Engaging Content
Engaging People

Addressing some challenges of scarce resources in Irish NLP

Teresa Lynn

ADAPT Centre, Dublin City University

Ireland's European Structural and Investment Funds Programmes 2014-2020

European Ur
European Regiona

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

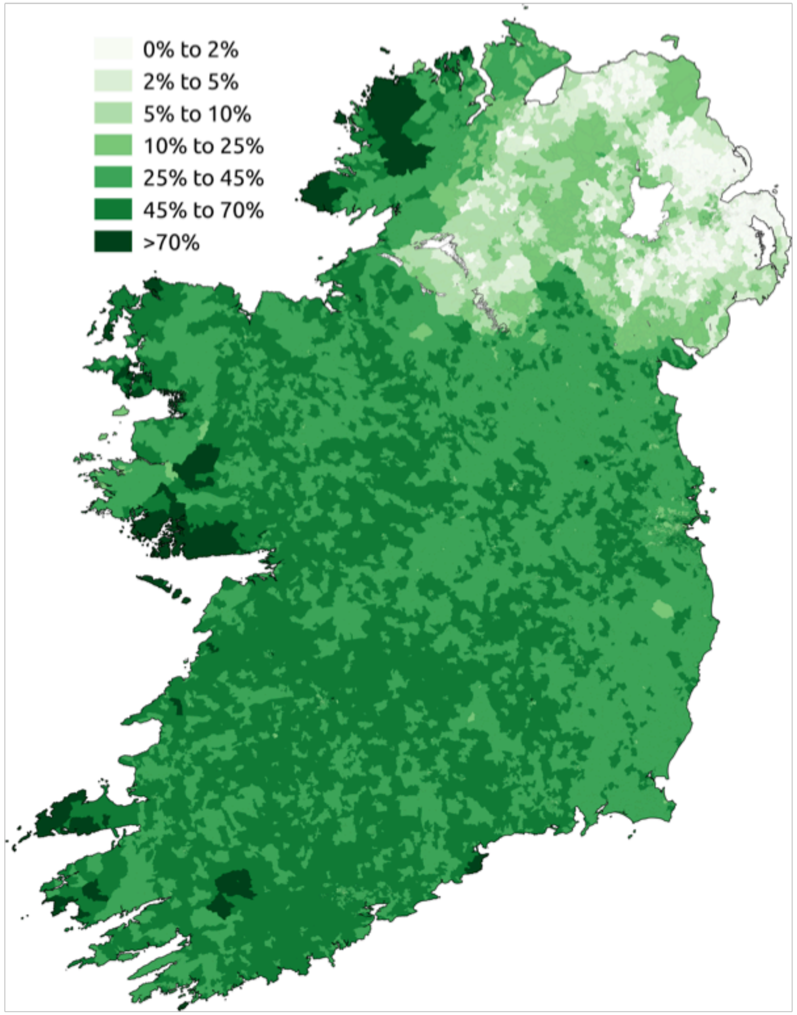
Outline

- 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/

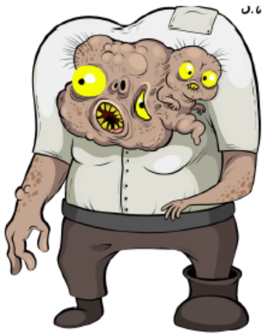
Irish Language Features

Word Order = Verb Subject Object

English:	`I saw the boy’			
Irish:	*Chonaic*	*mé*	*an*	*buachaill*
Gloss:	Saw	I	the	boy

Irish Language Features

iNitial mUtation



Eclipsis: *Form=Ecl*

Tír na ***n****Óg* ‘Land of the Youth’

i ***m****Béarla* ‘in English’

go ***m****bíonn* ‘that is’

ar an ***g****craobh* ‘on the branch’

Lenition: *Form=Len*

*sa c****h****eantar* ‘in the area’

*a t****h****uillfeadh* ‘that would earn’

*a d****h****eartháir* ‘his brother’

Irish Language Features

Vowel Harmony

*Caith**im*** – `**I** spend’
*Cas**aim*** – `**I** turn’

*Rith**finn*** – `**I would** run’
*D’íos**fainn*** – `**I would** eat’

Irish Language Features

Inflected Prepositions

(16 simple prepositions)

le – with

liom– `with **me**’

leat– `with **you**’

ag – at

agam– `at **me**’

agat – `at **you**’

faoi – about/under

fúm – ‘about/under **me**’

fút – ‘about/under **you**’

ó – from

uaim– `from **me**’

uait– `from **you**’

do – to

dom– to **me**’

duit– `to **you**’

ar – on

orm– ‘on **me**’

ort – ‘on **you**’

Irish Language Features

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 fheiceann siad í*
‘It’s **in the office** they see her’

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

*B’ **ag obai**r 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

(spoken by the minority)

