# Working with a small dataset - semi-supervised dependency parsing for Irish

**Teresa Lynn**[1,2], **Jennifer Foster**[1], **Mark Dras**[2] and **Josef van Genabith**[1]
[1]NCLT/CNGL, Dublin City University, Ireland
[2]Department of Computing, Macquarie University, Sydney, Australia
[1]{tlynn,jfoster,josef}@computing.dcu.ie
[2]{teresa.lynn,mark.dras}@mq.edu.au,

## Abstract

We present a number of semi-supervised parsing experiments on the Irish language carried out using a small seed set of manually parsed trees and a larger, yet still relatively small, set of unlabelled sentences. We take two popular dependency parsers – one graph-based and one transition-based – and compare results for both. Results show that using semi-supervised learning in the form of self-training and co-training yields only very modest improvements in parsing accuracy. We also try to use morphological information in a targeted way and fail to see any improvements.

## 1 Introduction

Developing a data-driven statistical parser relies on the availability of a parsed corpus for the language in question. In the case of Irish, the only parsed corpus available to date is a dependency treebank, which is currently under development and still relatively small, with only 803 gold-annotated trees (Lynn et al., 2012a). As treebank development is a labour- and time-intensive process, in this study we evaluate various approaches to bootstrapping a statistical parser with a set of unlabelled sentences to ascertain how accurate parsing output can be at this time. We carry out a number of different semi-supervised bootstrapping experiments using self-training, co-training and sample-selection-based co-training. Our studies differ from previous similar experiments as our data is taken from a work-in-progress treebank. Thus, aside from the current small treebank which is used for training the initial seed model and for testing, there is no additional gold-labelled data available to us to directly compare supervised and semi-supervised approaches using training sets of comparable sizes.

In the last decade, data-driven dependency parsing has come to fore, with two main approaches dominating – transition-based and graph-based. In classic transition-based dependency parsing, the training phase consists of learning the correct parser action to take given the input string and the parse history, and the parsing phase consists of the greedy application of parser actions as dictated by the learned model. In contrast, graph-based dependency parsing involves the non-deterministic construction of a parse tree by predicting the maximum-spanning-tree in the digraph for the input sentence. In our study, we employ Malt (Nivre et al., 2006), a transition-based dependency parsing system, and Mate (Bohnet, 2010), a graph-based parser.

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

The semi-supervised parsing experiments do not explicitly take into account the morphosyntactic properties of the Irish language. In order to examine the effect of this type of information during parsing, we carry out some orthogonal experiments where we



*Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11,
Seattle, Washington, USA, 18 October 2013. 

reduce word forms to lemmas and introduce morphological features in certain cases. These changes do not bring about an increase in parsing accuracy.

The paper is organised as follows. Section 2 is an overview of Irish morphology. In Section 3 our previous work carried out on the development of an Irish dependency treebank is discussed followed in Section 4 by a description of some of our prior parsing results. Section 5 describes the self-training, co-training and sample-selection-based co-training experiments, Section 6 presents the preliminary parsing experiments involving morphological features, and, finally, Section 7 discusses our future work.

## 2 Irish as a morphologically rich language

Irish is a Celtic language of the Indo-European language family. It has a VSO word order and is rich in morphology. The following provides an overview of the type of morphology present in the Irish language. It is not a comprehensive summary as the rules governing morphological changes are too extensive and at times too complex to document here.

Inflection in Irish mainly occurs through suffixation, but initial mutation through lenition and eclipsis is also common (Christian-Brothers, 1988). A prominent feature of Irish (also of Scottish and Manx), which influences inflection, is the existence of two sets of consonants, referred to as 'broad' and 'slender' consonants (Ó Siadhail, 1989). Consonants can be slenderised by accompanying the consonant with a slender vowel, either *e* or *i*. Broadening occurs through the use of broad vowels; *a*, *o* or *u*. For example, *buail* 'to hit' becomes *ag bu**a**l**a**dh* 'hitting' in the verbal noun form. In general, there needs to be vowel harmony (slender or broad) between stem endings and the initial vowel in a suffix.

