctuation, such as quotation marks (apostrophes), commas, colons, semi-colons, brackets, dashes, exclamation marks and question marks must be separated from any word to which they are adjoined, e.g. 'Seán,' becomes two tokens.

(24) ‘Cá bhfuil Seán, Máire agus Síle?’, arsa Liam.
‘Where is Seán, Máire and Síle?’, said Liam.

4.2.2 Sentence Boundaries

Usually punctuation such as '.' '!' and '?' mark the end of sentences, but this can not be relied on in all cases, particularly in the case of direct speech where '?' or '!' are not sentence-final, (24) or when a sentence contains an abbreviation with a full stop such as Dr. (25).

(25) Cá bhfuil Dr. Ó Ceallaigh?
Where is Dr. Ó Ceallaigh?

A particularly difficult case arises when an abbreviation is sentence final and the full stop performs two functions simultaneously. In the current implementation, if the abbreviation is specified in the tokenizer, the end-of-sentence marker will be lost. This will cause a loss of

information if the sentence is not followed by a linefeed/return character.[11] If the type of sentence-final abbreviation shown in (26) is not specified in the tokenizer, the punctuation will be treated as sentence final and `Teo` as an unknown word rather than an abbreviation.

```
(26)  Tá sé ag obair   in Teile Teo.
      Is he at working in Teile Ltd.
      'He is working in Teile Ltd.'
```

Alternatively, a non-deterministic tokenizer could be implemented where the punctuation in input text such as 'Dr.' could be given three analyses, 1) part of the previous word, 2) part of the previous word and an end-of-sentence mark, and 3) an end-of-sentence mark.

4.2.3 Word Internal Punctuation

The task of tokenizing punctuation is further complicated by the fact that some of the most common symbols (particularly full stop, comma and apostrophe) can also occur word-internally. While, in general, punctuation must be separated from words, there are several important exceptions to this rule. These include numbers which have internal punctuation e.g. '100,234' or '€12.50', abbreviations such as 'Mr.', 'cm.', 'I.N.T.O.' and XML tags like `<p>`, `</p>`. We must also handle e-mail addresses and web site addresses which have their own internal syntax, e.g. 'name@org.ie' and 'http://www.org.ie'.

Enumerated lists are an example of another type of string classified here under word-internal punctuation. Usually we separate brackets from adjoining text, however in the case of list items such as '(a)', 'b)', '(iii)', '(2)', '(IV)' etc., it does not make sense to separate the brackets, therefore we specify in the tokenizer that these items should be kept together.

As described in Section 1.3 (Cleanup of Newspapers/Periodicals), the problem of removing spurious hyphens which were inserted for type-setting reasons is largely dealt with at the pre-processing stage. Those that remain are treated as part of the word, including those that are related to initial mutation of Irish words, e.g. *an t-arán* 'the bread'.

Some lexical items contain more than one token. In the case of contractions such as *m'athair* 'my father' (as previously described), we must decide whether we want this to be one, two or three tokens. In this implementation *m'athair* is treated as two tokens, *m'* and *athair*, where *m'* is associated with the lemma *mo* 'my' (27). This method avoids duplication, as the bare noun and the reduced prefixes need only be encoded once in the lexicon, rather than

[11] The Evaluation Corpus is sentence delimited, but the NCII corpus from which it was randomly selected is not.

encoding the bare noun and also the noun plus its possible prefixes, e.g. *athair* 'father' and *m'athair* 'my father' etc.

(27)	Token	Lemma	Morpho-syntactic tags	Gloss[12]
	m'	mo	+Det+Poss+1P+Sg	'my'

Some Irish texts contain English proper nouns containing a possessive marker, e.g. 'Pete's Pizzas'. In these cases we treat a possessive noun containing an apostrophe as one token.

4.2.4 Multi-Word Expressions (MWE)

Multi-word expressions are those whose meaning is non-compositional, i.e. it cannot be discerned from the individual parts. For example *cé is móite* as a unit means 'except', whereas analysing the parts individually as in (28) would result in substantial ambiguity and an incorrect analysis. This can be avoided by treating it as a MWE as in (29). In order to do this *cé is móite* must be treated as a single token at the tokenization stage. In general the more MWEs that are identified at the tokenization stage, the less ambiguity there will be at the morphological analysis and disambiguation stages. There is great benefit to be derived from identifying MWEs at the earliest possible stage.

(28)	*Token*	*Lemma*	*Morpho-syntactic tags*	*Gloss*
	cé	cé	+Conj+Subord	'even'
	cé	cé	+Noun+Fem+Com+Sg	'quay'
	cé	cé	+Noun+Fem+Gen+Sg	
	cé	cé	+Noun+Fem+Gen+Sg+DefArt	
	cé	cé	+Pron+Q	'who'
	is	is	+Cop+Pres	copula
	is	is	+Cop+Pres+Rel	
	is	is	+Cop+Part+Sup	
	is	agus	+Conj+Coord	'and'
	móite	móite	+Guess+Verbal+Adj	no gloss

(29)	cé is móite	cé_is_móite	+Conj+Subord	'except'

In some cases a non-MWE solution is sufficient or even necessary. For example, traditional grammars (An Gúm, 1999) list a score of compound prepositions (comprising of a simple

[12] Gloss has been added for information but is not part of the analyser output.

preposition and a noun), e.g. *ar feadh* 'for (a period of time)' and state the rule that a noun following a compound preposition has genitive case (30), whereas a noun following a simple preposition has nominative case (dative case has been lost in all but a few fossilised phrases). However, this genitive case rule is a general consequence of two nouns occurring in succession, and therefore in this instance it is not essential that the compound preposition be encoded as an MWE.

(30)	Token	Lemma	Morpho-syntactic tags	Gloss
	ar	ar	+Prep+Simp	'on'
	feadh	feadh	+Noun+Masc+Com+Sg	'duration'
	míosa	mí	+Noun+Fem+Gen+Sg	'month'

Furthermore, there are cases where the tokens can be either idiomatic or literal depending on the context, e.g. the compound preposition *ar bhord*; meaning 'aboard/on board' could also be a prepositional phrase literally meaning 'on a table' in which case the MWE interpretation would be incorrect. In this case, in order to allow the possibility of individual tokens, we cannot encode this compound preposition as a MWE. In other cases elements of an expression can be inflected and it would be impractical to list all of the inflectional variations as MWEs.

The decision whether to analyse a sequence of tokens as a MWE or not, is, based on whether a) the particular sequence *always* has an idiomatic meaning (including named entities such as places and organisations consisting of multiple parts, e.g. *Baile Átha Cliath* 'Dublin', *Fianna Fáil*), or b) a compound preposition which always needs an NP complement, therefore ambiguity will be avoided by keeping the items together. If the phrase can sometimes have a literal (compositional) meaning or has numerous inflectional variations we do not encode it as an MWE. All things being equal our preference is for MWEs and we would wish to expand this part of the tokenizer considerably in the future, particularly in the area of named entities.

4.3 Implementation of the Finite-State Tokenizer

We have chosen to implement a rule-based finite-state tokenizer using Xerox Tools. Statistical machine learning techniques can also be used for tokenization, particularly for named entity recognition and sentence boundary detection (Mikheev, 2003), but as these methods require annotated training material which we do not have available to us, we prefer at present to use rule based methods.

The finite state tokenizer is developed using Xerox XFST Tools[13] and is based on a tokenizer by Anne Schiller (Grefenstette et al., 2000). Currently, the tokenizer is 6.2 MB in size, has 5221 states and 509,211 arcs. Initially, white-space (i.e. space, tab, new line character) and punctuation characters are defined. Everything other than punctuation and white-space is considered to be a character and a word is defined as a string of characters.

An apostrophe is treated as punctuation, and by default is separated from other characters. During the testing of this tokenizer using the Development Set data, we found that most misalignments related to instances where, contrary to the general rule, punctuation should not have been separated from the word, as in the case of contractions like *im'* meaning *i mo* 'in my' etc. Further examples include:

- contractions, e.g. *a's* (*agus* 'and'), *'un* (*chun* 'towards'), *a'* (*an* 'the' or *ag '*at')
- list item emumerators which should stay together, e.g. '(iii)', '(B)'
- e-mail & URLs, e.g. 'panceltic@eircom.net', 'www.oneworld.com'
- abbreviations, *gCo*., 'Co.' *Uimh.* 'No.', 'CD-ROM'
- proper names with English genitive, e.g. 'Madigan's', 'Pete's Pizzas'

The other areas for concern are multi-word units and typographical errors.

Punctuation and multi-word expression problems were addressed by adding specific regular expressions to the tokenizer to deal with these cases. Typographical errors in the raw text, other than misspellings, (e.g. a space in the middle of a word or a hyphenated word separated into two parts), which were missed during pre-processing were removed (Uí Dhonnchadha and van Genabith, 2006).

All cases where punctuation should not be separated from its adjacent characters, or where the text should not be segmented on white space are explicitly defined using regular expressions in the tokenizer, as follows:

- Contractions
- Abbreviations
- English possessives
- XML tags
- Numeric expressions
- Enumerated lists
- URL's and e-mail addresses

[13] Xerox Finite State Tools: see http:// xrce.xerox.com for details

- Hyphenation
- Multi-word expressions
- Named Entities

Contractions such as *m'*, *d'*, *b'*, *a's*, *a'm* (i.e. contractions of *mo* 'my', *do* 'your', *ba* 'was', *agus* 'and, *agam* 'at me', respectively) are defined in the tokenizer (Figure 10), enabling words such as *m'athair* 'my father' to be analysed as two tokens (*m'* and *athair*) and *a'm* to remain as one token (rather than three tokens: *a*, *'* and *m*).

```
define CONT [ s | {MB'} | {b'} | {B'} | {d'} | {D'} | {m'} |
{M'}
| { 's } # 's space in front to avoid Shea's
| {a'm}
| {a'at}
| {an-}
| {dod'}
| {s'} # s'againne
| {S'} # S'againne
| {'na }
| {'n } # 'n space must follow to avoid quoted words like 'nua'
| {ars'}
| {a's}
| {a'}
| {N'} # N'fheadar
| {n'} # n'fheadar
etc.
];
```

Figure 10 Tokenizer Definitions: Contractions

In Figure 11, some common abbreviations which include a full stop are defined (ABBR), followed by a more general definition of the way initials can be used in abbreviations (INIT) We allow one or more letters with full stops to be an abbreviation. As sentences such as *Chuala mé í* 'I heard her/it' which end in a single letter pronoun *í* or *é* ('her', 'him/it') followed by a full stop are common in Irish we exclude *é, í, É* and *Í* from being single-letter abbreviations.

```
define ABBR [ {Co.}
|{gCo.}
|{Dr.}
|{eag.}
|{e.g.}
etc.
];

# Letters are defined, then é,í,É and Í are excluded
define Letter
[A|Á|B|C|D|E|É|F|G|H|I|Í|J|K|L|M|N|O|Ó|P|Q|R|S|T|U|Ú|V|W|X|Y|Z|
a|á|b|c|d|e|é|f|g|h|i|í|j|k|l|m|n|o|ó|p|q|r|s|t|u|ú|v|w|x|y|z];
define Pron [é|í|É|Í];
```

```
define Notpron [\Pron]
define INIT [ Letter %. [Letter %.]+ ] | [ Notpron %.];
```

Figure 11 Tokenizer Definitions: Abbreviations

Irish texts often include English proper nouns. In the case of possessives we choose to keep the possessive apostrophe with the English proper noun, e.g. 'Pete's Pizzas'. In the definition in Figure 12, `Char` is previously defined as any character other than those defined as whilte space (spaces, tabs, newline markers) or punctuation.

```
define ENGWORD [ Char Char+ [' s]];
```

Figure 12 Tokenizer Definitions: English Possessive Apostrophe

In Figure 13 we define the XML Tags which can occur in the pre-processed text.

```
define TAG [ {<p>} | {</p>} | {<s>} | {</s>}
| {<title>} | {</title>}
| {<caption>} | {</caption>}
| {<gap desc='table'/>}
| {<poem>} | {</poem>}
| [%& a m p %;]
.... ];
```

Figure 13 Tokenizer Definitions: XML Tags

Numeric expressions and list item indicators of the form `(a)`, `(3)`, `(12)`, `(ii)`, `(IV)` (excluding forms such as `(ab)`, or `(123)`) are defined as shown in Figure 14.

```
define Digit [%0| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 ];
define NumOp [%- | %= | %+ | %* | %/ | %: ];
define NumSep [%. | %, ];
define NUM [ [Digit | NumOp | NumSep]+ [Digit]] | [%# Digit+];

define Roman  [ i | v | x | l | c ];
define URoman [ I | V | X | L | C ];
define ITEM [ %( [Letter|[Digit (Digit)]|Roman+|URoman+] %) ];
```

Figure 14 Tokenizer Definitions: Numeric Expressions and List Numbering

In Figure 15 we define URLs and email addresses.

```
define WEB [ [h t t p %: %/ %/]|[w w w %.] ];
define AT [%@];
define AlphaNum
[A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z|a|á|b|c|d|
```

```
e|é|f|g|h|i|í|j|k|l|m|n|o|ó|p|q|r|s|t|u|ú|v|w|x|y|z|1|2|3|4|5|6
|7|8|9|%0];
define EMAIL [ [AlphaNum+ (%. AlphaNum+) ]+ AT [AlphaNum+ %.]+
AlphaNum+ ]
|[ WEB [AlphaNum+ %.]+ AlphaNum+];
```

Figure 15 Tokenizer Definitions: URLs and E-mail Addresses

A hyphen is not defined as punctuation, therefore hyphenated words are not split e.g. *dea-mhéin* 'goodwill' etc. Exceptions such as the prefix *an-* 'very' (e.g. *an-mhaith* 'very good) are handled in contractions (`CONT`) above, and initial mutations *t-* and *n-* are dealt with separately in Figure 16.

```
define MUTWORD [[t %- ] | [n %-]][Char]+ ;
```

Figure 16 Tokenizer Definitions: Initial Mutation Hyphen

All multi-word expressions must be defined in the tokenizer, including compound prepositions, place names and organisation names. At present we have listed some commonly used Irish place names and organisations - but a more comprehensive list would be beneficial (see Section 5.2).

```
define MWE [ {ar feadh} |

# MWE Compound Prepositions
{ar fud} |
{ní ba} |
{ní b'} |
{os cionn} |
{os comhair} |
{tar éis} |
etc.

# MWE Quantifiers
{a lán} |       # lots
{a thuilleadh} |      # more
{go leor} |           # plenty
etc.

# MWE Adverbs
{ó thuaidh} |   # north
{ó dheas} |           # south
etc.

# MWE - Named Entities
# Political Organisations
{Sinn Féin} |
{Fianna Fáil}  |
{Lucht Oibre} |
etc.
```

```
# Place Names
{Baile Átha Cliath} | # Dublin
{Béal Feirste}        # Belfast
etc.
];
```

Figure 17 Tokenizer Definitions: Multi-Word Expressions

4.4 Evaluation of the Tokenizer

While there is substantial discussion in the literature on the problems associated with tokenization, very little is to be found on the subject of evaluation of tokenizers. However, two methods are suggested. The first method entails comparing the tokenizer output with gold standard tokenized texts. Grefenstette and Tapanainen (1994) use the Brown Corpus as a gold standard in their experiments. The second method is to compare the tokenizer output with the output of other tokenizers run on the same texts (Habert et al., 1998; He and Kayaalp, 2006).

As we do not have any other tokenizers for Irish available, we choose the first method for evaluation. Our tokenizer is assessed by comparing the alignment of the automatically tokenized output with the Gold Standard Corpus whose tokens were manually checked and tokenization was corrected where necessary.

After updating the tokenizer to deal with the problems encountered in the Development Set, the Gold Standard and the automatically tokenized texts were again compared. The results are given in Table 12.

Table 12 Tokenization Evaluation

	Gold Standard Tokens	Automatic Tokens	Difference
Development Set	50,166	50,152	-14
Test Set	24,588	24,584	-4

Precision (Dev. Set): $\frac{CorrectAutoTokens}{AllAutoTokens} \times \frac{100}{1} = \frac{50{,}151}{50{,}152} \times \frac{100}{1} = 99.99\%$

Recall (Dev. Set): $\frac{CorrectAutoTokens}{GoldTokens} \times \frac{100}{1} = \frac{50{,}151}{50{,}166} \times \frac{100}{1} = 99.97\%$

F-score (Dev. Set): $$\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{99.97\% \times 99.99\% \times 2}{99.97\% + 99.99\%} = 99.98\%$$

There is currently very little difference in the outcomes. In the case of the Development Set, the automatically tokenized text produced 14 fewer tokens, and in the case of the Test Set, 4 fewer tokens were produced. Results could include compensating errors, i.e. a number of incorrectly identified contractions could increase the number of tokens, while a number of incorrectly identified MWEs could reduce the number of tokens, thus cancelling each other out. We, therefore, decided to carry out a detailed analysis of the Development Set tokenization.

Table 13 shows that there is in fact only one compensating error. This is because in terms of the total number of tokens, the error in the first row of Table 13 is cancelled out by the error in the second row. The majority of errors are a result of Multi-Word Expressions which were added to the tokenizer after the Gold Standard was manually created. We, subsequently, updated the Gold Standard texts to incorporate these MWEs. We also added the new MWE *mór roinn* 'continent' to the tokenizer. Table 14 demonstrates the process involved. (Note, that there were multiple occurances in the data of the errors in Table 13). The final result is that the number of tokens for the Gold Standard Development Set and the Automatic Tokenization of the same sentences now agree (i.e. 50,151 tokens), giving precision, recall and f-score of 100%.

Table 13 Development Set: Error Analysis of Tokenization

Development Set	
Gold Standard Tokens	**Automatic Tokenization**
1 mór roinn	1 mór 2 roinn
2 in 3 ann	3 in ann
4 thar 5 a 6 bheith	4 thar a bheith
7 le 8 go	5 le go
9 chun 10 go	6 chun go

Table 14 Development Set: After Correction

Development Set	
Gold Standard Tokens	**Automatic Tokenization**
1 mór roinn	1 mór roinn
2 in ann	2 in ann
3 thar a bheith	3 thar a bheith
4 le go	4 le go
5 chun go	5 chun go

By extrapolation, we can say that the differences observed in the Test Set are, most likely, due to a small number of inconsistencies arising from the iterative nature of the development cycle, whereby changes (e.g. additions or deletions of MWEs) were implemented in the tokenizer after Gold Standard sentences where checked. While every effort was made to reflect these changes in the checked texts, inevitably some instances were missed. A further round of error checking would be required to eliminate such errors from the Gold Standard Test Set.

4.5 Summary

In this chapter we presented the development and evaluation of a finite-state tokenizer for Irish texts. We discussed some of the issues involved in the tokenization of Irish and presented a finite-state tokenizer modelled on that of Grefenstette *et al*. (2000). Currently, the tokenizer is 6.2 MB in size, has 5221 states and 509,211 arcs.

In our evaluation, which involved comparing manually and automatically tokenized versions of the Development Set data, we found that more than 99% of tokens aligned correctly. Any outstanding problems were due to a deficiency of MWEs and Named Entities (NE) in the automatic tokenizer. Therefore, further work in the area of tokenization should focus primarily on substantially increasing the number of MWEs and NE's defined in the tokenizer.

In the next chapter we will turn our attention to Finite-State Morphological analysis.

5 Finite-State Morphological Analysis

5.1 Introduction

The objective of finite-state morphological analysis is to produce all of the possible morphological analyses for each token in the input text. The analysis includes lemma, POS category, and other morphosyntactic features (tense, mood, gender, number etc.). Frequently more than one morphological analysis per token is found (60% of Irish tokens in our corpus have more than one analysis).

Morphological analysis of tokens for Irish is achieved in one of two ways:

1. Either the token or the root of the token exists in the Finite-State Morphology lexicons, or
2. The token's morphosyntactic features are predicted based on a range of measures including inflectional and derivational affixes, capitalisation, vowels in the final syllable, and the presence of foreign characters.

The work described in this chapter builds on an existing prototype finite-state morphology implementation for Irish (Uí Dhonnchadha, 2002), using Xerox Finite-State Tools (Beesley and Karttunen, 2003; Karttunen and Beesley, 1992). This Finite State Morphology (FSM) implemented all of the inflectional rules for Irish and contained a lexicon of approximately 1,500 lemmas, which included the lemmas associated with the 1,000 most frequently occurring word-forms in the ITÉ Reference Corpus of Irish (ITÉ, 2002) comprising of approximately 15 million words. The token recognition rate for this corpus was on average 81%.

In order to improve recognition rates in the existing FSM Analyser and obtain an analysis for all tokens in unrestricted text the following additional work was undertaken:

- The FSM lexicons were semi-automatically extended
- Derivational morphology rules were added and
- Morphological guessers were implemented.

We do not describe the development of the inflectional morphology FSM for Irish in this thesis as a detailed description may be found in (Uí Dhonnchadha, 2002).

In Sections 5.2 and 5.3 we describe the semi-automatic extension of the finite-state lexicons, followed by an evaluation of the results. In Sections 5.3 and 5.4 we describe the addition of rules for derivational morphology, followed by an evaluation of the results. In Sections 5.5 and 5.6 we give details of our compound recogniser, once again, followed by an evaluation

of the results. In Sections 5.7 and 5.8 we describe the finite-state guessers, and present details of our evaluation. Finally, in Section 5.9, we describe the morphological analysis lookup strategy, and in Section 5.10, we give an overview of token recognition rates.

5.2 Semi-Automatic Extension of FSM Lexicons

The finite-state morphology (FSM) lexicon was increased by semi-automatically converting a machine-readable dictionary (MRD) (An Roinn Oideachas, 1986) to Xerox *lexc* format.

Newspaper and web texts contain a high proportion of proper nouns. Therefore, lists of personal names and place names in printed resources were scanned and incorporated into the lexicon. Some lists of personal names were found on the Internet. After inclusion of the MRD headwords, as well as OCR and Internet named entities,[14] at least one analysis was returned for 93% of tokens in unrestricted text, i.e. the 30 million word NCII corpus. This is a 12% increase on the 81% result obtained for ITÉ (2002) corpus.

5.2.1 Organisation of FSM Lexicons

The lexicons in the Irish inflectional finite-state morphology engine (Uí Dhonnchadha, 2002; Uí Dhonnchadha et al., 2005) are organised in a hierarchical manner whereby a stem is associated with a lexical subclass (called continuation classes (Beesley and Karttunen, 2003; Karttunen and Beesley, 1992)) which in turn points to further continuation classes, which in an incremental manner, produce inflected surface forms and analyses associated with particular inflectional paradigms.

In order to add a new lexical item to the FSM, it is necessary to identify the appropriate top-level continuation class. In the case of verbs and adjectives this can be achieved with a high degree of accuracy by examining the surface form, but the morphology of nouns is far more complex and unpredictable. Traditional Irish grammars (An Gúm, 1999) describe 5 noun paradigms (declensions) for nouns based on the formation of the genitive singular. However, within these paradigms there is considerable variation in the manner in which plurals can be formed (over 20 varieties of plural are currently encoded). In our implementation, each of these 5 paradigms have been sub-divided on average ten times to reflect the various plural types, resulting in approximately 50 noun top-level continuation classes. In contrast, verbs and adjectives have 10 and 13 top-level continuation classes, respectively.

[14] Multi-word place names and organisations must also be included in the tokenizer.

5.2.2 Automatic Population of FSM Lexicons from Machine-Readable Dictionary Resources

Adding new words manually to the FSM of an inflected language is a slow and labour intensive process. For example, in order to locate the correct top-level lexical sub-category (continuation class) for an Irish noun, it is necessary to know its gender, as well as details of case and number formation. It is, therefore, highly beneficial to locate any machine-readable and printed wordlists available for the language. Ideally the lists should contain some grammatical information, which can be used to automate the process of FSM lexicon building. We were fortunate in obtaining permission to use a machine-readable version of a pocket Irish-English dictionary, *An Foclóir Póca* (An Roinn Oideachas, 1986), with about 15,000 Irish head-words (see Table 15).

Table 15 Summary of Foclóir Póca Data

POS	No. Headwords (approx.)
Noun	10700
Adjective	3020
Verb	1600
Other	340
Total	15660

Figure 18 gives an example of the type of plain-text data contained in the dictionary.

```
cabhair1 kaur' f, gs -bhrach help, assistance
cabhair2 kaur' vt, pres -bhraíonn vn -bhradh emboss, chase
cabhán kaua:n m1, ~ abhann yellow water-lily
cabhlach kauləx m1 fleet; navy
cabhrach kaurəx a1 helpful
cabhraigh kauri: vi help, ~ liom help me
```

Figure 18 Machine-Readable Dictionary Text

The source text file was cleaned up to produce a tab-separated file with 4 fields. Table 16 shows the four distinct types of information which can be used to automatically assign the headword to the appropriate FSM inflectional class.

Table 16 Sample of MRD Data

Headword	Phonetics	POS	Definition
cabhair	kaur'	f	gs -bhrach help, assistance
cabhair	kaur'	vt	pres -bhraíonn vn -bhradh emboss, chase
cabhán	kaua:n	m1	~ abhann yellow water-lily
cabhlach	kaulǝx	m1	fleet; navy
cabhrach	kaurǝx	a1	helpful
cabhraigh	kauri:	vi	help, ~liom help me

Table 16 shows headwords which include nouns, verbs or adjectives. The POS column provides the basic lexical classification for the headwords as well as gender in the case of nouns (f = feminine noun, m = masculine noun) and transitivity in the case of verbs (vt = transitive verb, vi = intransitive verb, vti = transitive and intransitive verb).[15] Some nouns and adjectives contain a number indicating a declensional class (e.g. m1, a1).

Further valuable information can be found in the definition column. For instance, "gs –*bhrach*" indicates that the genitive singular of the noun *cabhair* 'help', is formed by syncopation, i.e. dropping of vowels in the final unstressed syllable, and addition of the suffix –*ach* giving *cabhrach*. In the case of the verb *cabhair* 'help', "pres -*bhraíonn* vn –*bhradh*" indicates that the present tense is formed by syncopation of the final syllable and addition of the suffix -*aíonn*, giving *cabhraíonn*. Furthermore, the verbal noun (vn) is derived from *cabhair* by syncopation of the final syllable and addition of the -*adh* suffix, giving us *cabhradh*.

This information, together with the structure of the headword in terms of number of syllables and vowels in the final syllable, can be used in the majority of cases to automatically determine which category (continuation class) of verb, noun or adjective a particular headword should be assigned to in the FSM lexicon. The phonetic description could also be used as an aid to automatic assignment, although it was not used in this instance.

We implemented a Perl program to convert the machine-readable dictionary text to *lexc* format (Beesley and Karttunen, 2003) as shown in Figure 19. Each record is processed by first examining the POS field. In the case of verbs and adjectives, processing relies heavily

[15] The MRD has is no information about unaccusative verbs, i.e. where the verb has one NP specified but it is not the agent, e.g. Bhris an fhuinneog 'The window broke' where window experiences the action.

on the structure of the headword, whereas processing for nouns, which have a far more complex (and often unpredictable) morphology, relies on the additional morphological information found in the definition field. For example, in Figure 19, the headword *cabhair* 'help', points to continuation class `Nf5-2` which in turn points to other continuation classes that append the appropriate affixes, assign the appropriate morphological tags and insert inflectional triggers for this type of noun.

```
LEXICON Nouns
cabhair         Nf5-2;  ! Noun, feminine, class 5, sub-class 2
!!!!cabhán      Nm1-1;  ! Noun, masc., class 1, sub-class 1
!!!!cabhlach    Nm1-1;  ! Noun, masc., class 1, sub-class 1

LEXICON Verbs
cabhair   V2-BR-sync; ! Verb, conj. 2, broad stem, syncopate
cabhraigh V2-BR;      ! Verb, conjugation 2, broad stem

LEXICON Adjectives
cabhrach  Adj1-3;     ! Adjective, class 1, sub-class 3
```

Figure 19 Sample of lexc Compatible Input Automatically Derived from MRD

Despite the information available, over a third of the 10,700 nouns from *An Foclóir Póca* (An Roinn Oideachas, 1986), could not be assigned to a specific class with certainty, due in general to a lack of information about plural formation in this particular MRD. In these cases the headword was assigned the most likely sub-class given the structure of the headword, and the output was prefixed with "`!!!!`" which served to highlight the fact that the item required manual checking. At the same time it also comments out the line which causes the FST compiler *lexc* not to include it in the FSM. Overall, of the 15,000+ headwords in the MRD over 11,000 were automatically assigned to the correct FSM lexical class. On inspection of the remaining 4,000 headwords, (mainly nouns), further patterns were detected and the conversion program was amended and re-run. In the end, approximately 3,000 lemmas, (mainly nouns), had to be assigned manually using a larger paper dictionary (Ó Dónaill, 1977) for which an electronic version was not available at that time.

5.2.3 Scanning and Optical Character Recognition (OCR)

When suitable data is not available in electronic format, scanning of printed material and the use of Optical Character Recognition (OCR) software can be a viable alternative. This strategy was adopted in order to increase the number of proper nouns in the Irish FSM lexicons. Lists of towns and countries were scanned (Ó Siochfhrada, 1998), as well as a book of Irish surnames (Ó Droighneáin, 1991).

All scanned material was proof read, and scanning quality proved to be high despite the fact the OCR software was intended for Portuguese[16] rather than Irish. Approximately 5% of names contained an OCR error. Due to the nature of the material it was possible to automatically correct almost all errors: the most common errors involved a number in place of a letter, (and no numbers were expected in the input), e.g. '0' (zero) instead of 'O', '1' (one) instead of 'I', '6' (six) instead of 'ó' etc. Other common errors included 'm' in place of 'rn' and 'oh' in place of 'ch', and by searching for unusual letter combinations these were easily located and automatically corrected using the global replace operation in a word processor.

In the sample of name data in Figure 20, English surnames are followed by their Irish counterparts.

Abbott, Abóid

Acton, Ó Gnímh,

Adair, Ó Dáire

Figure 20 Sample of Scanned Data

Irish texts, especially newspapers, contain many English personal names as well as Irish names, therefore we created two lexicons, one containing Irish data and one English data (Figure 21).

```
LEXICON Names-Ir
Abóid     NP-Fam;
Gnímh     NP-Fam;
Dáire     NP-Fam;

LEXICON Names-En
Abbott    NP-Fam-en;
Acton     NP-Fam-en;
Adair     NP-Fam-en;
```

Figure 21 Sample of lexc Compatible Input Derived from Scanned Data

[16] The OCR software supplied with the scanner came with a choice of European languages, of which Portuguese was the only one containing all of the necessary diacritics, i.e. an acute accent on all vowels, both upper and lowercase.

5.2.4 Internet Sources

In a brief search, some personal names were located on the Internet[17] and included in the finite-state lexicons. The Internet is a resource which could be exploited with relatively little effort to increase the FSM lexicon, and this method merits further investigation.

5.3 Evaluation of Results of Semi-Automatic Population of Lexicons

Table 17 shows the total number of lexical items in the major part-of-speech categories after semi-automatic population had taken place. It also shows the number of surface (inflected) forms and morphological descriptions generated by inflectional rules from these headwords (stems). Surface forms in general have more than one morphological analysis. The category "Other" in Table 17 is the exception. This is made up of function words, most of which have one analysis per surface form, and in some cases, there are variant surface forms associated with the same morphological analysis.

Table 17 Extended FSM Lexicons

	Stems	Surface Forms	Morphological Descriptions
Verbs	1,630	105,000	305,100
Nouns (all):	22,100	166,100	350,600
Common nouns 10700			
Proper nouns 4200			
Proper N. (english) 7200			
Adjectives	3,035	14,100	43,900
Deverbal Nouns & Adjs.	3,220	5,305	6,436
Other	555	640	630
Total	**30,540**	**291,145**	**706,666**

[17] Symbols. http://www.symbols.net/names.htm. April 2005.

5.4 Addition of Derivational Morphology Rules

Our base-line system (Uí Dhonnchadha, 2002) implements Irish inflectional morphology. Examination of the word-forms not recognized by the FSM showed that many were derived from a root that was already in the lexicons and that the addition of derivational morphology would improve recognition rates.

In order to extract maximum benefit from the FSM lexicons, we check whether unrecognised tokens could be derived from items already contained in the FSM lexicons. In this section, we look at prefixing and suffixing of FSM lexical items, and in 5.5 we detail the evaluation of Morphological Analysis. Compounding of FSM lexical items is discussed in Section 5.6.

5.4.1 Diminutive Suffix

All nouns can accept a diminutive suffix *-ín* as in (31)a. If the final syllable of the noun is broad (i.e. ends in broad vowel *a, o, u, á, ó*, or *ú*), it must be slenderised by inserting a slender vowel i.e. *i* before attaching the slender suffix *-ín,* as in (31)b. This is achieved by including a slenderisation trigger (Uí Dhonnchadha, 2002) in the surface form which when composed with the relevant replace-rule FST will result in slenderisation taking place.

(31) a. buachaill "boy"; buachaillín "little boy"
b. rud "thing"; ruidín "little thing"

5.4.2 Emphatic Suffix

Similarly, all nouns and pronouns in Irish as well as verbs and prepositions which incorporate personal pronouns can accept an emphatic suffix. Broad and slender forms of the suffix exist, (see example (32)), therefore, rather than changing the stem, the appropriate suffix is chosen, e.g. in the case of the *-sa/-se* broad/slender pair, the *s* is added in the lexicon and either *a* or *e* is inserted by replace rule depending on the broad or slender nature of the previous syllable.

(32) a. mo theach "my house"; mo theachsa "**my** house"
b. mé "I"; mise "**I**"
c. déanaim é "I do it"; déanaimse é "**I** do it"
d. orm "on me"; ormsa "on **me**"

5.4.3 Verb and Agentive Noun Suffixes

All verb stems and agentive nouns, e.g. (33)a and (34)a, can accept one of a number of suffixes (and/or morphological processes) to create what is traditionally referred to as a

verbal noun (An Gúm, 1999, p193). Likewise, a (de)verbal adjective, e.g. (33)c, is derived from each verb stem.

(33) a. dún 'close' (verb)
b. dúnadh 'closure' (noun) 'closing' (verbal noun)
c. dúnta 'closed' (verbal adjective)

(34) a. aisteoir 'actor' (agentive noun)
b. aisteoireacht 'acting' (noun/verbal noun)

For the 1,600+ verb stems (see Table 17) in the FSM, 20 new continuation classes were included to account for the various ways in which (de)verbal nouns are derived. For the same set of verb stems, 14 new continuation classes were included to accommodate the various ways in which (de)verbal adjectives are derived. The fact that verb stems were already assigned to verbal continuation classes based on number and type of syllables speeded up the task of assigning the appropriate continuation class for (de)verbal nouns and (de)verbal adjectives, since in all cases the stem structure is relevant.

5.4.4 Other Derivational Suffixing

There are a number of other suffixes which derive verbs, nouns and adjectives from existing stems, e.g. (35)-(37). These phenomena have been implemented as guessers (due to the variety of possible combinations) and will be discussed in Section 5.7.

(35) banc 'bank' (noun)
bancáil 'bank' (verb) or 'banking' (verbal noun)

(36) aer 'air' (noun)
aereach 'airy' (adj.)

(37) ábalta 'able, capable' (adj.)
ábaltacht 'ability, strength' (noun)

5.4.5 Derivational Prefixing

Irish derivational morphology mainly involves prefixing of stems, (An Gúm, 1999), as well as some derivational suffixes already mentioned. Nouns, verbs and adjectives can all accept a range of standard prefixes, which in general do not change the lexical class.

(38) a. déan 'do/make', athdhéan 'redo/remake'
b. maith 'good', sármhaith 'excellent'
c. féasta 'feast', an-fhéasta 'great feast'

A regular relation containing over 250 common prefixes is defined. This is compiled and saved as a Prefix FST, which can be concatenated to the front of the noun FST. The boundary between the prefix and stem is marked by a boundary trigger in the surface form which when composed with the relevant replace-rule FST will result in the appropriate morphophonological processes taking place. In example (38) lenition takes place, i.e. when a prefix is joined to a stem, "h" is inserted after the initial consonant of the stem, (i.e. *déan -> dhéan*, *maith -> mhaith*, *féasta -> fhéasta*). The verb and adjective FSTs are also prefixed in the same manner.

5.5 Evaluation of Morphological Analysis Coverage

We evaluated the effects of adding additional lexicons and derivational rules to the FSM by measuring token recognition rates for the Gold Standard Evaluation Corpus (75,000 tokens approx., see Table 10). We do not carry out precision analysis on the FSM lexicons as a) they have either been hand-coded or converted from an existing MRD resources and b) they often produce multiple analyses per token. In addition, they were carefully tested at the time of their encoding and are therefore assumed to be of high quality.

Table 18 shows that the single biggest increase in token recognition is due to the use of the machine-readable dictionary, followed by the OCR scanned proper nouns.

Table 18 Coverage of Morphological Analysers

	Development Set		Test Set	
	Tokens Recognised	**% Increase**	**Tokens Recognised**	**% Increase**
FSM Transducers				
Test Lexicons	**82.79%**		**83.01%**	
MRD Lexicons	93.43%	10.64%	93.28%	10.27%
OCR Lexicons	94.54%	1.11%	94.55%	1.27%
Verbal N & Verbal Adj. Lexicons	**95.25%**	0.71%	**95.16%**	0.61%
Derivational Prefixes	95.95%	0.70%	95.84%	0.68%
Derivational Suffixes	95.98%	0.03%	95.89%	0.05%

The first four FSM transducers (Table 18) provide morphological analyses for 95% of tokens encountered in unrestricted Irish texts. Verbal noun and adjective lexicons (see Table 17), which are important in terms of POS tagging and syntactic analysis, do not have much impact on recognition rates as these word-forms have in most cases the same form as an inflected verb or noun. In effect, they provide an additional analysis to an already recognised word-form rather than providing a morphological analysis of a previously unrecognised word-form. The prefixing and suffixing morphology transducers attach to stems in the FSM lexicons. They increase recognition rates by under 1%, bringing the total to almost 96%.

So far, we have detailed the ways in which coverage of the prototype morphological analyser (Uí Dhonnchadha, 2002) was increased from 83% approx. to 96% (see Table 18) through extending the lexicons and the addition of derivational morphology rules. In the next sections we show how the use of compound identifiers and morphological guessers is an effective method for dealing with the remaining 4% approx. of unrecognised tokens.

5.6 Compound Recognition

New lexical items can be created through compounding, with nominal compounds being the most common type of compound. Irish compounds are always right-headed, therefore, the new compound word inherits the lexical features of the rightmost element.

```
(39)  a. domhainchomhrá 'deep conversation' (noun masculine)
      b. domhain 'deep' (adjective)
      c. domhain 'depth' (noun feminine)
      d. domhain 'worlds' (noun masc. common pl/gen. sg.)
      e. comhrá 'conversation' (noun masculine)
```

In (39)a we have a compound *domhainchomhrá* 'deep conversation' made up of two lexemes *domhain* 'deep' and *comhrá* 'conversation'. As is evident from the example, *domhain* is ambiguous with regard to meaning and part-of-speech. However, in this case, this is not a problem as the compound takes on the features of the rightmost element *comhrá* which is not ambiguous.

5.6.1 Compound Recogniser - Version 1

Initially, we treated nominal compounding as a very general form of prefixing, whereby all nouns in the FSM lexicons could be prefixed by any string of characters, and the compound acquires the features associated with the particular noun in question.

In Figure 22 below, we define a generic compound element, containing one or two syllables and the compound boundary marker `^CB`. This is compiled as a Compound FST and may be

concatenated to the front of the lexical noun FST to produce a new FST. Compounds which can be identified by this method are 1) not limited to the items in the FSM lexicon, as we allow any string to be prefixed to a noun lexicon item, and 2) they receive the correct morphological features from the rightmost feature.

As with previously described prefixed and suffixed FSTs (Section 5.4), inflectional rules are then composed with the new transducer to produce all possible inflected forms.

```
define V [a|o|u|á|ó|ú|e|i|é|í];         # vowels
define C [b|c|d|f|g|h|l|m|n|p|r|s|t|%-];      # consonants
define I [j|k|q|v|w|x|y|z];  # other cons. found in loanwords
define Syl1 [ (C|I) (C) (C) V (V) (V) (C) (C) (C) ];
define Syl2 [ (C) V (V) (V) (C) (C) (C) ];
define Compound [ Syl1 (Syl2) %^CB ]  ;
```

Figure 22 Extract from Compounding Regular Expression Script

5.6.2 Evaluation of Compound Recogniser - Version 1

In order to assess this method of detecting and analysing compounds we carried out both precision and recall tests using the Development Set data.

Development Set: Precision

To evaluate the precision of the compound recogniser the Development Set sentences were tokenized and analysed using the FSM and Guessers. All tokens which were tagged as being compounds (i.e. having `+GuessCmpd` tag) were extracted for inspection. This set contained 1,178 analyses relating to 159 unique tokens (i.e. types). Therefore, each token had an average of 7.4 analyses.

We took a closer look at the 159 analyses containing the `+GuessCmpd` tag, and categorised them according to whether they described genuine compounds (61 analyses), or erroneous compounds (98 analyses). This gives us a precision rate of 38% as shown below.

Precision (types): $$\frac{CorrectAutoCompounds}{AllAutoCompounds} \times \frac{100}{1} = \frac{61}{159} \times \frac{100}{1} = 38\%$$

Development Set: Precision Error Analysis

We further examined the set of incorrect analyses and sub-categorised them according to why they were misidentified as a compound (Table 19).

Table 19 Compound Recogniser 1: Error Analysis

Types tagged as Compounds	**Types**	**%**
Correctly tagged as Compound	61	38%
Incorrectly tagged as Compound	98	62%
	159	**100%**
Analysis of Incorrect Compounds		
Single Lexeme	34	35%
Typographical Error	28	29%
Dialectal Variant	26	27%
Neologism	8	8%
Foreign Word	2	1%
	98	**100%**

Most compounds identified in error (35%) were in fact single headwords not coded in the FSM lexicons, e.g. *ostáin* 'hotels' or *máguaird* 'surrounding', where the latter part of the string was identified as a headword in a FSM lexicon, i.e. (os)*táin* 'herd', (mágu)*aird* 'direction' or 'attention'.

The most effective way to prevent these types of error in future is to increase the FSM lexicons further, ideally, using MRDs. In addition, these types of error can be avoided, in many cases, by requiring that both parts of the compound be found in the FSM lexicons, e.g. in the case of *mágu-aird*, *mágu* is not a lexeme in Irish.

The category Typographical Error (29%) is difficult to remedy other than by implementing spelling correction rules. Dialectal Variant (27%) is a specific type of missing single lexeme. Neologisms (8%) (newly coined words) highlight the need for regular updating of the FSM lexicons. The number of Foreign Words identified as Irish compounds was negligible (1%).

As well as examining whether a token was a genuine compound or not, we also examined how successful this guesser was at assigning POS and feature tags (regardless of whether it was a genuine compound or not).

Table 20 shows that in the case of tokens which are genuine compounds our method assigns the correct features in the majority of cases (97%) (in two instances an adjective head was tagged as a noun). It is slightly less successful in identifying the head (95% correct) due to the fact that some compounds can be segmented in a variety of ways not all of which will be correct, i.e. the rightmost element identified as the head may contain too

much or too little material. Note also that of the incorrectly identified compounds, 50% of them received a correct POS tag. This means that overall the precision of the POS tagging was 68%, even though the success in identifying compounds was far less (38%).

Table 20 Analysis of POS and Feature Assignment to Compounds

Compound Analyses	Types	POS/Feat		Head	
		No. Correct	% Corr.	No. Correct	% Corr.
Compound	61	59	*97%*	58	95%
Non-compound	98	49	*50%*	4	4%
	159	**108**	**68%**	**62**	**39%**

Before presenting our Compound Recogniser Version 2, we present a recall-based analysis of Version 1 which further informs the development of Version 2.

Development Set: Recall

In order to assess the recall of the compound guesser we checked how many tokens which should have been recognised as compounds, were overlooked. All tokens which received a guessed analysis (1,453 types) other than compound (i.e. not `+GuessCmpd`) were examined to see if any of them should have been identified as compounds. A further 83 compounds were identified in addition to the 61 correctly identified compounds, giving a total of 144 actual compounds in the Development Set. This results in a recall rate of 42%.

Recall (types): $\frac{CorrectAutoCompounds}{ActualCompounds} \times \frac{100}{1} = \frac{61}{61+83} \times \frac{100}{1} = 42\%$

Development Set: Recall Error Analysis

The 83 unidentified compounds were then examined and categorised with a view to establishing how we could to correctly identify such compounds in future, see Table 21.

Table 21 Compound Recogniser 1: Analysis of Omitted Compounds

Types tagged as Compounds	**Types**	**%**
Compounds Recognised	61	42%
Compounds not Recognised	83	58%
	144	**100%**
Analysis of Compounds not Recognised		
Lookup Order	24	29%
Typographical Error	22	27%
Non-Noun Head	11	13%
Missing Lexeme	9	11%
Capital	9	11%
Dialectal Variant	6	7%
Neologism	2	2%
	83	**100%**

Table 21 shows that of the tokens which were not identified by the compound identifier, the most common cause (29%) was simply the order in which the guessers were run in the lookup script (see Section 5.9 for details). These tokens were analysed by a noun or adjective guesser and, therefore, never had the possibility of being identified as a compound. This can easily be rectified by changing the order in which the guessers are run. In fact some guessers needed to split into two parts, one which ran before the compound identifier and one which ran after it (i.e. Noun Guesser 1A and 1B, see page 105).

Typographical errors are the second most common reason for a compound not being identified. These misspellings prevent the stems being found in the lexicons and have the effect of making the token appear to belong to a different lexical category.

To date we have only looked for noun-headed compounds. An examination of the Non-Noun Head category shows that there are a number of other possible compounds, i.e. adjectival compounds, *idir-réaltach*[18] 'interstellar', verbal noun compounds, *idirghníomhú* 'interaction', verbal adjective compounds, *neamh-íochta*, 'non-paid'.

[18] The lexeme *idir* can be either a bound prefix morpheme meaning 'inter' or a free adjective lexeme meaning 'between'. If the Prefix Guesser had been run before the Compound Identifier both *idir-*

The other categories of Missing Lexeme, Dialectal Variant and Neologism all involve compounds in which the head element is not found in one of the FSM lexicons. The remaining category Capital contains tokens where a character other than the first letter was capitalised, e.g. *gCraobhchomórtas*, 'branch competition' or where a common noun, used as a proper noun (*óglach* 'volunteer'), was included in the compound e.g. *chorr-Óglach* 'occasional Volunteer'.

Each of these difficulties are addressed in the revised Compound Recogniser described in the next section.

5.6.3 Compound Recogniser - Version 2

The results with precision of 38% (Table 19) and recall of 42% (Table 21) are unacceptably low. As the evaluation demonstrates, the initial implementation of compound identification has several problems, most importantly the fact that it is too unconstrained.

A better approach is to ensure that both parts of the compound are found in the FSM lexicons. As before, the POS and features of the compound will be taken from the head - the rightmost element, and inflectional rules are applied to the compound as a whole (i.e. initial mutation will apply to the first element and final mutation to the second element). As iterative compounding (more than two elements) is unusual for Irish, this method captures the majority of compounds - any which have more than two stems will be handled by another guesser.

In 5.6.2, we only implemented a guesser for noun-headed compounds. The following, (40)-(44), are examples of compounds in which the head is not a noun. As a result we implement adjectival compounds. We also allow for the possibility of verb-headed compounds, although none were encountered in this test data[19].

(40) béaldúnta 'tightlipped'
béal 'mouth' (n) + dúnta 'closed' (verbal adj)

(41) buanchruthú 'stereotype'
buan 'permanent' (n) + cruthú 'creation' (verbal noun)

(42) glanghearrtha 'cleancut'
glan 'clean' (adj) + gearrtha 'cut' (v adj)

réatach 'interstellar' and *idirghníomhú* 'interaction'would have been analysed as prefixed adjective and verbal noun respectively, rather than as compounds.

[19] New verbs tend to use an existing stem with standard prefixes or else they are a neologism.

(43) mórmhéadaithe 'much increased'
mór (adj) + méadaithe 'increased' (v adj)

(44) tréithlag 'exhausted'
tréith 'weak/feeble' (adj) + lag 'weak' (adj)

In our test data, only two items with more than two elements, i.e. (45) and (46), were encountered; the first was an intensifier prefixed to a noun-adjective compound, the second was an adjective prefixed to a noun-noun compound. This phenomenon is marginal and we will not implement iterative compounding and prefixing.

(45) fíordhrochspite 'really bad spite'
fíor 'really (intensifier) + droch 'bad' (adj) + 'spite' (n)

(46) seanghráinghunna 'old shotgun'
sean 'old' (adj) + ghráin 'shot' (n) + gunna 'gun' (n)

In summary, in the second version of the compound identifier the following changes have been made:

a) compounds must consist of a head in the FSM lexicons prefixed by any word in the FSM lexicons,
b) heads may be nominal, adjectival or verbal,
c) non-initial capital letters are catered for and proper nouns are now included with nouns,
d) the lookup order has been modified.

5.6.4 Evaluation of Compound Recogniser - Version 2

The revised Compound Recogniser was also evaluated using the Development Part of the Gold Standard Corpus. The precision for Version 2 is 82% as shown below.

Precision (types): $\frac{CorrectAutoCompounds}{AllAutoCompounds} \times \frac{100}{1} = \frac{101}{123} \times \frac{100}{1} = 82\%$

Precision: Error Analysis

Table 22 Compound Recogniser 2: Error Analysis

Types tagged as Compounds	**Types**	**%**
Correctly tagged as Compound	101	82%
Incorrectly tagged as Compound	22	18%
	123	**100%**
Analysis of Incorrect Compounds		
Single Lexeme	8	36%
Typographical Error	7	32%
Proper Noun	5	23%
Dialectal Variant	2	9%
	22	**100%**

Table 22 shows that 82% of items identified as compounds are genuine compounds. The largest category of misdiagnosed compounds is still due to missing single lexemes, closely followed by typographical errors. We also count separately the missing single lexemes which are in fact proper nouns, e.g. Dubhghlas 'Douglas' analysed as *dubh* 'black' and *glas* 'green'. Most of the items which were described as dialectal variants in Table 19 (many of which are variant plurals) are now being picked up by other guessers due to the re-arranged lookup order.

Recall Test

We perform recall analysis, and find that the various changes outlined above have had a positive effect. Recall has risen from 42% to 70%.

Recall (types): $\frac{CorrectAutoCompounds}{ActualCompounds} \times \frac{100}{1} = \frac{101}{144} \times \frac{100}{1} = 70\%$

Recall: Error Analysis

We examined the omitted compounds with a view to making further improvements to the Compound Recogniser. A number of the categories present in Table 21, i.e. Look-up Order, Non-Noun Head and Capitals have been removed altogether. We are left, in Table 23, with the Missing Lexeme, Typographical Error, Proper Noun, Dialectal Variant and Neologism

categories which all require additions to the lexicons or spelling correction rules, in the case of typographical errors.

Table 23 Compound Recogniser 2: Analysis of Omitted Compounds

Types tagged as Compounds	**Types**	**%**
Compounds Recognised	101	70%
Compounds not Recognised	43	30%
	144	**100%**
Analysis of Compounds not Recognised		
Missing Lexeme	15	35%
Typographical Error	13	30%
Proper Noun	9	21%
Dialectal Variant	3	7%
Neologism	3	7%
	43	**100%**

In the next section we will look at Morphological Guessers, followed by an evaluation of their performance and their relative contributions to morphological analysis. We finish by describing the lookup strategy employed to utilise the compound identifier and the guessers.

5.7 Morphological Guessers

A lexicon of approximately 30K lemmas (see Table 17), while very useful, is still not very large. Living languages are constantly changing and acquiring new words, therefore, a method for dealing with unrecognised words will always be necessary. A Morphological Guesser makes use of distinctive suffixes, syllable structure, initial capitals and foreign characters in tokens, in order to identify possible verbs, adjectives, nouns, proper nouns, foreign words and compounds in the text which are not covered by the lexicon. We define a series of morphological guessers for Irish following Beesley and Karttunen (2003). In addition to guessing the part-of-speech, we also guess lemmas and lexical features such as gender, number and case (nouns and adjectives), or tense, number and person (verbs).

In this section we describe the various guessers and in the following Section 5.8, we present an evaluation of the guessers.

5.7.1 Verb Guessers

The inflectional suffixes of Irish verbs are distinctive in identifying verbs. Therefore, if a token is not recognized by the FSM transducer or one of the derivational transducers and it ends in one of these suffixes, we can confidently predict that it is a verb and has the verbal features associated with that suffix.

A verb is defined in terms of a generic stem to which one of the set of defined suffixes is attached. In Figure 23, `VPresentSuf` shows some of the inflectional suffixes for present tense verbs (indicative mood) and their associated person and number feature tags. The various suffixes (i.e. `VpresentSuf, VpastSuf, VfutSuf, etc.`) concatenated with `Stem` form a possible `Verb`.

```
define Stem [ Syl1 (Syl2) (Syl2) ];
define VPresentSuf [
[%+Guess %+Verb %+PresInd           .x. (e) a n n ] |
[%+Guess %+Verb %+PresInd %+1P %+Sg .x. (a) i m ] |
[%+Guess %+Verb %+PresInd %+1P %+Pl .x. (a) i m i d ] |
[%+Guess %+Verb %+PresInd %+Auto    .x.  t (e) a r ] ];
etc. etc .
define Verb [
Stem [VPresentSuf | VPastSuf | VFutSuf | VCondSuf | VImperSuf
etc. ];
```

Figure 23 Extract 1 from Verb Guesser Regular Expression Script

However, there are a small number of verb endings which are ambiguous. In (47) and (48) we show how *-aigh*, a common verbal ending, can also be an inflected form of a noun ending in *-ach*.

(47) tosaigh 'start, begin' tosaigh+Verb+Imper+2P+Sg

(48) tosaigh 'beginnings, forwards' tosach+Noun+Masc+Com+Pl
tosaigh 'beginning, forward' tosach+Noun+Masc+Gen+Sg

We handle these exceptions by generating the verb reading as usual, and by also including the alternative noun reading in the verb guesser, as shown in Figure 24.

```
#  Imperative
define VImperSuf [C [%+Guess %+Verb %+Imper %+1P %+Sg .x. a i m
]
# VERB ANALYSIS
              | [%+Guess %+Verb %+Imper %+2P %+Sg .x. a i g h ]
# NOUN ANALYSIS ALSO
              | [a c h %+Guess %+Noun %+Masc %+Com %+Pl .x. a i
g h]
              | [a c h %+Guess %+Noun %+Masc %+Gen %+Sg .x. a i
g h]
```

Figure 24 Extract 2 from Verb Guesser Regular Expression Script

5.7.2 Noun & Adjective Guessers

Irish nouns are less distinctive in their morphology than verbs. Two types of noun guessers have been implemented. The first one (i.e. Noun Guesser 1) uses stem endings and suffixes that are usually associated with a particular gender, number and case. The second type of guesser (Noun Guesser 2) (Figure 25) makes use of another generalisation, namely that words ending in a broad syllable (final vowel is either *a, á, o, ó, u* or *ú*) are usually masculine, and words ending in a slender syllable (final vowel is either *i, í, e* or *é*) are usually feminine. Nominal gender and number features are guessed based on the vowels in the last syllable in the word, and common case (nominative, accusative and dative) is assigned by default. We also attempt to distinguish between singular and plural based on the number of syllables in the word, assuming that a longer word has a suffixed plural morpheme, and by encoding some distinctive plural morphemes.

```
define CN  [b|c|d|f|g|h|l|m|n|p|r|s|t|%-|v]; # Consonants
define VL  [a|o|u|á|ó|ú|e|i|é|í]; # All vowels
define BV  [a|o|u|á|ó|ú] ;      # Broad Vowels
define SV  [e|i|é|í] ;               # Slender Vowels
define BSg [o|u|á|ó|ú] ;        # Broad Vowels singular excl. a
define SSg [e|i|é] ;            # Slender Vowels singular excl.
í

# Nouns include all strings that look like valid Irish roots.
# Allowe up to 5 consecutive consonants e.g. tonnchrith
'vibration'

define Syl    [ (CN)  CN) (CN) (VL) (VL) VL (CN) (CN) (CN)];
define BrSyl  [ (CN) (CN) (CN) (VL) (VL) BV (CN) (CN) (CN)];
define SlSyl  [ (CN) (CN) (CN) (VL) (VL) SV (CN) (CN) (CN)];

# Assume sing. nouns end in a consonant or a vowel other than a
or í
# Assume fem. nouns end in broad vowel, masc. in slender vowel
# Allow up to 3 syllables, i.e. 3 vowel clusters
# e.g. easportáil, liteagraf

define Nouns [(Syl) (Syl) (Syl) SlSyl [CN|SSg] ]
"+Guess+Noun+Fem+Com+Sg":0
```

```
           |[(Syl) (Syl) (Syl) BrSyl [CN|BSg] ]
"+Guess+Noun+Masc+Com+Sg":0 ;

# Assume plural nouns end in a or í
# allow up to 4 syllables plus plural suffix
# e.g. liteagrafanna 'lithographs', sagartóireachta
'priesthood' gen
define Femroot [(Syl) (Syl) (Syl) SlSyl];
define MascRoot [(Syl) (Syl) (Syl) BrSyl];
define NounsPl [FemRoot [%+Guess %+Noun %+Fem %+Com %+Pl .x.
[a|í]]]
         |[MascRoot [%+Guess %+Noun %+Masc %+Com %+Pl .x.
[a|í]]];
define CommNouns [Nouns | NounsPl ];
```

Figure 25 Extract from Noun Guesser Type 2 Regular Expression Script

In Figure 25 we define a broad syllable `BrSyl` and a slender syllable `SlSyl` and use these to assign either feminine or masculine gender. We assume that singular nouns can have up to 3 syllables (e.g. *cúlchistin*, 'back-kitchen') and end in a consonant, or a vowel other than *a* or *í*. We assume that plural nouns end in either *a* or *í* and we allow up to 5 vowel clusters (e.g. *cúlchistineacha* 'back-kitchens').

5.7.3 Other Guessers

We also developed guessers to handle proper nouns, abbreviations, foreign words. Finally, any remaining tokens are given the tag `+Unknown`.

5.8 Evaluation of Guessers

We evaluated the precision of each of the guessers as well as their impact on recognition rates. We automatically processed the Development Set, and all tokens which received a guessed analysis were extracted for inspection. This gave 1,929 unique tokens (i.e. types). Analysis of guesser results are given in Table 24. As shown below, the overall precision for all guessers (including the compound identifier) is 91%.

Precision Test

Precision (types): $\frac{CorrectGuesses}{AllGuesses} \times \frac{100}{1} = \frac{1{,}760}{1{,}929} \times \frac{100}{1} = 91\%$

In Table 24 we give a breakdown of the precision for each individual guesser.

Table 24 Development Set: Guesser Precision

Development Set	Tokens			Types		
Guessers	**Count**	**Correct**	**% Prec.**	**Count**	**Correct**	**% Prec.**
Prefixed Noun, Adj, Verb	351	325	**93%**	283	260	**92%**
Suffixed Noun, Adj, Verb	14	14	**100%**	12	12	**100%**
Compounds	120	98	**82%**	106	87	**82%**
Verbs	143	78	**55%**	127	69	**54%**
Nouns, Adjs 1A	179	166	**93%**	149	140	**94%**
Nouns, Adjs 1B	193	171	**87%**	164	147	**94%**
Nouns 2	323	308	**95%**	297	285	**96%**
Proper Nouns, incl. Abbrevs.	725	721	**99%**	629	625	**99%**
Verbal Nouns/Adjs	95	82	**86%**	87	75	**86%**
Foreign Words	48	44	**92%**	45	41	**91%**
Unknowns	33	20	**60%**	30	19	**63%**
	2224	**2027**	**91%**	**1929**	**1760**	**91%**

As the verb guesser is the only guesser which is not performing as well as expected, we take a closer look at these results.

Table 25 Verb Guesser: Error Analysis

Types guessed as Verbs	**Count**	**%**
Correctly tagged as verb	78	55%
Incorrectly tagged as verb	65	45%
TOTAL	**143**	**100%**
Analysis of Incorrectly guessed verbs		
Proper noun	35	54%
Noun (plural)	30	46%
	65	**100%**

We find that the data (Development Set) contains many Irish place names and personal names whose final syllable makes them look like a verb form, e.g. *Tor**aigh*** 'Tory', *Uisn**igh*** 'Uisnigh', *Chróc**aigh*** 'Croke', *Mal**ainn*** 'Malin', etc. This problem can be alleviated by generating both noun and verb analyses (see Figure 24, p100) and allowing the POS disambiguator at a later stage in the processing pipeline to decide which is the correct

analysis in the context. The addition of more Named Entities (place names and personal names) would also help alleviate this type of error.

The other category of misdiagnosed verb, is nouns with plurals ending in *-aí,* e.g. *aireachtaí* 'ministries', *liomatáistí* 'extents', some of which are non-standard plurals, e.g. *cantaireachtaí* 'chantings', *aighneachtaí* 'complaints'. The generation of non-standard plurals in the FSM lexicons would help alleviate this second type of problem.

We also evaluated on the Test Set, and calculate the overall precision rate as 85%.

Precision Test

Precision Test Set (types): $\frac{CorrectGuesses}{AllGuesses} \times \frac{100}{1} = \frac{867}{1{,}021} \times \frac{100}{1} = 85\%$

In Table 26 we give a breakdown of the precision for each individual guesser. The precision rates, as expected, are lower than for the Development Set, and are probably more representative of the corpus as a whole, as some problems specific to the Development Set data were rectified during the iterative development of the guessers. However, the Verb Guesser stands out as the guesser most underperforming. We believe that this is due to the influence of Proper Nouns and non-standard plurals, as seen in the Development Set. If the Verb Guesser numbers are omitted from Table 26, the overall precision is 88% (rather than 85%) which is closer to the Development Set precision of 91%.

Table 26 Test Set: Guesser Precision

Test Set	**Tokens**			**Types**		
Guessers	**Count**	**Correct**	**% Prec.**	**Count**	**Correct**	**% Prec.**
Prefixed Noun, Adj, Verb	171	164	**96**	149	142	95
Suffixed Noun, Adj, Verb	11	11	**100**	11	11	100
Compounds	40	30	**75**	34	24	71
Verbs	80	37	**46**	74	34	46
Noun, Adj 1A	72	69	**96**	65	63	97
Noun, Adj 1B	116	89	**78**	99	73	74
Noun 2	197	134	**68**	170	119	70
Proper Nouns	366	364	**99**	339	337	99
Verbal Nouns, Adjs	48	43	**89**	42	37	88
Foreign word	21	17	**81**	19	16	84
Unknown	23	11	**48**	19	11	59
	1145	**969**	**85%**	**1021**	**867**	**85%**

5.9 Morphological Analysis Lookup Strategy

The FSTs developed in this implementation are designed to be used with the Xerox *lookup* tool (Beesley and Karttunen, 2003). This tool allows the developer to specify a series of transducers and the order in which they should be used.

The order in which the guesser transducers are run in the script is very important as the Xerox lookup stops searching as soon as the first match is found. The FSM transducers are tried first and are followed by the guessers. Irish verbal endings are the most distinctive morphological feature, followed by suffixed and prefixed word forms, verbal nouns and verbal adjectives, etc. If none of these more specific transducers finds a match, the lookup script will continue on to try a more general guesser. If a loosely constrained noun guesser were tried first, it would match with most patterns including a string with a distinctive verb ending. Therefore, the most restrictive guesser should be tried first (i.e. verb guesser) followed by the next most restrictive and so on, ending with the most general transducer (i.e. foreign noun).

The lookup utility also enables "virtual composition" whereby a transducer can be composed with another transducer on the fly where necessary. In the Irish morphological analyser, all lexical items are defined in lowercase in the lexicons, except for proper nouns that are unlikely to be used without an initial uppercase character, e.g. 'Dublin', 'London', 'Paris'. If the lexical transducer does not recognize a word, it may be because the word occurs at the start of a sentence and has been capitalized. A transducer, following Grefenstette *et al*. (2000), is defined which maps the initial letter of word forms (49)a, or the second letter (49)b or third letter (49)c, to a capital letter, or all letters to capitals (49)d. These capitalizing transducers can be composed with the lexical transducer on the fly, and the resulting transducers are tried in the lookup script before going on to try the guesser transducers. Each guesser is also composed with the capitalisers to gain maximum benefit.

(49) a. *Uachtarán* 'President'– Initial capital
b. *na hÉireann* 'of Ireland' – Capital with initial mutation
c. *cúrsaí i bhFiontar* 'courses in Fiontar' - Capital with initial mutation
d. *TÁ* 'is'- All capitals

Currently, the following morphological analysers and guessers are run in the following order :

- FSM lexicons
- Numbers
- FSM lexicons with Capitals
- Prefixed Adj Lexicons
- Prefixed Adjs with Capitals
- Prefixed Noun Lexicons
- Prefixed Nouns with Capitals
- Prefixed Verb Lexicons
- Prefixed Verbs with Capitals
- Suffixed Noun Lexicons
- Suffixed Nouns with Capitals
- Verb Guesser
- Verb Guesser with Capitals
- Noun Guesser 1A
- Noun Guesser 1A with Capitals
- Compound Noun Guesser with Capitals
- Compound (Capitalised Head Noun) Guesser
- Proper Noun Guesser
- Compound Guesser with Capitals
- Verbal Noun/Adj Guesser
- Noun Guesser 1B
- Noun Guesser 1B with Capitals
- Compound Adj. Guesser
- Noun Guesser 2
- Noun Guesser 2 with Capitals
- Foreign word guesser
- Unrecognised forms

5.10 Summary of Token Recognition Rates

The effect of the additional Guesser FSTs on recognition rates is shown in Table 27. These coverage results were obtained using both Test and Development Sets of Gold Standard Corpus (for further details see Table 10 Composition of Gold Standard (3000) POS Corpus, p60).

Table 27 Summary of Token Recognition Rates

	Development Set		Test Set	
	Increase in Recognition	**Tokens Recognised**	**Increase in Recognition**	**Tokens Recognised**
FSM Transducers				
Initial Test Lexicons		**82.79%**		**83.01%**
MRD Lexicons	+10.64%	93.43%	+10.27%	93.28%
OCR Lexicons	**+1.11%**	94.54%	+1.27%	94.55%
Verbal N & Verbal Adj. Lexicons	+0.71%	95.25%	+0.61%	95.16%
Derivational Prefixes	+0.70%	95.95%	+0.68%	95.84%
Derivational Suffixes	+0.03%	95.98%	+0.05%	95.89%
Total Lexicon Increase	**+13.19**		**+12.88%**	
Morphological Guessers				
Verb Guesser	+0.33	96.31	+0.37	96.26
Noun Guesser 1A	+0.37	96.68	+0.30	96.56
Compound Recogniser (v2)	+0.25	96.93	+0.17	96.73
Proper Noun Guesser	**+1.68**	98.61	+1.67	98.40
Deverbal Noun/Adj Guesser	+0.17	98.78	+0.15	98.55
Noun Guesser 1B	+0.39	99.17	+0.47	99.02
Compound Adj Recogniser	+0.00	99.17	+0.00	99.02
Noun Guesser 2, incl. Abbrev.	+0.66	99.83	+0.79	99.81
Foreign Word Guesser	+0.10	99.93	+0.09	99.90
Unknown Item Guesser	+0.07	100.00	+0.10	100.00
Total Guesser Increase	**+4.02%**		**+4.11%**	
Total Increase	**+17.21%**		**+16.99%**	

It is worth noting that although the Proper Noun Guesser makes a bigger impact on recognition rates (1.68%) than the OCR lexicons (1.11%), the analysis provided from the OCR lexicons is more reliable in terms of morphological features (i.e. gender, number and case). Therefore, further work should concentrate on improving the 96% recognition rate

from the FSM lexicons through further use of MRDs[20] where possible and OCR where necessary.

5.11 Summary

In this chapter we first described the semi-automatic extension of the finite-state lexicons and the addition of rules for derivational morphology. We then turned our attention to finite-state compound recognition, and followed this with a description of the implementation of finite-state verb, noun and adjective guessers. The guesser transducers analyse tokens which are not defined in the finite-state morphology (FSM) lexicons. We ended by presenting the lookup strategy for morphological analysis and gave a summary of token recognition rates.

We evaluated the performance of the guesser tools using precision tests. We also carried out coverage tests by calculating token recognition rates.

While the lexicon-based morphological analyser is performing well (95% recognition rates), it could benefit from significantly increasing the finite-state lexicons through the conversion of additional MRDs, supplemented with scanned data and perhaps word lists sources from the Internet. An increased lexicon will have positive benefits for derivational morphology and compounding recognition, as well as reducing the burden on the morphological guessers.

Two problems which will not be alleviated by including MRDs etc. are typographical errors in the input text and dialectal variants. The performance of the guessers could be improved by including a spelling correction module to handle suspected typographical errors. The issue of dialectal variants could be addressed through the generation of regular variants in the lexicon, e.g. dialectal plural forms, and the inclusion of dialectal affixes in the guessers.

The average precision of the guessers is 85% on the Test Set data. The only guesser which showed an unexpected low precision rate was the Verb Guesser. It transpires that many Irish place names and some of the dialectal variants of plural nouns, end in a syllable identical with that of a verbal affix. These problems will be alleviated by the inclusion of further Named Entities and by making provision for dialectal variants, as mentioned above.

In the next section we will look at disambiguating the output of the morphological analysers and guessers in order to arrive (in most cases) at a single POS tag per token.

[20] Work has begun on converting an additional 30K headwords from a MRD (Ó Dónaill, 1977) which has recently become available. We expect this to significantly improve token recognition rates.

6 POS Tagging Using Morphosyntactic Disambiguation

6.1 Introduction

In order to create part-of-speech (POS) tagged text, we must associate one morphosyntactic tag with each token. In more than 60% of cases, the morphological analyser, presented in the previous sections, provides more than one possible analysis, and we must choose the correct analysis from the choice available, i.e. we must disambiguate.

We have developed hand-written, language specific, rules which use the local context within a sentence to determine the most likely POS tag from the choice presented by the morphological analyser. These rules follow the Constraint Grammar (CG) formalism (Karlsson et al., 1995).

CG is described by Karlsson (1995, p1) as "a language-independent formalism for surface-oriented, morphology-based parsing of unrestricted text". By 'language-independent' we mean that the program code executing the grammar rules, and the grammar rules themselves are clearly separated; the same program code can be used for any new language for which language specific grammar rules are written.

Morphology-based parsing in the context of CG consists of the assignment of morphological features and shallow syntactic structure (clause boundaries, grammatical function tags and dependency relations). CG handles the task of (shallow) parsing in 3 stages:

- morphological disambiguation (POS tagging)
- assignment of clause boundaries
- assignment of surface syntactic function labels and dependency relations

In Section 6.2, we provide an introduction to Constraint Grammar. In Section 6.3, we present examples of the various types of CG rules used for morphological disambiguation of Irish, and in Section 6.4 we highlight some of the more challenging aspects of the POS disambiguation of Irish texts. In Sections 6.5 and 6.6 we present our evaluation measures, and report the results.

The assignment of clause boundaries, surface syntactic function labels and dependency relations will be described in Part III of this dissertation.

6.2 Principles of Constraint Grammar

The main aims and principles of Constraint Grammar (CG) as described by Karlsson *et al.* (1995) are as follows:

- to assign the appropriate morphological and syntactic information according to the context of each token or larger structure in the text;
- to assign an analysis to every string in the input, bearing in mind that unrestricted text will contain typographical errors, non-sentential fragments, dialectal and colloquial material;
- if an ambiguity cannot be resolved, the alternative analyses are retained rather than forcing a (possibly incorrect) choice.

Disambiguation is achieved through either selecting the only possible analysis given the context, or alternatively rejecting all of the impossible/improbable analyses until only one presumably correct analysis remains. CG relies heavily on lexical and morphological features, as well as word order configurations. CG rules ideally are unordered and mutually independent, although this is not always possible in practice.

