Partial Dependency Parsing for Irish

Elaine Uí Dhonnchadha[1], Josef Van Genabith[2]

[1]Centre for Language and Communication Studies,
Trinity College, Dublin 2, Ireland.
[2]Centre for Next Generation Localisation,
Dublin City University, Glasnevin, Dublin 9, Ireland.
uidhonne@tcd.ie, josef@computing.dcu.ie

Abstract

In this paper we present a partial dependency parser for Irish, in which Constraint Grammar (CG) rules are used to annotate dependency relations and grammatical functions in unrestricted Irish text. Chunking is performed using a regular-expression grammar which operates on the dependency tagged sentences. As this is the first implementation of a parser for unrestricted Irish text (to our knowledge), there were no guidelines or precedents available. Therefore deciding what constitutes a syntactic unit, and how it should be annotated, accounts for a major part of the early development effort. Currently, all tokens in a sentence are tagged for grammatical function and local dependency. Long-distance dependencies, prepositional attachments or coordination are not handled, resulting in a partial dependency analysis. Evaluations show that the partial dependency analysis achieves an f-score of 93.60% on development data and 94.28% on unseen test data, while the chunker achieves an f-score of 97.20% on development data and 93.50% on unseen test data.

1. Introduction

This paper presents the development and evaluation of a partial dependency parser for Irish. Using Constraint Grammar (CG), dependency relation tags and grammatical function tags are applied to text which has already been morphosyntactically analysed and disambiguated (Uí Dhonnchadha and van Genabith, 2006) In addition, we annotate the dependency tagged text with chunk boundaries using a regular-expression grammar which operates on the dependency tags. Currently, all tokens in the surface structure of sentences are tagged for grammatical function or local dependency.

2. Methodology

We implemented a dependency parser rather than a constituency parser for a number of reasons. Firstly, there are a number of unresolved issues in the theoretical syntax of Irish which could present problems for a constituency analysis, e.g. the non-adjacency of verb and object in a VSO language, and the treatment of several periphrastic aspectual constructions in Irish. By carrying out a dependency analysis we can annotate functional dependencies and other grammatical notions such as subject, object etc. which are somewhat clearer.

A second reason for choosing a dependency analysis is that we can develop the parser using Constraint Grammar (1995; Tapanainen, 1996; 1999), which integrates well with previous work in POS tagging for Irish (Uí Dhonnchadha and van Genabith, 2006).

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 specific lexically-marked rules to handle special cases. Neither does it attempt to determine constituency structure.

The primary aim of this exploratory work is to account for as much of the linguistic phenomena of Irish as possible and to decide on an initial style guide for the partial syntactic annotation of the language. In developing a parser 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 begin, relevant grammatical and syntactic studies were consulted, (e.g. (Ó hUallacháin and Ó Murchú, 1981); (Doherty, 1996); (Stenson, 1981); (Biber et al., 2003)), and a number of short, grammatical, sample sentences, covering the main syntactic phenomena were devised. These sample sentences (225 approx.) constituted our initial Test Suite.

The Test Suite sentences were anotated using the Constraint Grammar Dependency Rules being developed. These annotated sentences were manually corrected and used as a gold standard in the iterative development and testing of the CG Dependency Rules (250 approx. to date).

The final results reported in the Abstract and Section 5 are based on unseen real data randomly selected from a 30 million word corpus of Irish (Kilgarriff *et al.*, 2007).

3. Dependency Parsing

Constraint Grammar (CG) is used to produce a dependency analysis of the already POS tagged sentences (Karlsson, 1995, p33). Surface syntactic dependency labels (highlighted in bold type) are appended to the existing morphosyntactic tags of each token, as shown in (1). By convention, the dependency tags all start with the @ symbol to distinguish them from tags which have already been appended to the token in the POS tagging stage.

```
(1) Fuair sé leabhar ins an  siopa
    Got   he book    in  the shop
    'He got a book in the shop'

    Fuair   faigh+Verb+VTI+PstInd+Len+@FMV
    sé      sé+Pro+Per+3P+Sg+Msc+Sbj+@SUBJ
    leabhar leabhar+Noun+Masc+Com+Sg+@OBJ
    ins     i+Prep+Art+Sg+@PP_ADVL
    an      an+Art+Sg+Def+@>N
    siopa   siopa+Noun+Msc+Cm+Sg+DefArt+@P<
    .       .+Punct+Fin.
```