A process known as syncopation also occurs when words with more than one syllable have a vowel-initial suffix added. For example *imir* 'to play' inflects as ***imr**ím* 'I play'.

**Nouns** While Old Irish employed several grammatical cases, Modern Irish uses only three: Nominative, Genitive and Vocative. The nominative form is sometimes regarded as the 'common form' as it is now also used to account for accusative and dative forms. Nouns in Irish are divided into five classes, or declensions, depending on the manner in which the genitive case is formed. In addition, there are two grammatical genders in Irish - masculine and feminine. Case, declension and gender are expressed through noun inflection. For example, *páipéar* 'paper' is a masculine noun in the first declension. Both lenition and slenderisation are used to form the genitive singular form: *p**h**áipé**i**r*. In addition, possessive adjectives cause noun inflection through lenition, eclipsis and prefixation. For example, *teach* 'house', *mo t**h**each* 'my house', *ár **d**teach* 'our house'; *ainm* 'name', *a **h**ainm* 'her name'.

**Verbs** Verbs can incorporate their subject, inflecting for person and number through suffixation. Such forms are referred to as synthetic verb forms. In addition, verb tense is often indicated through various combinations of initial mutation, syncopation and suffixation. For example, *scríobh* 'write' can inflect as *scríobh**aim*** 'I write'. The past tense of the verb *tug* 'give' is *t**h**ug* 'gave'. Lenition occurs after the negative particle *ní*. For example, *tugaim* 'I give'; *ní t**h**ugaim* 'I do not give'; *níor t**h**ug mé* 'I did not give'. Eclipsis occurs following clitics such as interrogative particles (*an, nach*); complementisers (*go, nach*); and relativisers (*a, nach*) (Stenson, 1981). For example, *an **d**tugann sé?* 'does he give?'; *nach **d**tugann sé?* 'does he not give?'.

**Adjectives** In general, adjectives follow nouns and agree in number, gender and case. Depending on the noun they modify, adjectives can also inflect. Christian-Brothers (1988) note eight declensions of adjectives. They can decline for genitive singular masculine, genitive singular feminine and nominative plural. For example, *bacach* 'lame' inflects as *baca**igh*** (Gen.Sg.Masc), *baca**í*** (Gen.Fem.Sg) and *bacach**a*** (Nom.PL). Comparative adjectives are also formed through inflection. For example, *láidir* 'strong', *níos láid**re*** 'stronger'; *déanach* 'late', *is déana**í*** 'latest'.

**Prepositions** Irish has simple and compound prepositions. Most of the simple prepositions can inflect for person and number (known as prepositional pronouns or pronominal prepositions), thus including a nominal element. For example, compare *bhí sé ag labhairt **le** fear* 'he was speaking **with** a man' with *bhí sé ag labhairt **leis*** 'he was speaking **with him**'. These forms are used quite fre-

quently, not only with regular prepositional attachment where pronominal prepositions operate as arguments of verbs or modifiers of nouns and verbs, but also in idiomatic use where they express emotions and states, e.g. *tá brón orm* (lit. 'be-worry-on me') 'I am worried' or *tá súil agam* (lit. 'be-expectation-with me') 'I hope'. Noted by Greene (1966) as a noun-centered language, nouns are often used to convey the meaning that verbs often would. Pronominal prepositions are often used in these types of structures. For example, *bhain mé geit* ***aisti*** (lit. extracted-I-shock-**from her**) 'I frightened her'; *bhain mé mo chóta* ***díom*** (lit. extracted-I-coat-**from me**) 'I took off my coat'; *bhain mé úsáid* ***as*** (lit. extracted-I-use-**from it**) 'I used it'; *bhain mé triail* ***astu*** (lit. extracted-I-attempt-**from them**)'I tried them'.

