Irish Treebanking and Parsing: A Preliminary Evaluation

Teresa Lynn[1,2], Özlem Çetinoğlu[3*] , Jennifer Foster[1], Elaine Uí Dhonnchadha[4], Mark Dras[2] and Josef van Genabith[1]

[1]NCLT/CNGL, Dublin City University, Ireland,
[2]Department of Computing, Macquarie University, Sydney, Australia,
[3]IMS, University of Stuttgart, Germany,
[4]Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
[1]{tlynn,jfoster,josef}@computing.dcu.ie
[2]mark.dras@mq.edu.au, [3]ozlem@ims.uni-stuttgart.de, [4]uidhonne@tcd.ie

Abstract

Language resources are essential for linguistic research and the development of NLP applications. Low-density languages, such as Irish, therefore lack significant research in this area. This paper describes the early stages in the development of new language resources for Irish – namely the first Irish dependency treebank and the first Irish statistical dependency parser. We present the methodology behind building our new treebank and the steps we take to leverage upon the few existing resources. We discuss language-specific choices made when defining our dependency labelling scheme, and describe interesting Irish language characteristics such as prepositional attachment, copula and clefting. We manually develop a small treebank of 300 sentences based on an existing POS-tagged corpus and report an inter-annotator agreement of 0.7902. We train MaltParser to achieve preliminary parsing results for Irish and describe a bootstrapping approach for further stages of development.

Keywords: Dependency, Treebank, Irish

1. Introduction

Despite enjoying the status of an official EU language, Irish is considered a minority language. To date, little research has been carried out on computational analysis or processing, resulting in a lack of important linguistic resources. In this project, we assess the feasibility of employing a bootstrapping approach to develop such resources for a low-density language.

As a verb-initial language, Irish has several features that are uncharacteristic of many languages previously studied in parsing research. Our work broadens the application of NLP methods to less-studied language structures and provides a basis on which future work in Irish NLP is possible.[1]

This paper is organised as follows. Section 2 provides some background linguistic information on the Irish language along with a brief discussion of treebank development. In Section 3, we discuss the few existing resources and how we use these as a basis for our work. The methodology behind treebank creation is also then discussed with consideration to the various decisions that are made and the motivation behind them. Section 4 gives a summary of our first inter-annotation agreement evaluation. We present and discuss preliminary parsing experiments in Section 5 and finally, in Section 6, we present some options available to us for further development of our treebank.

*Part of this work was done while the author was a member of NCLT/CNGL, Dublin City University.

[1]One of the long term objectives of this project is to evaluate the application of a successful Irish parser to English-Irish Machine Translation.

2. Background

2.1. The Irish Language

The Irish language is a Celtic language of the Indo-European language family. Irish shares distinct features with other Celtic languages (Stenson, 1981) such as Verb-Subject-Object (VSO) word order and rich morphology.
The Irish language presents interesting computational linguistic challenges. The morphological features of Irish affect many parts of speech. For example, synthetic verb forms incorporate a subject through inflection (e.g. *cuir* 'put', *cuir**im*** 'I put'). Grammaticalised nominal inflections often appear in the form of initial mutations such as lenition (e.g. *tuairim* 'opinion', *i mo **th**uairim* 'in my opinion') or eclipsis (e.g. *bord* 'table', *ar an **m**bord* 'on the table'). In addition, final mutations are realised (e.g. to indicate genitive case) through slenderisation and broadening (e.g. *leabhar* 'book', *teideal an leabhai**r*** 'the title of the book' and *dochtúir* 'doctor', *ainm an dochtú**ra*** 'the doctor's name'). Most simple prepositions can also be inflected for person and number. These are discussed in more detail in Section 3.4.
As noted by Uí Dhonnchadha (2009), there are still several issues in Irish theoretical syntax that have yet to be resolved. Some of these issues relate to the status of VP (verb phrase) in Irish, arising from insufficient research into VSO languages in general. Other theoretical issues such as the nature of periphrastic aspectual structures in Irish are unclear.