CG rules are applied to the output of the morphological analyser.[21] CG operates at sentence level. A sentence is described in terms of *cohorts*, *readings* and *tags*. A cohort consists of a token and all the possible readings for that token. Each reading consists of tags, which include the lemma and morphosyntactic tags. Figure 26 shows a sentence fragment *Tháinig an bháisteach*, 'The rain came' (see p55 for detailed gloss) which has three cohorts. The second cohort which contains the token *an* 'the' has three possible readings. In this case the lemma as well as the morphosyntactic tags is ambiguous, i.e. *an* or *is*.

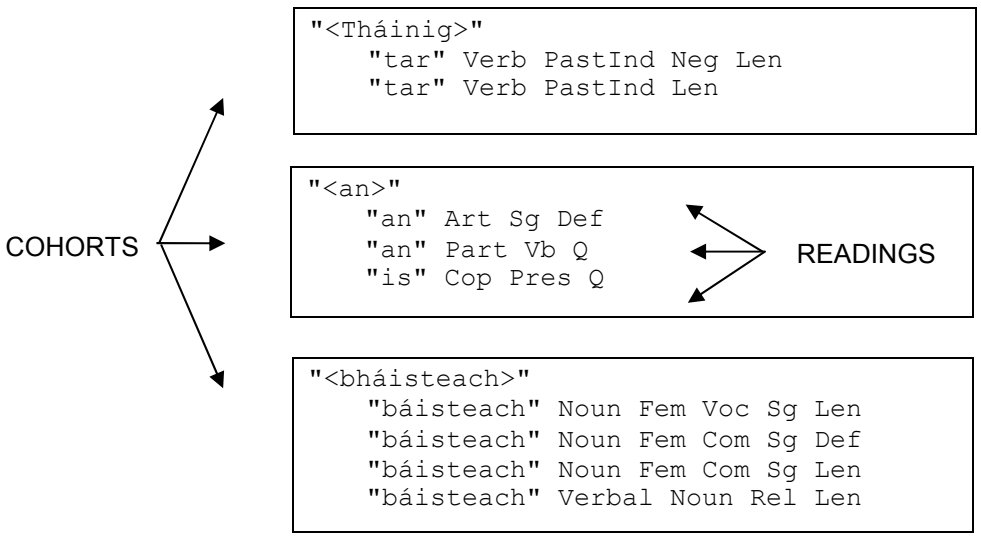


Figure 26 CG Cohorts and Readings

[21] The output of the morphological analyser is slightly modified using a Perl script to meet the required CG input format as shown in Figure 26.

In order to select the most likely morphological analysis for an ambiguous token, CG uses other cohorts within the sentence. A positional reference system is used, whereby the cohort under consideration is at position `0`, the following cohort is at position `1` and the previous cohort is at position `-1`, and so on. It is also possible to specify that a tag must exist somewhere to the left in the sentence `*-1`, somewhere to the right in the sentence `*1`, at the start of the sentence `@1`, or at the end `@-1`.

6.2.1 CG Rule Syntax

CG has two basic types of rule; '`SELECT`' and '`REMOVE`'. The input is disambiguated by either selecting one reading from a cohort based on the context to the left and/or right of the token or by removing impossible readings based on the context. The last remaining reading is never removed. Example (50) shows a rule where the noun reading is selected if the previous token is unambiguously (`C` means careful mode) an article (in Irish adjectives follow the noun), and in (51) the verb reading is removed if the previous token is unambiguously an article.

```
(50)  SELECT (Noun) IF (-1C (Art));
(51)  REMOVE (Verb) IF (-1C (Art));
```

The rule syntax is straightforward and rules are intuitive to encode. A rule can contain many conditions and each condition can refer to many tags including the word-form or lemma. Using the word-form is more restrictive than using the lemma. For example, in Figure 26 if we use the word-form `"<an>"` (52), this rule will only apply to the form *an* 'the (`Sg`)', whereas if we use the lemma `"an"` (53), this will include associated (inflected) forms such as *na* 'the (`Pl`)' also.

```
(52)  SELECT (Noun) IF (-1C ("<an>"));
(53)  SELECT (Noun) IF (-1C ("an"));
```

When specifying a rule, tags must be listed in the order in which they appear in a reading but not every tag must be specified, e.g. (`Noun Sg`) will include all singular nouns regardless of any other intervening tags, e.g. gender or case, which may be present, e.g., a rule specifying (`Noun Sg`) will include such readings as `"bháisteach" Noun Fem Com Sg Len`, ('rain').

```
(54)  REMOVE (Noun Sg) IF (-1C (Art Pl));
```

This allows the flexibility to write very general or very specific rules. (54) is a more specific than (50), in that it specifies a singular noun. It also means that CG rules are independent of many changes to the morphological analysis module, e.g. new tags can be introduced and

as long as the sequence of existing tags is maintained the existing CG rules will be unaffected.

Being able to specify a subset of tags is also an advantage when dealing with complex word-forms. Several tokens in Irish are contractions of the article and a preposition, e.g. *don* 'to the' which is a contraction of the preposition *do* 'to' and the article *an* 'the'. The morphological analyser tags *don* as `"do" Prep Art`. The rule in (50) above will apply to *don fhear* 'to the man' in the same way as it would apply to *an fear* 'the man'. The noun reading of *fear* 'man' will be selected as the preceding token in both cases (i.e. `"an" Art` and `"do" Prep Art`), contains the `Art` tag.

As already mentioned, rules can also be generalised by using the `*` operator, where `*-1` means somewhere to the left, and `*1` means somewhere to the right. The search can be constrained using the BARRIER function which prevents searching past a named tag type. The following rule, (55), states that the verb reading of the current word should be selected if there is no verb or copula to the left. There must not be a verb to the right (looking no further than a relative clause marker).

```
(55)  SELECT (Verb) IF
            (NOT *-1 (Verb))
            (NOT -1C (Cop))
            (NOT *1 (Verb) BARRIER (Rel));
```

Sets of tags or lexemes may be defined using the LIST keyword and used later in rules. For example we can list all the possible pre-modifiers of a noun as in (56) and use it to generalise a rule as in (57).

```
(56)  LIST NOUN-PREMODIFIER = (Art) (Det Poss) (Det Qty) (Num);

(57)  SELECT (Noun) IF (-1C NOUN-PREMODIFIER);
```

However, there are some weaknesses in the system. Because the rules are strictly positional in nature, it is currently not possible to use regular expression-like operators e.g. to specify that a noun can be followed by zero or more adjectives. Also, there is no elegant way of dealing with optional items e.g. quotes around a lexical item. For example, if we say that a verb can follow a relativizer such as *a* 'that' in (58) the rule will not hold as there is a quotation mark between *a* 'that' and the verb *ainmníonn* 'name'.

```
(58)  ...a    ' ainmníonn agus a    náiríonn ' iad
      ...that ' name      and  that shame    ' them
```

Also, although many conditions can follow the IF keyword, they are all "anded" together, i.e. IF (1=x) (2=y) (3=z) etc. But, although you can say IF (1= x OR y), there is no means of saying IF (1=x) OR (2=y). For example, relative clauses are introduced by a relative particle followed by the verb, but in one verb form the relative particle and verb are combined. Therefore, if we wish to specify a relative clause we may need to be able to express that either the relative verbal particle precedes the current verb OR that the current verb is a relative verb form. However the OR in statement (59) is not allowed in CG2,[22] the version of CG currently used in this work.

```
(59)  *IF (-1 (Part Vb Rel)) OR (0 (Verb Rel))
```

6.2.2 Rule Ordering and Rule Interaction

CG rules are intended, ideally, to be unordered and independent of each other. This is not always possible, and certainly if we are using any kind of heuristics we would wish them to apply after all of the safer rules have been applied. This can be achieved by grouping rules together in blocks called "sections". Sections are applied in order, and within a section certain priorities apply. Firstly, a global preference can be set at the outset with the `PREFERRED-TARGETS` variable, as in (60), where tags are listed in order of preference. If, for instance, there is a general ambiguity between the past tense form and the present subjunctive form of verbs, we may wish to favour the past tense. Using the `PREFERRED-TARGETS` variable, priority will be given to rules which 'select' the past tense or 'remove' the present subjunctive form, over rules which 'remove' the past tense or 'select' the present subjunctive form. Secondly, 'select' rules, in general, take precedence over 'remove' rules (when equally applicable) and apart from these priorities the rules are applied in the order in which they appear.

```
(60)  PREFERRED TARGETS = Past PresSubj Pron Noun;
```

Disambiguation may be carried out in cycles using sections. In the first cycle, the first section is applied alone; when no more disambiguation of the text can be carried out, the first and second sections are applied together; then the first, second and third sections are applied, and so on. By grouping the safest rules in the first block, and putting less safe rules, or rules which require some prior disambiguation, in subsequent blocks, the order of application of rules can be influenced. In CG2 syntax, a new section is created each time the keyword `CONSTRAINTS` is used in the code, as shown in Figure 27.

[22] This is implemented in CG3 (http://beta.visl.sdu.dk/cg3.html) the latest version of VISLCG.

```
# CONSTRAINT GRAMMAR (CG2) CODE OUTLINE
# ========================================================== #
# SENTENCE DELIMITERS
# ========================================================== #
DELIMITERS = "<.>" "<!>" "<?>" "<#>" "<</p>>" "<</s>>" "<...>";
PREFERRED-TARGETS = Pron Noun PastInd PresSubj ;
# ========================================================== #
# SETS
# ========================================================== #
SETS
LIST PUNCT-INT = (Punct Int) (Punct Bar) (Punct Brack);
LIST OBJ-PRON = "í" "é" "iad" ;
# ========================================================== #
# DISAMBIGUATION RULES
# ========================================================== #
CONSTRAINTS
# SECTION 1 - Definite Rules
SELECT (Noun Sg) IF (0 ("cor")) (1 ("ar")) (2 ("bith"));
SELECT (Verb) IF (1 (Sbj));
REMOVE (Verb Auto) IF (1C (Sbj));

CONSTRAINTS
# SECTION 2
# MORE RULES
CONSTRAINTS
# SECTION 3
# MORE RULES

#=====#
END #
#=====#
```

Figure 27 Example of CG2 Syntax

6.3 CG Morphosyntactic Disambiguation Rules for Irish

We developed over 425 disambiguation rules for Irish, which achieve an f-score of 95% on Development Set data and 94.35% on Test Data (see Section 6.6). The English Constraint Grammar, ENGCG, achieved 93-97% precision using 1,100 rules (Karlsson et al., 1995, p39; Tapanainen and Voutilainen, 1994; Voutilainen, 1995, p186). Approximately 40% were lexical rules rather than linguistic generalisations, i.e. they have a word-form/lemma as target rather than morphosyntactic features (Voutilainen, 1995, p179). A similar proportion holds true for the Irish rules.

In the current application to Irish, we have categorised our CG rules into the following 5 sections:

- Universal Safe Rules which always choose the correct analysis
- Qualified Safe Rules which always choose the correct analysis after some other possible analyses have been eliminated
- Idioms and lexically specific rules
- Strong Tendency Rules which are almost always correct
- Most Likely Rules which are correct more often than not.

Some examples of each type of rule are discussed below.

6.3.1 Universal Safe Rules

These are rules which can be guaranteed to always choose the correct morphological analysis given a particular context, regardless of other possible analyses for the token.

There are specific subject pronouns in Irish which always follow a verb, (61), as a finite verb[23] cannot be separated from its subject. Therefore we can confidently select the verb reading for *rinne* using rule (62), even though, *rinne* also has possible noun readings. We can also say with confidence that a particular verb cannot be in the autonomous form (unspecified subject), since it is followed by a subject pronoun (rule (63)).

```
(61)  Rinne siad      é
      Did   they-SUBJ it
      'They did it'.

(62)  SELECT (Verb) IF (1 (Pron Sbj));

(63)  REMOVE (Verb Auto) IF (1 (Pron Sbj));
```

There are many safe rules associated with particles. For instance, a token cannot be a numeral particle if it does not precede a numeral, as shown in (64) and (65). This rule is expressed in (66).

```
(64)  a        ceathair a  clog
      PART-NUM four     of clock
      'four o' clock'

(65)  Dé Satharn ar a        4
      Saturday   at PART-NUM 4
      'Saturday at 4'
```

[23] Excluding phrasal verbs where the subject is embedded in a prepositional phrase.

```
(66)  REMOVE (Part Nm) IF (NOT 1 NUM-COUNT OR (Num Dig));
```

6.3.2 Qualified Safe Rules

These rules also reliably choose the correct morphological analysis, but they may require some prior disambiguation of neighbouring tokens. They are characterised by the "C" (careful mode) flag. For example, we can say that a particular token cannot be a verb if it is directly followed by something which is unambiguously an adjective. In example (67), *líon* can be a verb 'fill' or a noun 'quantity or number', but the verb reading can be eliminated if it is followed by an unambiguous adjective such as *beag* 'small'. Many adjectives also have other readings (particularly noun readings) - but rule (68) only holds if other such reading have already been discarded, or the token was unambiguously an adjective to begin with.

```
(67)  líon       beag  tithe
      number small houses
      'a small number of houses'

(68)  REMOVE (Verb) IF (1C (Adj));
```

Likewise, we can select the adverbial particle reading of *go* (particle functioning equivalently to -ly in English), as in (70), if it is followed by an unambiguous adjective.

```
(69)  Rith siad      go      tapaidh
      Ran  they-SUBJ PART-AD quick
      'They ran quickly'.

(70)  SELECT (Part Ad) IF (1C (Adj));
```

6.3.3 Idioms and Lexically Specific Rules

There are many set phrases or idioms whose individual tokens are ambiguous but when they occur together we can analyse them with certainty. The idiomatic phrase *ar chor/cor ar bith* 'at all' is very common and contains ambiguous tokens, but we can resolve all its elements early on: e.g. rule (71) selects the singular noun reading for *cor* 'turn'. By using the lemma `"cor"` rather than the word-form `"<cor>"` we include inflected forms such as *chor*.

```
(71)  SELECT (Noun Sg) IF (0 ("cor")) (1 ("ar")) (2 ("bith"));
```

An example of a lexical rule is given in (72) and (73). The word *ann* is almost always a prepositional pronoun meaning 'in it' or 'there', except in the phrase *in ann* where it means

'able' and is analysed as a substantive noun *ann*[24] following the preposition *in* 'in'. (A substantive noun is a word which behaves like a noun but has no other inflected forms). If *ann* is preceded by the token *in* we select the noun reading and if it is not, we discard the noun reading.

```
(72)  REMOVE (Subst Noun) IF (0 ("ann")) (NOT -1 ("<in>"));
```

```
(73)  SELECT (Subst Noun) IF (0 ("ann")) (-1 ("<in>"));
```

6.3.4 Strong Tendency Rules

In this category we have rules which are almost always true, but there can be occasional exceptions.

Adverbs used as intensifiers must be followed by an adjective, e.g. *breá te* 'pleasantly hot', *sách ard* 'fairly high', so we select this interpretation whenever possible. For example, the word *sách* in (74) can be an adjective, adverb or noun, but if it is followed by a (possible) adjective such as *ard* (tall, high place) we choose the intensifying adverb reading.

```
(74)  sách   ard
      fairly tall
```

```
(75)  SELECT (Adv Its) IF (1 (Adj));
```

The reason that this rule may not always give the correct result is that the adjective can also have other readings (e.g. as a noun), in some less likely context *sách* could in fact be the noun meaning 'well-fed person'.

Rules which favour a very common interpretation over a very rare interpretation of a lexeme fall into the Strong Tendency category also.

```
(76)  SELECT (Verb) (0 ("abair")) (-1 ("<a>"));
```

Deir is usually the past tense of the verb lemma *abair* 'say', although it has a much rarer noun meaning of 'shingles' or 'herpes'. Theoretically *a deir* could mean 'her shingles or herpes', but there is a far greater probability that it simply means 'that said'.

[24] *in ann* 'able' was formerly *i n-ann* (O' Neill Lane, 1916), which was formerly *i n-ion* 'in fitness/worthiness/possibility', i.e. 'able' (Dineen, 1934).

6.3.5 Most Likely Rules

This category contains rules which are clearly heuristics. These rules deal with constructs that are difficult to resolve with certainty. As in example (77) *ar* is predominantly a preposition meaning 'on', but it can less commonly be used as a verb also meaning 'said' as in (78). Using rule (79) we remove the verb reading of *ar* except where it is preceded by punctuation signalling direct speech, such as a quotation mark or comma. This will be correct in the great majority of cases but we know that it is not universally true, i.e. such punctuation may not always be present, i.e. (77) could possible mean 'It/He fell said Liam'.

```
(77)  Thit sé ar Liam
      Fell it on Liam
      'It fell on Liam'

(78)  Cá    bhfuil sé?', ar    Liam
      Where is     he?', said Liam
      'Where is he?', said Liam
```

We risk making an occasional error in order to solve a great number of ambiguities through removing the unlikely verb reading and leaving the frequently occurring reading of preposition, e.g. *ar* 'on'.

```
(79)  REMOVE (Verb PastInd) IF (0 ("ar")) (NOT -1 (Punct Int) OR
      (Punct Quo));
```

6.3.6 Testing and Debugging

As more and more rules are added, testing and debugging becomes more of an issue. It can often be difficult to tell what combination of rules interacted to result in a particular (erroneous) analysis of a sentence. Using the `-trace` flag with the CG2 software, we can see exactly which subset of the rules was used to disambiguate a sentence. An example sentence fragment (80) and its morphological analysis is given below.

```
(80)  Labhair sé faoi   ...
      Spoke   he about ...
      'He spoke about ...'

"<Labhair>"
      "labhair" Verb VTI PastInd Len
      "labhair" Verb VTI PastInd Neg Len
      "labhair" Verb VTI PastInd NegQ Len
      "labhair" Verb VTI PastInd Q Len
```

```
      "labhair" Verb VTI Imper 2P Sg
      "labhair" Verb VTI Imper 2P Sg Neg
"<sé>"
      "is" Cop Pres Pron Pers 3P Sg Masc
      "sé" Prop Noun Masc Com Sg
      "sé" Num Card
      "sé" Num Card Ecl
      "sé" Noun Masc Com Sg
      "sé" Noun Masc Com Sg Ecl
      "sé" Noun Masc Com Sg DefArt
      "sé" Noun Masc Gen Sg
      "sé" Noun Masc Gen Sg Ecl
      "sé" Pron Pers 3P Sg Masc Sbj
"<faoi>"
      "faoi"     Prep Simp
      "faoi"     Pron Prep 3P Sg Masc
```

The following is a listing of the rules used to disambiguate the verb *labhair* 'spoke' in the sentence fragment in (80), presented in the order in which they were applied. This is a very useful facility for tracking rules which are interacting in an unexpected manner.

```
REMOVE TARGET (Verb Neg) IF (NOT -1 (Part Vb Neg)) (NOT 0 ("<níl>")
OR ("<Níl>")) ; # line 237
REMOVE TARGET (Verb NegQ) IF (NOT -1 (Part Vb NegQ)) ; # line 252
REMOVE TARGET (Verb Q) IF (NOT -1 (Q)) ; # line 257
REMOVE TARGET (Verb Imper) IF (1 (Sbj)) ; # line 262
```

The result of disambiguating all three tokens is given below:

```
"<Labhair>"
      "labhair" Verb VTI PastInd Len
"<sé>"
      "sé" Pron Pers 3P Sg Masc Sbj
"<faoi>"
      "faoi" Prep Simp
```

There is also a `-debug` flag which can be used to evaluate the accuracy of the rules. Rules can be run against a manually disambiguated text where the correct analyses have all been manually identified and marked with the `<Correct!>` tag. Any rules that would remove an

analysis marked as correct are highlighted. This is a very effective way of looking for possible problems in the rules.

A Perl script was used to append the `<Correct!>` tag to each analysis in the gold standard development corpus. (Note: each analysis was written out twice to force the CG parser into disambiguating, as tokens with only one analysis are ignored by the CG parser since they do not require disambiguation).

6.4 Disambiguation Challenges

In this section, we highlight three of the more challenging ambiguities in automatic POS tagging for Irish. Firstly, the particle *a* can have a great variety of functions, secondly, many of the most commonly used numbers are homonymous, and thirdly we touch upon some ambiguities associated with homonymous forms of the copula *'is'*.

6.4.1 Multi-Functional Particle *a*

The token *a* has more possible analyses than any other item in the Irish dictionary (Ó Dónaill, 1977). A list of 11 functions is given below. As a functional particle, *a* determines the type of phrase, and choosing the wrong analysis can have many knock-on effects for the sentence as a whole.

1. *a* in Noun Phrase

a [POSS DET FEM]	*teach* [NOUN]	'her/its house'
a [POSS DET MASC]	*theach* [NOUN]	'his/its house'
a [POSS DET PL]	*dteach* [NOUN]	'their house'
a [ART ABBREV]	*tí* [NOUN GEN]	'... the house'
a [VOC PART]	*Sheáin* [NOUN VOC]	'O Seán'

2. *a* in Infinitival Phrase

a [INFPART]	*dhéanamh* [VERBALNOUN]	'to do'

3. Relative Verb Phrase

a [DIRECTREL]	*bhris* [VERB]	'that broke'
a [INDIRECTREL]	*mbris* [VERB]	'that broke'
a [RELPRONOUN]	*bhí* [VERB]	'that which/who was'

4. Number Phrase

a [NUMPART]	*trí* [NUM]	'three' – counting, time phrases etc.

5. Focussed/Emphatic Clause

a [DEGPART] *géire* [COMPADJ] 'how sharp(ly)'

The token *a* comprises 4.2% (2,124 instances) of the 50K approx. tokens in the Development Set (Table 28). Even though the majority (96%) of instances of the particle *a* are automatically tagged correctly by our CG rules, the remaining 85 instances constitute the token which in terms of raw frequency is most frequently tagged incorrectly.

Table 28 Disambiguation: Error Analysis of Token *a*

Development Set Tokens	**Count**	**%**
Token *a*	2,124	4.2%
All other tokens	48,027	95.8%
	50,151	**100%**
Token *a* in Development Set		
Correctly tagged	2039	96%
Incorrectly tagged	85	4%
	2,124	**100%**

The problems associated with the incorrectly tagged instances of the token *a* are detailed in the confusion matrix in Table 29.

Table 29 Confusion Matrix for Particle *a*

a	Poss	Art	Voc	Inf	Rel	Num	Deg
Poss		1	2	**10**	**8**		**9**
Art			1				
Voc	3	1					
Inf	**17**	4			2		1
Rel	**8**			**9**			1
Num					1		
Deg	4	1					

Error Analysis

Infinitival Particle+VN tagged as Possessive Determiner +N (17)

In this category we have infinitival phrases tagged as determiner phrases, i.e. 'to VN' (infinitive) tagged as 'his/her/their/its N', e.g. *a scríobh* is tagged as 'his/her/their/its writing' rather than 'to write'.

On closer inspection the majority of problems related to items not found in the lexicon. Either the verbal noun token only featured in the lexicon as a common noun or did not exist in the lexicon at all and was wrongly guessed as a common noun. Some examples of the types of sentences in question are given below. In example (81) the verbal noun *choimead* is a misspelling of *choimeád* and was guessed as a noun, while in (82) the verbal noun *chónascadh* is in the lexicon only as a common noun.

(81) **a choimead** le chéile
to keep with other
'to keep together'

(82) tuiscint ar shuimiú a fhorbairt trí thacair **a chónascadh.**
understanding of summing to develop through sets **to conjoin**
'to develop an understanding of summing through joining sets.'

Possessive Determiner+N tagged as Infinitival Particle +VN (10)

In this category, we have determiner phrases tagged as infinitival phrases, i.e. 'his/her/their/its N' tagged as 'to VN' (infinitive), e.g. *a scríobh* is tagged as 'to write' rather than 'his/her/their/its writing'. In the majority of cases, a verbal noun in the genitive case is used with a determiner, e.g. *tír a dhéanta* 'country of its making'.

(83) ... líon na mbreiseán; tír **a dhéanta**, ...
... amount the additives; country its making, ...
'... amount of additives; country of manufacture, ...'

RelPart+V tagged as an Infinitival Particle +VN (9)

Here we have a verbal phrase tagged as an infinitival particle and verbal noun i.e. 'that + V' tagged as 'to + VN' (infinitive), e.g. *a scríobh* 'that wrote' is tagged as 'to write'.

Of the 9 incorrect occurrences, 6 related to the lexeme *scríobh* 'write' which has no surface realisation of lenition or eclipsis (initial mutation), with the result that both the particle *a* and the following token are highly ambiguous. Most of the instances where problems occur

contain fronted clauses rather than the default VSO sentence structure. Two examples of this type of sentence are given below.

(84) Cuid de na daoine **a scríobh** iad siúd táid marbh anois
Some of the people **that wrote** them there they-are dead now
'Some of the people that wrote those are dead now.'

(85) an Oifig Eolais, **a chraoladh** ainmneacha phríosúnaigh chogaidh
the Office Info, **that broadcast** names prisoners war
the Information Office, that broadcast names of war prisoners

Possessive Determiner tagged as DegPart +Adj (9)

In this category of error, a possessive determiner phrase is tagged as a degree particle and adjective. This is caused by either lexical gaps where the token features as an adjective or verbal noun in the lexicon and not as a noun also, as in (86), where *másaí* 'thighs' only appears in the lexicon as a comparative form of the adjective *másach* 'big-thighed'. There can also be tokenizing problems, where a token such as *c(h)uid*[25] as in (87), is split into 3 tokens: c, (h) and uid.

(86) a corróga, **a másaí** ...
her hips, her thighs ...

(87) ... **a c(h)uid** foghlama agus a (h)eispéiris
... his/her part learning and his/her aspirations
... his/her learning and his/her aspirations

RelPart+V tagged as Possessive Determiner+N (8)

These errors were mostly due to typographical errors where a misspelled verb is analysed as a noun, or where the particular form occurs in the lexicon as a noun, e.g. *bhéas* 'manner, moral conduct' (88) and *fritheadh* 'finding' (89), but their verbal forms are not in the lexicon.

(88) gurb iad cúrsaí Thuaisceart Éireann is mó **a bhéas** sa nuacht
that them matters North Ireland most that **will be** in-the news
'that it is mostly N.Ireland matters that will be in the news'

(89) Fear óg ... an teagascóir **a fritheadh** di
Man young the tutor **that was-procured** for-her
'A young man ...was the tutor that was procured for her'

[25] *a c(h)uid* 'her/his part' can be used as as an abbreviation for *a cuid* 'her part' or *a chuid* 'his part'

Possessive Determiner +N tagged as RelPart +V (8)

Here we have possessive determiner phrases being tagged as relative verbal phrases. The missing lexical items *cinn* 'head (gen)' (90) and *leath* 'half' (88) were guessed as verbs *cinn* 'decide' and *leath* 'spread/halve', respectively.

(90) ..agus is é a bhí bréan ag tochas **a chinn** de shíor
..and COP it that was bored at scratching **his head** for ever
'..and its he who was sick of always scratching his head'

(91) ..gur bréaga **a leath** dá bhfuil foghlama againn.
COP-REL lies **its half** that-which is learned by-us
'..that half of what we have learned is lies'.

Others: Dialectal variants, quotation marks

In the following two cases, dialectal variation caused a preposition and verbal noun to be tagged as a relative particle and verb. In (92) *a dh'* is a dialectal variant of the preposition *ag* 'at', while in (93) *leagadh* is a variant of the verbal noun *leagan* 'place'.

(92) ... d'imigh a thriúr mac **a dh'iascaireacht.**
... went his three son **at fishing**
'... his three sons went fishing'.

(93) ... béim **a leagadh** ar 10 a shuimiú le méaduithe de 10,
... emphasis **to place** on 10 to sum with multiples of 10,
'... to place emphasis on summing 10 with multiples of 10'

One issue which is difficult to resolve in CG in an elegant and comprehensive way, is the problem of a quotation mark token occurring between two other tokens, e.g. ...*a ' ainmníonn agus a náiríonn ' iad* '...that 'name and shame' them' as in (94).

(94) **a ' ainmníonn agus a náiríonn '** iad siúd a...
which ' name and which shame ' them those who...
'..which 'name and shame' those who...'

6.4.2 Numbers

Many of the forms used for numbers in Irish are homonymous. This leads to much POS ambiguity, as shown in Table 30. The context must be carefully examined in order to successfully disambiguate these forms.

Table 30 Homonymous Number Forms

token	Num Card.	Num Ord.	Noun	Det	Verbal Noun	Prep Pron.
aon	one		ace	any		
dó	two		burn		burning	to him
dá	two					to his/her/ their/it
dhá	two					to his/her/ their/it
trí	three					through
ceathair	four		quadriped			
sé	six		occasional			he
céad	hundred	first				
míle	thousand		mile			

There are several different forms corresponding to the number 'two', each of which have alternative meanings and part-of-speech categories. *Dá* 'two' is only used after the definite article *an* 'the' as in (95), whereas in all other cases *dá* is a prepositional pronoun form (96).

(95) an **dá** thicéad
the two ticket-SG
'the two tickets'

(96) le tabhairt **dá** n-athair
to giving to-their father
'to give to their father'

The singular form of nouns is usually used with numbers. This fact can be used in some instances to disambiguate between a homonymous form functioning as a number or a preposition. If a plural noun is used, as in (97) and (98) where we have *trí* 'three/through'. In (99) we can disambiguate *sé* 'six/he' as 'six' due to the initial mutation on the noun *duine* 'person', whereas the plural form in (100) allows us to disambiguate *sé* 'six/he' as 'he'.

(97) **trí** artaire
three/through artery-SG
'three arteries' OR 'through an artery'

(98) **trí** artairí
through artery-PL
'through arteries'

(99) meallann **sé** dhuine ...
charms six person-SG
'six people charm ...'

(100) meallann **sé** daoine
charms he person-PL
'he charms people'

Some of the more difficult homonymous number forms include *céad* meaning 'first' and 'hundred', and *míle* meaning 'thousand' and 'mile', (101)-(103).

(101) an **chéad** chéim eile
the first step other
'the next step'

(102) le linn **chéad** fiche bliain
during hundred twenty year
'for 120 years'

(103) Bhí sé **míle** míle ó bhaile
Was he thousand mile from home
'He was one thousand miles from home'

6.4.3 Other Challenging Ambiguities

Some phrases are difficult to disambiguate automatically using only the local context, though there is rarely a problem for the human interpreter who has access to the wider semantic context. In (104) *ní* can either be a copula or a noun meaning 'thing' (with an initial copula *is* elided).

(104) **Ní** beag é OR **Ní** beag é
COP-NEG small it Thing small it
'It is not small' 'A small thing'

In (105) and (106), there is ambiguity between the homonymous forms functioning as verbal nouns (with progressive aspect) or functioning as common nouns. In (105) *dlí* can be interpreted as 'law', a common noun preceded by the preposition *ag* 'at, by', alternatively in (105) *ag dlí* can be interpreted as 'deserving', a verbal noun preceded by an aspectual preposition *ag*. Similarly, in (106) *roinnt* can be interpreted as the quantifier 'some' or in (106) as a verbal noun 'dividing'.

(105) a) ... rud eile atá faoi rialú ag **dlí** poiblí ...
... thing other is under regulation at/by **law** public ...
'... another thing which is regulated by public law ...'

b) ... rud eile atá faoi rialú **ag dlí** poiblí ...
... thing other is under regulation **deserving** public ...
'...another thing which is regulated deserving public ...'

(106) a) **roinnt** diagairí
some theologians

b) **roinnt** diagairí
dividing theologians

6.5 Evaluation of POS Disambiguation Rate

We evaluate both the rate of POS disambiguation in a text, i.e. how much ambiguity remains and, in Section 6.6 we evaluate the quality of the disambiguation process.

On average 60% of tokens are ambiguous after morphological analysis. After applying CG rules (425 rules approx.), we find that approximately 98% of tokens are fully disambiguated in terms of POS or Lemma. However, when we include the additional morphological features in our evaluation, approximately 94% of tokens are fully disambiguated. In (107), *cailín* 'girl' is unambiguously a noun but its case feature remains ambiguous.

(107) "<cailín>" "cailín" Noun Masc Com Sg
"<cailín>" "cailín" Noun Masc Gen Sg

The results are summarised in Table 31.

Table 31 Development Set: Rate of Disambiguation

Development Set	Lemma		POS		POS+Features	
	count	%	count	%	count	%
Disambiguated	49,404	98.5%	48,994	97.7%	47,383	94.5%
Remains Ambiguous	747	1.5%	1,157	2.3%	2,768	5.5%
	50,151		50,151		50,151	

There are several different types of ambiguity associated with the morphological tags (lemma, POS, morphosyntactic features)

- Lemma ambiguity
- POS ambiguity
- Morphosyntactic feature ambiguity
    - Inflection ambiguity
    - Initial mutation ambiguity

In 1.5% of cases, there is lemma ambiguity, as in example (108). In this case a preposition and possessive determiner have been conflated and the original form of the preposition cannot be discerned from the token *dá*.

```
(108) "<dá>"
            "do_a" Prep Poss        ! to its/her/his
            "de_a" Prep Poss        ! of its/her/his
```

In 2.3% of cases there is part-of-speech (POS) ambiguity such as shown in (109).

```
(109) "<sin>"
            "sin" Det Dem           ! that (demonstrative)
            "sin" Pron Dem          ! that (pronoun)
```

Excluding the major POS category, there are two types of ambiguity which can occur in the morphosyntactic feature tags: a) inflection ambiguity, and b) initial mutation ambiguity. In (110) we have an example of case inflection ambiguity where the token could be either common or vocative case. In (111) we have an initial mutation ambiguity where a word beginning with a consonant which does not display overt initial mutation marking, e.g. the consonant *l* in this case could represent the eclipsed or lenited form (as opposed to (112) where the consonant *p* can be overtly eclipsed and lenited). These initial mutations are required for local morphosyntactic agreement and are important for disambiguation of adjacent tokens.

```
(110) "<bháisteach>"                                    ! rain
                "báisteach" Noun Fem Com Sg Len
                "báisteach" Noun Fem Voc Sg Len

(111) "<leanaí>"                                        ! children
            "leanbh" Noun Masc Gen Strong Pl
            "leanbh" Noun Masc Gen Strong Pl Ecl
            "leanbh" Noun Masc Gen Strong Pl Len
```

```
(112) "<páistí>"
            "páiste" Noun Masc Com Pl           ! children
      "<bpáistí>"
            "páiste" Noun Masc Gen Strong Pl Ecl
      "<pháistí>"
            "páiste" Noun Masc Com Pl Len
```

There are of course some cases where POS, lemma and feature tags all contain ambiguity, as in (113) where *dúnta* 'closed' could be either an adjective or a verbal noun in the genitive case.

```
(113) "<dúnta>"
            "dúnta" Adj Base            ! closed, secured, reticent
            "dúnadh" Verbal Noun Gen    ! (of/for) closing
```

6.6 Evaluation of POS Tagging

In this section, we evaluate the quality of the disambiguation process using precision, recall, and f-score measures. The evaluation is based on the Short Parole Tags (see Appendix A). as shown in Table 32. These truncated tags do not include the detailed morphological features used during the disambiguation process.

Development Set: Overall Results for POS Tagging

As not all corpus tokens are fully disambiguated, we generate more POS tags than there are tokens in the Gold Standard Corpus. Currently, the tagger achieves an overall POS precision of 93.85%, recall of 96.19%, and an f-score of 95.01% on the Development Set.

Overall Precision (Dev. Set): $\frac{CorrectAutoTags}{TotalAutoTags} \times \frac{100}{1} = \frac{47{,}297}{50{,}399} \times \frac{100}{1} = 93.85\%$

Overall Recall (Dev. Set): $\frac{CorrectAutoTags}{TotalGoldTags} \times \frac{100}{1} = \frac{47{,}297}{49{,}168} \times \frac{100}{1} = 96.19\%$

Overall F-score (Dev. Set) : $\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{96.19 \times 93.85 \times 2}{96.19 + 93.85} = 95.01\%$

Test Set: Overall Results for POS Tagging

We carry out the same calculations on the Test Set. The overall precision is 93.21%, recall is 95.5%, and f-score is 94.35% as presented below:

Overall Precision (Test Set): $\frac{CorrectAutoTags}{TotalAutoTags} \times \frac{100}{1} = \frac{23{,}321}{25{,}020} \times \frac{100}{1} = 93.21\%$

Overall Recall (Test Set): $\frac{CorrectAutoTags}{TotalGoldTags} \times \frac{100}{1} = \frac{23{,}321}{24{,}415} \times \frac{100}{1} = 95.52\%$

Overall F-score (Test Set) : $\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{95.52 \times 93.21 \times 2}{95.52 + 93.21} = 94.35\%$

Development Set: Detailed Results for POS Tagging

We also carry out a precision, recall and f-score analysis of each individual POS category, in order to see how they perform relative to one another. For example, we give the precision, recall and f-score calculations for common nouns below.

Precision (N common): $\frac{CorrectAutoNouns}{TotalAutoNouns} \times \frac{100}{1} = \frac{10{,}894}{11{,}445} \times \frac{100}{1} = 95.19\%$

Recall (N common): $\frac{CorrectAutoNouns}{GoldNouns} \times \frac{100}{1} = \frac{10{,}894}{11{,}510} \times \frac{100}{1} = 94.65\%$

F-score (N common) : $\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{94.65 \times 95.19 \times 2}{94.65 + 95.19} = 94.92\%$

Table 32 shows all POS categories in descending order of frequency in the Development Set data, i.e. Noun (common) is the first entry, as 23.41% of tokens are common nouns, whereas the last entry, Interjections, only account for 0.03% of tokens.

Table 32 Development Set: Detailed POS Tagging Results

POS	Tokens %	Gold	Auto	Correct	Precis.	Recall	Fscore	Weight. Fscore
N (com)	23.41	11510	11445	10894	95.19	94.65	94.92	22.22
Prep	14.55	7156	7118	6956	97.72	97.21	97.46	14.18
Punct.	10.37	5101	5101	5100	99.98	99.98	99.98	10.37
Verb	7.76	3813	3926	3727	94.93	97.74	96.32	7.47
Pronoun	6.32	3105	3192	3053	95.65	98.33	96.97	6.13
Art.	6.06	2978	2948	2945	99.90	98.89	99.39	6.02
Conj	5.61	2758	3011	2720	90.34	98.62	94.30	5.29
Adj.	4.43	2179	2293	2061	89.88	94.58	92.17	4.08
N (proper)	3.75	1843	1962	1745	88.94	94.68	91.72	3.44
N (verbal)	3.51	1726	1878	1572	83.71	91.08	87.24	3.06
Verb Prt.	3.26	1604	1610	1542	95.78	96.13	95.96	3.13
Det.	2.83	1393	1463	1315	89.88	94.40	92.09	2.61
Adverb	1.76	863	929	827	89.02	95.83	92.30	1.62
Copula	1.59	784	920	666	72.39	84.95	78.17	1.24
Particle	1.50	736	844	715	84.72	97.15	90.51	1.36
Numeral	1.44	706	773	672	86.93	95.18	90.87	1.31
Adj. (verbal)	0.72	355	493	343	69.57	96.62	80.90	0.58
N (subst.)	0.60	293	279	257	92.11	87.71	89.86	0.54
Abbrev.	0.28	140	142	130	91.55	92.86	92.20	0.26
Foreign	0.23	111	69	51	73.91	45.95	56.67	0.13
Interject.	0.03	14	9	6	66.67	42.86	52.17	0.02
Sub-total		**49,168**		**47,297**	**96.19**			**95.06**
Para Tags		983		983				
Totals		50,151		48,280				

The results in Table 32 show that most f-scores are in the 90's. Three POS categories have f-scores in the 80's, i.e. Noun (substantive) 89.86%, Noun (verbal) 87.24, Adjective (verbal) 80.90%. These are categories which contain many homonymous forms, as does the Copula whose f-score is 78.17%. The two lowest f-scores are Foreign 56.67% and Interjection 52.17%. However, these two categories combined only account for 0.26% of tokens and so have a negligible effect on overall performance, as shown by their weighted f-scores (WFScore). Note that difference between the sum of the weighted f-scores (95.06%) and the overall f-score (Dev. Set) previously calculated (95.01%), is due to rounding errors.

6.7 Summary

In this chapter, we have explained the principles behind CG and how it is applied in practice. We have described its application to Irish POS tagging, based on disambiguationg morphologically analysed text. We have highlighted some of the more challenging tasks in this approach.

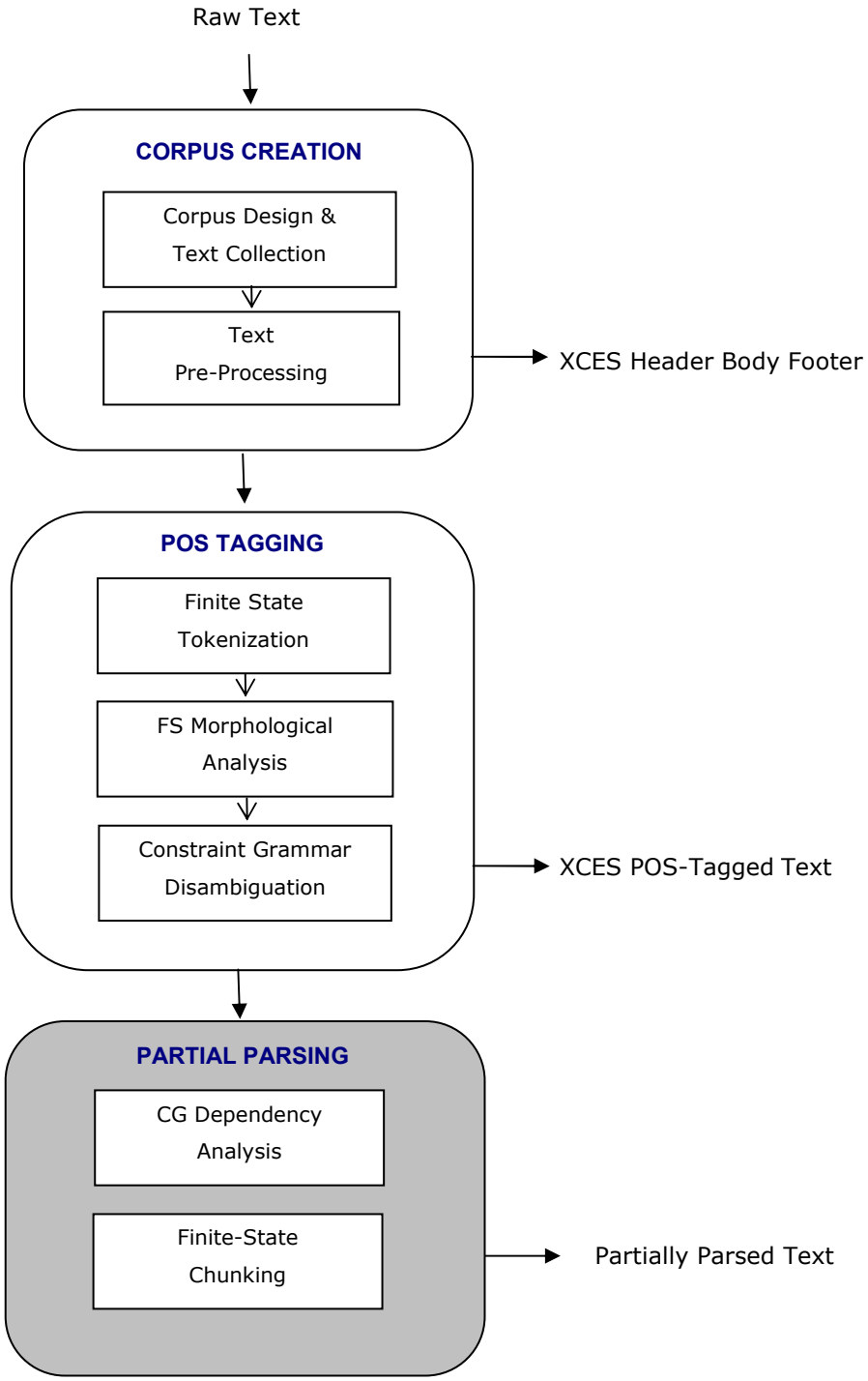
We find that 98% approx. of tokens are fully disambiguated as regards POS category and 94.5% of tokens are fully disambiguated with regard to morphosyntactic features.

In evaluating POS category only, the disambiguation process achieves a precision of 95.19%, recall of 94.65%, and f-score of 94.92% on the Development Set data. The corresponding figures for the Test Set are a precision of 93.21%, recall of 95.52%, and f-score of 94.35%.

These figures leave room for improvement as POS taggers, in general, currently achieve precision rates of 95-99%. In order to improve on our figures we hope to supplement the existing CG rules with rules automatically induced from the tagged Development Set of the Gold Standard Corpus, e.g. in a manner similar to that of Samuelsson *et al.* (1996).

In Part III of this thesis, we will look at partial parsing of the POS tagged text.

Part III Partial Parsing of Irish



7 Dependency Analysis of Irish

7.1 Introduction

As stated in Section 2.4, we have chosen to implement partial parsing of Irish in two stages. Firstly, we apply Dependency Analysis annotation tags to each word token (Chapter 7), and, secondly, we bracket the annotated sentence into phrase-like units described by Abney (1991) as 'chunks' (Chapter 8).

In automatically parsing a language for the first time, deciding what constitutes a syntactic unit, and how it should be annotated, accounts for a major part of the work. In order to do this, we constructed a dedicated Test Suite of 225 made-up, short, grammatical sample sentences, covering the main syntactic phenomena of Irish ((Biber et al., 2003); (Ó hUallacháin and Ó Murchú, 1981); (Doherty, 1996); (Stenson, 1981)). A selection of sentences from the Test Suite are used throughout this chapter.

We provide examples of the various syntactic structures in Irish, and present illustrative templates for the dependency analysis of each type of sentence. Using CG rules we automatically annotate the tokens with grammatical relation or unlabelled dependency tags.

The Test Suite sentences were automatically tagged using these rules, and manually corrected. The automatic tagging process is iteratively developed and tested using the made-up sentences of the Gold Standard Test Suite, and later using attested sentences in a Gold Standard Corpus of 250 sentences which were randomly selected from the NCII-based Gold Standard (3000) Corpus POS Tagged Corpus (see Table 11 p63 for details). All of the Test Suite sentences, together with their analyses are given in Appendix E.

The dependency analysis is shallow and partial, as it does not cover co-ordination, long-distance dependencies and prepositional and clausal attachments are not resolved. The result is a single deterministic analysis.

In Section 7.2, we describe grammatical function and dependency relation annotation as applied to Irish. In Section 7.3, we describe the annotation scheme we have developed. In Section 7.4, we present a set of abstract templates which illustrate the main syntactic patterns in Irish. In Section 7.5, we present the implementation of Dependency Analysis for Irish using Constraint Grammar and, finally, in Section 7.6, we present the results of our evaluation.

7.2 Grammatical Functions and Dependency Relations for Irish

In addition to using Constraint Grammar (CG) to disambiguate morphological analyses for POS tagging, we also use CG to produce a dependency based analysis of POS disambiguated sentences (Karlsson, 1995, p33), by assigning surface syntactic labels to each token.

There are a number of differences between CG and other parsing methodologies (Karlsson, 1995, p37). Unlike a context-free grammar, a Constraint Grammar does not attempt to define the set of grammatical sentences in a language. The CG philosophy is that everything is licensed which is not explicitly ruled out. This makes it more robust in handling unrestricted text. Also, it does not aim to produce a minimal set of general rules – a CG grammar can contain many lexically specific rules to handle special cases. Neither does it attempt to determine constituency structure.

In our Dependency Analysis of Irish, all tokens receive either a grammatical function tag, or an unlabelled dependency tag, i.e. they are identified as being either a head or a modifier of a head. Sentences are first divided into clauses. Within a clause, the verb (or non-verbal copula) and its arguments are annotated with grammatical labels such as verb, subject and object, or copula, subject and predicate. We, also, annotate various types of prepositional phrase. Subjects and objects correspond to NPs, while indirect objects correspond to PPs. The head of a PP is a preposition, with its direct dependent being the head of a noun phrase, which in turn may have dependent modifiers such as adjectives or determiners. When there is a possessive relation between NPs we consider the possessor noun to be a modifier of the possessed noun. Dependent modifiers can come before or after head, therefore the tag specifies the direction of the head they modify, e.g. `@>N` marks a noun premodifier, while `@N<` marks a noun postmodifier. In co-ordinated structures, we tag the second conjoint as being dependent on the conjunction as we are not in a position to determine the exact nature and extent of the co-ordinated elements.

In our Dependency Analysis, the grammatical function 'subject' is a surface syntactic subject. To identify the subject we use all available information, including 1) lexical cues, e.g. synthetic verbs, or special 3rd person subject pronouns, 2) morphological cues, e.g. transitivity information on verbs, 3) syntactic cues, e.g. word order and clause structure. In most cases the surface notion of subject equates to the traditional notion of a subject, i.e. in verbal constructions it is often the doer of the action, and in copular constructions we annotate a subject and predicate.

As a sentence with multiple clauses can have more than one subject, we use a number of different subject tags for processing purposes, in order to identify the appropriate head, i.e.

@SUBJ_INF, @SUBJ_ASP and @SUBJ_REL for infinitival, aspectual and relative clauses respectively (similarly for object labels).

In contrast to full parsers such as FDG (Tapanainen and Järvinen, 1997) or MaltParser (Nivre and Hall, 2005), in our analysis we do not explicitly mark the local head associated with a dependent. In full parsers, this information is encoded in terms of numerical (often positional) indices. In our analysis, although not explicitly represented, this information is largely recoverable from the tagset and marking of clause boundaries. For example, the '<' annotation in a dependency tag specifies that the local head is the first appropriate head located to the left; and @SUBJ_REL indicates that this token is the subject of the relative verb.

No abstract levels are inserted during the dependency analysis, i.e. no traces (e.g. to capture long-distance dependencies), elided items (ungrammatical structures, e.g. 'Seems we have a problem here') or ellipted items (grammatical structures, e.g. 'John can swim but Pat can't'). Only tokens present in the surface structure are tagged. Constituents are not explicitly marked, although in most cases there is a strong parallel between a head plus dependants and a constituent.

7.3 Annotation Scheme

A full list of the tags used in Irish dependency relation annotations, arranged in alphabetical order, is given in Table 33. While this tagset follows the style of tags described by Karlsson (1995) for English, as well as the tagsets used for Danish, Portuguese and other languages which are described on the VISL website,[26] there is not a prescribed list of tags. This flexibility allows one to tailor the tagset to the language under consideration. By convention, the dependency tags all start with the `@` symbol to distinguish them from morphological tags which have already been appended to the tokens (see Chapter 6). (114) shows the verb *inis* 'tell', to which the grammatical function tag `@FMV` has been appended, denoting that it is functioning as a finite main verb.

```
(114) "inis" Verb VTI PastInd Len @FMV
```

In (115) we illustrate the grammatical function and dependency labels which are appended to the POS-tagged tokens in a simple declarative sentence. In this sentence, the main verb, subject and object are tagged with `@FMV`, `@SUBJ` and `@OBJ`, respectively. We also have a pre-verbal particle tagged as `@>V`, a pre-modifying article tagged as `@>N`, and the final noun

[26] VISL website: http://visl.sdu.dk (last accessed 10 May 2008).

which is the object of the preceding preposition is tagged as @P<, denoting that it is dependent on the previous preposition. The prepositional phrase is tagged as @PP_OBL to indicate that it contains an oblique object of the verb.

(115) D'inis sí an scéal do Mháire
Told she the story to Mary
'She told the story to Mary'

```
"<D'>"      "do" Part Vb @>V                          Part.
"<inis>"    "inis" Verb VTI PastInd Len @FMV          told
"<sí>"      "sí" Pron Pers 3P Sg Fem Sbj @SUBJ        she
"<an>"      "an" Art Sg Def @N>                       the
"<scéal>"   "scéal" Noun Masc Com Sg DefArt @OBJ      story
"<do>"      "do" Prep Simp @PP_OBL                    to
"<Mháire>"  " Máire" Prop Noun Fem Com Sg Len @P<     Máire
"<.>"       "." Punct Fin                             .
```

All dependency labels are conditioned on the context within the sentence. For example, a noun premodifier will only be marked as @N> if it actually precedes a noun. A token such as *d'* will be tagged as @N> if it preceded a noun and @V> if it precededs a verb. In our implementation, dependent modifiers, i.e. those with directional labels, always refer to the first available head to the left or right as appropriate.

Table 33 Grammatical Function and Head/Modifier Dependency Labels

TAG	DESCRIPTION	EXAMPLE
@>ADJ	adverbial particle dependent on the adjective to the right	*go ciúin* 'quietly'
@>N	pre-modifier dependent on the first noun to the right	*an* 'the'
@>V	pre-verbal particle dependent on a verb to the right	*ní* 'not'
@ADVL	adverbial	*anocht* 'tonight'
@ADVL<	adverbial post modifier	
@AUG>SUBJ	augment pronoun dependent on subj. to the right	*Is é Seán* …, It/He, Seán is…
@CC	co-ordinating conjunction	*agus* 'and'
@CLB	clause boundary	e.g. *agus* 'and' when followed by a verb, and subordinating conjs.etc.
@COP	copula	
@COP_WH	interrogative copula	*cé leis an leabhar* 'whose is the book'
@COP_SUBJ	copula including subject	*Seo an fear...*'This is the man...'
@FAUX	finite auxiliary verb	*Tá sé ag cócaireacht* 'He is cooking'
@FAUX_REL	relative finite auxiliary verb	*atá siad* 'that/which they are'
@FAUX_REL_SUBJ	relative finite auxiliary verb including subject	*atáimid* 'that/which we are'
@FAUX_SUBJ	finite auxiliary verb including subject	*táimid* 'we are'
@FMV	finite main verb	*rith* 'run'
@FMV_REL	relative finite main verb	*a chuala mé*, 'that I heard'
@FMV_REL_SUBJ	relative finite main verb incl. subject	*a chualamar*, 'that we heard'
@FMV_SUBJ	finite main verb including subject	*ritheamar* 'we ran'
@INF	bare infinitive	*Ba mhaith liom fanacht* 'I would like to stay'
@N<	noun post-modifier	*teach mór* 'big house'
@NP	unlabelled noun head, e.g. list item, apposition, or fragment	*1) dathuithe, 2) leasaithigh*, '1) colours, 2) additives'

@OBJ	object	*Chonaic Seán Máire*, 'Seán saw Máire'
@OBJ_ASP	object of aspectual	*ag déanamh oibre*, 'doing work'
@OBJ_INF	object of infinitive	*bainne a ól*, 'to drink milk'
@PP_SUBJ	prep + subj pronoun	*D'éirigh liom*, 'I succeeded' i.e. success was with me'
@P<	noun dependent on the preceding prep.	*ag an doras* 'at the door'
@PC<	noun dependent on *compound preposition* is in genitive case	*tar éis na Nollag*, after Christmas
@PN<	pronoun post-mod.	*é féin* 'himself'
@PP_ADVL	adverbial PP head	*ag an doras* 'at the door'
@PP_ASP	aspectual PP head	*ag rith* '(at) running'
@PP_HAS	PP meaning has	*ag Seán*, 'Seán has' i.e. at Seán
@PP_NEG	negative preposition	*gan dul* 'without going'
@PP_OBL	oblique PP head	*do Mháire* 'to Máire'
@PP_PRED	predicative	*Is liom é* 'It is mine' i.e. Is with me it
@PP_STAT	stative	*ina rí* 'is a king' i.e. 'in his king(hood)'
@PP_SUBJ	prep with a subject	*D'éirigh liom* 'I succeeded', i.e. success (was) with me
@PRED	predicate	*Tá sé mór* 'It is big'
@PRED<	dependent on predicate	*Is deas an lá é* 'It is a nice day' i.e. Is nice the day it
@SUBJ	subject	*Chonaic Seán Máire*, 'Seán saw Máire'
@SUBJ_INF	subject of infinitive (intrans)	*an obair a bheith déanta* 'the work to be done'
@SUBJ_OR_OBJ	subject or obj. of relative clause	*a chonaic an bhean*, 'that the woman saw' OR 'that saw the woman'
@SUBJ_ASP	subject of aspectual phrase	bhí sé ag obair 'he was working'
@SUBJ_REL	subject of relative clause	*a rinne sé* 'that he made'

Before turning our attention to a selection of sample sentences and their dependency annotation templates, we will look in more detail at the labels which are used to tag verbs, nouns, prepositions, adverbs.

7.3.1 Verbs

We identify all verbs as finite main verbs, except *bí* 'to be' which we identify as a finite auxiliary (Ó hUallacháin and Ó Murchú, 1981, p146) when used in periphrastic aspectual constructions, and as a copula when used to describe states, emotions etc. All verbs have relative forms, and all verb-forms can have an incorporated[27] subject. Inflected verb forms which contain person and number, (as opposed to analytic verb forms where the subject is expressed as a separate noun or pronoun) are known as synthetic verb forms. It is ungrammatical to have a verb inflected for person/number and a separate subject noun or pronoun.

The following tags are used to tag verbs:

```
@FMV, @FMV_SUBJ, @FMV-REL, @FMV-REL_SUBJ

@FAUX, @FAUX_SUBJ @FAUX-REL, @FAUX-REL_SUBJ
```

We tag the finite main verb in a main clause as follows, e.g. *Labhair Seán* 'Seán spoke'

(116)

```
"<Labhair>" "labhair" Verb VTI PastInd Len @FMV            Spoke
"<Seán>"    "Seán" Prop Noun Masc Com Sg @SUBJ             Seán
```

Relative finite main verb forms are always introduced by a relative particle *a*[28], e.g. *a cheannaigh siad*, 'that/which they bought'.

(117)

```
"<a>"             "a" Part Vb Rel Direct @>V               that
"<cheannaigh>"    "ionsaigh" Verb VT PastInd Len @FMV_REL  bought
"<siad>"          "siad" Pron Pers 3P Pl Sbj @SUBJ         they
```

[27] We use the term 'incorporated subject' in the sense that there cannot be another subject in addition to the inflected verb form.

[28] The relative particle can be incorporated into another item, e.g. *len**a** cheannaigh siad* 'with-which they bought'

The following is an example of the verb *bí* 'to be', functioning as an auxiliary, with a non-finite aspectual complement, e.g. *Tá sé ag rith*, 'He is running'. Other auxiliaries include *caith* 'must' and *téigh* 'go'.

(118)

```
"<Tá>"      "bí" Verb VI PresInd @FAUX                    Is
"<sé>"      "sé" Pron Pers 3P Sg Masc Sbj @SUBJ           he
"<ag>"      "ag" Prep Simp @PP_ASP                        at
"<rith>"    "rith" Verbal Noun VTI @P<                    running
```

In the following example, we have a relative finite auxiliary verb which includes a subject, e.g. *a bhídís ag obair*, 'that they were working''

(119)

```
"<a>"       "a" Part Vb Rel Direct @>V                    that
"<bhídís>"  "bí" Verb VI PastImp 3P Pl Len @FAUX_REL_SUBJ  they-
                                                          were
"<ag>"      "ag" Prep Simp @PP_ASP                        at
"<obair>"   "obair" Verbal Noun NStem @P<                 working
```

7.3.2 Nouns

There are a number of ways in which the head noun of a noun phrase may be annotated, depending on its function in the sentence. We include extra information in the tags in order to identify the heads of the various subject and object roles described in the sections below. Dependent nouns are tagged as being dependent on a noun or preposition as appropriate using the `@N<` and `@P<` tags. Other nouns, appearing in lists or in apposition, or as unidentified conjuncts in a conjunctive phrase are simply tagged as `@NP`.

It is important to note the distinction in Irish between "verbal nouns" and all other types of noun, such as common or proper nouns. A verbal noun is a noun derived from a verb root (or agent noun) and it carries the same transitivity properties (being semantically related) as its associated verb. These verbal nouns appear in a range of aspectual roles, performing functions usually carried out by verb forms in many other languages. Although some literature on Irish syntax is ambivalent as to whether the verbal noun is verbal or nominal (Ó Siadhail, 1989; Stenson, 1981) and in other literature a verbal analysis is adopted (McCloskey, 1983),. In this implementation, because of the overt nominal qualities of verbal nouns, i.e. they appear with prepositions and their postposition objects are in the genitive

case, we treat them as nominal[29] structures dependent on a prepositional head (i.e. `@P<`). Their aspectual function is included in the grammatical function tag, e.g. `@PP_ASP`. Prepositional phrases will be discussed in more detail in Section 7.3.3. We also identify verbal nouns acting as infinitives with the `@INF` tag.

If we were to attempt to treat verbal nouns as verbal forms, we would have to duplicate each lexical form in the finite-state lexicon, as they can also function as pure nominals with a determiner, i.e. each form would have a verbal analysis as well as a nominal analysis. In addition to inherent inefficiency, it would also lead to enormous ambiguity in the verbal noun form, as well as in all of its dependants. Furthermore, in a verbal analysis, prepositions occurring with the verbal noun would also need to be duplicated in the lexicon as a type of pre-verbal particle, again leading to additional ambiguity on a large scale. We, therefore, feel that the most sensible course of action is to treat the verbal noun morphologically as noun (in POS tagging) and to identify its aspectual and infinitival functions in certain constructions at the syntactic level through dependency tags and/or labelled chunks. This is similar to our treatment of adverbials such as *go tobann* 'suddenly' where an adverbial particle *go* together with an adjective *tobann* 'sudden', which is still morphologically described as an adjective, functions as an adverbial.

The following tags are used to tag nouns:

`@SUBJ, @SUBJ_REL, @SUBJ_ASP, @SUBJ_INF`

`@OBJ, @OBJ_ASP, @OBJ_INF, @SUBJ_OR_OBJ`

`@N<, @P<, @NP, @INF`

Subject, Object

A noun or pronoun may be the subject or object of a simple declarative sentence, such as *Cheannaigh sé úll* 'He bought an apple'. For these nouns, we use the `@SUBJ` or `@OBJ` grammatical function tag as appropriate.

(120)

```
"<Cheannaigh>"    "ceannaigh" Verb VTI PastInd Len @FMV     Bought
"<sé>"            "sé" Pron Pers 3P Sg Masc Sbj @SUBJ       he
"<úll>"           "úll" Noun Masc Com Sg @OBJ               an-apple
```

[29] They are, however, modified by adverbs.

Subject of Relative Clause

A noun or pronoun in the main clause may be the subject of a relative clause as in *Chonaic Máire an fear a bhí ag iascaireacht*, 'Máire saw the man who was fishing'. We use the `@SUBJ_REL` tag to distinguish this usage from NPs which are simply the subject of the main verb.

(121)

```
"<Chonaic>"       "feic" Verb VTI PastInd Len @FMV          Saw
"<Máire>"         "Máire" Prop Noun Fem Com Sg @SUBJ        Máire
"<an>"            "an" Art Sg Def @>N                       the
"<fear>"          "fear" Noun Masc Com Sg DefArt @SUBJ_REL  man
"<a>"             "a" Part Vb Rel Direct @>V                that
"<bhí>"           "bí" Verb VI PastInd Len @FAUX_REL        was
"<ag>"            "ag" Prep Simp @PP_ASP                    at
"<iascaireacht>"  "iascaireacht" Verbal Noun NStem @P<      fishing
```

Subject of an Aspectual Complement

A noun or pronoun in the main clause may be the subject of a non-finite complement, e.g. a progressive, as in *Chonaic mé Séan ag oscailt an dorais*, 'I saw Seán opening the door'. We use the `@SUBJ_ASP` tag to indicate this role.

(122)

```
"<Chonaic>" "feic" Verb VTI PastInd Len @FMV              Saw
"<mé>"      "mé" Pron Pers 1P Sg @SUBJ                    I
"<Seán>"    "Seán" Prop Noun Masc Com Sg @SUBJ_ASP        Seán
"<ag>"      "ag" Prep Simp @PP_ASP                        at
"<oscailt>" "oscailt" Verbal Noun VTI @P<                 opening
"<an>"      "an" Art Sg Def @>N                           the
"<dorais>"  "doras" Noun Masc Gen Sg @OBJ_ASP             door
```

Subject of an Infinitive

A noun or pronoun in the main clause may be the subject of an infinitive, i.e. *Ní mór dúinn aonad a bheith againn*, 'It is necessary for us to have a unit'. We use the `@SUBJ_INF` tag to indicate this grammatical role.

(123)

```
"<Ní>"      "is" Cop Pres Neg @COP                          Not
"<mór>"     "mór" Adj Base @PRED                            big
"<dúinn>"   "do" Pron Prep 1P Pl @PP_ADVL                   to-us
"<aonad>"   "aonad" Noun Masc Com Sg @SUBJ_INF              unit
"<a>"       "a" Part Inf @>N                                to
"<bheith>"  "bheith" Verbal Noun VI Len @INF                be
"<againn>"  "ag" Pron Prep 1P Pl @PP_ADVL                   at-us
```

Ambiguous Subject/Object of Direct Relative Clause

In certain relative clauses, it is inherently ambiguous as to whether the noun is the subject or the object of the relative verb, e.g. *Seo an fear a chonaic an bhean* could mean either 'This is the man the woman saw' or 'This is the man that saw the woman. In these cases we use the @SUBJ_OR_OBJ tag. In doing so we avoid assigning two tags, and we explicitly identify the token as being syntactically ambiguous.

(124)

```
"<Seo>"     "seo" Cop Pro Dem @COP_SUBJ                     This
"<an>"      "an" Art Sg Def @>N                             the
"<fear>"    "fear" Noun Masc Com Sg DefArt @PRED            man
"<a>"       "a" Part Vb Rel Direct @>V                      that
"<chonaic>" "feic" Verb VTI PastInd Len @FMV_REL            saw
"<an>"      "an" Art Sg Def @>N                             the
"<bhean>"   "bean" Noun Fem Com Sg DefArt @SUBJ_OR_OBJ      woman
```

In cases where there is no ambiguity, the appropriate tag is used, i.e. where a) the relative verb incorporates the subject, or b) the verb is intransitive and, therefore, has no object or c) the relative verb is preceded by an adverbial or prepositional phrase.

Object of an Aspectual Verbal Noun

We identify the noun or pronoun which is functioning as the object of an aspectual verbal noun, as in (125), *Tá mé ag déanamh cáca*, 'I am making a cake', with the @OBJ_ASP tag.

(125)

```
"<Tá>"       "bí" Verb VI PresInd @FAUX                    Is
"<mé>"       "mé" Pron Pers 1P Sg @SUBJ_ASP                I
"<ag>"       "ag" Prep Simp @PP_ASP                        at
"<déanamh>" "déanamh" Verbal Noun VTI @P<                  making
"<cáca>"     "cáca" Noun Masc Gen Sg @OBJ_ASP              a cake
```

Infinitival Phrases

The verbal noun when functioning as an infinitive is tagged `@INF`. In the case of a transitive verbal noun, it is preceded by an infinitive marker, *a* (or *do*), e.g. *bainne a fháil,* 'to get milk'.

(126)

```
"<Chuaigh>" "téigh" Verb VTI PastInd Len @FAUX        Went
"<sé>"      "sé" Pron Pers 3P Sg Masc Sbj @SUBJ       he
"<amach>"   "amach" Adv Dir @ADVL                     out
"<chun>"    "chun" Prep Simp @PP_ASP                  to
"<bainne>"  "bainne" Noun Masc Com Sg @OBJ_INF        milk
"<a>"       "a" Part Inf @>N                          to
"<fháil>"   "fáil" Verbal Noun VT Len @INF            get
```

Object of an Infinitive

We identify the noun or pronoun immediately preceding a transitive infinitive as being the object of the infinitive, as in *Chuaigh sé amach chun bainne a fháil*, 'He went out to get milk'. The tag used is `@OBJ_INF`.

(127)

```
"<bainne>"  "bainne" Noun Masc Com Sg @OBJ_INF        milk
"<a>"       "a" Part Inf @>N                          to
"<fháil>"   "fáil" Verbal Noun VT Len @INF            get
```

Possessive Noun Phrases

Where an NP is in genitival relation to another NP, i.e. a noun (e.g. *ceantar* 'region') modifying a head noun (e.g. *teorainn* 'border') will be in the genitive case, we tag the modifying noun as being dependent on the head noun using the `@N<` tag, e.g. *ag teorainn an cheantair* 'at the border of the region'

(128)

```
"<ag>"            "ag" Prep Simp @PP_ADVL                  at
"<teorainn>"      "teorainn" Noun Fem Com Sg @P<           border
"<an>"            "an" Art Sg Def @>N                      the
"<cheantair>"     "ceantar" Noun Masc Gen Sg DefArt @N<    region
```

7.3.3 Prepositional Phrases

Several types of prepositional phrase are identified according to the function they perform. These include oblique (or indirect) objects, adverbial adjuncts and the important class of aspectual complements. In each case, the head of the noun complement is tagged as being dependent on the preposition which precedes it, using the @P< tag.

The following tags are used to tag prepositional heads:

@PP_ADVL, @PP_OBL, @PP_NEG

@PP_ASP, @PP_STAT, @PP_HAS, @PP_PRED, PP_SUBJ

Adverbial Phrases

A preposition heading an locative adverbial phrase, e.g. *ins an siopa* 'in the shop', is tagged using the @PP_ADVL tag. This could alternatively be tagged as @PP_LOC but we have not as yet distinguished between differerent sub-types of adverbial, e.g. locative, manner, temporal etc.

(129)

```
"<ins>"     "i" Prep Art Sg @PP_ADVL                 in
"<an>"      "an" Art Sg Def @>N                      the
"<siopa>"   "siopa" Noun Masc Com Sg DefArt @P<      shop
```

Prepositional pronouns (also known as conjugated prepositions), which are a combination of preposition and pronoun, are always tagged as @PP_ADVL in this implementation, as in *ag plé léi*, 'discussing with her'.

(130)

```
"<ag>"      "ag" Prep Simp @PP_ASP                          at
"<plé>"     "plé" Verbal Noun VTI @P<                     discussing
"<léi>"     "le" Pron Prep 3P Sg Fem @PP_ADVL               with-her
```

Oblique/Indirect Object Phrases

The `@PP_OBL` tag is used on prepositions indicating indirect objects of ditransitive verbs, e.g. *do Mháire* 'to Mary' when used with a verb such as *tabhair* 'give'.

(131)

```
"<do>"      "do" Prep Simp @PP_OBL                          to
"<Mháire>"  "Máire" Prop Noun Fem Com Sg Len @P<            Máire
```

Aspectual Phrases

The preposition *ag* 'at', preceding a verbal noun, functions as a progressive aspectual marker; in (132) as a progressive e.g. *ag iascaireacht* '(at) fishing'.

(132)

```
"<ag>"            "ag" Prep Simp @PP_ASP                    at
"<iascaireacht>"  "iascaireacht" Verbal Noun NStem @P<      fishing
```

Stative Aspectual Phrases

While the preposition *i* 'in' with a possessive determiner *a*, i.e. *ina* 'in his' a, can of course be used locatively, it is also used with the copular verb *bí* 'to be' to denote a state, e.g. *Tá sé ina chodladh* 'He is asleep', i.e. in his sleep. These prepositional phrases, involving a verbal noun, are indicated using the `@PP_STAT` tag.

(133)

```
"<Tá>"            "bí" Verb VI PresInd @FAUX                Is
"<sé>"            "sé" Pron Pers 3P Sg Masc Sbj @SUBJ       he
"<ina>"           "i" Prep Poss 3P Pl @PP_STAT              in-his
"<chodladh>"      "codladh" Verbal Noun VI Len @P<          sleep
```

The preposition *ar* 'on' with a verbal noun is used to denote a progressive state, e.g. *Tá sé ar snámh* 'It is floating'.