**Derivational morphology** There are also some instances of derivational morphology in Irish. Uí Dhonnchadha (2009) notes that all verb stems and agentive nouns can inflect to become verbal nouns. Verbal adjectives are also derived from verb stems through suffixation. For example, the verb *dún* 'close' undergoes suffixation to become *dúnadh* 'closing' (verbal noun) and *dúnta* 'closed' (verbal adjective). An emphatic suffix *-sa/-se* (both broad and slender form) can attach to nouns or pronouns. It can also be attached to any verb that has been inflected for person and number and also to pronominal prepositions. For example *mo thuairim* 'my opinion' → *mo thuairim**se*** '**my** opinion; *tú* 'you'(sg) → *tu**sa*** '**you**'; *cloisim* 'I hear' → *cloisim**se*** '**I** hear'; *liom* 'with me' → *liom**sa*** 'with **me**'. In addition, the diminutive suffix *-ín* can attach to all nouns to form a derived diminutive form. The rules of slenderisation apply here also. For example, *buachaill* 'boy' becomes *buachai**llín*** 'little boy', and *tamall* 'while' becomes *tamai**llín*** 'short while'.

## 3 The Irish Dependency Treebank

Irish is the official language of Ireland, yet English is the primary language for everyday use. Irish is therefore considered an EU minority language and is lacking in linguistic resources that can be used to develop NLP applications (Judge et al., 2012).

Recently, in efforts to address this issue, we have begun work on the development of a dependency treebank for Irish (Lynn et al., 2012a). The treebank has been built upon a gold standard 3,000 sentence POS-tagged corpus[1] developed by Uí Dhonnchadha (2009). Our labelling scheme is based on an 'LFG-inspired' dependency scheme developed for English by Çetinoğlu et al. (2010). This scheme was adopted with the aim of identifying functional roles while at the same time circumventing outstanding, unresolved issues in Irish theoretical syntax.[2] The Irish labelling scheme has 47 dependency labels in the label set. The treebank is in the CoNLL format with the following fields: `ID`, `FORM`, `LEMMA`, `CPOSTAG`, `POSTAG`, `HEAD` and `DEPREL`. The coarse-grained part of speech of a word is marked by the label `CPOSTAG`, and `POSTAG` marks the fine-grained part of speech for that word. For example, prepositions are tagged with the `CPOSTAG Prep` and one of the following `POSTAG`s: `Simple`: *ar* 'on', `Compound`: *i_ndiaidh* 'after', `Possessive`: *ina* 'in its', `Article`: *sa* 'in the'.

At an earlier stage of the treebank's development, we carried out on an inter-annotator agreement (IAA) study. The study involved four stages. (i) The first experiment (IAA-1) involved the assessment of annotator agreement following the introduction of a second annotator. The results reported a Kappa score of 0.79, LAS of 74.4% and UAS of 85.2% (Lynn et al., 2012a). (ii) We then held three workshops that involved thorough analysis of the output of IAA-1, highlighting disagreements between annotators, gaps in the annotation guide, shortcomings of the labelling scheme and linguistic issues not yet addressed. (iii) The annotation guide, labelling scheme and treebank were updated accordingly, addressing the highlighted issues. (iv) Finally, a second inter-annotator agreement experiment (IAA-2) was carried out presenting a Kappa score of 0.85, LAS of 79.2% and UAS of 87.8% (Lynn et al., 2012b).

We found that the IAA study was valuable in the development of the treebank, as it resulted in im-

[1] A tagged, randomised subset of the NCII, (New Corpus for Ireland - Irish http://corpas.focloir.ie/), comprised of text from books, news data, websites, periodicals, official and government documents.

[2] For example there are disagreements over the existence of a VP in Irish and whether the language has a VSO or an underlying SVO structure.

provement of the quality of the labelling scheme, the annotation guide and the linguistic analysis of the Irish language. Our updated labelling scheme is now hierarchical, allowing for a choice between working with fine-grained or coarse-grained labels. The scheme has now been finalised. A full list of the labels can be found in Lynn et al. (2012b). The treebank currently contains 803 gold-standard trees.