Table 1 shows a sample of the Dependency Tags which are used in this dependency analysis (for full details see (Uí Dhonnchadha, 2009; 2010)). While this tagset follows the style of tags described for English (Karlsson, 1995), and for Danish (Bick, 2003),[1] there is not a prescribed list of tags for CG. This allows us to tailor the tagset to the language under consideration.

A dependency analysis consists of a root, and leaf nodes, without intermediate levels, therefore, the tokens present in the input string are annotated without introducing any abstract categories (phrasal nodes or ellipted or elided items). Clause boundaries and head-modifier dependencies within clauses are identified, as well as the grammatical functions of subject, object, predicate, and various types of prepositional phrase, e.g. adverbial, aspectual, predicative, etc.

Dependent modifiers can come before or after the head, therefore the tag specifies the direction of the head they modify, e.g. the tag @>N marking a noun pre-modifier, points to a head noun to the right, whereas @N< points to a head noun to the left.

As a sentence with multiple clauses can have more than one subject, a number of different subject tags are used, i.e. @SUBJ_INF, @SUBJ_ASP and @SUBJ_REL for subjects of infinitival, aspectual and relative clauses respectively (similarly for object labels).

In (2), the subject of the main clause is annotated as @SUBJ, whereas the subject of the subordinate relative clause is annotated as @SUBJ_REL. This sentence also distinguishes the finite main verb, @FMV, from the relative finite auxiliary verb, @FAUX_REL.

TAG	DESCRIPTION	EXAMPLE
@>N	pre-modifier dependent on the first noun to the right	*an* 'the'
@CLB	clause boundary	e.g. *agus* 'and' when followed by a verb, and subordinating conjunctions. etc.
@COP	copula	
@FAUX	finite auxiliary verb	*Tá sé ag cócaireacht* 'He is cooking'
@FMV	finite main verb	*rith* 'run'
@N<	noun post-modifier, e.g. adj.	*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'
@PP_SUBJ	prep + subject 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'
@PP_ADVL	adverbial PP head	*ag an doras* 'at the door'
@PP_ASP	aspectual PP head	*ag rith* '(at) running'
@PP_PRED	predicative PP	*Is liom é* 'It is mine' i.e. Is with me it
@SUBJ	subject	*Chonaic Seán Máire*, 'Seán saw Máire'
@SUBJ_REL	subject of relative clause	*a rinne sé* 'that he made'

Table 1 Selected Dependency Tags for Irish

```
(2) Chonaic Máire an fear a   bhí ag ithe
    Saw     Máire the man  that was at eating
    'Máire saw the man who was eating'

    Chonaic feic+Verb+VTI+PastInd+Len+@FMV
    Máire   Máire+Prop+Noun+Fem+Cm+Sg+@SUBJ
    an      an+Art+Sg+Def+@>N
    fear    fear+Noun+Masc+Com+Sg+@SUBJ_REL
    a       a+Part+Vb+Rel+Direct+@>V
    bhí     bí+Verb+VI+PastInd+Ln+@FAUX_REL
    ag      ag+Prep+Simp+@PP_ASP
    ithe    ithe+Verbal+Noun+@P<
    .       .+Punct+Fin
```

Sentence (2) also illustrates the treatment of the progressive aspect in Irish, where *ag ithe* '(at) eating' is

[1] Other languages are also detailed on the VISL website: http://visl.sdu.dk/corpus_linguistics.html (last accessed on 29 Oct 2009).