(134)

```
"<Tá>"          "bí" Verb VI PresInd Len @FAUX          Is
"<sé>"          "sé" Pron Pers 3P Sg Masc Sbj @SUBJ     it/he
"<ar>"          "ar" Prep Simp @PP_STAT                 on
"<snámh>"       "snámh" Verbal Noun VTI @P<             floating
```

Currently, we only implement this distinction for verbal nouns as indicated by the POS tag in (134), although the same construction is used with common nouns, and ideally *Tá sé ina mhúinteoir* 'He is a teacher' should be tagged as stative, as shown in (135):

(135)

```
"<Tá>"          "bí" Verb VI PresInd @FMV               Is
"<sé>"          "sé" Pron Pers 3P Sg Masc Sbj @SUBJ     he
"<ina>"         "i" Prep Poss 3P Sg Masc @PP_STAT       in-his
"<mhúinteoir>"  "múinteoir" Noun Masc Com Sg Len @P<    teacher
```

However, additional noun subcategorisation information (e.g. professions) would be required to distinguish between stative and locative constructions involving common nouns, such as the locative predicate *Tá sé ina theach* 'He is in his house', shown below, as opposed to the stative aspect of the preceding example. We currently tag these common nouns as the object of predicative prepositional phrases with the verb *bí* 'to be' used as a copular verb.

(136)

```
"<Tá>"          "bí" Verb VI PresInd @FMV               Is
"<sé>"          "sé" Pron Pers 3P Sg Masc Sbj @SUBJ     he
"<ina>"         "i" Prep Poss 3P Sg Masc @PP_PRED       in-his
"<theach>"      "teach" Noun Masc Com Sg Len @P<        house
```

This analysis can also be used for such constructions as *Tá sé ar buille* 'angry'/*ar meisce* 'drunk'/*thar cinn* 'excellent'/*thar fóir* 'excessive' which all involve prepositional phrases.

Negative Marker

A negative marker on a noun phrase, e.g. *gan airgead* 'without money', is tagged with the `@PP_NEG` tag.

(137)

```
"<gan>"      "gan" Prep Simp @PP_NEG                    without
"<airgead>" "airgead" Noun Masc Com Sg @P<              money
```

It is used in the same manner with verbal nouns as in *gan stad* 'without stopping' in (138).

(138)

```
"<gan>"      "gan" Prep Simp @PP_NEG                    without
"<stad>"     "stad" Verbal Noun VTI @P<                 stopping
```

'Gan' can also be used as a negative marker on non-finite clauses as in *gan an bainne a fháil* 'without getting the milk' (139).

(139)

```
"<gan>"      "gan" Prep Simp @PP_NEG                    without
"<an>"       "an" Art Sg Def @>N                        the
"<bainne>"   "bainne" Noun Masc Com Sg @OBJ_INF         milk
"<a>"        "a" Part Inf @>N                           to
"<fháil>"    "fáil" Verbal Noun VT Len @INF             get
```

Predicative Prepositional Phrases

The preposition *le* 'with' is used (in conjunction with a noun, or as a conjugated preposition) as a predicate in copular constructions such as *Is le Dónal an teach mór* 'Dónal owns the big house' to denote ownership (further described in Section 7.4.7)

(140)

```
"<Is>"       "is" Cop Pres @COP                         Is
"<le>"       "le" Prep Simp @PP_PRED                    with
"<Dónal>"    "Dónal" Prop Noun Masc Com Sg @P<          Dónal
```

```
"<an>"      "an" Art Sg Def @>N                          the
"<teach>"   "teach" Noun Masc Com Sg DefArt @SUBJ        house
"<mór>"     "mór" Adj Masc Com Sg @N<                    big
```

Prepositional Phrases denoting Possession

The preposition *ag* 'at' preceding a noun (other than a verbal noun) together with the substantive verb *bí* 'to be' equates to the verb 'has' in English, e.g. *Bhí an t-airgead ag Séan* 'Seán had the money' lit. 'The money was at Seán' (see also Section 7.4.5).

(141)

```
"<Bhí>"          "bí" Verb VI PastInd Len @FMV            Was
"<an>"           "an" Art Sg Def @>N                      the
"<t-airgead>"    "airgead" Noun Masc Com Sg DefArt @SUBJ  money
"<ag>"           "ag" Prep Simp @PP_HAS                   at
"<Seán>"         "Seán" Prop Noun Masc Com Sg @P<         Seán
```

However, the above structure is indistinguishable from the locative structure in (142). To avoid incorrectly tagging prepositional phrases with the `@PP_HAS` tag, we only apply this tag where the dependent noun is a proper noun or a pronoun. The disadvantage of this decision is that we do not properly account for a minority of cases involving common noun possessors in these type of structures. To resolve this difficulty we would need subcategorisation information denoting nouns as animate and human.

(142)

```
"<Bhí>"          "bí" Verb VI PastInd Len @FMV            Was
"<Seán>"         "Seán" Prop Noun Masc Com Sg @SUBJ       Seán
"<ag>"           "ag" Prep Simp @PP_ADVL                  at
"<an>"           "an" Art Sg Def @>N                      the
"<doras>"        "doras" Noun Masc Com Sg DefArt @P<      door
```

7.3.4 Adverbial Phrases

Adverbs and their modifiers are tagged with the following tags:

`@ADVL, @>ADJ`

Bare adverbials

Apart from prepositional adverbial phrases already mentioned, an adverbial phrase may consist of a bare adverb, e.g. *Tháining sé abhaile*, 'He came home(wards)', and as such is tagged as @ADVL.

(143)

```
"<Tháinig>" "tar" Verb VI PastInd Len @FMV                  Came
"<sé>"      "sé" Pron Pers 3P Sg Masc Sbj @SUBJ             he
"<abhaile>" "abhaile" Adv Dir @ADVL                  home(wards)
```

An adverbial particle with an adjective functions as an adverb, as in *Labhair go soiléir*, 'Speak clearly', as in (144).

(144)

```
"<Labhair>" "labhair" Verb VTI Imper 2P Sg @FMV_SUBJ        Speak
"<go>"      "go" Part Ad @>ADJ                              (part.)
"<soiléir>" "soiléir" Adj Base @ADVL                        clear
```

We also use adverbial tags to handle such adjuncts as *áiteanna eile chomh maith*, 'other places as well'.

(145)

```
"<áiteacha>"      "áit" Guess Noun Fem Com Pl @P<           places
"<eile>"          "eile" Det Dem @N<                       other
"<chomh>"         "chomh" Adv Its @>ADJ                    as
"<maith>"         "maith" Adj Base @ADVL                   well
```

Compound prepositions consist of a preposition and noun used idiomatically (and are, therefore, treated as multi-word expressions). They usually take an NP complement in the genitive. However, they are occasionally used without a noun complement. In such cases we tag them as @ADVL. This enables us to handle cases where they directly precede another prepositional phrase, as we do not wish to have a PP with a PP complement. In the following example we have *in aice le Brondesbury Park* 'next to Brondesbury Park'.

(146)

```
"<in aice>"        "in_aice" Prep Cmpd @ADVL                    beside
"<le>"             "le" Prep Simp @PP_ADVL                      with
"<Brondesbury>"  "Brondesbury" Prop Noun Masc Com Sg @P< Brondesbury
"<Park>"           "Park" Prop Noun Masc Com Sg @N<             Park
```

7.3.5 Predicates

Both the copula *is* 'is' and the verb *bí* 'to be' have arguments which consist of a subject and a predicate. In the case of the copula, the predicate is either an adjective or a noun phrase, and in the case the verb *bí* 'to be' the predicate can be an adjective or PP but not an NP.

Predicates of the copula *is* 'is' and substantive verb *bí* 'to be' are tagged as follows:
`@PRED, @PRED<`

The following two examples illustrate the use of the `@PRED` and `@PRED<` tags. In the first example (147), we have the verb *bí* with an adjectival predicate, *tá na daoine fairsing* 'the people are numerous'. In the second example (148), we have an inverted copular construction, i.e. the predicate comes before the subject, *Is náireach an scéal é* 'It is a shameful story'. In the copular construction, we interpret the word order of *náireach an scéal* 'shameful the story' as being a fronted form of *scéal náireach* 'shameful story'. We handle the unusual situation of the adjective coming before the noun, by tagging the noun as being dependent on the adjectival predicate, using the `@PRED<` tag.

(147)

```
"<Tá>"             "bí" Verb VI PresInd @FMV                    Are
"<na>"             "na" Art Pl Def @>N                          the
"<daoine>"         "duine" Noun Masc Com Pl Def @SUBJ           people
"<fairsing>"       "fairsing" Adj Base @PRED                    numerous
```

(148)

```
"<Is>"             "is" Cop Pres @COP                           Is
"<náireach>"       "náireach" Adj Base @PRED                    shameful
"<an>"             "an" Art Sg Def @>N                          the
"<scéal>"          "scéal" Noun Masc Com Sg DefArt @PRED<       story
"<é>"              "é" Pron Pers 3P Sg Masc @SUBJ               it
```

7.3.6 Adjectives

Adjectives are tagged according to whether the are used attributively or predicatively in this dependency analysis. Attributive adjectives are tagged as noun dependents, i.e. `@N<`, and predicative adjectives are tagged as predicates, i.e. `@PRED`.

7.4 Sentence Templates for Dependency Analysis

In this section, we present a set of abstract templates, which we use to illustrate the sentence patterns covered by our dependency tagging of Irish. These templates are not directly used in the implementation of the dependency analysis, but rather act as guidelines for applying dependency annotations using Constraint Grammar rules.

7.4.1 Introduction

Abney (1991) described a 'chunk' as a "single content word surrounded by a constellation of function words, matching a fixed template". We extend the usage of the word 'template', in this context, to describe a typical clause pattern in terms of a series of chunks. An NP in our implementation of Dependency Analysis correlates directly with this notion of a chunk, i.e. it contains a grammatical function item together with any possible dependants such as determiners and adjectives. An NP, therefore, can range from a single bare noun or pronoun to a complex NP which includes other modifying noun(s).

In Figure 28, we show a template which defines a simple sentence as consisting of at least a verb and a noun phrase, possibly followed by another noun phrase and/or prepositional phrase (depending on verb transitivity) as well as zero or more adjuncts. We use round brackets to denote zero or one instance and '*' to denote zero or more instances.

V	**NP**	**(NP)**	**(PP)**	**Adjunct***
@FMV	@SUBJ	@OBJ	@PP_OBL	@PP_ADVL @ADVL

Figure 28 Template for Sentence with Finite Main Verb (Analytic)

In each column of the table, we show the grammatical function or dependency relation tags which can be used in this position in the sentence pattern, i.e. items on seperate rows represent choice. In Figure 28, the optional adjunct(s) could either be prepositional phrases used adverbially or other bare adverbials. Each column position in the table can be assumed to also include dependants of the head.

VS	(NP)	(PP)	Adjunct*
@FMV_SUBJ	@OBJ	@PP_OBL	@PP_ADVL @ADVL

Figure 29 Template for Sentence with Finite Main Verb (Synthetic)

Alternatively, in the case of synthetic verbs, the verb and noun phrase are contained in one VS phrase, as shown in Figure 29. In general, we combine the distinctions between V and VS into one template table as in Figure 30, whenever possible.

7.4.2 Sentence Templates

A simple declarative sentence consists of a main clause only, with a complex sentence having a main clause and one or more subordinate clauses. There are several types of subordinate clause, the most important of which are complement clauses, relative clauses and adverbial clauses (Trask, 1992, p268). In the following sub-sections we introduce templates for the following syntactic constructions:

- Finite clauses with main verbs
- Finite clauses using the substantive verb *bí*
- Non-finite complement clauses using the verb *bí* as an auxiliary
- Copular constructions
- Infinitives
- Relative Clauses
- Complementizers
    - verbal
    - copular
- Other constructions such as:
    - Wh-questions
    - Passives
    - Phrasal verbs
    - Adverbial adjuncts
    - Conjunctions

7.4.3 Finite Main Clauses

The template in Figure 30 is used for simple sentences, whether declarative, negative or interrogative. The notation '`V(S) (NP)`' in the first row indicates that the subject may be either a separate NP or incorporated in an inflected verb form.

V(S)	(NP)	(NP)	(PP)	Adjunct*
@FMV	@SUBJ	@OBJ	@PP_OBL	@PP_ADVL
@FMV_SUBJ				@ADVL

Figure 30 Template for Sentence with Finite Main Verb

In the following examples, we have simple declarative, negative and interrogative sentences:

(149) Labhraíomar.
Spoke-1PL
@FMV_SUBJ
'We spoke'

(150) Níor labhair Seán.
NEG spoke Seán
@>V @FMV @SUBJ
'Seán didn't speak'

Yes/No questions are answered in Irish by repeating the verb, but not the subject (except for emphasis).

(151) Ar labhair Seán?
Q spoke Seán
@>V @FMV @SUBJ
'Did Seán speak?'

Affirmative answer:

(152) Labhair.
Spoke @FMV
[He] spoke

Negative answer:

(153) Níor labhair.
NEG spoke
@>V @FMV
[He] didn't speak

7.4.4 Finite Complement Clauses

Complement clauses complement some element of the main clause and, are usually, though not always, introduced by functional elements known as complementizers, some examples of which are listed below:

- *go/gur* - that
- *nach*/*nár* - that-NEG
- *a* - that, who, which
- *agus* - and
- *ó* - since

We annotate complementizers as clause boundaries, `@CLB`.

V(S)	**(NP)**	**(PP)**	**Cmpl.**	**V(S)**	**(NP)**	**Adjunct***
@FMV	@SUBJ	@PP_ADVL	@CLB	@FMV	@SUBJ	@PP_ADVL
@FMV_SUBJ				(+SUBJ)	@OBJ	@ADVL

```
(154) Dúirt sé    go   rachadh  sé.
      Said  he    that would-go he
      @FMV  @SUBJ @CLB @FMV     @SUBJ
      'He said that he would go'
```

7.4.5 Substantive Verb bí (to be)

The substantive verb *bí* (to be) is used to express various notions (An Gúm, 1999, p167; Christian Brothers, 1988, p117):

- state, including feelings and emotions
- possession
- location
- existence

The complement of the substantive verb is never a bare NP. It can be a predicative adjective or adverb and is frequently a PP, as shown in the following template.

V(S)	(NP)	Predicate	Adjunct*
@FMV	@SUBJ	@PRED	@PP_ADVL
@FMV_SUBJ		@ADVL	@ADVL
		@PP_ADVL	
		@PP_HAS	

Figure 31 Template for Substantive Verb *bí* 'to be'

7.4.5.1 State

(156) and (155) illustrate some of the ways in which states are expressed using the substantive verb.

(155) Tá an leabhar go maith.
Is the book PRT good
@FMV @>N @SUBJ @>ADJ **@ADVL**
'The book is good'

(156) Tá sé mór.
Is he big
@FMV @SUBJ **@PRED**
'He is big'

(157) expresses a comparative state. This particular type of predicate requires a conjoined subject, which we tag as @NP (see also (185)).

(157) Tá sliabh níos airde ná cnoc
Is mountain thing-PRT higher than hill
@**FMV** @SUBJ @**PRED** @N< @CC @NP
'A mountain is higher than a hill'

The following construction is used to express emotions and states, such as joy, sadness, hunger, thirst etc., as in (158): These nouns should be marked as 'abstract' in the lexicon to differentiate them from common nouns, e.g. *cóta* 'coat' (*Tá cóta orm* 'There is a coat on me', i.e. 'I am wearing a coat'). The abstract noun in these constructions cannot have any type of determiner such as *an* ' the' or *mo* 'my'.

(158) Tá áthas orm.
Is happiness on-me
@FMV @SUBJ **@PP_PRED**
'I am happy'

Adjectives can be used either predicatively or attributively. Predicative adjectives are used with definite subjects, i.e. a pronoun, proper noun, common noun with definite article, or synthetic verbs, and they are never inflected. If the subject is indefinite, e.g. *bríste* 'trousers' (160), the adjective will be attributive, i.e. dependent on the noun, and will be inflected to agree with the noun. In (159) we have a predicative adjective tagged as `@PRED`. In (160), we have two examples of attributive adjectives; each is tagged as `@N<`.

```
(159) Bhíomar   tinn  inné.
      Was-1PL   sick  yesterday
      @FMV_SUBJ @PRED @ADVL
      'We were sick yesterday'
```

```
(160) Bhíodh bríste   fada ann   chomh    maith le     bríste glúine
      Was    trousers long there    as     well  with   trousers knee
      @FMV   @SUBJ    @N<  @PP_ADVL @ADJ> @ADVL @PP_ADVL @P<    @N<
      'There were long trousers as well as knee-length trousers'
```

7.4.5.2 Possession

The combination of the verb *bí* (inflected for past tense as *bhí*) and the preposition *ag* 'at' is used to convey the meaning 'have' in Irish as in (161) below:

```
(161) Bhí  an  t-airgead ag       Seán.
      Was  the money     at       Seán
      @FMV @>N @SUBJ     @PP_HAS  @P<
      'Seán had the money'
```

7.4.5.3 Location

The predicate of the substantive verb can also be a prepositional phrase functioning locatively, as in (162), or we can have the prepositional pronoun *ann* 'in it' meaning 'there' which is used to express existence, as in (163). These prepositional phrases are not optional (as an adverbial phrase would be) therefore we tag them as prepositional predicates.

```
(162) Tá   an  carr  sa       gharáiste.
      Is   the car   in-the   garage
      @FMV @>N @SUBJ @PP_PRED @P<
      'The car is in the garage'
```

7.4.5.4 Existence

(163) Bhí rí ann fadó.
Was king in-it long-ago
@FMV @SUBJ **@PP_PRED** @ADVL
'There was a king, long ago.'

7.4.6 Non-Finite Complement Clauses with Verb bí 'to be' as Auxiliary

Non-finite complements involving the verbal noun are very common in Irish as they perform various aspectual functions as well as functioning as infinitives. The always occur with a finite auxiliary verb, most commonly the verb *bí* 'to be'. We provide a template, and some illustrative examples for the following aspectual uses:

- Progressive aspect
- Passive Progressive aspect
- Stative aspect
- Prospective aspect
- After Perfect aspect

7.4.6.1 Progressive Aspect

We propose the following template, Figure 32, for progressive aspectuals occurring with a finite auxiliary verb.

V(S)	**(NP)**	**Aspectual**	**NP***	**Adjunct***
@FAUX	@SUBJ	@PP_ASP	@OBJ_ASP	@PP_ADVL
@FAUX_SUBJ		@PP_STAT	@INF	@ADVL
				@PP_HAS

Figure 32 Template for Progressive Aspect

The verb *bí* 'to be' is used as an auxiliary verb with a non-finite complement, as in (164), where the non-finite (progressive) complement is tagged with the @PP_ASP tag.

(164) Tá sé ag iascaireacht.
Is he at fishing
@FAUX @SUBJ **@PP_ASP @P<**
'He is fishing'

In (165) - (168), we also indicate the object of the non-finite clause using the @OBJ_ASP tag. In progressive aspectual constructions, the aspectual object usually follows the verbal noun in the genitive case, *an dorais* 'the door' in (165), or as a prepositional complement, *liom* 'with me' in (166).

(165) Tá Seán ag oscailt an dorais
Is Seán at opening the door
@FAUX @SUBJ **@PP_ASP** **@P<** @>N **@OBJ_ASP**
'Seán is opening the door'

(166) Tá sé ag cabhrú liom
Is he at helping with-me
@FAUX @SUBJ **@PP_ASP** @P< **@PP_ADVL**
'He is helping me'

However, the aspectual object may also occur before the verbal noun in the case of pronominal objects. In this case, it is realised as a possessive pronoun *mo* 'my' (167).

(167) Tá sé do mo chabhrú
Is he to my helping
@FAUX @SUBJ **@PP_ASP** **@OBJ_ASP** @P<
'He is helping me'

The aspectual object of verbal nouns such as *dul* 'going', may also be an infinitive, *a chodladh* 'to sleep' in (168).

(168) Tá sé ag dul a chodladh
He is at going to sleep
@FAUX @SUBJ **@PP_ASP** @P< @>N **@INF**
'He is going to sleep'

In (169), the progressive aspectual clause *Seán ag oscailt an dorais* 'Seán opening the door' is the complement of a finite verb of perception, *chonaic* 'saw'. We indicate the subject and object of the non-finite clause using the @SUBJ_ASP and @OBJ_ASP tags. (This construction, having a finite main verb rather than auxiliary, is more suited to the template in Figure 31).

(169) Chonaic mé Seán ag oscailt an dorais
Saw I Seán at opening the door
@FMV @SUBJ **@SUBJ_ASP** **@PP_ASP** **@P<** @>N **@OBJ_ASP**
'I saw Seán opening the door'

7.4.6.2 Passive Progressive Aspect

In the case of the passive progressive, the aspectual object is realised as the subject of the auxiliary verb. The aspectual preposition changes from *ag* 'at' to *á* 'to' with an incorporated pronoun, e.g. *á* 'to its' in (170).

```
(170) Tá    cáca  á       dhéanamh agam
      Is    cake  to-its  making   at-me
      @FAUX @SUBJ @PP_ASP @P<      @PP_HAS
      'A cake is being made by me'
```

7.4.6.3 Stative Aspect

In the case of stative aspect, the aspectual preposition changes from *ag* 'at' to *ar* 'on', as in (171), or *i* 'in' with an incorporated pronoun, e.g. *ina* 'in-his', in (172).

```
(171) Tá    an  doras ar       oscailt
      Is    the door  on       opening
      @FAUX @>N @SUBJ @PP_STAT @P<
      'The door is open'
(172) Tá   sé    ina      chodladh.
      Is   he    in-his   sleeping
      @FMV @SUBJ @PP_STAT @P<
      'He is asleep'
```

7.4.6.4 Prospective Aspect

In the case of prospective aspect, the aspectual preposition *le* 'with', or *chun* 'towards' is used with an infinitive, to express the meaning of an intended future action. In (173), we have *le* 'with' and an intransitive verbal noun *fanacht* 'waiting'.

```
(173) Tá    sé    le      fanacht
      Is    he    with    waiting
      @FAUX @SUBJ @PP_ASP @INF
      'He is going to wait'
```

In (174), we have the object, *cáca* 'cake', of the infinitival complement occurring in its usual position before the verbal noun *déanamh* 'making'. The aspectual object *cáca* 'cake' is also the subject of the finite auxiliary, *tá* 'is', or more correctly, the aspectual clause *cáca le déanamh* 'cake to make' is the subject of the finite auxiliary. (The logical subject 'I' is incorporated in the prepositional pronoun *agam* 'at me' which forms part of the *tá ... ag* ' is ... at' construction meaning 'has', see Section 7.4.5.2).

(174) Tá cáca le déanamh agam
Is cake to making at-me
@FAUX **@OBJ_ASP @PP_ASP @INF @PP_HAS**
'I have to make a cake' lit. 'A cake is to be made by me'

Alternatively, we can have an overt subject, e.g. *mé* 'I' in (175), with the infinitive and its preposed object forming the complement of the aspectual preposition *chun* 'towards'.

(175) Tá mé chun cáca a dhéanamh inniu.
Is I towards cake to make today
@FAUX @SUBJ **@PP_ASP @OBJ_INF @>N @INF** @ADVL
'I am going to make a cake today'

7.4.6.5 After Perfect

The compound prepositions *tar éis* 'after' and *i ndiaidh* 'after', used with a verbal noun, express the meaning of an action recently completed. Considering that the verbal noun, functioning as an infinitive, takes a preposed object, we tag the verbal noun as an infinitive, `@INF`, in this type of construction, as shown in (176).

(176) Tá mé tar éis cáca a dhéanamh
Is I after cake to make
@FAUX @SUBJ **@PP_ASP @OBJ_INF @>N @INF**
'I am after making a cake' OR 'I have just made a cake'

7.4.7 Copula is (to be)

The following description of the uses of the copula *is* follows Doherty (1996) and also the New Irish Grammar (Christian Brothers, 1988, p122-5).

The copula is widely used in Irish and performs a number of functions. In order to parse the variety of copular constructions we propose templates for each of the following usages:

- Identity (Equative) Constructions — Figure 33
- Classificatory Constructions — Figure 34
- Ownership Constructions — Figure 35
- Comparative Constructions — Figure 36
- Fronted Constructions — Figure 37, Figure 38
- Idiomatic Constructions — Figure 39
- Copular Complements — Figure 40

7.4.7.1 Identity Sentences

Identity sentences follow the pattern copula-subject-predicate. In these sentences, both subject and predicate must be definite NPs and these sentences generally have the meaning "subject is predicate". (Christian Brothers, 1988, p124). Definite subjects and predicates include proper nouns, pronouns or common nouns with the definite article. The predicate is a complement of the subject.

The following template is used for Identity sentences:

COP	Definite NP	Definite Predicate
@COP	@SUBJ	@PRED

Figure 33 Template for Identity Copula

An example of an identity use of the copula is illustrated in (177).

```
(177) Ní      mise   an  múinteoir
      COP-NEG I-EMPH the teacher
      @COP    @SUBJ  @>N @PRED
      'I am not the teacher'
```

When the subject is in the 3rd person it is preceded by an augment pronoun. Such constructions are known as Augmented Copular Constructions (Adger and Ramchand, 2003; Doherty, 1997). Example (178) illustrates the use of the augment pronoun with a 3rd person subject in an identity/equative construction.

```
(178) An    iad       na  daoine siúd  na  buaiteoirí?
      COP-Q them      the people those the winners
      @COP  @AUG>SUBJ @>N @SUBJ  @N<   @>N @PRED
      Are those people the winners?
```

7.4.7.2 Classificatory Sentences (Inverted Copular Constructions)

These sentences follow the pattern copula-predicate-subject. They are known as classification sentences as the 'subject' is said to be a member of the class 'predicate'. The predicate complement must be an indefinite noun or an adjective. These constructions are also known as Inverted Copular Constructions (Adger and Ramchand, 2003) as the predicate comes before the subject.

COP	Indefinite Predicate	NP
@COP @COP_WH	@PRED	@SUBJ

Figure 34 Template for Classificatory Copula

In (179), we have an example of a widely used inverted copular construction.

(179) Is lá deas é
COP day nice it
@COP @PRED @N< @SUBJ
'It is a nice day'

In (180), where the adjective comes before the noun, i.e. *deas an lá* 'nice the day', we are treating this as an alternative (fronted) version of the predicate *lá deas* 'a nice day' in (179).

(180) Is deas an lá é
COP nice the day it
@COP @PRED @>N @PRED< @SUBJ
'It is nice (that)the day is' i.e. 'It is a nice day'

(180) is possibly a fronted copular version of (181), see also Figure 30.

(181) Tá an lá go deas
Is the day PRT nice
@FMV @>N @SUBJ @>ADJ @ADVL
'The day is nice'

7.4.7.3 Ownership (is-le Constructions)

The copula *is* together with the preposition *le* 'with' denotes ownership. The item which is owned must be a definite NP (Ó Siadhail, 1989, p233).

COP	Le + Definite Predicate	Definite NP.
@COP @COP_WH	@PP_PRED	@SUBJ

Figure 35 Template for Ownership Copula

In this type of construction, it can be difficult to determine where the subject is. In (182) it is not obvious whether *teach* 'house' or *Dónal* is the subject. We follow the New Irish Grammar (Christian Brothers, 1988, p125) in assigning the subject role to *an teach* 'the house', and tag the prepositional phrase *le Dónal* 'with Dónal' as `@PP_PRED`.

```
(182) Is   le       Dónal  an  teach
      COP  with     Dónal  the house
      @COP @PP_PRED @PRED< @>N @SUBJ
      'The house is Donal's OR
      'The house belongs to Dónal OR 'Dónal owns the house'
```

Note that when the definite predicate is a pronoun, it combines with the preposition *le* 'with' to form a prepositional pronoun.

```
(183) Ní      liomsa       an  t-airgead
      COP-NEG with-me-EMPH the money
      @COP    @PP_PRED     @>N @SUBJ
      'The money is not mine'

(184) Cé      leis     an  teach?
      COP-WH  with-it  the house
      @COP_WH @PP_PRED @>N @SUBJ
      'Who's is the house?' lit. 'With whom is the house?'
```

Note that in (184) *cé* 'who' is tagged as `@COP_WH`, as the interrogative pronoun is used as the question form of the ownership copula (Christian Brothers, 1988, p124).

7.4.7.4 Comparatives

The copula *is* together with the comparative form of an adjective can be used in making comparisons. (185) is an alternative to (and perhaps a fronted form of) the type of comparative structures using the verb *bí* 'to be' in (157).

COP	Adj. Predicate	Indefinite Conjoined NPs.
@COP	@PRED	@SUBJ

Figure 36 Template for Comparative Copula

```
(185) Is   airde  sliabh   ná   cnoc
      COP  higher mountain than hill
      @COP @PRED  @SUBJ    @CC  @NP
      'A mountain is higher than a hill'
```

7.4.7.5 Fronting (Preposing) for emphasis

Any phrase type, including VNP clauses,, can be fronted using a copula, predicate, and relative verb.

COP	Fronted NP/PP	Relative Verb	Remaining Constituents
@COP	@PRED @PP_OBL @PP_ASP	@CLB	@SUBJ @OBJ @PP_OBL @PP_ADVL

Figure 37 Template for Fronting Using a Copula

The various arguments of the verb *tabhair* 'give' (*thug* in the past tense) in (186), are fronted in examples (187)-(189).

(186) Thug sí leabhar do Mháire.
Gave she book to Máire
@FMV @SUBJ @OBJ @PP_OBL @P<
'She gave a book to Máire'

(187) Is ise a thug leabhar do Mháire
COP she-EMPH that gave book to Máire
@COP **@PRED** @>V @FMV_REL @SUBJ_OR_OBJ @PP_OBL @P<
'It is she that gave a book to Mary'

(188) Is leabhar a thug sí do Mháire
COP book that gave she to Máire
@COP **@PRED** @>V @FMV_REL @SUBJ @PP_OBL @P<
'It is a book that she gave to Máire'

(189) Is do Mháire a thug sí leabhar.
COP to Máire that gave she book
@COP **@PP_OBL** @P< @>V @FMV_REL @SUBJ @OBJ
'It is to Máire that she gave a book'

In (190), we have a copular construction expressing a fronted version of the substantive verb *bí* 'to be' and its non-finite complement (191):

(190) Is ag iascaireacht atá sé
COP at fishing REL-is he
@COP **@PP_ASP** @P< @FAUX_REL @SUBJ
'Tis fishing he is.

```
(191) Tá    sé    ag      iascaireacht.
      Is    he    at      fishing
      @FAUX @SUBJ @PP_ASP @P<
      'He is fishing'
```

In (192), we have a copular construction itself being fronted, e.g. a fronted classificatory copular construction of (193).

Indef. Predicate	COP	Definite NP
@PRED	@COP	@SUBJ

Figure 38 Template for Fronted Copular Construction

```
(192) Cailín is   ea        í
      Girl   COP  PRON      she
      @PRED  @COP @AUG>SUBJ @SUBJ
      'A girl is what she is'

(193) Is   cailín í
      COP  girl   she
      @COP @PRED  @SUBJ
      'She is a girl'
```

7.4.7.6 Idiomatic Use

Copular constructions are used to express feelings or desires in an idiomatic manner (Mac Congáil, 2002, p165):

- *Is maith liom* 'I like', i.e. It is good with me
- *Is fearr liom* 'I prefer', i.e. It is better with me
- *Is aoibhinn liom* 'I love/enjoy', i.e. It is delightful with me
- *Is oth liom* 'I regret', i.e. It is regretted by me
- *Is fuath liom* 'I hate', i.e. It is hated by me
- *Is léir dom* 'It is clear to me'
- *Is eol dom* 'I know', i.e. It is known to me
- *Is dócha* 'I suppose', i.e. It is likely/probable
- *Is mian liom* I wish'

COP	Adj	PP	(NP)	NP
@COP	@PRED	@PP-SUBJ	@OBJ_INF	@INF @OBJ

Figure 39 Template for Idiomatic Use of the Copula

In sentences such as *Is maith liom milseáin* 'I like sweets' (194), it is difficult to decide the location of the subject, as can be seen from the alternative translations. However, in the interests of semantic interpretation, we have decided to tag the prepositional phrase *liom* 'with me' as the subject, and *milseáin* 'sweets' as the object. This decision is supported by evidence from emphatic responses in Irish where it is the subject, *liom* 'with me', which is retained while the object, *milseáin* 'sweets', is dropped (195). (A non-emphatic response would be *Is maith* 'Like').

(194) Is maith liom milseáin
COP good with-me sweets
@COP @PRED @PP_SUBJ @OBJ
'I like sweets' OR 'Sweets are good with me'

(195) An maith leat milseáin?
COP good with-you sweets
@COP @PRED @PP_SUBJ @OBJ
'Do you like sweets?'

Is maith liom ...
COP good with-me ...
@COP @PRED @PP_SUBJ
'I like indeed'

Example (196) demonstrates the use of a copular construction with an infinitival complement, used to express a wish or desire.

(196) Ba mhaith liom teach a cheannach
COP-COND good with-me house to buy
@COP @PRED @PP_SUBJ @OBJ_INF @>N INF
'I would like to buy a house' OR 'Is would be good with me to buy a house'

7.4.7.7 Copular Complements

V(S)	NP	(PP)	COP	NP	(PP)	NP
@FMV	@SUBJ	@PP_ADVL	@CLB	@PRED	@PP_ADVL	@SUBJ
@FMV_SUBJ						

Figure 40 Template for FMV introducing Copular Complements

A copular complement can be introduced by either a finite main verb, as in (197), or another copula, as in (198).

(197) Dúirt sé gur múinteoir é
Said he COP teacher he?
@FMV @SUBJ @CLB @PRED @SUBJ
'He said that he is a teacher'

COP	NP	COP	NP	(PP)	NP
@COP	@PRED	@CLB	@PRED	@PP_ADVL	@SUBJ

Figure 41 Template for Copula introducing Copular Complements

(198) Ní hé nár mhaith liom é
COP-NEG it COP-NEG good with-me it
@COP @PRED @CLB @PRED @PP_ADVL @SUBJ
'It is not that I did not like it'

7.4.8 Infinitives

Infinitives are formed using the verbal noun. Infinitival objects precede the verbal noun.

7.4.8.1 Infinitives with Auxiliary Verb

V	NP	(Aspectual)	(NP)	NP	Adjunct*
@FAUX	@SUBJ	@PP_ASP	@OBJ_INF	@INF	@PP_HAS @PP_ADVL

Figure 42 Template for Infinitive with Auxiliary Verb

In (199), we have an intransitive infinitive, *fanacht* 'wait', while in (200), we have a transitive infinitive *déanamh* 'make/do' preceded by its object é 'it' and the infinitival particle *a*.

(199) Caithfidh mé fanacht
Must I stay
@FAUX @SUBJ **@INF**
'I must stay'

(200) Caithfidh mé é a dhéanamh
Must I it PRT do/make
@FAUX @SUBJ **@OBJ_INF** @>N **@INF**
'I must do/make it'

7.4.8.2 Infinitives with the Copula

In (201), we have the commonly occurring construction of copula and infinitive. This construction is similar to (196), except that this example, (201), also includes the negative preposition *gan* 'not/without'. This sentence, therefore, fits the template in Figure 39.

(201) B' fhearr liom gan fanacht
COP better with-me NEG stay
@COP @PRED @PP_ADVL **@PP_NEG** **@INF**
'I would prefer not to stay'

7.4.9 Relative Clauses

Relative clauses are usually post modifiers of a noun phrase[30] in the main clause. This noun phrase may be the subject or the object of the relative clause (Trask, 1992, p238). There are two types of relative clause in Irish; direct and indirect. Information on Irish relative clauses is available in a number of sources (An Gúm, 1999, p265; Christian Brothers, 1988, p143; McCloskey, 1979; 1985; Ó Baoill and Ó Tuathaill, 1992; Ó Siadhail, 1989, p311).

7.4.9.1 Direct Relative

As shown in the template in Figure 43, a direct relative clause can be introduced by a main clause containing a finite main verb, a copula, or by an adverbial.

[30] Except for instances of fronted adverbial noun phrases in the main clause, e.g. 'It was at 3 o'clock that Mary came home.'

Main Clause	Rel V(S)	NP
@FMV+@SUBJ	@FMV_REL	@SUBJ
@COP+@SUBJ		@OBJ
@ADVL		@SUBJ_OR_OBJ
@PP_ADVL		

Figure 43 Template for Direct Relative Clauses

In direct relatives, the subject of the main clause is either the subject (202), or the object (203), of the relative clause. This subject or object is elipted leaving a 'gap' in the relative clause. In the following examples the gap indicating the elipted constituent (which is co-referential with the subject of the main clause) is denoted by an underscore, '_'.

(202) D' fhág an fear a d' ionsaigh _ **iad.**
PRT Left the man that PRT attacked them
@>V @FMV @>N **@SUBJ** @>V @>V @FMV_REL **@OBJ**
'The man that attacked them left'

(203) D' fhág an fear a d' ionsaigh **siad** _.
PRT Left the man that PRT attacked they
@>V @FMV @>N **@SUBJ** @>V @>V @FMV_REL **@SUBJ**
'The man they attacked left'

Although the surface word order in (202) and (203) is the same, we can tell from the form of the pronoun in the relative clause whether a subject or an object has been elipted in the relative clause. In (203), the subject pronoun, *siad* 'they' indicates that the object has been elipted, whereas, in (202), the non-subject pronoun, *iad* 'them' is used which indicates that the subject has been elipted.

While this distinction can be seen in pronouns, this subject-object distinction is not overtly marked on nouns. In (204), the subject of the main clause, *fear* 'man', is the subject of the relative clause, whereas in (205), the subject of the main clause, *bád* 'boat' is the object of the relative clause. We rely on the lexical (or semantic) properties of the verb *chonaic* 'saw' in order to interpret the sentence i.e. that this verb requires an animate subject.

(204) D' fhág an fear a chonaic _ an bád.
PRT Left the man that saw the boat
@>V @FMV @>N @SUBJ @>V @FMV_REL @>N @OBJ
'The man that saw the boat left'

(205) D' fhág an bád a chonaic an fear _ .
PRT Left the boat that saw the man
@>V @FMV @>N @SUBJ @>V @FMV_REL @>N @SUBJ
'The boat that the man saw left'

When both the subject and object are animate nouns, there is inherent ambiguity, as we cannot tell whether the elipted constituent is the subject or the object of the relative clause, as in (206).

(206) D' fhág an fear a chonaic ? an bhean ? .
PRT Left the man that saw the woman
@>V @FMV @>N @SUBJ @>V @FMV_REL @>N **@SUBJ_OR_OBJ**
'The man that the woman saw left'
OR
'The man that saw the woman left'

This inherent ambiguity in relative clauses is one of the most difficult problems to solve in Irish parsing which is why we use the category **@SUBJ_OR_OBJ**. In the current dependency analysis, we can handle (202) and (203), where morphologically distinct pronouns are used. We would need to introduce subcategorization frames for verbs and semantic classes for nouns, in order to interpret (204) and (205) correctly. Example (206) is even more difficult, in that we would need wider contextual information which goes beyond the scope of the sentence in order to resolve the ambiguity.

As with simple declarative sentences, we can have a synthetic verb-form in the relative clause, e.g. in (207), we have an autonomous verb-form. This type of relative clause presents no problem as the subject is morphologically marked on the verb-form.

(207) An lá a cuireadh Butt ...
The day that put-AUTO Butt ...
@>N @ADVL @>V @FMV_REL_SUBJ **@OBJ**
'The day that Butt was buried ...

In the following type of direct relative, the subject follows the auxiliary verb in the embedded clause. The object of the main clause, *obair* 'work', is also the object of the progressive complement *a dhéanamh* 'doing' in the relative clause.

(208) Chonaic mé an obair a bhí Seán a dhéanamh _
Saw I the work that was Seán at-its doing
@FMV @SUBJ @>N **@OBJ_ASP** @>V @FAUX_REL **@SUBJ @PP_ASP @P<**
'I saw the work that Seán was doing'

7.4.9.2 Indirect Relative

All of the examples in (209)-(212), are covered by the template in Figure 44.

Main Clause	Rel V(S)	NP	(NP)	(PP)
@FMV+@SUBJ @COP+@SUBJ @ADVL @PP_ADVL	@FMV_REL	@SUBJ	@OBJ	@PP_ADVL

Figure 44 Template for Indirect Relatives

In the case of indirect relatives, the subject of the relative clause is not the same as the subject of the main clause. Example (204) is extended in (209), to introduce an indirect subject *a mhac* 'his son' in the relative clause.

(209) D' fhág an fear a[31] chonaic a mhac an bád.
PRT Left the man that saw his son the boat
@>V @FMV @>N @SUBJ @>V @FMV_REL @>N @SUBJ @>N @OBJ
'The man whose son saw the boat left'

Note that the ambiguity of (206) is resolved in (210), when an indirect subject *a mhac* 'his son' is introduced.

(210) D' fhág an fear a chonaic a mhac an bhean.
PRT Left the man that saw his son the woman
@>V @FMV @>N @SUBJ @>V @FMV_REL @>N @SUBJ @>N @OBJ
'The man whose son saw the woman left'

Relatives with resumptive pronouns in the embedded clause, are always indirect relatives. (211) is an example of a relative clause with a resumptive pronoun *é* 'it' (Ó Baoill and Ó Tuathaill, 1992, p213), while (212) is an example of an indirect relative with the resumptive prepositional pronoun *air* 'on it' (Christian Brothers, 1988, p144).

(211) Chonaic mé an crann a bhuail an tintreach é.
Saw I the tree that hit the lightening it
@FMV @SUBJ @>N @OBJ @>V @FMV_REL @>N @SUBJ @OBJ
'I saw the tree that the lightening hit.'

[31] The relativizer for the past tense is usually *ar* 'that', however, in these examples *a* is used as the verb *feic*' to see' (past tense *chonaic* 'saw') is irregular.

```
(212) Chonaic mé    an  crann a    bhfuil   na  húlla  air.
      Saw     I     the tree  that is       the apples on-it
      @FMV    @SUBJ @>N @OBJ  @>V  @FMV_REL @>N @SUBJ  @PP_ADVL
      'I saw the tree that the apples are on'
```

The three types of indirect clause above are described in *Úrchúrsa Gaeilge* (Ó Baoill and Ó Tuathaill, 1992, p213), as genitive (210), accusative (211) and dative (212) indirect relatives, respectively.

7.4.9.3 Pronominal Relative

In the following examples, an object pronoun is understood to be included in the relativizer *a* 'that' or 'which'. This type of relative is covered by the template in Figure 44.

```
(213) Íocfaidh mé    as       a          gceannóidh tú
      Will-pay I     out      that-which will-buy   you
      @FMV     @SUBJ @PP_ADVL @>V        @FMV_REL   @SUBJ
      'I will pay for what (that which) you buy'

(214) Sin       a          bhfuil   ann
      That-is   that-which is       in-it
      @COP_SUBJ @>V        @FMV_REL @PP_ADVL
      'That is all there is'
```

7.4.10 Other Syntactic Constructions

In this section, we introduce a range of constructions, i.e. Wh-Questions, Passives, Phrasal Verbs, Adverbial Clauses, Dative Shift, Conjunctions, and NP Fragments.

7.4.10.1 Wh-Questions

Interrogative	**Rel V(S)**	**(NP)**	**(NP)**	**(PP)**
@SUBJ @OBJ @PP_ADVL	@FMV_REL	@SUBJ	@OBJ	@PP_OBL

Figure 45 Template for Wh-Questions

Interrogatives (which require an answer other than yes or no), consist of an interrogative pronoun or an adverbial such as *cathain* 'when' followed by a relative verb construction. Examples (215)-(219) demonstrate some common wh-question constructions, all of which fit the template in Figure 45.

(215) Cé a labhair?
Who REL spoke
@SUBJ @>V @FMV
'Who spoke?'

(216) Cé nár labhair?
Who REL-NEG spoke
@SUBJ @>V @FMV
'Who didn't speak?'

(217) Cad a thug sí do Mháire?
What REL gave she to Máire
@OBJ @>V @FMV @SUBJ @PP_OBL @P<
'What did she give to Máire?'

(218) Cé dó a thug sí an leabhar?
Who to-him REL gave she the book
@COP @PP_OBL @>V @FMV @SUBJ @>N @OBJ
'To whom did she give the book?'

(219) Cathain a thug sí an leabhar do Mháire?
When REL gave she the book to Máire
@ADVL @>V @FMV @SUBJ @>N @OBJ @PP_OBL @P<
'When did she give the book to Máire?'

7.4.10.2 Passive Constructions

In addition to the passive progressive aspect (see 7.4.6.2), there are two passive-like constructions in Irish. The first uses an impersonal (autonomous) verb form, as in Figure 46, while the second construction uses a verbal adjective with the substantive verb, as in Figure 47. The former focuses on the action while the latter focuses on the state.

Autonomous VS Chunk	(NP)	Adjunct*
@FMV_SUBJ	@OBJ	@PP_ADVL

Figure 46 Template for Passive Using Autonomous Verb Form

The impersonal (autonomous) form of a transitive verb corresponds most closely to the passive form in other languages (An Gúm, 1999, p166). Intransitive verbs may also be used in this way.

(220) Deisíodh an rothar.
Fixed-AUTO the bicycle
@FMV_SUBJ @>N @OBJ
'One fixed the bicycle' (i.e. The bicycle was fixed)

The impersonal form may not be used in combination with an animate agent, i.e. with a synthetic verb you can't have another subject, e.g. (221) is not allowed. Where an animate agent is required, it must be expressed in the active voice (222).

(221) *Deisíodh an rothar ag Seán.
Fixed-AUTO the bicycle at Seán.
@FMV_SUBJ @>N @OBJ @PP_ADVL @P<
*One fixed the bicycle at/by Seán'

(222) Dheisigh Seán an rothar.
Fixed Seán the bicycle
@FMV @SUBJ @>N @OBJ
'Seán fixed the bicycle.

The impersonal form, may however, be used with an inanimate agent (i.e. instrument), e.g. *stoirm* 'storm' (223) or instrument, e.g. *clocha* 'stones' (224) (Ó Baoill and Ó Tuathaill, 1992, p64-5).

(223) Briseadh an fhuinneog leis an stoirm.
Broke-AUTO the window with the storm
@FMV_SUBJ @>N @OBJ @PP_ADVL @>N @P<
'The window was broken by the storm'

(224) Líonadh an poll le clocha.
Filled-AUTO the hole with stones
@FMV_SUBJ @>N @OBJ @PP_ADVL @P<
'The hole was filled with stones'

Alternatively, a verbal adjective with the substantive verb *bí* 'to be' may be used in a manner that is similar to the passive, except that it describes a state rather than an action.

V(S)	(NP)	Verbal Adj.	Adjunct*
@FMV	@SUBJ	@PRED	@PP_ADVL

Figure 47 Template for Passive Using Verbal Adjective

```
(225) Bhí  an  geata dúnta.
      Was  the gate  closed
      @FMV @>N @SUBJ @PRED
      'The gate was closed' (i.e. The gate was in a closed state)
```

The combination of the verb *bí* and the preposition *ag* are used to convey the meaning 'have' in Irish, (see (161), page 157). This combination together with a verbal adjective, has been translated as a passive perfective by Ó Siadhail (1989, p299). In (226), we show the verbal adjective *léite* 'read'.

```
(226) Tá   an  leabhar léite agam.
      Is   the book    read  at-me.
      @FMV @>N @SUBJ   @PRED @PP_HAS
      'I have read the book'
```

It is not clear whether the following usage is entirely grammatical or not, (i.e. the inclusion of an external agent in a stative construction) but should it occur in texts our system will tag it as shown in (227).

```
(227) ?Bhí  an  geata dúnta  ag      Seán.
       Was  the gate  closed at      Seán
       @FMV @>N @SUBJ @PRED @PP_HAS @P<
      'Seán had the gate closed' (i.e. The gate was in the state of
      having been closed by Seán)
```

The resultative aspect is sometimes regarded as synonymous with the perfect aspect. Dahl (1985) (quoted in Trask (1992, p240)) argues that the resultative focuses on the present state (e.g. He is gone) while the perfective focuses more on the action which has lead to the present state (e.g. He has gone). If we accept this distinction then this use of the verbal adjective could be described as a resultative.

7.4.10.3 Phrasal Verbs

Phrasal verb constructions, i.e. verb-preposition combinations, are treated similarly to other finite main verb constructions, except that we treat the subject as being part of a prepositional phrase.

V Chunk	(Adverbial)	PP	(NP)	(PP)	Adjunct*
@FMV	@ADVL	@PP_SUBJ	@OBJ	@PP_OBL	@PP_ADVL @ADVL

Figure 48 Template for Sentence with Finite Phrasal Verb

Phrasal verbs are constructions in which the verb together with a particle (usually a preposition) has an idiomatic meaning. In the following example, *éirigh* 'rise' and *leis* 'with' together mean 'succeed'. In cases where the preposition and the subject are separate, we could analyse the preposition as being dependent on the verb as a post modifier as in (228)a. However, the fact that a pronoun subject combines with the preposition (229) (as a prepositional pronoun), means that this option is not feasible, as we would be including the subject with the preposition as a post modifier of the verb. As subjects have not otherwise been treated as modifiers we have rejected this option. Instead, we have opted to analyse the preposition as a PP head with subject using the tag `@PP_SUBJ` as shown in (229).

(228) D' éirigh leis an mac léinn sa scrúdú.
PRT rose with the student in_the exam
a)*@>V @FMV @V< @>N @SUBJ @N< @PP_ADVL @P<
b) @>V @FMV **@PP_SUBJ** @>N @P< @N< @PP_ADVL @P<
'The student passed the exam'

(229) D' éirigh liom sa scrúdú.
PRT rose with-me in_the exam
@>V @FMV **@PP_SUBJ** @PP_ADVL @P<
'I succeeded in the exam' i.e. 'I passed the exam'

Apart from idiomatic verbs, there are many other instances of verbs with prepositions, where the preposition combines with a pronoun, e.g. *dar leis* 'according to him'.

(230) Dar leis tá an teach réidh
According to-him is the house ready
@FMV **@PP_SUBJ** @FMV @>N @SUBJ @PRED
'According to him the house is ready'

A distinction can be made between phrasal verbs and prepositional verbs (Trask, 1992, p215). The preposition is more closely bound to the verb in phrasal verbs. In (228), the subject cannot intervene between the verb and the preposition, though a non-prepositional adverb seems acceptable. In (231), the adverb *go maith* 'well' comes between the verb and the prepositional subject.

(231) D' éirigh go maith leis an mac léinn sa scrúdú.
PRT rose PRT well with the student in_the exam
@>V **@FMV** @>ADJ @ADVL **@PP_SUBJ** @>N @SUBJ @N< @PP_ADVL @P<
'The student succeeded well in the exam'

In (232), when we insert the subject, *an mac léinn*, 'the student', between the verb and the prepositional subject we lose the idiomatic meaning of *éirigh le* 'rise with = succeed'. The extra argument, as wll as the incorrect word order, has the effect of making this sentence ungrammatical.

(232) *D' éirigh an mac léinn leis sa scrúdú.
PRT rose the student with-it in-the exam
@>V @FMV @>N @SUBJ @N< **@PP_ADVL** @PP_ADVL @P<
? 'The student rose with it/him in the exam

In (233), inserting the prepositional adverbial *sa scrúdú*, 'in the exam', between the verb and preposition, results in an ungrammatical structure, with no apparent subject. We give two alternative analyses; a) the preposition *leis* 'with' is interpreted as a simple preposition, or b) *leis* 'with him/it' is interpreted as a prepositional pronoun.

(233) *D' éirigh sa scrúdú leis an mac léinn.
a) PRT rose in-the exam with the student
@>V @FMV @PP_ADVL @P< **@PP_ADVL @>N @P< @N<**
*Rose in the exam with the student
b) PRT rose in-the exam with-it the student
@>V @FMV @PP_ADVL @P< **@PP_ADVL** @>N @NP @N<
*Rose in the exam with it/him the student

7.4.10.4 Dative Shift

In English, a sentence like (234) can be expressed as (235), where the indirect object Mary can come before the direct object and lose its preposition.

(234) John gave a book to Mary

(235) John gave Mary a book

This has no counterpart in Irish as there cannot be more than two NP's per clause outside of prepositional phrases (Stenson, 1981, p65).

```
(236) Thug Seán  leabhar do      Mháire
      Gave Seán  book    to      Mary
      @FMV @SUBJ @OBJ    @PP_OBL @P<
      Seán gave a book to Máire
```

(237) is not a valid structure in Irish but would be tagged as follows:

```
(237) *Thug Seán  Máire leabhar
       Gave Seán  Máire leabhar
       @FMV @SUBJ @N<   @OBJ
      ? Seán Máire gave a book
```

7.4.10.5 Adverbial Clauses

Adverbial clauses elaborate on the main clause as a whole, or some element of it, by providing information on manner, place, time, reason etc. (Brown and Miller, 1991, p93; Ó Siadhail, 1989, p267).

An adverbial clause does not have to contain an actual adverb; we can have PPs or NPs functioning adverbially as in (238). There can be several adverbials in the same sentence, and they can appear in any order.

```
(238) Tar_éis  trí   lá  tháinig sé    abhaile.
      After    three day came    he    home
      @PP_ADVL @>N   @P< @FMV    @SUBJ @ADVL
      'After three days he came home.'
```

(239) and (240) exemplify different types of adverbial clause.

```
(239) Tóg  go    bog   é
      Take PRT   soft  it
      @FMV @>ADJ @ADVL @SUBJ
      'Take it easy'

(240) D'  fhan   sé    ansin le       fiche  bliain.
      PRT Stayed he    there for      twenty years
      @>V @FMV   @SUBJ @ADVL @PP_ADVL @>N    @P<
      'He stayed there for twenty years'
```

7.4.10.6 Conjunctions

There are a great variety of constructions which can be conjoined using coordinate conjunctions. In (241), we have conjoined prepositional phrases.

(241) go hAlbain agus go Sasana
to Scotland and to England
@PP_ADVL @P< **@CC** @PP_ADVL @P<
'to Scotland and to England'

The following is an example of coordinated independent clauses. We use the @CLB tag to denote the clause boundary; in this case it is attached to the coordinating conjunction *agus* 'and'. (All subordinating conjunctions are tagged as clause boundaries.).

(242) Cheannaigh Seán leabhar agus léigh sé é
Bought Seán book and read he it
@FMV @SUBJ @OBJ **@CLB** @FMV @SUBJ @OBJ
'Seán bought a book and he read it'

In conjoined sentences, where the subject of each is the same, it is often elipted in the second clause as illustrated in (243).

(243) Thug sé freagra orm go múinte agus shiúil leis
Gave he an-answer on-me PRT polite and walked with-him
@FMV @SUBJ @OBJ @PP_ADVL @>ADJ @ADVL **@CLB** @FMV @PP_ADVL
'He answered me politely and left.'

7.4.10.7 NP Fragments

We tag nouns which are not functioning as subject or object etc. as @NP. (244)-(246) illustrate some uses of the tag @NP.

Vocative Case

(244) A mhná uaisle
PRT women-VOC noble
@>N @NP @N<
'O noble women'

Apposition

(245) an duine uasal seo Marstrander
an person noble this Marstrander
@>N @SUBJ @N< @N< @NP
this noble person Marstrander

Lists

```
(246) bialann, siopa, srl.
      @NP,     @NP,   @NP
      restaurant, shop, etc.
```

7.5 Implementation

One particular difference between our implementation of dependency tagging and that described in Karlsson *et al* (1995), is that we do not introduce ambiguity at the dependency annotation level. In Karlsson *et al* (1995), if a verb could be either a main verb or an auxiliary, both tags are appended. Likewise if a noun could be either a subject or an object, both tags are appended. Select and Remove rules are then used (as in POS tagging) to eliminate ambiguity where possible. We have chosen instead to only ever apply one dependency tag per token, (using our detailed morphosyntactic information), and we thereby avoid having to disambiguate dependency tags.

7.5.1 Automatic Dependency Analysis

In order to determine the structure of a sentence, we have developed the following divide-and-conquer approach to dependency annotation, Figure 49.

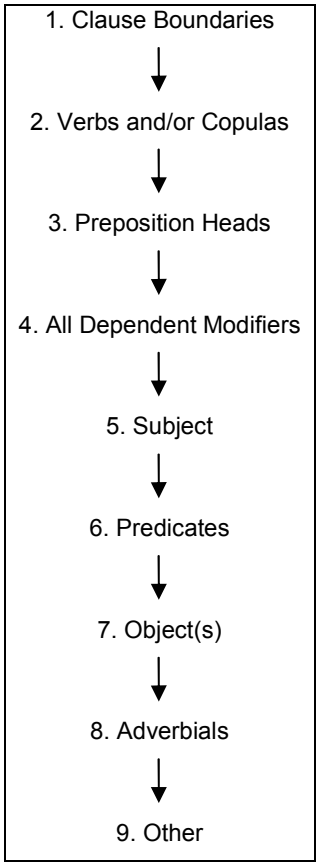


Figure 49 Dependency Analysis Flowchart

7.5.1.1 Clause Boundaries

Firstly, we label the clause boundaries, as these will limit the search space for identifying subsequent grammatical functions, e.g. verb, subject and object, and their dependants.

7.5.1.2 Verbs and/or Copulas

Next, within the clause, we locate the verb or copula. In the case of verbs, we determine whether it is being used as a finite, auxiliary or relative. This will have a bearing on the location of the subject and object. We also at this stage mark cases where the verb and subject are combined in a single word form.

7.5.1.3 Preposition Heads

We next identify prepositional phrases. We do this at this stage, as it rules out a number of NPs from being a subject or direct object. Several types of prepositional heads are distinguished, e.g. adverbial, aspectual, etc. We do not attempt to distinguish between certain types of adverbial phrase (in brackets below), as we are unable to do so without additional subcategorization information, e.g.

- *i mbosca le* 'in a box with' (a locative PP)
- *i gcomhairle le* 'in cnsultation with' (idiomatic PP)

and

- *Bhí sé ina theach* 'He was in his house' (locative PP - common noun)
- *Bhí sé ina rí* 'He was a king' (stative PP - animate human noun)
- *Bhí sé ina thost* 'He was silent' (stative PP - abstract noun)

7.5.1.4 All Dependent Modifiers

Following PPs, we mark up modifiers of nouns and verbs. In the case of verbs, we have preverbal particles, and prepositions which are part of phrasal verbs. Nouns may be modified by prenominal modifiers (determiners and numerals), postnominal modifiers (adjectives and demonstratives) or another NP, i.e. possessive NPs.

7.5.1.5 Subjects and Predicates

We are now in a position to try to identify the subject. In the case of finite main or auxiliary non-relative verbs (which do not incorporate a subject), this will normally be the first NP following the verb. We also label predicates of a copula or substantive verb.

7.5.1.6 Objects, Adverbials and Other

Using transitivity information on verbs, we attempt to locate direct and indirect objects. Finally adverbials are marked up, and any remaining noun phrases are tagged as NPs (e.g. lists or appositions).

7.5.2 Constraint Grammar Dependency Annotation Rules

Over 250 CG dependency and grammatical function rules have been developed to date, in order to annotate Irish sentences with grammatical function and dependency tags. The CG `MAP` statement is used to append dependency tags to the already morphosyntactically annotated tokens. The general format of the `MAP` statement is as follows:

```
MAP (@TAG) TARGET (POS) IF (CONDITION(S));
```

The grammatical function or dependency tag to be applied is specified following the `MAP` keyword. This is followed by the keyword `TARGET`, and the token type to which the tag should be applied. Finally, one or more conditions can optionally be specified using the keyword `IF`.

```
MAP (@SUBJ) TARGET (Pron) IF (*-1 (@FMV) BARRIER NOUN-OR-PRO);
```

In the CG `MAP` statement above, a pronoun will receive the `@SUBJ` tag if the specified condition is fulfilled. In this case, the tag should only be applied if there is a finite main verb, `@FMV`, somewhere to the left, using `*-1`. By using the `BARRIER` keyword we ensure that searching to the left stops if a noun or pronoun is encountered, before we encounter `@FMV`. The term `NOUN-OR-PRO` is a user-defined term, which can be defined as follows, using the `LIST` statement:

```
LIST NOUN-OR-PRO = (Noun) (Pron Pers) (Pron Dem) (Pron Idf);
```

A full listing of the CG mapping rules may be found in Appendix F. We present some illustrative examples of CG mapping rules in the following subsections.

7.5.2.1 Clause Boundaries

We consider a clause to be a verb or copula and its arguments. Finite complements and coordinated independent sentences are marked with a clause boundary. We do not insert a clause boundary for relative clauses, as their arguments may be distributed over the main and relative clauses.

As shown in the code snippet in Figure 50, a clause boundary tag `(@CLB)` is appended (using the `MAP` statement), to a token's existing list of morphosyntactic tags, if the token is a

co-ordinating conjunction followed by a non-relative verb-form (247), a subordinating conjunction (248), or a dependent (subordinate) form of the copula (249).

(247) Cheannaigh Seán leabhar agus léigh sé é
Bought Seán book and read he it
@FMV @SUBJ @OBJ @CLB @FMV @SUBJ @OBJ
'Seán bought a book and he read it'

(248) Dúirt sé go rachadh sé.
Said he that would-go he
@FMV @SUBJ @CLB @FMV @SUBJ
'He said that he would go'

(249) Dúirt sé gur múinteoir é
Said he COP teacher he?
@FMV @SUBJ @CLB @PRED @SUBJ
'He said that he is a teacher'

```
# Part 1 - Clause Boundaries
# ========================================================== #
SETS
LIST PUNCT = (":");
# ========================================================== #
MAPPINGS
MAP (@CLB) TARGET (Cop Dep);            # Dúirt sé [gur] Seán
MAP (@CLB) TARGET (Conj Subord);        # e.g. nuair
MAP (@CLB) TARGET (Conj Coord) IF (1 (Verb)); # [agus] bhí
MAP (@CLB) TARGET (Conj Coord) IF       # [agus] is léir;
    (1 (Cop Pres) OR (Cop Past) OR (Cop Pron) OR (Cop Q));
MAP (@CLB) TARGET (Conj Coord) IF       # . [agus] ná déan siúd
    (1 (Part Vb)) (NOT 1 (Part Vb Rel)) (2 (Verb));
MAP (@CLB) TARGET PUNCT;                # e.g. [:] Ar an maidin
```

Figure 50 Dependency Annotation: Clause Boundaries

7.5.2.2 Verbs and Copulas

The code snippet in Figure 51 illustrates how some finite main verbs are labelled. As illustrated, we have defined some sets which are subsequently used in the rules. In the example below we define synthetic verbs (`VSYNTH`) as those having the morphological tags `Verb` and person features (`1P`, `2P`, `3P`, `Auto`). Similarly, auxiliary verbs (`AUX`) are defined by listing the lemmas which can function as auxiliaries. Set members can be defined using POS tags, lemmas or word forms, or any combination of the three.

```
# Non-Relative Finite Main Verbs
#   Analytic (@FMV), Synthetic (@FMV_SUBJ),
# ========================================================== #
SETS
LIST VSYNTH = (Verb 1P) (Verb 2P) (Verb 3P) (Verb Auto) ;
LIST AUX = ("bí") ("téigh") ("tosaigh") ("tosnaigh") ("féad")
("caith") ("féach");
LIST RELPART = (Vb Rel) (Prep Rel) ;
MAPPINGS
MAP (@FMV) TARGET (Verb) IF       # e.g. Chuaigh an bhean amach
    (NOT 0 VSYNTH OR AUX)
    (NOT -1 RELPART)
    (NOT -2 RELPART);
MAP (@FMV_SUBJ) TARGET (Verb) IF  # e.g. Chuamar amach
    (0 VSYNTH )
    (NOT 0 AUX)
    (NOT -1 RELPART) ;
```

Figure 51 Dependency Annotation: Finite Main Verbs

7.5.2.3 Prepositional Phrases

In Figure 52, we give some rules for tagging prepositional phrases which are functioning aspectually with the verbal noun.

```
# PP - STATIVE
# ========================================================== #
MAP (@PP_STAT) TARGET (Prep Simp) IF # ar oscailt, open
    (0 ("ar"))
    (1 (Verbal Noun));
# ========================================================== #
# PP - ASPECTUAL
# ========================================================== #
MAP (@PP_ASP) TARGET (Prep Simp) IF # ag gearradh, cutting
    (NOT 0 ("ar"))
    (1 (Verbal Noun));
MAP (@PP_ASP) TARGET (Prep Simp) IF # do mo ghearradh, cutting
me
    (1 (Det Poss))
    (2 (Verbal Noun));
```

Figure 52 Dependency Annotation: Prepositional Phrases

7.5.2.4 Dependent Modifiers

Figure 53 shows some of the rules which are used to map noun premodifiers (`@>N`), noun postmodifiers (`@N<`), and verbal nouns dependent on prepositional aspectual heads (`@P<`).

```
MAP (@>N) TARGET (Part Voc);
MAP (@>N) TARGET (Det);
MAP (@N<) TARGET (Num Dig) IF (-1 ("Euro") OR ("euro"));
MAP (@P<) TARGET (Verbal Noun) IF # á dhéanamh, tar éis dul
  (-1 (Prep Simp) OR (Prep Poss) OR (Prep Cmpd) OR (Det Poss));
```

Figure 53 Dependency Annotation: Dependent Modifiers

7.5.2.5 Subjects

In general, clauses contain at most one subject, other than comma separated lists and conjoined subjects. In main declarative clauses, the subject is the first NP after the verb, if it is not a synthetic verb form, (i.e. already includes a subject). We show an example of this type of rule in Figure 54.

```
# SUBJECT of FMV
# ========================================================== #
SETS
LIST NOUN-OR-PRO = (Noun) (Pron Pers) (Pron Dem) (Pron Idf);
LIST NOUN-NOM = (Noun Com) (Subst Noun) (Prop Noun) (Abr)
(Unk);
# ========================================================== #
MAPPINGS
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FMV) BARRIER NOUN-OR-
PRO);
```

Figure 54 Dependency Annotation: Subjects 1

In Figure 55, in progressive aspectual clauses, the subject will be the NP preceding the aspectual preposition (250).

```
(250) ... agus é          ag      caitheamh airgid
      ... and  he         at      spending  money
          @CC  @SUBJ_ASP @PP_ASP @P<       @OBJ_ASP
      '... and he spending money'
```

```
MAP (@SUBJ_ASP) TARGET NOUN-OR-PRO IF
    (1 ("ag" Prep Simp))
    (2 (Verbal Noun));
```

Figure 55 Dependency Annotation: Subjects 2

In the case of transitive infinitives, the object will be the NP preceding the infinitival particle. However, with intransitive infinitives (i.e. a verbal noun without an infinitival particle), the subject immediately precedes the verbal noun (251).

```
(251) ... ar   mhian leo       caitheamh anuas
      ... on   wish  with-them throw     down
          @COP @PRED @PP_SUBJ  @INF      @ADVL
      '... they who wished to criticise'
```

Some irregular verbs, including the substantive verb *bí* 'to be', which although intransitive, optionally occur with an infinitival particle. These are handled in the code in Figure 56.

```
# eagla a bheith orthu, aonad a bheith againn
MAP (@SUBJ_INF) TARGET NOUN-NOT-VN IF
  (NOT 0 (Noun Gen))
  (*1 (Part Inf) BARRIER NOUN-OR-PRO LINK 1 (Verbal Noun VI) );
```

Figure 56 Dependency Annotation: Subjects 3

If a relative verb is followed by a possessive determiner and a noun, then the subject precedes the relative verb, as in (252). This is implemented in the code snippet in Figure 57.

```
(252) ... an  fear  a    bhfuil    a   mhac      ag      imeacht
      ... the man   that is        his son       at      leaving
          @>N @SUBJ @>V  @FAUX_REL @>N @SUBJ_ASP @PP_ASP @P<
      '... the man whose son is leaving'
```

```
# an fear a bhfuil a mhac ag imeacht
MAP (@SUBJ) TARGET NOUN-OR-PRO IF
    (NOT 0 (Cop))
    (*1 (Part Vb Rel) LINK 1 (Verb) LINK 1 (Det Poss));
```

Figure 57 Dependency Annotation: Subjects 4

7.5.2.6 Objects

In main declarative clauses, the direct object is the second NP after the verb or the first NP if the subject is combined with the verb, as shown in the code snippet in Figure 58.

```
# rinneamar é
MAP (@OBJ) TARGET (Pron Pers) IF (-1 VSYNTH);
# ná déan seo agus ná déan siúd
MAP (@OBJ) TARGET (Pron Dem) IF (-1 VSYNTH);
```

Figure 58 Dependency Annotation: Objects 1

With transitive infinitives, as mentioned above in relation to (251), the direct object always precedes the particle *a* and the infinitive.

```
(253) d'iarr mé    ar       an  bhfear an  doras    a   dhúnadh
      asked  I     on       the man    the door     to  close
      @FMV   @SUBJ @PP_ADVL @>N @P<    @>N @OBJ_INF @>N @INF
      'I asked the man to close the door'
```

```
LIST NOUN-NOT-VN = (Noun Sg) (Noun Pl) (Abr) (Unknown);
LIST TRANSVN = (Verbal VT) (Verbal VTI) (Verbal VD) ;
MAP (@OBJ_INF) TARGET NOUN-NOT-VN IF
 (NOT 0 (Noun Gen))
 (*1 (Part Inf) BARRIER (Noun) OR (Pron Pers) LINK 1 TRANSVN );
```

Figure 59 Dependency Annotation: Objects 2

In simple sentences, where a verb is marked as ditransitive (VD), then the first prepositional phrase after the verb usually contains the indirect object, as shown in Figure 60.

```
# "Thug sé an leabar do Mháire"
MAP (@PP_OBL) TARGET (Prep Simp) IF
    (*-1 (VD) BARRIER (Prep Simp));
    (*1 NOUN-NOT-VN BARRIER (Noun) or (Verb) OR (Cop)):
```

Figure 60 Dependency Annotation: Objects 3

7.5.2.7 Predicates

In Figure 61, we have a rule which tags adjectives as predicates if the are not attributive adjectives (i.e. not inflected for agreement with the noun), and they occur with the substantive verb *bí* 'to be'.

```
LIST ADJ-ATTR = (Adj Sg) (Adj Pl) (Adj Len) (Adj Ecl)
# Bhíomar tinn inné
MAP (@PRED) TARGET (Adj) IF
    (NOT 0 ADJ-ATTR)
    (-1 ("bí") BARRIER (@CLB));
```

Figure 61 Dependency Annotation: Predicates

7.5.2.8 Time Adverbials

In example (254), the fronted NP *An lá* 'the day' is functioning as a temporal adverbial clause. We have created a set called `TIME` which lists lemmas such as *mí* 'month', *bliain* 'year', *lá* 'day', which can occur in temporal adverbial adjuncts, as shown in Figure 62.

```
(254) An  lá    a    cuireadh      Butt ...
      The day   that put-AUTO      Butt ...
      @>N @ADVL @>V  @FMV_REL_SUBJ @OBJ
      'The day that Butt was buried ...
```

```
# TIME ADVERBIAL
# ========================================================== #
LIST TIME-PERIOD = "mí" "bliain" "lá" "ráithe" "uair"
"seachtain";
LIST TIME = "inné" "inniú" "amárach" "arú" "anocht" "aréir"
"istíoche" "tráthnóna" "ardtráthnóna" "Dé" "Déardaoin";
MAP (@ADVL) TARGET TIME;
```

Figure 62 Dependency Annotation: Temporal Adverbials

7.5.2.9 Other Nouns

If a noun or other nominal item has not already been tagged, it will now be tagged with the general purpose `@NP` tag, as shown in Figure 63.

```
MAP (@NP) TARGET (Pron Pers); # Iad/NP uile faoi shuan ..
MAP (@NP) TARGET (Abr) IF (NOT -1 (Prop)) (NOT 1 (Prop));
```

Figure 63 Dependency Annotation: Other Nouns

7.6 Evaluation

Test Suite Results for Dependency Analysis

Our first set of evaluation results for Dependency Analysis is based on the 225 made-up Test Suite Sentences. We calculate the precision of our automatic dependency tagging against Gold Standard dependency tagged Test Suite Sentences. The results are as follows:

Overall Precision (Test Suite): $\frac{CorrectAutoTags}{TotalAutoTags} \times \frac{100}{1} = \frac{1{,}212}{1{,}241} \times \frac{100}{1} = 97.66\%$

Overall Recall (Test Suite) : $\frac{CorrectAutoTags}{TotalGoldTags} \times \frac{100}{1} = \frac{1{,}212}{1{,}241} \times \frac{100}{1} = 97.66\%$

Overall F-score (Test Suite) : $\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{97.66 \times 97.66 \times 2}{97.66 + 97.66} = 97.66\%$

As the number of automatically tagged tokens equals the number of Gold tagged tokens (i.e. each token has one and only one tag), precision, recall and f-score have the same value. The f-score, 97.66%, is high due to the fact that the Test Suite contains only short, grammatical sentences (the longest sentence has 20 tokens, excluding punctuation).