## 4 Preliminary Parsing Experiments

In our previous work (Lynn et al., 2012a), we carried out some preliminary parsing experiments with MaltParser and 10-fold cross-validation using 300 gold-standard trees. We started out with the feature template used by Çetinoğlu et al. (2010) and examined the effect of omitting `LEMMA`, `WORDFORM`, `POSTAG` and `CPOSTAG` features and combinations of these, concluding that it was best to include all four types of information. Our final LAS and UAS scores were 63.3% and 73.1% respectively. Following the changes we made to the labelling scheme as a result of the second IAA study (described above), we re-ran the same parsing experiments on the newly updated seed set of 300 sentences - the LAS increased to 66.5% and the UAS to 76.3% (Lynn et al., 2012b).

In order to speed up the treebank creation, we also applied an active learning approach to bootstrapping the annotation process. This work is also reported in Lynn et al. (2012b). The process involved training a MaltParser model on a small subset of the treebank data, and iteratively, parsing a new set of sentences, selecting a 50-sentence subset to hand-correct, and adding these new gold sentences to the training set. We compared a passive setup, in which the parses that were selected for correction were chosen at random, to an active setup, in which the parses that were selected for correction were chosen based on the level of disagreement between two parsers (Malt and Mate). The active approach to annotation resulted in superior parsing results to the passive approach (67.2% versus 68.1% LAS) but the difference was not statistically significant.

## 5 Semi-Supervised Parsing Experiments

In order to alleviate data sparsity issues brought about by our lack of training material, we experiment with automatically expanding our training set using well known semi-supervised techniques.

### 5.1 Self-Training

#### 5.1.1 Related Work

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

#### 5.1.2 Experimental Setup

The labelled data available to us for this experiment comprises the 803 gold standard trees referred to in Section 3. This small treebank includes the 150-tree development set and 150-tree test set used in experiments by Lynn et al. (2012b). We use the same development and test sets for this study. As for the remaining 503 trees, we remove any trees that have more than 200 tokens. The motivation for this is two-fold: (i) we had difficulties training Mate parser with long sentences due to memory resource issues, and (ii) in keeping with the findings of Lynn et al. (2012b), the large trees were sentences from legislative text that were difficult to analyse for automatic parsers and human annotators. This leaves us with 500 gold-standard trees as our seed training data set.

For our unlabelled data, we take the next 1945 sentences from the gold standard 3,000-sentence

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

Figure 1: Self-training algorithm

POS-tagged corpus referred to in Section 3. When we remove sentences with more than 200 tokens, we are left with 1938 sentences in our unlabelled set.

The main algorithm for self-training is given in Figure 1. We carry out two separate experiments using this algorithm. In the first experiment we use Malt. In the second experiment, we substitute Mate for Malt.[3]

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

#### 5.1.3 Results

The results of our self-training experiments are presented in Figure 2. The best Malt model was trained on 2115 trees, at the 5th iteration (70.2% LAS). UAS scores did not increase over the baseline (79.1%). The improvement in LAS over the baseline is not statistically significant. The best Mate model was trained on 1792 trees, at the 4th iteration (71.2% LAS, 79.2% UAS). The improvement over the baseline is not statistically significant.