2.2. Treebanks

Many data-driven NLP applications rely heavily on parsed corpora (treebanks) as training data for development and as

language	size	source
Slovene	2,000	Džeroski et al. (2006)
Danish	5,540	Kromann (2003)
Finnish	7,076	Haverinen et al. (2011)
Turkish	5,635	Eryiğit et al. (2008)
Czech	90,000	Hajič (2005)

Table 1: Dependency treebanks, size in number of sentences

test data for evaluation. There has been much interest in the application of dependency grammar to the development of treebanks for use in data-driven dependency parsing: e.g. Turkish (Oflazer et al., 2003), Czech (Hajič, 1998), Danish (Kromann, 2003), Slovene (Džeroski et al., 2006) and Finnish (Haverinen et al., 2010). Table 1 provides an overview of the size of these treebanks, with the dates indicating the year these figures were published. Once a treebank of significant size is available for Irish, it will be possible to induce statistical dependency parsing models using transition-based approaches, e.g. MaltParser (Nivre et al., 2006) and graph-based approaches, e.g. MSTParser (McDonald et al., 2005). Parsing experiments on thirteen treebanks have shown that reasonably accurate parsing models can be learned from training set sizes as small as 1500 sentences (Nivre, 2008).

3. Methodology — where to start?

3.1. Develop upon existing NLP tools

In recent years, some progress has been made in the collection and development of linguistic resources for Irish. A 30 million word corpus of Modern Irish text (NCII)[2] was developed in 2004 for *Foras na Gaeilge*.[3] In addition, corpus annotation tools, namely a morphological analyser (Uí Dhonnchadha et al., 2003), a part-of-speech (POS) tagger (Uí Dhonnchadha and van Genabith, 2006) and a chunker (Uí Dhonnchadha and van Genabith, 2010) have been developed.

A 3,000-sentence gold standard POS-annotated corpus was produced as a by-product of this work. These sentences were randomly selected from the NCII corpus and consist of text from books, newspaper, websites and other media, which form a solid representation of real Modern Irish language data.

Uí Dhonnchadha (2009) also made available a small corpus of 225 chunked Irish sentences. These sentences represented a Test Suite for a shallow parser which is based on Constraint Grammar Dependency Mapping Rules (Karlsson, 1995) and implemented using Xerox Finite State Tools.[4] The shallow nature of this chunking parser means that the dependency analysis does not extend to cover coordination, prepositional attachment, long-distance dependencies or clausal attachment. However, these 225 invented

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

Figure 1: Example of chunked output for *D'fhan siad ansin le fiche bliain* 'They stayed there for twenty years'

sentences cover the main syntactic phenomena of Irish and provided a valuable starting point for this treebank development.

Our first step involved reviewing the dependency analysis defined by Uí Dhonnchadha (2009) and adapting it to fit our chosen dependency scheme, which is discussed in Section 3.3. We modified and extended the parses in this small corpus to produce deep, full syntactic parses. Figure 1 is example output from the chunker for *D'fhan siad ansin le fiche bliain* 'They stayed there for twenty years'. The sentence is parsed into 4 chunks; V (Verb), NP (Noun Phrase), AD (Adverb) and PP (Prepositional Phrase). This output also indicates the kind of data available to us in the POS-tagged corpus that we use in our treebank, i.e. surface form, lemma, coarse-grained POS-tag and fine-grained POS-tag. Figure 2 presents our extended parse analysis for the same sentence, showing in particular adverbial and prepositional attachment.

We then added these fully parsed sentences to the 3,000 gold standard POS-tagged corpus and subsequently randomised the data so that the relatively easy 225 sentences would not form a misrepresentative chunk of Irish data. The first 300 (manually parsed) sentences in this new text collection form the basis for the discussions and experiments outlined in this paper. It can be noted that these sentences average 23 tokens in length and that 18 of them are from the original 225 invented sentences.

As an important factor for any treebank development, an annotation guide is being developed concurrently. This document will continue to evolve as we encounter new linguistic phenomena and analyse results of inter-annotation assessments (this will be discussed further in Section 4). It serves as a reference point to ensure consistency during annotation and will also facilitate the involvement of additional annotators in the future.