Development and Test Set: Overall Results for Dependency Analysis

In order to assess performance on real-world data we use a Gold Standard (250) Corpus randomly extracted from the larger Gold Standard (3,000) Corpus (see Chapter 3 for details). These 250 sentences consist of 150 Development Set sentences and 100 Test Set sentences.

In Table 34, we present details of the overall precision of Dependency Analysis tagging, based on the automatic dependency tagging of the Development Set (150 sentences) and Test Set (100 sentences). As with the Test Suite, precision, recall and f-score have the same value, as the number of automatic tags equals the number of gold tags.

The overall f-score, for the 150 Development Set sentences is 93.60%, and for the Test Set sentences is 94.28%, as presented in Table 34.

Table 34 Dependency Annotation: Overall Evaluation Results

Gold Standard Development Set (150 Sentences)						
Tot Tokens	*Punct. Tokens*	*Tokens*	*Correct*	*Incorrect*	*% Precision*	*F-Score*
4403	444	3959	3706	253	**93.60**	**93.60**
Gold Standard Test Set (100 Sentences)						
Tot Tokens	*Punct. Tokens*	*Tokens*	*Correct*	*Incorrect*	*% Precision*	*F-Score*
2555	282	2273	2143	130	**94.28**	**94.28**

Development: Detailed Results for Dependency Analysis

We also generate precision, recall and f-score analysis for each of the individual grammatical function and dependency tags in the 150 sentence Development Set, as shown in Table 35. For example, the precision, recall and f-score for adverbial prepositions is as calculated below:

Precision (PP_ADVL): $\frac{CorrectAutoPP\_ADVL}{TotalAutoPP\_ADVL} \times \frac{100}{1} = \frac{539}{548} \times \frac{100}{1} = 98.36\%$

Recall (PP_ADVL): $\frac{CorrectAutoPP\_ADVL}{GoldPP\_ADVL} \times \frac{100}{1} = \frac{539}{545} \times \frac{100}{1} = 98.90\%$

F-score (PP_ADVL) : $\frac{Recall \times Precision \times 2}{Recall + Precision} = \frac{98.90 \times 98.36 \times 2}{98.90 + 98.36} = 98.63\%$

The F-score for `@PP_ADVL` is 98.63%, and the weighted F-score is 13.58%, as calculated below.

W F-score (PP_ADVL) : $F - Score \times \frac{GoldPP\_ADVL}{TotalGoldTags} = 98.63 \times \frac{545}{3{,}959} \times \frac{100}{1} = 13.58\%$

In Table 35, the results are ordered according to the frequency with which the various dependency tags occur in the data. We begin with the most common tag `@PP_ADVL` (occurring 545 times), followed by nouns dependent on prepositions `@P<` (529 occurrences). There are fewer dependent NPs as prepositional pronouns (prepositions inflected for person) are tagged as `@PP_ADVL`. Following this, we have noun pre-modifiers (`@>N`) and nouns post-modifiers (`@N<`). (Note that the sum of the weighted f-scores (93.65%) differs slightly from the previously calculated overall f-score (93.60%) due to rounding errors).

Table 35 Development Set (150): Dependency Annotation Results

Tag	Gold	Auto	Correct	Precis.	Recall	F-Score	WF-Score
PP_ADVL	545	548	539	98.36	98.90	98.63	13.58
P<	529	506	493	97.43	93.19	95.27	12.73
>N	497	500	494	98.80	99.40	99.10	12.44
N<	484	463	455	98.27	94.01	96.09	11.75
FMV	267	269	256	95.17	95.88	95.52	6.44
SUBJ	226	215	194	90.23	85.84	87.98	5.02
NP	200	234	164	70.09	82.00	75.58	3.82
>V	153	152	151	99.34	98.69	99.02	3.82
CC	149	146	146	100.00	97.99	98.98	3.72
CLB	135	138	134	97.10	99.26	98.17	3.35
INF	105	114	105	92.11	100.00	95.89	2.54
ADVL	90	77	75	97.40	83.33	89.82	2.04
PRED	88	87	74	85.06	84.09	84.57	1.88
OBJ	86	80	65	81.25	75.58	78.31	1.70
PP_ASP	70	67	65	97.01	92.86	94.89	1.68
FAUX	69	63	59	93.65	85.51	89.39	1.56
OBJ_INF	55	60	47	78.33	85.45	81.74	1.14
COP	45	46	45	97.83	100.00	98.90	1.13
>ADJ	32	32	32	100.00	100.00	100.00	0.81
PP_HAS	21	18	18	100.00	85.71	92.31	0.49
SUBJ/OBJ	19	18	16	88.89	84.21	86.49	0.42
PN<	14	13	13	100.00	92.86	96.30	0.34
AUG>SUBJ	12	11	11	100.00	91.67	95.65	0.29
SUBJ_REL	11	15	4	26.67	36.36	30.77	0.09
SUBJ_ASP	11	12	10	83.33	90.91	86.96	0.24
PP_STAT	10	13	10	76.92	100.00	86.96	0.22
PP_SUBJ	7	9	7	77.78	100.00	87.50	0.16
PP_NEG	5	5	5	100.00	100.00	100.00	0.13
OBJ_ASP	5	9	4	44.44	80.00	57.14	0.07
COP_W	3	3	3	100.00	100.00	100.00	0.08
PRED<	1	0	0	0.00	0.00	0.00	0.00
							93.65

In general, the tagging of unlabelled dependency relations is reasonably straightforward. Tagging of grammatical functions is more problematic. One of the main difficulties is ambiguity regarding subjects and objects, particularly in relative verb constructions. In order to resolve many of these constructions, additional semantic and pragmatic information would be necessary.

As is evident from the confusion matrix in Table 36, the most common tagging error is a dependent noun (N<) in the Gold Standard which has been tagged as an NP in the automatic annotation (20 occurrences). In the main, this is due to proper nouns in titles, particularly English titles, where the words have no case marking, or are marked as Foreign at the POS tag level. If one proper noun directly follows another, we can assume that the second is dependent on the first, although, this can cause problems in the following type of sentence, where, in fact, there are two separate NPs (255).

(255) Chuir **O' Neill Lennon** isteach san aicsean
Put O' Neill Lennon into in-the action
@FMV **@SUBJ** **@OBJ** @ADVL @PP_ADVL @P<
'O' Neill put Lennon into the action'

Table 36 Dependency Annotation Confusion Matrix

Tag	N<	NP	SUBJ	P<	OBJ	PRED	OBJ_INF
N<		**20**	5			6	
NP			6		**10**	2	2
SUBJ							
P<		9					**11**
OBJ		3	5				5
PRED	6						
OBJ_INF		8		4			

The second most frequent problem (11 occurrences) relates to NPs which can either be dependent on the following infinitive, @OBJ_INF, or on the preceding preposition @P<. In (256), we have *chluiche ceannais a bhuachan* 'the final game to win', whereas in (257) *an Aire a bheith riachtanach* 'the Minister to be necessary' would be incorrect as it is the 'directives' and not the Minister which is necessary

(256) ag na foirne sin ar **chluiche ceannais** a bhuachan
at the teams those on **game** **final** to win
@PP_ADVL @>N @P< @N< @PP_ADVL **@OBJ_INF** @N< @>N @INF
'those teams ... at winning the final'

```
(257) forálacha  is   dóigh    leis     an  Aire      a   bheith riachtanach
      directives COP  consider with    the Minister  to  be     necessary
      @NP        @COP @PRED    @PP_ADVL @>N @P<      @>N @INF   @PRED
      'directives which the Minister considers to be necessary'
```

The third most frequent problem (10 occurrences) which we will highlight is where NPs `@NP` are incorrectly tagged as objects `@OBJ`. Many verbs are tagged `VTI` meaning they can function transitively or intransitively, resulting in some cases with an available bare NP being tagged as an object when, in fact, the verb is being used intransitively.

7.7 Summary

In this chapter, we introduce Dependency Analysis for Irish. We describe in detail the tagset used to tag grammatical functions and unlabelled dependency relations. We present the main syntactic structures for Irish using sentence templates and examples of each type of structure.

The dependency analysis is shallow and partial, as it does not cover co-ordination, long-distance dependencies and prepositional and clausal attachments are not resolved. The result is a single deterministic analysis.

In the implementation section, we describe the order in which dependency annotation rules are applied and give illustrative examples of each type of rule. The Dependency Analysis for all of the examples in this chapter are given in Appendix E.

Finally, we evaluate the automatic tagging using Test Suite sentences and Gold Standard data. The f-score for the Development Set data is 93.60% and for the Test Set data is 94.28%.

These results can be improved by extending the dependency tagging rules, as well as enhancing the finite-state lexicons by adding verb subcategorization information and semantic properties of nouns (animate, inanimate, human, animal, abstract etc.). Upgrading from the CG2 version of Constraint Grammar, currently used, to CG3 will allow for greater modularisation through the use of templates. This will allow us to combine several rules into one and, thereby, reducing the chance of accidental errors and omissions, (e.g. changing a rule relating to nouns and omitting to make a similar change in rules involving pronouns etc.).

In the next chapter we describe chunking, the final stage of linguistic annotation in our current implementation of partial parsing.

8 Chunking

8.1 Introduction

As mentioned in Chapter 2, dependency mark-up does not contain any phrasal nodes, i.e. all mark-up is attached to individual tokens (terminal nodes). However, for linguistic analysis and NLP applications both constituency based and functional annotation are necessary. Most recently constructed treebanks use a combination of both types of mark-up.

Consequently, we implement chunking of the dependency marked-up text using finite-state transducers compiled from regular expressions using Xerox finite-state tools. This bracketing overlays the dependency marked-up data. For example, in order to decide where one noun phrase ends and the next begins, e.g. the subject and object (in VSO word order) we make use of the dependency and functional tags. We use the longest-match operator to bracket the maximum length noun phrases, taking into account case marking.

While we implement several levels of nesting, we do not include prepositional phrase attachment or resolve co-ordinated items. There is no recursion, i.e. no chunk contains a chunk of the same type as itself, or a higher- level phrase, i.e a level 2 chunks contain level 1 chunks, but not vice versa, see Table 37. To facilitate the implementation of nesting using regular expressions, chunk labels have matching end brackets. Example (259) shows the chunked representation of (258), where an NP is nested in a PP. Note that our definition of NP includes adjectival modifiers as shown in (259).

```
(258) den    chuid  is  mó
      of-the part   PRT most
      'for the most part'

(259) [PP den de+Prep+Art+Sg+@PP_ADVL
        [NP chuid cuid+Noun+Fem+Com+Sg+Def+@P<
            is is+Part+Sup+@>ADJ mó mór+Adj+Comp+@N< NP] PP]
```

In Section 8.2, we describe our annotation scheme for labelling chunks and we define the levels of nesting which we currently implement. In Section 8.3, we present the implementation of the Finite State Chunker using regular expressions and Xerox Finite-State Tools. Finally in Section 8.4, we give details of our evaluation results and error analysis.

8.2 Annotation Scheme for Nested Chunking

In Table 37, we list the chunk labels we use in our annotation, together with an example of each. In a chunk label "`..`" represents text. Please note that chunks which currently are not

nested within higher level chunks (other than overall sentence brackets `[S .. S]`), end with an unlabelled end bracket, i.e. `[V .. ]`, as opposed to `[NP .. NP]` which can be nested within higher level chunks. As Table 37 shows, we separate the chunks according to their level of nesting.

Table 37 Bracketed Chunk Labels

Nesting Level	Chunk Type	Chunk Label	Example
1	Verb	`[V .. ]`	[V *Labhair*] *Seán*, 'Seán spoke'
	Verb+Subj	`[VS .. ]`	[VS *Labhaíomar*], 'We spoke'
	Copula	`[COP ..]`	[COP *Is*] *maith liom*, 'I like' i.e. 'Is good with me'
	Adverbial	`[AD .. ]`	[AD *amárach*] 'tomorrow'
	Predicate	`[PRED .. ]`	*Tá sé* [PRED *mór*], 'He/It is big'
	Noun	`[NP .. NP]`	[NP *teorainn an cheantair* NP] 'border of the region'
	Obj of Asp	`[OA .. OA]`	*ag déanamh* [OA *cáca* OA], 'making a cake'
	Infinitive	`[I ..I]`	*cáca* [I *a dhéanamh* I] 'to make a cake'
	Obj of Inf	`[OI .. OI]`	[OI *cáca* OI] *a dhéanamh* 'to make a cake'
	Prep.	`[PP .. PP]`	[PP *liom* PP] 'with me'
2	Prep.	`[PP .. [NP] PP]`	[PP *ins* [NP *an siopa* NP] PP] 'in the shop'
	Asp. Prep.	`[PP-ASP .. [NP] PP-ASP]`	[PP-ASP *ag* [NP *déanamh* NP] PP] 'making'
	Infinitival Phrase	`[INF (..) ([OI]) [I] INF]`	[INF *gan* [OI *cáca* OI] [I *a dhéanamh* I] INF] 'not to make a cake'
3	Aspectual Prep. Phr	`[ASP [PP-ASP] [OA] ASP]`	[ASP [PP-ASP *ag* [NP *déanamh* NP] PP-ASP] [OA *cáca* OA] ASP] 'making a cake'
	Aspectual Infinitival	`[ASP [PP] ([OI]) [INF] ASP]`	[ASP *chun* [INF [OI *cáca* OI] [I *a dhéanamh* I] INF] ASP] 'to make a cake'
4	Conjoint	`[CJ2 .. [?] CJ2]`	*úlla* [CJ2 *agus* [NP *oráistí* NP] CJ2] 'apples and oranges'
5	Sentence	`[S [?]+ S]`	[S [VS *Labhaíomar*] S],

In order to bracket the chunks shown in Table 37, we use the dependency labels attached to tokens. In Table 38, we have grouped the dependency labels from Table 33, according to the chunks to which they can belong. In general, a chunk will have only one chunk head from the list of possible heads shown, i.e. a verb chunk head can have any one of the eight grammatical labels listed as a verb head in the table. Chunks may have zero or more of the associated pre- and post-modifiers. The finite state regular expressions which define the chunks are the subject of the Section 8.3.

Table 38 Chunk Dependency Tags

Chunk Type	Tag type	Tag	Description
Verb	PreMod	@>V	pre-verbal particle dependent on a verb to the right
	Head	@FAUX	finite auxiliary verb
		@FAUX_REL	relative finite auxiliary verb
		@FAUX_SUBJ	finite auxiliary verb including subject
		@FMV	finite main verb
		@FMV_REL	relative finite main verb
		@FAUX_REL_SUBJ	relative finite auxiliary verb including subject
		@FMV_REL_SUBJ	relative finite main verb incl. subject
		@FMV_SUBJ	finite main verb including subject
Noun	PreMod	@>N	pre-modifier dependent on the first noun to the right
		@AUG>SUBJ	augment pronoun dependent on subj. to the right
	Head	@INF	bare infinitive clause
		@NP	noun phrase; in list, in apposition, or fragment
		@OBJ	object
		@OBJ_ASP	object of aspectual clause
		@OBJ_INF	object of infinitive clause
		@SUBJ	subject
		@SUBJ_INF	subject of infinitive (intrans)
		@SUBJ_OR_OBJ	subject or obj. of relative clause
		@SUBJ_ASP	subject of. aspectual phrase, e.g.progres., stative
		@SUBJ_REL	subject of relative clause
		@P<	noun dependent on the preceding prep.
		@PC<	noun dependent on comp'd prep.
	PostMod	@N<	noun post-modifier
		@PN<	pronoun post-mod.

Copula	Head	@COP	copula
		@COP_WH	interrogative pronoun + copula
		@COP_SUBJ	copula including subject
Pred.	Head	@PRED	predicate
	PostMod	@PRED<	dependent on predicate
Prep.	Head	@PP_ADVL	adverbial prepositional phrase
		@PP_NEG	negative marker
		@PP_OBL	oblique prepositional phrase
		@PP_PRED	predicative
		@PP_STAT	stative
		@PP_ASP	aspectual prepositional phrase
		@PP_HAS	has prepositional phrase
		@PP_SUBJ	prep + subj pronoun
	PostMod	@P<	pronoun post-mod.
		@PN<	pronoun post-mod.
Adverbial	PreMod	@>ADJ	adverbial particle dependent on the adjective to the right
	Head	@ADVL	adverbial
	PostMod	@ADVL<	dependent on an adverbial
Conjunct.	Head	@CC	co-ordinating conjunction
		@CS	subordinating conjunction

8.3 Implementation of the Finite-State Chunker

The finite-state chunker is run on the dependency annotated data. For example, the output from POS tagging and dependency mark-up for the sentence fragment *Bhéimnigh sé freisin an t-easpa seirbhísí ar nós HEMS agus an bealach...*'He emphasised also the lack of services such as HEMS and the way ...' is as follows:

(260)

```
"<Bhéimnigh>"     "béimnigh" Verb PastInd Len @FMV    Emphasised
"<sé>"            "sé" Pron Pers 3P Sg Masc Sbj @SUBJ       he
"<freisin>"       "freisin" Adv Gn @ADVL                    also
"<an>"            "an" Art Sg Def @>N                       the
"<t-easpa>"       "easpa" Noun Fem Com Sg @OBJ              lack
"<seirbhísí>"     "seirbhís" Noun Fem Gen Strong Pl @N<     services
"<ar nós>"        "ar nós" Prep Cmpd @PP_ADVL               such_as
"<HEMS>"          "HEMS" Guess Abr @P<                      HEMS
"<agus>"          "agus" Conj Coord @CC                     and
"<an>"            "an" Art Sg Def @>N                       the
"<bealach>"       "bealach" Noun Masc Com Sg Def @NP        way
...                                                         ...
```

This is converted to sentence-per-line format, as shown in (261), with each token followed by its tag string. A tag string consists of the lemma, morphological tags and functional or dependency tag, e.g. "`Bhéimnigh béimnigh+Verb+PastInd+Len+@FMV`" represents a token and tagstring pair.

(261)

```
Bhéimnigh béimnigh+Verb+PastInd+Len+@FMV sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ freisin freisin+Adv+Gn+@ADVL an
an+Art+Sg+Def+@>N t-easpa easpa+Noun+Fem+Com+Sg+@OBJ seirbhísí
seirbhís+Noun+Fem+Gen+Strong+Pl+@N< ar_nós ar+nós+Prep+Cmpd+@PP_ADVL
HEMS HEMS+Guess+Abr+@P< agus agus+Conj+Coord+@CC an
an+Art+Sg+Def+@>N bealach bealach+Noun+Masc+Com+Sg+Def+@NP
```

Using regular expressions and *xfst* we insert chunk boundaries using the longest match operator. Phrases are bracketed as shown below. (Note that the morphological and dependency tags have been removed from this example for readability).

(262)

```
[S [V Bhéimnigh ] [NP sé NP] [AD freisin ] [NP an t-easpa seirbhísí
NP] [PP ar_nós [NP HEMS NP] PP] [CJ2 agus [NP an bealach NP]] ...
```

The above sentence illustrates the difficulty associated with selecting the correct chunks to associate with a conjunction.

Presently, we use the `[CJ2 .. CJ2]` labels to associate a conjunction with the following chunk (which may have embedded chunks). Frequently, this is correct, as in the example given in Table 37, *úlla agus oráistí* 'apples and oranges' which is bracketed as `[NP úlla NP] [CJ2 agus [NP oráistí NP] CJ2]`. In this case it would be straightforward to add another level of bracketing to associate the two parts of the conjoined phrase as follows:

`[CONJ [NP úlla NP] [CJ2 agus [NP oráistí NP] CJ2] CONJ]`. However, because of the difficulties inherent in bracketing the correct conjoined elements in many sentences, including our example sentence, *an t-easpa seirbhísí ar nós HEMS agus an bealach...*'the lack of services such as HEMS and the way ...' `[CONJ [NP an t-easpa seirbhísí NP] [PP ar_nós [NP HEMS NP] PP] [CJ2 agus [NP an bealach NP]] ... CONJ]`, we have not included a full implementation of conjunctions in this chunker.

As Irish is a VSO language, the subject and object NPs are usually adjacent, e.g. *Chuir gach imreoir fáilte roimh...* 'Every player welcomed...' (263). This makes it difficult to chunk without adequate case marking. However, the detailed dependency and functional tags attached to each token, make the actual bracketing, which would otherwise be a very difficult task, an almost trivial exercise.

```
(263) [VP Chuir] [NP gach imreoir NP] [NP fáilte NP] [PP roimh

"<Chuir>"          "cuir" Verb PastInd Len @FMV        Put
"<gach>"           "gach" Det Qty @>N                  every
"<imreoir>"        "imreoir" Noun Masc Com Sg @SUBJ    player
"<fáilte>"         "fáilte" Noun Fem Com Sg @OBJ       welcome
"<roimh>"          "roimh" Prep Simp @PP_ADVL          before
```

In the following code snippets, we show the regular expressions which are used to implement the `V` and `VS` chunks.

In Figure 64, we begin by defining the alphabets used for tokens and lemmas (`Alpha`), for morphological tags (`MAlpha`), and for dependency tags (`DAlpha`). We also define a token and lemma string (`TokLem`) and a string of morphological tags (`MTag`). These are combined to define a general purpose token-lemma-morphtags definition (`TokLemMTag`) which is the used in all of the subsequent chunk definitions.

```
# Input format "token lemma+MTags+@DTag token lemma+MTags+@DTag
############################################################
# Alphabet used for tokens and lemmas
define Alpha
[a|á|b|c|d|e|é|f|g|h|i|í|j|k|l|m|n|o|ó|p|q|r|s|t|u|ú|v|w|x|y|z|
A|Á|B|C|D|E|É|F|G|H|I|Í|J|K|L|M|N|O|Ó|P|Q|R|S|T|U|Ú|V|W|X|Y|Z|1
|2|3|4|5|6|7|8|9|%0|%.|%,|%-|%+|%*|%/|%>|%<|%?|%:|'|''|%_|%@];

# Alphabet used for Morphological Tags
define MAlpha
[a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z|A|B|C|D|E|
F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z|1|2|3|%_];
```

```
# Alphabet used for Dependency Tags
define DAlpha [A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|
Y|Z|%<|%>|%_];

# Whitespace
define SP [" "|"\n"|"\t"]+ ;

# Token/Lemma/Tag definitions
define TokLem     [Alpha+ SP Alpha+ ];  # chuir cuir
define MTag       [%+ MAlpha+];         # +Verb
define TokLemMTag [TokLem MTag+ %+];    # chuir cuir +Verb+Past
+
```

Figure 64 Chunker Definitions: General

Figure 65 shows how verb chunks are implemented. We have already defined the general form of token, lemma and morphological tags. In this section, we define the dependency tags specific to verb chunks. Firstly we define verb functional tags (`VTag`), synthetic verb functional tags, i.e. verb and subject, (`VSTag`) and pre-verbal dependency tags (`PreVTag`). Next we define a pre-verbal string (`PreVStr`) and a verb string (`VStr`). In the case of `PreVStr`, this consists of the concatenation of a `TokLemMTag`, `PreVTag` and a space (`SP`). `VStr` is defined in a similar manner. A verb chunk (`VChunk`) is defined as zero or more `PreVStr*` followed by a verb string, `VStr`. Finally, we define a bracketed verb chunk, (`VChunkBr`), by surrounding the verb chunk with the labelled and unlabelled brackets "`[V` and "` ]`", using the longest match operator (`@->`). A bracketed synthetic verb chunk (`VSChunk`) is defined in a similar manner.

All other Level 1 chunks follow the same methodology.

```
# Verb Dependency Tags
############################################################
define VTag       [%@FAUX|%@FAUX%_REL|%@FMV|%@FMV%_REL];
define VSTag      [%@FAUX%_SUBJ|%@FAUX%_REL%_SUBJ|
                   %@FMV%_SUBJ|%@FMV%_REL%_SUBJ];
define PreVTag    [%@%>V];

# Verb Pre Modifiers
define PreVStr    [TokLemMTag PreVTag SP];

# Verb Chunk
define VStr       [TokLemMTag VTag SP];
define VChunk     [PreVStr* VStr];
define VChunkBr   [VChunk @-> "[V " ... "] "];
# Verb_Subject Chunk
define VSStr      [TokLemMTag VSTag SP];
define VSChunk    [PreVStr* VSStr];
define VSChunkBr  [VSChunk @-> "[VS " ... "] "];
```

Figure 65 Chunker Definitions: Verb Chunks

Higher level chunks, i.e. chunks containing other chunks, are created by defining how chunks may be combined. In Figure 66, we give the regular expressions used to define prepositional phrases. We begin by defining aspectual preposition dependency tags (`PPASTag`) and other preposition dependency tags (`PPADTag`). We use these to define three types of prepositional phrase. The first, `PPChunkBr1`, which brackets prepositional pronouns, is really a Level 1 chunk as it does not contain an embedded `NP`. Prepositional pronouns (also known as conjugated prepositions) are prepositions which incorporate a pronoun e.g. *leis* 'with him/it'.

The second type of PP consists of a preposition with an embedded complement NP. This bracketed chunk (`PPChunkBr2`) is defined as a preposition string followed by an NP chunk, which is to be surrounded by PP labelled brackets, i.e.`[PPSimpStr "[NP " ?+ " NP] "] @> "[PP " ... " PP] ";` Note that we do not use the longest match (`@->`) operator in this case, as we wish to include only one NP in the PP brackets, and there may be a several NPs following the preposition.

The third type of PP, an aspectual PP, embeds a verbal noun NP and possibly a preposed aspectual pronoun. This chunk, `PPChunkBr3`, is defined as follows: `[PPASSimpStr ("[OA " ?+ " OA] ") "[NP " ?+ " NP] "] @> "[PP-ASP " ... " PP-ASP] ";` where we have an aspectual preposition followed by a possible aspectual object (`[OA]`) chunk, followed by an `[NP]`, all of which are surrounded by `[PP-ASP` and `PP-ASP]` labelled brackets.

```
# Prepositional Phrases
#############################################################
# Aspectual Preposition Dependency Tags
define PPASTag       [%@PP%_ASP|%@PP%_STAT];
define PPASStr       [TokLemMTag PPASTag SP];

# Other Preposition Dependency Tags
define PPADTag       [%@PP%_ADVL|%@PP%_HAS|%@PP%_NEG|%@PP%_OBL|
                      %@PP%_PRED|%@PP%_SUBJ];
define PPADStr       [TokLemMTag PPADTag SP];

# 1) Prepositional Pronouns (a.k.a Conjugated Prepositions)
# These preps. incorporate a pronoun, leis = with him/it.
# Therefore PP has no nested NP complement.
# It can have a reflexive pronoun leis féin = with him/itself
define PostNStr0   [TokLemMTag PostNTag SP];  # féin (self)
define PPronTag    [%+Pron%+Prep];              # liom = with
me
define TokLemPPTag [TokLem PPronTag MTag+ %+];  # liom le Tags
+
define PPPronStr   [TokLemPPTag PPADTag SP];
define PPChunkBr1  [PPPronStr PostNStr0*  @-> "[PP " ... " PP]
"];
```

```
# 2) Simple and Compound Prepositions with NP complement
define PSimpTag  [%+Prep%+Simp]|  # le(Prep Simp) = with;
                 [%+Prep%+Poss]|  # lena(Prep Poss) = with its;
                 [%+Prep%+Cmpd]|  # ar nós (Prep Cmpd) = such
as
                 [%+Prep%+CmpdNoGen]|  # maidir le = regarding
                 [%+Prep%+Art]];  # sa (Prep Art) = in the
define TokLemPSTag [TokLem PSimpTag MTag* %+];  #
define PPSimpStr   [TokLemPSTag PPADTag SP];
define PPChunkBr2  [[PPSimpStr "[NP " ?+ " NP] "] @> "[PP " ...
" PP] "];

# 3) Aspectual Prepositions with NP Complement
# e.g. ag/p cabhrú/np (helping)
# OR possible pre-posed object do/p mo/oa chabhrú/np (helping
me)
define PPASSimpStr   [TokLemPSTag PPASTag SP];
define PPChunkBr3   [[PPASSimpStr ("[OA " ?+ " OA] ") "[NP " ?+
" NP] "] @> "[PP-ASP " ... " PP-ASP] "];
```

Figure 66 Chunker Definitions: Prepositional Chunks

For our final example, we will look at an aspectual chunk (Level 3 nesting). This may consist of just a `[PP-ASP]` chunk, but in the case of progressives it may include an non-pronominal aspectual object `[OA]` which always follows the verbal noun `[NP]`. Alternatively, the aspectual chunk may consist of an aspectual preposition `(PPSimpAStr)`, with a possible infinitival object, followed by an infinitival complement `[INF]`. This is defined as in Figure 67.

```
# Aspectual Phrases
################################################################
# [ASP [PP_ASP [[PP ag déanamh cáca PP] [OA cáca OA] PP-ASP]
ASP] # 'making a cake'
# [ASP tar_éis [INF imeacht] ASP] 'after leaving'
# [ASP tar_éis [OI cáca] [INF a dhéanamh] ASP] 'after making a
# cake' i.e. after a cake to make

define ASPChunkBr1 ["[PP-ASP " ?+ " PP-ASP] " ("[OA " ?+ " OA]
") @-> "[ASP " ... " ASP] "] ;

define ASPChunkBr2 [PPSimpAStr ("[OI " ?+ " OI] ") "[INF " ?+ "
INF] " @> "[ASP " ... " ASP] "] ;
```

Figure 67 Chunker Definitions: Aspectual Chunks

A full listing for the Finite-State Chunker may be found in Appendix G.

8.4 Evaluation

In this section we present results of evaluating the Finite-State Chunker, firstly against 225 made-up Test Suite sentences, and then using the NCII-based Gold Standard 250 Data, i.e. 150 Development Set sentences and 100 Test Set sentences.

8.4.1 Test Suite

We developed the Finite-State Chunker using the 225 sentences in the Test Suite which was used for Dependency Annotation development (see Appendix E). We automatically chunked the manually corrected dependency annotated Test Suite sentences and found the chunking to be 100% correct (using the *evalb* program[32]), in chunking these sentences.

Before running *evalb*, it is necessary to convert our data to the required input format, i.e. we must convert square brackets to round brackets, remove labels from the closing brackets, and enclose each token tagstring pair in round brackets. For our earlier sample sentence (260), the result is as follows:

```
(S
  (V (Bhéimnigh béimnigh+Verb+PastInd+Len+@FMV))
  (NP (sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ))
  (AD (freisin freisin+Adv+Gn+@ADVL))
  (NP (an an+Art+Sg+Def+@>N)(t-easpa easpa+Noun+Fem+Com+Sg+@OBJ)
      (seirbhísí seirbhís+Noun+Fem+Gen+Strong+Pl+@N<))
  (PP (ar_nós ar+nós+Prep+Cmpd+@PP_ADVL)
    (NP (HEMS HEMS+Guess+Abr+@P<)))
  (CJ2 (agus agus+Conj+Coord+@CC)
    (NP (an an+Art+Sg+Def+@>N) (bealach
         bealach+Noun+Masc+Com+Sg+Def+@CC<))
```

We also evaluated the Chunker using the same 225 POS tagged Test Suite sentences with automatic dependency tagging, but without manual correction of the dependency annotations. The resulting F-Score is still 100%, although this time, 26 sentences contain errors, as shown in the evalb output in Table 39. These errors are all the result of differences in dependency tags between the automatically tagged data and the Gold Standard. However, in all cases the chunk still receives the correct label, i.e. a noun may have the wrong grammatical function tag, or a verb may be tagged as a main verb rather than an auxiliary etc., but they will still fall within a correctly labelled NP or V chunk, and, therefore, do not affect bracketing recall or precision.

For the Test Suite, coverage is 100%, i.e. all text is included in a particular chunk.

[32] Downloadable from http://nlp.cs.nyu.edu/evalb/ (last accessed 30 June 2008).

Table 39 Test Suite (225): EVALB Bracket Scoring Summary

ALL SENTENCES (Len<40)	
Number of sentence	225
Number of Error sentence	**26**
Number of Skip sentence	0
Number of Valid sentence	199
Bracketing Recall	**100.00**
Bracketing Precision	**100.00**
Bracketing FMeasure	**100.00**
Complete match	100.00
Average crossing	0.00
No crossing	100.00
2 or less crossing	100.00
Tagging accuracy	100.00

8.4.2 Gold Standard Development Set Data

For a more realistic evaluation, using naturally occurring data, we evaluate the Chunker against the 150 Development Set sentences of the NCII-based Gold Standard Dependency Annotated (250) Corpus, also using the *evalb* program.

In the early stages of development, some chunking errors were identified, which were clearly the result of errors in the Gold Standard Dependency Data. Upon inspection, these in turn were a result of POS tagging errors relating to noun case marking and attributive versus predicative marking on adjectives. These problems were corrected in the Gold Standard Dependency Data (Development Set) before re-running the *evalb* evaluation.

The output of the *evalb* program, in Table 40, shows an overall precision of 98.15%, and for sentences with less than 40 words, precision is 98.57%. The reasons for this high result are twofold, 1) we have detailed manually corrected grammatical and dependency information available from the dependency analysis which greatly facilitates accurate and elegant chunking, and 2) we have postponed the more difficult aspects of parsing, i.e. co-ordination, prepositional attachment and long-disance dependencies to a later stage. Nevertheless this is a very positive result.

Table 40 Development Set (150): EVALB Bracket Scoring Summary

ALL SENTENCES		**SENTENCES Len<40**	
Number of sentence	150	Number of sentence	120
Number of Error sentence	0	Number of Error sentence	0
Number of Skip sentence	0	Number of Skip sentence	0
Number of Valid sentence	150	Number of Valid sentence	120
Bracketing Recall	**96.26**	Bracketing Recall	97.31
Bracketing Precision	**98.15**	Bracketing Precision	98.57
Bracketing FMeasure	**97.20**	Bracketing FMeasure	97.94
Complete match	68.42	Complete match	80.00
Average crossing	0.05	Average crossing	0.03
No crossing	96.71	No crossing	99.17
2 or less crossing	99.34	2 or less crossing	99.17
Tagging accuracy	100.00	Tagging accuracy	100.00

8.4.3 Gold Standard Test Set Data

We also evaluate the Chunker against the 100 sentence Test Set of the Gold Standard Dependency Annotated (250) Corpus, using the *evalb* program. The precision in this case 94.12%. This and other measures are presented in Table 41.

Table 41 Test Set (100): EVALB Bracket Scoring Summary

ALL SENTENCES		**SENTENCES Len<40**	
Number of sentence	100	Number of sentence	85
Number of Error sentence	0	Number of Error sentence	0
Number of Skip sentence	0	Number of Skip sentence	0
Number of Valid sentence	100	Number of Valid sentence	85
Bracketing Recall	**92.89**	Bracketing Recall	**94.09**
Bracketing Precision	**94.12**	Bracketing Precision	**94.09**
Bracketing FMeasure	**93.50**	Bracketing FMeasure	**94.09**
Complete match	61.39	Complete match	67.06
Average crossing	0.21	Average crossing	0.18
No crossing	83.17	No crossing	85.88
2 or less crossing	100.00	2 or less crossing	100.00
Tagging accuracy	100.00	Tagging accuracy	100.00

Precision, recall and f-score (FMeasure) results are lower for the Test Set than the Development Set. This is probably due at least in part to errors and underspecified items in the underlying data, e.g. names and titles where the relationships and dependencies between nouns are not explicitly shown, as well as errors in the Gold Standard Dependency Annotations.

8.4.4 Error Analysis

In order to assess the coverage of the Chunker, we run the 150 sentence Development Set chunked text through a Perl program which outputs any material not belonging to a chunk. Apart from punctuation and XML tags, which we currently ignore, a number of fragments of text were not part of any chunk.

In Table 42, we categorise the various types of fragment which are omitted from chunks. Coordinated elements are by far the most common fragment (41%) to be omitted from a chunk. While some have a simple cause such as a comma intervening between the conjunction and the following chunk, most require more comprehensive handling of coordinate structures.

In (264), we have a prepositional phrase whose object is a complex noun phrase, in which conjoined genitive nouns *réamhléitheoireachta agus réamh-scríbhneoireachta* 'pre-reading and pre-writing' modify the head noun *ngníomhaíochtaí* 'activities' (which is itself a genitive noun following a compound preposition *le linn* 'during'). As we have not implemented coordination in a comprehensive manner, *réamh-scríbhneoireachta* 'pre-writing' appears to be a genitive noun without a head and therefore does not fall within our definition of a noun chunk.

(264) le_linn na ngníomhaíochtaí réamhléitheoireachta **agus réamh-scríbhneoireachta**
during the activities pre-reading and pre-writing
PREP-CMPD ART NOUN-GEN NOUN-GEN CONJ NOUN-GEN
'during pre-reading and pre-writing activities'

In (265), we have a similar problem, except that in this case, we have conjoined prepositions, *ar an agus ón* 'on the and from the'. In this case, the first conjoint *ar an* 'on the' appears not to have a complement and therefore it is excluded from the following prepositional phrase *ón lá* 'from the day'.

(265) **ar** **an** agus ón lá
on the and from-the day
PREP ART CONJ PREP NOUN
'as and from the day'

The remaining items in Table 42, are unproblematic, and only require relatively straightforward extensions to the Finite-State Chunker's regular expressions to handle these additional structures.

Table 42 Chunker: Development Set Error Analysis

	Analysis of Unchunked text		
1	Coordinate structures	15	41%
2	Compound Prepositions without an object	6	16%
3	Preps. with Infinitival Object	5	14%
4	List Items	5	14%
5	Proper Nouns	3	8%
6	Text in Quotes	2	5%
7	Gen. NP including Number	1	2%
		37	**100%**

8.5 Summary

In this chapter, we describe our method of chunking for Irish. We describe the annotation scheme we use for labelling chunks and the levels of embedding which we have implemented. We also describe the regular expression implementation of the Chunker. Finally, we present the results of our evaluation and error analysis. Currently, chunking of dependency annotated text achieves an f-score of 97.2% on Development Set Data and 93.5% on Test Set Data. The difference between 93.5% on Gold Standard Corpus Test Data and 100% on the manually composed Test Suite data highlights the necessity for testing on real world corpus data.

9 Conclusion

In this thesis, we describe the design, implementation and evaluation of a POS tagger and Partial Parser for Irish. Through this work, we provide a valuable set of tools for Irish NLP, as well as a platform for further research. To our knowledge, these are the only such tools for Irish.[33] In addition to these tools, we provide a useful linguistic resource in our Gold Standard Corpus which can be used for both linguistic research and machine-learning applications.

In this the final Chapter, we summarise the preceding chapters, highlight our main contributions to research, and outline possible directions for future research.

9.1 Summary

This dissertation is arranged in three parts: Part I: Background, Part II: POS Tagging and Part III: Partial Parsing. In Part I, we present the development of a corpus of Irish texts, followed by a discussion of techniques for POS tagging and Partial Parsing. We finish Part I with a description of our Gold Standard Corpus and evaluation measures. In Part II, we present our method of POS Tagging for Irish, and in Part III, we present a method for the Partial Parsing of Irish using Dependency Analysis and a Finite-State Chunker.

Part I: Background

The main focus of the thesis is on the development of text processing tools for Irish. However, in order to develop such tools, a large body of texts is required for development and testing purposes. Indeed, this body of texts is a valuable resource in its own right. We, therefore, began by describing our involvement in the creation of a 30 million word corpus of Irish texts (NCII). We, briefly, describe corpus design and text collection, and then go into more detail about the task of text preparation. Text preparation is vital, as the quality of the raw text in a corpus has a bearing on every subsequent step in the linguistic annotation process, as well as on the utility of the annotated corpus for the end user.

Next, we discuss the main techniques for POS tagging, i.e. Statistical Data-Driven Tagging, Rule based Tagging, and Transformation based Tagging. In the area of parsing, we discuss both constituency based annotation and Dependency Analysis annotation.

[33] Dr. Kevin Scannell, St. Louis University has carried out related work in developing a grammar checker for Irish. See http://borel.slu.edu/nlp.html for more details.

In Part I, we also describe the development of a 3,000 sentence Gold Standard Corpus, as well as the evaluation measures, i.e. precision, recall and f-score, which are used in Parts II and III.

Part II: POS Tagging of Irish

The first step in processing a corpus of texts is tokenization. This entails dividing the input stream into separate tokens which will be passed on to the morphological analyser. By default, a token is a sequence of characters bounded by white-space. Multi-word expressions which we wish to keep together (e.g. idioms, place names etc.) and contractions which we wish to divide (e.g. *d'fhéach* 'looked', *m'aghaidh* 'my face' etc.) must be explicitly defined. By default, punctuation is separated from words, and any exceptions to this general rule (e.g. abbreviations, titles, mathematical formulae etc.) must also be explicitly defined.

Next, we describe the scaling up of a prototype finite-state morphological analyser (Uí Dhonnchadha, 2002) for use on unrestricted text. This involved extending the basic lexicon, the addition of named entities (names, places, organisations etc.), and the addition of derivational morphology rules. Coverage was increased by more than 12%, resulting in over 95% of tokens receiving at least one analysis.

To account for the unrecognized tokens (5% of tokens), we developed a series of morphological guessers. The guessers make use of stems, prefixes and suffixes in the lexicon to identify possible compounds and derived words. The remaining unrecognized tokens are analysed according to any distinguishing characteristics which they may have, e.g. characters and syllables which are indicative of a part-of-speech category, or other morphological features such as gender, number, tense, person etc.

The morphological analyser outputs multiple analyses per token, in two thirds of cases on average. The challenge in POS tagging is to choose the appropriate analysis for the token based on its context in the text. This disambiguation task is achieved using Constraint Grammar rules, which use a combination of the token's morphosyntactic properties and its local context within the sentence, in order to select the correct analysis. Based on comparison with a manually verified evaluation corpus (i.e. a gold standard), the tagger chooses the correct POS analysis in 95% approx. of cases.[34]

[34] See https://www.cs.tcd.ie/Elaine.UiDhonnchadha/irish.htm for a demonstration of Irish POS Tagging.

Part III: Partial Parsing of Irish

The next step in our linguistic annotation process is partial parsing. This means grouping the tokens in a sentence into larger syntactic units, known as chunks. Chunks may contain more than one phrasal head, i.e. an NP chunk may contain adjectives which could also be considered phrasal heads. In parsing a language for the first time, deciding what those syntactic units are, and how they should be annotated, constitutes a major part of the work.

There are two main schools of thought regarding syntactic annotation of corpora, i.e. a constituency based analysis, or a dependency based analysis, and some parsed corpora (treebanks) combine elements of both. There is a substantial overlap between both types of analysis and one can be mapped on to the other to a large degree.

Our primary aim in this exploration of partial parsing of Irish is to account for as much of the linguistic phenomena as possible and to decide on an initial style guide for the partial syntactic annotation of the language. In order to do so, we have used a dependency analysis overlaid with chunk boundaries. In our dependency analysis, we only tag the tokens present in the input string, i.e. we do not posit abstract or elipted categories. In our chunking, we have not implemented recursion. This results in a partial rather than full parse of the sentences.

The dependency analysis currently achieves an f-score of 93.60% on Gold Standard POS tagged Development Data and 94.28% on unseen Gold Standard POS tagged Test Data. The chunker achieves an f-score of 97.20% on the Development data and 93.50% on the unseen Test Data.

9.2 Main Contributions

The main achievements described in this thesis, include the development of NLP tools and annotated corpora for Irish. Other useful resources include, a) a set of morphological continuation classes for the analysis and generation of Irish nouns, verbs and adjectives, b) guidelines for manual POS tagging (Appendix C), and c) an exploratory set of syntactic labels and classes for parsing of Irish sentences (Chapters 7 & 8).

Partial parsing of Irish presents a number of challenges. As this is the first attempt at implementing a partial parser for Irish, (to our knowledge), there were no guidelines or precedents available, and, therefore, many decisions had to be made. The fact that Irish is a VSO language i.e. the subject occurs between the verb and its object, means that the standard SVO definition of VP does not apply. In addition, many aspectual functions are carried out using nominal rather than verbal constructions. Furthermore, in common with the

other Celtic languages, Irish has the unusual phenomenon of prepositions which are inflected for person and number.

The tools and corpus resources which constitute the main contribution are summarised below.

9.3 NLP tools for Irish

- **Tokenizer and Morphological Analyser and Generator**

We have developed a full-scale finite-state implementation of tokenization and morphological analysis for Irish. The finite-state lexicons contain 30K lemmas and this is currently being extended by a further 30K lemmas. We are not aware of any other such tools for the language.

- **A POS tagger**

We have developed a POS tagger for Irish which currently achieves an f-score of 95% on development data and 94.35% on unseen test data. This POS tagger has been used to tag a 30-million word corpus of Irish, which will be used in a government funded project to develop the first ever corpus-based English-Irish Dictionary (Kilgarriff, Rundell and Uí Dhonnchadha, 2007). It is currently being used in Irish Text-to-Speech Synthesis research[35] in Trinity College Dublin, and has also been used in the WISPR (Welsh and Irish Speech Processing Resources) Project (Prys et al., 2004).

- **A Partial Parser for Irish**

The partial parser for Irish, uses dependency analysis and finite-state chunking. The dependency analysis currently achieves an f-score of 93.60% on development data and 94.28% on unseen test data. The chunker which uses information provided in the dependency tags achieves an f-score of 97.20% on development data and 93.50% on unseen test data.

9.4 Linguistic Resources for Irish

As well as the tools themselves, the following linguistic resource are now available:

[35] See http://www.abair.ie (last accessed 30 June 2008) for a demonstration of Irish Text-to-Speech Synthesis

- **A 30-million word automatically POS tagged NCII Corpus**

The NCII corpus is sponsored and managed by Foras na Gaeilge, the government body in charge of promoting the Irish language on the island of Ireland. Plans are under way to make it publicly available on the Internet. This will be of enormous benefit to scholars of Irish and of linguistics, and to commercial bodies interested in developing language applications, as well as to interested members of the public.

- **A 3,000 Sentence Gold Standard POS Tagged Corpus, a 250 Sentence Gold Standard Dependency Analysis Corpus, and a 250 Sentence Gold Standard Chunked Corpus**

Creating a manually verified gold standard resource is a time-consuming, tedious and error-prone task, but once completed provides a very valuable resource for a variety of further research. For example, the Gold Standard POS Tagged Corpus of Irish has recently been used as training data in machine-learning algorithms to learn morphological features and lemmatization classes (Chrupala, 2008). This data could also be used as training data for a statistical POS tagger.

9.5 Future Research

We hope that this preliminary work on syntactic parsing for Irish, provides a basis for further research in this area, and, in particular, we would like to develop an Irish Treebank. In order to proceed in this direction, research into subcategorization frames for Irish verbs as well as semantic classes for nouns would be an beneficial. The issues of PP-attachment, long-distance dependencies and co-ordination must also be addressed.

We hope to investigate the automatic induction of Constraint Grammar rules using the Gold Standard Corpus, in order to further improve the precision of the POS tagger.

We hope to generate morphological analyses in a form compatible with the CHILDES CHAT format to facilitate the study of first language acquisition of Irish.

We hope to continue collaborating with our colleagues in integrating POS tagging and Chunking into Text-to-Speech Synthesis, and also Automatic Speech Recognition.

We would like to use the POS tagged corpora (Gold Standard and/or NCII) to train a Brill POS tagger for Irish.

Glossary of Terms

Clause: Any constituent dominated by the larger structure S; usually divided into two types - main and subordinate (Trask, 1992, p44).

Complement: Any constituent which is subcategorized for by a lexical head, e.g. In Lisa put the book on the table, the NP 'the book' and the locative phrase 'on the table' are complements of the verb 'put', while 'the table' is the complement of the preposition 'on'(Trask, 1992, p51).

Complementizer: A type of subordinator that begins a complement phrase, e.g. 'that' in 'I said that I wasn't perfect' (Biber et al., 2003).

Constituent: Any part of a sentence which behaves as a syntactic unit within the structure of the sentence, with respect to displacement, coordination, ellipsis or pro-form replacement (Trask, 1992, p57).

Elision: A general term for the omission of material which is required to complete a syntactic structure, e.g. 'Seems we have a problem', where the initial it has been elided (Trask, 1992, p89).

Elipsis: A construction where some material is omitted, but which is immediately recoverable from the context, e.g. 'John can speak Irish but Pat can't' (Trask, 1992, p89).

Extraposition: Dummy *it* fills subject slot, and complement (that) clause is placed after predicate, e.g. 'It is clear that it will not be simple' (Biber et al., 2003).

Finite-State: A finite-state machine is a model of computation, defined in terms of an initial state and one or more transitions, resulting in one or more final states. A finite-state transducer is a two-level finite-state machine.

Fronting (Preposing): Any construction in which a constituent which is typically found elsewhere is brought forwarded to the front (preposed) of the sentence, e.g. 'carefully' is a preposed adverb in 'Carefully she decanted the wine' (Trask, 1992, p10).

Matrix clause: Any clause which contains an embedded clause (Trask, 1992, p168).

Predicate: Logical centre of a clause - can be verb (I thought) or copula+adj (I'm sure) (Biber et al., 2003).

Predicative: A clause element that characterizes the referent of some other clause element, e.g. subject (i.e. subject-predicative), or object (i.e. object-predicative) (Biber et al., 2003).

Predicate Complement: A category occurring in a complement which is interpreted as describing or referring to another NP in the sentence, e.g. in 'Lisa is a translator', translator describes the subject Lisa (Trask, 1992, p51).

Predicate Object-Complement: In 'He called me a fool', 'fool' describes the object 'me' and is therefore an object-complement (Trask, 1992, p51).

Predicate Subject-Complement: In 'Lisa is a translator', 'translator' describes the subject 'Lisa' and is therefore a subject-complement (Trask, 1992, p51).

Preposing: see Fronting.

Relativizer: A grammatical form which introduces a relative verb clause (Stenson, 1981, p32).

Subordinator: A lexical category whose members introduce adverbial clauses, e.g. *because* it was amazing; *if* he is going with me (Biber et al., 2003, p268).

Publications Resulting from Research Reported in Dissertation

Kilgarriff, A., Rundell, M., Uí Dhonnchadha, E., (2005). Corpus creation for lexicography In: Proceedings of AsiaLex 2005, Singapore.

Kilgarriff, A., Rychly, P., Chu-Ren, H., Smith, S., Tugwell, D., Uí Dhonnchadha, E, (2005). Word sketches for Irish and Chinese. In Proc: Corpus Linguistics 2005 Birmingham July 2005.

Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E., (2007). Efficient corpus creation for lexicography. *Language Resources and Evaluation Journal*.

Prys, D., Williams, B., Hicks, B., Jones, D., Ní Chasaide, A., Gobl, C., Carson-Berndsen, J., Cummins, F., Ní Chiosáin, M., McKenna, J., Scaife, R., Uí Dhonnchadha, E., (2004). WISPR: Speech Processing Resources for Welsh and Irish. In Proc: First Steps in Language Documentation for Minority Languages. SALTMIL Workshop in association with LREC 2004, Lisbon.

Uí Dhonnchadha, E, and van Genabith, J., (2006). A Part-of-speech tagger for Irish using Finite-State Morphology and Constraint Grammar Disambiguation. Paper presented at *LREC 2006*, Genoa.

Uí Dhonnchadha, E., Van Genabith, J., (2007). Scaling an Irish FST morphology engine for use on unrestricted text. In: Lecture Notes in Artificial Intelligence (LNAI): Proceedings of the FSMNLP 2005 (Eds: Lauri Karttunen, Juhani Karhumäki, and Anssi Yli-Jyrä). Springer Publications

Uí Dhonnchadha, E. (2003). Finite-State Morphology and Irish. In: Proceedings of EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics. Budapest, Hungary.

References

Abney, S. 1991. Parsing by Chunks. In *Principle-Based Parsing*, eds. Robert Berwick, Stephen Abney and Carol Tenny. Dordrecht: Kluwer Academic Publishers.

Abney, S. 1996b. Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2:337-344.

Adger, D., and Ramchand, G. 2003. Predication and Equation. *Linguistic Inquiry* 34.

An Gúm. 1999. *Graiméar Gaeilge na mBráithre Críostaí*. Baile Átha Cliath: An Gúm.

An Roinn Oideachas. 1986. *Foclóir Póca English-Irish/Irish-English Dictionary*. Baile Átha Cliath: An Gúm.

Attia, M. 2000. A Large-Scale Computational Processor of the Arabic Morphology,and Applications, Computing Engineering, Cairo University.

Banko, M., and Moore, R. 2004. Part of Speech Tagging in Context. Paper presented at *COLING 2004*, Geneva, Switzerland.

Beesley, K. 1998. Arabic Morphology Using Only Finite-State Operations. Paper presented at *Workshop On Computational Approaches To Semitic Languages*.

Beesley, K., and Karttunen, L. 2003. *Finite State Morphology*. California: CSLI Publications.

Biber, D., Conrad, S., and Leech, G. 2003. *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Bick, E. 2006. Turning a Dependency Treebank into a PSG-style Constituent Treebank. Paper presented at *5th. Conference on Language Resources and Evaluation*, Genoa, Italy.

Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project.

Brants, T., and Franz, A. *Web 1T 5-gram Version 1*. 2006. Linguistic Data Consortium, Philadelphia.

Brants, T., Skut, W., and Uszkoreit, H. 2003. Syntactic Annotation of a German Newspaper Corpus. In *Treebanks: Building and Using Parsed Corpora*, ed. Anne Abeillé. Dordrecht, Boston, London: Kluwer Academic Publishers.

Bresnan, J. 2001. *Lexical Functional Syntax*: Blackwell.

Brill, E. 1995a. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics* 21:543-565.

Brill, E. 1995b. Unsupervised learning disambiguation rules for part of speech tagging. Paper presented at *ACL Third Workshop on Very Large Corpora*, Cambridge, MA.

Brown, K., and Miller, J. 1991. *Syntax: A Linguistic Introduction to Sentence Structure*. London and New York: Routledge.

Carnie, A., and Guilfoyle, E. eds. 2000. *The Syntax of Verb Initial Languages*. Oxford: Oxford University Press.

Chanod, J.-P., and Tapanainen, P. 1995a. Tagging French – comparing a statistical and a constraint-based method. Paper presented at *EACL'95: Seventh Conference of European Chapter of Association of Computational Linguistics*, Dublin.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 1988. *Lectures on government and binding : the Pisa lectures*. Dordrecht Foris.

Christian Brothers. 1988. *New Irish Grammar*. Dublin: C J Fallon.

Chrupala, G. 2008. Towards a Machine-Learning Architecture for Lexical Functional Grammar Parsing, School of Computing, Dublin City University.

CLAWS. URL: <http://www.comp.lancs.ac.uk/ucrel/claws/>.

Cook, W. A. 1989. *Case grammar theory*. Washington, D.C: Georgetown University Press.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. 1992. A practical part-of-speech tagger. Paper presented at *3rd Conference on Applied Natural Language Processing*, Trento, Italy.

Dahl, Ő. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Diab, M., Hacioglu, K., and Jurafsky, D. 2005. Tagging of Arabic Text: From raw text to Base Phrase Chunks. Paper presented at *HLT-NAACL 2004*.

Dineen, R. P. S. 1934. *Foclóir Gaeilge agus Béarla* Dublin & Cork: The Educational Company of Ireland.

Doherty, C. 1996. Clausal structure and the Modern Irish Copula. *Natural Language and Linguistic Theory* 14:1-46.

Doherty, C. 1997. The Pronominal Augment in Irish Identificational Sentences. In *Dán do Oide*, eds. Anders Ahlqvist and Vera Čapková. Dublin: Institiúid Teangeolaíochta Éireann.

Duffield, N. 1995. *Particles and Projections in Irish Syntax*. Dordrecht: Kluwer.

EAGLES. 1996 "Synopsis and Comparison of Morphostntactic Phenomena Encoded in Lexicons and Corpora". URL: <www.ilc.cnr.it/EAGLES96/morphsyn/node12.html>.

Fillmore, C. J. 1968. The Case for Case. In *Universals in linguistic theory*, eds. Emmon Bach and Robert T Harms. New York, London: Holt, Rinehart and Winston.

Garside, R. ed. 1987. *The CLAWS word-tagging system*. *The Computational Analysis of English: a corpus-based approach.* London: Longman.

Garside, R. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report. In *Spoken English on the computer: transcription, mark-up and application*, eds. Geoffrey Leech, Greg Myers and Jenny Thomas. Essex: Longman.

Grefenstette, G., Schiller, A., and S, A.-M. 2000. Recognizing Lexical Patterns in Text. In *Lexicon Development for Speech and Language Processing*, eds. F van Eynde and Dafidd Gibbon. Dordrecht: Kluwer Academic Publishers.

Grefenstette, G., and Tapanainen, P. 1994. What is a word, what is a sentence? Problems of tokenization. Paper presented at *The 3rd International Conference on Computational Lexicography*, Budapest.

Guthmann, N., Krymolowski, Y., Milea, A., and Winter, Y. 2009. Automatic Annotation of Morpho-Syntactic Dependencies in a Modern Hebrew Treebank. Paper presented at *7th. International Workshop on Treebanks and Linguistic Theories (TLT) 2009*, Groningen.

Habash, N., and Rambow, O. 2006. A Morphological Analyzer and Generator for the Arabic Dialects

Paper presented at *Coling-ACL*, Syndey, Australia.

Habert, B., Adda, G., Adda-Decker, M., Boula de Marëuil, P., Ferrari, S., Ferret, O., Illouz, G., and Paroubek, P. 1998. Towards tokenization evaluation. In , editors, , volume I, pages , Granada, May 1998. Paper presented at *International Conference on Language Resources and Evaluation*, Grenada.

Hajič, J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues in Valency and Meaning. Studies in Honour of Jarmila Panevová*, ed. Eva Hajičová, 106-132. Prague: Charles University Press.

He, Y., and Kayaalp, M. 2006. A Comparison of 13 Tokenizers on MEDLINE. Bethesda, MD: The Lister Hill National Center for Biomedical Communications.

Hindle, D. 1993. A parser for text corpora. In *Computational Approaches to the Lexicon*, eds. B. T. S. Atkins and Antonio Zampolli. Oxford: Oxford University Press.

Hudson, R. A. 2007. *Language networks: the new word grammar*. Oxford: Oxford University Press.

Ide, N., Bonhomme, P., and Romary, L. 2000. XCES: An XML-based Standard for Linguistic Corpora. Paper presented at *2nd Language Resources and Evaluation Conference*, Athens.

Ide, N., and Suderman, J. 2002 "XCES: Corpus Encoding Standard for XML". URL: <http://www.ces-xml.org>. Date Accessed: Oct 2007.

ITÉ. 2001. Parole Corpus of Irish: ITÉ.

ITÉ. 2002. Reference Corpus of Irish: ITÉ.

ITÉ. 2003. Corpus Náisiúnta na Gaeilge. Baile Átha Cliath: ITÉ.

Järvinen, T. 2003. Bank of English and Beyond. In *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, eds. Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila, 430. Berlin - New York: Mouton de Gruyter.

Johansson, S. 1986. The Tagged LOB Corpus Users Manual: Norwegian Computing Centre for the Humanities, Bergen.

Jurafsky, D., and Martin, J. H. 2000. *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing*. Saddle River, N.J.: Prentice Hall.

Karlsson, F. 1995. Designing a parser for unrestricted test. In *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, eds. Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila, 430. Berlin - New York: Mouton de Gruyter.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. eds. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. vol. 4. Berlin - New York: Mouton de Gruyter.

Karttunen, L., and Beesley, K. 1992. Two-Level Rule Compiler. Palo Alto: Xerox PARC.

Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E. 2007. Efficient corpus creation for lexicography. *Language Resources and Evaluation Journal*.

Krauwer, S. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. Paper presented at *Speech and Computer (SPECOM-2003)*.

Kroeger, P. 2004. *Analysing Syntax: A lexical-functional approach*: Cambridge University Press.

Kuĉera, H., and Francis, W. N. 1967. Computational Analysis of Present-day American English. Providence, RI: Brown University Press.

Leech, G., Garside, R., and Bryant, M. 1994. CLAWS4: The tagging of the British National Corpus. Paper presented at *COLING 94 - 15th International Conference on Computational Linguistics*.

LMC. 2004 "Design Principles for the New Corpus for Ireland (NCI) Version 2 ". URL: <http://www.focloir.ie/pdf/TaskH_corpus%20design%20principles_Final.pdf>.

Mac Congáil, N. 2002. *Leabhair Gramadaí Gaeilge*. Indreabhán, Co. na Gaillimhe: Cló Iar-Chonnachta.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Franz, A., Katz, K., and Schasberger, B. 1994. The Penn Treebank: Annotating Predicate Argument Structure.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19:313-330.

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 *Order, Concord and Constituency*, eds. G Gazdar, E Klein and G Pullum, 9-55. Dordrecht: Foris.

McCloskey, J. 1985. The Modern Irish Double Relative and Syntactic Binding. *Ériu* 36:45-84.

Mel'čuk, I. A. 1988. *Dependency Syntax: Theory and Practice*. Albany: State university of New York Press.

Meyer, C. F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Mikheev, A. 2003. Text Segmentation. In *The Oxford Handbook of Computational Linguistics*, ed. Ruslan Mitkov, 201-218. Oxford: Oxford University Press.

Mittendorf, I., and Sadler, L. 2006. A Treatment of Welsh Initial Mutation. Paper presented at *LFG06*, Konstanz.

Multext. 1996 "Multext". URL: <http://www.lpl.univ-aix.fr/projects/multext/>.

Nivre, J. 2006. *Inductive Dependency Parsing*: Text, Speech and Language Technology. Dordrecht: Springer.

Nivre, J. 2007. Dependency Grammar and Dependency Parsing. In *ESSLLI 2007: 19th European Summer School in Logic, Language and Information (Course:Introduction to Data-Driven Dependency Parsing)*. Trinity College Dublin, Ireland.

Nivre, J., and Hall, J. 2005. MaltParser: A language-independent system for data-driven dependency parsing. Paper presented at *4th. International Workshop on Treebanks and Linguistic Theories (TLT) 2009*.

Nolan, B. 2001. A Study of Valency in Modern Irish, The Centre for Language and Communication Studies, University of Dublin, Trinity College.

O' Neill Lane, T. 1916. *Lanes's Larger English-Irish Dictionary* Dublin & Belfast: The Educational Company of Ireland.

Ó Baoill, D., and Ó Tuathaill, É. 1992. *Úrchúrsa Gaeilge*: Institiúid Teangeolaíochta Éireann.

Ó Cróinín, D., and Uí Dhonnchadha, E. 1998. LE-PAROLE and Corpus Náisiúnta na Gaeilge. Paper presented at *Language Resources and Evaluation (LREC)*, Grenada, Spain.

Ó Dónaill, N. 1977. *Foclóir Gaeilge Béarla*. Baile Átha Cliath: Oifig an tSoláthair.

Ó Droighneáin, M. 1991. *An Sloinnteoir Gaeilge agus an tAinmneoir*. Baile Átha Cliath: Coiscéim.

Ó hUallacháin, C., and Ó Murchú, M. 1981. *Irish Grammar*: University of Ulster Coleraine.

Ó Siadhail, M. 1989. *Modern Irish: Grammatical structure and dialectal variation*. Cambridge: Cambridge University Press.

Ó Siochfhrada, N. 1998. *Foclóir Gaeilge/Béarla - Béarla/Gaeilge*. Baile Átha Cliath: An Comhlacht Oideachais, Cló Thalbóid.

PARGRAM. "Parallel Grammar and Parallel Semantics Projects". URL: <http://www2.parc.com/isl/groups/nltt/pargram/>.

Perlmutter, D. M., and Rosen, C. G. eds. 1984. *Studies in relational grammar. 2*. Chicago; London: University of Chicago Press.

Pollard, C., and Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Prys, D., Williams, B., Hicks, B., Jones, D., Ní Chasaide, A., Gobl, C., Carson-Berndsen, J., Cummins, F., Ní Chiosáin, M., McKenna, J., Scaife, R., and Uí Dhonnchadha, E. 2004. WISPR: Speech Processing Resources for Welsh and Irish. Paper presented at *First Steps in Language Documentation for Minority Languages. SALTMIL Workshop in association with LREC 2004*, Lisbon.

Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. Paper presented at *EMNLP*.

Sampson, G. 1993. The Susanne Corpus. Release 2.

Samuelsson, C., Tapanainen, P., and Voutilainen, A. 1996. Inducing Constraint Grammars. In *Grammatical Inference: Learning Syntax from Sentences*, eds. L. Miclet and C. de la Higuera: Springer.

Scannell, K. 2007 "Natural Language Processing". URL: <http://borel.slu.edu/nlp.html>.

Sima'an, K., Itai, A., Winter, Y., Altman, A., and Nativ, N. 2001. Building a Tree-Bank of Modern Hebrew Text. In *Traitment Automatique des Langues*.

Starosta, S. 1988. *The case for lexicase: an outline of lexicase grammatical theory*. London: Pinter.

Stenson, N. 1981. *Studies in Irish Syntax*: Ars Linguistica. Tübingen: Gunter Narr Verlag.

Tapanainen, P. 1996. The Constraint Grammar Parser CG-2. Publication No. 27: University of Helsinki.

Tapanainen, P. 1999. Parsing in two frameworks: finite-state and functional dependency grammar, University of Helsinki: Ph.D. Thesis.

Tapanainen, P., and Järvinen, T. 1997. A non-projective dependency parser. Paper presented at *5th. Conference on Applied Natural Language Processing*, Washington D.C.

Tapanainen, P., and Voutilainen, A. 1994. Tagging accurately - Don't guess if you know. Paper presented at *5th. Conference on Applied Natural Language Processing (ANLP'94)*, Stuttgart.

Taylor, A., Marcus, M., and Santorini, B. 2003. The Penn Treebank: An Overview. In *Trebanks: Building and Using Parsed Corpora*, ed. Anne Abeillé. Dordrecht; Boston; London: Kluwer Academic Publishers.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Librairie Klincksieck.

Trask, R. L. 1992. *A Dictionary of Grammatical Terms in Linguistics*. London & New York: Routledge.

Uí Dhonnchadha, E. 2002. An Analyser and Generator for Irish Inflectional Morphology using Finite State Transducers, School of Computing, Dublin City University: Unpublished MSc Thesis.

Uí Dhonnchadha, E., Nic Pháidín, C., and van Genabith, J. 2005. Design, Implementation and Evaluation of an Inflectional Morphology Finite-State Transducer for Irish. *MT Journal - Special Issue on Finite State Language Resources and Language Processing*.

Uí Dhonnchadha, E., and van Genabith, J. 2006. A Part-of-speech tagger for Irish using Finite-State Morphology and Constraint Grammar Disambiguation. Paper presented at *LREC 2006*, Genoa.

Van Valin, R. D. 2001. *An Introduction to Syntax*. Cambridge: Cambridge University Press.

Voutilainen, A. 1995. Morphological Disambiguation. In *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, eds. Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila, 165-284. Berlin - New York: Mouton de Gruyter.

Voutilainen, A., Heikkilä, J., and Anttila, A. 1992. Constraint Grammar of English. A Performance-Oriented Introduction. Helsinki: Department of General Linguistics, University of Helsinki.

Wallis, S. 2003. Completing Parsed Corpora. In *Treebanks: Building and Using Parsed Corpora*, ed. Anne Abeillé. Dordrecht; Boston; London: Kluwer Academic Publishers.

Wigger, A. 2007. Advances in the lexicography of Modern Irish verbs. Paper presented at *Poznań Linguistic Meeting*, Gniezno, Poland.

Wintner, S. 2008. Strengths and weaknesses of finite-state technology: a case study in morphological grammar development. *Natural Language Engineering* 14.

Wintner, S., and Yona, S. 2003. Resources for Processing Hebrew. Paper presented at *MT-Summit IX workshop on machine translation for semitic languages*, New Orleans.

SOFTWARE

Xerox Finite-State Tools (tools: lexc, xfst, twolc; operating system: Linux/Solaris). For details contact: Xerox Research Centre Europe, Attn: Licensing of Finite-State Programming Languages, 6 chemin de Maupertuis, 38240 Meylan, France. See also: http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html (last accessed 10 May 2008).

VISL CG For information see: http://beta.visl.sdu.dk/constraint_grammar.html and for source code see: http://beta.visl.sdu.dk/cg3.html. (last accessed 10 May 2008).

EVALB Software for bracketing evaluation. Available from: http://nlp.cs.nyu.edu/evalb/ (last accessed 10 May 2008).

Appendix A: Parole Morphosyntactic Descriptions for Irish

APPENDIX A: PAROLE MORPHOSYNTACTIC DESCRIPTIONS FOR IRISH 1

PAROLE MORPHOSYNTACTIC TAGSET FOR IRISH (REVISED 2004) 3
PAROLE SHORT TAGS (POS ONLY).....9

Parole Morphosyntactic Tagset For Irish (Revised 2004)

1. NOUN								
1. N	**2. Type** c = common p = proper s = substantive v = verbal	**3. Gender** f = fem m = masc	**4. Number** s = sing. p = pl.	**5. Case** c = common g = genitive v = vocative d = dative	**6. Sem-Gender** n/a	**7. Contrast* e=emphatic	**8 Derived* v = de-verbal (assumed) n = de-nominal	

Type

- v = verbal or action noun - can be de-verbal or de-nominal, added to Noun POS category and removed from Verb category
  e.g. déanamh 'making' - de-verbal (déan 'make/do'), bádóireacht 'boating' - de-nominal (bádóir 'boatman/boatwoman')
- s = substantive - this is a term traditionally used for any single item that functions syntactically like a noun, but which does not have any other inflected forms in the nominal paradigm, e.g. *son* 'wellbeing', (in) *ann* 'there', both of which occur after a preposition in idiomatic phrases, i.e. *ar son* 'on behalf of', and *in ann* 'capable of' or 'able'.

Case

- c = common - same morphological form for nominative, accusative and dative
- d = dative - case is marked only where there is a distinct morphological form

Contrast

- All common nouns can have an emphatic form,
  e.g. mo theach 'my house'; mo theach*sa* '*my* house'

Derived

- New feature added to distinguish de-nominal from de-verbal verbal nouns

2. VERB								
1. V	**2. Type** m = main	**3. Mood** i = indic. s = subj. m = imper c = cond.	**4. Tense** p = pres. s = past h = past hab f = future g = pres. hab	**5. Person** 1 = first 2 = sec. 3 = third 0 = free	**6. Number** s = sing p = pl.	**7. Gender** n/a	**8. Dependency* d = dependant r = relative n = neg	**9. Contrast* e=emphatic

Type

- m = main - all verbs including substantive verb *bí* which can be referenced by its lemma when necessary

Contrast

- All verbs inflected for person can have an emphatic form, e.g. táim*se* im' chodladh '*I*'m asleep

Dependency

- d = dependant – only suppletive forms are explicitly marked, e.g. bhí versus (go) raibh

3. ADJECTIVE								
1. A	**2. Type** q = qualificator v = verbal	**3. Degree** p = positive c = comparative a = attributive	**4. Gender** f = fem. m = masc.	**5. Number** s = sing p = pl.	**6. Case** c = com. g = gen. v = voc.	**7. Contrast** e=emphatic (contrastive)		

Type

- v = verbal adjective - added to Adjective POS category and removed from Verb category
- No possessive adjectives; mo, do, a etc – see Determiner - possessive

4. PRONOUN								
1. P	**2. Type** p = personal x = reflexive i = indefinite r=prepositional d=demonstrative	**3. Person** 1 = first 2 = sec. 3 = third	**4. Gender** f = fem. m = masc.	**5. Number** s = sing. p = pl.	**6. Case** n/a s= subject only	**7. Posessor** e=emphatic (contrastive)		

Type

- No possessive pronouns; mo, do, a etc – see Determiner - possessive.

