Addressing some challenges of scarce resources in Irish NLP

Teresa Lynn
ADAPT Centre, Dublin City University
Outline

- Irish Language
- Status of Irish language technology
- A closer look at Irish parsing
- Universal Dependencies
- Conclusion
Outline

- Irish Language
  - Status of Irish language technology
  - A closer look at Irish parsing
  - Universal Dependencies
- Conclusion
Irish – a minority language

National Language
First Official Language

Census: 2016
Population: 4,761,865
Ability to speak: 1,761,420 people
Daily usage: 73,803 people

Source: https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/
Word Order = Verb Subject Object

**English:**  `I saw the boy`

**Irish:**  *Chonaic mé an buachaill*

**Gloss:**  Saw  I  the  boy
Initial Mutation

Eclipsis: Form=Ecl
Tír na nÓg ‘Land of the Youth’
i mBéarla ‘in English’
go mbíonn ‘that is’
ar an gcraobh ‘on the branch’

Lenition: Form=Len
sa cheantar ‘in the area’
a thuillfeadh ‘that would earn’
a dheartháir ‘his brother’
Vowel Harmony

*Caithim* – ‘I spend’
*Casaim* – ‘I turn’

*Rithfinn* – ‘I would run’
*D’íosfainn* – ‘I would eat’
## Inflected Prepositions

(16 simple prepositions)

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Meaning</th>
<th>Inflected Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>with</td>
<td>liom– ‘with me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>leat– ‘with you’</td>
</tr>
<tr>
<td>ag</td>
<td>at</td>
<td>agam– ‘at me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>agat – ‘at you’</td>
</tr>
<tr>
<td>faoi</td>
<td>about/under</td>
<td>fúm – ‘about/under me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fút – ‘about/under you’</td>
</tr>
<tr>
<td>ó</td>
<td>from</td>
<td>uaim– ‘from me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uait– ‘from you’</td>
</tr>
<tr>
<td>do</td>
<td>to</td>
<td>dom– to me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>duit– ‘to you’</td>
</tr>
<tr>
<td>ar</td>
<td>on</td>
<td>orm– ‘on me’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ort – ‘on you’</td>
</tr>
</tbody>
</table>
Prevalent use of clefting/fronting

*Creidtear gur go *mailíseach* a tosaíodh an tine*
‘It is believed that it was *mailiciously* that the file was started’

*Is san oifig a fhéiceann siad í*
‘It’s *in the office* they see her’

*B’ *ise a chonaic siad*
‘It is *she* whom they saw’

*B’ *ag obair a bhí sí nuair a chonaic muid í*
‘It is *working* that she was when I saw her’ (she was working when I saw her)
Outline

- Irish Language
- Status of Irish language technology
- A closer look at Irish parsing
- Universal Dependencies
- Conclusion
Some Terminology Issues

Irish = minority language
(speaking by the minority)

Irish = low/lesser-resourced language
(lacking language tools and resources)

BUT

Does “low-resourced” always mean “minority”??
Tagalog (Philippines)

- 21 million L1 speakers
- 50 million L2 speakers

Not a minority language...

...but is considered low-resourced
Examples of existing resources

- Speech synthesizer/Screen Reader
- Multiple electronic dictionaries, terminology DBs
- POS tagger/Morphological analyser/stemmer
- POS tagged corpus, Dependency treebank, Spoken Corpus, Parallel Data, Monolingual corpus (30 million words), Vicipéid (48k articles), DBpedia
- POS tagged Twitter corpus, POS-tagger for Irish tweets
- Chunking parser, statistical parser
- Basic CALL systems
- 2x Machine Translation systems (one in use by Government translators)
Examples of unfunded contributions
(Kevin Scannell)

- Spell-checker, Grammar Checker
- Localisation of: GNU/Linux, Mozilla, Open Office, Gmail, Facebook, Twitter
- Web-corpus collection
- English Irish SMT/ Irish-Scots Gaelic SMT
- Indigenous Tweets website
- Irish Web crawler
- WordNet for Irish
- Code.org in Irish
- Predictive Text Tool for Irish
Irish = A minority European Language
Irish = A low-resourced European Language
META-NET white paper series (Judge et al., 2012)