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

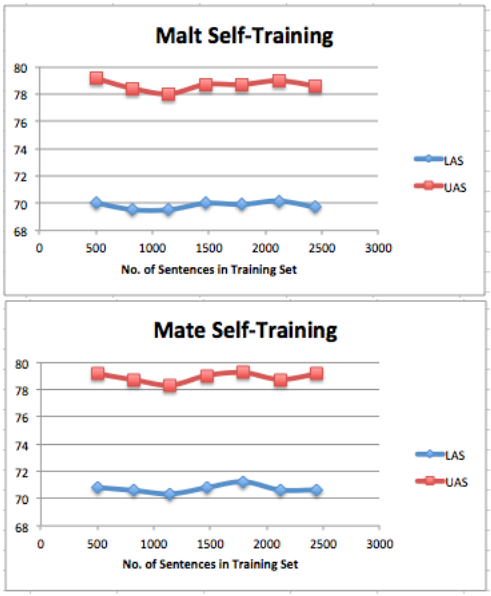


Figure 2: Self-Training Results on the Development Set

### 5.2 Co-Training

#### 5.2.1 Related Work

Co-training involves training a system on the output of a different system. Co-training has found more success in parsing than self-training, and it is not difficult to see why this might be the case as it can be viewed as a method for combining the benefits of individual parsing systems. Steedman et al. (2003) directly compare co-training and self-training and find that co-training outperforms self-training. Sagae and Tsujii (2007) successfully employ co-training in the domain adaption track of the CoNLL 2007 shared task on dependency parsing.

#### 5.2.2 Experimental Setup

In this and all subsequent experiments, we use both the same training data and unlabelled data that we refer to in Section 5.1.2.

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

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

Figure 3: Co-training algorithm

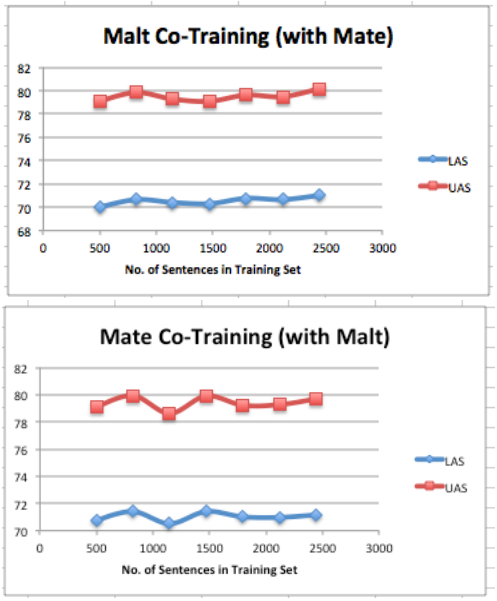


Figure 4: Co-Training Results on the Development Set

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

#### 5.2.3 Results

The results of our co-training experiment are presented in Figure 4. The best Malt model was trained on 2438 trees, at the final iteration (71.0% LAS and 79.8% UAS). The improvement in UAS over the baseline is statistically significant. Mate's best model was trained on 823 trees on the second iteration (71.4% LAS and 79.9% UAS). The improvement over the baseline is not statistically significant.

### 5.3 Sample-Selection-Based Co-Training

#### 5.3.1 Related Work

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

In Active Learning research, the Query By Committee selection method (Seung et al., 1992) is used to choose items for annotation – if a committee of two or more systems disagrees on an item, this is evidence that the item needs to be prioritised for manual correction (see for example Lynn et al. (2012b)). Steedman et al. (2003) discuss a sample selection approach based on differences between parsers – if parser A and parser B disagree on an analysis, parser A can be improved by being retrained on parser B's analysis, and vice versa. In contrast, Ravi et al. (2008) show that parser *agreement* is a strong in-

dicator of parse quality, and in parser domain adaptation, Sagae and Tsujii (2007) and Le Roux et al. (2012) use agreement between parsers to choose which automatically parsed target domain items to add to the training set.

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

#### 5.3.2 Agreement-Based Co-Training

**Experimental Setup** The main algorithm for agreement-based co-training is given in Figure 5. Again, Malt and Mate are used. However, this algorithm differs from the co-training algorithm in Figure 3 in that rather than adding the full set of 323 newly parsed trees ($P_A^i$ and $P_B^i$) to the training set at each iteration, selected subsets of these trees ($P_A^i{\prime}$ and $P_B^i{\prime}$) are added instead. To define these subsets, we identify the trees that have 85% or higher **agreement** between the two parser output sets. As a result, the number of trees in the subsets differ at each iteration. For iteration 1, 89 trees reach the agreement threshold; iteration 2, 93 trees; iteration 3, 117 trees; iteration 4, 122 trees; iteration 5, 131 trees; iteration 6, 114 trees. The number of trees in the training sets is much smaller compared with those in the experiments of Section 5.2.