5. DETERMINER								
1. D	**2. Type** d = demonstrative p = possessive q = quantifier c = contextual w = interrogative	**3. Person** 1 = first 2 = sec. 3 = third	**4. Gender** f = fem. m = masc.	**5. Number** s = sing p = pl.	**6. Case** n/a	**7. Posessor** n/a		

6. ARTICLE								
1. T	**2. Type** d = definite	**3. Gender** f = fem. m = masc.	**4. Number** s = sing p = pl.	**5. Case** c = com. g = gen.				

7. ADVERB								
1. R	**2. Type** g = general d = direction i = intensifier q = interrogative r = relative t = temporal l = locative	**3. Degree** b = base c = comparative s = superlative	**4. Function** m = modifier s = specifier	**5. Wh-ness** n/a				

8. ADPOSITION								
1. **S**	**2. Type** p = preposition	**3. Formation** p = pronoun c = compound a = with article i=infinitive	**4. Gender** n/a	**5. Number** s = sing p = pl.				

9. CONJUNCTION								
1. **C**	**2. Type** c = coordinate s = subordinative	**3. Ctype** w = with copula q=interrog r=relative	**4. Coord-pos** s=past tense					

Type

- For *a, nach, nár, ar* see Verbal Particle where affirm/neg, and dir/indir rel can more easily be encoded

10. NUMERALS								
1. **M**	**2. Type** c = cardinal o = ordinal n = number r = roman	**3. Gender** n/a	**4. Number** n/a	**5. Case** n/a				

11. INTERJECTION								
1. PoS **I**								

12. UNIQUE MEMBERSHIP CLASS								
1. PoS **U**	**2. Particle Type** a = adverbial r = relative d=degree v = vocative m = numeral c=comparative s=superlative p = patronym o = other	**3. B-Function**						

13. RESIDUALS								
1. PoS **X**	**2 Type** f = foreign s = symbol t = toponym a = acronym d = date x = unknown							

14. PUNCTUATION								
1. PoS **F**	2. Type e = sentence final i = sen. Internal a = quote/par init. z = quote/par fin. b = hyphen/ underscore/ dash u = ! q = ? x = apostrophe							

15. ABBREVIATION								
1. PoS **Y**								

16. COPULA								
1. W (was Vmi)	**2. Tense/Mood** p = pres./fut s = cond/past	**3. Clause Type** i = independant d = dependant r = relative (dir) s = relative (indir)	**4. Mood** i = indic. s = subj. q = interrogative	5. Neg n = neg a = affirm				

Table 16

- New category Copula, whose members were formerly with verbs. This has been created as

- a copula has features not found on verbs:
    - direct and indirect relatives
    - combined mood and tense e.g. conditional-past
    - same form for present and future
- a verb has features not found with the copula
    - person/number/gender
    - emphatic form

17 VERBAL PARTICLE								
1. PoS Q	**2. Type** q=interrogative n=negative a=affirmative x=neg. interog.	**3. Mood** s=subjunct. m=imperative	**4. Tense** s=past (other tenses are unmarked)					

Table 17

- New category created as there is a number of pre-verbal particles which bear the features of type (affirmative, negative, interogative), mood (subjunctive, imperative) and tense (past and non-past).
- Removed from the Unique Membership Class which is for types where only one lexical item exists, e.g. pre-adverbial particle, vocative particle.

Parole Short Tags (POS Only)

Short POS = First 2 Characters of Parole Tags

POS	Description
Aq	Adj - qualificator
Av	verbal adjective
C	conjunction
Cc	coord. conj.
Cs	subbord. conj.
Dd	demonstrative determiner
Di	indefinite determiner
Dp	possessive determiner
Dq	quantifier
Dw	interrogative det.
Fa	quote initial
Fb	hyphen (bar)
Fe	sentence final
Fi	sentence internal
Fp	other punctuation, e.g. brackets etc
Fz	quote final
I	interjection/exclamation
Mc	numeral - aon, dó, trí …
Mn	actual numbers 1, 2, 3
Mo	ordinal
Mr	roman numerals
Nc	common noun
Np	proper noun
Ns	substantive noun (not declined)
Nv	verbal noun
Pd	demonstrative pronoun
Pi	indefinite pron.
Pp	personal pron
Pr	prepositional pronoun
Px	reflexive pronoun
Q	verbal particles
R	adverb
Sp	preposition
Td	article (definite only)
U	unique membership class e.g voc., adv. particles etc
Vm	main verbs
W	copula - is
X	residuals
Y	abbreviation

Appendix B: Finite-State Morphological Feature Tags for Irish

Morphological Feature Tags

The following tables contain the tags used in Irish Finite-State Morphology.

0. General — **1. Noun**

2. Verb — **3. Adjective**

4. Pronoun — **5. Determiner**

6. Article — **7. Adverb**

8. Preposition — **9. Conjunction**

10. Numeral — **11. Interjection**

12. Unique Membership Class — **13. Residuals**

14. Punctuation — **15. Abbreviation**

16. Copula — **17. Verbal Particle**

Table 0. General Tags

Tag	*Description*
+CM	canúint na Mumhan, Munster dialect
+CC	canúint Chonnacht, Connaught dialect
+CU	canúint Uladh, Ulster dialect
+Len	séimhiú (lenition)
+Ecl	urú (eclipsis)
+hPref	prefixed vowel
+Emph	nouns, verbs, adjectives can all take an emphatic suffix
+Guess, +GuessCmpd	any token which is not in the F-S lexicon will receive a guessed analysis

Table 1. Noun Tags

Tag	*Description*
+Noun	noun
+Prop	proper noun
+Pers	proper noun - personal name
+Fam	proper noun - family name
+Place	proper noun - placename
+Verbal	verbal/action noun; mostly de-verbal but some derived from agentive nouns
+Subst	substantive; words functioning as a noun but lacking full paradigm
+Fem	feminine
+Masc	masculine

+Com	common case (nominative/accusative/most datives)
+Gen	genitive case
+Voc	vocative case
+Dat	dative case (where exists)
+Sg	singular in number
+Pl	plural in number
+DefArt	definite noun e.g. preceded by an article
+Idf	indefinite noun e.g. not preceded by an article
+Strong	strong plural (same plural for common, gen. and voc. cases)
+Weak	weak plural (different com, gen, voc plurals)
+Emph	emphasis: *ár dteach**sa*** ‘our house’, *ár bpáirc**se*** ‘our field’
+Len	lenition e.g. after simple prep. e.g. *ar t**h**ír* ‘on land’
+Ecl	eclipsis e.g. after compound prep e.g.*ar an **g**cat* ‘on the cat’
+NStem	verbal noun which is de-nominal rather than de-verbal

Table 2. Verb Tags

Tag	*Description*
+Verb	verb
+1P +2P +3P	first, second and third person
+Auto	autonomous form
+Sg +Pl	singular and plural
+PresInd	present indicative
+PastInd	past indicative
+PastIndDep	past indicative dependant form (irregular verbs)
+PastImp	past imperfect indicative
+FutInd	future indicative
+Cond	conditional
+PresSubj	present subjunctive
+PastSubj	past subjunctive
+Imper	imperative
+Neg	negative form
+Q	interrogative form
+NegQ	negative interrogative form
+Rel	relative

Table 3. Adjective Tags

Tag	*Description*
+Adj	adjective
+Base	base form; a.k.a. positive form
+Comp	comparative and superlative form
+Masc	masculine gender

+Fem	feminine gender
+Com	common case
+Gen	genitive case
+Voc	vocative case
+Sg	singular
+Pl	plural
+Strong	an adj. qualifying a strong plural noun will also have the same plural form in all cases
+Weak	an adj. qualifying a weak plural noun, in the gen.case, is not inflected
+Slender	adj qualifying a plural noun ending in a slender consonant
+NotSlen	adj. qualifying a plural noun ending in a broad consonant or a vowel
+Len	adjectives with nouns which are lenited, e.g. a masc noun after prepositions (e.g. *ag an* 'at the', *ar an* 'on the', *as an* 'out of the' etc.), is either lenited or eclipsed according to preference/dialect.
+Verbal	de-verbal adjective
+Its	intensifier

Table 4. Pronoun Tag

Tag	*Description*
+Pron	pronoun
+Prep	with preposition; e.g. liom 'with me', leat 'with you'
+Emph	emphatic (contrastive) form of personal pronoun
+Ref	reflexive
+Idf	indefinite
+1P +2P +3P	first, second or third person
+Fem	feminine gender
+Masc	masculine gender
+Sg +Pl	singular or plural in number
+VerbSubj	pronoun as verb subject, e.g. *Chuaigh **sí** amach* 'She went out'

Table 5. Determiner Tags

Tag	*Description*
+Det	determiner
+Dem	demonstrative: seo, sin, eile
+Poss	possessive: mo do, a etc.
+Qty	quantifier
+Idf	indefinite quantifier:aon,
+Def	definite quantifier: gach, uile
+1P +2P +3P	first, second or third person
+Fem	feminine gender

+Masc	masculine gender
+Sg +Pl	singular or plural in number

Table 6. Article Tags	
Tag	*Description*
+Art	article
+Def	definite
+Fem	feminine gender
+Gen	genitive case
+Sg	singular
+Pl	plural

Table 7. Adverb Tags	
Tag	*Description*
+Adv	adverb
+Gn	general, e.g. *(go) tapaidh*, quickly, *fadó, fós*
+Its	intensifier, e.g. ***sách*** *tapaidh*, 'fairly quickly'
+Dir	direction: suas, thart
+Q	interrogative, e.g. ***cá*** *bhfuil sé* 'where is it/he'
+Loc	location: anseo, ansin
+Temp	temporal: inniu, anocht

Table 8. Preposition Tags	
Tag	*Description*
+Prep	preposition
+Simp	simple
+Cmpd	compound, e.g. *tar éis*
+Emph	emphatic form of prep pronoun
+Art	with article: *den, sna*
+Rel	with relative: *ina (mbíonn sé)*
+Poss	with possessive, e.g. *ina* 'in his' *inár* 'in our'
+Obj	with object pronoun, e.g. *á (de + a) mbualadh*
+Deg	with degree particle, e.g. *dá (de +a) airde an sliabh...*
+1P +2P +3P	first, second or third person
+Fem	feminine gender
+Masc	masculine gender
+Sg +Pl	singular or plural in number

Table 9. Conjunction Tags	
Tag	*Description*
+Conj	conjunction
+Coord	co-ordinate, e.g. *agus* 'and'
+Subord	subordinate, e.g. *ach* 'but'
+Past	e.g. gur tharla sé

+Cop	copula

Table 10. Numeral Tags

Tag	*Description*
+Num	numeral
+Card	cardinal, e.g. *aon dó trí...* 'one, two, three'
+Ord	ordinal, e.g. *céad dara tríú...* 'first, second, third'
+Pers	personal, e.g. *duine, beirt, triúr* 'one person, two people, three people '
+Rom	roman numerals: iii, IV
+Op	operator; +, -, *, / etc
+Def	form following definite article, e.g. *an t-aon*

Table 11. Interjection Tags

Tag	*Description*
+Itj	interjection, e.g. á 'aah', *faraor* 'unfortunately'

Table 12. Particle Tags (Unique Membership Class)

Tag	*Description*
+Part	particle
+Ad	adverbial, e.g. ***go*** *holc* ' badly'
+Nm	numeral, e.g. ***a*** *haon* 'one'
+Comp	comparative degree, e.g. ***níos*** *fearr* 'better'
+Pat	patronym, e.g. ***Ó*** *Beirn*, ***Ní*** *Bheirn*, ***Uí*** Bheirn
+Voc	vocative particle, e.g. ***a*** *Mháire* 'Mary!'
+Deg	degree particle, e.g. ***a*** *géire a labhair sé* 'how sharply he spoke'
+Cp	copular particle
+Cmpl	complementizer, ***go*** *ndéanfadh sé é* 'that he would do it'

Table 13. Residuals Tags

Tag	*Description*
+Foreign	foreign words
+Dig	digits, e.g. 123,000 10.12
+Cur	currency symbols
+PC	per cent sign e.g. %
+Item	list item e.g. a) iv) (3)
+Time	am pm
+Email	e-mail addresses
+Web	website addresses

Table 14. Punctuation Tags

Tag	*Description*
+Fin	sentence final punctuation, e.g. !?.
+Q	question mark i.e. ?
+Int	sentence internal punctuation, e.g. ,;:()
+Quo	quotation marks, e.g. ' "
+Bar	hyphen, underscore e.g. - _

Table 15. Abbreviation Tags	
Tag	*Description*
+Abr	abbreviation, e.g. ***lch.*** *(leathanach)* ‘page’

Table 16. Copula Tags	
Tag	*Description*
+Cop	verb
+Sg +Pl	singular and plural
+Pres	present / future
+PresSubj	present subjunctive
+Past	past / conditional
+Dep	dependant clause
+Neg	negative form
+Cop	copula
+Q	interrogative form
+NegQ	negative interrogative form
+Rel	relative (direct)
+RelInd	relative indirect
+VF	form before vowel or f word e.g. ab (fhusa)
+Pro	with pronoun, e.g. *sea (is + ea), sé (is + é), sí (is + í)*

Table 17. Verbal Particle Tags	
Tag	*Description*
+Part	particle
+Vb	verbal particle
+Neg	negative, e.g. ***ní*** *raibh* ‘was not’
+Q	interrogative verbal particle, e.g. ***an*** *raibh* ‘was?’
+Subj	subjunctive, e.g. ***go*** *raibh maith agat* ‘thank you’
+Imp	imperative, e.g. ***ná*** *déan*, ‘don’t do it’
+Past	past tense verbal particle, e.g. ***an*** *raibh sé* ‘was he?'***ar*** *chuala sé* ‘did he hear’
+Fut	future tense, e.g. ***an*** *mbeidh tú ann*? ‘will you be there?’
+Pres	present tense, e.g. ***an*** *bhfuil tú ann?* ‘are you there’
+Cond	conditional, e.g. ***má*** *bhíonn tú ann* ‘if you would be there’
+Rel	relative particle, *a, ar*
+Direct	direct relative, e.g. *an fear* ***a*** *bhíonn tinn* ‘the man who is (habitually) sick’
+Pro	relative particle with pronoun, e.g. *gach* ***a*** *tharla* ‘all that which happened’

Appendix C: Guidelines for Manual POS Disambiguation

Table of Contents

1. Nouns 2
2. Verbs 5
3. Adjectives 6
4. Pronouns 8
5. Determiners 9
6. Articles 9
7. Adverbs 10
8. Prepositions 10
9. Conjunction 11
10. Numerals 12
11. Copula 12
12. Verbal Particle 13
13. Notes on Common Ambiguous Lexical Items 13

The following guidelines are intended to aid manual disambiguation in cases where the choice of POS in a particular context is not obvious. Recommendations in the text below are highlighted with an arrow as follows ⇨ .

1. Nouns

Four types of noun are distinguished in the Finite-State Morphology (FSM):

- common (`Noun`)
- proper (`Prop Noun`)
- substantive (`Subst Noun`)
- verbal (`Verbal Noun`)

Common Nouns

If it is not clear whether or not a token is functioning as a common noun:

⇨ Check whether it can be used with the definite article in the context in which it is found

⇨ Check whether it can be modified by an adjective in the context in which it is found, e.g. whether or not *brí* 'meaning' functioning a noun in (1), (Stenson, 1981, p63)

(1) de bhrí nach raibh fhios aige ...
of meaning NEG was knowledge at-him
'because he didn't know ...'

(2) *de bhrí maith nach raibh fhios aige ...
of reason good NEG was knowledge at-him

The fact that *brí* 'meaning' cannot be modified by an adjective such as maith 'good' in (2) suggests that *de bhrí* is an idiom and should be handled as a multi-word expression (MWE).

⇨ Use `DefArt` rather than `Len` form of the noun whenever it follows the definite article e.g. *an chathaoir* 'the chair'

(3) "<an>"
"an" Art Sg Def
"<c**h**athaoir>"
"cathaoir" Noun **Fem** Com Sg **DefArt**

Verbal Nouns

Most verbal nouns are derived or are related (semantically) to a verb. The verbal noun has same transitivity as its corresponding verb, as in (4) and (5). A few verbal noun nouns are derived from agentive nouns, e.g. (6).

(4) déan (V transitive)
'make/do'
ag déanamh (VN) cáca
at making a-cake
'making a cake'

(5) fan (V intransitive)
'stay'
ag fanacht (VN)
at staying
'staying'

(6) siopadóir (agent N)
'shopkeeper'
ag siopadóireacht (VN)
at shopping
'shopping'

Verbal Nouns are commonly used with a preposition indicating aspect in non-finite phrases, and are always accompanied by a auxiliary verb in the sentence.

(7) **a choinneáil ag sodarnaíl**
keeping(VN) at trotting (VN)
'continuing to trot'

(8) a_lán **dul** chun cinn
a lot going (VN) to head
'a lot of progress' (headway)

(9) Tá sé **ina** **chodladh**
Is he in-his sleeping (VN)
'He is asleep'

⇨ In all cases the verbal noun is treated as noun. It may of type de-verbal or de-nominal.

⇨ For the purposes of POS tagging we do not try to distinguish between a preposition used locatively from the same preposition used aspectually. Both are tagged `Prep Simp` or `Prep Cmpd`, as the case may be.

⇨ For consistency reasons also, "*a"* before a verbal noun is tagged as a preposition, (except with an infinitival uses of the verbal noun, where it is tagged `Part Inf`) although it currently has no non-aspectual prepositional use, unlike the other prepositions used aspectually. (However, there is evidence that the preposition *do* became *a* before verbal nouns (Williams, 1994, p461).

The various functions of the verbal noun in non-finite clauses such as the progressive and infinitive are not distinguished at the POS level. This will be handled at the phrasal level only.

Verbal Noun vs. Common Noun Ambiguity

We rely on the type of noun (common or verbal) and the context in which it is used to make the distinction between locative and aspectual or infinitival use. For this reason it is important to resolve verbal noun and common noun ambiguity.

All verbal nouns, as well as taking part in non-finite clauses, can function as common nouns. This can lead to ambiguity, as it can be difficult to distinguish between locative prepositional phrases and non-finite phrases, e.g. (10) could be interpreted as 'He was dancing' or 'He was at a dance' as there is no indefinite article in Irish.

(10) Bhí sé **ag damhsa**
Was he at dancing(VN)/a-dance (N)
'He was dancing' OR 'He was at a dance'

The same ambiguity applies to the copular version (11) where 'dancing' has been fronted for emphasis. In this case it could be interpreted as 'It's dancing he was' or 'It's at a dance he was'.

(11) Is **ag damhsa** a bhí sé
COP at dancing (VN)/a-dance (N) that was he
'It's dancing he was'

However, in practice, the verbal noun when used as a common noun usually occurs with the definite article as in (12).

(12) Bhí sé ag **an damhsa**
Was he at the dance (N)
'He was at the dance'

⇨ Therefore, we choose the verbal noun reading whenever it occurs immediately after a preposition, as in examples (10) and (11).

⇨ We choose the common noun reading only when it is preceded by an article as in (12).

⇨ Note that the verbal noun reading should be chosen after a possessive determiner (13) following a preposition.

(13) Tá sé do **mo** chabhrú
Is he to my helping-VN
'He is helping me'

2. Verbs

Ó hUallacháin and Ó Murchú (1981, p10) distinguish 4 types of verb:

- copula "is" (`Cop`)
- substantive verb *bí* (`Verb VI`)
- intransitive verbs (`Verb VI, Verb VTI`)
- transitive verbs (`Verb VT, Verb VTI, Verb VD`)

In this implementation, the copula is not tagged as a verb as it takes part in different syntactic constructions and has different morphological features (see Section 11 and Appendix A). No distinction is made between the substantive verb *bí* 'to be' and other verbs, all of which are tagged as `Verb`, along with an additional tag indicating transitivity, as shown below.

- VI (intransitive)
- VT (transitive)

- VTI (transitive and intransitive) or
- VD (ditransitive).

Many synthetic forms of verbs listed in (Dillon and Cróinín, 1961) which are not part of the current spelling standard (Rannóg an Aistriúcháin, 1958) and Graiméar Gaeilge na mBráithre Críostaí (An Gúm, 1999) are included in the finite-state morphology as they are to be found in older texts.

(14) "<bhfuilir>"
"bí" Verb PresInd **2P Sg Dep Ecl CM**

3. Adjectives

For consistent POS tagging, adjectives are noun post-modifiers only. Therefore *mo, do* (my, your) etc. which precede the noun, are tagged as possessive determiners rather than possessive adjectives (Christian Brothers, 1988, p82) or possessive pronouns (Doyle, 2001: p69).

⇨ In texts, adjectives can follow any one of the following categories: noun, pronoun, adjective, conjunction, copula or punctuation (e.g. comma).

⇨ Attributive adjectives should agree with their noun for gender case and number. Whenever a choice exists, choose the adjective with matching feature tags e.g. *leigheas iomlán* 'total cure', otherwise use `Adj Base` e.g. *cat buí* 'yellow cat'.

(15) leigheas Noun **Masc Com Sg**
iomlán Adj **Masc Com Sg**

⇨ Note that adjectives ending in a vowel are always tagged as `Adj Base` as they have no distinct inflected forms for gender, case or number.

(16) cat Noun Masc Com Sg
buí Adj Base

Attributive adjectives are inflected for gender, case and number to agree with the noun which they modify (17). Predicative adjectives are not inflected (18). No distinction in function is made at the POS level.

(17) Chuaigh an bhean **bheag** amach
Went the woman small out
'The small woman went out'

(18) Tá an bhean **beag**
Is the woman small
'The woman is small'

⇨ Predicative adjectives should be tagged as `Adj Base` or `Adj Base Len` when initial lenition is present.

Verbal Adjectives

Every verb has an associated de-verbal adjective. As with non-deverbal adjectives they are used both attributively (19) and predicatively (20).

(19) Tá an chathaoir **bhriste** agam
Is the chair broken (ATTR) at-me
'I have the broken chair'

(20) Tá an chathaoir **briste** agam
Is the chair broken (PRED) at-me
'I have broken the chair'

Verbal Adjective vs. Verbal Noun Genitive Case Ambiguity

The verbal adjective and the verbal noun in the genitive case share the same form, b) & e) of (21) to (23), except in the case of borrowed verbs ending in *–áil* (24),

(21) a) Verb: imir - to play
b) Verbal Adjective: **imeartha** - played
c) Noun: an imirt – the playing
d) Verbal Noun: ag imirt – playing
e) Verbal Noun Gen.: páirc **imeartha** – playing field

(22) a) Verb: bris - to break
b) Verbal Adjective: **briste** - broken
c) Noun: an briseadh – the break
d) Verbal Noun: ag briseadh – breaking
e) Verbal Noun Gen.: I rith an **bhriste** – during the break

(23) a) Verb: cláraigh - to register
b) Verbal Adjective: **cláraithe** - registered
c) Noun: an clárú – the registration
d) Verbal Noun: ag clárú – registering
e) Verbal Noun Gen.: lá an **chláraithe** – the day of registration

(24) a) Verb: pleanáil – to plan
b) Verbal Adjective: an t-aistriú **pleanáilte** – the planned transfer
c) Noun: an pleanáil – the planning
d) Verbal Noun: ag pleanáil – planning
e) Verbal Noun Gen. : An Bord **Pleanála** – The Planning Board

In order to decide whether the form is functioning as a verbal adjective or a verbal noun in the genitive case, we propose the following guidelines:

⇨ If the head (modified) noun undergoes the action, i.e. has a patient role, the modifier is a verbal adjective, e.g. (25).

(25) Na Stáit Aontaithe
the states (Patient) united (VA)
'The United States'

(26) Ar dhéanmhais chosanta
on structures (Patient) protected (VA)
'on protected structures'

⇨ If the head (modified) noun is the agent or facilitator of the action, then the modifier is a verbal noun in the genitive case, e.g. (27) and (28).

(27) Cailín deas crúite na mbó
girl (Agent) nice milking (VNg) the cows
'pretty milking maid'

(28) páirc imeartha
field (Facilitator) playing (VNg)
'playing field'

(29) Binse Fiosraithe
Board (Facilitator) Investigating (VNg)
'Investigating Board' i.e. 'Tribunal'

⇨ If the modifying noun is clearly functioning as a common noun in genitival noun phrases, i.e. is preceded by a determiner *an* 'the', e.g. part e) of (21) and (23) repeated below, then the modifier is a verbal noun in the genitive case.

(30) I rith an bhriste
in run the breaking (VNg)
'during the break'

(31) lá an chláraithe
day the registering (VNg)
'the day of registration'

4. Pronouns

The following types of pronoun are encoded:

- Personal (Pron Pers): *mé* 'me', *tú* 'you', *sí* 'she', *sé* 'he', *sibh* 'you' (pl.), *siad* 'them'

- Reflexive `(Pron Ref)`: *féin* 'self'
- Indefinite `(Pron Idf)`: *ceachtar* 'either', *cibé* 'whoever'
- Interrogative `(Pron Q)`: *cad* 'what', *cé* 'who', *cén* 'which one' etc.
- Demonstrative `(Pron Dem)`: *seo* 'this', *sin* 'that' etc.
- Prepositional `(Pron Prep)`: *agam* 'at me', *ort* 'on you', *leo* 'with them' etc.

⇨ A pronoun can substitute for a noun phrase (NP) and cannot co-occur with a definite article.

Prepositional Pronoun

The class of prepositional pronouns (or conjugated prepositions) are classified under pronoun, e.g. *di* (to her), as number/person features are encoded for pronouns but not prepositions in the Parole tagset (Appendix A).

```
(32) "<di>"
          "do" Pron Prep 3P Sg Fem
```

5. Determiners

The following types of determiner are encoded:

- Possessive (`Det Poss`): *mo* 'my', *do* 'your', *a* 'hers, his, theirs', *bhur* 'your', *ár* 'our'
- Indefinite Quantifier (`Det Qty Idf`): *aon* 'any', *cibé* 'whichever' etc.
- Definite Quantifier (`Det Qty Def`): *gach* 'every', *uile* 'each' etc.
- Demonstrative (`Det Dem`): *seo* 'this', *sin* 'that', *úd* 'those' etc.

⇨ A determiner cannot co-occur with an article

⇨ A determiner must qualify a noun. All determiners except demonstratives must precede the noun.

⇨ Demonstrative post-determiners, *seo* 'this', *sin* 'that', *úd* 'those' etc. can only occur with a pronoun or a definite noun, i.e. the noun must be preceded by either an article or a possessive determiner, or an article incorporated into a preposition, e.g. *ina* 'in-his/her/their'.

6. Articles

The singular and plural definite articles *an* and *na* respectively, as well as the article which precedes a feminine genitive noun are tagged as shown below. There is no indefinite article.

- `Art Sg Def` *an*
- `Art Sg Def Fem Gen` *na*
- `Art+Pl+Def` *na*

7. Adverbs

The following types of adverb are encoded in the lexicon.

- Manner (`Adv Gn`):, e.g. *fós* 'yet', *déanach* 'late'
- Directional (`Adv Dir`): e.g. suas 'upwards'*, timpeall* 'around'
- Locative (`Adv Loc`):, e.g. *anseo* 'here'*, laistigh* 'within' *thuas* 'above'
- Temporal (`Adv Time`):, e.g. *inniu* 'today'*, aréir* 'last night'*, istoíche* 'tonight'
- Interrogative (`Adv Q`):, e.g. *cá* 'where'*, cathain* 'when'*, conas* 'how'
- Intensifiers (`Adv Its`):, e.g. *iontach* 'wonderful'*, measartha* 'middling', sách 'fairly'

⇨ Many of these forms are used adjectivally but we tag them in all cases as adverbs to avoid unnecessary duplication in the lexicon.

```
(33) Chonaic mé é maidin  inniu
     saw     I it morning today (ADV)
     'I saw it this morning'
```

Adverbial Particle

⇨ The particle *go* with an adjective is used to form an adverb, which is equivalent to the -ly class of adverbs in English.

```
(34) go    híontach
     PART. wonderful (A)
     'wonderful(ly)'
```

8. Prepositions

Types of preposition encoded:

- simple, (`Prep Simp`): e.g. *ar* 'on' *ag* 'at' *faoi* 'under', *i* 'in'
- compound (`Prep Cmpd`): e.g. *i measc* 'among', *go dtí* 'as far as'
- with article (`Prep Art`): *san* 'in the'
- with possessive determiner (`Prep Poss`): *lena* 'with his/her/their'
- with degree particle (`Prep Deg`): *dá airde* (of height)

⇨ A preposition may only precede a NP (noun, verbal noun, pronoun, determiner, number, abbreviation).

```
(35) gan         moill
     without (P) delay (N)
```

(36) seachas cinntí
except (P) decisions (N)

(37) ar airde
on (P) height (N)
'high' OR 'in height'

⇨ Note that "go" before an adjective is tagged as an adverbial particle - not a preposition.

9. Conjunction

⇨ Co-ordinating conjunctions can join a word, phrase or sentence.

Two types of conjunction are encoded:

- coordinating, (`Conj Coord`): e.g. *agus* 'and', *nó* 'or'.
- subordinating (`Conj Subord`): e.g. *ó* 'since', *ach* 'but', *má* 'if', *dá* 'if'

There are several compound subordinating conjunctions, all of which are tagged with the `Conj Subord` tags:

(38) más
má is (COP)

(39) ós
ó is (COP)

(40) mura
mur a (REL)

(41) sula
sul (REL)

(42) dá
do (REL)

There are a number of multi-word expressions (MWE) which are also tagged with the `Conj Subord tags`:

(43) nuair a
'when'

(44) go dtí go
'until'

(45) cé go
'even though'

(46) cé is móite
'however'

Although an item such as *má* 'if', occur only before verbs, we have not tagged it as a verbal particle as it also combines with copula *is* to form más.

(47) má bhíonn sé
if is it
'if it is'

(48) más rud é
if-COP thing it
'if it's a fact that'

10. Numerals

The following types of numbers are encoded:

- Cardinal (Num+Card): *aon* 'one'
- Ordinal (Num+Ord): *céad* 'first'
- Operator (Num+Op): +/-
- Digit (Num+Dig): 1, 2, 3
- Roman Numeral (Num+Rom): i, ii, iv
- Currency (Num+Cur): €1,000,000
- Per cent (Num+PC): 100%

11. Copula

⇨ Copula can be followed by a noun, pronoun, or adjective predicate (Ó Dónaill, 1977).

(49) Is fear maith é
COP man good he
'He is a good man'

(50) Is mise Briain
COP I Briain
'I am Briain'

(51) Is maith liom é
COP good with-me it
'I like it'

⇨ When the copula follows a fronted noun predicate it is followed by the 3rd. person singular neuter pronoun *ea*.

(52) Fear maith is ea é
Man good COP PRON he
'A good man is what he is'

⇨ The copula is always used before *féidir* 'possible'

(53) is féidir an rialtas a athhrú
COP possible the government to change
'it is possible to change the government'

12. Verbal Particle

⇨ A verbal particle (Part Vb) may only precede a verb.

(54) ní raibh
NEG-PART was
'was not'

(55) ní féidir
COP possible
'not possible'

13. Notes on Common Ambiguous Lexical Items

Seo 'this', *sin* 'that', *siúd* 'those' (demonstrative pronoun vs. demonstrative determiner)

⇨ If a pronoun is removed it will render the sentence syntactically or semantically ungrammatical.

(56) Rinne sé **sin**
Did he that
'He did that '

(57) *Rinne sé
Did he

⇨ A demonstrative determiner is optional; if removed the sentence will still be syntactically and semantically complete

(58) D'fhág sé **an teach sin**
Left he the house that
'He left that house'

(59) D'fhág sé an teach
Left he the house
'He left the house'

⇨ A Demonstrative Determiner can be found following a noun, pronoun or prepositional pronoun:

- Noun
- with Definite Article : an fear sin ‘that man’
- with Possessive Determiner: mo theach seo ‘this house of mine’
- with Preposition and Possessive Pronoun: lena charr sin ‘with that car of his’;
- Pronoun

(60) tabhair dom **é sin**
give to-me it that
'give me that';

(61) ní **hé sin** amháin
not it that only
‘not only that’

- Prepositional Pronoun

(62) ach fiú leis sin
but even with that’

go: conjunction, proposition, adverbial particle

⇨ Before a Verb *go* is `Part Vb`, and when it comes before a Subjunctive verb, it is a `Part Vb Subj`, e.g. go raibh maith agat.

⇨ Before a Noun *go* should be tagged as a preposition (`Prep Simp`).

⇨ Before an Adjective *go* should be tagged an adverbial particle (`Part Ad`).

Mac, Ó, Ní, Uí, de as Particles

Before a surname, these are tagged as patronymic particles i.e. `Part Pat`.

a as preposition

Before an aspectual verbal noun *a* is tagged as a preposition (`Prep Simp`)

(63) an rud a bhí sé **a dhéanamh**
the thing that was he at doing
'the thing he was **doing**'

a as infinitival particle

Before infinitival verbal noun *a* is tagged as a preposition (`Part Inf`)

(64) Ba mhaith liom é **a dhéanamh**
Is good with-me it to do
'i would like **to do** it'