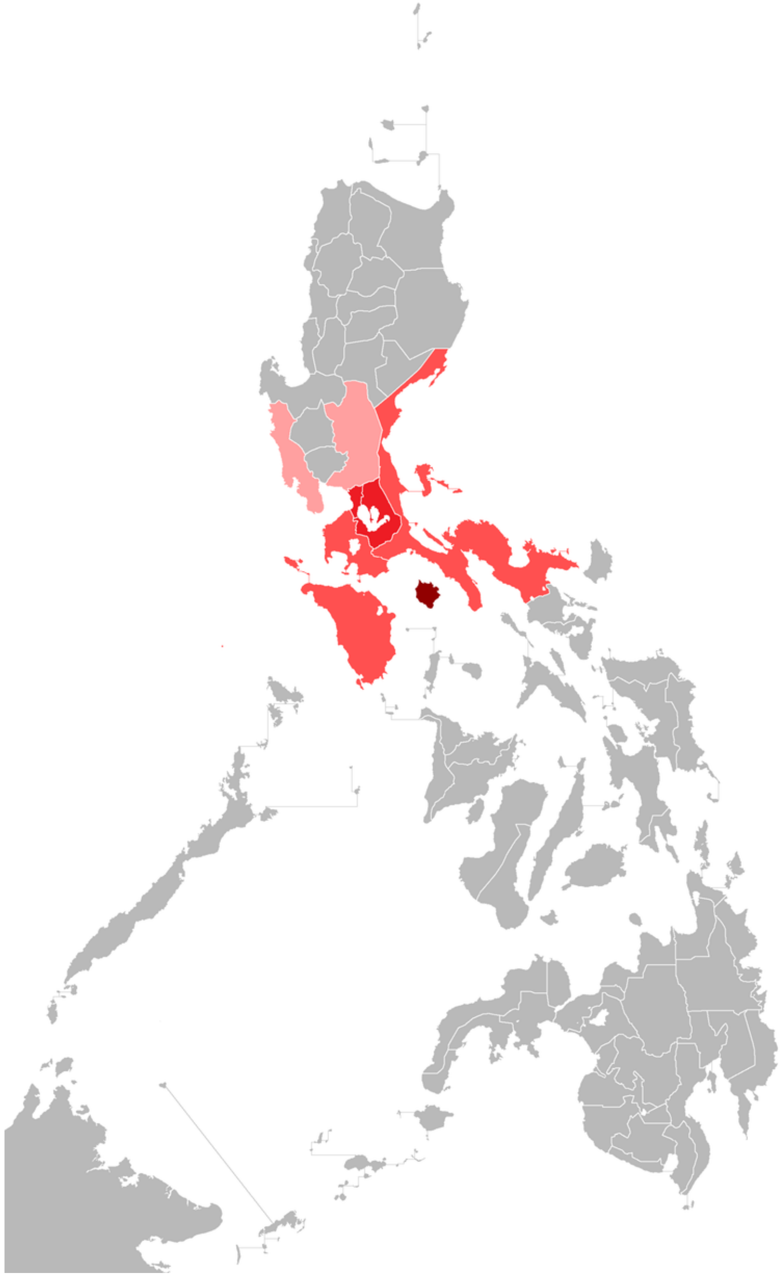
Irish = low/lesser-resourced language

(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

Recognised as low-resourced

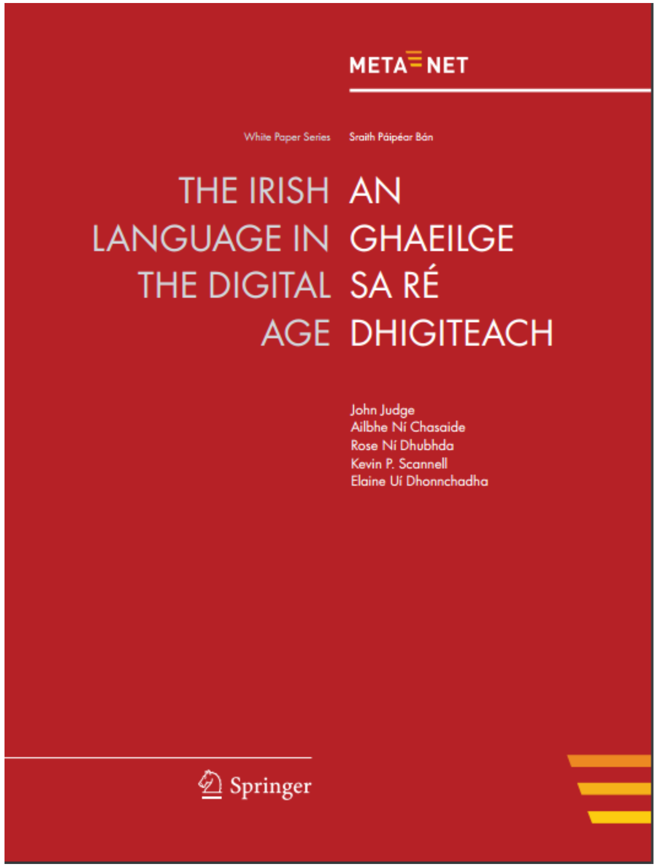
Irish = A minority European Language

Irish = A low-resourced European Language

Irish language technology survey

META-NET white paper series (Judge et al., 2012)

- ○ EU-led study
- ○ Survey of 31 EU languages
- ○ Language resources and technologies



META=NET

White Paper Series

Sraith Páipéar Bán

THE IRISH LANGUAGE IN THE DIGITAL AGE

AN GHAEILGE SA RÉ DHIGITEACH

John Judge
Ailbhe Ní Chasaide
Rose Ní Dhubhda
Kevin P. Scannell
Elaine Uí Dhonnchadha

Springer

MT

excellent	good	moderate	fragmentary	weak or no support
	English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, **Irish**, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish, Welsh

Text Analysis

excellent	good	moderate	fragmentary	weak or no support
	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, **Irish**, Latvian, Lithuanian, Maltese, Serbian, Welsh

Speech

excellent	good	moderate	fragmentary	weak or no support
	English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, **Irish**, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian, Welsh

Resources

excellent	good	moderate	fragmentary	weak or no support
	English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, **Irish**, Latvian, Lithuanian, Maltese, Welsh

Language at Risk – in Digital Age

“Printing Press resulted in the extinction of many minority and regional languages”

Will technology have the same impact on Irish?

Language at Risk – in Digital Age

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

Source: http://www.leuphana.de/institute/ies/llt2015.html

What is the government doing?

What is the government doing?

Digital Strategy for the Irish Language 2018

An Roinn
Ealaíon, Oidhreachta agus Gaeltachta
Department of
Arts, Heritage and the Gaeltacht

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*

Digital Strategy for the Irish Language 2018

Topics Covered:

- Linguistic Resources
- Corpora
- Knowledge Bases
- NLP Tools
- NLG Tools
- Speech Models
- Speech Synthesis
- Speech Recognition
- Spoken Dialogue Systems
- Machine Translation
- Information Retrieval
- State and Public Use
- CALL
- Disability and Access
- Synergies (Industry and Public)

Current Irish LT Projects at DCU

GaelTech Project (2017-2021)

An Roinn
Cultúir, Oidhreachta agus Gaeltachta

Department of
Culture, Heritage and the Gaeltacht

- 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

Translate

- European Language Resource Coordination

European Language Resource Coordination
Connecting Europe Facility

- Universal Dependencies for Irish

Outline

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

Parsing = who is doing what?

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*"

Parsing = who is doing what?

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*"

Parsing = who is doing what?

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*"

Parsing = who is doing what?

Sentence = a string/sequence of characters:

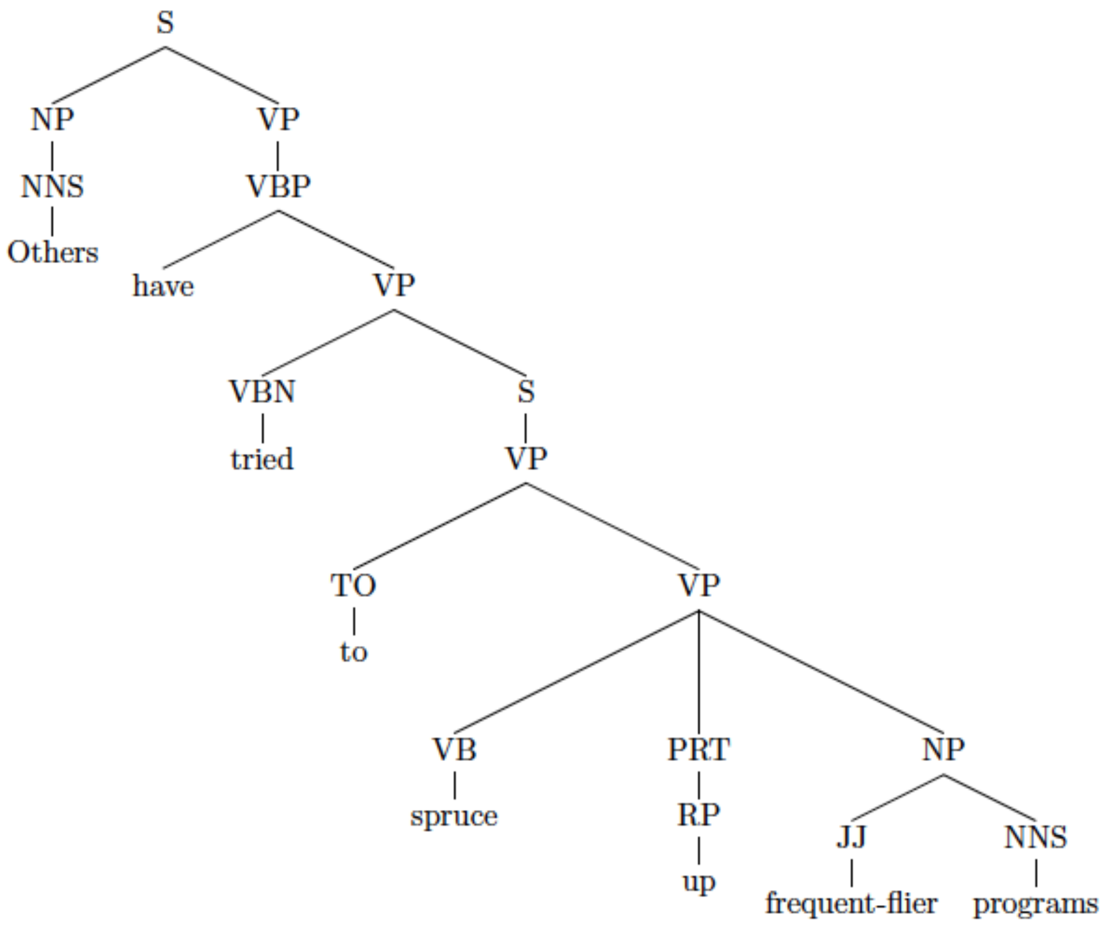
"*True self-control is waiting until the movie starts to eat your popcorn*"