- EU-led study
- Survey of 31 EU languages
- Language resources and technologies
<table>
<thead>
<tr>
<th>MT</th>
<th>English</th>
<th>good</th>
<th>moderate</th>
<th>fragmentary</th>
<th>weak or no support</th>
</tr>
</thead>
<tbody>
<tr>
<td>French, Spanish</td>
<td></td>
<td></td>
<td></td>
<td>Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian</td>
<td>Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td>Text Analysis</td>
<td>English</td>
<td>good</td>
<td>moderate</td>
<td>fragmentary</td>
<td>weak or no support</td>
</tr>
<tr>
<td>Dutch, French, German, Italian, Spanish</td>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish</td>
<td>Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian, Swedish</td>
</tr>
<tr>
<td>Speech</td>
<td>English</td>
<td>good</td>
<td>moderate</td>
<td>fragmentary</td>
<td>weak or no support</td>
</tr>
<tr>
<td>Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish</td>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish</td>
<td>Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian, Welsh</td>
</tr>
<tr>
<td>Resources</td>
<td>English</td>
<td>good</td>
<td>moderate</td>
<td>fragmentary</td>
<td>weak or no support</td>
</tr>
<tr>
<td>Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish</td>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene</td>
<td>Icelandic, Irish, Latvian, Lithuanian, Maltese, Welsh</td>
</tr>
</tbody>
</table>
“Printing Press resulted in the extinction of many minority and regional languages”

Will technology have the same impact on Irish?
Need to ensure **continuing** language usage through technology

- Edutainment/ CALL packages
- Word processing tools
- Webpage translation
- Mobile platform support
- Search engines
- Games
- Social media
  - Sociolinguistic studies
  - Track misuse

What is the government doing?
What is the government doing?

Digital Strategy for the Irish Language 2018

Contributors:

- Teresa Lynn  
  *Dublin City University*
- John Judge  
  *Dublin City University*
- Elaine Uí Dhonnchadha  
  *Trinity College Dublin*
- Neasa Ní Chiaráin  
  *Trinity College Dublin*
- Ailbhe Ní Chasaide  
  *Trinity College Dublin*
Topics Covered:

<table>
<thead>
<tr>
<th>Linguistic Resources</th>
<th>Corpora</th>
<th>Knowledge Bases</th>
<th>NLP Tools</th>
<th>NLG Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Models</td>
<td>Speech Synthesis</td>
<td>Speech Recognition</td>
<td>Spoken Dialogue Systems</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>State and Public Use</td>
<td>CALL</td>
<td>Disability and Access</td>
<td>Synergies (Industry and Public)</td>
</tr>
</tbody>
</table>
GaelTech Project (2017-2021)

- Automatic Identification of Multiword Expressions
- NLP for Irish User-Generated Content
- Dependency Treebank(s) expansion (parsing)
Current Irish LT Projects at DCU

- Tapadóir SMT project
- European Language Resource Coordination
- Universal Dependencies for Irish
Outline

- Irish Language
- Status of Irish language technology
- A closer look at Irish parsing
- Universal Dependencies
- Conclusion
Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”
Parsing = who is doing what?

Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”
Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”
Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”
Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”

“True self-control is waiting until the movie starts to eat your popcorn”
Syntactic Parsing – Phrase Structure tree

S
  NP
    NNS
    Others
  VP
    VBP
    have
    VP
      VBN
      tried
      S
      VP
        TO
        to
        VP
        VB
        spruce
        PRT
        up
        NP
        JJ
        frequent-flier
        NNS
        programs
Problems with Constituency parsing

• Tied to linguistic theories such as transformational grammar

• Led by English-speaking linguists with English data (e.g. Chomsky)

• Word order is central to constituency grammars

• No close links to semantics
Problems with Constituency parsing

- Tied to linguistic theories such as **transformational grammar**
- Led by **English**-speaking linguists with English data (e.g. Chomsky)
- **Word order** is central to constituency grammars
- **No** close links to **semantics**
"We have already discussed this question at some length."
“Others have tried to spruce up frequent-flier programs”
Advantages of Dependency parsing

• Better handling of free word order (less-Anglo-centric)

• Node simplicity

• Clean mapping to semantic predicate-argument structure

• Easier to develop multilingual systems
Advantages of Dependency parsing

For Irish:

Disagreements in theoretical constituency syntax …
  o Flat VSO vs underlying SVO
  o Particles vs complementisers
  o Copula – linking element? Verb? Particle?
Sentence = a string/sequence of characters:

“True self-control is waiting until the movie starts to eat your popcorn”
"True self-control is waiting until the movie starts to eat your popcorn"

You are waiting
You will eat your popcorn
“True self-control is waiting until the movie starts to eat your popcorn”

True self-control is waiting
The movie will eat your popcorn
• Collection of parsed sentences *(trees)*

• Annotated with a pre-defined *part-of-speech tagset* *(Noun, Verb, etc)*

• Pre-defined *annotation scheme*  
  *(list of prescribed labels)*

• Pre-defined *linguistic* structure

• Used to develop *statistical parsers* *(train, test, and bootstrap)*
• Built upon gold POS-tagged corpus (Ui Dhonnchadha 2009)

• Newly-defined annotation scheme
  (list of prescribed labels)

• Inspired by LFG and Stanford dependencies (adapted for Irish)

• Currently 1020 trees

• Current parsing accuracy: LAS 71.4  UAS 80.1

Teresa Lynn, Irish Dependency Treebanking and Parsing. PhD Thesis 2016, Dublin City University and Macquarie University, Sydney
Resource-poor

• Lack of funding

• Lack of text resources

• Lack of skilled annotators

• Time-intensive
Basic Bootstrapping Approach (Passive Learning)

Training data (gold)

Train parser

Correct Parse Trees

Parse new text
Active Learning: Query-by-Committee

- Training Data (gold)
- Manually correct parses x1
- Parse New Data x2
- Select “informative” trees
- Train parser x2
## Active Learning vs Passive Learning

Lynn, Teresa, Jennifer Foster, Mark Dras and Elaine Uí Dhonnchadha, Active Learning and the Irish Treebank, ALTA 2012, Dunedin, NZ, December 2012

<table>
<thead>
<tr>
<th>Experiment</th>
<th>baseline</th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
<th>Iteration 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive LAS</td>
<td>65.86</td>
<td>65.36</td>
<td>66.89</td>
<td>68.39</td>
<td>68.71</td>
</tr>
<tr>
<td>Active LAS</td>
<td>65.86</td>
<td>66.5</td>
<td>68.46</td>
<td>68.81</td>
<td>67.92</td>
</tr>
<tr>
<td>Passive UAS</td>
<td>75.6</td>
<td>75.11</td>
<td>76.81</td>
<td>77.67</td>
<td>77.49</td>
</tr>
<tr>
<td>Passive UAS</td>
<td>75.6</td>
<td>76.74</td>
<td>78.34</td>
<td>78.49</td>
<td>78.2</td>
</tr>
</tbody>
</table>
Active Learning vs Passive Learning

UAS

LAS
Outline

- Irish Language
- Status of Irish language technology
- A closer look at Irish parsing
- Universal Dependencies
- Conclusion
Dependency Treebanks – variations

Varying labelling conventions:
Dependency Treebanks – variations

Varying labelling conventions:
Varying structural analyses:
Varying structural analyses:
**Problems** with variations:

- Difficult to do cross-lingual analysis
- Difficult to compare parser performance
- Difficult to do cross-lingual transfer
  (using data from one language to help another)
- Difficult to build and evaluate multilingual systems
**Premise:**
no Universal Grammar, but:
“all languages share fundamental similarities” (linguistic universals)

**Goals:**
- develop a set of harmonised dependency treebanks
- design a universal annotation scheme
- enable comparison of treebanks
- enable comparison of parsing results
- improve multilingual processing
Manning’s Law

The secret to understanding the design of UD is to realize that it is a very subtle compromise between approximately 6 things:

1. UD needs to be satisfactory on linguistic analysis grounds for individual languages.
2. UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be suitable for computer parsing with high accuracy.
5. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, ...).

It’s easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.
Solution: Universal Dependencies Project

Part-of-speech tags

Morphological features

Slide credit: Chris Manning, Stanford University
Part-of-Speech Tags

<table>
<thead>
<tr>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADP</td>
<td>PUNCT</td>
</tr>
<tr>
<td>ADV</td>
<td>AUX</td>
<td>SYM</td>
</tr>
<tr>
<td>INTJ</td>
<td>CCONJ</td>
<td>X</td>
</tr>
<tr>
<td>NOUN</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>PROPN</td>
<td>NUM</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>PART</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCONJ</td>
<td></td>
</tr>
</tbody>
</table>

Taxonomy of 17 universal part-of-speech tags, expanding on the Google Universal Tagset (Petrov et al., 2012)

All languages use the same inventory, but not all tags have to be used by all languages.