Appendix D: CG POS Disambiguation Rules for Irish

Listing of CG POS Disambiguation Rules for Irish

```
# ============================================================== #
# I R I S H  P O S  D I S A M B I G U A T I O N
# CONSTRAINT GRAMMAR CG2
# ============================================================== #
# Elaine Uí Dhonnchadha 2008
# ============================================================== #
# "ar" => LEMMA, "<ar>" => WORDFORM
# ============================================================== #
# SENTENCE DELIMITERS
# ============================================================== #
DELIMITERS = "<.>" "<!>" "<?>" "<#>" "<</p>>" "<</s>>" "<...>";
PREFERRED-TARGETS = Pron Noun PastInd PresSubj ;
# ============================================================== #
# SETS
# ============================================================== #
SETS
LIST BOS = (>>>) "<p>" "<s>" (ChildesID);
LIST EOS = (<<<); # end and beg. of sentence. for vislcg.
LIST COMMA = "<,>" ;
LIST PUNCT-INT = (Punct Int) (Punct Bar) (Punct Brack);
LIST CLB = (Rel) (Coord) (Subord) (Cmpl) ;
SETS
# attributive adj. set
LIST ADJ-ATTR = (Adj Com) (Adj Gen) (Adj Voc) ;
LIST ADJ-NOT-VA = (Adj Com) (Adj Gen) (Adj Voc) (Adj Base) (Guess
Adj) ;
# adjectives follow nouns but the following is
# a list of the few adjectives can precede a noun
LIST ADJ-PRENOM = "droch" "sean" "príomh" "fíor" "iontach" "dearg"
"leath" "corr" "gnáth" "mór";

# Any noun other than verbal-noun
# the are several types of Noun: +Noun, Subst+Noun, Prop+Noun,
Verbal+Noun,
# Guess+Noun, but all nouns except verbal nouns have number (even
guess nouns)
LIST NOUN-NOT-VN = (Noun Sg) (Noun Pl) ;

# a list of items which can precede a noun
LIST NOUN-PREMOD = (Art) (Det Poss) (Det Qty) (Num) ADJ-PRENOM ;

# a list of items which can follow a simple preposition
# (art def is used to exclude "sa" e.g. "shuigh sé faoi sa
chathaoir"
# rel clause: an rud as ar/Part Vb Rel(not Cop) tháinig
# thar/Prep a/Prep bheith/VNoun
# mar iad/Pron Pers
LIST POST-PREP = (Noun) (Art Def) (Det) (Pron) (Num) (Part Nm) ADJ-
PRENOM (Part Vb Rel) (Prep Simp) (Punct Quo);

# "a" functions as a simple prep in following phrases
LIST A-PREP-PHR = "chlog" "chois" "chóir" "dhíth" ;

# the genitive follows some simple prepositions and partitives,  as
well as another noun, verbal noun or compound preposition
LIST GEN-SIMP-PREP = "chun" "trasna" "timpeall" "fearacht" "dála"
"cois" ;
LIST GEN-PART = "roinnt" "cuid" "morán" "lán" "méid" "dosaen"
"péire" "scór" ;
```

```
# lemmas include emphatic forms
LIST OBJ-PRON = "í" "é" "iad" ;

# wordform rather than lemma is used as we do not want to include
thrí or dtrí etc.
LIST NUM-COUNT = "<haon>" "<dó>" "<trí>" "<ceathair>" "<cúig>"
"<sé>" "<seacht>" "<hocht>" "<naoi>" "<deich>" "<hAon>" "<Dó>"
"<Trí>" "<Ceathair>" "<Cúig>" "<Sé>" "<Seacht>" "<hOcht>" "<Naoi>"
"<Deich>";

LIST NUM-LEN = "aon" "<chéad>" "<dhá>" "trí" "<ceithre>" "cúig" "sé"
"beirt" "uile";
LIST NUM-ECL = "seacht" "ocht" "naoi" "deich" ;
LIST NUM-PL-ADJ = "<dhá>" "trí" "ceithre" "cúig" "sé" "seacht"
"ocht" "naoi" "deich" "beirt" ;

# prepositions commonly used before verbal nouns
# "ar" => lemma, "<ar>" => wordform
LIST PREP-VN = "<ag>" "<le>" "<gan>" ("<a>" Part Inf) "<á>" ("<ar>"
Prep) "tar éis" "chun" "le" "i ndiaidh" "ar tí" "roimh" "<ina>";

# this type of verbal noun can be modified by attributive adj.
# e.g. "ag mothú tinn" but not "ag déanamh mór", "a bheith tanaí"
LIST SENSORY-VN = "<bheith>" ("mothú" Verbal) ("breathnú" Verbal)
("fáil" Verbal) ("aireachtáil" Verbal) "<éirí>";

# titles are nouns but often dont have gen. case on following noun
# some do ... e.g a thiarna easpaig
# e.g. an tUrramhach James, ár dTiarna Íosa
LIST TITLE = "urramach" "bantiarna" "tiarna" "uasal" ;
LIST DAYS = "Luan" "Máirt" "Céadaoin" "Déardaoin" "Aoine" "Satharn"
"Domhnach" ;
LIST TIME = "mí" "bliain" "lá" "ráithe" "uair" "seachtain";
LIST TIME-PERIOD = "linn" "feadh" ;
LIST MEASURE = "slat" "orlach" "míle";
LIST THING = "ceann" "rud";

LIST VERB-WITH-SUBJ = (Verb 1P) (Verb 2P) (Verb 3P) (Verb Auto) ;

LIST PREP-ECL = "<i>" ; # etc. etc.
LIST PREP-LEN = "<de>" "<do>"; # etc. etc.

# ================================================================ #
# DISAMBIGUATION RULES
# ================================================================ #
# ================================================================ #
# SECTION 1 - Definite Rules
CONSTRAINTS
# ================================================================ #
# ---------------------------------------------------------------- #
# S1 IDIOMS
# ---------------------------------------------------------------- #
# maille le
SELECT (Prep) IF (0 ("<maille>")) (1 ("<le>"));

# as seo/sin amach
SELECT (Pron Dem) IF (-1 ("as")) (1 ("amach"));
SELECT (Prep Simp) IF (1 (Pron Dem)) (2 ("amach"));
# amach = Adv Dir not Adj
#SELECT (Adj) IF (-2 ("as")) (-1 (Adj));

#sé = Noun only in phrase "sé nó seachrán" or after Art
REMOVE (Noun) IF (0 ("<sé>")) (NOT 2 ("seachrán"));
# ann = Noun only in "in ann"
```

```
REMOVE (Noun) IF (0 ("ann")) (NOT -1 ("<in>")) ;
SELECT (Noun) IF (0 ("ann")) (-1 ("<in>")) ;

# ar/agus/ná a chumas/cumas/gcumas
SELECT (Noun Com) IF (0 ("cumas")) (-2 ("ar") OR ("agus") OR ("ná"))
(-1 ("a"));

# mar gheall ar/air
SELECT (Noun Com) IF (0 ("geall")) (1 ("ar")) (-1 ("mar"));
SELECT (Prep Simp) IF (1 ("geall")) (2 ("ar")) (0 ("mar"));

# ar chor/cor ar bith
SELECT (Noun Sg) IF (0 ("cor")) (1 ("ar")) (2 ("bith"));
# aon chor/cor
SELECT (Noun Sg) IF (0 ("cor")) (-1 ("aon"));
SELECT (Det Qty) (1 ("cor")) (0 ("aon"));

# ar mhaithe le
SELECT (Noun Sg Len) IF (0 ("maithe")) (-1C ("<ar>")) (1 ("le"));

# cé as a conjunction
SELECT (Conj) IF (0 ("cé")) (1 ("go") OR ("gur") OR ("nach") OR
("nár"));

# le cúnamh/cuidiú Dé
REMOVE (Verbal Noun) IF (0 ("cúnamh") OR ("cuidiú")) (1 ("<Dé>"));

# let us assume that ba = bó(pl) must have a pl article preceding it
# not strictly true but ba=cop in most cases ...
REMOVE (Noun) IF (0 ("<ba>") OR ("<Ba>")) (NOT -1 (Art Pl));

# go deimhin is usually adverbial; certainly
SELECT (Part Ad) IF (0 ("go")) (1 ("<deimhin>"));
SELECT (Adj Base) IF (-1 ("go")) (0 ("<deimhin>"));

# dar as a verb must be followed by prep "le", i.e. dar le =
according to
# dar/V leis/léi/liom ,
SELECT (Verb) IF (0 ("<dar>") OR ("<Dar>")) (1 ("le"));
REMOVE (Verb) IF (0 ("<dar>") OR ("<Dar>")) (NOT 1 ("le"));

# dar as verb particle only occurs before a PastInd verb
# dar/Q tháinig/V-PastInd
# but dar/!=Q díobh/V-Imper OR dar/!=Q dtaltaí/V-Ecl
REMOVE (Verb) IF (-1 ("<dar>") OR ("<Dar>")) (NOT 0 (Verb PastInd));
REMOVE (Verb Ecl) IF (-1 ("<dar>") OR ("<Dar>"));
REMOVE (Part Vb) IF (0 ("<dar>") OR ("<Dar>")) (NOT 1 (Verb
PastInd));
SELECT (Part Vb) IF (0 ("<dar>") OR ("<Dar>")) (1 (Verb PastInd));

# dar as a copula occurs with prep pron "de" i.e dar díobh = was of
them
SELECT (Conj Cop) IF (0 ("<dar>") OR ("<Dar>")) (1 ("de" Pron
Prep));
# dar as copula occurs with Adj e.g. dar léir siúd
SELECT (Conj Cop) IF (0 ("<dar>") OR ("<Dar>")) (1C (Adj));
# dar as copula occurs with go + Adj e.g. dar go deimhin
SELECT (Conj Cop) IF (0 ("<dar>") OR ("<Dar>")) (1 (Part Ad)) (2
(Adj));

# dar mo chonsias...dar fia ... dar Crom ... dar an leabhair ...
# before all else it must ne a prep simp ...
SELECT (Prep Simp) IF (0 ("<dar>") OR ("<Dar>")) (1C (Noun) OR NOUN-
PREMOD);
```

```
# ar sise/seisean/
# dont use Prop Noun as too many many Prep + Placenames/Orgn.
SELECT (Verb) IF (0 ("<ar>") OR ("<Ar>")) (1 (Sbj));
REMOVE (Verb) IF (0 ("<ar>") OR ("<Ar>")) (NOT 1 (Sbj) OR ("mé"));

# a bheag nó/ná a mhór
SELECT (Det Poss) (1 ("mór")) (0 ("a"));
SELECT (Noun) (0 ("mór")) (-1 ("a"));

# Cop: is/ní dóigh
# Prep: ar/sa/ón ndóigh/dhóigh/dóigh
# Conj: ach/agus dóigh
SELECT (Noun) IF (0 ("<dóigh>") OR ("<dhóigh>") OR ("<ndóigh>"))
(NOT 1 (Sbj)) (-1 (Cop) OR (Prep) OR (Conj));

# some De Names commonly found in the corpus ...
SELECT (Part Pat) (0 ("de")) (1 ("<Valera>") OR ("<Brún>") OR
("<hÍde>") OR ("<Niro>") OR ("<Róiste>") OR ("<Cléir>"));

# TYPE 1
# "ann" is only subst noun in phrase "in ann"
REMOVE (Subst Noun) IF (0 ("ann")) (NOT -1 ("i"));
SELECT (Subst Noun) IF (0 ("ann")) (-1 ("i"));

# "go bhfuil" is invariably verbal ...
SELECT (Verb PresInd Dep Ecl) (0 ("<bhfuil>")) (-1 ("<go>"));

# "rinne" is usually verbal ...
# an éileamh a rinne Hitler ...
SELECT (Verb) (0 ("<rinne>")) (-1 ("<a>")) (1C (Noun) OR NOUN-
PREMOD);
SELECT (Verb) (0 ("<Rinne>")) (-1 BOS);

# "cruth an duille, leagan/!=VN amach na mbláthanna
# <leagan amach> to be made into MWE (multi word expression)
REMOVE (Verbal Noun) IF (0 ("<leagan>")) (1 ("<amach>"));

# a deir/!=noun ... usually
SELECT (Verb) (0 ("abair")) (-1 ("<a>"));

# the verb reading will only be removed for ambiguous items
# genuine unambiguous verbs will survive this rule
# e.g. " an cuid atá/V
# but " cuid mhaith/!=V
REMOVE (Verb) IF (-1 ("<cuid>"));

# idiom scun scan/!=V
SELECT (Subst) IF (0 ("<scan>")) (-1 ("<scun>"));

# i gcomhair, faoi chomhair, os comhair
SELECT (Subst) IF (0 ("comhair")) (-1 (Prep Simp));

# os ár comhair
SELECT (Subst) IF (0 ("comhair")) (-2 (Prep Simp)) (-1 (Det Poss));

# dála an scéil
SELECT (Subst) IF (0 ("dála")) (2 ("<scéil>"));

# mo/do/a dhála féin, ár ndála féin
SELECT (Subst) IF (0 ("dála")) (-1 (Det Poss));

# mac léinn: Mac is not patronymic (part of name) if followed by
"léinn"
```

```
SELECT (Noun Com) IF (0 ("mac")) (1 ("<léinn>"));

# mac léinn: "léinn" is not a verb if it follows "mac"
SELECT (Noun Gen) IF (-1 ("mac")) (0 ("<léinn>"));

# leath uair=Com a chloig,
# vs. leath an ama=Gen
SELECT (Det) IF (0 ("leath")) (1 (Noun Com));

# ar mhaith leat; is maith liom etc.
SELECT (Adj) IF (0 ("maith")) (-1 (Cop)) (1 ("le")) ;
SELECT (Cop) IF  (1 ("maith")) (2 ("le")) ;

# ================================================================ #
# ---------------------------------------------------------------- #
# S1 VERB
# ---------------------------------------------------------------- #

SELECT (Verb) IF (1 (Sbj));
REMOVE (Verb Auto) IF (1C (Sbj));

# TYPE 1
# it is a negative form of verb if it is preceded by a neg verb part
SELECT (Verb Neg) IF (-1C (Part Vb Neg));

# TYPE 1
# it is not a negative form of verb if it is not preceded by a neg
verb part
# unless it is "níl"
REMOVE (Verb Neg) IF (NOT -1 (Part Vb Neg)) (NOT 0 ("<níl>") OR
("<Níl>"));

# neg particle for pres subj is nár or ná, i.e. a neg particle like
ní implies that it is not pres subj
# note will deal with "ar mhaithe le" where "mhaithe" could be
mistaken for a verb PresSubj rather than noun
REMOVE (Verb PresSubj Neg) IF ( NOT -1 ("ná") OR ("nár")) ;
REMOVE (Verb PresSubj) IF ( NOT -1 ("ná") OR ("go")) ;

# it is a verb if it is followed by a pronoun form tagged as verb
subject
# it is an interrogative form of verb if it is preceded by interr
adv e.g. cá
# an bhfuil, an raibh, ar chuala
SELECT (Verb Q) IF (-1C (Adv Q) OR (Part Q));
SELECT (Verb NegQ) IF (-1C (Part NegQ));

# it is not a neg. interog. form of verb if it is not preceded
# by a neg interog. verb part
REMOVE (Verb NegQ) IF (NOT -1 (Part Vb NegQ));

# it is not an  interog. form of verb if it is not preceded
# by a interog. verb part / interog. adv etc.
REMOVE (Verb Q) IF (NOT -1 (Q));

# e.g. rith sé amach: sé=verbsubj pron => rith = verb
# NOTE that rith can not be the imperative form
REMOVE (Verb Imper) IF (1 (Sbj));   # 1 SEP

# neg particle for imperative form is "ná", not "níor" etc.
REMOVE (Verb Imper Neg) IF ( NOT -1 ("ná")) ;

# Ní fear a bhí ann ...
REMOVE (Verb Imper) IF ( -1 ("ní")) ;
```

```
#Fear maith a bhí ann
REMOVE (Verb Imper) IF ( *1 (Part Vb Rel) BARRIER (Noun)) ;

# Fear maith is ea é
REMOVE (Verb Imper) IF ( *1 (Cop) LINK 1 ("ea") BARRIER (Noun)) ;

# Ith PastInd Len not possible unless preceded by "do" verbal
particle
REMOVE (Verb Vow PastInd) IF (NOT -1 (Part Vb)) ;
#
# TYPE 2
# let us assume that imperatives are at the start of a sentence or
that
# they are preceded by some punctuation such as a quotation mark or
comma
# Déan é. Ná déan é.
# A Sheáin, déan é.
# (a) Déan an rud seo
# ... agus déan an rud sin
REMOVE (Verb Imper) IF (NOT -1 BOS OR (Part Vb Imp) OR (Punct) OR
(Item) OR (Conj Coord))  ;

# TYPE 3
# where form is ambiguously n/v after prep and noun; select noun gen
# in Irish Chumann/N-gen Staire is Seandálaíochta Chiarraí ..
# in many cases the form is unambiguously a verb
# e.g. ar maidin dúnann an t-ollmhargadh
REMOVE (Verb) IF (NOT 0 ("<rinne>")) (-1C NOUN-NOT-VN) (-2C (Prep));

# Type 4 - true after disamb in earlier cycles
# Ba mhaith/V/N/Adj an rud ...
REMOVE (Verb) IF (-1C (Cop));

# it cannot be a verb if followed by an unambigguous noun-gen
# "Meandar sula dtuigim nach bloscadh/!=verb toirní atá ann ...
REMOVE (Verb) IF (1C (Noun Gen));

# liath/!=v nó bán
# fuaim ná fothram
REMOVE (Verb) IF (1C (Conj)) (NOT 2 (Verb) OR (Part Vb));

# póg ar a leiceann./!=verb
# BUT: eg ..sin a bhfuil.  ...Pól a d'aistrigh.
# ================================================================ #
# ---------------------------------------------------------------- #
# S1 PRONOUNS
# ---------------------------------------------------------------- #
# TYPE 3
# sé/sí/siad after a verb is usually a pronoun, unless followed by a
noun
# number etc.
# chuaigh sé amach vs. chuaigh sé fhear amach, chuaigh sé chéad
fhear amach
# thóg sé trí úll
SELECT (Sbj) IF (-1C (Verb)) (NOT 1 (Noun) OR NUM-COUNT);

# TYPE 3
# D'imigh sé leis; Bhailigh siad leo etc
SELECT (Sbj) IF (-1C (Verb)) (1 (Pron Prep));

# ================================================================ #
# ---------------------------------------------------------------- #
# S1 NOUN
```

```
# ------------------------------------------------------------- #
# "an bhfuil" cannot be a noun unless preceded by preposition
# ar an bhfuil
REMOVE (Noun Ecl) IF (-1 ("an")) (NOT -2 (Prep));

# it cannot have the DefArt reading if not preceded by an Art (or
Prep Art etc.)
REMOVE (Noun DefArt) IF ( NOT -1 (Art));

# Unlikely to have emphatic noun without possessive determiner
# mo theachsa, a ngalfchúrsa
REMOVE (Noun Emph) IF ( NOT -1 (Det Poss));

# it cannot have the Dat reading if not preceded by a Preposition
# Use Prep Simp eventhough there are some exceptions
# e.g. Chraith chuile dhuine acu láimh/NDat leis : but in these
cases the
# only reading (e.g. for láimh) is Dat - and so will not be removed
# but this helps with dtig and cois where there are multiple
readings
REMOVE (Noun Dat) IF (NOT -1 (Prep Simp));

# Dé Luain, Dé Sathairn etc Gen form follows Dé in days of week
SELECT (Subst) IF (0 ("Dé")) (1 DAYS);
REMOVE (Subst) IF (0 ("Dé")) (NOT 1 DAYS);
SELECT (Gen) IF (-1 ("Dé" Subst));

# gach a raibh le déanamh,
# "gach" is not a substantive if it is followed by a noun/det/num
# gach rud, gach aon rud etc.
SELECT (Subst Noun) IF (0 ("gach")) (1 (Rel) OR (RelInd));
REMOVE (Subst Noun) IF (0 ("gach")) (NOT 1 (Rel) OR (RelInd));

# ag ardú a chinn ... vs nuair a chinn sí ...
SELECT (Noun Gen) IF (0 ("chinn")) (-1 (Noun));

# Verb "a" X; where X is either noun or verb choose noun;
# e.g. Ní raibh a fear céile...; Bhí a leath i bhfolach ...
# Note: in majority of cases X is unambiguously a verb
# e.g. thuigfeadh a raibh i gceist; na postanna atá a gcailleadh
REMOVE (Verb) IF (-1 ("a")) (-2C (Verb));

# ina theannta/measc siúd vs. Déan siúd
SELECT (Noun) IF (-1 (Prep)) (1 ("<siúd>"));

#  "ina" X; where X is either noun or verb choose noun;
# e.g. ina measc, ina bhás,
# Note: bhfuil is an exception
# Note: in majority of cases X is unambiguously a verb
SELECT (Noun) IF (-1 ("<ina>")) (NOT 0 ("<bhfuil>"));

# Proper Names
SELECT (Noun) IF (1 ("<Ó>") OR ("<Ní>") OR ("<Uí>")) (2 (Prop
Noun));
SELECT (Noun) IF (-1 ("<Ó>") OR ("<Ní>") OR ("<Uí>")) (-2 (Prop
Noun));
SELECT (Part Pat) (0 ("<Ó>") OR ("<Ní>") OR ("<Uí>")) (-1C (Noun))
(1C (Noun));

# disambiguate Len/Ecl tags on non-mutable initial vowel/cons
# i.e. vowels, l, n, r, etc.
# e.g. aicme laoch roinnt etc.

# a verb does not cause a following noun to be lenited or eclipsed
```

```
# e.g. bhíodh aicme laoch ...
REMOVE (Noun Len) IF ( -1C (Verb));
REMOVE (Noun Ecl) IF ( -1C (Verb));

# e.g. i roinn; this would be eclipsed if it were possible
SELECT (Noun Ecl) IF (0C NOUN-NOT-VN) (-1C PREP-ECL);

# e.g. de roinn; this would be lenited if it were possible
SELECT (Noun Len) IF (0C NOUN-NOT-VN) (-1C PREP-LEN);

# we should not have a gen noun form directly following a verb
REMOVE (Noun Gen) IF (-1C (Verb));
# ================================================================ #
# ---------------------------------------------------------------- #
# S1 ARTICLE
# ---------------------------------------------------------------- #
# a = an is only allowed in time phrases such as "trí a chlog" etc.
REMOVE (Art Sg) IF (0 ("<a>")) (NOT 1 ("clog"));
# ================================================================ #
# ---------------------------------------------------------------- #
# S1 POSSESSIVE DETERMINER
# ---------------------------------------------------------------- #
# a != Det Poss if immediately preceded by definite NP
# an té a cheap, fear a sheas
# NOTE this does not take into account longer NPs with Adj etc.
# OR: "b'fhearr le na cailíní a/Det? leithéid"
# TEST this:
REMOVE (Det Poss) IF (NOT 1 (Noun));

# ================================================================ #
# ---------------------------------------------------------------- #
# S1 PREPOSITION
# ---------------------------------------------------------------- #
# a != Prep Simp unless followed by a Verbal Noun
# an té a cheap, fear a sheas
REMOVE (Prep Simp) IF (0 ("<a>")) (NOT 1 (Verbal Noun));

# go != Prep Simp unless followed by a Noun
# go Meiriceá, go doran an tí, go 91 vóta
REMOVE (Prep Simp) IF (0 ("<go>")) (NOT 1 (Noun) OR NOUN-PREMOD);
SELECT (Prep Simp) IF (0 ("<go>")) (1C (Noun) OR (Num) OR (Det));

# TYPE 1
# go/PartAd is used with adjectives to form an adverb
# e.g. go leor, go maith etc.
# "go" is tagged as Part Ad so that preps will only ever precede an
NP
REMOVE (Prep Simp) IF (0 ("go")) (1C (Adj));


# "a" before "chlog/cloig" is a prep
SELECT (Prep Simp) IF (0 ("<a>")) (1C ("clog"));
SELECT (Prep Simp) IF (0 ("<a>")) (1 (Punct)) (2C ("clog"));

# it is a prep. if followed by the dative case
# e.g. ó Éirinn
SELECT (Prep Simp) IF (1 (Noun Dat));

# TYPE 1
# d'éirigh leo: "leo" is prep not noun
SELECT (Prep) IF (0 ("le")) (-1 ("éirigh"));

# ================================================================ #
# ---------------------------------------------------------------- #
```

```
# S1 COPULA
# ------------------------------------------------------------------ #
# TYPE 1
# ea is only used with copula
SELECT (Cop) IF (1 ("ea"));
# ba dh'ea
SELECT (Cop) IF (1 ("<dh'>")) (2 ("ea"));

# TYPE 1
# féidir is only used with the copula
SELECT (Cop) IF ( 1  ("féidir"));

# e.g. Ba iad, is é ... perhaps this is too broad
# an é , nach é, ní hí
SELECT (Cop) IF (1 ("é") OR ("í") OR ("iad") OR ("hé") OR ("hí") OR
("hiad"));

# Is/Ní gá: is/ní are most likely to be a copula if followed by "gá"
# but exclude "an" as this could be either Art or Cop e.g  an gá
SELECT (Cop) IF (NOT 0 ("<an>")) (1 ("gá"));

# If it is at the start of a sentence Is is more likely to be Cop
than Conj
SELECT (Cop) IF (0 ("<Is>")) (NOT 1 (Verb) OR (Part Vb) OR (Cop));

#Arbh é Seán a bhí ann - arbh!=dependant at start of sentence
REMOVE (Cop Dep) IF ( -1 (>>>));
# ================================================================== #
# ------------------------------------------------------------------ #
# S1 PARTICLES - Verbal
# ------------------------------------------------------------------ #

# it cannot be a verbal particle if it is not followed by a verb
# or verb particle (Vb) such as d' e.g. a d'fhreagair Máire
REMOVE (Part Vb) IF ( NOT 1 (Verb) OR (Vb));

# it cannot be a subjunctive verbal particle if it is not followed
by a subj. verb
REMOVE (Part Vb Subj) IF ( NOT 1 (Verb PresSubj));

# it cannot be an imperative verbal particle if it is not followed
by an imper. verb
REMOVE (Part Vb Imp) IF ( NOT 1 (Verb Imper));

# cé go/nár/gur/nach
# e.g. cé go raibh,
REMOVE (Part Vb Subj) IF ( -1 ("cé"));

# it is a verb particle if it is followed by an unambiguous
(C=careful) verb
# except go=Conj e.g. Tá sé soiléir go raibh ...
# go=verb part only in the case of subjunctives
SELECT (Part Vb Subj) IF (1C (Verb PresSubj));
SELECT (Part Vb) IF (1C (Verb));

SELECT (Part Vb NegQ) IF (1C (Verb Q));
SELECT (Part Vb Q) IF (1C (Verb Q));

# Ní mór
# Ní raibh
# but: níor glacadh
REMOVE (Part Vb Neg) IF (0 ("ní")) (NOT 1 (Verb Len));

# nach bhfuil vs. nach raibh
```

```
REMOVE (Part Vb Past) IF (NOT 1 (PastInd) OR (PastImp) OR
(PastSubj));

# d' before a possible verb is most likely to be a verb particle
# e.g. d'fhág is part + Verb rather than prep + noun
SELECT (Part Vb) IF (0 ("do")) (1 (Verb PastInd));

# "a" is an indirect relative, if the following verb is followed by
# poss det mo/do/a
# e.g. Indirect: an fear a raibh a mhac san ospidéal
# Direct  : an fear a bhí san ospidéal
SELECT (Part Vb Rel Indirect) IF ( 1C (Verb)) (2 (Det Poss));

# verb is lenited after direct rel "a" and eclipsed after indir rel
"a"
# use 1C here as "a" could be Prep Simp with Nv e.g. a dhíol
SELECT (Part Vb Rel Direct)   IF (0 ("a")) ( 1C (Verb Len)) ;
SELECT (Part Vb Rel Indirect) IF (0 ("a")) ( 1C (Verb Ecl)) ;
# dependent forms also follow rel particles
SELECT (Part Vb Rel Indirect) IF ( 1 (Verb Dep));

# Ar shíl/inis/thomhais tú etc. where Verb/Noun ambiguity - choose
verb if preceded by Ar at start of sentence (i.e. starts with cap)
SELECT (Part Vb Q Past) IF (0 ("<Ar>")) (1 (Verb)) (2 (Pron Pers))
(NOT 2 ("í") OR ("é"));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Copula
# ------------------------------------------------------------------ #
# it can't be  a copular particle if it is not followed by a
(possible) copula
REMOVE (Part Cp Rel) IF (NOT 1 (Cop));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Vocative
# ------------------------------------------------------------------ #
# it cannot be a vocative particle if it is not followed by a proper
(voc) noun
# not all prop nouns are marked as such ... but all will have Voc
# but voc not always used ... e.g , a Tom.
REMOVE (Part Voc) IF ( NOT 1 (Noun Voc) OR (Prop Noun) OR (Guess
Noun));

# "A dhiabhail" or "A bhó chairdiúil"
#SELECT (Part Voc) IF (@1 (Part Voc)) (@2 (Noun Voc)) (@3 (Punct) OR
(Noun) OR (Adj) OR ("<na>") OR (Itj));
SELECT (Part Voc) IF (-1 BOS)  (1 (Noun Voc)) (2 (Punct) OR (Noun)
OR (Adj) OR ("<na>") OR (Itj));
SELECT (Part Voc) IF (-2 BOS) (-1 (Itj) OR (Punct) OR (XMLTag))  (1
(Noun Voc)) (2 (Punct) OR (Noun) OR (Adj) OR ("<na>") OR (Itj));
# , a Bheartla
# <p> A Mháire
# ' A Lia,
# Ó a Rí na trua, tóg dínn ...
# .. in this above example Ó is taken to be a preposition :-(
REMOVE (Part Voc) IF (NOT @1 (Part Voc)) ( NOT -1 (Punct) OR
(XMLTag) OR (Itj)) ;
SELECT (Part Voc) IF (-1 (Itj)) (1 (Noun Voc));
# it is a voc particle if it is followed by a proper noun
SELECT (Part Voc) IF (1 (Prop));

# ", a chroí" ", a mhac" etc
# BUT not if the noun could be a verbal noun
# e.g. ", a chothú...", ", a dheisiú..." ", a chodladh.."
```

```
# OR even a verb
# e.g. ", a shíl mé..."
# OR if followed directly be a functional category like Conj or Prep
# vocative examples:
# a dhuine cóir, a Bheartla, a Mhicil chroí, a ghiolla na
léitheoireachta,
# a chúil fáinneach na dtrioplaí siar, a iníon ó, a mhaca Uisnigh,
SELECT (Part Voc) IF (-1 (",")) (1 (Noun Voc)) (NOT 1 (Verbal Noun)
OR (Verb)) (2 (Punct) OR (Noun) OR (Adj) OR ("<na>") OR (Itj));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Patronymic (names)
# ------------------------------------------------------------------ #
# it cannot be a patronymic particle if it is not followed by a
proper noun
# soften the following rule to noun rather than proper noun
# some surnames are also common nouns

REMOVE (Part Pat) IF ( NOT 1 (Noun));

# restrict de as Part Pat to cases where followed and preceeded by a
proper
# noun to avoid unnecessary ambiguity
REMOVE (Part Pat) IF (0 ("<de>")) ( NOT -1 (Prop Noun)) (NOT 1 (Prop
Noun));

# Let us assume that in a name (Part Pat) Mac will have uppercase
REMOVE (Part Pat) IF (0 ("<mac>")) (NOT 1 (Prop)) ;

# it is a patronymic particle if it is followed by a proper noun
# but note that de is quite often a Prep before Prop noun
# e.g. 31ú lá de Nollaig
SELECT (Part Pat) IF (1 (Prop)) (NOT 0 ("de"));
SELECT (Part Pat) IF (-1 (Prop))  (0 ("de")) (1 (Prop));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Numeric
# ------------------------------------------------------------------ #
# it cannot be a numeral particle if it is not followed by a numeral
# a ceathair a clog, Dé  Satharn ar a 4
REMOVE (Part Nm) IF ( NOT 1 NUM-COUNT OR (Num Dig));
# a chéad rogha
REMOVE (Part Nm) IF ( 1C (Num Len) OR (Num Ecl));
SELECT (Part Nm) IF ( 1 (Num Card));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Adjectival
# ------------------------------------------------------------------ #
# it cannot be an comparative or superlative particle if it is
# not followed by a comp adj
# e.g. is mó, níos lú, ba mhó
# AND Verbal Adj   ...níos spreagtha ...
# e.g. ...beartais a sholáthar agus ionchur níos sonraithe/VA a
áirithiú...
# BUT some Adj Comp are not recognised and are guessed as something
else ...
# also ní ba mhó
REMOVE (Part Comp) IF ( NOT 1 (Adj Comp) OR (Verbal Adj) OR (Guess)
OR (Cop));
REMOVE (Part Sup) IF ( NOT 1 (Adj Comp) OR (Verbal Adj) OR (Guess));
REMOVE (Part Deg) IF ( NOT 1 (Adj Comp) OR (Verbal Adj) OR (Guess));

# it cannot be an degree particle if it is not followed by a comp.
adj
```

```
# or abstract noun and a rel particle
# e.g a géire (agus) a labhair sí, a dhonnacht (is) a bhí sé
REMOVE (Part Deg) IF ( NOT 1 (Adj Comp) ) (NOT 2 (Rel) OR (Coord));
# select comparative or superlative particle if followed by
Comparative Article
SELECT (Part Sup) IF (1  (Adj Comp));
SELECT (Part Comp) IF (1  (Adj Comp));
# ní ba mhó;
SELECT (Part Comp) IF (1  ("<ba>")) (2 (Adj Comp));

# ------------------------------------------------------------------ #
# S1 PARTICLES - Adverbial
# ------------------------------------------------------------------ #
# "go" is adverbial particle if followed by adjective
SELECT (Part Ad) IF ( 1  (Adj));
# ================================================================== #
# ------------------------------------------------------------------ #
# S1 NUMERALS
# ------------------------------------------------------------------ #
# an chéad bhliain, sa chéad leath, den chéad uair
# a chéad phost
# mar chéad fhocal, but NOT faoi chéad
REMOVE (Num Ord) IF (0 ("<chéad>")) (NOT -1 (Art) OR (Det Poss) OR
("<mar>"));
SELECT (Num Ord) IF (0 ("<chéad>")) (-1 (Art) OR (Det Poss) OR
("<mar>"));
SELECT (Num Card) IF (0 ("míle")) (1C (Num) OR (Noun) OR (Punct
Int));
REMOVE (Num Card) IF (0 ("míle")) (-1 ("míle"));
# dhá chéad,
# faoi chéad,
# le linn chéad fiche bliain, ar feadh chéad bliain
# ceithre mhíle siar ón mbaile => míle = Noun NOT Num in this
context
SELECT (Num Card) IF (-1C (Num) OR TIME-PERIOD OR ("<faoi>")) (1
(Noun) OR (Num));
# mar aon chéad
# BUT Tógann sé dhá uair a chloig ...
SELECT (Num Card) IF (1C (Num Card)) (NOT -1 ("sé"));

# an seachtú reisimint
REMOVE (Num Ord) IF (NOT -1 (Art) OR (Det Poss) OR (Prep Simp));
SELECT (Num Ord) IF (-1 (Art) OR (Det Poss) OR (Prep Simp));

# an dá cheann vs dá réir sin
REMOVE (Num Card) IF (0 ("<dá>")) (NOT -1 (Art));
SELECT (Num Card) IF (0 ("<dá>")) (-1 (Art));

# deoch nó dhó vs d'inis mé (féin) dhó
SELECT (Num Card) IF (0 ("<dhó>")) (-1 ("nó") OR ("faoi"));
REMOVE (Num Card) IF (0 ("<dhó>")) (NOT -1 ("nó") OR ("faoi"));

# deoch/com nó dó vs di nó dó
REMOVE (Num Card) IF (0 ("<dó>")) (-1 ("nó")) (-2 ("do" Pron Prep));
# i gceann nóiméid/gen nó dó
SELECT (Num Card) IF (0 ("<dó>")) (-1 ("nó")) (-2 (Noun));
# rud beag nó dó
SELECT (Num Card) IF (0 ("<dó>")) (-1 ("nó")) (-2 (Adj)) (-3
(Noun));

# aon nó dó; deich nó dó dhéag
SELECT (Num Card) IF (0 ("<dó>")) (-1 ("nó")) (-2 (Num Card));

# a dó/Num a chlog vs ar fáil dó/Pn vs tar éis an dó/N vs ag dó/Vn
```

```
REMOVE (Num Card) IF (0 ("<dó>")) (NOT -1 ("a"));
SELECT (Num Card) IF (0 ("<dó>")) (-1 ("a"));

# na trí chineál rud vs a trí a chlog vs trí mheán
#REMOVE (Num Card) IF (0 ("<trí>")) (NOT -1 (Art) OR ("a"));
SELECT (Num Card) IF (0 ("<trí>")) (-1 (Art) OR ("a"));

#trí mhí, trí chinn,
SELECT (Num Card) IF (0 ("<trí>")) (1 TIME OR MEASURE OR THING);

# níos mó ná deich ...
SELECT (Num Card) IF (3 ("<níos>")) (2 ("<mó>") OR ("<lú>")) (1
("<ná>"));

# trí artairí, trí théamh
REMOVE (Num Card) IF (0 ("<trí>")) (1 (Verbal Noun));

# an dó dhéag
SELECT (Num Card) IF (1 ("déag"));

#trí nó ceathair
SELECT (Num Card) IF (1 (Conj)) (2 (Num Card));

# a trí nó a ceathair
# an trí nó an ceathair
SELECT (Num Card) IF (1 (Conj)) (2 ("<an>") OR ("<a>")) (3 (Num
Card));

# the simple case
# e.g. aon dream amháin ..
SELECT (Num Card) IF (0 ("aon")) (2 ("amháin"));
# NEED TO ACCOUNT FOR COMPLEX NPs HERE
# let us say that a clause is bounded by internal punctuation or the
start of a PP
SELECT (Num Card) IF (0 ("aon")) (*1 ("amháin") BARRIER PUNCT-INT OR
(Prep Simp)) ;
# remove Num reading for "aon" if there is not "amháin" somewhere to
the right
# except " a haon ..."
# BUT ní raibh ach aon ceann mór, tiubh amháin fágtha.
REMOVE (Num Card) IF (0 ("aon")) (NOT *1 ("amháin") BARRIER PUNCT-
INT OR (Prep Simp)) (NOT -1 ("a")) ;

# a sé/NumCard a chlog
SELECT (Num Card) IF (1 ("<a>")) (2 ("clog"));
# an dó chlog ar maidin
SELECT (Num Card) IF (1 ("clog"));
# a sé/NumCard a ' chlog
SELECT (Num Card) IF (2 (Punct)) (3 ("clog"));
SELECT (Num Card) IF (-1C (Part Nm));

# ================================================================ #
# ---------------------------------------------------------------- #
# S1 CONJUNCTIONS
# ---------------------------------------------------------------- #
# e.g. má bhíonn &rl.
SELECT (Conj) IF (0 ("má") OR ("dá")) (1C (Verb));
# if má/dá at start of sentence they are invariably a Conj
SELECT (Conj) IF (0 ("má") OR ("dá")) (-1 BOS) (1 (Verb));

# ================================================================ #
# SECTION 2 - More Definite Rules
CONSTRAINTS
# ================================================================ #
```

```
# ------------------------------------------------------------------ #
# S2 ADVERBS
# ------------------------------------------------------------------ #
# intensifiers must be followed by an adjective
# e.g. breá te, sách ard
REMOVE (Adv Its) IF (NOT 1 (Adj));
SELECT (Adv Its) IF (1 (Adj));

# Anois preceded by Prep is usually the Newspaper .. otherwise it is
an adverb
SELECT (Adv) IF (0 ("<Anois>")) (NOT -1 (Prep Simp));

# ag rith timpeall/Adv
SELECT (Adv) IF (-1C (Verbal Noun));
# ================================================================== #
# ------------------------------------------------------------------ #
# S2 ADJECTIVE
# ------------------------------------------------------------------ #
# only adjectives follow an intensifier
# e.g. breá te, sách ard, chomh maith
SELECT (Adj Base) IF (-1 (Adv Its)) ;

# níos is only used before comparative form of adjectives
SELECT (Adj Comp) IF (-1 ("<níos>")) ;
# ní ba mheasa, ní b´áille
SELECT (Adj Comp) IF (-2 ("<ní>")) ;

# adjectives follow a noun/pron (tháinig/v fear/n mór/a..., tá/v
sé/pn mór/a)
# another adj (fear/n mór/a ramhar/a)
# a comma (fear/n mór/a, ramhar/a, saibhir/a)
# an intensifier (sách te)
# adverbial "go" (go maith)
# copula "is/cop maith/adj liom "
# ag mothú/SENSORY tinn
# conjunction
# buan agus lán-aimseartha
REMOVE ADJ-NOT-VA IF (NOT -1 NOUN-NOT-VN OR (Pron) OR (Adj) OR (Adv
Int) OR (Part Ad) OR COMMA OR (Cop) OR SENSORY-VN OR (Conj));

# PrepPron cant be followed by Adj other than Verbal Adj ??
# an cuid is mó acu déanta ... acu
REMOVE ADJ-NOT-VA IF (-1 (Pron Prep));

# it is not an attributive adj (i.e. inflected) unless preceded by a
noun
# or another adj
REMOVE ADJ-ATTR IF ( NOT -1 (Noun) OR ADJ-ATTR);

# it is not a comparative adj unless preceded by a comparat. or
superl. part.
# or degree particle (e.g. a géire )
# or "ba shéimhe...
REMOVE (Adj Comp) IF ( NOT -1 (Part Comp) OR (Part Sup) OR (Part
Deg) OR (Cop));

# NOTE: adj can be followed immediately by a verb
# e.g. Nuair a bhí an poll lán(adj) dhéantaí (verb) é a chlúdach
# it is not an adj if it is preceded immediately by an unambiguous
verb in
# declarative clauses - but this does not hold for relatives
# e.g. "...an líne a/rel bhí/verb díreach"
# e.g. "Sin fear atá/rel-verb cliste ...
# Tá buailte agam ar an bhfear sin
```

```
# ...tá ceangailte orm dul agus an éagóir sin a chosc ...
# therefore use lemma bí rather than atá
REMOVE (Adj) IF ( -1C (Verb)) (NOT -1 ("bí")) (NOT -2 (Part Rel));
# Adj should match number of previous noun (unless numbers are
involved)
# avoid the tagging "ainmneacha" as Adj in following "a lán
ainmneacha"
# dhá(num) bhád(n-sg) bheaga(a-pl)
# trí(num) long(n-sg) déag(n-sg) mhóra(a-pl) (NIG p78)
REMOVE (Adj Sg) IF ( -1 (Noun Pl));
REMOVE (Adj Pl) IF ( -1 (Noun Sg)) (NOT -1 ("déag")) (NOT -2 (Num))
;

# for the case of vowel ending adjs which now only have Base and
Base Len
# le galar nua
REMOVE (Adj Base Len) IF (-1C (Noun)) (NOT -1 (Noun Len) OR (Noun
Fem) OR (Noun Masc Gen));

# Adjective should match noun in gender, number and case
# of the prev noun: exclude the PART-GEN nouns e.g. roinnt, cuid etc
# let us assume that they are followed by Noun Gen rather than Adj
# - this more likely but not essential ...

SELECT (Adj Fem) IF (0C ADJ-NOT-VA) (-1C (Fem))  ;
SELECT (Adj Masc) IF (0C ADJ-NOT-VA) (-1C (Masc)) ;
SELECT (Adj Sg) IF (0C ADJ-NOT-VA) (-1C (Sg)) (NOT -1 ("déag")) (NOT
-2 (Num))  ;
SELECT (Adj Pl) IF (0C ADJ-NOT-VA) (-1C (Pl))  ;
SELECT (Adj Pl) IF (0C ADJ-NOT-VA) (-1C ("déag"))  ;
SELECT (Adj Pl) IF (0C ADJ-NOT-VA) (-2 (Num))  ;
SELECT (Adj Com) IF (0C ADJ-NOT-VA) (-1C (Com) OR (Dat))  ;
SELECT (Adj Gen) IF (0C ADJ-NOT-VA) (-1C (Gen))  ;
SELECT (Adj Voc) IF (0C ADJ-NOT-VA) (-1C (Voc))  ;

# buama/n sách/Its éasca/a ..
SELECT (Adj Fem) IF (0C ADJ-NOT-VA) (-1 (Adv Its)) (-2 (Fem))  ;
SELECT (Adj Masc) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Masc)) ;
SELECT (Adj Sg) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Sg))  ;
SELECT (Adj Pl) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Pl))  ;
SELECT (Adj Com) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Com) OR
(Dat))  ;
SELECT (Adj Gen) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Gen))  ;
SELECT (Adj Voc) IF (0C ADJ-NOT-VA)(-1 (Adv Its))  (-2 (Voc))  ;

# select the comparative if preceded by "is" or "níos" or "a"
SELECT (Adj Comp) IF (-1C (Part Sup) OR (Part Comp) OR (Part Deg));
SELECT (Adj Comp) IF (-2C (Part Comp)) (-1 (Cop));

# e.g. na hAifrice Thuaigh ...
# but not roinnt airgid
SELECT (Adj) IF (-1 (Noun) OR (Adj) OR COMMA) (NOT -1 GEN-PART) ;

# TYPE 4 Rule - after disamb
# adjectives used adverbially
# select adj if preceded by adverbial particle,
# e.g. go maith, go daingean etc...
SELECT (Adj) IF (-1C (Part Ad));

# Bá mhór, ní mór, nach mór etc
SELECT (Adj) IF (0 ("mór")) (-1 (Cop));

# Is iontach, (like Is breá)
REMOVE (Adj Len) IF (0C (Adj)) (-1C (Cop Pres));
```

```
# "Ba láidir" like "Ba mhaith",
SELECT (Adj Len) IF (0C (Adj)) (-1C (Cop Past));

# comparative adjective and degree particle are always followed by a
relative clause
# a géire a labhair sí,
# ba ea a dhéine a bhí sé á breathnú ...
SELECT (Adj Comp) IF (-1 ("<a>")) (1 (Rel)) (2C (Verb));
# a chiúine is a bhí sé
SELECT (Adj Comp) IF (-1 ("<a>")) (1 (Coord)) (2 (Rel)) (3C (Verb));
# a riachtanaí atá sé
SELECT (Adj Comp) IF (-1 ("<a>")) (1 ("<atá>")) ;
# a thabhachtaí is atá sé
SELECT (Adj Comp) IF (-1 ("<a>")) (1 (Coord)) (2 ("<atá>")) ;
# a indéanta agus a úsáidtí atá
SELECT (Adj Comp) IF (-1 ("<a>")) (1 (Coord)) (2 ("<a>")) (3 (Adj
Comp)) (4 (Rel));
# ================================================================= #
# ----------------------------------------------------------------- #
# S2 VERBS
# ----------------------------------------------------------------- #
# a verb follows relative particles like "inar (Prep Rel) dhiúltaigh
(Verb) sé"
# and "a (Vb Rel) deir (Verb)"
# but not "inar ghearr (Noun) go raibh sé"
# note "ar/cop mhaithe/n le" mhaithe = V presSubj should already be
dealth with
# as not preceded by go or nár
SELECT (Verb) IF (-1 (Prep Rel)) (NOT 1 (Part Vb) OR (Verb) OR
("<go>"));

# BUT:  as salann a dhíol sa cheantar etc.
SELECT (Verb Len) IF (-1 (Part Vb Rel Direct)) (NOT 1 (Punct Fin));
SELECT (Verb Ecl) IF (-1 (Part Vb Rel Indirect)) (NOT 1 (Punct
Fin));

# it is not a verb if it is followed immediately by "í", "é", "iad"
or their emphatic forms, and verb is not a synthethic verb form
(i.e. includes
# person/number)
# e.g. can't have Tá iad ...
# Unless verb is imperative, e.g.  Déan é
# NOTE relative constructions
# e.g a/Rel fhorlíonann/V é "...agus mise a/Rel chonaic/v é/obj..."
REMOVE (Verb) IF (NOT 0 VERB-WITH-SUBJ OR (Imper)) (1 OBJ-PRON)
(NOT -1C (Rel));

# remove reading of "ar" as a verb (e.g. , ar Seán. ) except where
preceded
# by a quotation mark or comma
# this may not always be true but it prevents the unlikely verb
reading from
# constantly appearing ...
REMOVE (Verb PastInd) IF (0 ("ar")) (NOT -1 (Punct Int) OR (Punct
Quo)) ;

# a verb is not usually preceded by a noun/pronoun/art
# BUT sula rith siad/Pron d'fhéach/Verb siad thart
# '... óna chaint/n thuigeas/v go raibh ...
# e.g. "chuir siad deireadh go deo -> deireadh not verb
# na cinn chloiche -> cinn not verb
#REMOVE (Verb) IF (-1 (Pron) OR (Noun) OR (Art));
REMOVE (Verb) IF (-1C (Art));
```

```
# if it could be autonomous (unspecified person) and is not followed
by a pron or noun
# then select autonomous reading
# transitive verbs need an obj NP e.g. 40,000 laoch
# e.g. Maraíodh 40,000 laoch "40,000 warriors were killed": killer
not specified
# it is a not a verb  if preceded by an unambiguous prep (C=careful
mode)
# tháinig siad chun cinn -> cinn is not a verb in this context
# note "inar (Prep Rel) dhiúltaigh (Verb)"
REMOVE (Verb) IF (-1C (Prep)) (NOT -1 (Prep Rel));

# it is probably not a verb if preceded by an unambiguous verb
(C=careful mode)
# d'fhéadfadh líon mór tithe ... -> líon is not a verb in this
context
REMOVE (Verb) IF (-1C (Verb));

# it is probably not a verb if followed by an unambiguous verb
(C=careful mode)
# forbairt nó táirgeadh/!=v atá; Moltaí/!=v atá uaim
# BUT: Sna blianta a lean/v chuir/v sí go mór le litríocht
# mar dá dtiocfadh bheadh an sagart
# ach má phrioc thug sé priocadh chomh géar leis uaidh
REMOVE (Verb) IF (1C (Verb)) (NOT -1 ("a") OR ("má") OR ("dá"));

#it is probably not a verb if followed by an unambiguous adj
# e.g. líon beag tithe ... -> líon is not a verb in this context
REMOVE (Verb) IF (1C (Adj));

# "aon" is usually followed by an NP not a verb (the "ace" meaning
is not
# very common ...)
# e.g. "ní raibh aon cheapadh/N agam go dtabharfainn cuairt ar an
áit
# "Is é Pádraic Ó Conaire an t-aon fhear/N a raibh an t-eolas aige .
# but NOTE "Taca a ceathair thagadh/V an saoiste chugainn"
# AND "Fan go gcloisimid é dhá shéanadh."
# so we do not extend this rule to Num Card in general
REMOVE (Verb) IF (-1 ("<aon>") OR ("<t-aon>"));
# "Ní raibh aon briseadh anseo"
REMOVE (Verbal Noun) IF (-1 ("<aon>") OR ("<t-aon>"));

# If a verb reading is possible and there is no verb to the left
# or if there is no verb to the right (looking no farther than a
relative particle) .. then select the verb reading
# e.g. "(a) Déan liosta .. " here we want to select verb rather than
noun reading for déan
# Ní mór dúinn - mór is not a verb here as preceded by Ní/Cop
# Ní fear aon leabhar a bhí ann ...
#NOTE Níor rith, Níor glacadh
SELECT (Verb) IF (NOT *-1 (Verb)) (NOT *-1 (Cop)) (NOT *1 (Verb));
# ná beadh, etc.
SELECT (Verb Cond) IF (0C (Verb)) (-1 ("ná"));

# labhair go soiléir OK but not labhair sé/Séan/an fear etc
SELECT (Verb Imper) IF (NOT 1 (Noun) OR (Pron Pers) OR NOUN-PREMOD);
REMOVE (Verb Imper) IF (1C (Prop Noun));
# ------------------------------------------------------------------ #
# S2 VERBAL NOUNS
# ------------------------------------------------------------------ #
# a VN following "ag" is not followed by attrib adj
# unless it is a sensory vn
```

```
# e.g. "ag dlí (noun)  poiblí (adj)" but "ag mothú(vn) tinn (adj)"
REMOVE (Verbal Noun) IF (-1 ("<ag>")) (NOT 0 SENSORY-VN) (1C ADJ-
NOT-VA);

# verbal noun must be lenited after "a"
# a dhéanamh(VN?) vs a déanamh (!=Vn)
# also Ecl: "...ní raibh ionsaithe a ndéanamh orthu..."
REMOVE (Verbal Noun) IF (-1 ("<a>")) (NOT 0 (Verbal Noun Len) OR
(Verbal Noun Ecl) OR (Guess Verbal Noun));

# e.g "agus a rá" where choice of rá=VN/VN_Len/VN_Ecl remains,
choose VN_Len
SELECT (Verbal Noun Len) IF (-1C ("a" Prep Simp))  (0C (Verbal
Noun));

# a dtaithí, a ndéanamh: "a" is Det Poss rather than Simp Prep ...
REMOVE (Verbal Noun Ecl) IF (-1C ("a" Det Poss 3P Pl)) ;

# ar crochadh/VN ar siúl/VN vs. ar chrochadh/N
REMOVE (Verbal Noun Len) IF (-1C ("<ar>"))  (NOT 0 ("<bheith>"));

# if one of the possible readings is Verbal Noun and it follows a
simple prep
# - then select this reading
# but not a dtaithí where a/DetPoss - see previous rule ...
SELECT (Verbal Noun) IF ( -1 PREP-VN) (NOT 0 (Verbal Noun Gen));

# verbal noun follows "á" - lemma "do" (a)
SELECT (Verbal Noun) IF ( -1 ("<á>")) ;

# bheith is almost invariably a verbal noun
# e.g. ..ar a chumas bheith páirteach ...
SELECT (Verbal Noun) IF (0 ("<bheith>")) ;

# e.g. bain úsáid as -> úsáid = common noun
# e.g. sa (=ins an) troid
# NOTE i mbun an fhiosraithe/VNgen
# NOTE fear a/rel theastaigh/v fanacht/vn ...
# ... nach/rel bhféadfaí/v glacadh/vn leis na rialacháin ...
REMOVE (Verbal Noun) IF ( -1 (Verb) OR (Prep Art)) (NOT -2 (Rel)) ;

# a verbal noun is (often) preceded by a verb or copula,
# therefore remove VN reading otherwise
# e.g. Roinnt bliain ó shin ... roinnt is not VN in this context
# BUT SEE Seasamh amuigh i lár an bhóthair ... => use ADV amuigh
# NOTE Ag fanacht/vn sa bhaile a bhí/v sé.
# Ag tuairisciú don Chomhairle, táimid freagrach as ...
# Cén mhaith caint/VN leis go nochtfaidh/V ...
REMOVE (Verbal Noun) IF (NOT *-1 (Verb) OR (Cop) ) (NOT *1 (Rel) OR
COMMA OR ("go") LINK 1C (Verb)) (NOT 1 (Adv));

# VN doesnt start a sentence without some prep before it ????
# Rith/!=VN siad amach
# Ag rith/VN amach dóibh chonaic siad ...
# but can start a clause: "Éilófar uirthi, afách, taisteal/VN chuig
an Ionad...
REMOVE (Verbal Noun) IF (-1 BOS);
# ------------------------------------------------------------- #
# S2 NOUNS
# ------------------------------------------------------------- #
# TYPE 1/2
# nouns are not eclipsed after "go"
# eclipsed noun only in phrases like "scian go mbuaidh" where go
means "with"
```

```
# but this an exception
REMOVE (Noun Ecl) IF (-1 ("go"));

# eliminate the unlikely leo=noun rather than common leo=prep-pron
# unless preceded by a poss det
# note can have "an leo é?" where an could be mistaken for the
article ...
REMOVE (Noun) IF (0 ("<leo>")) (NOT -1 NOUN-PREMOD);

# it cannot have the DefArt tag if it is not preceded by the
appropriate article
REMOVE (Noun Pl DefArt) IF ( NOT -1 (Art Pl));
REMOVE (Noun Sg DefArt) IF ( NOT -1 (Art Sg));

# remove Noun Gen reading if it follows a verb
# Diúltaíonn (na) mná tí ...
REMOVE (Noun Gen) IF ( -1C (Verb));
REMOVE (Noun Gen) IF ( -1C (Art)) (-2C (Verb));

# Cheannaigh Seán leabhar!=Gen i measc na sluaite=Gen
REMOVE (Noun Gen) IF ( -2C (Verb VT) OR (Verb VTI)) (-1C (Noun Com))
(NOT *1 (Noun Com) BARRIER CLB) ;

# if gen pl is preceded by an article it must be a pl art
# remove Noun Gen Pl reading if it follows "an"
# Shín an sáirsint/n an guthán/n dearg chuig Dunne
REMOVE (Noun Gen Pl) IF ( -1 ("<an>") OR ("<An>"));

# gen doesnt follow titles e.g. an tUrramhach James
REMOVE (Noun Gen) IF ( -1C TITLE);
SELECT (Noun) IF ( -1C TITLE);

# it cannot be a vocative noun if it is not preceded by a voc
particle
REMOVE (Noun Voc) IF ( NOT -1 (Part Voc));
SELECT (Noun Voc) IF ( -1C (Part Voc));

# can have Art Noun Noun
# e.g. chonaic an fear bean
# e.g. An Roinn Oideachas, Na Gardaí Síochána

# NOTE: siopa mór Sheáin N A N-gen
# it cannot be a genitive noun if it is not preceded by a
# noun or verbal noun or compound prep or some simp preps (see LIST
GENITIVE)  or noun+article
# ag baint as aimhréidhe
# added second Prep Cmpd
# e.g. i mbun an fhiosraithe ...
# Det Poss: tar éis a/detPoss mbreithe/gen
# Conj Coord: Cumann Seandálaíochta/gen is/conj Staire/gen: OMITTED
for the moment as it causes too much ambiguity ..."chaith a
chloigeann san aer agus rinne/!=gen gáire ...
# e.g. Is iomaí Gael/!=gen ...

# imithe chun/gen_part a/DetPoss pósta/gen
SELECT (Noun Gen) IF (-1 (Det Poss)) (-2C GEN-SIMP-PREP);

# saolré na Chéad Dála
REMOVE (Noun Gen) IF (NOT -1 GEN-SIMP-PREP OR GEN-PART OR (Prep
Cmpd) OR (Noun) OR (Adj) OR (Art) OR (Det Poss) OR (Num Ord)) ;

# remove noun-gen if there is not a noun or compound prep etc.
somewhere to the left
# noun should be in gen following certain preps and partitive nouns
```

```
# therefore remove common and voc cases ...
REMOVE (Noun Voc) IF (-1 GEN-SIMP-PREP OR GEN-PART OR (Prep Cmpd)) ;
REMOVE (Noun Com) IF (-1 GEN-SIMP-PREP OR GEN-PART OR (Prep Cmpd))
(NOT -1C (Guess))  ;

# leis: noun or prep pron?
# remove the noun reading of "leis" unless it is obviously a noun
REMOVE (Noun) IF (0 ("<leis>")) (NOT -1 NOUN-PREMOD);

# dó: noun or prep pron?
# remove the noun reading of "dó" unless it is obviously a noun
REMOVE (Noun) IF (0 ("<dó>")) (NOT -1 NOUN-PREMOD) ;

# sí: noun or pronoun?
# remove the noun reading of "sí" unless it is obviously a noun
REMOVE (Noun) IF (0 ("<sí>")) (NOT -1 NOUN-PREMOD);

# úd: noun or determiner?
# remove the noun reading of "úd" unless it is obviously a noun
REMOVE (Noun) IF (0 ("<úd>")) (NOT -1 NOUN-PREMOD);

# dul: noun or verbal noun?
# remove the noun reading of "dul" unless it is obviously a noun
REMOVE (Noun) IF (0 ("dul")) (NOT -1 NOUN-PREMOD);

# roinnt: noun or verbal noun?
# remove the noun reading of "roinnt" if not followed by a noun
# rud le roinnt/vn vs le roinnt/n mínonna
REMOVE (NOUN-NOT-VN) IF (0 ("roinnt")) (-1 (Prep Simp)) (NOT 1
(Noun));

# céanna: noun or adj?
# céanna/chéanna (almost) always an adj
# gcéanna is noun in "mar an/a gcéanna"
# remove the noun reading of "céanna" unless it is obviously a noun
REMOVE (Noun) IF (0 ("céanna")) (NOT -1 (Art Sg));
SELECT (Noun) IF (0 ("céanna")) (-1 (Art Sg));
SELECT (Art Sg) IF (1 ("céanna")) ;

# ar uain, d´uain
SELECT ("uain" Noun Com) IF (0 ("<uain>")) (-1 (Prep Simp));
# select uain/com unless it could possible be genitive of uan
SELECT ("uain" Noun Com) IF (0 ("<uain>")) (-1 (Art Def)) (NOT -2
(Noun Com));

# chuir sé a lámha timpeall/adv uirthi/prep-pron
REMOVE (Noun) IF (0 ("<timpeall>")) (1C (Pron Prep));
# Ní raibh aon bhac orthu gabháil/nv timpeall/adv
REMOVE (Noun) IF (0 ("<timpeall>")) (-1C (Verbal Noun));
#
# bí cannot take a bare noun complement
# e.g. Bhí sé dorcha!=Noun
REMOVE (Noun) IF (*-1 ("bí") BARRIER (Rel) OR (>>>) LINK *1 (Noun
Com) OR (Pron Pers));
# ------------------------------------------------------------- #
# S2 NUMERALS
# ------------------------------------------------------------- #
# aon MEANS ONE WHEN AMHÁIN FOLLOWS IT
# aon can mean one/Num, any/Det or ace/Noun, but before another noun
it is unlikely to mean ace/Noun
# e.g. Ní dhearna sé aon dochar ...
REMOVE (Noun) IF (0 ("aon")) ( 1 (Noun));

# 2 - 1  vs agus - agus
```

```
REMOVE (Num Op) IF (NOT -1 (Num Dig));

# here  aon = Det (any); not aon/Num (one) or aon/Noun (ace)
# e.g. ní raibh aon siopaí
REMOVE (Num Card) IF (0 ("aon")) (1C (Noun Pl));

# cardinal num should be followed by a noun or "a chlog" only???
# e.g. "ní raibh sé mór" vs. "ní raibh sé ubh ann"
# le dhá chéad bhliain
# naoi gcéad seachtó's/Unknown a dó ...
REMOVE (Num Card) IF (NOT 1 (Noun) OR ("a" Prep Simp) OR (Num Card)
OR (Unknown));

# ------------------------------------------------------------------ #
# S2 NOUNS
# ------------------------------------------------------------------ #

# it is a gen noun if preceded by another noun/verbal noun/comp prep
etc

# cois chladaigh
SELECT (Noun Gen) IF (-1 GEN-SIMP-PREP);
# cois na trá
SELECT (Noun Gen) IF (-1 (Art Def)) (-2 GEN-SIMP-PREP);

# roinnt ama
SELECT (Noun Gen) IF (-1 GEN-PART);
# roinnt bheag ama
SELECT (Noun Gen) IF (-2 GEN-PART) (-1 (Adj));

# tar éis scoile
SELECT (Noun Gen) IF (-1 (Prep Cmpd));
# tar éis na scoile
SELECT (Noun Gen) IF (-1 (Art Def)) (-2 (Prep Cmpd));

# cá bhfios don arm bocht ... dont want to include the Art in "don"
# BUT not "An fíor an ráiteas... "Léiríonn an pictiúir an méid ...
SELECT (Noun Gen) IF (-1C (Art Def) ) (-2C (Noun)) (NOT -3 (Art
Def));

# it is a noun (not verbal) if preceded by a numeral
# BUT exclude sé which is commonly a pronoun (he/it)
SELECT NOUN-NOT-VN IF (-1 (Num)) (NOT -1 ("sé" Pron Pers));

# it is a noun if preceded by an unambiguous prep (C=careful mode)
SELECT (Noun) IF (-1C (Prep));

# noun  should match following adj in gender, number and case
SELECT (Noun Fem) IF (0C (Noun)) (1 (Adj Fem));
SELECT (Noun Masc) IF (0C (Noun)) (1 (Adj Masc));
SELECT (Noun Sg) IF (0C (Noun)) (1 (Adj Sg));
SELECT (Noun Pl) IF (0C (Noun)) (1 (Adj Pl));
SELECT (Noun Com) IF (0C (Noun)) (1 (Adj Com));
SELECT (Noun Gen) IF (0C (Noun)) (1 (Adj Gen));
SELECT (Noun Voc) IF (0C (Noun)) (1 (Adj Voc));

# noun  should match preceding article in gender, number and case
# an = Art Sg Def / Art Gen Sg Def Fem
# na = Art Pl Def
SELECT (Noun Fem Gen Sg) IF (-1C (Art Gen Fem));
SELECT (Noun Sg) IF (-1C (Art Sg));
SELECT (Noun Pl) IF (-1C (Art Pl));

# TYPE 4 - after earlier disambiguation
```

```
# seomra suí
# BUT Chonaic fear bean (if bean could be gen)
# remove Com? exclude VN?  NO
# le lucht imirce
SELECT (Noun Gen) IF (-1C (Noun Com)) (-2C (Prep Simp));
# ------------------------------------------------------------- #
# S2 DETERMINERS (demonstrative and poss.)
# ------------------------------------------------------------- #
# demonstrative determiners follow an NP not a VP
# e.g. Thaitin sin leis => pronoun not det.
REMOVE (Det Dem) IF (-1C (Verb));

# san = ins an(prep) vs. san = sin (det dem)
# sonraithe san ordachán, bagairt san Anschluss
REMOVE (Det Dem) IF (0 ("<san>")) (1C (Noun));

# do do leithéadh, do mo chrá: in these cases first "do" is Prep
second "do"
# is possessive determiner
# mar a d'éirigh leo where d' is Part Vb
REMOVE (Det Poss) IF (1 (Det Poss) OR (Part Vb));

# seo (this) or sin (that) can be either a pronoun or a
demonstrative det.
# IT IS DEM IF FOLLOWS A *DEF* NOUN
# e.g. "rinne an madra seo (def)" but "rinne an madra é seo (pron)
é"
# e.g. "í seo" This seo is most likely demonstrative rather than a
pronoun >>NO
# NOTE the case for pron. will require transitivity info to resolve
it
# rinne an madra (é) seo (the dog did this) = VP NP NP
# an fear sin
SELECT (Det Dem) IF (-2 (Art)) (-1C (Noun));
# an dá fear sin
SELECT (Det Dem) IF (-3 (Art)) (-2 (Num)) (-1C (Noun));
# an fear beag sin
# an deireadh seachtaine seo
# an liosta seiceála (VN-gen?) sin
SELECT (Det Dem) IF (-3 (Art)) (-2C (Noun)) (-1C (Adj) OR (Gen)) ;
# an dá fear beag sin
SELECT (Det Dem) IF (-4 (Art)) (-3 (Num)) (-2C (Noun)) (-1C (Adj) OR
(Gen)) ;

# corollary of above
# e.g. Sin cuid den fhadhb ...
# Is iad/Pron sin/Det na sainceisteanna ..
REMOVE (Det Dem) IF (NOT -1 (Noun) OR (Adj) );
# iad siúd/Det
# díobh/leo siúd/Pron
#SELECT (Det Dem) IF (-1 (Pron));
SELECT (Det Dem) IF (-1 (Pron Pers));

# is ionann sin ..
 REMOVE (Det Dem) IF (-2 (Cop)) (-1 (Adj) OR (Pron));

# an saghas sin cainte
SELECT (Det Dem) IF (1C (Noun Gen));
# ------------------------------------------------------------- #
# S2 ARTICLE
# ------------------------------------------------------------- #
# TYPE 1
# e.g. an bhfuil - if bhfuil is a verb not a noun then an cannot be
an art
```

```
REMOVE (Art) IF (1C (Verb));

# TYPE 1
# Art "na" is not a feminine article unless the following noun is
Fem
REMOVE (Art Fem) IF ( NOT 1 (Noun Fem));

# TYPE 1
# Art "na" is not the plural article unless the following noun is Pl
REMOVE (Art Pl) IF ( NOT 1 (Noun Pl));

# NOTE
# an 19ú haois déag .. an t-aon ceann amháin, an droch rud ná, an
CCEA
REMOVE (Art) IF ( NOT 1 (Noun) OR (Num) OR (Det) OR (Abr) OR ADJ-
PRENOM );

# it is an article if followed by definite noun
# problem: ar an seilp -> Def forms should be "ar an seilpe" or "an
tseilp"
# remove DefArt on Noun from rule for robustness
# also remove Com on Noun  e.g. An fíor é? an=cop
SELECT (Art Sg) IF (1C (Noun Sg));
SELECT (Art Pl) IF (1 (Noun Pl));

# it is a fem gen article if followed by definite fem gen noun
SELECT (Art Gen Fem)  IF (1 (Noun Fem Gen DefArt));

# an article can precede a numeral
# an chéad rud eile, an dara dul suas etc.
# an t-aon dráma
# ar an 1 Aibreán
SELECT (Art) IF (1 (Num));

# "an" is is most likely an Art (rather than Cop) if  preceded by a
prep (c)
# and followed be a noun
# e.g. ar an gcuma san
# ar an gcéad dul síos
SELECT (Art) IF (-1C (Prep Simp)) (1 (Noun) OR NOUN-PREMOD);

# san = ins an (prep art)  / sin (det)
# san Earrach => san = prep art
# # e.g. ar an machnamh san dó/!=N
# an/Cop fear/N atá ann?
# scéalaíochta san na ridireachta
# san = "sin" not "ins an"
REMOVE (Prep Art) IF (1C (Art));
# den chéad uair
REMOVE (Prep Art) IF (NOT 1C (Noun) OR (Num Card));

# ------------------------------------------------------------------ #
# S2 PREPOSITION
# ------------------------------------------------------------------ #
# TYPE 1
# a is prep always before bheith/VNoun
SELECT (Prep Simp) IF (1C ("<bheith>"));

# TYPE 1
# ar is prep always before ...
SELECT (Prep Simp) IF (1C ("<chlé>") OR ("<dheis>"));

# TYPE 1
```

```
# dhá dhath vs dhá choinneál dúinte: dhá usually means two and can
only be a
# prep possesssive before a (potential) verbal noun.
REMOVE (Prep Poss) IF (0 ("<dhá>")) (NOT 1 (Verbal Noun));

# TYPE 1
# e.g. i/prep do/poss theach, "do" is not a simp prep in this case
...
# e.g. do/prep do/poss  mháthair,
# do/prep do/poss chrá, do/prep mo/poss chrá, dá/prep_poss crá
# but: thar a bheith,
REMOVE (Prep Simp) IF (0 ("<do>")) (-1C (Prep Simp));

# TYPE 1
# it can't be a simp prep if not followed by an NP
REMOVE (Prep Simp) IF (NOT 1 POST-PREP);

# TYPE 2
# e.g. "ar nó roimh"
# a chruthú trí/Prep agus/C ar/Prep ghréas
# NOT le A nó le B; omit "a"
SELECT (Prep) IF (NOT 0 ("a")) (1C (Conj Coord)) (2C (Prep)) (NOT 2
("a"));
SELECT (Prep) IF (NOT 0 ("a")) (-1C (Conj Coord)) (-2C (Prep)) (NOT
-2 ("a"));

# le (simp prep) becomes leis before "an"
# otherwise leis is complex prep or occasionally a noun
# leis/PrepSimp an fhírinne a rá ...

# leis/PrepSimp sin, d'oscail an doras
REMOVE (Prep Simp) IF (0 ("<leis>")) (NOT 1 (Art) OR (Dem));
# new
SELECT (Prep Simp) IF (0 ("<leis>")) (1C (Art) OR (Dem));

# "a" functions as a prep in some phrases listed in A-PREP-PHR
REMOVE (Prep Simp) IF (0 ("a")) ( NOT 1 A-PREP-PHR OR (Verbal Noun
Len) OR (Guess Verbal Noun));
# not necessarily a prep if followed by any type of noun
# e.g. "ar mhaith leat?" cop noun/adj prep-pron

# usually it is a (simple?) prep if followed by an article/det/num
??? or by a noun
# "de/PrepSimp chuid/Noun na/Art gCeilteach/Noun
# "as/PrepSimp Halltsatt/Noun
# BUT:  trí sheisiún could be prep or num
# "a" could be Det Poss or Prep Simp before Noun
# e.g. a shúil ar an pheata madaidh ...
SELECT (Prep Simp) IF (1C NOUN-PREMOD OR (Noun)) (NOT 0 ("<trí>") OR
("<a>"));

# TYPE 4 after diasamb of VN
# it is a (simple?) prep if followed by an verbal noun
# except if preceded by simp prep e.g. do do mholadh
# or i do chodladh
SELECT (Prep Simp) IF (1 (Verbal Noun)) (NOT -1 (Prep Simp));
# chun a fháil amach
SELECT (Part Inf) IF (1 (Verbal Noun)) (-1 ("<chun>"));
# "a (Rel) mbíonn" vs "a (Poss) hintinn"
SELECT (Prep Poss) IF (1C (Noun)) (NOT -1 (Prep Simp));
# but note " le dhá(Num/PrepPoss) fhichid..."
# le(Prep Simp) dhá(Num) bhliain
# BUT do(PrepSimp) mo(Det Poss) chrá
REMOVE (Prep Poss) IF (-1 (Prep Simp));
```

```
# TYPE 1/2
# in copular constructions like Is féidir X; X is frequently a Prep.
# e.g. is féidir linn: linn is Prep not Noun
SELECT (Prep) IF (-1C ("féidir"));

# Thart/timpeall ar etc.
SELECT (Prep Simp) IF (1 (Adv Dir)) ;
# ------------------------------------------------------------------ #
# S2 CONJUNCTIONS
# ------------------------------------------------------------------ #
# Agus is ceart do ...
REMOVE (Conj) IF (0 ("agus")) (-1 ("agus")) ;
# tuairim is 300 bliain: is = conj in this context ...
SELECT (Conj) IF (0 ("agus")) (-1 ("tuairim")) ;
# bán is buí
SELECT (Conj) IF (0 ("agus")) (-1 (Adj)) (1 (Adj));
# Briain is Cormac
SELECT (Conj) IF (0 ("agus")) (-1 (Prop Noun)) (1 (Prop Noun));

# in "mar a bhíodh" mar is conj rather than prep
SELECT (Conj) IF (1 (Part Vb)) (2 (Verb));
# nó "mar/nuair is eol duit", "mar/agus ba cheart"
SELECT (Conj) IF (1 (Cop));

# Ó tháinig ann dó
SELECT (Conj) IF (0 ("ó")) (1C (Verb));

# e.g. óir beidh / óir ní bheidh
SELECT (Conj) IF (0 ("óir")) (1C (Verb));
SELECT (Conj) IF (0 ("óir")) (1 (Part Vb)) (2C (Verb));

# thall is abhus
SELECT (Conj) IF (0 ("agus")) (1 ("abhus"));

# amach is amach
SELECT (Conj) IF (0 ("agus")) (-1 ("amach")) (1 ("amach"));

# trí chéad is a trí, cúig céad is 50
# BUT is 50 duine a bhí ann; include -1 Num also
SELECT (Conj) IF (0 ("agus")) (1 (Part Nm) OR (Num)) (-1 (Num));

# COMMA SEPARATED LISTS
SELECT (Noun) IF (1 (",")) (2 (Noun)) (3 (",")) (4 (Noun));
SELECT (Noun) IF (-1 (",")) (-2 (Noun)) (1 (",")) (2 (Noun));
SELECT (Noun) IF (-1 (",")) (-2 (Noun)) (-3 (",")) (-4 (Noun));

SELECT (Verb) IF (1 (",")) (2 (Verb)) (3 (",")) (4 (Verb));
SELECT (Verb) IF (-1 (",")) (-2 (Verb)) (1 (",")) (2 (Verb));
SELECT (Verb) IF (-1 (",")) (-2 (Verb)) (-3 (",")) (-4 (Verb));

SELECT (Adj) IF (1 (",")) (2 (Adj)) (3 (",")) (4 (Adj));
SELECT (Adj) IF (-1 (",")) (-2 (Adj)) (1 (",")) (2 (Adj));
SELECT (Adj) IF (-1 (",")) (-2 (Adj)) (-3 (",")) (-4 (Adj));

# ------------------------------------------------------------------ #
# S2 DETERMINERS (possessive) -
# ------------------------------------------------------------------ #
# a = "theirs" eclipses following noun,
# a = "his" lenites following noun
# a = "hers" no initial mutation to following noun
# Poss includes Det Poss and Prep Poss
# e.g. lena = Le+Prep+Poss+3P etc...
# á úsáid ...
```

```
# Ecl and Len are not always present on Verbal Nouns - so omit ..
REMOVE (Poss 3P Pl) IF ( NOT 1 (Noun Ecl) OR (Verbal Noun));
REMOVE (Poss 3P Sg Masc) IF ( NOT 1 (Noun Len) OR (Verbal Noun));

SELECT (Poss 3P Pl) IF ( 1C (Noun Ecl) OR (Verbal Noun));
SELECT (Poss 3P Sg Masc) IF ( 1C (Noun Len) OR (Verbal Noun)) (NOT
1C (Noun Voc));
# á hól
SELECT (Poss 3P Sg Fem) IF ( 1C (Noun hPref) OR (Verbal Noun));

# a possessive determiner is not followed by a verb e.g. mo, do etc
# (can be followed by verbal noun - e.g. do mo chrá)
REMOVE (Poss) IF (1C (Verb));

# select the posessive det reading (rather than noun) if followed by
a noun
# e.g. rinneamar ár ndícheall
# (if followed by gen. noun then could be the noun reading of ár)
SELECT (Poss) IF (1C (Noun Com) OR (Verbal Noun));

# ------------------------------------------------------------------ #
# S2 PRONOUNS
# ------------------------------------------------------------------ #
# can't have a pronoun following an article
# e.g tar éis an dó
REMOVE (Pron) IF (-1 (Art));

# cé is interog. pron rather than N or Conj
SELECT (Pron Q) IF (0 ("cé")) (1 ("hé") OR ("hí") OR ("hiad"));

# cé acu/aige/leis/air etc.
SELECT (Pron Q) IF (0 ("cé")) (1 (Pron Prep));

# cé a rinne, a ndéanfadh, a mbíonn etc.
SELECT (Pron Q) IF (1 ("a"));

# cé ba mhó
SELECT (Pron Q) IF (1 (Cop));

# TYPE 1
# Déan seo/sin/siúd
# san  (=sin) can be confused with the prep san so we exclude it
SELECT (Pron Dem) IF (-1C (Verb)) (NOT 0 ("<san>"));
# ------------------------------------------------------------------ #
# S2 INTERROGATIVES
# ------------------------------------------------------------------ #

# select interog reading if at start of sentence
# and sentence ends in a ?
# includes Cop Q and Pron Q ...
# e.g. Cé/Q a bhí ann?
# Nach/NegQ raibh sé ann?
# NOTE: An/Q bhfuil tú a rá nach/Neg raibh sé ann? nach=Neg not NegQ
SELECT (Q) IF (-1 BOS) (*1 (Punct Q));
SELECT (NegQ) IF (-1 BOS) (*1 (Punct Q));

REMOVE (Q)    IF  (NOT *1 (Punct Q));
# following doesnt work over long distances .. so limit it with
punctuation
# also don't have more than two interrogatives ...
# e.g. Cad(=Q) chuige nach n-abrófá amhráin agus tú an fear ceol is
fearr ar(!=Q) an mbaile?
SELECT (Q)    IF  (*1 (Punct Q) BARRIER (Punct Int)) (NOT *-1C (Q));
```

```
# ar/Q chuala/V tú nach/Vb raibh/V sé ann?
# is/Cop maith an rud nár/Vb tháinig/V tú.
SELECT (Vb Neg Rel) IF (*-1C (Verb) OR (Cop)) (1C (Verb)) ;

# Cén duine nár fhoghlaim ar scoil ... ?
# Cén caisleán mór ar chas Bruce Springsteen ann ?
SELECT (Verb) IF (-1 (Rel)) (@1 (Pron Q)) (NOT *-1 (Verb)) (NOT *1
(Verb));

# ------------------------------------------------------------- #
# S2 COPULA
# ------------------------------------------------------------- #
# COP INDEPENDANT = is, an, ar etc
# COP DEPENDANT = gur, nár, arbh etc

# Indep copula doesnt directly FOLLOW a verb
# e.g. cheannaigh sé ar/!=Cop an gcuma sin
# or verbal noun e.g. ag léamh an/!=Cop leabhair
# BUT dependant copular forms which introduce subordinate clauses
# e.g. creidim/V+S gur/Cop fear é
# or verb subject pron (sí, sé, siad)
# e.g. má mheastar/Verb gur/Cop dóigh leis
# a cheapadh/VN gur/Cop chol ceathracha iad
# e.g. má mheas/Verb sí gur/Cop dóigh leis
REMOVE (Cop) IF (NOT 0 (Cop Dep)) (-1C (Verb) OR (Sbj) OR (Verbal
Noun));

# copula doesnt PRECEDE a verb
# or verb subject pron (sí, sé, siad)
REMOVE (Cop) IF (1C (Verb) OR (Sbj) OR (Verbal Noun));

# ============================================================= #
# SECTION 3
CONSTRAINTS
# ============================================================= #

# ------------------------------------------------------------- #
# S3 VERBS
# ------------------------------------------------------------- #
# if one of the possible readings is Verb and it is followed by a
personal
# pronoun e.g. mé, tú etc - then select this reading
# e.g. rinne mé
# Dá ndéanfaí é
# can have genuine nouns before pron e.g. chonaic sé ar scoil mé ...
# Note Ní cabhair é ...
SELECT (Verb) IF (1 (Pron Pers)) (NOT -1 (Cop));
REMOVE (Verb) IF (1 (Pron Pers)) (-1C (Cop));
# ------------------------------------------------------------- #
# S3 VERBAL ADJS
# ------------------------------------------------------------- #
# e.g. déanta de leathar, déanta domA
# but not "imithe chun a pósta/!=VA ar/Prep Thomás"
# "Faoi cheal an oiread sin d'fhios a labhartha/!=VA a/Prep bheith
orra,"
SELECT (Verbal Adj) IF ( 1 (Prep)) (NOT -1 (Det Poss));

SELECT (Verbal Adj) IF ( -1 (Verbal Noun)) ;

# ------------------------------------------------------------- #
# S3 COPULA
# ------------------------------------------------------------- #

# Níl bean ar liosta: ar = prep not cop
```

```
REMOVE (Cop) IF (-2C (Verb)) (-1 (Noun) OR (Sbj));

# Is maith liom
# BUT not nach bhfuil le déanamh
SELECT (Cop) IF ( 1C  (Noun) OR (Adj) ) (2 ("le"));

# It is a copula if there is no verb to the left ...
# and if there is no verb to the right ...

SELECT (Cop) IF (NOT *-1 (Verb) ) (NOT *1 (Verb) );

# if the sentence starts with any form of copula "is" choose this
# reading as long as it is not followed by a verb
# Ní hé la na gaoth ... If it is at the start of a sentence choose
Cop
SELECT (Cop) IF (-1 BOS) (NOT 1 (Verb));

# It is a relative copula if preceded by the subject or direct
object
# e.g.dir. rel: an áit is deise ar domhain , an bhean ab óige
# e.g. indir rel: fear nach cuimhin leis é, an duine ar leis an
teach

SELECT (Cop Rel) IF (0C (Cop)) (*-1 (Noun) OR (Pron) OR (Adj) );
REMOVE (Cop Rel) IF (NOT *-1 (Noun) OR (Pron) OR (Adj) );

# ------------------------------------------------------------------ #
# S3 PREPOSITION
# ------------------------------------------------------------------ #
# e.g. mar b'fhoireann - this mar is a Conj
# mar is gnáth
REMOVE (Prep) IF (1C (Cop));

# some forms are both prep simp and prep pron e.g. faoi, leis etc.
# "chuala sé faoi=prep-pron", but "chuala sé faoi=prep rud=noun"
# de/PrepSimp cheachtar/PronIdf
# this includes possibly "de céard", faoi seo,  etc ...
# liom/leat/leis/Pron Prep féin/PronRef

REMOVE (Pron Prep) IF (1 (Noun) OR (Pron Idf) OR (Pron Dem) OR (Pron
Q));
# de 'thíreolaíocht'
REMOVE (Pron Prep) IF (1 (Punct)) (2 (Art) OR (Noun) OR (Pron Idf)
OR (Pron Dem) OR (Pron Q));

# bhain sé de/PronPrep a chóta BUT  bhain sé geit as/PrepSimp a mhac
SELECT (Pron Prep) IF (NOT 1 (Art) OR (Noun) OR (Pron) OR (Prep) OR
(Det Poss));
SELECT (Pron Prep) IF (1 (Punct)) (NOT 2 (Noun) OR (Pron) OR (Prep)
OR (Det Poss));
# ------------------------------------------------------------------ #
# S3 CONJUNCTION
# ------------------------------------------------------------------ #

SELECT (Conj Coord) IF (1C (Verb) OR (Part Vb));

# e.g. mar atá, mar ba mhaith linn etc.
SELECT (Conj) IF (0 ("mar")) (1 (Rel) OR (Cop));

#Thuig siad go/Conj mbeadh/VerbCond
SELECT (Conj) IF (1 (Verb Cond));

# e.g. cé chomh maith
```

```
# this comes after interogatives are tried "Cé chomh minic a rinne
tú é?"
SELECT (Conj) IF (0 ("cé")) (1 ("chomh")) (NOT *1 (Punct Q));
# ================================================================ #
# SECTION 4
CONSTRAINTS
# ================================================================ #
# ---------------------------------------------------------------- #
# S4 INTERJECTIONS
# ---------------------------------------------------------------- #
# Lets say that interjections like Á must be at the start of a
sentence
# or followed by punct e.g. Á, Ó! etc
REMOVE (Itj) IF (NOT 1 (Punct)) ;
SELECT (Itj) IF (1 (Punct)) ;

# ---------------------------------------------------------------- #
# S4 VERBAL NOUNS
# ---------------------------------------------------------------- #
# ---------------------------------------------------------------- #
# S4 VERBS
# ---------------------------------------------------------------- #
# TYPE 4
# it is a form of verb if it is preceded by a verb part(C)
# d' éag ... -> éag= verb not noun
# a ghabhann ...
SELECT (Verb) IF (-1C (Part Vb));

# ---------------------------------------------------------------- #
# S4 NOUNS
# ---------------------------------------------------------------- #
# TYPE 4
# e.g. dúil acu sa troid - select the noun reading for troid rather
than verb reading
SELECT (Noun) IF (-1C (Prep));
SELECT (NOUN-NOT-VN) IF (-1C (Art));

# TYPE 4
# nouns take com case after prep except for list of preps which take
# gen case, (dat case is handled earlier)
SELECT (Noun Com) IF (-1C (Prep Simp)) (NOT -1 GEN-SIMP-PREP);

# TYPE 4
# arm na Róimhe => arm = Noun Com not Gen
SELECT (Noun Com) IF (1C (Art)) (2C (Noun Gen));

# TYPE 4
# it is a definite noun if preceded by an article
# including e.g. san (Prep Art)
SELECT (Noun DefArt) IF (-1C (Art));

# a gcuid dúnta : dúnta is more likely to be Noun Gen than Adj
# after cuid; so select this reading
SELECT (Noun Gen) IF (-1C ("cuid")) (-2 (Det Poss));

# Cé Árann
SELECT (Prop Noun) IF (1C (Prop Noun)) (NOT 1 (Guess Prop));
# ---------------------------------------------------------------- #
# S4 ADJECTIVE
# ---------------------------------------------------------------- #
# e.g. cultúr na gCeilteach beo: gCeilteach is Gen Weak -> beo is
Adj Weak
SELECT (Adj Weak) IF (-1C (Noun Weak)) ;
SELECT (Adj Strong) IF (-1C (Noun Strong)) ;
```

```
# e.g. ... a bheith cinnte
# BUT give preference to Verbal Adj
SELECT (Adj Base) IF (NOT 0 (Verbal Adj)) (-1C (Verbal Noun)) ;

# Ní mór .. nach maith , nach mór
# ================================================================ #
# SECTION 5
CONSTRAINTS
# ================================================================ #

# ---------------------------------------------------------------- #
# S5 PREPOSITIONS
# ---------------------------------------------------------------- #
# e.g. mar sin/seo = Simp Prep + Pron Dem
# leis sin, as sin amach
SELECT (Prep Simp) IF (1 (Pron Dem));
SELECT (Pron Dem) IF (-1 (Prep Simp));

# ---------------------------------------------------------------- #
# S5 VERBS
# ---------------------------------------------------------------- #
# If there are a choice of inflected verb forms remaining - choose
the autonomous one ....
SELECT (Verb Auto) IF (0C (Verb)) (NOT 1 (Sbj) OR (Prop Noun));
# ---------------------------------------------------------------- #
# S5 NOUNS
# ---------------------------------------------------------------- #
# common noun is lenited only after
# poss dets (mo/do/a theach)
# copula (ba dhuine mór é)
# numeral (dhá theach)
# prep simp (de chrann)
# prep poss (lena stór)
REMOVE (Noun Com Len) IF (NOT -1 (Poss Sg) OR (Cop) OR NUM-LEN OR
(Prep Simp));
# ar an ngaoth
REMOVE (Noun Com Ecl) IF (NOT -1 (Poss Pl) OR NUM-ECL OR ("i") OR
("<an>"));

# ar an ngalfchúrsa
REMOVE (Noun Gen) IF (-1 ("an")) (-2 (Prep Simp));

SELECT (Noun Com Len) IF (-1 (Det Poss Sg Masc));
SELECT (Noun Com Ecl) IF (-1 (Det Poss Pl));

# "a deir" - direct relative verbal particle
SELECT (Direct) IF (1C ("abair"));

# ag aisteoireacht, ag léim etc where lenited and unlenited are same
REMOVE (Len) IF (0C (Verbal Noun)) (NOT -1 (Poss) OR ("a"));
REMOVE (Ecl) IF (0C (Verbal Noun)) (NOT -1 (Poss) OR ("a"));

# ---------------------------------------------------------------- #
# S5 ADJS
# ---------------------------------------------------------------- #
# when there is a choice following a Cop it is "usually" Adj
# Is deas/noun/adj/subst an lá
SELECT (Adj) IF (-1C (Cop));

# ---------------------------------------------------------------- #
# S5 TIDY UP
# ---------------------------------------------------------------- #
# Mí as a proper noun is always preceded by "na"
```

```
# e.g. muintir na Mí, contae na Mí etc.
REMOVE (Prop Noun) IF (0 ("<Mí>")) (NOT -1 ("<na>"));
SELECT (Prop Noun) IF (0 ("<Mí>")) (-1 ("<na>"));

# Ní náire feacadh i láthair Dé... remove the N and VN reading for
ní when followed by another noun ... this is not necessarily always
correct ...
REMOVE (Noun) IF (0 ("<Ní>")) (1C (Noun));

#=====#
END #
#=====#
```

Appendix E: Test Suite Sentences

Table of Contents

Test Suite Sentences 2
Gold Standard Dependency Annotated and Chunked Test Suite Sentences 6

Test Suite Sentences

1. An fíor é?
2. An í Eilís an bainisteoir?
3. An iad na daoine siúd na buaiteoirí?
4. An lá a bádh é.
5. An lá a cuireadh Butt ...
6. An leabhar.
7. An leat an teach?
8. An leatsa an teach?
9. An tusa an múinteoir?
10. An tusa Briain?
11. Ar ith sí an dinnéar?
12. Ar labhair Seán?
13. Ar mhaith leat teach a cheannach.
14. Ar thug sí an leabhar do Mháire?
15. Arbh é é a bhí ann?
16. Arbh é Seán a bhí ann?
17. Ba mhaith liom cáca a dhéanamh.
18. Ba mhaith liom fanacht.
19. Ba mhaith liom gan cáca a dhéanamh.
20. Ba mhaith liom gan fanacht.
21. Ba mhaith liom teach a cheannach.
22. B'fhearr liom é.
23. B'fhearr liom gan cáca a dhéanamh.
24. B'fhearr liom gan fanacht.
25. Bhí an fear ag an doras.
26. Bhí an geata dúnta ag Seán.
27. Bhí an geata dúnta.
28. Bhí an t-airgead ag Seán.
29. Bhí rí ann fadó.
30. Bhí sé ar snámh.
31. Bhí sé thar cinn.
32. Bhí sí ar buille agus ar mire.
33. Bhíodh bríste fada ann chomh maith le bríste glúnach.
34. Bhíomar tinn inné.
35. Briseadh an fhuinneog leis an stoirm.
36. Cá bhfuil sé ag tógáil na móna?
37. Cá ndeachaigh sé?
38. Cad a d'ith sí?
39. Cad a thug sí do Mháire?
40. Cailín is ea í.
41. Cár cheannaigh sé an leabhar?
42. Cé a d'ith an leon?
43. Cé a dtug an leabhar do Mháire?
44. Cé a labhair?
45. Cé dó a thug sí an leabhar?
46. Cé leis an teach?
47. Cé nár ith an dinnéar?
48. Cé nár labhair?
49. Cé nár thug an leabhar do Mháire?
50. Cén chaoi a rinne sé é?
51. Cén fáth a ndeachaigh sé amach?
52. Chaoin sé le háthas.

53. Cheannaigh sé leabhar áit a bhí sé ar fáil.
54. Cheannaigh sé leabhar áit a bhí siad ar fáil.
55. Cheannaigh sé leabhar anseo.
56. Cheannaigh sé leabhar ins an siopa.
57. Cheannaigh sé úll mór agus oráiste beag.
58. Cheannaigh Seán leabhar agus léigh sé é.
59. Chonaic Máire an fear a bhí ag iascaireacht.
60. Chonaic Máire gur ag iascaireacht a bhí an fear.
61. Chonaic mé Seán ag oscailt an dorais.
62. Chuaigh sé abhaile nuair a bhí an cóisir thart.
63. Chuaigh sé amach chun bainne a fháil.
64. Chuaigh sé isteach.
65. Chuaigh sí amach faoi dheifir.
66. Conas a chaoin sé?
67. Conas a labhair sé?
68. Conas atá sé?
69. Conas atá sé ag rith?
70. Dá mba mise thú ní dhéanfainn é.
71. Daoine nach iad.
72. *D'éirigh an mac léinn leis sa scrúdú.
73. D'éirigh go maith leis an mac léinn sa scrúdú.
74. D'éirigh leis an mac léinn sa scrúdú.
75. D'éirigh sa scrúdú leis an mac léinn.
76. Deisíodh an rothar ag Seán.
77. Deisíodh an rothar.
78. D'fhág an bád a chonaic mac an fhir.
79. D'fhág an fear a chonaic a mhac an bád.
80. D'fhan an fear a bhuaigh an crannchur.
81. D'fhan an fear a d'ionsaigh iad.
82. D'fhan an fear a d'ionsaigh siad.
83. D'fhan sé ansin go ciúin ins an seomra ar feadh leath uair a chloig nuair a bhí tuirse air.
84. D'fhan sé ansin inné.
85. D'fhan sé ansin le fiche bliain.
86. D'fhan sé ansin nuair a bhí sé dorcha.
87. Dheisigh Seán an rothar.
88. D'ith sí an dinnéar.
89. D'ith.
90. D'itheamar an dinnéar.
91. Dúirt sé go dtabharfaidh an bhean an leabhar do Mháire.
92. Dúirt sé go rachadh sé.
93. Dúirt sé gur múinteoir é.
94. Dúirt sé gur thug sí an leabhar do Mháire.
95. Dúirt sé nach múinteoir é.
96. Dúirt sé nár múinteoir é.
97. Dúirt sé nár thug sí an leabhar do Mháire.
98. Dúirt siad nach bhfeiceann siad an cineál seo chomh minic sin.
99. Fear maith is ea é.
100. go deo, go bás, go brách.
101. Íocfaidh mé as a gceannóidh tú.
102. Is ag cabhrú liom atá sé.
103. Is ag déanamh cáca atá mé.
104. Is ag iascaireacht atá sé.
105. Is airde sliabh ná cnoc.
106. Is amhlaidh a bhídís ag obair do na feirmeoirí.
107. Is an leabhar a thug sí do Mháire.
108. Is beag planda a fhásann i dteocht faoi bhun 4C.
109. Is cáca atá a dhéanamh agam.
110. Is cailín í.
111. Is deas an lá é.
112. Is deas an lá.
113. Is do Mháire a thug sí an leabhar.
114. Is do Mháire a thug sí leabhar.

115. Is eisean atá ag cabhrú liom.
116. Is eisean atá do mo cabhrú.
117. Is fear maith é.
118. Is fearr liom úlla ná oráistí.
119. Is í an líne glas teorainn an cheantair.
120. Is ise a thug an leabhar do Mháire.
121. Is ise a thug leabhar do Mháire.
122. Is lá deas é.
123. Is le Dónal an teach.
124. Is leabhar a thug sí do Mháire.
125. Is maith liom úlla agus oráistí.
126. Is mise atá ag déanamh cáca.
127. Is mise Briain.
128. Is múinteoir é.
129. Is múinteoir Seán.
130. Ith an dinnéar.
131. Labhair go soiléir.
132. Labhair sé os ard.
133. Labhair Seán.
134. Labhair.
135. Labhraíomar.
136. Labhraítear go soiléir.
137. le tamall, le fada, le seachtain.
138. Líonadh an poll le clocha.
139. Má bhíonn an t-am agat, déan é.
140. Máire.
141. Nach é é a bhí ann?
142. Nach tusa an múinteoir?
143. Nár ith sí an dinnéar?
144. Nár labhair Seán.
145. Nár thug sí an leabhar do Mháire?
146. Ní gorm atá sé.
147. Ní hé Briain an múinteoir.
148. Ní hé nár mhaith liom é.
149. Ní liomsa an t-airgead.
150. Ní mór dúinn aonad a bheith againn.
151. Níl an cinneadh déanta fós.
152. Níor ith sí an dinnéar.
153. Níor ith.
154. Níor labhair Seán.
155. Níor labhair.
156. Níor tháinig sé go fóill.
157. Níor thug sí an leabhar do Mháire.
158. Níor thug.
159. Níorbh é.
160. Rinne sé é go maith.
161. Rith sé le luas lasrach.
162. Rud ab fhusa a dhéanamh ...
163. Seán.
164. Seo an bád a chonaic an fear.
165. Seo an fear a bhuaigh an crannchur.
166. Seo an fear a chonaic an bád.
167. Seo an fear a chonaic an bhean.
168. Seo an fear a d'ionsaigh iad.
169. Seo an fear a d'ionsaigh siad.
170. Sin a bhfuil ann.
171. Sin an fear a bhfuil a mhac ag imeacht.
172. Sin an fear a bhfuil a mhac tinn.
173. Sin an fear a chuireann síol.
174. Sin an fear a phléasc.
175. Sin an gort a cuireadh an síol ann.
176. Sin an síol a chuireann fear.
177. Sin an síol a cuireadh.

178. Sin an té a itheann feoil.
179. Sin an teach a raibh sé ina chónaí ann.
180. Sise.
181. Tá an cáca arna dhéanamh agam.
182. Tá an carr sa gharáiste.
183. Tá an doras ar oscailt.
184. Tá an leabhar go maith.
185. Tá an leabhar léite agam.
186. Tá an pictiúir péinteáilte ag Mary.
187. Tá áthas orm.
188. Tá cáca á dhéanamh agam.
189. Tá cáca le déanamh agam.
190. Tá cuimhne mhaith agam chomh cruaidh agus a bhí sé.
191. Tá mé ag déanamh cáca.
192. Tá mé i ndiaidh cáca a dhéanamh.
193. Tá mé tar éis cáca a dhéanamh.
194. Tá ocras orm.
195. Tá sé ag cabhrú liom.
196. Tá sé ag caoineadh gan stad.
197. Tá sé ag dul a chodladh.
198. Tá sé ag iascaireacht.
199. Tá sé ag rith go tapaidh.
200. Tá sé ag tógáil isteach na móna.
201. Tá sé ag tógáil na móna isteach.
202. Tá sé déanta.
203. Tá sé do mo chabhrú.
204. Tá sé go hálainn.
205. Tá sé ina chodladh.
206. Tá sé ina mhúinteoir.
207. Tá sé le teacht.
208. Tá sé mór.
209. Tá sé thíos staighre.
210. Tabhair an leabhar do Mháire.
211. Táim chun cáca a dhéanamh inniu.
212. Tar éis trí lá tháinig sé abhaile.
213. Tháinig sé abhaile an oíche sin.
214. Tháinig sé abhaile tar éis trí lá.
215. Thaistil Eoin ní ba mhó ná aon duine eile.
216. Thóg sé isteach an mhóin.
217. Thug Seán Máire leabhar.
218. Thug Seán Ó Broin leabhar do Mháire.
219. Thug sí an leabhar do Mháire.
220. Thug sí leabhar do Mháire.
221. Thug.
222. Thugamar an leabhar do Mháire.
223. Títhe lucht oibre ba mhó a bhí ann.
224. Tóg go bog é.
225. Tuigeann Nollaig níos mó ná Seán.