"*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



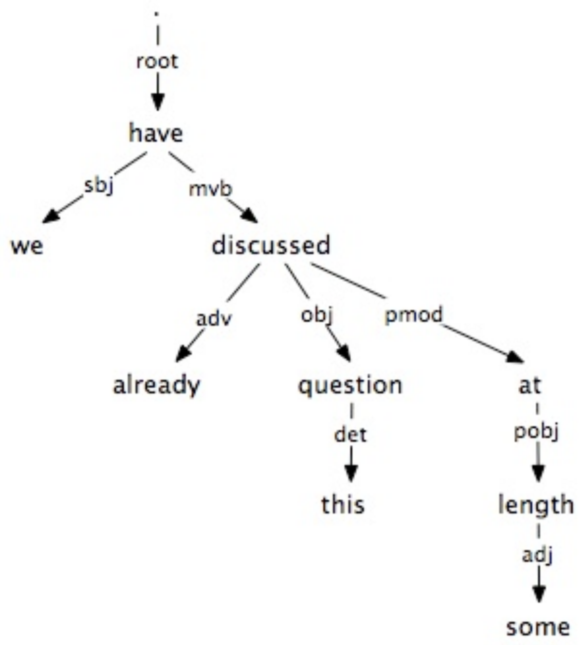
Problems with Constituency parsing

- 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

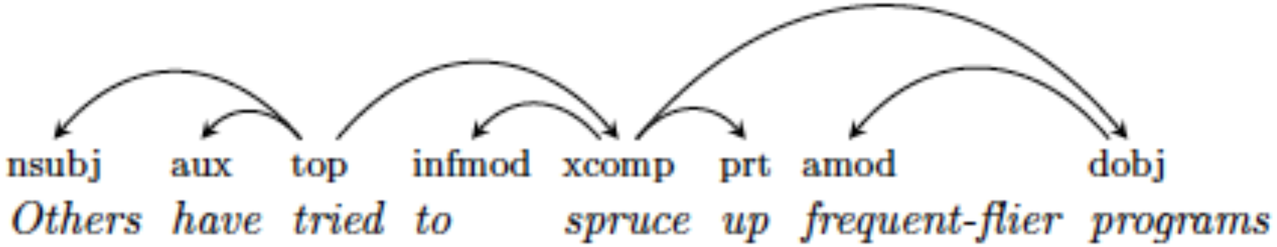
- Tied to linguistic theories such as **transformational grammar**
- Led by **English**-speaking linguists with English data (e.g. Chomsky)
- **Word order** is central to constituency grammars
- **No** close links to **semantics**

Syntactic Parsing - Dependency trees



“We have already discussed this question at some length.”

Syntactic Parsing - Dependency trees


nsubj
aux
top
infmod
xcomp
prt
amod
dobj
Others *have* *tried* *to* *spruce* *up* *frequent-flier* *programs*

“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 …

- Flat VSO vs underlying SVO
- Particles vs complementisers
- Copula – linking element? Verb? Particle?

What does a machine know about language?

Sentence = a string/sequence of characters:

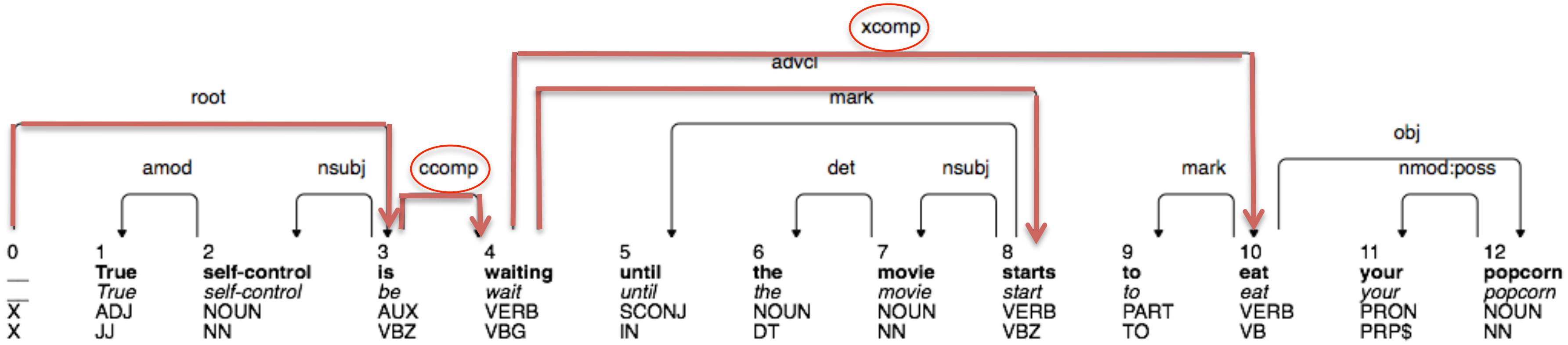
“*True self-control is waiting until the movie starts to eat your popcorn*”

Parsing = who is doing what?

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

You are waiting

You will eat your popcorn

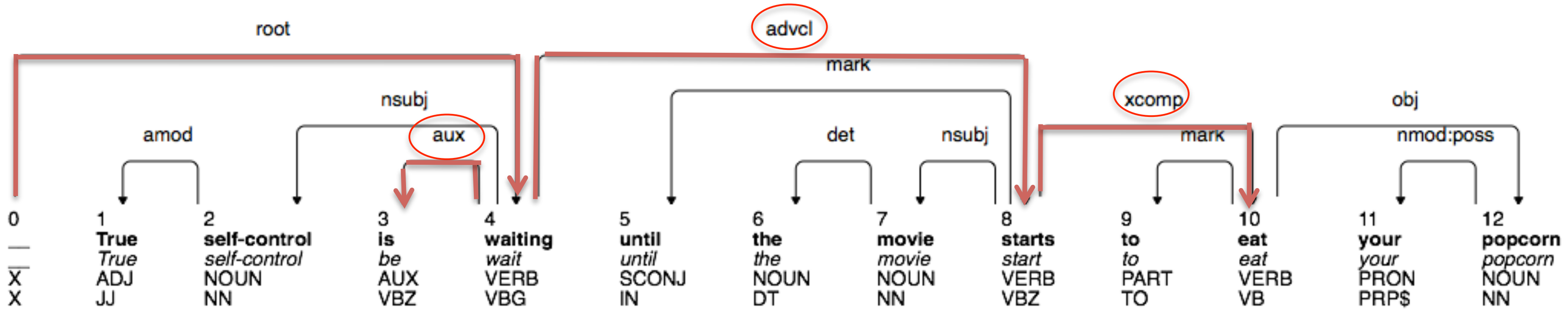


Parsing = who is doing what?