**Results** The results for agreement-based co-training are presented in Figure 6. Malt's best model was trained on 1166 trees at the final iteration (71.0% LAS and 79.8% UAS). Mate's best model was trained on 1052 trees at the 5th iteration (71.5% LAS and 79.7% UAS). Neither result represents a statistically significant improvement over the baseline.

#### 5.3.3 Disagreement-based Co-Training

**Experimental Setup** This experiment uses the same sample selection algorithm we used for agreement-based co-training (Figure 5). For this experiment, however, the way in which the subsets of trees ($P_A^i{\prime}$ and $P_B^i{\prime}$) are selected differs. This time we choose the trees that have 70% or higher **disagreement** between the two parser output sets.

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

Figure 5: Sample selection Co-training algorithm

Again, the number of trees in the subsets differ at each iteration. For iteration 1, 91 trees reach the disagreement threshold; iteration 2, 93 trees; iteration 3, 73 trees; iteration 4, 74 trees; iteration 5, 68 trees; iteration 6, 71 trees.

**Results** The results for our disagreement-based co-training experiment are shown in Figure 7. The best Malt model was trained with 831 trees at the 4th iteration (70.8% LAS and 79.8% UAS). Mate's best models were trained on (i) 684 trees on the 2nd iteration (71.0% LAS) and (ii) 899 trees on the 5th iteration (79.4% UAS). Neither improvement over the baseline is statistically significant.

#### 5.3.4 Non-Iterative Agreement-based Co-Training

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

**Experimental Setup** Again, we also follow the algorithm for agreement-based co-training as presented in Figure 5. However, two different ap-

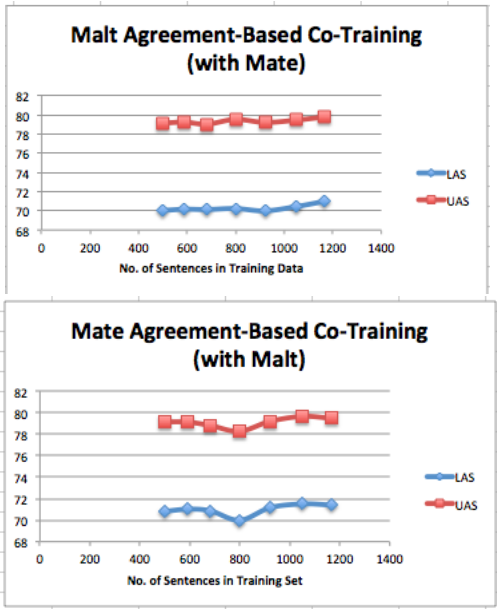


Figure 6: Agreement-based Co-Training Results on the Development Set

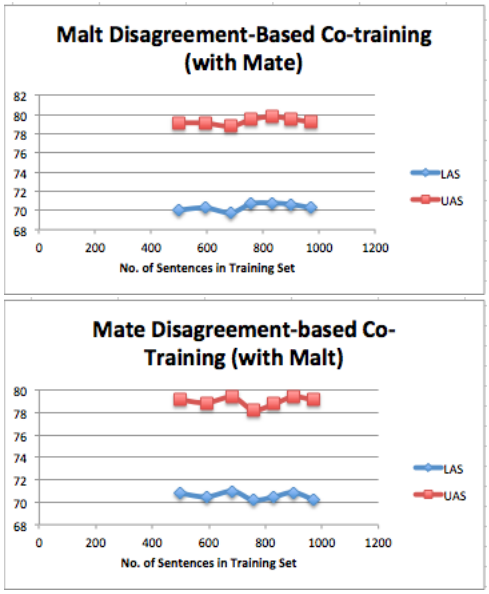


Figure 7: Disagreement-based Co-Training Results on the Development Set

proaches are taken this time, involving only one iteration in each. For the first experiment (ACT1a), the subsets of trees ($P_A^i{\prime}$ and $P_B^i{\prime}$) that are added to the training data are chosen based on an agreement threshold of 85% between parsers, and are taken from the *full* set of unlabelled data (where $U^i = U$), comprising 1938 trees. In this instance, the subset consists of 603 trees, making a final training set of 1103 trees.