Slide credit: Chris Manning, Stanford University
• Content words are related by dependency relations
• Function words attach to the content word they further specify
• Punctuation attaches to head of phrase or clause
• **40** universal grammatical relations (de Marneffe et al., 2014) (aim to address linguistic universals across languages)

• Language-specific **subtypes** may be added (e.g. Irish UD: *csubj:cleft*)
• Standardized inventory of morphological features, based on the Interset system (Zeman, 2008)
• Languages select relevant features and can add language-specific features or values with documentation

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Inflectional Nominal</th>
<th>Inflectional Verbal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PronType</td>
<td>Gender</td>
<td>VerbForm</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
<td>Mood</td>
</tr>
<tr>
<td>Poss</td>
<td>Number</td>
<td>Tense</td>
</tr>
<tr>
<td>Reflex</td>
<td>Case</td>
<td>Aspect</td>
</tr>
<tr>
<td></td>
<td>Definite</td>
<td>Voice</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polarity</td>
</tr>
</tbody>
</table>
Features – CoNLL-U format

```
# sent_id = 904
# text = Creidtear gur go mailiseach a tosaíodh an tine.
1  Creidtear  creid  VERB  VTI  Mood=Ind|Tense=Pres|Voice=Auto  0  root  _
2   gur     is AUX  Cop  Tense=Pres|VerbForm=Cop  4  cop  _  _
3   go     go PART  Ad  PartType=Ad  4  mark:pst  _  _
4  mailiseach mailiseach ADJ  Adj  Degree=Pos  1  ccomp  _  _
5    a    a PART  Vb  PartType=Vbl|PronType=Rel  6  mark:pst  _
6  tosaíodh tosaígh  VERB  VTI  Mood=Ind|Tense=Past|Voice=Auto  4  csubj:cleft
7   an     an DET  Art  Definite=Def|Number=Sing|PronType=Art  8  det  _
8   tine   tine NOUN  Noun  Case=NomAcc|Gender= Fem|Number=Sing  6  obj  _
9   .     . PUNCT  .   _  1  punct  _  _
```

Source: Irish Universal Dependencies Treebank
Timeline of UD project to date

Release of annotation guidelines (v1): October 2014
- 10 treebanks: January 2015
- 18 treebanks: May 2015
- 37 treebanks: November 2015
- 54 treebanks: May 2016
- 64 treebanks: November 2016

- 70 treebanks (50 languages): March 2017
- 102 treebank (60 languages): November 2017
- 122 treebanks (71 languages): July 2018
<table>
<thead>
<tr>
<th>Language</th>
<th>Count</th>
<th>POS</th>
<th>đề</th>
<th>Người</th>
<th>OK</th>
<th>NX</th>
<th>WX</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>254K</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English–ESL</td>
<td>97K</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English–LinES</td>
<td>82K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonian</td>
<td>234K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>181K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish–FTB</td>
<td>159K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>390K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galician</td>
<td>138K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>293K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gothic</td>
<td>56K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td>59K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>115K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>351K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>42K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>121K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irish</td>
<td>23K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>252K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese–KTC</td>
<td>267K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kazakh</td>
<td>4K</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>-</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Introduction**
- **Tokenization**
- **Morphology**
  - General principles
  - Irish POS tags (single document)
  - Irish features (single document)
- **Syntax**
  - General principles
  - Specific constructions
  - Irish relations (single document)
Outline

- Irish Language
- Status of Irish language technology
- A closer look at Irish parsing
- Universal Dependencies
- Conclusion
Conclusion

When you have limited resources…

- Make use of Bootstrapping / leveraging approaches
- Involvement in larger projects (COST, UD)
- Organise workshops for sharing knowledge/collaborations/networking
- Crowdsourcing (empower the language community)
- Seek Government support
Influence Government Policy …

- Analysing online language use
- Empirically demonstrating evolution of language
- Starting off with pilot systems and demonstrate the benefits of LT
- Teaming up with other (similar) minority languages
- Involvement with public engagement – pop science
Conclusion

All this can lead to:

Understanding of **need** for language technology
#GRMA

Go raibh maith agaibh
Thank you (pl)