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

True self-control is waiting

The movie will eat your popcorn



Treebanks

- 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)

Irish Dependency Treebank

- 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

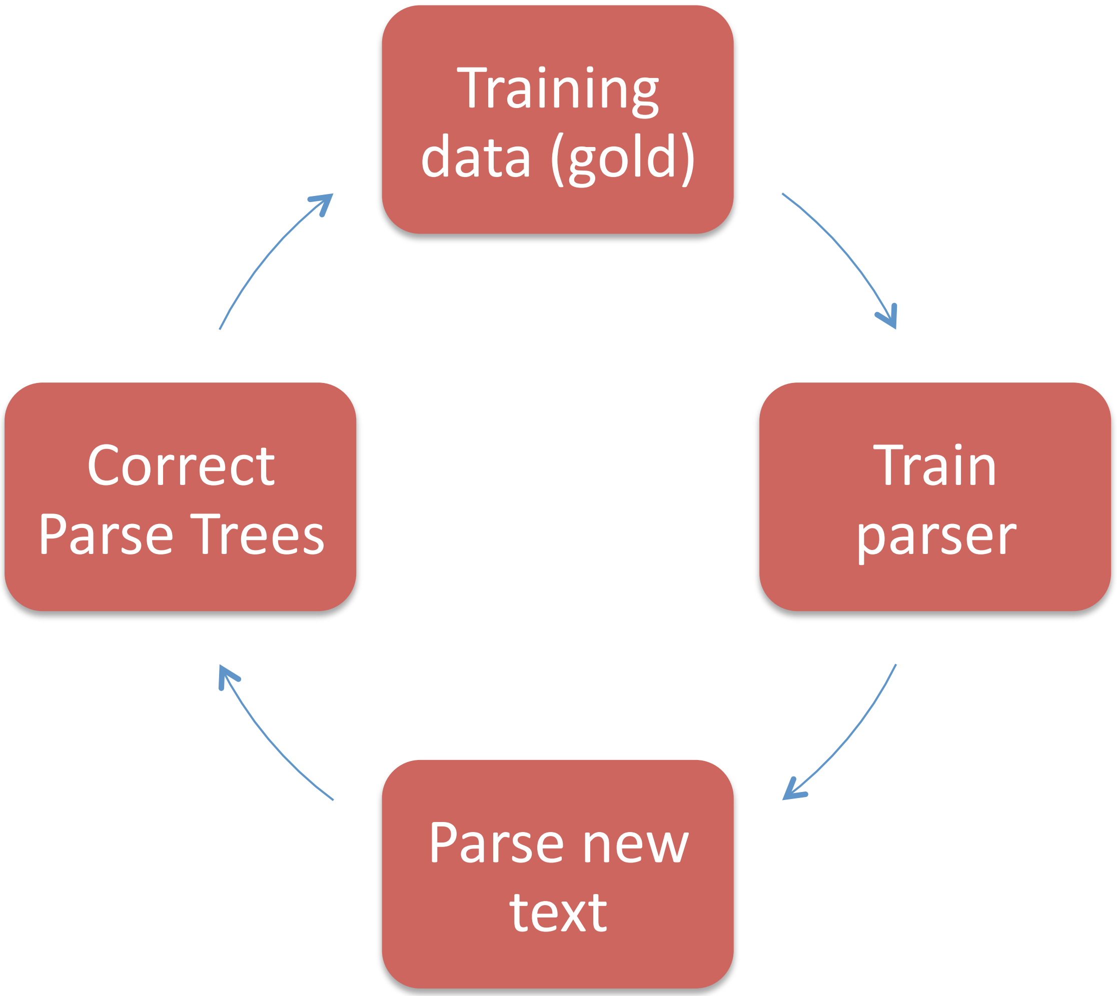
Teresa Lynn, Irish Dependency Treebanking and Parsing. PhD Thesis 2016, Dublin City University and Macquarie University, Sydney

Challenges for Irish Treebank Development

Resource-poor

- Lack of funding
- Lack of text resources
- Lack of skilled annotators
- Time-intensive

Basic Bootstrapping Approach (Passive Learning)



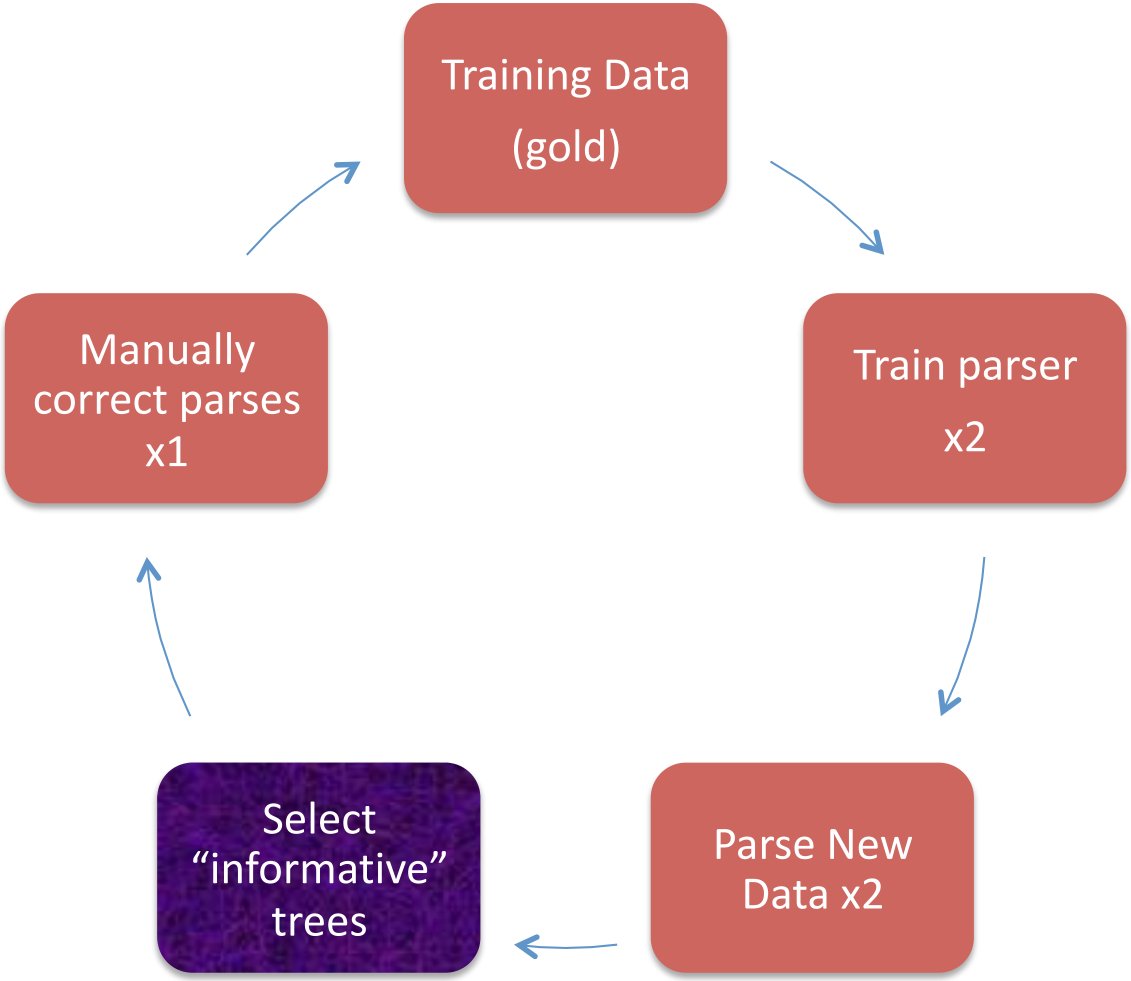
Training data (gold)