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

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

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

### 5.4 Analysis

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

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

I saw yesterday/ the girl who saw me yesterday'.[4] (iii) There is no passive verb form in Irish. The autonomous form is most closely linked with passive use and is used when the agent is not known or mentioned. A 'hidden' or understood subject is incorporated into the verbform. *Casadh eochair i nglas* 'a key was turned in a lock' (lit. somebody turned a key in a lock). In this sentence, *eochair* 'key' is the object.

- For both parsers, there is some confusion between the labelling of `obl` and `padjunct`, both of which mark the attachment between verbs and prepositions. Overall, Malt's confusion decreases over the 6 iterations of self-training, but Mate begins to incorrectly choose `padjunct` over `obl` instead. Mixed results are obtained using the various variants of co-training.
- Mate handles coordination better than Malt.[5] It is not surprising then that co-training Malt using Mate parses improves Malt's coordination handling whereas the opposite is the case when co-training Mate on Malt parses, demonstrating that co-training can both eliminate and introduce errors.
- Other examples of how Mate helps Malt during co-training is in the distinction between `top` and `comp` relations, between `vparticle` and `relparticle`, and in the analysis of `xcomps`.
- Distinguishing between relative and cleft particles is a frequent error for Mate, and therefore Malt also begins to make this kind of error when co-trained using Mate. Mate improves using sample-selection-based co-training with Malt.
- The sample-selection-based co-training variants show broadly similar trends to the basic co-training.

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

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

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

Table 1: Results for best performing models

### 5.5 Test Set Results

The best performing parsing model for Malt on the development set is in the final iteration of the basic co-training approach in Section 5.2. The best performing parsing model for Mate on the development set is the non-iterative 85% threshold agreement-based co-training approach described in Section 5.3.4. The test set results for these optimal development set configurations are also shown in Table 1. The baseline model for Malt obtains a LAS of 70.2%, the final co-training iteration a LAS of 70.8%. The baseline model for Mate obtains a LAS of 71.9%, and the non-iterative 85% agreement-based co-trained model obtains a LAS of 73.1%.

## 6 Parsing Experiments Using Morphological Features

As well as the size of the dataset, data sparsity is also confounded by the number of possible inflected forms for a given root form. With this in mind, and following on from the discussion in Section 5.4, we carry out further parsing experiments in an attempt to make better use of morphological information during parsing. We attack this in two ways: by reducing certain words to their lemmas and by including morphological information in the optional FEATS (features) field. The reasoning behind reducing certain word forms to lemmas is to further reduce the differences between inflected forms of the same word, and the reasoning behind including morphological information is to make more explicit the similarity between two different word forms inflected in the same way. All experiments are car-

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

Table 2: Results with morphological features on the development set

ried out with MaltParser and our seed training set of 500 gold trees. We focus on two phenomena: prepositional pronouns or pronominal prepositions (see Section 2) and verbs with incorporated subjects (see Section 2 and Section 5.4).

In the first experiment, we include extra morphological information for pronominal prepositions. We ran three parsing experiments: (i) replacing the value of the surface form (FORM) of pronominal prepositions with their lemma form (LEMMA), for example *agam*→*ag*, (ii) including morphological information for pronominal prepositions in the FEATS column. For example, in the case of *agam* 'at me', we include `Per=1P|Num=Sg`, (iii) we combine both approaches of reverting to lemma form and also including the morphological features. The results are given in Table 2.

In the second experiment, we include morphological features for verbs with incorporated subjects: imperative verb forms, synthetic verb forms and autonomous verb forms such as those outlined in Section 5.4. For each instance of these verb types, we included `incorpSubj=true` in the FEATS column. The results are also given in Table 2.