tagged as a preposition and verbal noun at the POS level, but the preposition, which is functioning aspectually in this instance, is annotated as such using the dependency analysis tag @PP_ASP. The copular verb *bí* 'is' is tagged as a finite auxiliary, @FAUX, as the verbal noun *ithe* 'eating' carries the semantic content. Example (3) shows a progressive aspectual construction with an object.

```
(3) Tá mé ag déanamh cáca.
    Is I  at making  cake.
    'I am making a cake.'

    Tá      bí+Verb+VI+PresInd+@FAUX
    mé      mé+Pron+Pers+1P+Sg+@SUBJ_ASP
    ag      ag+Prep+Simp+@PP_ASP
    déanamh déanamh+Verbal+Noun+VTI+@P<
    cáca    cáca+Noun+Masc+Gen+Sg+@OBJ_ASP
    .       .+Punct+Fin+<<<
```

Only one dependency tag is applied to each token, thereby avoiding the re-introduction of ambiguity to the disambiguated POS output, e.g. rather than applying both @SUBJ and @OBJ tags to head nouns (i.e. (Karlsson *et al.*, 1995)) and later removing the inappropriate tag, we attempt to apply the correct tag in the first instance.

In contrast to full dependency parsers such as FDG (Tapanainen and Järvinen, 1997) or MaltParser (Nivre and Hall, 2005), this implementation does not explicitly mark the head associated with a dependent (usually encoded in terms of numerical indices). Currently this information is largely recoverable from the tagset and the marking of clause boundaries. All modifier dependency tags include either '<' or '>' indicating the direction of the head, which in the current implementation will always be the first such head in the indicated direction. All heads have a grammatical relation tag, and where appropriate additional information is included in the tag indicating the clause, e.g. in (4) the tag @OBJ_ASP indicates that the token is the object of the aspectual construction. All other tokens are dependent on a head, and therefore receive a dependency relation tag.

4. Chunking

As already stated, constituents are not marked in a dependency analysis, but using the dependency annotations and a regular expression grammar (implemented using Xerox Finite-State Tools[2]) we can identify phrase-like structures, described by Abney (1991) as 'chunks'. In (4), we show the results of chunking the sentence in (3).

```
(4)  [S
      [V Tá bí+Verb+VI+PresInd+@FAUX  V]
      [NP mé mé+Pron+Pers+1P+Sg+@SBJ_ASP NP]
      [ASP
       [PP-ASP ag ag+Prep+Simp+@PP_ASP
        [NP déanamh déanamh+Verbal+Noun+@P<
         NP]
        PP-ASP]
       [OA cáca cáca+Nn+Msc+Gen+Sg+@OBJ_ASP
        OA]
       ASP]
     ..+Punct+Fin+<<<
    S]
```

In order to see how the analysis performs on real corpus data, a sentence from the 250 sentence Gold Standard Evaluation Corpus is given in (5).

```
(5)  Ach sin an toradh is measa a fhéadfadh
     tarlú don pháirtí agus déarfaidís leat
     nár cóir an iomad airde a thabhairt do
     na pobalbhreitheanna nach raibh riamh
     fabhrach do na páirtithe beaga.
     'But that is the worst possible result
     for the party and they would say to you
     that it is not right to pay too much
     attention to the opinion polls that
     were never favourable to small
     parties.'

[S
[CONJ  Ach ach+Conj+Subord+@CLB ]
[COP   Sin sin+Cop+Pro+Dem+@COP_SUBJ ]
[NP    an an+Art+Sg+Def+@>N
       toradh
       toradh+Noun+Msc+Com+Sg+DefArt+@PRED
       is is+Part+Sup+@>ADJ
       measa olc+Adj+Comp+@N< NP]
[VP    a a+Part+Vb+Rel+Direct+@CLB
       fhéadfadh
       féad+Verb+VTI+Cond+Len+@FAUX_REL ]
[INF   tarlú tarlú+Verbal+Noun+VTI+@INF
INF]
[PP    don do+Prep+Art+Sg+@PP_ADVL
[NP    pháirtí
       páirtí+Noun+Masc+Com+Sg+Len+@P< NP]
PP]
[CB    agus agus+Conj+Coord+@CLB ]
[V     déarfaidís
       abair+Verb+VTI+Cond+3P+Pl+@FMV+SUBJ
[PP    leat le+Pron+Prep+2P+Sg+@PP_ADVL PP]
[COP   nár is+Cop+Past+Rel+Neg+@CLB ]
[PRED  cóir cóir+Adj+Base+@PRED ]
[INF   an an+Art+Sg+Def+@>N
       iomad iomad+Subst+Noun+Sg@OBJ_INF
```