Train parser

Parse new text

Correct Parse Trees

Active Learning: Query-by-Committee



Training Data (gold)

Train parser x2

Parse New Data x2

Select “informative” trees

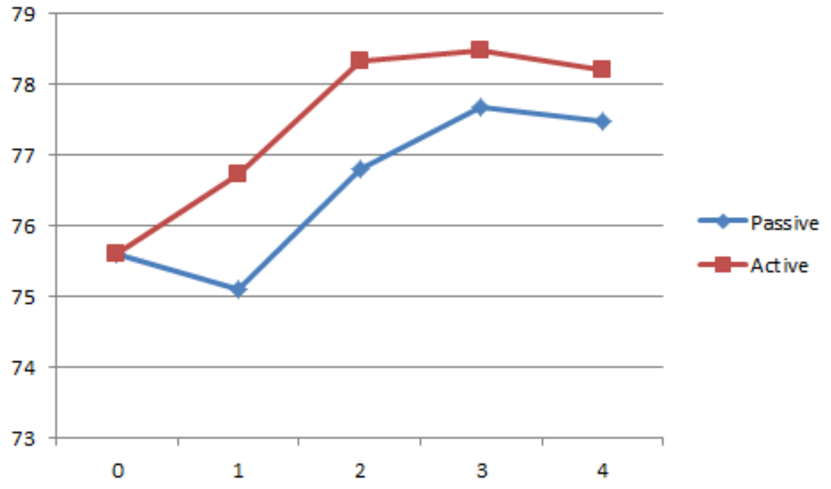
Manually correct parses x1

Active Learning vs Passive Learning

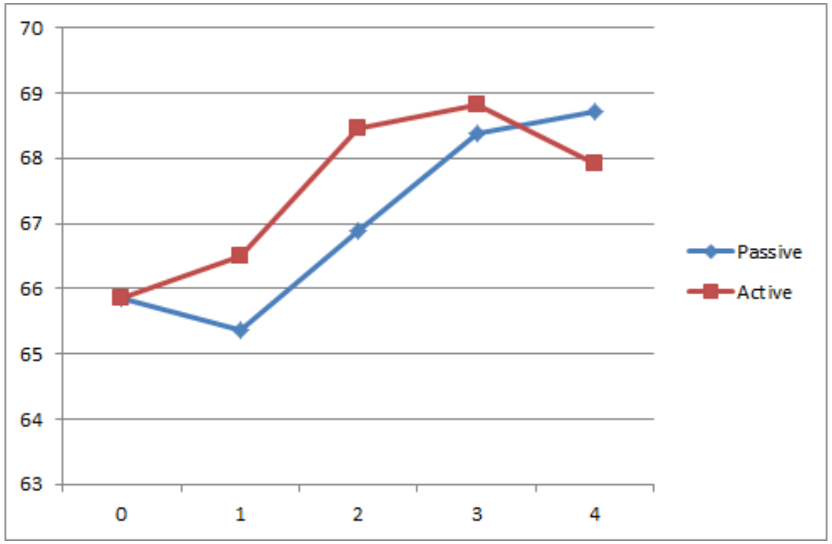
Experiment	baseline	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Passive LAS	65.86	65.36	66.89	68.39	68.71
Active LAS	65.86	66.5	68.46	68.81	67.92
Passive UAS	75.6	75.11	76.81	77.67	77.49
Passive UAS	75.6	76.74	78.34	78.49	78.2