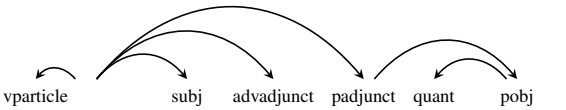
3.2. Choice of syntactic representation

We have chosen to build a dependency rather than constituency-based treebank for Irish. An examination of the existing literature in Irish theoretical syntax (including, but not limited to McCloskey (1979); Stenson (1981) and Carnie (2005), shows a lack of sufficient agreement on the syntactic representation of some fundamental linguistic phenomena. Their syntactic analyses differ even at a basic

[2]New Corpus for Ireland - Irish. See http://corpas.focloir.ie

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

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



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

Figure 2: The fully parsed sentence of Figure 1

level i.e. a flat VSO structure versus an underlying SVO structure. Discussions on topics such as deep structure or movement (represented by traces), for example, would have been highly relevant for a constituency-based treebank, as would the question of the existence of a VP constituent. It is more feasible therefore, in our treebank development, to identify the functional relationships within sentences (dependencies) than to try to address all the unsolved complexities of Irish syntax. While this approach requires fewer theoretical assumptions, we are still left with a significantly challenging task. A few of these challenges are discussed in Section 3.4.

From a practical point of view, dependency representations, which focus on grammatical functions and roles of elements in language, are ideal tools for abstracting away from more structural constituency information. Parsers based on these representations play an important role in the development of applications such as Question-Answering, Machine Translation and Sentiment Analysis systems (see, for example, Hermjakob (2001), Quirk et al. (2005) and Johansson and Moschitti (2010)).

3.3. Choice of dependency labelling scheme

We have elected to base our dependency labelling scheme on that of Çetinoğlu et al. (2010). This scheme was inspired by the functional relations defined within Lexical Functional Grammar (Bresnan, 2001), a theory that incorporates c(onstituent) and f(unctional) structures. It is relatively language-independent due to the abstract nature of the f-structure component, which is the main motivation behind the LFG ParGram project (Butt et al., 2002).

Thus, although their scheme was designed to describe English sentences, its roots in LFG theory make it a good starting point for developing resources for a language such as Irish with syntactic structures that are significantly different to English. We draw on both our previous expertise in this domain and relevant Irish LFG research[5] to develop this labelling scheme, e.g. Asudeh (2002) - an analysis of Irish preverbal particles and adjunction; Attia (2008) - an analysis of copula constructions taking Irish as an example; Sulger (2009) - an analysis of Irish cleft constructions. The 32 dependency labels in our current tagset are presented in Table 2. Our tagset is more fine-grained than that of Çetinoğlu et al. (2010), which has 25 labels and less fine-grained than Stanford Typed dependencies (de Marneffe and Manning, 2008), which have 53 labels.

[5]Only a limited range of Irish linguistic phenomena has been covered to date.

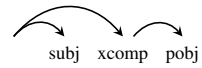
dependency label	**function**
top	root
punctuation	internal and final punctuation
subj	subject
obj	object
obl	oblique object
obl2	second oblique object
pobj	object of preposition
vnobj	object of verbal noun
det	determiner
det2	post or pre-determiner
dem	demonstrative pronoun
poss	possessive pronoun
aug	augment pronoun
quant	quantifier
coord	coordinate
relmod	relative modifier
relparticle	relative particle
advparticle	adverbial particle
vparticle	verb particle
vocparticle	vocative particle
adjunct	adjunct
adjadjunct	adjectival modifier
advadjunct	adverbial modifier
nadjunct	nominal modifier
padjunct	prepositional modifier
subadjunct	subordinate conjunction
toinfinitive	infinitive verb marker
app	noun in apposition
addr	addressee
focus	focus
xcomp	open complement
comp	closed complement

Table 2: Irish Treebank dependency tagset

3.4. Language specific choices

It should be noted that many of the common and well-known LFG-inspired analyses for English, familiar from the literature e.g. (Dalrymple, 2001; Bresnan, 2001), were just a starting point for our project and needed to be adapted considerably to Irish. With this, many interesting questions have arisen:

Prepositional attachment With Uí Dhonnchadha (2009)'s chunker, prepositional phrases are not attached to other constituents or phrases. While much work has already been done on prepositional attachment in other languages, Irish possesses some unusual prepositional behaviour. Irish has simple and compound prepositions. Most of the simple prepositions can inflect for person and number (prepositional pronouns/ pronominal prepositions), thus including a nominal element (often with a semantic role of experiencer). For example, comparing *leat* 'with you' and *leis an bhfear* 'with the man' it is clear that such inflection creates sparseness within the parser training data. A parser will therefore not have enough data to accurately learn the two patterns of prepositional phrases — those with and without an overt object. Data sparsity represents a major challenge for data-driven NLP in general but it



Tá sé ag iascaireacht
Is he at fishing
'He is fishing'

Figure 3: LFG-inspired dependency tree for Irish progressive aspectual phrase

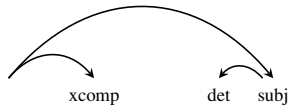
is a particular problem for morphologically rich languages (MRLs). There has been much interest in the parsing community in recent years in developing approaches to overcoming the data sparsity problem in MRL parsing (Tsarfaty et al., 2010).

Progressive aspectual phrases represent another interesting preposition function. As argued by Uí Dhonnachadha, these types of phrases, such as *tá sé ag iascaireacht* 'he is fishing' are constructed using the substantive verb *tá* 'is' as an auxiliary, along with a non-finite complement (a prepositional phrase (PP) consisting of a preposition *ag* 'at' and a verbal noun *iascaireacht* 'fishing'). The verbal nature of these types of PPs means that they cannot be labelled as adjuncts, which is often the case for prepositional attachment. Instead, we regard them as a predicates. Non-verbal predicates such as PPs can be labelled as open complements (XCOMPs) in LFG (Bresnan, 2001). While examples of PP predicates can be found in English, they are used in limited circumstances. In contrast, periphrastic constructions involving PP predicates occur frequently in Irish. We can see in Figure 3 how this would be represented in a dependency tree.

Irish copula In some languages the copula is regarded as a verb (copular verb). In Irish, however, there is a distinction between the substantive verb *bí* 'to be' and the copula *is*. Copula constructions present interesting questions when being defined by dependency relations. The order of elements is in general: copula, predicate (new/focussed information), and subject. The example in Figure 4 translates to English as 'You are the teacher'. However non-intuitive to an English speaker, our analysis identifies *tusa* 'you' as the predicate and *múinteoir* 'teacher' as the subject.

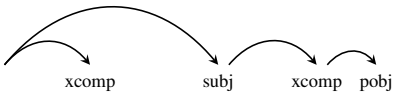
This role-labelling is explained by the fact that it answers the question "Who is the teacher?" (Christian-Brothers, 1988). The answer in Irish reads literally as 'The teacher is you'. It may be worth noting here that in some analyses (e.g. Carnie (1997)), the Irish copula in this construction is regarded as a complementizer particle which equates two noun phrases. Neither element, therefore, is labelled as the subject or predicate. We do not consider this unlabelled analysis, since, according to the dependency scheme we have adopted, all relations must be labelled.

Clefting is a commonly employed linguistic construction in Irish. The copula *is* is used to allow fronting of a word or clause followed by the rest of the sentence, which is in the form of a relative clause. E.g. *Is mise atá ag ithe* 'It is me who is eating'. In an LFG analysis, Sulger (2009) compares this cleft construction to the *identity* copula constructions in Irish. This analysis (Figure 5) fits in well with



Is tusa an múinteoir
COP you-EMPH the teacher
'You are the teacher'

Figure 4: LFG-inspired dependency tree for Irish copula identity construction



Is mise atá ag ithe
COP me -EMPH who-is at eating
'It is me who is eating'

Figure 5: LFG-inspired dependency tree for Irish cleft construction

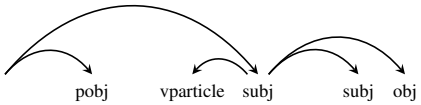
other analyses within our dependency scheme.