Gold Standard Dependency Annotated and Chunked Test Suite Sentences

1. [S [COP An is+Cop+Pres+Q+@COP_WH] [PRED fíor fíor+Adj+Base+@PRED] [NP é é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]
2. [S [COP An is+Cop+Pres+Q+@COP_WH] [NP í í+Pron+Pers+3P+Sg+Fem+@AUG>SUBJ Eilís Eilís+Prop+Noun+Fem+Com+Sg+@SUBJ NP] [PRED an an+Art+Sg+Def+@>N bainisteoir bainisteoir+Noun+Masc+Com+Sg+DefArt+@PRED] ? ?+Punct+Fin+Q+<<< S]
3. [S [COP An is+Cop+Pres+Q+@COP_WH] [NP iad iad+Pron+Pers+3P+Pl+@AUG>SUBJ na na+Art+Pl+Def+@>N daoine duine+Noun+Masc+Com+Pl+DefArt+@SUBJ siúd siúd+Det+Dem+@N< NP] [PRED na na+Art+Pl+Def+@>N buaiteoirí buaiteoir+Noun+Masc+Com+Pl+DefArt+@PRED] ? ?+Punct+Fin+Q+<<< S]
4. [S [AD An an+Art+Sg+Def+@>N lá lá+Noun+Masc+Com+Sg+DefArt+@ADVL] [VS a a+Part+Vb+Rel+Indirect+@>V bádh báigh+Verb+VT+PastInd+Auto+@FMV_REL_SUBJ] [NP é é+Pron+Pers+3P+Sg+Masc+@OBJ NP] . .+Punct+Fin+<<< S]
5. [S [AD An an+Art+Sg+Def+@>N lá lá+Noun+Masc+Com+Sg+DefArt+@ADVL] [VS a a+Part+Vb+Rel+Indirect+@>V cuireadh cuir+Verb+VTI+PastInd+Auto+@FMV_REL_SUBJ] [NP Butt Butt+Prop+Noun+Masc+Com+Sg+@OBJ NP] ... ...+Punct+Fin+<<< S]
6. [S [NP An an+Art+Sg+Def+@>N leabhar leabhar+Noun+Masc+Com+Sg+DefArt+@NP NP] . .+Punct+Fin+<<< S]
7. [S [COP An is+Cop+Pres+Q+@COP_WH] [PP leat le+Pron+Prep+2P+Sg+@PP_PRED PP] [NP an an+Art+Sg+Def+@>N teach teach+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]
8. [S [COP An is+Cop+Pres+Q+@COP_WH] [PP leatsa le+Pron+Prep+2P+Sg+Emph+@PP_PRED PP] [NP an an+Art+Sg+Def+@>N teach teach+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]
9. [S [COP An is+Cop+Pres+Q+@COP_WH] [NP tusa tú+Pron+Pers+2P+Sg+Emph+@SUBJ NP] [PRED an an+Art+Sg+Def+@>N múinteoir múinteoir+Noun+Masc+Com+Sg+DefArt+@PRED] ? ?+Punct+Fin+Q+<<< S]
10. [S [COP An is+Cop+Pres+Q+@COP_WH] [NP tusa tú+Pron+Pers+2P+Sg+Emph+@SUBJ NP] [PRED Briain Briain+Prop+Noun+Masc+Com+Sg+@PRED] ? ?+Punct+Fin+Q+<<< S]
11. [S [V Ar ar+Part+Vb+Q+Past+@>V ith ith+Verb+VTI+Vow+PastInd+Len+@FMV] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an an+Art+Sg+Def+@>N dinnéar dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] ? ?+Punct+Fin+Q+<<< S]
12. [S [V Ar ar+Part+Vb+Q+Past+@>V labhair labhair+Verb+VTI+PastInd+Q+Len+@FMV] [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]
13. [S [COP Ar is+Cop+Past+RelInd+@COP] [PRED mhaith maith+Adj+Base+Len+@PRED] [PP leat le+Pron+Prep+2P+Sg+@PP_SUBJ PP] [INF [OI teach teach+Noun+Masc+Com+Sg+@OBJ_INF OI] [I a a+Part+Inf+@>N

```
          cheannach ceannach+Verbal+Noun+VTI+Len+@INF  I]  INF] .
          .+Punct+Fin+<<< S]
14.  [S  [V Ar ar+Part+Vb+Q+Past+@>V thug
          tabhair+Verb+VD+PastInd+Len+@FMV  ] [NP sí
          sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [NP an
          an+Art+Sg+Def+@>N leabhar
          leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] [PP do
          do+Prep+Simp+@PP_OBL [NP Mháire
          Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] ?
          ?+Punct+Fin+Q+<<< S]
15.  [S  [COP Arbh is+Cop+Past+Q+VF+@COP_WH ] [NP é
          é+Pron+Pers+3P+Sg+Masc+@AUG>SUBJ é
          é+Pron+Pers+3P+Sg+Masc+@SUBJ  NP] [V a
          a+Part+Vb+Rel+Direct+@>V bhí
          bí+Verb+VI+PastInd+Len+@FMV_REL  ] [PP ann
          i+Pron+Prep+3P+Sg+Masc+@PP_ADVL  PP] ? ?+Punct+Fin+Q+<<< S]
16.  [S  [COP Arbh is+Cop+Past+Q+VF+@COP_WH ] [NP é
          é+Pron+Pers+3P+Sg+Masc+@AUG>SUBJ Seán
          Seán+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] [V a
          a+Part+Vb+Rel+Direct+@>V bhí
          bí+Verb+VI+PastInd+Len+@FMV_REL  ] [PP ann
          i+Pron+Prep+3P+Sg+Masc+@PP_ADVL  PP] ? ?+Punct+Fin+Q+<<< S]
17.  [S  [COP Ba is+Cop+Cond+@COP ] [PRED mhaith
          maith+Adj+Base+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF [OI cáca
          cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
          dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF] .
          .+Punct+Fin+<<< S]

18.  [S  [COP Ba is+Cop+Cond+@COP ] [PRED mhaith
          maith+Adj+Base+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF [I fanacht
          fanacht+Verbal+Noun+VI+@INF  I]  INF] . .+Punct+Fin+<<< S]
19.  [S  [COP Ba is+Cop+Cond+@COP ] [PRED mhaith
          maith+Adj+Base+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF gan
          gan+Prep+Simp+@PP_NEG [OI cáca
          cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
          dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF] .
          .+Punct+Fin+<<< S]
20.  [S  [COP Ba is+Cop+Past+Rel+@COP ] [PRED mhaith
          maith+Adj+Base+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF gan
          gan+Prep+Simp+@PP_NEG [I fanacht
          fanacht+Verbal+Noun+VI+@INF  I]  INF] . .+Punct+Fin+<<< S]
21.  [S  [COP Ba is+Cop+Cond+@COP ] [PRED mhaith
          maith+Adj+Base+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF [OI teach
          teach+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
          cheannach ceannach+Verbal+Noun+VTI+Len+@INF  I]  INF] .
          .+Punct+Fin+<<< S]
22.  [S  [COP B' is+Cop+Past+VF+@COP ] [PRED fhearr
          maith+Adj+Comp+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [NP é
          é+Pron+Pers+3P+Sg+Masc+@OBJ  NP] . .+Punct+Fin+<<< S]
23.  [S  [COP B' is+Cop+Past+VF+@COP ] [PRED fhearr
          maith+Adj+Comp+Len+@PRED ] [PP liom
          le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [INF gan
          gan+Prep+Simp+@PP_NEG [OI cáca
          cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
          dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF] .
          .+Punct+Fin+<<< S]
```

24. [S [COP B' is+Cop+Past+VF+@COP] [PRED fhearr maith+Adj+Comp+Len+@PRED] [PP liom le+Pron+Prep+1P+Sg+@PP_SUBJ PP] [INF gan gan+Prep+Simp+@PP_NEG [I fanacht fanacht+Verbal+Noun+VI+@INF I] INF] . .+Punct+Fin+<<< S]
25. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [PP ag ag+Prep+Simp+@PP_ADVL [NP an an+Art+Sg+Def+@>N doras doras+Noun+Masc+Com+Sg+DefArt+@P< NP] PP] . .+Punct+Fin+<<< S]
26. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N geata geata+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [PRED dúnta dúnta+Verbal+Adj+@PRED] . .+Punct+Fin+<<< S]
27. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N geata geata+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [PRED dúnta dúnta+Verbal+Adj+@PRED] [PP ag ag+Prep+Simp+@PP_HAS [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@P< NP] PP] . .+Punct+Fin+<<< S]
28. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N t-airgead airgead+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [PP ag ag+Prep+Simp+@PP_HAS [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@P< NP] PP] . .+Punct+Fin+<<< S]
29. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP rí rí+Noun+Masc+Com+Sg+@SUBJ NP] [PP ann i+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] [AD fadó fadó+Adv+Gn+@ADVL] . .+Punct+Fin+<<< S]
30. [S [V Bhí bí+Verb+VI+PastInd+Len+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ar ar+Prep+Simp+@PP_STAT [NP snámh snámh+Verbal+Noun+VTI+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
31. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [PP thar thar+Prep+Simp+@PP_ADVL [NP cinn ceann+Noun+Masc+Com+Pl+@P< NP] PP] . .+Punct+Fin+<<< S]
32. [S [V Bhí bí+Verb+VI+PastInd+Len+@FMV] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [PP ar ar+Prep+Simp+@PP_ADVL [NP buille buille+Noun+Masc+Com+Sg+@P< NP] PP] [CJ2 agus agus+Conj+Coord+@CC [PP ar ar+Prep+Simp+@PP_ADVL [NP mire mire+Noun+Fem+Com+Sg+@P< NP] PP] CJ2] . .+Punct+Fin+<<< S]
33. [S [V Bhíodh bí+Verb+VI+PastImp+Len+@FMV] [NP bríste bríste+Noun+Masc+Com+Sg+@SUBJ fada fada+Adj+Base+@N< NP] [PP ann i+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] [AD chomh chomh+Adv+Its+@>ADJ maith maith+Adj+Base+@ADVL] [PP le le+Prep+Simp+@PP_ADVL [NP bríste bríste+Noun+Masc+Com+Sg+@P< glúnach glúnach+Guess+Adj+Base+@N< NP] PP] . .+Punct+Fin+<<< S]
34. [S [VS Bhíomar bí+Verb+VI+PastInd+1P+Pl+Len+@FMV_SUBJ] [PRED tinn tinn+Adj+Base+@PRED] [AD inné inné+Adj+Base+@ADVL] . .+Punct+Fin+<<< S]
35. [S [VS Briseadh bris+Verb+VTI+PastInd+Auto+@FMV_SUBJ] [NP an an+Art+Sg+Def+@>N fhuinneog fuinneog+Noun+Fem+Com+Sg+DefArt+@OBJ NP] [PP leis le+Prep+Simp+@PP_ADVL [NP an an+Art+Sg+Def+@>N stoirm stoirm+Noun+Fem+Com+Sg+DefArt+@P< NP] PP] . .+Punct+Fin+<<< S]
36. [S [AD Cá cá+Adv+Q+@ADVL] [V bhfuil bí+Verb+VI+PresInd+Dep+Q+Ecl+@FAUX] [NP sé

sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag
ag+Prep+Simp+@PP_ASP [NP tógáil tógáil+Verbal+Noun+VTI+@P<
NP] PP-ASP] [OA na na+Art+Gen+Sg+Def+Fem+@>N móna
móin+Noun+Fem+Gen+Sg+DefArt+@OBJ_ASP OA] ASP] ?
?+Punct+Fin+Q+<<< S]

37. [S [AD Cá cá+Adv+Q+@ADVL] [V ndeachaigh
téigh+Verb+VTI+PastInd+Dep+Q+Ecl+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] ? ?+Punct+Fin+Q+<<<
S]
38. [S [NP Cad cad+Pron+Q+@OBJ NP] [V a a+Part+Vb+Rel+Direct+@>V
d' do+Part+Vb+@>V ith ith+Verb+VTI+Vow+PastInd+Len+@FMV_REL
] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] ?
?+Punct+Fin+Q+<<< S]
39. [S [NP Cad cad+Pron+Q+@OBJ NP] [V a a+Part+Vb+Rel+Direct+@>V
thug tabhair+Verb+VD+PastInd+Len+@FMV_REL] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] ?
?+Punct+Fin+Q+<<< S]
40. [S [PRED Cailín Cailín+Prop+Noun+Masc+Com+Sg+@PRED] [COP is
is+Cop+Pres+Rel+@COP] [NP ea ea+Pron+Pers+3P+Sg+@AUG>SUBJ
í í+Pron+Pers+3P+Sg+Fem+@SUBJ NP] . .+Punct+Fin+<<< S]
41. [S [AD Cár cá+Adv+Q+Past+@ADVL] [V cheannaigh
ceannaigh+Verb+VTI+PastInd+Q+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] ?
?+Punct+Fin+Q+<<< S]
42. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V a
a+Part+Vb+Rel+Direct+@>V d' do+Part+Vb+@>V ith
ith+Verb+VTI+Vow+PastInd+Len+@FMV_REL] [NP an
an+Art+Sg+Def+@>N leon
leon+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ NP] ?
?+Punct+Fin+Q+<<< S]
43. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V a
a+Part+Vb+Rel+Direct+@>V thug
tabhair+Verb+VD+PastInd+Len+@FMV_REL] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] ?
?+Punct+Fin+Q+<<< S]
44. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V a
a+Part+Vb+Rel+Direct+@>V labhair
labhair+Verb+VTI+PastInd+Len+@FMV_REL] ?
?+Punct+Fin+Q+<<< S]
45. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [PP dó
do+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] [V a
a+Part+Vb+Rel+Indirect+@>V dtug
tabhair+Verb+VD+PastInd+Ecl+@FMV_REL] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] ?
?+Punct+Fin+Q+<<< S]
46. [S [COP Cé cé+Cop+Pro+Q+@COP_WH] [PP leis
le+Pron+Prep+3P+Sg+Masc+@PP_PRED PP] [NP an
an+Art+Sg+Def+@>N teach teach+Noun+Masc+Com+Sg+DefArt+@SUBJ
NP] ? ?+Punct+Fin+Q+<<< S]
47. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V nár
nár+Part+Vb+Neg+Rel+Past+@>V ith
ith+Verb+VTI+Vow+PastInd+Neg+Len+@FMV_REL] [NP an
an+Art+Sg+Def+@>N dinnéar

dinnéar+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ NP] ?
?+Punct+Fin+Q+<<< S]
48. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V nár
nár+Part+Vb+Neg+Rel+Past+@>V labhair
labhair+Verb+VTI+PastInd+NegQ+Len+@FMV_REL] ?
?+Punct+Fin+Q+<<< S]
49. [S [NP Cé cé+Pron+Q+@SUBJ_OR_OBJ NP] [V nár
nár+Part+Vb+Neg+Rel+Past+@>V thug
tabhair+Verb+VD+PastInd+Neg+Len+@FMV_REL] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] ?
?+Punct+Fin+Q+<<< S]
50. [S [COP Cén cé+Cop+Pro+Q+Art+Sg+@COP_WH] [PRED chaoi
caoi+Noun+Fem+Com+Sg+DefArt+@PRED] [V a
a+Part+Vb+Rel+Direct+@>V rinne
déan+Verb+VT+PastInd+Len+@FMV_REL] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP é
é+Pron+Pers+3P+Sg+Masc+@OBJ NP] ? ?+Punct+Fin+Q+<<< S]
51. [S [COP Cén cé+Cop+Pro+Q+Art+Sg+@COP_WH] [PRED fáth
fáth+Noun+Masc+Com+Sg+DefArt+@PRED] [V a
a+Part+Vb+Rel+Indirect+@>V ndeachaigh
téigh+Verb+VTI+PastInd+Dep+Ecl+@FMV_REL] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD amach
amach+Adv+Dir+@ADVL] ? ?+Punct+Fin+Q+<<< S]
52. [S [V Chaoin caoin+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [PP le
le+Prep+Simp+@PP_ADVL [NP háthas
áthas+Noun+Masc+Com+Sg+hPref+@P< NP] PP] .
.+Punct+Fin+<<< S]
53. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP leabhar
leabhar+Noun+Masc+Com+Sg+@OBJ NP] [CB áit
áit+Conj+Subord+@CLB] [V a a+Part+Vb+Rel+Direct+@>V bhí
bí+Verb+VI+PastInd+Len+@FAUX_REL] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ar
ar+Prep+Simp+@PP_STAT [NP fáil fáil+Verbal+Noun+VT+@P< NP]
PP-ASP] ASP] . .+Punct+Fin+<<< S]
54. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP leabhar
leabhar+Noun+Masc+Com+Sg+@OBJ NP] [CB áit
áit+Conj+Subord+@CLB] [V a a+Part+Vb+Rel+Direct+@>V bhí
bí+Verb+VI+PastInd+Len+@FAUX_REL] [NP siad
siad+Pron+Pers+3P+Pl+Sbj+@SUBJ NP] [ASP [PP-ASP ar
ar+Prep+Simp+@PP_STAT [NP fáil fáil+Verbal+Noun+VT+@P< NP]
PP-ASP] ASP] . .+Punct+Fin+<<< S]
55. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP leabhar
leabhar+Noun+Masc+Com+Sg+@OBJ NP] [AD anseo
anseo+Adv+Loc+@ADVL] . .+Punct+Fin+<<< S]
56. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP leabhar
leabhar+Noun+Masc+Com+Sg+@OBJ NP] [PP ins
i+Prep+Art+Sg+@PP_ADVL [NP an an+Art+Sg+Def+@>N siopa
siopa+Noun+Masc+Com+Sg+DefArt+@P< NP] PP] .
.+Punct+Fin+<<< S]
57. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP úll
úll+Noun+Masc+Com+Sg+@OBJ mór mór+Adj+Masc+Com+Sg+@N< NP]
[CJ2 agus agus+Conj+Coord+@CC [NP oráiste
oráiste+Noun+Masc+Com+Sg+@OBJ beag beag+Adj+Masc+Com+Sg+@N<
NP] CJ2] . .+Punct+Fin+<<< S]

58. [S [V Cheannaigh ceannaigh+Verb+VTI+PastInd+Len+@FMV] [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@SUBJ NP] [NP leabhar leabhar+Noun+Masc+Com+Sg+@OBJ NP] [CB agus agus+Conj+Coord+@CLB] [V léigh léigh+Verb+VTI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [NP é é+Pron+Pers+3P+Sg+Masc+@OBJ NP] . .+Punct+Fin+<<< S]
59. [S [V Chonaic feic+Verb+VTI+PastInd+Len+@FMV] [NP Máire Máire+Prop+Noun+Fem+Com+Sg+@SUBJ NP] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ_REL NP] [V a a+Part+Vb+Rel+Direct+@>V bhí bí+Verb+VI+PastInd+Len+@FAUX_REL] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP iascaireacht iascaireacht+Verbal+Noun+NStem+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
60. [S [V Chonaic feic+Verb+VTI+PastInd+Len+@FMV] [NP Máire Máire+Prop+Noun+Fem+Com+Sg+@SUBJ NP] [CB gur is+Cop+Past+Dep+@CLB] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP iascaireacht iascaireacht+Verbal+Noun+NStem+@P< NP] PP-ASP] ASP] [V a a+Part+Vb+Rel+Direct+@>V bhí bí+Verb+VI+PastInd+Len+@FAUX_REL] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] . .+Punct+Fin+<<< S]
61. [S [V Chonaic feic+Verb+VTI+PastInd+Len+@FMV] [NP mé mé+Pron+Pers+1P+Sg+@SUBJ NP] [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@SUBJ_ASP NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP oscailt oscailt+Verbal+Noun+VTI+@P< NP] PP-ASP] [OA an an+Art+Sg+Def+@>N dorais doras+Noun+Masc+Gen+Sg+@OBJ_ASP OA] ASP] . .+Punct+Fin+<<< S]
62. [S [V Chuaigh téigh+Verb+VTI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD abhaile abhaile+Adv+Dir+@ADVL] [CB nuair nuair+Conj+Subord+@CLB] [V a a+Part+Vb+Rel+Direct+@>V bhí bí+Verb+VI+PastInd+Len+@FMV_REL] [NP an an+Art+Sg+Def+@>N cóisir cóisir+Noun+Fem+Com+Sg+@SUBJ NP] [AD thart thart+Adv+Dir+@ADVL] . .+Punct+Fin+<<< S]
63. [S [V Chuaigh téigh+Verb+VTI+PastInd+Len+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD amach amach+Adv+Dir+@ADVL] [ASP chun chun+Prep+Simp+@PP_ASP [INF [OI bainne bainne+Noun+Masc+Com+Sg+@OBJ_INF OI] [I a a+Part+Inf+@>N fháil fáil+Verbal+Noun+VT+Len+@INF I] INF] ASP] . .+Punct+Fin+<<< S]
64. [S [V Chuaigh téigh+Verb+VTI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD isteach isteach+Adv+Dir+@ADVL] . .+Punct+Fin+<<< S]
65. [S [V Chuaigh téigh+Verb+VTI+PastInd+Len+@FMV] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [AD amach amach+Adv+Dir+@ADVL] [PP faoi faoi+Prep+Simp+@PP_ADVL [NP dheifir deifir+Noun+Fem+Com+Sg+Len+@P< NP] PP] . .+Punct+Fin+<<< S]
66. [S [AD Conas conas+Adv+Q+@ADVL] [V a a+Part+Vb+Rel+Direct+@>V chaoin caoin+Verb+VTI+PastInd+Len+@FMV_REL] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]
67. [S [AD Conas conas+Adv+Q+@ADVL] [V a a+Part+Vb+Rel+Direct+@>V labhair labhair+Verb+VTI+PastInd+Len+@FMV_REL] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] ? ?+Punct+Fin+Q+<<< S]

```
68. [S  [AD Conas conas+Adv+Q+@ADVL  ] [V atá
        bí+Verb+VI+PresInd+Rel+@FMV_REL  ] [NP sé
        sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] ? ?+Punct+Fin+Q+<<<
        S]
69. [S  [AD Conas conas+Adv+Q+@ADVL  ] [V atá
        bí+Verb+VI+PresInd+Rel+@FAUX_REL  ] [NP sé
        sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [ASP [PP-ASP ag
        ag+Prep+Simp+@PP_ASP [NP rith rith+Verbal+Noun+VTI+@P<  NP]
        PP-ASP]  ASP] ? ?+Punct+Fin+Q+<<< S]
70. [S  [CB Dá dá+Conj+Subord+@CLB ] [COP mba is+Cop+Cond+Ecl+@COP
        ] [NP mise mé+Pron+Pers+1P+Sg+Emph+@SUBJ  NP] [PRED thú
        tú+Pron+Pers+2P+Sg+Len+@PRED ] [VS ní ní+Part+Vb+Neg+@>V
        dhéanfainn déan+Verb+VT+Cond+1P+Sg+Neg+Len+@FMV_SUBJ  ] [NP
        é é+Pron+Pers+3P+Sg+Masc+@OBJ  NP] . .+Punct+Fin+<<< S]
71. [S  [PRED Daoine duine+Noun+Masc+Com+Pl+@PRED ] [COP nach
        is+Cop+Pres+Rel+Neg+@COP ] [NP iad
        iad+Pron+Pers+3P+Pl+@SUBJ  NP] . .+Punct+Fin+<<< S]
72. [S  [V D' do+Part+Vb+@>V éirigh
        éirigh+Verb+VI+Vow+PastInd+Len+@FMV  ] [NP an
        an+Art+Sg+Def+@>N mac mac+Noun+Masc+Com+Sg+DefArt+@SUBJ
        léinn léann+Noun+Masc+Gen+Sg+@N<  NP] [PP leis
        le+Pron+Prep+3P+Sg+Masc+@PP_ADVL  PP] [PP sa
        i+Prep+Art+Sg+@PP_ADVL [NP scrúdú
        scrúdú+Noun+Masc+Com+Sg+DefArt+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
73. [S  [V D' do+Part+Vb+@>V éirigh
        éirigh+Verb+VI+Vow+PastInd+Len+@FMV  ] [AD go
        go+Part+Ad+@>ADJ maith maith+Adj+Base+@ADVL  ] [PP leis
        le+Prep+Simp+@PP_SUBJ [NP an an+Art+Sg+Def+@>N mac
        mac+Noun+Masc+Com+Sg+DefArt+@P< léinn
        léann+Noun+Masc+Gen+Sg+@N<  NP]  PP] [PP sa
        i+Prep+Art+Sg+@PP_ADVL [NP scrúdú
        scrúdú+Noun+Masc+Com+Sg+DefArt+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
74. [S  [V D' do+Part+Vb+@>V éirigh
        éirigh+Verb+VI+Vow+PastInd+Len+@FMV  ] [PP leis
        le+Prep+Simp+@PP_SUBJ [NP an an+Art+Sg+Def+@>N mac
        mac+Noun+Masc+Com+Sg+DefArt+@P< léinn
        léann+Noun+Masc+Gen+Sg+@N<  NP]  PP] [PP sa
        i+Prep+Art+Sg+@PP_ADVL [NP scrúdú
        scrúdú+Noun+Masc+Com+Sg+DefArt+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
75. [S  [V D' do+Part+Vb+@>V éirigh
        éirigh+Verb+VI+Vow+PastInd+Len+@FMV  ] [PP sa
        i+Prep+Art+Sg+@PP_ADVL [NP scrúdú
        scrúdú+Noun+Masc+Com+Sg+DefArt+@P<  NP]  PP] [PP leis
        le+Prep+Simp+@PP_ADVL [NP an an+Art+Sg+Def+@>N mac
        mac+Noun+Masc+Com+Sg+DefArt+@P< léinn
        léann+Noun+Masc+Gen+Sg+@N<  NP]  PP] . .+Punct+Fin+<<< S]
76. [S  [VS Deisíodh deisigh+Verb+VT+PastInd+Auto+@FMV_SUBJ  ] [NP
        an an+Art+Sg+Def+@>N rothar
        rothar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] [PP ag
        ag+Prep+Simp+@PP_ADVL [NP Seán
        Seán+Prop+Noun+Masc+Com+Sg+@P<  NP]  PP] . .+Punct+Fin+<<<
        S]
77. [S  [VS Deisíodh deisigh+Verb+VT+PastInd+Auto+@FMV_SUBJ  ] [NP
        an an+Art+Sg+Def+@>N rothar
        rothar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] . .+Punct+Fin+<<<
        S]
78. [S  [V D' do+Part+Vb+@>V fhág fág+Verb+VTI+PastInd+Len+@FMV  ]
        [NP an an+Art+Sg+Def+@>N bád
        bád+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [V a
        a+Part+Vb+Rel+Direct+@>V chonaic
```

feic+Verb+VTI+PastInd+Len+@FMV_REL] [NP mac mac+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ an an+Art+Sg+Def+@>N fhir fear+Noun+Masc+Gen+Sg+Len+@N< NP] . .+Punct+Fin+<<< S]

79. [S [V D' do+Part+Vb+@>V fhág fág+Verb+VTI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [V a a+Part+Vb+Rel+Indirect+@>V chonaic feic+Verb+VTI+PastInd+Len+@FMV_REL] [NP a a+Det+Poss+3P+Sg+Masc+@>N mhac mac+Noun+Masc+Com+Sg+Len+@SUBJ NP] [NP an an+Art+Sg+Def+@>N bád bád+Noun+Masc+Com+Sg+DefArt+@OBJ NP] . .+Punct+Fin+<<< S]
80. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [V a a+Part+Vb+Rel+Direct+@>V bhuaigh buaigh+Verb+VTI+PastInd+Len+@FMV_REL] [NP an an+Art+Sg+Def+@>N crannchur crannchur+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ NP] . .+Punct+Fin+<<< S]
81. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [V a a+Part+Vb+Rel+Direct+@>V d' do+Part+Vb+@>V ionsaigh ionsaigh+Verb+VTI+Vow+PastInd+Len+@FMV_REL] [NP iad iad+Pron+Pers+3P+Pl+@OBJ NP] . .+Punct+Fin+<<< S]
82. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP an an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [V a a+Part+Vb+Rel+Direct+@>V d' do+Part+Vb+@>V ionsaigh ionsaigh+Verb+VTI+Vow+PastInd+Len+@FMV_REL] [NP siad siad+Pron+Pers+3P+Pl+Sbj+@SUBJ NP] . .+Punct+Fin+<<< S]
83. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD ansin ansin+Adv+Loc+@ADVL] [AD go go+Part+Ad+@>ADJ ciúin ciúin+Adj+Base+@ADVL] [PP ins i+Prep+Art+Sg+@PP_ADVL [NP an an+Art+Sg+Def+@>N seomra seomra+Noun+Masc+Com+Sg+DefArt+@P< NP] PP] [PP ar_feadh ar_feadh+Prep+Cmpd+@PP_ADVL [NP leath leath+Det+Qty+@>N uair uair+Noun+Fem+Com+Sg+@P< a an+Art+Sg+Def+@>N chloig clog+Noun+Masc+Gen+Sg+DefArt+@N< NP] PP] [CB nuair nuair+Conj+Subord+@CLB] [V a a+Part+Vb+Rel+Direct+@>V bhí bí+Verb+VI+PastInd+Len+@FMV_REL] [NP tuirse tuirse+Noun+Fem+Com+Sg+@SUBJ NP] [PP air ar+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] . .+Punct+Fin+<<< S]
84. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD ansin ansin+Adv+Loc+@ADVL] [AD inné inné+Adv+Temp+@ADVL] . .+Punct+Fin+<<< S]
85. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP sé sé+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]
86. [S [V D' do+Part+Vb+@>V fhan fan+Verb+VI+PastInd+Len+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD ansin ansin+Adv+Loc+@ADVL] [CB nuair nuair+Conj+Subord+@CLB] [V a a+Part+Vb+Rel+Direct+@>V bhí bí+Verb+VI+PastInd+Len+@FMV_REL] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [PRED dorcha dorcha+Adj+Base+@PRED] . .+Punct+Fin+<<< S]
87. [S [V Dheisigh deisigh+Verb+VT+PastInd+Len+@FMV] [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@SUBJ NP] [NP an

an+Art+Sg+Def+@>N rothar
rothar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] . .+Punct+Fin+<<<
S]
88. [S [V D' do+Part+Vb+@>V ith ith+Verb+VTI+Vow+PastInd+Len+@FMV
] . .+Punct+Fin+<<< S]
89. [S [V D' do+Part+Vb+@>V ith ith+Verb+VTI+Vow+PastInd+Len+@FMV
] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N dinnéar
dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] . .+Punct+Fin+<<<
S]
90. [S [VS D' do+Part+Vb+@>V itheamar
ith+Verb+VTI+Vow+PastInd+1P+Pl+Len+@FMV_SUBJ] [NP an
an+Art+Sg+Def+@>N dinnéar
dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] . .+Punct+Fin+<<<
S]
91. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [CB go
go+Part+Vb+Cmpl+@CLB] [V dtabharfaidh
tabhair+Verb+VD+FutInd+Ecl+@FMV] [NP an an+Art+Sg+Def+@>N
bhean bean+Noun+Fem+Com+Sg+DefArt+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]
92. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [CB go
go+Part+Vb+Cmpl+@CLB] [V rachadh
téigh+Verb+VTI+Cond+Ecl+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] . .+Punct+Fin+<<< S]
93. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [CB gur
is+Cop+Pres+Dep+@CLB] [PRED múinteoir
múinteoir+Noun+Masc+Com+Sg+@PRED] [NP é
é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]
94. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [CB gur
gur+Part+Vb+Cmpl+Past+@CLB] [V thug
tabhair+Verb+VD+PastInd+Len+@FMV] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]
95. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [COP nach
is+Cop+Pres+Rel+Neg+@COP] [PRED múinteoir
múinteoir+Noun+Masc+Com+Sg+@PRED] [NP é
é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]
96. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [COP nár
is+Cop+Past+Rel+Neg+@COP] [PRED múinteoir
múinteoir+Noun+Masc+Com+Sg+@PRED] [NP é
é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]
97. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [CB nár
nár+Part+Vb+Neg+Cmpl+Past+@CLB] [V thug
tabhair+Verb+VD+PastInd+Neg+Len+@FMV] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire

Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

98. [S [V Dúirt abair+Verb+VTI+Vow+PastInd+@FMV] [NP siad siad+Pron+Pers+3P+Pl+Sbj+@SUBJ NP] [CB nach nach+Part+Vb+Neg+Cmpl+@CLB] [V bhfeiceann feic+Verb+VTI+PresInd+Ecl+@FMV] [NP siad siad+Pron+Pers+3P+Pl+Sbj+@SUBJ NP] [NP an an+Art+Sg+Def+@>N cineál cineál+Noun+Masc+Com+Sg+DefArt+@OBJ seo seo+Det+Dem+@N< NP] [AD chomh chomh+Adv+Its+@>ADJ minic minic+Adj+Base+@ADVL] [NP sin sin+Pron+Dem+@NP NP] .
.+Punct+Fin+<<< S]

99. [S [PRED Fear fear+Noun+Masc+Com+Sg+@PRED maith maith+Adj+Masc+Com+Sg+@N<] [COP is is+Cop+Pres+Rel+@COP] [NP ea ea+Pron+Pers+3P+Sg+@AUG>SUBJ é é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]

100. [S [PP go go+Prep+Simp+@PP_ADVL [NP deo deo+Subst+Noun+Sg+@P< NP] PP] , ,+Punct+Int [PP go go+Prep+Simp+@PP_ADVL [NP bás bás+Noun+Masc+Com+Sg+@P< NP] PP] , ,+Punct+Int [PP go go+Prep+Simp+@PP_ADVL [NP brách brách+Subst+Noun+Sg+@P< NP] PP] . .+Punct+Fin+<<< S]

101. [S [V Íocfaidh íoc+Verb+VTI+Vow+FutInd+@FMV] [NP mé mé+Pron+Pers+1P+Sg+@SUBJ NP] [PP as as+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] [V a a+Part+Vb+Rel+Indirect+Pro+@>V gceannóidh ceannaigh+Verb+VTI+FutInd+Ecl+@FMV_REL] [NP tú tú+Pron+Pers+2P+Sg+@SUBJ NP] . .+Punct+Fin+<<< S]

102. [S [COP Is is+Cop+Pres+@COP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP cabhrú cabhrú+Verbal+Noun+VI+@P< NP] PP-ASP] ASP] [PP liom le+Pron+Prep+1P+Sg+@PP_ADVL PP] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] . .+Punct+Fin+<<< S]

103. [S [COP Is is+Cop+Pres+@COP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP déanamh déanamh+Verbal+Noun+VTI+@P< NP] PP-ASP] [OA cáca cáca+Noun+Masc+Gen+Sg+@OBJ_ASP OA] ASP] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [NP mé mé+Pron+Pers+1P+Sg+@SUBJ NP] . .+Punct+Fin+<<< S]

104. [S [COP Is is+Cop+Pres+@COP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP iascaireacht iascaireacht+Verbal+Noun+NStem+@P< NP] PP-ASP] ASP] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] . .+Punct+Fin+<<< S]

105. [S [COP Is is+Cop+Pres+@COP] [PRED airde ard+Adj+Comp+@PRED] [NP sliabh sliabh+Noun+Masc+Com+Sg+@SUBJ NP] [CJ2 ná ná+Conj+Coord+@CC [NP cnoc cnoc+Noun+Masc+Com+Sg+@NP NP] CJ2] . .+Punct+Fin+<<< S]

106. [S [COP Is is+Cop+Pres+@COP] [AD amhlaidh amhlaidh+Adv+Gn+@ADVL] [VS a a+Part+Vb+Rel+Direct+@>V bhídís bí+Verb+VI+PastImp+3P+Pl+Len+@FAUX_REL_SUBJ] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP obair obair+Verbal+Noun+NStem+@P< NP] PP-ASP] ASP] [PP do do+Prep+Simp+@PP_ADVL [NP na na+Art+Pl+Def+@>N feirmeoirí feirmeoir+Noun+Masc+Com+Pl+DefArt+@P< NP] PP] .
.+Punct+Fin+<<< S]

107. [S [COP Is is+Cop+Pres+@COP] [PRED an an+Art+Sg+Def+@>N leabhar leabhar+Noun+Masc+Com+Sg+DefArt+@PRED] [V a a+Part+Vb+Rel+Direct+@>V thug tabhair+Verb+VD+PastInd+Len+@FMV_REL] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [PP do do+Prep+Simp+@PP_OBL [NP Mháire Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

108. [S [COP Is is+Cop+Pres+@COP] [PRED beag beag+Adj+Base+@PRED] [NP planda planda+Noun+Masc+Com+Sg+@SUBJ NP] [V a a+Part+Vb+Rel+Direct+@>V fhásann fás+Verb+VTI+PresInd+Len+@FMV_REL] [PP i i+Prep+Simp+@PP_ADVL [NP dteocht teocht+Noun+Fem+Com+Sg+Ecl+@P< NP] PP] [PP faoi_bhun faoi_bhun+Prep+Cmpd+@PP_ADVL [NP 4C 4C+Guess+Abr+@P< NP] PP] . .+Punct+Fin+<<< S]
109. [S [COP Is is+Cop+Pres+@COP] [OA cáca cáca+Noun+Masc+Com+Sg+@OBJ_ASP OA] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [ASP [PP-ASP á do+Prep+Poss+3P+Pl+Obj+@PP_ASP [NP dhéanamh déanamh+Verbal+Noun+VTI+Len+@P< NP] PP-ASP] ASP] [PP agam ag+Pron+Prep+1P+Sg+@PP_ADVL PP] . .+Punct+Fin+<<< S]
110. [S [COP Is is+Cop+Pres+@COP] [PRED cailín cailín+Noun+Masc+Com+Sg+@PRED] [NP í í+Pron+Pers+3P+Sg+Fem+@SUBJ NP] . .+Punct+Fin+<<< S]
111. [S [COP Is is+Cop+Pres+@COP] [PRED deas deas+Adj+Base+@PRED an an+Art+Sg+Def+@>N lá lá+Noun+Masc+Com+Sg+DefArt+@PRED<] . .+Punct+Fin+<<< S]
112. [S [COP Is is+Cop+Pres+@COP] [PRED deas deas+Adj+Base+@PRED] [NP an an+Art+Sg+Def+@>N lá lá+Noun+Masc+Com+Sg+DefArt+@SUBJ NP] [NP é é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]
113. [S [COP Is is+Cop+Pres+@COP] [PP do do+Prep+Simp+@PP_OBL [NP Mháire Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] [V a a+Part+Vb+Rel+Direct+@>V thug tabhair+Verb+VD+PastInd+Len+@FMV_REL] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an an+Art+Sg+Def+@>N leabhar leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] . .+Punct+Fin+<<< S]
114. [S [COP Is is+Cop+Pres+@COP] [PP do do+Prep+Simp+@PP_OBL [NP Mháire Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] [V a a+Part+Vb+Rel+Direct+@>V thug tabhair+Verb+VD+PastInd+Len+@FMV_REL] [NP sí sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP leabhar leabhar+Noun+Masc+Com+Sg+@OBJ NP] . .+Punct+Fin+<<< S]
115. [S [COP Is is+Cop+Pres+@COP] [NP eisean é+Pron+Pers+3P+Sg+Masc+Emph+@SUBJ NP] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP cabhrú cabhrú+Verbal+Noun+VI+@P< NP] PP-ASP] ASP] [PP liom le+Pron+Prep+1P+Sg+@PP_ADVL PP] . .+Punct+Fin+<<< S]
116. [S [COP Is is+Cop+Pres+@COP] [NP eisean é+Pron+Pers+3P+Sg+Masc+Emph+@SUBJ NP] [V atá bí+Verb+VI+PresInd+Rel+@FAUX_REL] [ASP [PP-ASP do do+Prep+Simp+@PP_ASP [OA mo mo+Det+Poss+1P+Sg+@OBJ_ASP OA] [NP chabhrú cabhrú+Verbal+Noun+VI+Len+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
117. [S [COP Is is+Cop+Pres+@COP] [PRED fear fear+Noun+Masc+Com+Sg+@PRED maith maith+Adj+Masc+Com+Sg+@N<] [NP é é+Pron+Pers+3P+Sg+Masc+@SUBJ NP] . .+Punct+Fin+<<< S]
118. [S [COP Is is+Cop+Pres+@COP] [PRED fearr maith+Adj+Comp+@PRED] [PP liom le+Pron+Prep+1P+Sg+@PP_SUBJ PP] [NP úlla úll+Noun+Masc+Com+Pl+@OBJ NP] [CJ2 ná ná+Conj+Coord+@CC [NP oráistí oráiste+Noun+Masc+Com+Pl+@OBJ NP] CJ2] . .+Punct+Fin+<<< S]
119. [S [COP Is is+Cop+Pres+Rel+@COP] [NP í í+Pron+Pers+3P+Sg+Fem+@AUG>SUBJ an an+Art+Sg+Def+@>N líne líne+Noun+Fem+Com+Sg+DefArt+@SUBJ glas glas+Adj+Masc+Com+Sg+@N< NP] [PRED teorainn

```
        teorainn+Noun+Fem+Com+Sg+@PRED an an+Art+Sg+Def+@>N
        cheantair ceantar+Noun+Masc+Gen+Sg+DefArt+@N< ] .
        .+Punct+Fin+<<< S]
120. [S [COP Is is+Cop+Pres+@COP ] [PRED ise
        í+Pron+Pers+3P+Sg+Fem+Emph+@PRED ] [V a
        a+Part+Vb+Rel+Direct+@>V thug
        tabhair+Verb+VD+PastInd+Len+@FMV_REL  ] [NP an
        an+Art+Sg+Def+@>N leabhar
        leabhar+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ  NP] [PP do
        do+Prep+Simp+@PP_OBL [NP Mháire
        Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
121. [S [COP Is is+Cop+Pres+@COP ] [PRED ise
        í+Pron+Pers+3P+Sg+Fem+Emph+@PRED ] [V a
        a+Part+Vb+Rel+Direct+@>V thug
        tabhair+Verb+VD+PastInd+Len+@FMV_REL  ] [NP leabhar
        leabhar+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ  NP] [PP do
        do+Prep+Simp+@PP_OBL [NP Mháire
        Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
122. [S [COP Is is+Cop+Pres+Rel+@COP ] [PRED lá
        lá+Noun+Masc+Com+Sg+@PRED deas deas+Adj+Masc+Com+Sg+@N< ]
        [NP é é+Pron+Pers+3P+Sg+Masc+@SUBJ  NP] . .+Punct+Fin+<<<
        S]
123. [S [COP Is is+Cop+Pres+@COP ] [PP le le+Prep+Simp+@PP_PRED [NP
        Dónal Dónal+Guess+Prop+Noun+Masc+Com+Sg+@P<  NP]  PP] [NP
        an an+Art+Sg+Def+@>N teach
        teach+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] . .+Punct+Fin+<<<
        S]
124. [S [COP Is is+Cop+Pres+@COP ] [PRED leabhar
        leabhar+Noun+Masc+Com+Sg+@PRED ] [V a
        a+Part+Vb+Rel+Direct+@>V thug
        tabhair+Verb+VD+PastInd+Len+@FMV_REL  ] [NP sí
        sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [PP do
        do+Prep+Simp+@PP_OBL [NP Mháire
        Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] .
        .+Punct+Fin+<<< S]
125. [S [COP Is is+Cop+Pres+@COP ] [PRED maith maith+Adj+Base+@PRED
        ] [PP liom le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [NP úlla
        úll+Noun+Masc+Com+Pl+@OBJ  NP] [CJ2 agus
        agus+Conj+Coord+@CC [NP oráistí
        oráiste+Noun+Masc+Com+Pl+@OBJ  NP] CJ2] . .+Punct+Fin+<<<
        S]
126. [S [COP Is is+Cop+Pres+@COP ] [NP mise
        mé+Pron+Pers+1P+Sg+Emph+@SUBJ  NP] [V atá
        bí+Verb+VI+PresInd+Rel+@FAUX_REL  ] [ASP [PP-ASP ag
        ag+Prep+Simp+@PP_ASP [NP déanamh
        déanamh+Verbal+Noun+VTI+@P<  NP]  PP-ASP] [OA cáca
        cáca+Noun+Masc+Gen+Sg+@OBJ_ASP  OA]  ASP] . .+Punct+Fin+<<<
        S]
127. [S [COP Is is+Cop+Pres+@COP ] [NP mise
        mé+Pron+Pers+1P+Sg+Emph+@SUBJ  NP] [PRED Briain
        Briain+Prop+Noun+Masc+Com+Sg+@PRED ] . .+Punct+Fin+<<< S]
128. [S [COP Is is+Cop+Pres+@COP ] [PRED múinteoir
        múinteoir+Noun+Masc+Com+Sg+@PRED ] [NP é
        é+Pron+Pers+3P+Sg+Masc+@SUBJ  NP] . .+Punct+Fin+<<< S]
129. [S [COP Is is+Cop+Pres+@COP ] [PRED múinteoir
        múinteoir+Noun+Masc+Com+Sg+@PRED ] [NP Seán
        Seán+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] . .+Punct+Fin+<<< S]
130. [S [VS Ith ith+Verb+VTI+Vow+Imper+2P+Sg+@FMV_SUBJ  ] [NP an
        an+Art+Sg+Def+@>N dinnéar
        dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] . .+Punct+Fin+<<<
        S]
```

```
131. [S  [VS Labhair labhair+Verb+VTI+Imper+2P+Sg+@FMV_SUBJ  ] .
         .+Punct+Fin+<<< S]
132. [S  [VS Labhair labhair+Verb+VTI+Imper+2P+Sg+@FMV_SUBJ  ] [AD
         go go+Part+Ad+@>ADJ soiléir soiléir+Adj+Base+@ADVL  ] .
         .+Punct+Fin+<<< S]
133. [S  [V Labhair labhair+Verb+VTI+PastInd+Len+@FMV  ] [NP sé
         sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [PP os
         os+Prep+Simp+@PP_ADVL [NP ard ard+Noun+Masc+Com+Sg+@P<  NP]
         PP] . .+Punct+Fin+<<< S]
134. [S  [V Labhair labhair+Verb+VTI+PastInd+Len+@FMV  ] [NP Seán
         Seán+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] . .+Punct+Fin+<<< S]
135. [S  [VS Labhraíomar
         labhair+Verb+VTI+PastInd+1P+Pl+Len+@FMV_SUBJ  ] .
         .+Punct+Fin+<<< S]
136. [S  [VS Labhraítear labhair+Verb+VTI+Imper+Auto+@FMV_SUBJ  ]
         [AD go go+Part+Ad+@>ADJ soiléir soiléir+Adj+Base+@ADVL  ] .
         .+Punct+Fin+<<< S]
137. [S  [PP le le+Prep+Simp+@PP_ADVL [NP tamall
         tamall+Noun+Masc+Com+Sg+@P<  NP]  PP] , ,+Punct+Int [PP le
         le+Prep+Simp+@PP_ADVL [NP fada fad+Noun+Masc+Com+Sg+@P<
         NP]  PP] , ,+Punct+Int [PP le le+Prep+Simp+@PP_ADVL [NP
         seachtain seachtain+Noun+Fem+Com+Sg+@P<  NP]  PP] .
         .+Punct+Fin+<<< S]
138. [S  [VS Líonadh líon+Verb+VTI+PastInd+Auto+@FMV_SUBJ  ] [NP an
         an+Art+Sg+Def+@>N poll poll+Noun+Masc+Com+Sg+DefArt+@OBJ
         NP] [PP le le+Prep+Simp+@PP_ADVL [NP clocha
         cloch+Noun+Fem+Com+Pl+@P<  NP]  PP] . .+Punct+Fin+<<< S]
139. [S  [CB Má má+Conj+Subord+@CLB ] [V bhíonn
         bí+Verb+VI+PresImp+Len+@FMV  ] [NP an an+Art+Sg+Def+@>N t-
         am am+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [PP agat
         ag+Pron+Prep+2P+Sg+@PP_HAS  PP] , ,+Punct+Int [VS déan
         déan+Verb+VT+Imper+2P+Sg+@FMV_SUBJ  ] [NP é
         é+Pron+Pers+3P+Sg+Masc+@OBJ  NP] . .+Punct+Fin+<<< S]
140. [S  [NP Máire Máire+Prop+Noun+Fem+Com+Sg+@NP  NP] .
         .+Punct+Fin+<<< S]
141. [S  [COP Nach is+Cop+Pres+NegQ+@COP ] [NP é
         é+Pron+Pers+3P+Sg+Masc+@AUG>SUBJ é
         é+Pron+Pers+3P+Sg+Masc+@SUBJ  NP] [V a
         a+Part+Vb+Rel+Direct+@>V bhí
         bí+Verb+VI+PastInd+Len+@FMV_REL  ] [PP ann
         i+Pron+Prep+3P+Sg+Masc+@PP_ADVL  PP] ? ?+Punct+Fin+Q+<<< S]
142. [S  [COP Nach is+Cop+Pres+NegQ+@COP ] [NP tusa
         tú+Pron+Pers+2P+Sg+Emph+@SUBJ  NP] [PRED an
         an+Art+Sg+Def+@>N múinteoir
         múinteoir+Noun+Masc+Com+Sg+DefArt+@PRED ] ?
         ?+Punct+Fin+Q+<<< S]
143. [S  [V Nár nár+Part+Vb+NegQ+@>V ith
         ith+Verb+VTI+Vow+PastInd+NegQ+Len+@FMV  ] [NP sí
         sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [NP an
         an+Art+Sg+Def+@>N dinnéar
         dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] ?
         ?+Punct+Fin+Q+<<< S]
144. [S  [V Nár nár+Part+Vb+NegQ+@>V labhair
         labhair+Verb+VTI+PastInd+NegQ+Len+@FMV  ] [NP Seán
         Seán+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] . .+Punct+Fin+<<< S]
145. [S  [V Nár nár+Part+Vb+Neg+Rel+Past+@>V thug
         tabhair+Verb+VD+PastInd+Neg+Len+@FMV_REL  ] [NP sí
         sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [NP an
         an+Art+Sg+Def+@>N leabhar
         leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] [PP do
         do+Prep+Simp+@PP_OBL [NP Mháire
         Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] ?
         ?+Punct+Fin+Q+<<< S]
```

```
146. [S  [COP Ní is+Cop+Pres+Neg+@COP ] [PRED gorm
         gorm+Adj+Base+@PRED ] [V atá
         bí+Verb+VI+PresInd+Rel+@FMV_REL  ] [NP sé
         sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] . .+Punct+Fin+<<< S]
147. [S  [COP Ní is+Cop+Pres+Neg+@COP ] [NP hé
         é+Pron+Pers+3P+Sg+Masc+hPref+@AUG>SUBJ Briain
         Briain+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] [PRED an
         an+Art+Sg+Def+@>N múinteoir
         múinteoir+Noun+Masc+Com+Sg+DefArt+@PRED ] . .+Punct+Fin+<<<
         S]
148. [S  [COP Ní is+Cop+Pres+Neg+@COP ] [NP hé
         é+Pron+Pers+3P+Sg+Masc+hPref+@SUBJ  NP] [COP nár
         is+Cop+Past+Rel+Neg+@COP ] [PRED mhaith
         maith+Adj+Base+Len+@PRED ] [PP liom
         le+Pron+Prep+1P+Sg+@PP_SUBJ  PP] [NP é
         é+Pron+Pers+3P+Sg+Masc+@OBJ  NP] . .+Punct+Fin+<<< S]
149. [S  [COP Ní is+Cop+Pres+Neg+@COP ] [PP liomsa
         le+Pron+Prep+1P+Sg+Emph+@PP_PRED  PP] [NP an
         an+Art+Sg+Def+@>N t-airgead
         airgead+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] .
         .+Punct+Fin+<<< S]
150. [S  [COP Ní is+Cop+Pres+Neg+@COP ] [PRED mór mór+Adj+Base+@PRED
         ] [PP dúinn do+Pron+Prep+1P+Pl+@PP_SUBJ  PP] [NP aonad
         aonad+Noun+Masc+Com+Sg+@SUBJ_INF  NP] [INF [I a
         a+Part+Inf+@>N bheith bheith+Verbal+Noun+VI+Len+@INF  I]
         INF] [PP againn ag+Pron+Prep+1P+Pl+@PP_ADVL  PP] .
         .+Punct+Fin+<<< S]
151. [S  [V Níl bí+Verb+VI+PresInd+Neg+@FMV  ] [NP an
         an+Art+Sg+Def+@>N cinneadh
         cinneadh+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [PRED déanta
         déanta+Verbal+Adj+@PRED ] [AD fós fós+Adv+Gn+@ADVL  ] .
         .+Punct+Fin+<<< S]
152. [S  [V Níor níor+Part+Vb+Neg+Past+@>V ith
         ith+Verb+VTI+Vow+PastInd+Neg+Len+@FMV  ] . .+Punct+Fin+<<<
         S]
153. [S  [V Níor níor+Part+Vb+Neg+Past+@>V ith
         ith+Verb+VTI+Vow+PastInd+Neg+Len+@FMV  ] [NP sí
         sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [NP an
         an+Art+Sg+Def+@>N dinnéar
         dinnéar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] . .+Punct+Fin+<<<
         S]
154. [S  [V Níor níor+Part+Vb+Neg+Past+@>V labhair
         labhair+Verb+VTI+PastInd+Neg+Len+@FMV  ] . .+Punct+Fin+<<<
         S]
155. [S  [V Níor níor+Part+Vb+Neg+Past+@>V labhair
         labhair+Verb+VTI+PastInd+Neg+Len+@FMV  ] [NP Seán
         Seán+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] . .+Punct+Fin+<<< S]
156. [S  [V Níor níor+Part+Vb+Neg+Past+@>V tháinig
         tar+Verb+VI+PastInd+Neg+Len+@FMV  ] [NP sé
         sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD go
         go+Part+Ad+@>ADJ fóill fóill+Adj+Base+@ADVL  ] .
         .+Punct+Fin+<<< S]
157. [S  [V Níor níor+Part+Vb+Neg+Past+@>V thug
         tabhair+Verb+VD+PastInd+Neg+Len+@FMV  ] . .+Punct+Fin+<<<
         S]
158. [S  [V Níor níor+Part+Vb+Neg+Past+@>V thug
         tabhair+Verb+VD+PastInd+Neg+Len+@FMV  ] [NP sí
         sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ  NP] [NP an
         an+Art+Sg+Def+@>N leabhar
         leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] [PP do
         do+Prep+Simp+@PP_OBL [NP Mháire
         Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] .
         .+Punct+Fin+<<< S]
```

```
159. [S  [COP Níorbh is+Cop+Past+Neg+VF+@COP ] [PRED é
          é+Pron+Pers+3P+Sg+Masc+@PRED ] . .+Punct+Fin+<<< S]
160. [S  [V Rinne déan+Verb+VT+PastInd+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [NP é
          é+Pron+Pers+3P+Sg+Masc+@OBJ  NP] [AD go go+Part+Ad+@>ADJ
          maith maith+Adj+Base+@ADVL  ] . .+Punct+Fin+<<< S]
161. [S  [V Rith rith+Verb+VTI+PastInd+Len+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [PP le
          le+Prep+Simp+@PP_ADVL [NP luas luas+Noun+Masc+Com+Sg+@P<
          lasrach lasair+Noun+Fem+Gen+Sg+@N<  NP]  PP] .
          .+Punct+Fin+<<< S]
162. [S  [INF [OI Rud rud+Noun+Masc+Com+Sg+@OBJ_INF ab
          is+Part+Sup+@>ADJ fhusa furasta+Adj+Comp+Len+@N<  OI] [I a
          a+Part+Inf+@>N dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF
          I]  INF] ... ...+Punct+Fin+<<< S]
163. [S  [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@NP  NP] .
          .+Punct+Fin+<<< S]
164. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N bád bád+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V chonaic
          feic+Verb+VTI+PastInd+Len+@FMV_REL  ] [NP an
          an+Art+Sg+Def+@>N fear
          fear+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ  NP] .
          .+Punct+Fin+<<< S]
165. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V bhuaigh
          buaigh+Verb+VTI+PastInd+Len+@FMV_REL  ] [NP an
          an+Art+Sg+Def+@>N crannchur
          crannchur+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ  NP] .
          .+Punct+Fin+<<< S]
166. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V chonaic
          feic+Verb+VTI+PastInd+Len+@FMV_REL  ] [NP an
          an+Art+Sg+Def+@>N bád
          bád+Noun+Masc+Com+Sg+DefArt+@SUBJ_OR_OBJ  NP] .
          .+Punct+Fin+<<< S]
167. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V chonaic
          feic+Verb+VTI+PastInd+Len+@FMV_REL  ] [NP an
          an+Art+Sg+Def+@>N bhean
          bean+Noun+Fem+Com+Sg+DefArt+@SUBJ_OR_OBJ  NP] .
          .+Punct+Fin+<<< S]
168. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V d' do+Part+Vb+@>V ionsaigh
          ionsaigh+Verb+VTI+Vow+PastInd+Len+@FMV_REL  ] [NP iad
          iad+Pron+Pers+3P+Pl+@OBJ  NP] . .+Punct+Fin+<<< S]
169. [S  [COP Seo seo+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Direct+@>V d' do+Part+Vb+@>V ionsaigh
          ionsaigh+Verb+VTI+Vow+PastInd+Len+@FMV_REL  ] [NP siad
          siad+Pron+Pers+3P+Pl+Sbj+@SUBJ  NP] . .+Punct+Fin+<<< S]
170. [S  [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ ] [V a
          a+Part+Vb+Rel+Indirect+@>V bhfuil
          bí+Verb+VI+PresInd+Dep+Ecl+@FMV_REL  ] [PP ann
          i+Pron+Prep+3P+Sg+Masc+@PP_ADVL  PP] . .+Punct+Fin+<<< S]
171. [S  [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ ] [PRED an
          an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED ]
          [V a a+Part+Vb+Rel+Indirect+@>V bhfuil
          bí+Verb+VI+PresInd+Dep+Ecl+@FAUX_REL  ] [NP a
```

a+Det+Poss+3P+Sg+Masc+@>N mhac
mac+Noun+Masc+Com+Sg+Len+@SUBJ_ASP NP] [ASP [PP-ASP ag
ag+Prep+Simp+@PP_ASP [NP imeacht imeacht+Verbal+Noun+VI+@P<
NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]

172. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED]
[V a a+Part+Vb+Rel+Indirect+@>V bhfuil
bí+Verb+VI+PresInd+Dep+Ecl+@FMV_REL] [NP a
a+Det+Poss+3P+Sg+Masc+@>N mhac
mac+Noun+Masc+Com+Sg+Len+@SUBJ NP] [PRED tinn
tinn+Adj+Masc+Com+Sg+@PRED] . .+Punct+Fin+<<< S]

173. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED]
[V a a+Part+Vb+Rel+Direct+@>V chuireann
cuir+Verb+VTI+PresInd+Len+@FMV_REL] [NP síol
síol+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ NP] . .+Punct+Fin+<<<
S]

174. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N fear fear+Noun+Masc+Com+Sg+DefArt+@PRED]
[V a a+Part+Vb+Rel+Direct+@>V phléasc
pléasc+Verb+VTI+PastInd+Len+@FMV_REL] . .+Punct+Fin+<<<
S]

175. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N gort gort+Noun+Masc+Com+Sg+DefArt+@PRED]
[VS a a+Part+Vb+Rel+Indirect+@>V cuireadh
cuir+Verb+VTI+PastInd+Auto+@FMV_REL_SUBJ] [NP an
an+Art+Sg+Def+@>N síol síol+Noun+Masc+Com+Sg+DefArt+@OBJ
NP] [PP ann i+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] .
.+Punct+Fin+<<< S]

176. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N síol síol+Noun+Masc+Com+Sg+DefArt+@PRED]
[V a a+Part+Vb+Rel+Direct+@>V chuireann
cuir+Verb+VTI+PresInd+Len+@FMV_REL] [NP fear
fear+Noun+Masc+Com+Sg+@SUBJ_OR_OBJ NP] . .+Punct+Fin+<<<
S]

177. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N síol síol+Noun+Masc+Com+Sg+DefArt+@PRED]
[VS a a+Part+Vb+Rel+Direct+@>V cuireadh
cuir+Verb+VTI+PastInd+Auto+@FMV_REL_SUBJ] .
.+Punct+Fin+<<< S]

178. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N té té+Noun+Masc+Com+Sg+DefArt+@PRED] [V
a a+Part+Vb+Rel+Direct+@>V itheann
ith+Verb+VTI+Vow+PresInd+Len+@FMV_REL] [NP feoil
feoil+Noun+Fem+Com+Sg+@SUBJ_OR_OBJ NP] . .+Punct+Fin+<<<
S]

179. [S [COP Sin sin+Cop+Pro+Dem+@COP_SUBJ] [PRED an
an+Art+Sg+Def+@>N teach teach+Noun+Masc+Com+Sg+DefArt+@PRED
] [V a a+Part+Vb+Rel+Indirect+@>V raibh
bí+Verb+VI+PastInd+Dep+Ecl+@FAUX_REL] [NP sé
sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ina
i+Prep+Poss+3P+Sg+Masc+@PP_STAT [NP chónaí
cónaí+Verbal+Noun+VI+Len+@P< NP] PP-ASP] ASP] [AD ann
ann+Adv+Loc+@ADVL] . .+Punct+Fin+<<< S]

180. [S [NP Sise sí+Pron+Pers+3P+Sg+Fem+Sbj+Emph+@SUBJ NP] .
.+Punct+Fin+<<< S]

181. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [OA an an+Art+Sg+Def+@>N
cáca cáca+Noun+Masc+Com+Sg+DefArt+@OBJ_ASP OA] [ASP [PP-
ASP arna arna+Prep+Cmpd+@PP_ASP [NP dhéanamh
déanamh+Verbal+Noun+VTI+Len+@P< NP] PP-ASP] ASP] [PP
agam ag+Pron+Prep+1P+Sg+@PP_ADVL PP] . .+Punct+Fin+<<< S]

```
182. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP an an+Art+Sg+Def+@>N
         carr carr+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [PP sa
         i+Prep+Art+Sg+@PP_ADVL [NP gharáiste
         garáiste+Noun+Masc+Com+Sg+Len+@P<  NP]  PP] .
         .+Punct+Fin+<<< S]
183. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP an an+Art+Sg+Def+@>N
         doras doras+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [ASP [PP-ASP
         ar ar+Prep+Simp+@PP_STAT [NP oscailt
         oscailt+Verbal+Noun+VTI+@P<  NP]  PP-ASP]  ASP] .
         .+Punct+Fin+<<< S]
184. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP an an+Art+Sg+Def+@>N
         leabhar leabhar+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [PRED go
         go+Part+Ad+@>ADJ maith maith+Adj+Base+@PRED ] .
         .+Punct+Fin+<<< S]
185. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP an an+Art+Sg+Def+@>N
         leabhar leabhar+Noun+Masc+Com+Sg+DefArt+@SUBJ  NP] [PRED
         léite léite+Verbal+Adj+@PRED ] [PP agam
         ag+Pron+Prep+1P+Sg+@PP_HAS  PP] . .+Punct+Fin+<<< S]
186. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP an an+Art+Sg+Def+@>N
         pictiúir pictiúr+Noun+Masc+Com+Pl+@SUBJ  NP] [PRED
         péinteáilte péinteáilte+Verbal+Adj+@PRED ] [PP ag
         ag+Prep+Simp+@PP_HAS [NP Mary Mary+Prop+Noun+Fem+Com+Sg+@P<
         NP]  PP] . .+Punct+Fin+<<< S]
187. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP áthas
         áthas+Noun+Masc+Com+Sg+@SUBJ  NP] [PP orm
         ar+Pron+Prep+1P+Sg+@PP_ADVL  PP] . .+Punct+Fin+<<< S]
188. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP cáca
         cáca+Noun+Masc+Com+Sg+@SUBJ  NP] [ASP [PP-ASP á
         do+Prep+Poss+3P+Sg+Masc+Obj+@PP_ASP [NP dhéanamh
         déanamh+Verbal+Noun+VTI+Len+@P<  NP]  PP-ASP]  ASP] [PP
         agam ag+Pron+Prep+1P+Sg+@PP_ADVL  PP] . .+Punct+Fin+<<< S]
189. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [OA cáca
         cáca+Noun+Masc+Com+Sg+@OBJ_ASP  OA] [ASP [PP-ASP le
         le+Prep+Simp+@PP_ASP [NP déanamh
         déanamh+Verbal+Noun+VTI+@P<  NP]  PP-ASP]  ASP] [PP agam
         ag+Pron+Prep+1P+Sg+@PP_ADVL  PP] . .+Punct+Fin+<<< S]
190. [S  [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP cuimhne
         cuimhne+Noun+Fem+Com+Sg+@SUBJ mhaith
         maith+Adj+Fem+Com+Sg+@N<  NP] [PP agam
         ag+Pron+Prep+1P+Sg+@PP_HAS  PP] [AD chomh
         chomh+Adv+Its+@>ADJ cruaidh crua+Adj+Base+@ADVL  ] [CJ2
         agus agus+Conj+Coord+@CC [V a a+Part+Vb+Rel+Direct+@>V bhí
         bí+Verb+VI+PastInd+Len+@FMV_REL  ] CJ2] [NP sé
         sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] . .+Punct+Fin+<<< S]
191. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP mé
         mé+Pron+Pers+1P+Sg+@SUBJ_ASP  NP] [ASP [PP-ASP ag
         ag+Prep+Simp+@PP_ASP [NP déanamh
         déanamh+Verbal+Noun+VTI+@P<  NP]  PP-ASP] [OA cáca
         cáca+Noun+Masc+Gen+Sg+@OBJ_ASP  OA]  ASP] . .+Punct+Fin+<<<
         S]
192. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP mé
         mé+Pron+Pers+1P+Sg+@SUBJ  NP] [ASP i_ndiaidh
         i_ndiaidh+Prep+Cmpd+@PP_ASP [INF [OI cáca
         cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
         dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF]  ASP] .
         .+Punct+Fin+<<< S]
193. [S  [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP mé
         mé+Pron+Pers+1P+Sg+@SUBJ  NP] [ASP tar_éis
         tar_éis+Prep+Cmpd+@PP_ASP [INF [OI cáca
         cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
         dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF]  ASP] .
         .+Punct+Fin+<<< S]
```

194. [S [V Tá bí+Verb+VI+PresInd+@FMV] [NP ocras ocras+Noun+Masc+Com+Sg+@SUBJ NP] [PP orm ar+Pron+Prep+1P+Sg+@PP_ADVL PP] . .+Punct+Fin+<<< S]
195. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP cabhrú cabhrú+Verbal+Noun+VI+@P< NP] PP-ASP] ASP] [PP liom le+Pron+Prep+1P+Sg+@PP_ADVL PP] . .+Punct+Fin+<<< S]
196. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP caoineadh caoineadh+Verbal+Noun+VTI+@P< NP] PP-ASP] ASP] [PP gan gan+Prep+Simp+@PP_NEG [NP stad stad+Verbal+Noun+VTI+@P< NP] PP] . .+Punct+Fin+<<< S]
197. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP dul dul+Verbal+Noun+VTI+@P< NP] PP-ASP] ASP] [INF [I a a+Part+Inf+@>N chodladh codladh+Verbal+Noun+VTI+Len+@INF I] INF] . .+Punct+Fin+<<< S]
198. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP iascaireacht iascaireacht+Verbal+Noun+NStem+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
199. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP rith rith+Verbal+Noun+VTI+@P< NP] PP-ASP] ASP] [AD go go+Part+Ad+@>ADJ tapaidh tapaidh+Adj+Base+@ADVL] . .+Punct+Fin+<<< S]
200. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP tógáil tógáil+Verbal+Noun+VTI+@P< NP] PP-ASP] ASP] [AD isteach isteach+Adv+Dir+@ADVL] [OA na na+Art+Gen+Sg+Def+Fem+@>N móna móin+Noun+Fem+Gen+Sg+DefArt+@OBJ_ASP OA] . .+Punct+Fin+<<< S]
201. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ag ag+Prep+Simp+@PP_ASP [NP tógáil tógáil+Verbal+Noun+VTI+@P< NP] PP-ASP] [OA na na+Art+Gen+Sg+Def+Fem+@>N móna móin+Noun+Fem+Gen+Sg+DefArt+@OBJ_ASP OA] ASP] [AD isteach isteach+Adv+Dir+@ADVL] . .+Punct+Fin+<<< S]
202. [S [V Tá bí+Verb+VI+PresInd+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [PRED déanta déanta+Verbal+Adj+@PRED] . .+Punct+Fin+<<< S]
203. [S [V Tá bí+Verb+VI+PresInd+@FAUX] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP do do+Prep+Simp+@PP_ASP [OA mo mo+Det+Poss+1P+Sg+@OBJ_ASP OA] [NP chabhrú cabhrú+Verbal+Noun+VI+Len+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
204. [S [V Tá bí+Verb+VI+PresInd+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [AD go go+Part+Ad+@>ADJ hálainn álainn+Adj+Base+hPref+@ADVL] . .+Punct+Fin+<<< S]
205. [S [V Tá bí+Verb+VI+PresInd+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ina i+Prep+Poss+3P+Pl+@PP_STAT [NP chodladh codladh+Noun+Masc+Com+Sg+Len+@P< NP] PP-ASP] ASP] . .+Punct+Fin+<<< S]
206. [S [V Tá bí+Verb+VI+PresInd+@FMV] [NP sé sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ NP] [ASP [PP-ASP ina i+Prep+Poss+3P+Sg+Masc+@PP_STAT [NP mhúinteoir

```
          múinteoir+Noun+Masc+Com+Sg+Len+@P<  NP]  PP-ASP]  ASP] .
          .+Punct+Fin+<<< S]
207. [S   [V Tá bí+Verb+VI+PresInd+@FAUX  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [ASP le
          le+Prep+Simp+@PP_ASP [INF [I teacht
          teacht+Verbal+Noun+VI+@INF  I]  INF]  ASP] .
          .+Punct+Fin+<<< S]
208. [S   [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [PRED mór
          mór+Adj+Base+@PRED ] . .+Punct+Fin+<<< S]
209. [S   [V Tá bí+Verb+VI+PresInd+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD thíos
          thíos+Adv+Dir+@ADVL  ] [NP staighre
          staighre+Noun+Masc+Com+Sg+@NP  NP] . .+Punct+Fin+<<< S]
210. [S   [VS Tabhair tabhair+Verb+VD+Imper+2P+Sg+@FMV_SUBJ  ] [NP an
          an+Art+Sg+Def+@>N leabhar
          leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ  NP] [PP do
          do+Prep+Simp+@PP_OBL [NP Mháire
          Máire+Prop+Noun+Fem+Com+Sg+Len+@P<  NP]  PP] .
          .+Punct+Fin+<<< S]
211. [S   [VS Táim bí+Verb+VI+PresInd+1P+Sg+@FAUX_SUBJ  ] [ASP chun
          chun+Prep+Simp+@PP_ASP [INF [OI cáca
          cáca+Noun+Masc+Com+Sg+@OBJ_INF  OI] [I a a+Part+Inf+@>N
          dhéanamh déanamh+Verbal+Noun+VTI+Len+@INF  I]  INF]  ASP]
          [AD inniu inniu+Adv+Temp+@ADVL  ] . .+Punct+Fin+<<< S]
212. [S   [PP Tar_éis tar_éis+Prep+Cmpd+@PP_ADVL [NP trí
          trí+Num+Card+@>N lá lá+Noun+Masc+Com+Sg+Len+@P<  NP]  PP]
          [V tháinig tar+Verb+VI+PastInd+Len+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD abhaile
          abhaile+Adv+Dir+@ADVL  ] . .+Punct+Fin+<<< S]

213. [S   [V Tháinig tar+Verb+VI+PastInd+Len+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD abhaile
          abhaile+Adv+Dir+@ADVL  ] [AD an an+Art+Sg+Def+@>N oíche
          oíche+Noun+Fem+Com+Sg+DefArt+@ADVL sin sin+Det+Dem+@N<  ] .
          .+Punct+Fin+<<< S]
214. [S   [V Tháinig tar+Verb+VI+PastInd+Len+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD abhaile
          abhaile+Adv+Dir+@ADVL  ] [PP tar_éis
          tar_éis+Prep+Cmpd+@PP_ADVL [NP trí trí+Num+Card+@>N lá
          lá+Noun+Masc+Com+Sg+Len+@P<  NP]  PP] . .+Punct+Fin+<<< S]

215. [S   [V Thaistil taistil+Verb+VTI+PastInd+Len+@FMV  ] [NP Eoin
          Eoin+Prop+Noun+Masc+Com+Sg+@SUBJ  NP] [AD ní_ba
          ní_ba+Part+Comp+@>ADJ mhó mór+Adj+Comp+Len+@ADVL  ] [CJ2 ná
          ná+Conj+Coord+@CC [NP aon aon+Det+Qty+Idf+@>N duine
          duine+Noun+Masc+Com+Sg+@NP eile eile+Det+Dem+@N<  NP] CJ2]
          . .+Punct+Fin+<<< S]
216. [S   [V Thóg tóg+Verb+VTI+PastInd+Len+@FMV  ] [NP sé
          sé+Pron+Pers+3P+Sg+Masc+Sbj+@SUBJ  NP] [AD isteach
          isteach+Adv+Dir+@ADVL  ] [NP an an+Art+Sg+Def+@>N mhóin
          móin+Noun+Fem+Com+Sg+DefArt+@OBJ  NP] . .+Punct+Fin+<<< S]
217. [S   [V Thug tabhair+Verb+VD+PastInd+Len+@FMV  ] .
          .+Punct+Fin+<<< S]
218. [S   [V Thug tabhair+Verb+VD+PastInd+Len+@FMV  ] [NP Seán
          Seán+Prop+Noun+Masc+Com+Sg+@SUBJ Máire
          Máire+Prop+Noun+Fem+Com+Sg+@N< leabhar
          leabhar+Noun+Masc+Gen+Weak+Pl+@N<  NP] . .+Punct+Fin+<<< S]
219. [S   [V Thug tabhair+Verb+VD+PastInd+Len+@FMV  ] [NP Seán
          Seán+Prop+Noun+Masc+Com+Sg+@SUBJ Ó ó+Part+Pat+@>N Broin
          Broin+Prop+Noun+Masc+Com+Sg+@N<  NP] [NP leabhar
          leabhar+Noun+Masc+Com+Sg+@OBJ  NP] [PP do
          do+Prep+Simp+@PP_OBL [NP Mháire
```

Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

220. [S [V Thug tabhair+Verb+VD+PastInd+Len+@FMV] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP an
an+Art+Sg+Def+@>N leabhar leabhar+Noun+Masc+Com+Sg+@OBJ
NP] [PP do do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

221. [S [V Thug tabhair+Verb+VD+PastInd+Len+@FMV] [NP sí
sí+Pron+Pers+3P+Sg+Fem+Sbj+@SUBJ NP] [NP leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

222. [S [VS Thugamar tabhair+Verb+VD+PastInd+1P+Pl+Len+@FMV_SUBJ]
[NP an an+Art+Sg+Def+@>N leabhar
leabhar+Noun+Masc+Com+Sg+DefArt+@OBJ NP] [PP do
do+Prep+Simp+@PP_OBL [NP Mháire
Máire+Prop+Noun+Fem+Com+Sg+Len+@P< NP] PP] .
.+Punct+Fin+<<< S]

223. [S [NP Títhe Títhe+Guess+Prop+Noun+Masc+Com+Sg+@SUBJ lucht
lucht+Noun+Masc+Com+Sg+Len+@N< oibre
obair+Noun+Fem+Gen+Sg+@N< ba is+Part+Sup+@>ADJ mhó
mór+Adj+Comp+Len+@N< NP] [V a a+Part+Vb+Rel+Direct+@>V bhí
bí+Verb+VI+PastInd+Len+@FMV_REL] [PP ann
i+Pron+Prep+3P+Sg+Masc+@PP_ADVL PP] . .+Punct+Fin+<<< S]

224. [S [VS Tóg tóg+Verb+VTI+Imper+2P+Sg+@FMV_SUBJ] [AD go
go+Part+Ad+@>ADJ bog bog+Adj+Base+@ADVL] [NP é
é+Pron+Pers+3P+Sg+Masc+@OBJ NP] . .+Punct+Fin+<<< S]

225. [S [V Tuigeann tuig+Verb+VTI+PresInd+@FMV] [NP Nollaig
Nollaig+Prop+Noun+Fem+Com+Sg+@SUBJ NP] [AD níos
níos+Part+Comp+@>ADJ mó mór+Adj+Comp+@ADVL] [CJ2 ná
ná+Conj+Coord+@CC [NP Seán Seán+Prop+Noun+Masc+Com+Sg+@NP
NP] CJ2] . .+Punct+Fin+<<< S]