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

Active Learning vs Passive Learning



UAS



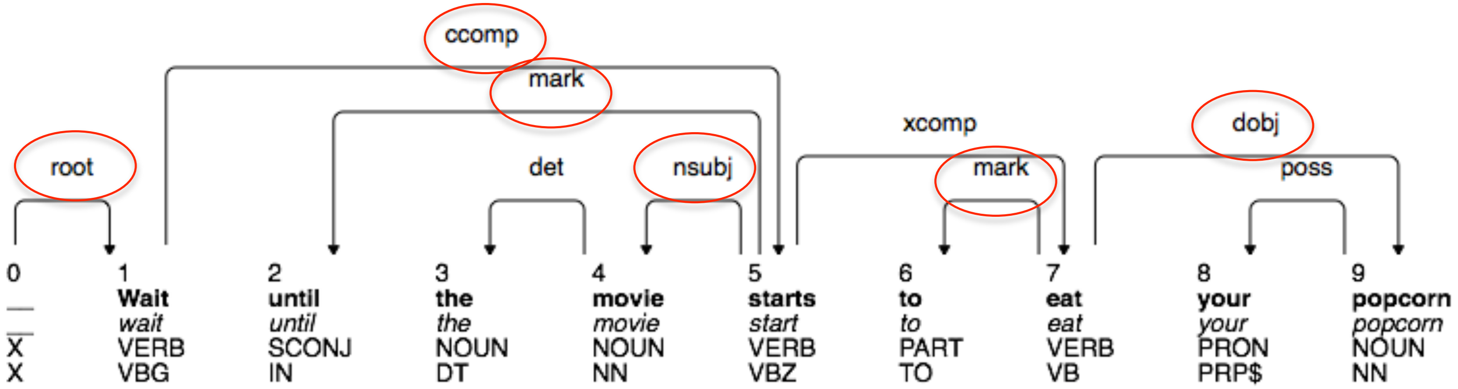
LAS

Outline

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

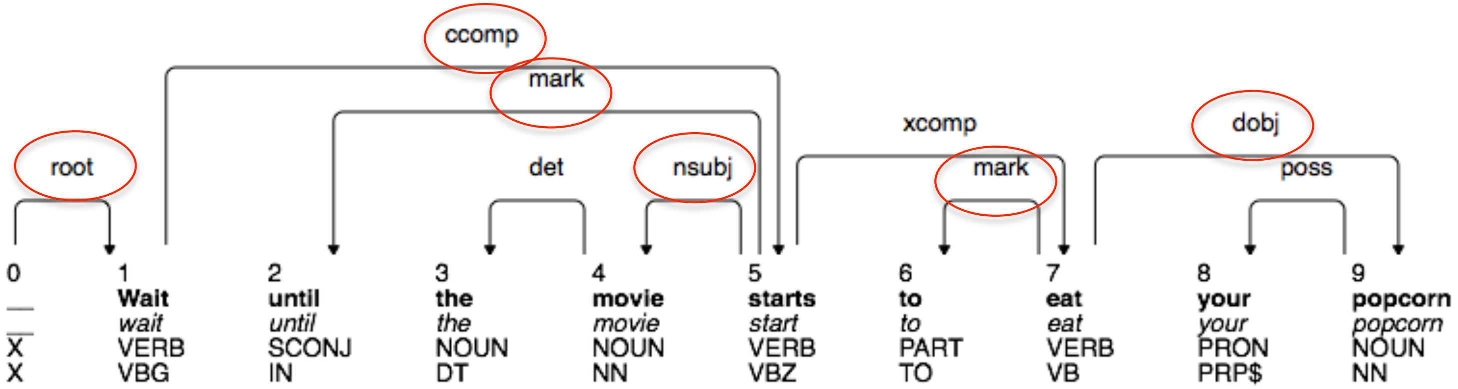
Dependency Treebanks – variations

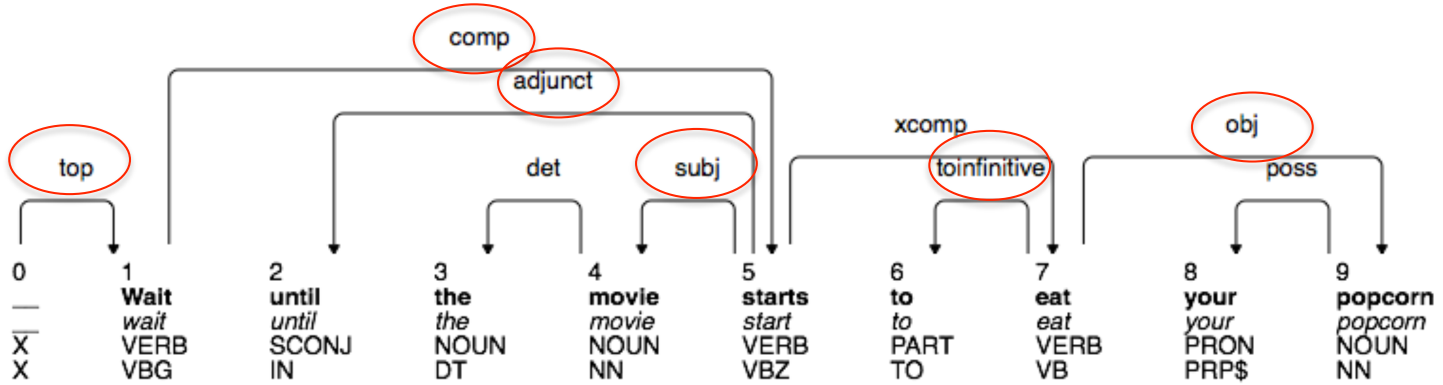
Varying labelling conventions:



Dependency Treebanks – variations

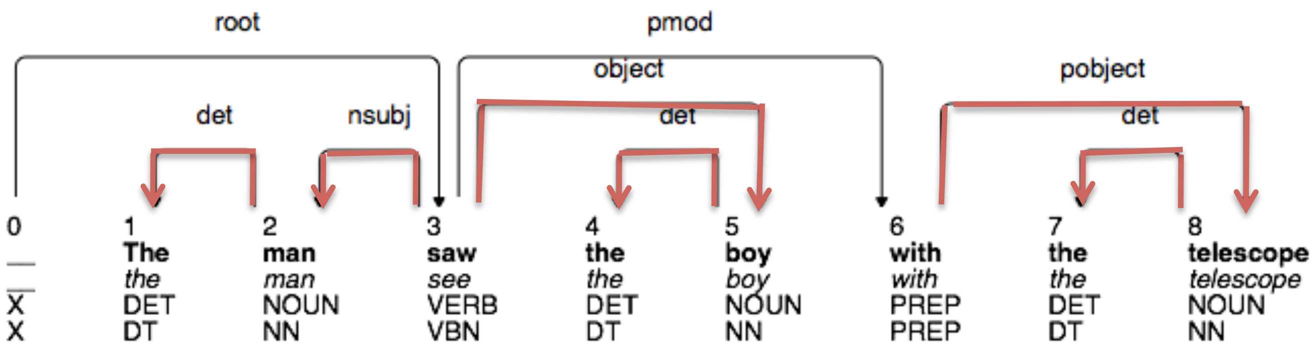
Varying labelling conventions:





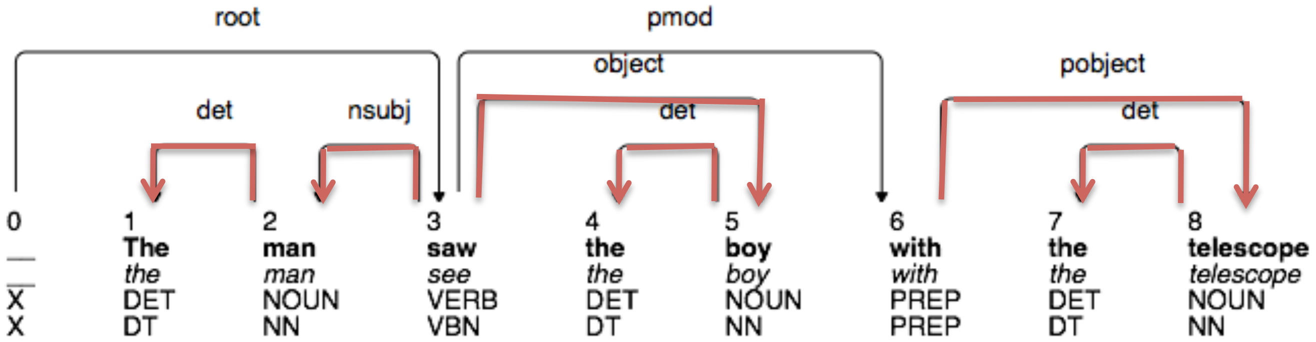
Dependency Treebanks – variations

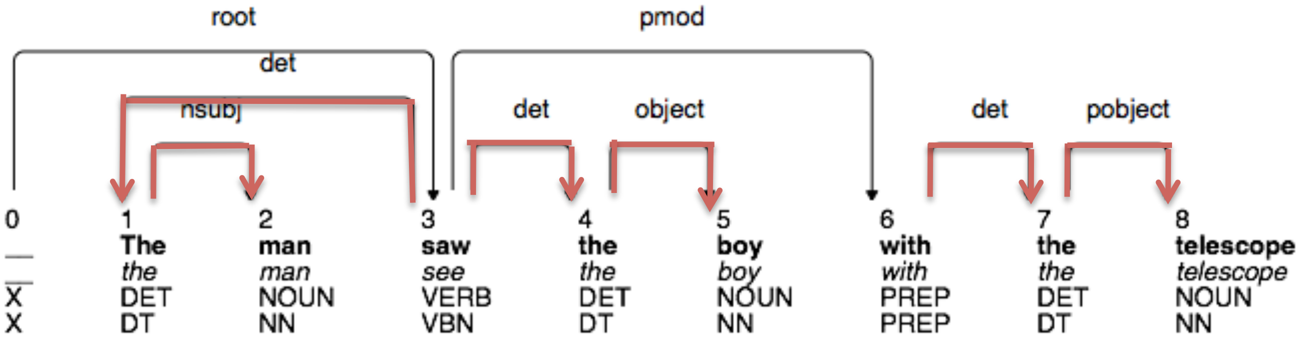
Varying structural analyses:



Dependency Treebanks – variations

Varying structural analyses:





Treebanks – variations

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

Solution: Universal Dependencies Project

Premise:

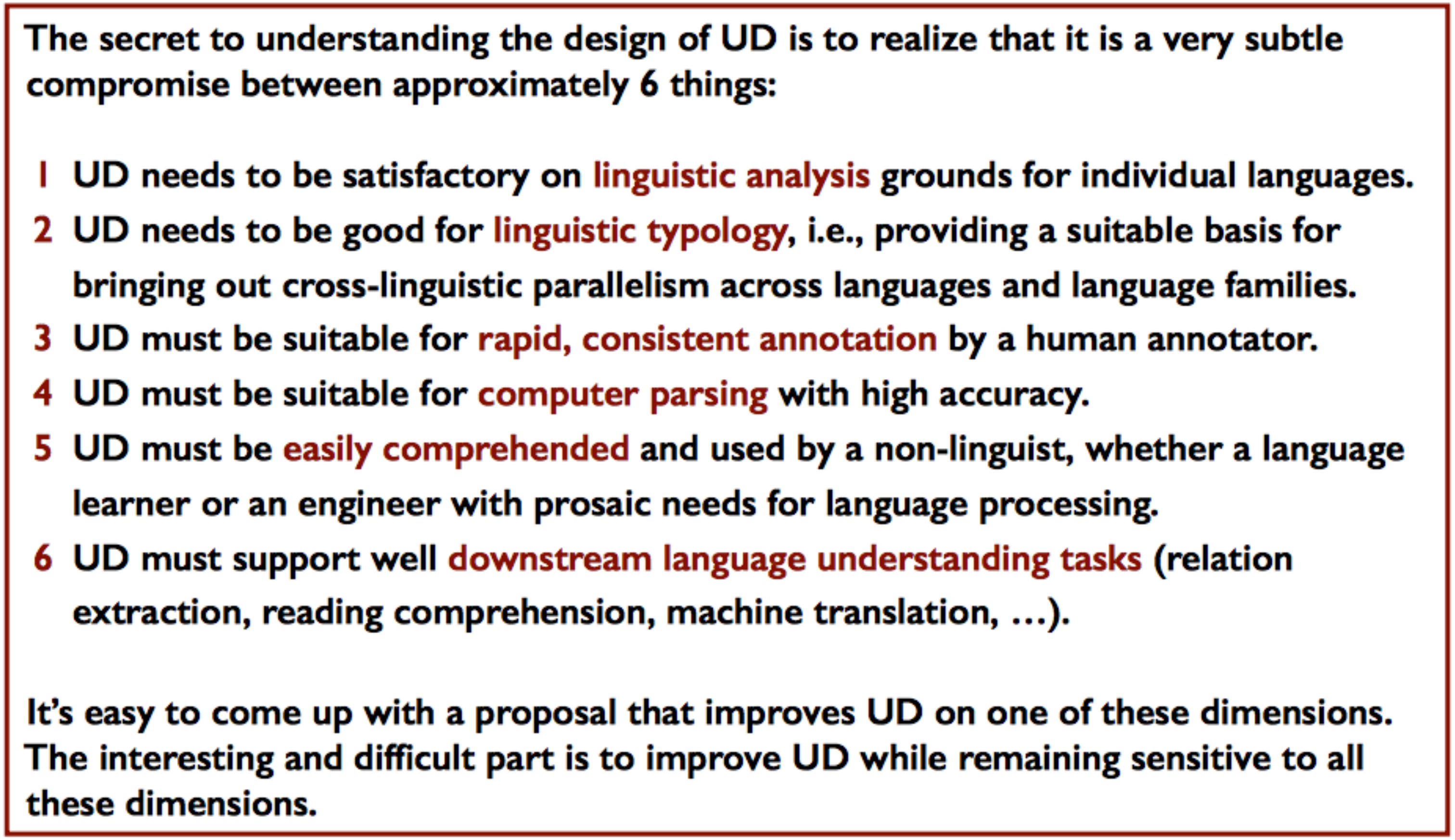
no Universal Grammar, but:
“all languages share fundamental similarities” (linguistic universals)

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

Solution: Universal Dependencies Project

Manning's Law



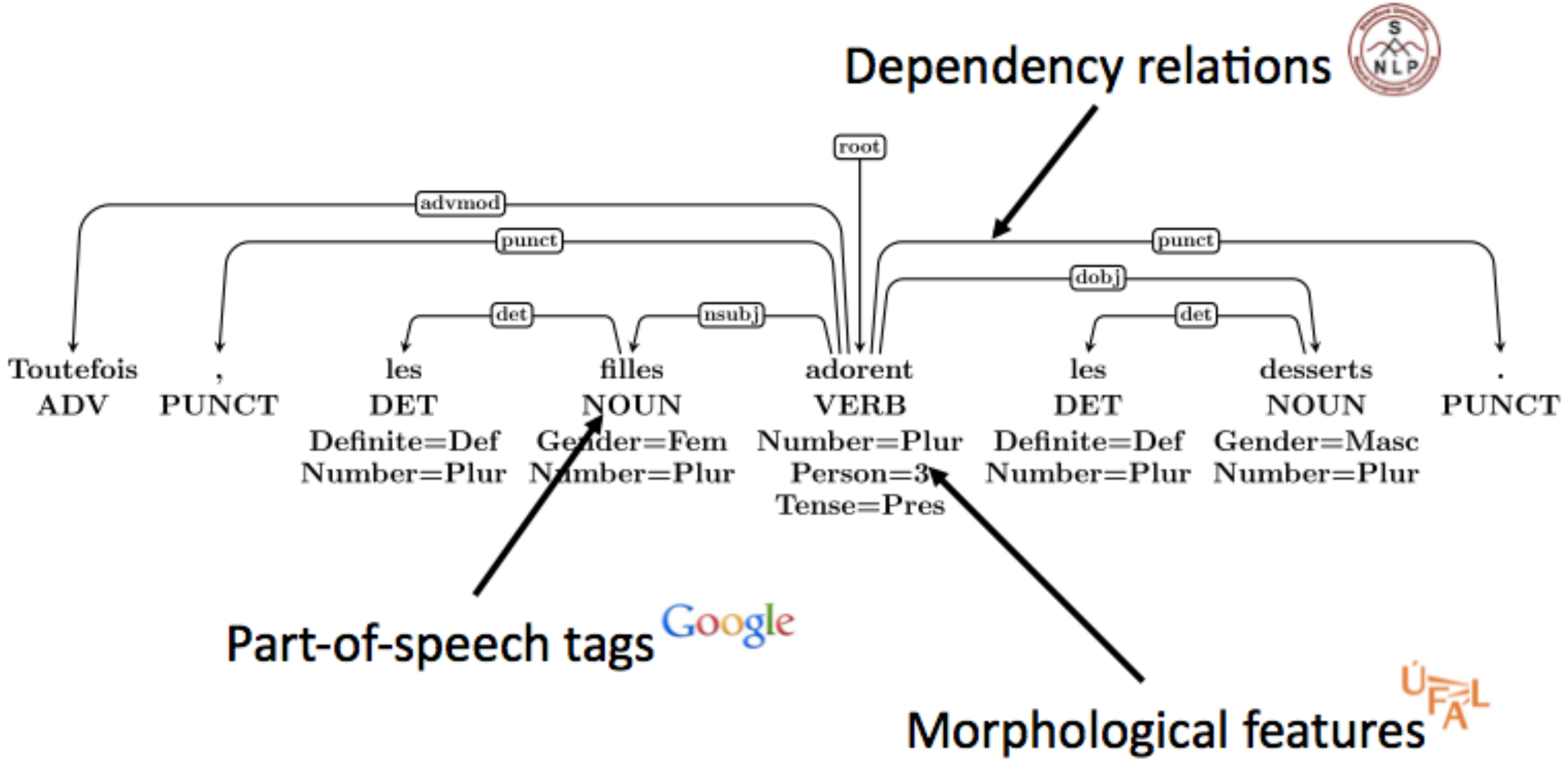
The secret to understanding the design of UD is to realize that it is a very subtle compromise between approximately 6 things:

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.