[2] For details see http://www.cis.upenn.edu/~cis639/docs/xfst.html (Accessed 29/10/2009)

```
       airde aird+Noun+Fem+Gen+Sg+@N<

[I     a a+Prep+Simp+@PP_INF

       thabhairt
       tabhairt+Verbal+Noun+VTI+Len+@P<
I] INF]

[PP    do do+Prep+Simp+@PP_ADVL

[NP    na na+Art+Pl+Def+@>N

       pobalbhreitheanna
       pobalbhreith+Noun+Fem+Com+Pl+@P<
 NP]PP]

[V     nach nach+Part+Vb+Neg+Rel+@CLB

       raibh
       bí+Verb+PastInd+Neg+Len+@FMV_REL ]

[PRED  riamh riamh+Adv+Its+@>ADJ ]

       fabhrach fabhrach+Adj+Base+@PRED ]

[PP    do do+Prep+Simp+@PP_ADVL

[NP    na na+Art+Pl+Def+@>N

       páirtithe
       páirtí+Noun+Masc+Com+Pl+DefArt+@P<

       beaga beag+Adj+Com+NotSlen+Pl+@N<
 NP] PP]
. +Punct+Fin

S]
```

Identifying the relationships between certain chunks (i.e. PP attachment and co-ordination) is beyond the scope of the current work, as are issues relating to long-distance dependencies.

5. Evaluation

As the short grammatical sentences of the Test Suite may not cover all the basic linguistic phenomena and as they do not take into account the complexity of naturally occurring language or frequency of usage, a Gold Standard Dependency evaluation corpus based on real text was constructed. This consists of 250 attested sentences (average length 25.4 words), randomly selected from 3,000 sentences of the Gold Standard POS-tagged Corpus (Uí Dhonnchadha, 2009; 2010), which were randomly selected from a 30 million word corpus of Irish (Kilgarriff *et al*., 2007). The partial dependency analysis currently achieves an f-score of 93.60% on development data (150 sentences) and 94.28% on unseen test data (100 sentences). The chunker achieves an f-score of 97.20% on the development data and 93.50% on unseen the test data.

6. Future Work

It is hoped to extend this framework to full parsing, by addressing PP attachment, co-ordination and long-distance dependencies, as well as moving from CG2 to CG3.[3] Although the current parsing scheme is partial in nature, we hope that it will provide a basis for future work in the parsing of Irish.

[3] CG3 from VISL http://beta.visl.sdu.dk/cg3.html (Accessed 29/10/2009)

7. References

Abney, S. 1991. Parsing by Chunks. In *Principle-Based Parsing*, eds. Robert Berwick, Stephen Abney and Carol Tenny. Dordrecht: Kluwer Academic Publishers.

Biber, D., Conrad, S., and Leech, G. 2003. *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Bick, E. 2003. A CG & PSG Hybrid Approach to Automatic Corpus Annotation. Paper presented at *SProLaC2003 Corpus Linguistics Conference*, Lancaster.

Doherty, C. 1996. Clausal structure and the Modern Irish Copula. *Natural Language and Linguistic Theory* 14:1-46.

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.

Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E. 2007. Efficient corpus creation for lexicography. *Language Resources and Evaluation Journal*.

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*.

Ó hUallacháin, C., and Ó Murchú, M. 1981. *Irish Grammar*: University of Ulster Coleraine.

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.

Uí Dhonnchadha, E. 2009. Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar, School of Computing, Dublin City University: Unpublished PhD Thesis.

Uí Dhonnchadha, E. 2010. *Natural Language Processing Tools: Developing a Part-of-Speech Tagging and Partial Parsing for Irish*. Köln: LAP Lambert Academic Publishing.

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.