It is grammatically acceptable to sometimes drop the copula, whether the predicate is nominal, prepositional, adjectival or adverbial. When the structure is *(copula) predicate subject*, we are left with the question of what to mark as the root of the sentence, should the copula be dropped. In these cases, we promote the head of the predicate (XCOMP) to the root position of the sentence. Figure 6 shows our analysis for copula-drop in PP fronting.

4. Inter-annotator agreement

The first stage of our treebank development will involve parsing the 3,225 POS-tagged Irish sentences made available by Uí Dhonnchadha (2009). Currently there are 557 manually parsed sentences, that is, 225 invented sentences and 332 from the 3,000 NCII-based POS-tagged corpus. As a result of the randomisation step discussed in Section 3.1, the 225 invented sentences were dispersed throughout the new text collection and further manual parsing took place sequentially. To begin development, the first author was the main annotator. Discussions were held frequently with the third and fourth authors to establish a concrete annotation scheme for Irish. Once an annotation guide became available, the third author was subsequently trained.

We have calculated an inter-annotator agreement (IAA)



Sa pháirc a chonaic mé é
In-the field REL saw I him
'(It was) in the field that I saw him'

Figure 6: LFG-inspired dependency tree for copula-drop construction

Kappa (labels)	LAS	UAS
0.7902	74.37%	85.16%

Table 3: Inter-annotation agreement and accuracy results

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

Table 4: Landis and Koch's interpretation of Cohens's Kappa

measure on 50 sentences to assess consistency between both annotators. The Kappa coefficient of agreement is widely regarded as a standard for calculating IAA for corpus annotation tasks (Di Eugenio and Glass, 2004; Artstein and Poesio, 2008). This method of measurement has been adopted for assessing the objectivity of annotators in tasks such as discourse annotation (Poesio, 2004), word-sense annotation (Bruce and Wiebe, 1998) and POS annotation (Mieskes and Strube, 2006), for example. However, although agreement scores are reported in some of the dependency treebank literature (e.g. Uria et al. (2009); Gupta et al. (2010); Voutilainen and Purtonen (2011)), there does not appear to be a standard approach to measuring IAA for dependency parse annotation. This task differs somewhat to other annotation tasks in that the agreement of the (head, label) pair of a dependency annotation cannot be measured in the same way as agreement of single-value tags (e.g. POS-tags, discourse units, word-senses).

We have decided to divide the assessment into two measurements: (i) calculation of accuracy on (head, label) pair values through LAS/ UAS scores,[6] taking the primary annotator's set as gold-standard data and (ii) calculation of agreement on dependency tags (label values) through Cohen's Kappa coefficient measurement (Cohen, 1960). Our calculations do not take punctuation into account.

The Kappa statistic is defined as:

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

where $P(A)$ is the proportion of observed agreement among annotators, and $P(E)$ is the proportion of expected agreement. By correcting for $P(E)$, this measurement accounts for the fact that the annotators are expected to agree a proportion of times just by chance.

We use Landis and Koch (1977)'s metric shown in Table 4 for interpretation of our Kappa results, which are presented in Table 3.

[6]LAS - Labelled Attachment Score, UAS - Unlabelled Attachment Score.

5. Preliminary parsing experiments

For quality assurance, human input is fundamental in treebank development. However, there are steps that can be taken to semi-automate the annotation task so as to reduce the manual effort required. In order to therefore speed up the treebank development process, we will apply the following bootstrapping approach (similar to that of Judge et al. (2006) and more recently, Seraji et al. (2012)):

1. Create dependency analyses for a seed set of n sentences.
2. Train a baseline parsing model on the seed set.
3. Parse m sentences with the baseline model and manually correct the output.
4. Add the m automatically parsed and manually corrected trees to the training set and train a new model.
5. Parse another m sentences with the new model and manually correct the output.
6. Repeat steps 4 and 5 until the treebank is complete.