Appendix F: CG2 Dependency Mapping Rules

Listing of Dependency Mapping Rules for Irish

```
# ================================================================#
# IRISH DEPENDENCY MAPPING  PART 1
# CONSTRAINT GRAMMAR CG2
# ================================================================
# Elaine Uí Dhonnchadha 2008
# ================================================================
# Delimiters
# Sets
# Disambiguation rules
# ================================================================
# SENTENCE DELIMITERS
# ================================================================
DELIMITERS = "<.>" "<!>" "<?>" "<#>" "<</p>>" "<</s>>" ;
# ================================================================
# SETS
# ================================================================
#
SETS
LIST BOS = "<<p>>" "<<s>>" (>>>);
LIST EOS = (<<<); # end of sentence for vislcg.
LIST COMMA = "<,>" ;
SETS
# Any noun other than verbal-noun
# the are several types of Noun: +Noun, Subst+Noun, Prop+Noun,
Verbal+Noun,
# Guess+Noun, but all nouns except verbal nouns have number (even
guess nouns)
LIST NOUN-NOT-VN = (Noun Sg) (Noun Pl) ;
# a list of items which can precede a noun
LIST NOUN-PREMOD = (Art) (Det Poss) (Det Qty) (Num) ADJ-PRENOM ;
# a list of items which can follow a simple preposition
# (art def is used to exclude "sa" e.g. "shuigh sé faoi sa
chathaoir"
# rel clause: an rud as ar/Part Vb Rel(not Cop) tháinig
# thar/Prep a/Prep bheith/VNoun
# mar iad/Pron Pers
LIST POST-PREP = (Noun) (Art Def) (Det) (Pron) (Num) ADJ-PRENOM
(Part Vb Rel) (Prep Simp) (Punct Quo);
LIST OBJ-PRON = "í" "é" "iad" "iadsan" "ise" "eisean" ;
# this type of verbal noun can be modified by attributive adj.
# e.g. "ag mothú tinn" but not "ag déanamh mór", "a bheith tanaí"
LIST SENSORY-VN = "bí" "mothú" "breathnú" "éirí" ;
LIST VSYNTH = (Verb 1P) (Verb 2P) (Verb 3P) (Verb Auto) ;
LIST N-OR-REL = (Noun) (Rel)  ;
LIST TIME = "mí" "bliain" "lá" "ráithe" "uair" "seachtain";
LIST NOUN-OR-PRO = (Noun) (Pron Pers);
LIST PUNCT = (":");
# ================================================================
# MAPPINGS
# ================================================================
#
MAPPINGS
# CLB-SCOMP
MAP (@CLB) TARGET (Cop Dep); # Dúirt sé [gur Seán ...
MAP (@CLB) TARGET (Conj Subord); # ... [nuair ...
MAP (@CLB) TARGET (Part Vb Cmpl); # ... [nach mbíonn
MAP (@CLB) TARGET PUNCT;      # ... : Ar an maidin dár gcionn
MAP (@CLB) TARGET (Conj Coord) IF (1 (Verb)); # ... [mar bhí OR ...
[mar atá
# ...agus is léir; nó cad faoi
```

```
MAP (@CLB) TARGET (Conj Coord) IF (1 (Cop Pres) OR (Cop Past) OR
(Cop Pron) OR (Cop Q));
MAP (@CLB) TARGET (Conj Coord) IF (1 (Part Vb)) (NOT 1 (Part Vb
Rel)) (2 (Verb)); # ... [agus ná déan siúd
#============#
  END PART 1 #
#============#
# ================================================================
# IRISH DEPENDENCY MAPPING PART 2
# ================================================================
# SETS
# ============================================================ #
SETS
LIST PUNCT = (":");
LIST AUX = ("bí") ("téigh") ("tosaigh") ("tosnaigh") ("féad")
("caith") ("féach");
LIST MOD-AUX = ("féad") ("caith");
# a list of items which can precede a noun
LIST NOUN-PREMOD = (Art) (Det Poss) (Det Qty) (Num) ADJ-PRENOM ;
# Any noun other than verbal-noun
# the are several types of Noun: +Noun, Subst+Noun, Prop+Noun,
Verbal+Noun,
# Guess+Noun, but all nouns except verbal nouns have number (even
guess nouns)
LIST NOUN-NOT-VN = (Noun Sg) (Noun Pl) ;
LIST NP = (Noun Sg) (Noun Pl) (Pron Pers Sbj) (Pron Dem) (Abr);
LIST RELPART = (Vb Rel) (Prep Rel) ;
LIST OBJ-PRON = "í" "é" "iad" "iadsan" "ise" "eisean" ;
# ============================================================#
# MAPPINGS
# ============================================================
MAPPINGS
MAP (@CC) TARGET (Conj Coord) (NOT 0 (@CLB));
####################################################################
# Copula
####################################################################
MAP (@COP_SUBJ) TARGET (Cop Pro Dem) (NOT 0 (@CLB)) ; # copula; seo,
sin
MAP (@COP_WH) TARGET (Cop Q)  ; # copula: cad, céad, cén
MAP (@COP) TARGET (Cop) (NOT 0 (Cop Pro)) (NOT 0 (Cop Q))  (NOT 0
(@CLB) ) ; # copula
####################################################################
# Verbal Particles
####################################################################
MAP (@>V) TARGET (Part Vb) IF (NOT 0 (@CLB));
MAP (@>V) TARGET (Prep Rel); # lena n-áirítear
####################################################################
# Finite  Auxilliary (with Verbal Noun)
####################################################################
# FAUX Relatives + Synthetic
# there are two versions of the REL rules to cater for the synthetic
atá form of bí
# ag feabhsú atáimid ...
MAP (@FAUX_REL_SUBJ) TARGET AUX IF (0 (Verb Rel)) (0 VSYNTH) (*1
(Verbal Noun)) ;
MAP (@FAUX_REL_SUBJ) TARGET AUX IF (0 (Verb Rel)) (0 VSYNTH) (*-1
(Verbal Noun)) ;
MAP (@FAUX_REL_SUBJ) TARGET AUX IF (-1 RELPART) (0 VSYNTH)  (*1
(Verbal Noun));
MAP (@FAUX_REL_SUBJ) TARGET AUX IF (-1 RELPART) (0 VSYNTH)  (*-1
(Verbal Noun));
####################################################################
# FAUX Relatives + Analytic
# ag laghdú atá an daonra ...
```

```
# not a thabhairt do na pobalbhreitheanna atá fabhrach ...
MAP (@FAUX_REL) TARGET AUX IF (0 (Verb Rel)) (*-1 (Verbal Noun)
BARRIER (@CLB) OR NOUN-NOT-VN) (NOT 1 (Prep));
# conas atá sé ag rith?
# include Is mise atá ag déanamh cáca
MAP (@FAUX_REL) TARGET AUX IF (0 (Verb Rel)) (*1 (Verbal Noun)
BARRIER (@CLB));
# ag laghdú a bhí an daonra ...
# not a thabhairt do na pobalbhreitheanna nach raibh fabhrach ...
MAP (@FAUX_REL) TARGET AUX IF (-1 RELPART )  (*-1 (Verbal Noun)
BARRIER (@CLB) OR (Prep)) (NOT 1 (Prep));
# ag laghdú atá an daonra
MAP (@FAUX_REL) TARGET AUX IF (0 (Verb Rel))  (*-1 (Verbal Noun)
BARRIER (@CLB) OR (Prep)) (NOT 1 (Prep));
# conas a bhí sé ag rith?
MAP (@FAUX_REL) TARGET AUX IF (-1 RELPART )  (*1 (Verbal Noun)
BARRIER (@CLB)) (NOT 1 (Verbal Noun)) (NOT 1 (Prep));
# aux followed by np followed by vn => np = subj
# an lá a bhí/faux an fear/np ag snámh/vn
# but cant have vn before the np as in:
# daoine a bhíonn/faux ag gníomhú ...
MAP (@FAUX_REL) TARGET AUX IF (-1 RELPART )  (*1 NP BARRIER (Verbal
Noun) LINK *1 (Verbal Noun) BARRIER (@CLB) OR RELPART OR COMMA);
# a bhíonn á lorg
# is í teorainn an cheantar atá le feiceáil ...
# sin an fear atá a mhac ag déanamh na hoibre
MAP (@FAUX_REL) TARGET AUX IF (0 (Verb Rel)) (*-1 (Noun Com)
BARRIER (@CLB))   (*1 (Verbal Noun) BARRIER (@CLB)) (NOT 1 (Verbal
Noun));
# e.g. sin an fear a mbíonn a mhac ag déanamh na hoibre
# not atá iarraidh/vn
MAP (@FAUX_REL) TARGET AUX IF (-1 RELPART LINK *-1 (Noun Com)
BARRIER (@CLB))  (*1 (Verbal Noun) BARRIER (@CLB) OR (Verb)) (NOT 1
(Verbal Noun));
# an rud is measa a fhéadfadh (chaithfidh) tarlú ...
MAP (@FAUX_REL) TARGET AUX IF (0 ("féad") OR ("caith")) (-1 RELPART
LINK *-1 (Noun Com) BARRIER (@CLB))  (*1 (Verbal Noun) BARRIER
(@CLB));
#####################################################################
# FAUX Non-Relatives + Synthetic
# beimid ag imeacht le chéile
MAP (@FAUX_SUBJ) TARGET AUX IF (0 VSYNTH) (*1 (Verbal Noun) BARRIER
(@CLB)) ;
MAP (@FAUX_SUBJ) TARGET AUX IF (0 VSYNTH) (*-1 (Verbal Noun) BARRIER
(@CLB)) ;
# an rud is measa a fhéadfadh tarlú
# tá cáca le déanamh agam
# tá sé tar éis cáca a dhéanamh
# not: mar atá réamhráite, ar gníomhartha ...
#####################################################################
# FAUX Non-Relatives + Analytic
# not vn gen e.g go dtéann lucht eagraithe
MAP (@FAUX) TARGET AUX IF (NOT 0 VSYNTH) (*2 (Verbal Noun) BARRIER
(@CLB) OR (Punct) OR RELPART OR (Verb) ) ;
# other verbs e.g. théadh sé ag obair ...
# no NP between VN and AUX
# aird a thabhairt/vn do na pobalbhreitheanna nach raibh/fmv riamh
fabhrach do na páirtithe beaga
MAP (@FAUX) TARGET AUX IF (NOT 0 VSYNTH) (*-1 (Verbal Noun) BARRIER
NP OR (@CLB)) ;
#####################################################################
# Finite Main Verb
#####################################################################
# FMV Relatives + Synthetic
```

```
# a cuireadh tús le sraith
MAP (@FMV_REL_SUBJ) TARGET (Verb) IF (0 VSYNTH) (-1 RELPART) (NOT 0
AUX);
# a bhíodh ag Seán
MAP (@FMV_REL_SUBJ) TARGET (Verb) IF (0 VSYNTH) (-1 RELPART) (0 AUX)
(NOT *-1 (Verbal Noun)) (NOT *1 (Verbal Noun));
#  atáimse i dteagmáil
MAP (@FMV_REL_SUBJ) TARGET (Verb) IF (0 VSYNTH) (0 (Rel));
################################################################
# FMV Relatives + Analytic
# nach raibh riamh fabhrach/adv ...
# nach smaoiníonn ach/conj ..
# a chum é
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (Prep) OR (Adv) OR
(Conj) OR (Adj) OR (Pron Pers) OR COMMA) (-1 RELPART);
# a d´ ith
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (-1 (Part Vb)) (-2
RELPART);
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (Prep) OR (Adv) OR
(Conj) OR (Adj) OR COMMA) (0 (Verb Rel));
# comhlachais foirne atá ionadaitheach do mhúinteoirí
# an fear a bhfuil a mhac tinn
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (Det Poss)) (-1
RELPART);
# an fear atá a mhac tinn
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (Det Poss)) (0
(Verb Rel));
# an fear a phléasc
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (<<<)) (-1
RELPART);
# sin mar atá
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (1 (<<<)) (0 (Verb
Rel));
# an fear a phléasc
# is ansin go díreach a las solas dearg na gréine
# an brú/obj a chuir an GPA/subj orthu
# na factóirí a bhfuil (T) ina ndiaidh
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (NOT 1 (Prep)) (-1
(Indirect)) ;
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (NOT 1 (Prep)) (-1
RELPART) (NOT *-2 (Noun Com) BARRIER (@CLB)) ;
# mar atá sé déanta
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (NOT 1 (Prep)) (0
(Verb Rel)) (NOT *-1 (Noun Com) BARRIER (@CLB)) ;
# next one not so safe ... assume the subj is to the right of the
verb ...
# cigireachtaí a chinnfidh an tAire
MAP (@FMV_REL) TARGET (Verb) IF (NOT 0 VSYNTH) (NOT 1 (Prep) OR
(Adj) OR (Adv)) (-1 RELPART) (*1 NP BARRIER (@CLB)) ;
################################################################
# FMV Non-Relatives + Synthetic
# exception Tá´s (=tá fios) agam
MAP (@FMV_SUBJ) TARGET (Verb Noun) IF (NOT 0 (Rel)) ;
# beimid ar an bhfarraige
MAP (@FMV_SUBJ) TARGET (Verb) IF (0 VSYNTH ) (NOT 0 AUX) (NOT -1
RELPART) ;
MAP (@FMV_SUBJ) TARGET (Verb) IF (0 VSYNTH ) (0 AUX) (NOT -1
RELPART) (NOT 0 (Verb Rel)) (NOT *1 (Verbal Noun) LINK *1 (@CLB))
(NOT *-1 (Verbal Noun) BARRIER (Verb) OR (@CLB)) ;
################################################################
# FMV Non-Relatives + Analytic
# tá an carr sa gharáiste
MAP (@FMV) TARGET (Verb) IF (NOT 0 VSYNTH OR AUX) (NOT -1 RELPART)
(NOT -2 RELPART);
```

```
MAP (@FMV) TARGET (Verb) IF (0 AUX) (NOT 0 VSYNTH) (NOT -1 RELPART)
(NOT 0 (Verb Rel)) (NOT *-1 (Verbal Noun) BARRIER (@CLB) OR COMMA)
(NOT *1 (Verbal Noun) BARRIER (@CLB) OR COMMA OR (Verb));
#============#
  END PART 2 #
#============#
# ================================================================
# IRISH DEPENDENCY MAPPING PART 3
# ================================================================
# SETS
# ================================================================
SETS
LIST COMMA = "<,>" ;
SETS
# the genitive follows some simple prepositions and partitives,  as
well as another noun, verbal noun or compound preposition
LIST GEN-SIMP-PREP = "chun" "trasna" "timpeall" "fearacht" "dála"
"cois" ;
LIST GEN-PREP = "chun" "trasna" "timpeall" "fearacht" "dála" "cois"
(Prep Cmpd) ;
LIST GEN-PART = "roinnt" "cuid" "morán" "lán" "méid" "dosaen"
"péire" "scór" ;
LIST OBJ-PRON = "í" "é" "iad" "ea" ;
LIST NUM-COUNT = "haon" "dó" "trí" "ceathair" "cúig" "sé" "seacht"
"hocht" "naoi" "deich";
LIST NUM-PERS = "beirt" "triúr" "ceathrar" "cúigear" "seisear"
"seachtar" "ochtar" "naonúir" "deichnúir" ;
# prepositions commonly used before verbal nouns
# "ar" => lemma, "<ar>" => wordform
LIST PREP-VN = "ag" "gan" "a" "<á>" "ar" "tar éis" "chun" "le" "i
ndiaidh" "ar tí" "roimh" "<ina>";
LIST TITLE = "Uas." "Uas" "Dr." "Dr" "Mr" "Mr." "Mrs." "Mrs" "Miss"
"Misses" "Ms." "Ms" "<Athair>" "<tAthair>";
# this type of verbal noun can be modified by attributive adj.
# e.g. "ag mothú tinn" but not "ag déanamh mór", "a bheith tanaí"
LIST SENSORY-VN = "bí" "mothú" "breathnú" "éirí" ;
# Any noun other than verbal-noun
# the are several types of Noun: +Noun, Subst+Noun, Prop+Noun,
Verbal+Noun,
# Guess+Noun, but all nouns except verbal nouns have number (even
guess nouns)
LIST NOUN-NOT-VN = (Noun Sg) (Noun Pl) (Abr) ;
LIST NOUN-OR-VN = (Noun Sg) (Noun Pl) (Verbal Noun) ;
# a list of items which can precede a noun
LIST PRENOM = (Art) (Det Poss) (Det Qty) (Num) ;
# a list of items which can follow a simple preposition
# (art def is used to exclude "sa" e.g. "shuigh sé faoi sa
chathaoir"
# rel clause: an rud as ar/Part Vb Rel(not Cop) tháinig
# thar/Prep a/Prep bheith/VNoun
# mar iad/Pron Pers
LIST POST-PREP = (Noun) (Art Def) (Det) (Pron) (Num) (Part Vb Rel)
(Prep Simp) (Punct Quo);
# tá´s = tá fios = (Verb Noun)
LIST VSYNTH = (Verb 1P) (Verb 2P) (Verb 3P) (Verb Auto) (Verb Noun);
LIST TRANSV = (Verb VT) (Verb VTI) (Verb VD) ;
LIST TRANSVN = (Verbal VT) (Verbal VTI) (Verbal VD) ;
LIST N-OR-REL = (Noun) (Rel)  ;
LIST TIME-PERIOD = "mí" "bliain" "lá" "ráithe" "uair" "seachtain" ;
LIST TIME = "inné" "inniú" "amárach" "arú" "anocht" "aréir"
"istíoche" "tráthnóna" "ardtráthnóna" "Dé" "Déardaoin";
LIST ATTR-ONLY = "céanna" "amháin"; # not used predicatively
LIST ADJ-ATTR = (Adj Sg) (Adj Pl) (Adj Len) (Adj Ecl) "céanna"
"amháin"; # not used predicatively
```

```
LIST NOUN-OR-PRO = (Noun) (Pron Pers) (Pron Dem) (Pron Idf);
#Idf=ceachtar
LIST SUBJECT = (@SUBJ) (@FMV_SUBJ) (@FMV_REL_SUBJ) (@FAUX_SUBJ)
(@FAUX_REL_SUBJ) (@PP_SUBJ) (@COP_SUBJ) (@SUBJ_REL) (@SUBJ_OR_OBJ);
# e.g. an GPA(Abr)
LIST NOUN-NOM = (Noun Com) (Subst Noun) (Prop Noun) (Abr) (Unknown);
LIST VERB-REL-O = (VT @FMV_REL) (VTI @FMV_REL);
LIST VERB-SUBJ-O = (VT @FMV_SUBJ) (VT @FMV_REL_SUBJ) (VT @FAUX_SUBJ)
(VT @FAUX_REL_SUBJ) (VTI @FMV_SUBJ) (VTI @FMV_REL_SUBJ) (VTI
@FAUX_SUBJ) (VTI @FAUX_REL_SUBJ) ;
LIST ANY = (Noun) (Pron) (Abr) (Adv) (Adj) (Prep);
LIST ANY-NOT-ADJ = (Noun) (Pron) (Abr) (Adv) (Verb) (Prep);
LIST VERB-PREP = "ag" "ar" "as" "chun" "de" "do" "faoi" "i" "idir"
"ionsar"
"le" "ó" "roimh" "seach" "thar" "trí" "um";
LIST AUX = ("bí") ("téigh") ("tosaigh") ("tosnaigh") ("féad")
("caith") ("féach");
# ================================================================
# MAPPINGS
# ================================================================
MAPPINGS
MAP (@NP) TARGET (Noun Voc);
################################################################
# VERB + PREP = Phrasal Verb
################################################################
MAP (@PP_SUBJ) TARGET (Prep Simp) IF (0 VERB-PREP) (-1 (Verb)) (NOT
-1 VSYNTH OR AUX) (NOT *-1 (Rel)); # laghdaigh ar a neart
# d´eirigh go maith leis an mac léinn
MAP (@PP_SUBJ) TARGET (Prep Simp) IF (0 VERB-PREP) (-3 (Verb)) (-2
("go")) (-1 (Adj)) (NOT -1 VSYNTH OR AUX) (NOT *-1 (Rel));
#MAP (@V<+SUBJ) TARGET (Pron Prep) IF (-3 (Verb)) (-2 ("go")) (-1
(Adj)) (NOT -1 VSYNTH) (NOT *-1 (Rel));
MAP (@PP_SUBJ) TARGET (Pron Prep) IF (-3 (Verb)) (-2 ("go")) (-1
(Adj)) (NOT -1 VSYNTH OR AUX) (NOT *-1 (Rel));
# Prep Simp ??? NO d'éirigh leo
MAP (@PP_SUBJ) TARGET (Pron Prep) IF (-1 (Verb)) (NOT -1 VSYNTH)
(NOT *-1 (Rel)); # laghdaigh ar a neart
# ================================================================
# Pronouns
# ================================================================ #
# rith sí go tapadh
MAP (@SUBJ) TARGET (Pron Sbj) ; # sí/sé/siad
# tá mé tinn
# NOT Rinneamar é
# NOT a rinne é
# NOT tá sé/subj_asp ag déanamh cáca
MAP (@SUBJ) TARGET (Pron Pers) IF (-1 (Verb)) (NOT 1 (Prep Simp))
(NOT -1 VSYNTH) (NOT 0 OBJ-PRON);
# Buailfidh sé thú
# NOT Dá mba mise thú ...
MAP (@OBJ) TARGET (Pron Pers Len) IF (0 ("<thú>")) (*-1 (Verb)
BARRIER (Cop));
# ================================================================
# Bare Infinitive: ba mhaith liom fanacht; d´iarr sé orm fanacht
# ================================================================ #
# not do mo chabhrú
# not ina chónaí; i gcónaí
# not (Rácáil agus Scuabadh)
MAP (@INF) TARGET (Verbal Noun VI) IF (NOT -1 (Prep Simp) OR (Det
Poss) OR ("i")) (-1 (Pron) OR (Noun) OR (Verb)); #
# rud a rá
MAP (@>N) TARGET (Part Inf) IF (1 (Verbal Noun));
MAP (@INF) TARGET (Verbal Noun) IF (-1 (Part Inf));
# le/chun/taréis teacht/fanacht=INF
```

```
# (compare le bainne a ól vs ag ól bainne
MAP (@INF) TARGET (Verbal Noun) IF
      (-1 ("<le>") OR ("<chun>") OR ("<tar éis>")
         OR ("<i ndiaidh>") OR ("<gan>"));
#####################################################################
# Interrogatives
#####################################################################
# Cén cuma/fáth/chaoi a rinne sé
MAP (@PRED) TARGET NOUN-OR-PRO IF (-1 (Cop Q)) (NOT 0 OBJ-PRON);
# Cé a rinne é?
MAP (@SUBJ_OR_OBJ) TARGET (Pron Q);
#####################################################################
# COP + PREP(le)  OWNERSHIP: IS + LE
#####################################################################
# Is le Seán an rothar=subj.
MAP (@PP_PRED) TARGET (Prep Simp) IF (0 ("le")) (-1 (Cop)) (1 (Art
Def) OR (Prop) OR (Pron)) (2 (Art Def) OR (Prop) OR (Pron));
# Is liomsa é=subj. An leatsa é?
MAP (@PP_PRED) TARGET (Pron Prep) IF (0 ("le")) (-1 (Cop)) (1 (Art
Def) OR (Prop) OR (Pron));
# Is le Seán an rothar=subj.
# not Ní liomsa Seán
# not Is le Denise a bhí an Daibhéadach ..
MAP (@P<) TARGET (Noun) IF (-1 (@PP_PRED)) (-1 (Prep Simp)) (1
(Noun) OR (Pron) OR PRENOM);
#####################################################################
# COP + PREP (other)
#####################################################################
# b´fhearr liom é, is maith liom, is aoibhinn liom , is fearr dom
...
# ní mór dúinn
# ba=cop fearr=pred liom = pp_subj é=obj
MAP (@PP_SUBJ) TARGET (Pron Prep) IF (-1 (Adj)) (-2 (Cop));
#####################################################################
# COP + PRED
#####################################################################
MAP (@PRED) TARGET (Subst Noun Sg) IF (0 ("féidir")); # nach féidir,
is féidir, b'fhéidir
MAP (@PRED) TARGET (Adj Base) IF (0 ("cuma")) (-1 (Cop)); # is cuma,
ba chuma
MAP (@PRED) TARGET (Adj Base Len ) IF (-1 (Cop)); # ba chuma
MAP (@PRED) TARGET (Adj Comp Len ) IF (-1 (Cop)); # b´fhearr liom
...
#####################################################################
# AUGMENTED COPULA CONSTRUCTIONS (ACC)
#####################################################################
# IDENTIFICATION: DEFINITE NP i.e. def noun, prop noun or pronoun
# An é Seán a bhí ann: é = aug
# An é an carr atá mór?
# níorbh é!=aug a rinne é
# Sin=Cop+Pron é=subj
MAP (@AUG>SUBJ) TARGET OBJ-PRON IF (-1 (Cop)) (NOT -1 (Pron)) (*1
(Noun) OR (Pron Pers) BARRIER (@CLB) OR (ADJ) OR (Prep) OR (Verb));
# mise, é
# Is é é
MAP (@SUBJ) TARGET NOUN-OR-PRO IF (*-1 (@AUG>SUBJ) BARRIER NOUN-OR-
PRO );
# Is mise an múinteoir (id)
# NOT Is mise a thug ...
MAP (@SUBJ) TARGET (Pron Pers) IF (-1 (Cop)) (NOT 1 (Rel)); # mise,
é
# Is (é) Brian an múinteoir (id)
# anaphora if there is a previous noun
# muintir Cúige Uladh iad féin
```

```
# NOT Iad féin agus Séan ...
#####################################################################
# PREPOSITIONAL PHRASES
#####################################################################
# tá airgead agam; tá airgead ag Máire
# not Bhí seán ag an doras => obj of Prog must be animate (unlike
doras)
# use only Pron Prep or Prop Noun
MAP (@PP_HAS) TARGET ("ag" Prep) IF
      (1 (Prop Noun))
      (NOT -1 ("súil"))
      (NOT 1 (Verbal Noun))
      (*-1 ("bí") BARRIER (Verbal Noun) OR (@CLB)) ;
MAP (@PP_HAS) TARGET ("ag" Pron Prep) IF
      (NOT -1 ("súil"))
      (NOT 1 (Verbal Noun))
      (*-1 ("bí") BARRIER (Verbal Noun) OR (@CLB)) ;
# ===================================================================
# PREP PRON + NP = ADVERBIAL PHRASE
# ================================================================= #
MAP (@PP_ADVL) TARGET (Pron Prep);
# tá cáca le déanamh agam!=obj-i
# bhí an dinnéar ite agam=loc
# bhí rí ann=loc fadó
# arbh é é a bhí ann=loc
# ===================================================================
# PREP SIMP + NP
# ================================================================= #
MAP (@PP_NEG) TARGET (Prep Simp) IF (0 ("gan"));
# Níl ann ach spórt (negative polarity item
MAP (@PP_NEG) TARGET (Prep Simp) IF (0 ("ach"));
# ar oscailt
MAP (@PP_STAT) TARGET (Prep Simp) IF (0 ("ar")) (1 (Verbal Noun));
# ag gearradh
MAP (@PP_ASP) TARGET (Prep Simp) IF (NOT 0 ("ar")) (1 (Verbal
Noun));
# do mo ghearradh
MAP (@PP_ASP) TARGET (Prep Simp) IF (1 (Det Poss)) (2 (Verbal
Noun));
# chun/asp léitheoireacht/n a/inf fhoghlaim/vn
MAP (@PP_ASP) TARGET GEN-SIMP-PREP IF (*1 (Verbal Noun) BARRIER
(Noun Gen));
# chun/advl an tí: when chun is locative it is followed by noun in
the genitive
MAP (@PP_ADVL) TARGET GEN-SIMP-PREP IF (*1 (Noun Gen) BARRIER
(Noun));
# le seo; ó shin
MAP (@PP_ADVL) TARGET (Prep Simp) IF (1 (Pron Dem) );
# but: ... a labhair leis an mbean (not pp-loc if immed. after verb
)
# also Ind Obj e.g. thug sé an leabhar do Mháire
MAP (@PP_ADVL) TARGET (Prep Simp) IF (1 (Art) OR NOUN-NOT-VN OR ("a"
Det Poss 3P)) (NOT -1 (Verb)) (NOT *-1 (VD) BARRIER (Prep));
# mar amhránaí
# also i CFL (Abr)
# but not "thug sé an leabar do Mháire"
MAP (@PP_ADVL) TARGET (Prep Simp) IF (1 NOUN-NOT-VN) (NOT *-1 (VD))
(NOT *1 (VD));
# "thug sé an leabar do Mháire"
MAP (@PP_OBL) TARGET (Prep Simp) IF (*1 NOUN-NOT-VN BARRIER (Noun)
or (Verb) OR (Cop)) (*-1 (VD) BARRIER (Prep Simp));
#
MAP (@PP_ADVL) TARGET (Prep Simp) IF (1 PRENOM) (2 NOUN-NOT-VN); #
do mo mhamaí / le haon dream
```

```
# le trí chead agus a haon bliain
MAP (@PP_ADVL) TARGET (Prep Simp) IF (*1 NOUN-NOT-VN BARRIER (Verbal
Noun) OR (Verb));
MAP (@PP_ADVL) TARGET (Prep Simp) IF (1 (Num Dig)); # i 1977
# Líonadh an poll le clocha
# ================================================================
# PREP + ART
# ================================================================ #
MAP (@PP_ADVL) TARGET (Prep Art);
# ================================================================
# CMPD PREP
# ================================================================ #
# tá sé tar éis cáca a dhéanamh
MAP (@PP_ASP) TARGET (Prep Cmpd) IF (*1 NOUN-NOM BARRIER (Verbal
Noun)) (*-1 (@FAUX)  OR (@FAUX_SUBJ) BARRIER (Prep));
# tar éis dul i gcomhairle
MAP (@PP_ASP) TARGET (Prep Cmpd) IF (1 (Verbal Noun));
# tar_éis diúltiú
MAP (@PP_ASP) TARGET (Prep Cmpd) IF (*1 (Verbal Noun)) (*-1 (@FAUX)
OR (@FAUX_SUBJ) BARRIER (Prep));
MAP (@PP_ADVL) TARGET (Prep Cmpd);
MAP (@PP_ADVL) TARGET (Prep CmpdNoGen);
# ================================================================
# PREP POSS - STATIVE
# ================================================================ #
# stative: bhí sé ina mhúinteoir
# stative: bhí sé ina chodladh
MAP (@PP_STAT) TARGET (Prep Poss) IF (0 ("<ina>")) (1 NOUN-OR-VN)
(*-1 ("bí") BARRIER (Verb) OR (Verbal Noun)); #
#tá tú i do mhúinteoir/ i mo / in ár/ i bhur ...
MAP (@PP_STAT) TARGET (Prep Simp) IF (0 ("i")) (1 (Det Poss)) (2
NOUN-NOT-VN) (*-1 ("bí")); #
MAP (@PP_ADVL) TARGET (Prep Poss) IF (1 NOUN-NOT-VN); # ina dhiaidh
MAP (@PP_ADVL) TARGET (Prep Poss) IF (*1 NOUN-NOT-VN BARRIER
(Noun)); # ina ´gcúiseanna teanga´
MAP (@PP_ASP) TARGET ("do" Prep Poss) IF (1 (Verbal Noun)); # á
ngearradh
################################################################
# Noun Dependants
################################################################
# NOT tar éis cáca a dhéanamh
# Bryan Mc Fadden; Dr. O' Meara
# chun na scoile
# NOT chun cáca a dhéanamh
MAP (@P<) TARGET (Noun Gen) (*-1 GEN-PREP BARRIER (Noun) OR (Prep))
(NOT *1 (Part Inf) BARRIER (@CLB) OR (Rel)) ;
# allow for os_comhair an Bhreithimh
MAP (@N<) TARGET (Prop Noun) (-1 (Prop Noun));
MAP (@N<) TARGET (Prop Noun) (-1 (Part Pat)) (-2 (Noun) OR TITLE);
MAP (@N<) TARGET (Noun) (-1 (Part Pat)) (-2 (Noun) OR TITLE);
MAP (@N<) TARGET (Prop Noun) (-1 TITLE);
# ag déanamh gíoscáin
MAP (@OBJ_ASP) TARGET (Noun Gen) (-1 (Verbal Noun));
MAP (@OBJ_ASP) TARGET (Noun Gen) (-1 (Art Def)) (-2 (Verbal Noun) );
# NOT a eisiúint laistigh den tréimshe/!obj
# NOT gníomhaíochtaí grúpála agus athainmnithe
MAP (@OBJ_ASP) TARGET (Noun Gen) (*-1 (Verbal Noun) BARRIER (Noun)
OR (Prep) LINK NOT 0 (Gen));
# halla an bhaile
MAP (@N<) TARGET (Noun Gen) (*-1 NOUN-NOT-VN BARRIER (Prep Cmpd));
# possessive gen: hata Sheán;  tithe lucht/len oibre
# beirt fhear ?
MAP (@N<) TARGET (Noun Len) (-1 NOUN-NOT-VN);
################################################################
```

```
# SUBJ OF INFINITIVE
# should come before @P<
# ar an gcuntas a bheith ...-> prefer cuntas to be SUBJ_INF than P<
################################################################
# eagla a bheith orthu, aonad a bheith againn
MAP (@SUBJ_INF) TARGET NOUN-NOM IF
      (NOT 0 (Noun Gen))
      (*1 (Part Inf) BARRIER NOUN-NOM OR (Verbal Noun) OR (Pron
Pers) OR (Prep) LINK 1 (Verbal Noun VI) );
# iadsan a bheith ar an ...
MAP (@SUBJ_INF) TARGET (Pron Pers) IF
      (NOT 0 (Noun Gen))
      (*1 (Part Inf) BARRIER NOUN-NOM OR (Verbal Noun) OR (Pron
Pers) OR (Prep) LINK 1 (Verbal Noun VI) );
################################################################
# tá mé tar éis cáca a dhéanamh
# object of an transitive (VT/VTI/VD) infinitive ...
# na breoslaí seo a úsáid
# NOTE barrier = noun + pron pers, and applies to do + a
MAP (@OBJ_INF) TARGET NOUN-NOT-VN IF
      (NOT 0 (Noun Gen))
      (*1 (Part Inf) BARRIER (Noun) OR (Pron Pers) LINK 1 TRANSVN );
      #(*1 ("a" Prep Simp) OR ("do" Prep Simp) BARRIER (Noun) OR
(Pron Pers) LINK 1 TRANSVN );
# tá mé tar éis cáca a dhéanamh
MAP (@OBJ_INF) TARGET NOUN-OR-VN IF
      (-1 GEN-PREP)
      (*1 (Part Inf) BARRIER (Noun) OR (Pron Pers) LINK 1 TRANSVN );
      #(*1 ("a" Prep Simp) OR ("do" Prep Simp) BARRIER (Noun) OR
(Pron Pers) LINK 1 TRANSVN );
# object of an infinitive ...
# iad a glacadh; iad fhéin a scaoileadh
MAP (@OBJ_INF) TARGET (Pron Pers) IF
      (*1 (Part Inf) BARRIER (Noun) OR (Pron Pers) LINK 1 TRANSVN );
################################################################
# PP DEPENDANTS
################################################################
# aon amhras ann/pp-advl nach mbeadh na hathruithe seo ...
# also Abr: i CFL
# á dhéanamh, tar éis dul
MAP (@P<) TARGET (Verbal Noun) IF (-1 (Prep Simp) OR (Prep Poss) OR
(Prep Cmpd) OR (Det Poss));
MAP (@P<) TARGET (Pron Dem)  IF (-1 (@PP_ADVL));
MAP (@PC<) TARGET (Pron Idf)  IF (-1 (Prep Cmpd)); #de_bharr
ceachtar
MAP (@P<) TARGET (Pron Idf)  IF (-1 (Prep Simp)); #de_bharr ceachtar
MAP (@P<) TARGET NOUN-NOT-VN IF (-1 (@PP_ADVL)  OR (@PP_STAT) OR
(@PP_HAS)) (NOT -1 (Pron Prep) );
# Handle Quotes: rangíodh ina ´gcúiseanna teanga'iad
MAP (@P<) TARGET NOUN-NOT-VN IF (*-1 (Prep Poss) BARRIER (Noun) OR
(@P<));
# Handle pre-mods on np after prep phrase: idir thuas dhá sheamair
MAP (@P<) TARGET NOUN-NOT-VN IF (*-1 (Prep Simp) BARRIER (Noun) OR
(Rel) OR (@P<));
MAP (@P<) TARGET NOUN-NOT-VN IF (-1 (@PP_NEG));
MAP (@P<) TARGET NOUN-NOT-VN IF (-1 PRENOM) (-2 (@PP_ADVL)) (NOT -2
(Pron Prep));
MAP (@P<) TARGET NOUN-NOT-VN IF (-1 PRENOM) (-2 PRENOM) (-3
(@PP_ADVL)) (NOT -3 (Pron Prep));
################################################################
# NP DEPENDANTS
################################################################
# is ise a thug an  leabhar do Mháire!=scomp
# Is ionann sin agus a rá
```

```
MAP (@SUBJ) TARGET (Pron Pers) IF (*-1 (Cop) BARRIER (N-OR-REL) LINK
NOT *1 (Verb)); # mise, é
# An fíor é
# Dá mba mise/s thú=pred ní dhéanfainn/s é/obj
MAP (@SUBJ) TARGET (Pron Pers) IF (*-1 (@PRED) BARRIER (Noun) OR
(Verb) OR (Cop));
# iadsan atá faoi ionsaí
# TO BE TESTED
MAP (@SUBJ) TARGET (Pron Pers) IF (1 (RELPART) OR (Verb Rel));
# small clause: agus é ag caint leis féin
# but not sé: sé ag caint
# is there always a conj before ????
MAP (@SUBJ_ASP) TARGET NOUN-OR-PRO IF (1 ("ag" Prep Simp)) (2
(Verbal Noun));
MAP (@SUBJ) TARGET (Pron Dem) IF (-1 (@PRED) LINK *-1 (Cop));
################################################################
# Is mise=subj Brian=pred (identity sentence)
MAP (@PRED) TARGET (Noun) IF (*-1 (@SUBJ) BARRIER (@CLB) OR (Part
Rel) LINK *-1 (Cop) BARRIER (Verb) OR (Noun)) (NOT *-1 (@PRED)
BARRIER (@CLB)) (NOT *1 (@PRED) BARRIER (@CLB));
# Sin an fear ...
MAP (@PRED) TARGET (Noun) IF (*-1 (@COP_SUBJ) BARRIER (@CLB) OR
(Part Rel)) (NOT *-1 (@PRED) BARRIER (@CLB)) (NOT *1 (@PRED) BARRIER
(@CLB));
# Is an leabhar a thug ...
MAP (@PRED) TARGET (Noun) IF (*-1 (@COP) BARRIER (@CLB) OR (Part
Rel)) (*1 (Rel) BARRIER NOUN-OR-PRO) (NOT *-1 (@PRED) BARRIER
(@CLB)) (NOT *1 (@PRED) BARRIER (@CLB));
MAP (@PRED) TARGET (Pron Pers) IF (*-1 (@COP) BARRIER (@CLB) OR
(Part Rel)) (*1 (Rel) BARRIER NOUN-OR-PRO) (NOT *-1 (@PRED) BARRIER
(@CLB)) (NOT *1 (@PRED) BARRIER (@CLB));
# Dá mba mise thú (Is mise thú)
MAP (@PRED) TARGET (Pron Pers) IF (*-1 (@SUBJ) BARRIER (@CLB) OR
(Part Rel) LINK *-1 (Cop) BARRIER (Verb) OR (Noun)) (NOT *-1 (@PRED)
BARRIER (@CLB)) (NOT *1 (@PRED) BARRIER (@CLB));
# Is é Seán
MAP (@SUBJ) TARGET (Noun) IF (*-1 (@AUG>SUBJ) BARRIER N-OR-REL);
# CLASSIFICATION: INDEFINITE
# Is múinteoir=pred é=subj-classificatory
# Is deas=pred an lá=subj
MAP (@SUBJ) TARGET (Noun) IF (*-1 (@PRED)  BARRIER (@CLB) OR (Rel))
(NOT *1 (@SUBJ)) (NOT *-1 (@SUBJ));
MAP (@SUBJ) TARGET (Noun) IF (*-1 (@PP_PRED) BARRIER (@CLB)) (NOT *1
(@SUBJ)) (NOT *-1 (@SUBJ));
# Is fear Seán, Is deas an leabhar, Is fear maith é
# not Is leabhar!=pred a thug sé di
# ní caomhnóir láidir a bhí ...
# pred must be indef, subj must be def
MAP (@PRED) TARGET (Noun) IF (-1 (Cop)) (NOT 0 (Prop Noun)) (*1
(Prop Noun) OR (Pron Pers) OR (Adj) BARRIER (Noun) OR (Vb Rel));
MAP (@PRED) TARGET (Noun) IF (-1 (Cop)) (NOT 0 (Prop Noun)) (*1
(Art) LINK 1 (Noun) BARRIER (Noun) OR (Vb Rel));
# cailín is ea í
MAP (@PRED) TARGET (Noun) IF (*1 (Cop) LINK 1 ("ea") BARRIER
(Noun));
#Is deas an lá
#Ní móide go ndéantar ...
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (-1 (Cop)) ;
# tá sé fada
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (-1 (Pron Pers) LINK *-
1 ("bí") BARRIER (@CLB) OR (Prep));
# tá an bóthar fada
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (-1 (Noun) LINK *-1
("bí") BARRIER (@CLB) OR (Prep)) (-2 (Art));
```

```
# Tá an geata dúnta
# tá ... ag na Stáit Aontatithe
MAP (@PRED) TARGET (Verbal Adj) IF (-1 NOUN-NOM OR (Pron Pers) LINK
*-1 ("bí") BARRIER (@CLB) OR (Prep));
# Tá an geata sin dúnta
MAP (@PRED) TARGET (Verbal Adj) IF (*-1 NOUN-NOM OR (Pron Pers) LINK
*-1 ("bí") BARRIER (@CLB) OR (Prep));
# a bheith cláruithe
MAP (@PRED) TARGET (Verbal Adj) IF (-1 (Verbal Noun) LINK *-1 ("bí")
BARRIER (@CLB) );
# Bhíomar tinn inné
# may need base see previous ????
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (-1 ("bí") BARRIER
(@CLB));
# it is predicative if the noun is definitite and the verb is "bí"
# a bhfuil a mhac tinn
# NOTE is + Adj Com = @N<
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (NOT -1 (Part Sup)) (*-
1 ("bí") BARRIER (Rel) OR (@CLB)) (*-1 (Noun) LINK -1 (Art Def) OR
(Det));
# Bhí louis sásta
MAP (@PRED) TARGET (Adj) IF (NOT 0 ADJ-ATTR) (NOT -1 (Part Sup)) (*-
1 ("bí") BARRIER (Rel) OR (@CLB)) (*-1 (Prop Noun) BARRIER (Verb));
# mar atá réamhráite
MAP (@PRED) TARGET (Verbal Adj) IF (-1 ("bí") BARRIER (@CLB));
# inné etc which have Adj tag
MAP (@ADVL) TARGET TIME;
# uair éigin
MAP (@ADVL) TARGET TIME-PERIOD IF (1 ("éigin"));
# an lá sin/seo
MAP (@ADVL) TARGET TIME-PERIOD IF (1 (Det Dem));
# an lá a cuireadh ...
MAP (@ADVL) TARGET TIME-PERIOD IF (1 (Part Rel));
# bliain ó shin
MAP (@ADVL) TARGET TIME-PERIOD IF (1 ("ó")) (2 ("sin"));
# aon uair = anytime
MAP (@ADVL) TARGET TIME-PERIOD IF (-1 ("aon")) ;
# inar luigh slán
MAP (@ADVL) TARGET (Adj Base) IF (-1 (Verb)) ;
# bíonn gais ghlasa
# Sin lá deas
# not Is airde=comparative sliabh ná cnoc
# not Is deas an lá
# not chomh=Its deas
# not Tá sé déanta
# not Tá [an|a] mac tinn/Pred
MAP (@N<) TARGET ADJ-ATTR IF
      (NOT 0 (Comp))
      (NOT -1 (Part Ad) OR (Cop) OR (Its) OR (Pron)) ;
#####################################################################
# VERB DEPENDANTS
#####################################################################
MAP (@SUBJ) TARGET (Pron Pers Sbj) IF (-1 (Verb)) (NOT -1 VSYNTH);
#TEST
# ciallaíonn sin do raibh ...
MAP (@SUBJ) TARGET (Pron Dem) IF (-1 (Verb)) (NOT -1 VSYNTH); #
# rinneamar é
MAP (@OBJ) TARGET (Pron Pers) IF (-1 VSYNTH); #
# ná déan seo agus ná déan siúd
MAP (@OBJ) TARGET (Pron Dem) IF (-1 VSYNTH); #
#a d´ionsaigh iad
MAP (@OBJ) TARGET OBJ-PRON IF (-1 (@FMV_REL)); #
# rinne sé é
MAP (@OBJ) TARGET (Pron Pers) IF (*-1 (VT) OR (VTI)) (-1 (@SUBJ));
```

```
#generalise it: scrúdóidh an cigire sin é
MAP (@OBJ) TARGET (Pron Pers) IF (*-1 (VT) OR (VTI)) (*-1 (@SUBJ)
BARRIER (Verb) OR (@CLB));
# tóg go bog é
MAP (@OBJ) TARGET (Pron Pers) IF (*-1 (VT @FMV_SUBJ) OR (VTI
@FMV_SUBJ));
# it is ind obj if it is preceded by prep which is preceded by VD
(ditrans.)
MAP (@P<) TARGET (Noun Com) IF (-1 (Prep Simp) LINK *-1 (VD)) ;
MAP (@P<) TARGET (Noun Dat) IF (-1 (Prep Simp) LINK *-1 (VD)) ;
MAP (@P<) TARGET (Noun Com) IF (-1 (Prep Simp) LINK *1 (VD) LINK -1
(Part Rel)) ;
MAP (@P<) TARGET (Noun Dat) IF (-1 (Prep Simp) LINK *1 (VD) LINK -1
(Part Rel)) ;
####################################################################
# bhí louis sásta
# bíonn na bláthanna bán
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FMV) BARRIER NOUN-OR-PRO);
MAP (@SUBJ) TARGET (Item) IF (*-1 (@FMV) BARRIER NOUN-OR-PRO);
# ... nuair a bhí tuirse air ...
# an fear a chonaic an bhean
# bean could be subj or obj ...
# not subj_or_obj if verb is synthetic (i.e. includes subject)
# not subj_or_obj if verb is intransitive (i.e. has no object)
# not subj_or_obj if verb is preceded byy an adverbial only (i.e. no
subj or obj)
MAP (@SUBJ_OR_OBJ) TARGET NOUN-NOM IF
      (*-1 (Part Vb Rel Direct)  BARRIER NOUN-OR-PRO LINK *-1 NOUN-
OR-PRO)
      (NOT *-1 ("bí") BARRIER (Rel))
      (NOT *-1 VSYNTH BARRIER (Rel))
      (NOT *-1 (VI)  BARRIER (Rel));
# cé nár ith an dinnéar?
MAP (@SUBJ_OR_OBJ) TARGET NOUN-NOM IF (*-1 ("nár")  BARRIER NOUN-OR-
PRO);
MAP (@SUBJ) TARGET NOUN-NOM IF (-1 PRENOM) (-2 (@FMV_REL) BARRIER
(@CLB));
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FMV_REL)  BARRIER (@CLB) OR
SUBJECT OR NOUN-NOM) (NOT *1 SUBJECT BARRIER (@CLB));
# iad a  bhfuil (T)/Item ina ndiaidh ...
MAP (@SUBJ) TARGET (Item) IF (*-1 (@FMV_REL)  BARRIER (@CLB) OR
SUBJECT OR NOUN-NOM) (NOT *1 SUBJECT BARRIER (@CLB));
# this is (probably) the subj if there is a rel. verb to the right
# with subj to its left and there is no other subj in the clause
MAP (@SUBJ) TARGET NOUN-NOM IF (*1 (@FMV_REL)  BARRIER (@CLB) OR
SUBJECT) (NOT *-1 SUBJECT BARRIER (@CLB));
# sin a bhfuil ann
MAP (@SUBJ) TARGET (Pron Dem) IF (*1 (@FMV_REL)  BARRIER (@CLB) OR
SUBJECT) (NOT *-1 SUBJECT BARRIER (@CLB));
# NOT ó réim an Ombudsman
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FMV_REL) BARRIER SUBJECT OR
(Prep Simp));
################################################
# BÍ AUX
# bhí Seán ag fáil airgead
# bhí go_leor eile ag fáil airgead
#MAP (@OBJ_ASP) TARGET NOUN-NOM IF (*-1 (@FAUX) BARRIER NOUN-OR-PRO)
(*1 (Verbal Noun) BARRIER NOUN-OR-PRO);
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FAUX) BARRIER NOUN-OR-PRO) (*1
(Verbal Noun) BARRIER NOUN-OR-PRO);
MAP (@SUBJ) TARGET NOUN-NOM IF (*-1 (@FAUX) BARRIER NOUN-OR-PRO);
MAP (@SUBJ) TARGET NOUN-NOM IF (-1 (@FAUX_REL) BARRIER (@CLB));
MAP (@SUBJ) TARGET NOUN-NOM IF (-1 PRENOM) (-2 (@FAUX_REL) BARRIER
(@CLB));
```

```
# chonaic máire an fear/rel_subj a bhí ag iascaireacht
MAP (@SUBJ_REL) TARGET NOUN-NOM IF (*1 (@FAUX_REL) BARRIER (@CLB) OR
(Verb) OR SUBJECT);
# subject before rel verb if intrans i.e. prep following verb => no
dir. obj.
MAP (@SUBJ_REL) TARGET NOUN-NOM IF (NOT *-1 SUBJECT BARRIER (@CLB))
(NOT *1 SUBJECT BARRIER (@CLB)) (*1 (Part Rel) LINK 1 (VI) OR (VTI)
LINK 1 (Prep));
# cúig caibidil atá i CFL ...
MAP (@SUBJ) TARGET NOUN-NOM IF (NOT *-1 SUBJECT) (NOT *1 SUBJECT)
(*1 (Verb Rel) LINK 1 (Prep));
# an fear a bhfuil a mhac ag imeacht
# => fear = subj if "a" Det Poss follows verb
# exclude Cop Pron Dem as possible SUBJ as already tagged COP_SUBJ
MAP (@SUBJ) TARGET NOUN-OR-PRO IF (NOT 0 (Cop)) (*1 (Part Vb Rel)
LINK 1 (Verb) LINK 1 (Det Poss));
####################################################################
# dúirt sé gur múinteoir!=obj-d é
# D'inis sí an scéal
# Cheannaigh sí leabhair áit!=obj-d a bhí siopa ann
# not Suíonn Timi agus Ronna
MAP (@OBJ) TARGET NOUN-NOM  IF
 (*-1 TRANSV BARRIER (Verb))
 (*-1 SUBJECT BARRIER (Verb) OR (Prep) OR (Cop))
 (NOT *-1 (@OBJ) BARRIER (Verb))
 (NOT *-1 (Coord) BARRIER (Noun));
# NOTE: d'fhoilsigh sé féin agus Eoin => do not allow part of
conjoint to be an
# obj if the first part is not an obj
# do rule for conjoints ...
#MAP (@OBJ) TARGET NOUN-NOM IF (*-1 SUBJECT BARRIER (Verb) OR (Prep)
OR (Cop)) (NOT *-1 (@OBJ))(*-1 TRANSV) (NOT *-1 (Coord) BARRIER
NOUN-NOM);
MAP (@OBJ) TARGET NOUN-NOM  IF
 (*-1 TRANSV BARRIER (Verb))
 (*-1 SUBJECT BARRIER (Verb) OR (Prep) OR (Cop))
 (*-1 (@OBJ) LINK *1 (Coord) BARRIER NOUN-NOM OR (Punct));
MAP (@OBJ) TARGET NOUN-NOM IF
 (*-1 SUBJECT BARRIER (Verb) OR (Prep) OR (Cop))
 (NOT *-1 (@OBJ))
 (*1 TRANSV) ;
# cloigeann capaill a fheictear in armas Marsh ...
MAP (@OBJ) TARGET NOUN-NOM IF (*1 VERB-REL-O LINK *1 SUBJECT BARRIER
(@CLB));
# an lá a cuireadh Butt ...
MAP (@OBJ) TARGET NOUN-NOM IF (*1 VERB-SUBJ-O BARRIER (@CLB)) (NOT
*-1 (@OBJ)) (NOT *1 (@OBJ));
# an brú/obj a chuir an GPA/subj orthu
MAP (@OBJ) TARGET NOUN-NOM IF (*1 (VT @FMV_REL) OR (VTI @FMV_REL) OR
(VD @FMV_REL) BARRIER (Noun)) ;
# an teach a raibh sé ina chónaí ann=resumptive pron=obj
MAP (@OBJ) TARGET NOUN-NOM IF (*1 (@FAUX_REL) BARRIER (Noun) LINK *1
(Prep Pron)) ;
MAP (@N<) TARGET (Det Dem) IF (*-1 (Noun) BARRIER (Pron));
# iad sin
MAP (@PN<) TARGET (Det Dem) IF (*-1 (Pron) BARRIER (Noun));
# na Ballstáit uile
MAP (@N<) TARGET (Det Qty) (0 ("uile")) IF (*-1 (Noun) BARRIER
(Pron));
# iad uile
MAP (@PN<) TARGET (Det Qty) (0 ("uile")) IF (*-1 (Pron) BARRIER
(Noun));
MAP (@>N) TARGET (Art) (NOT 0 (Cop)); # cén = cop pron art;
MAP (@>N) TARGET (Poss);
```

```
MAP (@>N) TARGET (Det);
MAP (@>N) TARGET (Item) IF (1 (Noun));
# list items will be NPs
# (8) Déanfaidh Cígire ...
MAP (@NP) TARGET (Item) IF (NOT *1 (Noun) BARRIER (Punct) OR
(Verb));
MAP (@>N) TARGET (Num Dig PC) IF (1 ("de"));
MAP (@N<) TARGET (Num Dig) IF (-1 ("Euro") OR ("euro"));
MAP (@P<) TARGET (Num Dig) IF (-1 ("i")) (NOT *1 (Noun) BARRIER
(@CLB) OR (Rel));
MAP (@>N) TARGET (Nm);
MAP (@>N) TARGET (Num) IF (1 (Noun));
MAP (@N<) TARGET (Num) IF (-1 (Noun));
MAP (@NP) TARGET (Num) IF (NOT -1 (Noun)) (NOT 1 (Noun));
# chun a trí, tar_éis a dó
MAP (@P<) TARGET (Num) IF (-1 (Nm)) (-2 (Prep));
MAP (@>N) TARGET NUM-PERS IF (1 NOUN-NOT-VN); #do bheirt iníon
MAP (@>N) TARGET (Part Pat);
MAP (@>N) TARGET (Part Voc);
# á dhéanamh, an déanamh, tar éis dul, do mo chabhrú
# tarlú
# le mím agus gluaiseacht!=INF bailé=Gen
MAP (@INF) TARGET (Verbal Noun) IF (NOT -1 (Prep Simp) OR (Prep
Poss) OR (Art) OR (Prep Cmpd) OR (Det Poss))(-1 (Pron) OR (Noun) OR
(Verb)) (NOT 1 (Noun Gen));
MAP (@PN<) TARGET (Pron Ref) IF (-1 (Pron)); #é féin
MAP (@N<) TARGET (Pron Ref) IF (-1 (Noun)); #an tiarna féin
MAP (@N<) TARGET (Adj) IF (-1 (@CC)) (-2 (Adj)); # glas agus bán
MAP (@PN<) TARGET (Pron) IF (-1 (@CC)) (-2 (Pron)); # dúirt sé agus
é ag caint
MAP (@ADVL) TARGET (Adj) IF ( -1C (Part Ad));
# Bí go deas
MAP (@>ADJ) TARGET (Part Ad);
# deich mbliana nó níos mó
MAP (@>ADJ) TARGET (Part Comp) IF (-1 ("<nó>"));
MAP (@>ADJ) TARGET (Part Sup);
MAP (@>ADJ) TARGET (Adv Its);
MAP (@>ADJ) TARGET (Deg);
# tag all the comp. adjs which are use PRED as ADVL like chomh maith
# níos déanta
MAP (@ADVL) TARGET (Adj) IF (-1 (Part Comp));
# tag all superlatives as noun modifiers @N<
# daoine eile is cuí/ ba chuí
# is déanta; an ceann is fearr, fadhbanna is mó
MAP (@N<) TARGET (Adj) IF (-1 (Part Sup));
MAP (@ADVL) TARGET (Adv) IF (NOT 0 (Its));
# chomh/Its minic/A sin
MAP (@ADVL) TARGET (Adj) IF (-1 (Adv Its)) ;
# chomh/Its minic/A sin=le_sin=>ADVL
MAP (@ADVL) TARGET (Pron Dem) IF (-1C (Adj)) ;
MAP (@N<) TARGET (Prop Noun) (-1 (Noun)) (-2 (Prep Simp)); #
MAP (@N<) TARGET (Prop Noun) (-2 (Noun)) (-1 (Art Def)) (-3 (Prep
Simp)); #
MAP (@NP) TARGET (Noun) (NOT 0 (Verb)); # Tá's = tá/v + fios/n
#Liam P. Ó Riain
MAP (@N<) TARGET (Abr) IF (-1 (Prop));
# used for Adj ending in a vowel which only have Adj Base tags and
# therefore did not get tagged as attributive adjectives
MAP (@N<) TARGET (Adj Base) IF (-1 (Noun) OR (Adj));
# used for Verbal Adj used attributively
# ina dhéanmhas cosanta
MAP (@N<) TARGET (Verbal Adj) IF (-1 (Noun) OR (Adj));
MAP (@>N) TARGET (Abr) IF (1 (Prop));
# in Óglaigh an IRA
```

```
MAP (@N<) TARGET (Abr) IF (-2 (Noun)) (-1 (Art Def)) (-3 (Prep
Simp));
MAP (@N<) TARGET (Abr) IF (-1 (Noun)) (-2 (Prep Simp));
MAP (@NP) TARGET (Abr) IF (NOT -1 (Prop)) (NOT 1 (Prop));
# maidir_le ginmhilleadh
MAP (@ADVL) TARGET (Prep CmpNoGen);
MAP (@PC<) TARGET (Noun Com) IF (-1 (Prep CmpNoGen));
MAP (@NP) TARGET (Pron Pers); # Iad/NP uile faoi shuan ...
MAP (@NP) TARGET (Pron Idf); # cibé acu
MAP (@COM) TARGET ("," Punct Int);
#============#
  END PART 3 #
#============#
```

Appendix G: Finite-State Chunker Regular Expresions

Irish Chunker Regular Expressions

```
########################################################################
# This regex file is for chunking dependency mapped sentences.
# To be used with Xerox XFST Tools
# Input format "token lemma+MTags+@DTag token lemma+MTags+@DTag etc.
# PART 1
########################################################################
# Alphabet used for tokens and lemmas
define Alpha
[a|á|b|c|d|e|é|f|g|h|i|í|j|k|l|m|n|o|ó|p|q|r|s|t|u|ú|v|w|x|y|z|A|Á|B
|C|D|E|É|F|G|H|I|Í|J|K|L|M|N|O|Ó|P|Q|R|S|T|U|Ú|V|W|X|Y|Z|1|2|3|4|5|6
|7|8|9|%0|%.|%,|%-|%+|%*|%/|%>|%<|%?|%:|'|''|%_|%@];
# Alphabet used for Morphological Tags
define MAlpha
[a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z|A|B|C|D|E|F|G|H
|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z|1|2|3|%_];
# Alphabet used for Dependency Tags
define DAlpha
[A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z|%<|%>|%_];
# define whitespace
define SP [" "|"\n"|"\t"]+ ;
########################################################################
# Tag/Token/Lemma definitions
define TokLem     [Alpha+ SP Alpha+ ];    # chuir cuir
define MTag       [%+ MAlpha+];           # +Verb
define TokLemMTag [TokLem MTag+ %+];      # chuir cuir|+Verb+Past|+
define DTag       [%@ DAlpha+];           # @FMV
# Quo has no dependency tag at present - this may change
define QuoTag     [%+Punct%+Quo];
define TokLemQTag [TokLem QuoTag];        # ' '+Punct+Quo
########################################################################
# Space followed by optional Quote
define SPQ [SP (TokLemQTag SP)];
########################################################################
# Verb Dependency Tags
define VTag       [%@FAUX|%@FAUX%_REL|%@FMV|%@FMV%_REL];
define VSTag
[%@FAUX%_SUBJ|%@FAUX%_REL%_SUBJ|%@FMV%_SUBJ|%@FMV%_REL%_SUBJ];
define PreVTag    [%@%>V];
define PostVTag   [%@V%<];
# Verb Pre & Post Modifiers
define PreVStr    [TokLemMTag PreVTag SP];
# Verb Chunk
define VStr       [TokLemMTag VTag SP];
define VChunk     [PreVStr* VStr]; # old VChunk [PreVStr* VStr
PostVStr*];
define VChunkBr   [VChunk @-> "[V " ... " ] "];
# Verb_Subject Chunk
define VSStr      [TokLemMTag VSTag SP];
define VSChunk    [PreVStr* VSStr]; # old VSChunk [PreVStr* VSStr
PostVStr*];
define VSChunkBr  [VSChunk @-> "[VS " ... " ] "];
########################################################################
# Infinitive
define ITag       [%@INF];
define IStr       [TokLemMTag ITag SP];
define PreITag    [%@%>N];
define PreIStr    [TokLemMTag PreITag SP];
define IChunk     [(PreIStr) IStr];
define IChunkBr   [IChunk @-> "[I " ... " I] "];
########################################################################
```

```
# Noun Dependency Tags
define NTag       [%@NP| %@OBJ|

%@SUBJ|%@SUBJ%_ASP|%@SUBJ%_INF|%@SUBJ%_REL|%@SUBJ%_OR%_OBJ|
              %@P%<|%@PC%<
             ];
define PreNTag    [%@%>N|%@AUG%>SUBJ];
define PostNTag   [%@N%<|%@PN%<];
# Adjectival modifier separators
define SepTag1    [%@COM];          # camóg (comma)
define SepTag2    [%@CC];           # agus, nó (and, or)
# Noun Pre & Post Modifiers
define PatTag     [%+Part%+Pat];        # Ó, Ní, Uí, Mac etc. in
names
define TokLemPTag [TokLem PatTag];        # Ó ó+Part+Pat
define SupTag     [%+Part%+Sup%+%@%>ADJ];        # superlative "is"
define TokLemSTag [TokLem SupTag];        # is is+Part+Pat
define SupStr     [TokLemSTag SP];  # ,
define ArtTag     [["an an+Art"]|["na na+Art"]|["a
an+Art"]|TokLemPTag];
define TokLemATag [ArtTag MTag* %+];
define ArtStr     [TokLemATag PreNTag SP];
define PreGStr    [TokLemMTag PreNTag SPQ];     # doesnt include Art
define PreNStr    [[TokLemATag|TokLemMTag] PreNTag SPQ]; # includes
Art
define SepStr1    [TokLemMTag SepTag1 SPQ];     # ,
define SepStr2    [TokLemMTag SepTag2 SPQ];     # agus (and)
define PostNStr0  [TokLemMTag PostNTag SPQ];    # ard (tall)
define PostNStr1  [SepStr1 PostNStr0];          # , tanaí (, thin)
define PostNStr2  [SepStr2 PostNStr0];          # agus caol (and
narrow)
define PostNStr3  [SupStr PostNStr0];           # is mó
define PostNStr   [PostNStr0|PostNStr3|[PostNStr0 PostNStr1+
(SepStr1) (PostNStr2)]];
# Noun Chunk
define NStr       [TokLemMTag NTag SPQ];
define GHead      [NStr PostNStr*];
define GChunk     [(ArtStr) PostNStr*];
define NGChunk    [GHead GChunk];
define NChunk     [PreNStr* NStr PostNStr* ((ArtStr PreGStr*)
PostNStr+)];
define NChunkBr1  [NChunk @-> "[NP " ... " NP] "];
######################################################################
# NP: Object of Aspectual
define OATag       [%@OBJ%_ASP];
define OAStr       [TokLemMTag OATag SP];
define OAChunk     [PreNStr* OAStr PostNStr* ((ArtStr PreGStr*)
PostNStr+)];
define OAChunkBr   [OAChunk @-> "[OA " ... " OA] "];
######################################################################
# NP: Object of Infinitive
define OITag       [%@OBJ%_INF];
define OIStr       [TokLemMTag OITag SP];
define OIChunk     [PreNStr* OIStr PostNStr* ((ArtStr PreGStr*)
PostNStr+)];
define OIChunkBr   [OIChunk @-> "[OI " ... " OI] "];
######################################################################
# Preposition Dependency Tags
define PPASTag       [%@PP%_ASP|%@PP%_STAT];
define PPASStr       [TokLemMTag PPASTag SP];
define PPADTag
[%@PP%_ADVL|%@PP%_HAS|%@PP%_NEG|%@PP%_OBL|%@PP%_PRED|%@PP%_SUBJ];
define PPADStr       [TokLemMTag PPADTag SP];
```

```
# Simple and Compound Prepositions with NP complement
define PSimpTag     [[%+Prep%+Simp]|  # le(Prep Simp) = with;
                [%+Prep%+Poss]|  # lena(Prep Poss) = with its;
                [%+Prep%+Cmpd]|  # ar nós (Prep Cmpd) = such as
                [%+Prep%+CmpdNoGen]|  # maidir le (Prep Cmpd) =
regarding
                [%+Prep%+Art]];  # sa (Prep Art) = in the
define TokLemPSTag  [TokLem PSimpTag MTag* %+];  #
define PPSimpStr    [TokLemPSTag PPADTag SP];
define PPChunkBr2   [[PPSimpStr "[NP " ?+ " NP] "] @> "[PP " ... "
PP] "];
# Aspectual PPs
define PPASSimpStr   [TokLemPSTag PPASTag SP];
# do mo chabhrú
define PPChunkBr3   [[PPASSimpStr ("[OA " ?+ " OA] ") "[NP " ?+ "
NP] "] @> "[PP-ASP " ... " PP-ASP] "];
# Conjugated Prepositions
# These preps. incorporate a pronoun. Therefore PP has no nested NP
complement.
define PPronTag    [%+Pron%+Prep];              # liom = with me
define TokLemPPTag [TokLem PPronTag MTag+ %+];  # liom le Tags +
define PPPronStr   [TokLemPPTag PPADTag SP];
# can have "leis(Pron Prep) féin(Pron Ref)" = with himself
define PPChunkBr1   [PPPronStr PostNStr0*  @-> "[PP " ... " PP] "];
######################################################################
# Adverbial Dependency Tags
define ADTag       [%@ADVL];
define PreADTag    [%@%>ADJ];
define PostADTag   [%@ADVL%<];
# Adverbial Pre & Post Modifiers
define PreADStr    [TokLemMTag PreADTag SP];
define PostADStr   [TokLemMTag PostADTag SP];
# Adverbial Chunk
define ADStr       [TokLemMTag ADTag SP];
define ADChunk     [PreADStr* ADStr PostADStr*];
define NADChunk    [PreNStr* ADStr PostNStr*];  # an tseachtain seo
- this week
define ADChunkBr   [[ADChunk|NADChunk] @-> "[AD " ... " ] "];
######################################################################
# Copula Dependency Tags
define COPTag       [%@COP|%@COP%_WH|%@COP%_SUBJ];
define COPStr       [TokLemMTag COPTag SP];
define COPChunkBr   [COPStr @-> "[COP " ... "] "];
######################################################################
# Clause Boundary Dependency Tags
define CBTag       [%@CLB];   # go
define PreCBTag    [%@CC];    # ná go
define PreCBStr    [TokLemMTag PreCBTag SP];
define CBStr       [TokLemMTag CBTag SP];
define CBChunkBr   [PreCBStr* CBStr @-> "[CB " ... "] "];
######################################################################
# Predicate Dependency Tags
define PRTag       [%@PRED];
define PostPRTag   [%@PRED%<];
define PRStr       [TokLemMTag PRTag SP];
define PostPRStr   [TokLemMTag PostPRTag SP];
# go maith (@ADJ> @PRED)
define PRChunk     [PreADStr* PRStr PosPRStr*];
# na breiseán bia
# teorainn an cheantair
define NPRChunk    [PreNStr* PRStr PostPRStr* ((ArtStr PreGStr*)
PostNStr+)];
# is mór an trua
```

```
define NPRChunk2    [PreNStr* PRStr (ArtStr PostPRStr) PostNStr*];

define PRChunkBr   [[PRChunk|NPRChunk|NPRChunk2] @-> "[PRED " ... "]
"];
#####################################################################
# 2nd Conjoint Dependency Tags
define CJTag     [%@CC|%@CS];
define CJ2Str    [TokLemMTag CJTag SP];
define ConjStr   ["[NP " ?+ " NP] "
                 | "[PP " ?+ " PP] "
                 | "[V " ?+ " ] "
                 | "[VS " ?+ " ] "
                 | "[AD " ?+ " ] "
                 | "[COP " ?+ " ] "
                 | "[PRED " ?+ " ] " ];
define CJ2ChunkBr1   [[CJ2Str ConjStr] @> "[CJ2 " ... "CJ2] "];
#####################################################################
# Bracketed Sentence
define Sen   [?* @-> "[Z " ... " Z]"];
#####################################################################
# Define Chunker1
define Chunker [VchunkBr .o. VSChunkBr .o. COPChunkBr .o.
CBChunkBr .o. IchunkBr .o. OIChunkBr .o. OAChunkBr .o.
NChunkBr1 .o. PRChunkBr .o. ADChunkBr .o. PPChunkBr1 .o.
PPChunkBr2 .o. PPChunkBr3 .o. CJ2ChunkBr1 .o. Sen ];

#####################################################################
# PART 2
# Input format "token lemma+MTags+@DTag token lemma+MTags+@DTag etc.
#####################################################################

#####################################################################
# Infinitival Phrases
define INFChunkBr   [(PPSimpNStr) [("[OI " ?+ " OI] ") "[I " ?+ " I]
"]  @> "[INF " ... " INF] " ];
#####################################################################
define ASPChunkBr1   [PPSimpAStr ("[OI " ?+ " OI] ") "[INF " ?+ "
INF] " @> "[ASP " ... " ASP] "] ;
define ASPChunkBr2   ["[PP-ASP " ?+ " PP-ASP] " ("[OA " ?+ " OA] ")
@-> "[ASP " ... " ASP] "] ;
#####################################################################
# 2nd Conjoint Dependency Tags
define CJTag     [%@CC|%@CS];
define CJ2Str    [TokLemMTag CJTag SP];
define ConjStr   [ "[ASP " ?+ " ASP] "
                 | "[PP-ASP " ?+ " PP-ASP] "
                 | "[I " ?+ " I] "
                 | "[OI " ?+ " OI] "
                 | "[INF " ?+ " INF] "  ];
define CJ2ChunkBr2   [[CJ2Str ConjStr] @> "[CJ2 " ... "CJ2] "];
#####################################################################
# Bracketed Sentence
define Sen   [?* @-> "[S " ... " S]"];
#####################################################################
define Chunker2 [INFChunkBr
.o. ASPChunkBr1 .o. ASPChunkBr2
.o. CJ2ChunkBr2
.o. Sen
];
```

Appendix H: Finite-State To Parole Tag Mappings

FS Morphology Tags	Parole Tags	Description
Nouns		**incl. Verbal Nouns (VN)**
Noun	N	
Com	c	Common case
Dat	d	Dative case
Emph	e	Emphatic form
Fem	f	Feminine
Gen	g	Genitive case
Masc	m	Masculine
Nstem	n	VN Nominal stem
Pl	p	Plural
Sg	s	Singular
Verbal	v	VN Verbal Stem
Voc	v	Vocative case
Verbs		
Verb	V	
1P	1	1st. Person
2P	2	2nd. Person
3P	3	3rd. Person
Auto	0	Autonomous
Cond	c	Conditional
Dep	d	Dependent form
Fut	f	Future
FutInd	if	Future indicative
Imper	m	Imperative
Ind	i	Independent
Neg	n	Negative
Past	s	Past
PastImp	ih	Past Imperfect
PastInd	is	Past Indicative
PastIndDep	isd	Past Indicative Dependent
PastSubj	ss	Past Subjunctive
Pres	p	Present
PresImp	ig	Present Imperfect
PresInd	ip	Present Indicative
PresSubj	sp	Present Subjunctive
Rel	r	Relative
Subj	s	Subjunctive
Adjectives		
Adj	A	Adjective
Base	p	Base form
Comp	c	Comparative form
Pronouns		
Pron	P	Pronoun
Dem	d	Demonstrative
Idf	i	Indefinite
Pers	p	Personal
Prep	p	Prepositional
Q	q	Interogative
Ref	x	Reflexive
Sbj	s	Subject form
Determiner		
Det	D	Determiner
Poss	p	Possessive

Q	w	Interrogative
Qty	q	Quantifier
Article		
Art	T	Article
Adverb		
Adv	R	Adverb
Dir	d	Directional
Gn	g	General
Its	i	Intensifier
Loc	l	Locative
Temp	t	Temporal
Preposition		
Prep	S	Preposition
Art	a	Article
Cmpd	c	Compoud
Deg	d	Degree
Obj	o	Obj
Poss	p	Poss
Conjunction		
Conj	C	
Coord	c	co-ordinating
Cop	w	incl. Copula
Subord	s	subordinating
Numerals		
Num	M	numeral
Card	c	ordinal
Dig	n	digit
Op	s	operator
Ord	o	ordinal
Rom	r	roman
UniMember		
Ad	a	adverbial
Cp	w	comparative
Nm	m	numeral
Pat	p	patronymic
Sup	s	superlative
Punctuation		
Punct	P	
Bar	b	hyphen, dash etc
Fin	e	final
Int	i	internal
Quo	a	quote
Copula		
Cop	W	
Cond	s	Conditional
Neg	n	Neg
NegQ	nq	NegQ
Past	si	Past
Pres	pi	Pres
PresSubj	ps	PresSubj
Pron	3	Pron
RelInd	s	RelInd
Pro	p	Pronoun

Verbal Particle		
Part+Vb	U	
Part	Q	Part
Direct	r	Direct
Indirect	i	Indirect
Alphabetical Index		
1P	1	
2P	2	
3P	3	
Ad	a	
Adj	A	
Adv	R	
Art	A	(Adposition)
Art	T	(Article)
Auto	0	
Bar	b	
Base	p	
Card	c	
Cmpd	c	
Com	c	
Comp	C	
Cond	c	(Verbs)
Cond	S	(Copula)
Conj	C	
Coord	c	
Cop	w	(Conjunction)
Cop	W	(Copula)
Cp	w	
Dat	d	
Deg	d	
Dem	d	
Dep	d	
Det	D	
Dig	n	
Dir	d	
Direct	r	
Emph	e	
Fem	f	
Fin	e	
Fut	f	
FutInd	if	
Gen	g	
Gn	g	
Idf	i	
Imper	m	
Ind	i	
Indirect	i	
Int	i	
Its	i	
Loc	l	
Masc	m	
Neg	n	
NegQ	nq	

Nm	m	
Noun	N	
NSStem	n	
Num	M	
Obj	o	
Op	s	
Ord	o	
Part	Q	
Part|Vb	U	
Past	s	(Verb)
Past	si	(Copula)
PastImp	ih	
PastInd	is	
PastIndDep	isd	
PastSubj	ss	
Pat	p	
Pers	p	
Pl	p	
Poss	p	
Prep	p	(Pronouns)
Prep	S	(Adposition)
Pres	p	(Verbs)
Pres	pi	(Copula)
PresImp	ig	
PresInd	ip	
PresSubj	ps	(Copula)
PresSubj	sp(Verbs)	
Pro	p	
Pron	3	(Copula)
Pron	P	(Pronoun)
Punct	P	
Q	q	(Pronouns)
Q	w	(Determiner)
Qty	q	
Quo	a	
Ref	x	
Rel	r	
RelInd	s	
Rom	r	
Sg	s	
Subj	s	
Subord	s	
Sup	s	
Temp	t	
Verb	V	
Verbal	v	(Verbal Nouns)
VerbSubj	s	
Voc	v	