The experiments on the pronominal prepositions show a drop in parsing accuracy while the experiments carried out using verb morphological information showed no change in parsing accuracy.[6] In the case of inflected prepositions, perhaps we have not seen any improvement because we have not focused on a phenomenon which is critical for parsing. More experimentation is necessary.

## 7 Concluding Remarks

We have presented two sets of experiments which aim to improve dependency parsing performance for a minority language with a very small treebank. In the first set of experiments, the main focus of the paper, we tried to overcome the limited treebank size by increasing the parsers' training sets using automatically parsed sentences. While we do manage to achieve statistically significant improvements in some settings, it is clear from the results that the gains in parser accuracy through semi-supervised bootstrapping methods are fairly modest. Yet, in the absence of more gold labelled data, it is difficult to know now whether we would achieve similar or improved results by adding the same amount of gold training data. This type of analysis will be interesting at a later date when the unlabelled trees used in these experiments are eventually annotated and corrected manually.

The second set of experiments tries to mitigate some of the data sparseness issues by exploiting morphological characteristics of the language. Unfortunately, we do not see any improvements but we may get different results if we repeat these experiments using the larger semi-supervised training sets from the first set of experiments.

There are many directions this parsing research could take us in the future. Our unlabelled data consisted of sentences annotated with gold POS tags. In the future we would like to take advantage of the fully unlabelled, untagged data in the New Corpus for Ireland – Irish, which consists of 30 million words. We would also like to experiment with a fully unsupervised parser using this dataset. Our Malt feature models are manually optimised – it would be interesting to experiment with optimising them using MaltOptimizer (Ballesteros, 2012). An additional avenue of research would be to exploit the hierarchical nature of the dependency scheme to arrive at more flexible way of measuring agreement or disagreement in sample selection.

## Acknowledgements

We thank the three anonymous reviewers for their helpful feedback. 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.

[6] Although the total number of correct attachments are the same, the parser output is different.

## References


Miguel Ballesteros. 2012. Maltoptimizer: A system for maltparser optimization. In *Proceedings of the Eighth International Conference on Linguistic Resources and Evaluation (LREC)*, pages 2757–2763, Istanbul, Turkey.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*.

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*.

Eugene Charniak and Mark Johnson. 2005. Course-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of the 43rd ACL*.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI*.

Eugene Charniak. 2000. A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*.

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

David Greene. 1966. *The Irish Language*. Dublin: The Three Candles.

Zhongqiang Huang and Mary Harper. 2009. Self-training PCFG grammars with latent annotations across languages. In *Proceedings of EMNLP*.

John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. 2012. *The Irish Language in the Digital Age*. Springer Publishing Company, Incorporated.

Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasoul Samed Zadeh Kaljahi, and Anton Bryl. 2012. DCU-Paris13 systems for the sancl 2012 shared task. In *Working Notes of SANCL*.

Teresa Lynn, Özlem Çetinoğlu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith. 2012a. Irish treebanking and parsing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 1939–1946.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Uí Dhonnchadha. 2012b. Active learning and the Irish treebank. In *Proceeedings of the Australasian Language Technology Workshop (ALTA)*, pages 23–32.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, June. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. When is self-training effective for parsing? In *Proceedings of COLING*.

Joakim Nivre and Ryan McDonald. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*.

Mícheál Ó Siadhail. 1989. *Modern Irish: Grammatical structure and dialectal variation*. Cambridge: Cambridge University Press.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of EMNLP*, Hawaii.

Ines Rehbein. 2011. Data point selection for self-training. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2011)*, Dublin, Ireland.

Roi Reichart and Ari Rappaport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of ACL*.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*.

Kenji Sagae. 2010. Self-training without reranking for parser domain adapation and its impact on semantic role labelling. In *Proceedings of the ACL Workshop on Domain Adaptation for NLP*.

Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*.

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 331–338, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Elaine Uí Dhonnchadha. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.