This kind of iterative approach to parsing a corpus allows us to take advantage of the presence of repetition in the data. The parser learns more at each iteration as a result of exposure to repetitive syntactic structures. All the parsed data is reviewed. However, we expect the parser will correctly annotate the frequently encountered and learned structures, leaving us with the manual correction of only the infrequent, previously unseen or difficult parses each time. Through the addition of newly parsed data to the training material at each iteration, the learning process becomes quicker.

We will employ MaltParser (Nivre et al., 2006) as our bootstrapping parser.[7] MaltParser is a multilingual transition-based parsing system which provides several efficient deterministic parsing algorithms capable of producing a dependency tree in linear or quadratic time. We will employ the *stacklazy* algorithm which can directly handle non-projective structures.[8] For our preliminary experiments, we test a variety of feature models which make use of various combinations of the following information extracted from the 3,000-sentence POS-tagged corpus: form, lemma, fine-grained and coarse-grained POS tags. We train on a set of 300 manually annotated sentences using 10-fold cross-validation. The results are shown in Table 5.

The size of our seed set means that the differences between the various models are not statistically significant. Nevertheless, we choose the best-performing model as our baseline model in the bootstrapping process. This model uses information from the word form, lemma and both POS tags.

6. Future Work

Treebank building is a time-consuming effort, even with semi-automated processes like our bootstrapping approach.

[7]Since this paper focuses on the linguistic choices made during treebank development, we conduct our experiments with a single parser instead of trying to achieve the best parser performance by evaluating different parsers.

[8]Initial analyses show that the Irish data contains some non-projective structures.

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

Table 5: Preliminary parsing results with MaltParser

This cost arises from the requirement for extensive manual input. However, in step 3 of our process described in Section 5, there is scope to reduce the time-cost of manual correction through techniques such as Active Learning (AL). With Active Learning, problematic parses are identified and prioritised for manual correction in order to speed up improvement of the parser's performance. There is currently much research in this area and we plan to consider some of these approaches in the future.

Recent work by Mirroshandel and Nasr (2011) demonstrates how relevant substrings within identified problematic sentences can be isolated for correction. Other work focuses on developing methods for automatically detecting errors in dependency parses (e.g. Dickinson (2010); Dickinson and Smith (2011)). In fact, the latter approach is noted as being particularly beneficial to low-density languages.

Bootstrapping lesser-resourced languages through the use of parallel texts is also possible. This involves exploiting tools of the more highly-resourced language of the language pair e.g. Hwa et al. (2005); Wróblewska and Frank (2009). This approach may be a possibility for our treebank development as there is a large number of English-Irish parallel official documents available from both Irish and European Parliament proceedings. In addition, a parallel corpus developed by Scannell (2005) would also prove a valuable resource if we were to consider this approach.

7. Acknowledgements

This work is supported by Science Foundation Ireland (Grant No 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. Özlem Çetinoğlu is partly funded by the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG) via project D2 of SFB 732 "Incremental Specification in Context". The authors would like to thank the reviewers for their insightful comments and suggestions.

8. References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Ash Asudeh. 2002. The syntax of preverbal particles and adjunction in Irish. In *Proceedings of the LFG '02 Conference*.

Mohammad Attia. 2008. A unified analysis of copula constructions in LFG. In *Proceedings of the LFG '08 Conference*.

Joan Bresnan. 2001. *Lexical Functional Syntax*. Oxford: Blackwell.

Rebecca Bruce and Janyce Wiebe. 1998. Word sense distinguishability and inter-coder agreement. In *Proceedings of 3rd Empirical Methods in Natural Language Processing (EMNLP-98)*.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*.

Andrew Carnie. 1997. Two types of non-verbal predication in Modern Irish. *Canadian Journal of Linguistics*.

Andrew Carnie. 2005. Flat Structure, Phrasal Variability and VSO. *Journal of Celtic Linguistics*.

Özlem Çetinoğlu, Jennifer Foster, Joakim Nivre, Deirdre Hogan, Aoife Cahill, and Josef van Genabith. 2010. LFG without c-structures. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT9)*.

Christian-Brothers. 1988. *New Irish Grammar*. Dublin: C J Fallon.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46. Durham.

Mary Dalrymple. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, New York.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Workshop on Crossframework and Cross-domain Parser Evaluation (COLING2008)*.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101, March.

Markus Dickinson and Amber Smith. 2011. Detecting dependency parse errors with minimal resources. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*.

Markus Dickinson. 2010. Detecting errors in automatically-parsed dependency relations. In *Proceedings of the 48th Annual Meeting on Association for Computational Linguistics (ACL2010)*.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of turkish. *Computational Linguistics*, 34(3):357–389, September.

Mridul Gupta, Vineet Yadav, Samar Husain, and Dipti Misra Sharma. 2010. Partial Parsing as a Method to Expedite Dependency Annotation of a Hindi Treebank. In *Proceedings of the 7th International Confer-*

ence on Language Resources and Evaluation (LREC 2010).

Jan Hajič. 2005. Complex corpus annotation: The Prague dependency treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague dependency treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking Finnish. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.

Katri Haverinen, Filip Ginter, Veronika Laippala, Samuel Kohonen, Timo Viljanen, Jenna Nyblom, and Tapio Salakoski. 2011. A dependency-based analysis of treebank annotation errors. In *Proceedings of International Conference on Dependency Linguistics (Depling 2011)*.

Ulf Hermjakob. 2001. Parsing and question classification for question answering. In *Proceedings of the ACL Workshop on Open-Domain Question Answering*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*.

Richard Johansson and Alessandro Moschitti. 2010. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the 14th Conference on Computational Natural Language Learning*.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *In Proceedings of the 21st International Conference on Computational Linguitics and 44th Annual Meeting of the ACL*.

Fred Karlsson. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, volume 4. Berlin - New York: Mouton de Gruyter.

Matthias Kromann. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings from the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. In *Biometrics*, volume 33, pages 159–174. International Biometric Society.

James McCloskey. 1979. *Transformational syntax and model theoretic semantics: a case in Modern Irish*. Dordrecht: Reidel.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Association for Computational Linguistics*.

Margot Mieskes and Michael Strube. 2006. Part-of-speech tagging of transcribed speech. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Seyed Abolghasem Mirroshandel and Alexis Nasr. 2011. Active learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34:513–553, December.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In Anne Abeille, editor, *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, pages 72–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*.

Kevin Scannell. 2005. Applications of parallel corpora to the development of monolingual language technologies.

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. 2012. Bootstrapping a Persian Dependency Treebank. *Linguistic Issues in Language Technology*, 7.

Nancy Stenson. 1981. *Studies in Irish Syntax*. Gunter Narr Verlag Tübingen.

Sebastian Sulger. 2009. Irish clefting and information-structure. In *Proceedings of the LFG '09 Conference*.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, June.

Elaine Uí Dhonnchadha and Josef van Genabith. 2006. A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Elaine Uí Dhonnchadha and Josef van Genabith. 2010. Partial dependency parsing for Irish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Elaine Uí Dhonnchadha, Caoilfhionn Nic Pháidín, and Josef van Genabith. 2003. Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, 18:173–193. 10.1007/s10590-004-2480-9.

Elaine Uí Dhonnchadha. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Trans-*

ducers and Constraint Grammar. Ph.D. thesis, Dublin City University.

Larraitz Uria, Ainara Estarrona, Izaskun Aldezabal, Maria Jesús Aranzabe, Arantza Díaz De Ilarraza, and Mikel Iruskieta. 2009. Evaluation of the syntactic annotation in EPEC, the reference corpus for the processing of Basque. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, (CICLing '09), pages 72–85, Berlin, Heidelberg. Springer-Verlag.

Atro Voutilainen and Tanja Purtonen. 2011. A double-blind experiment on interannotator agreement: the case of dependency syntax and Finnish. In Gunta Nešpore Bolette Sandford Pedersen, Inguna Skadiņa. Bolette Sandford Pedersen, Gunta Nešpore, and Inguna Skadiņa, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*.

Alina Wróblewska and Anette Frank. 2009. Cross-lingual Projection of LFG F-Structures: Building an F-Structure Bank for Polish. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*.