Slide credit: Chris Manning, Stanford University

Solution: Universal Dependencies Project



Slide credit: Chris Manning, Stanford University

Part-of-Speech Tags

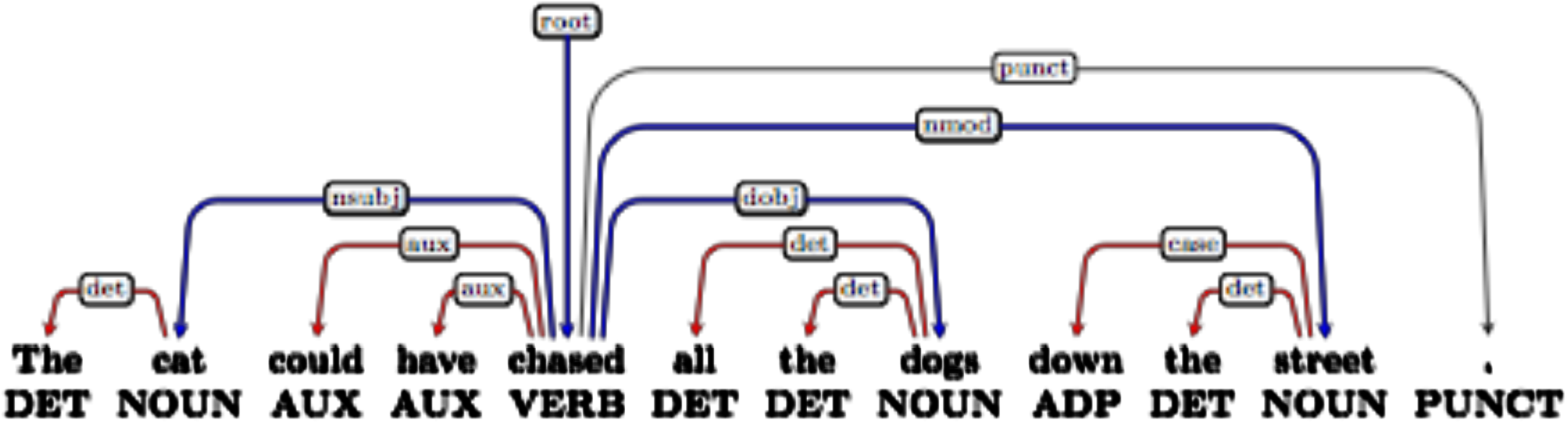
Open	Closed	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

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

Syntax



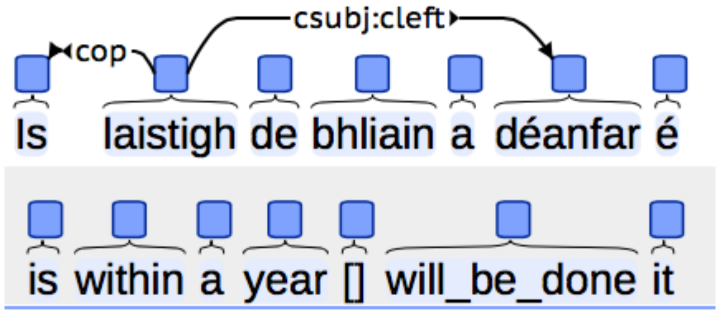
- Content words are related by dependency relations
- Function words attach to the content word they further specify
- Punctuation attaches to head of phrase or clause

Slide credit: Chris Manning, Stanford University

Dependency Relations

- **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*)


csubj:cleft
cop
Is laistigh de bhliain a déanfar é
is within a year [] will_be_done it

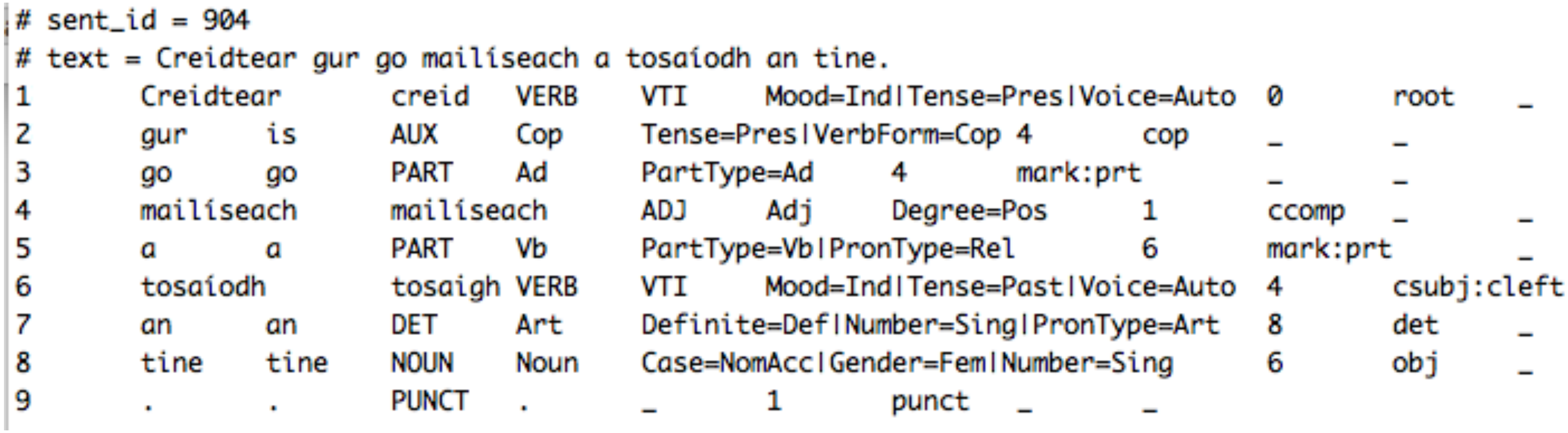
Features

Lexical	Inflectional Nominal	Inflectional Verbal
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
	Definite	Voice
	Degree	Person
		Polarity

- 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

Slide credit: Chris Manning, Stanford University

Features – CoNLL-U format



```
# sent_id = 904
# text = Creidtear gur go mailíseach 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:prt        _       _
4       mailíseach      mailíseach      ADJ     Adj     Degree=Pos      1       ccomp   _       _
5       a       a       PART    Vb      PartType=Vb|PronType=Rel        6       mark:prt        _
6       tosaíodh        tosaigh 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

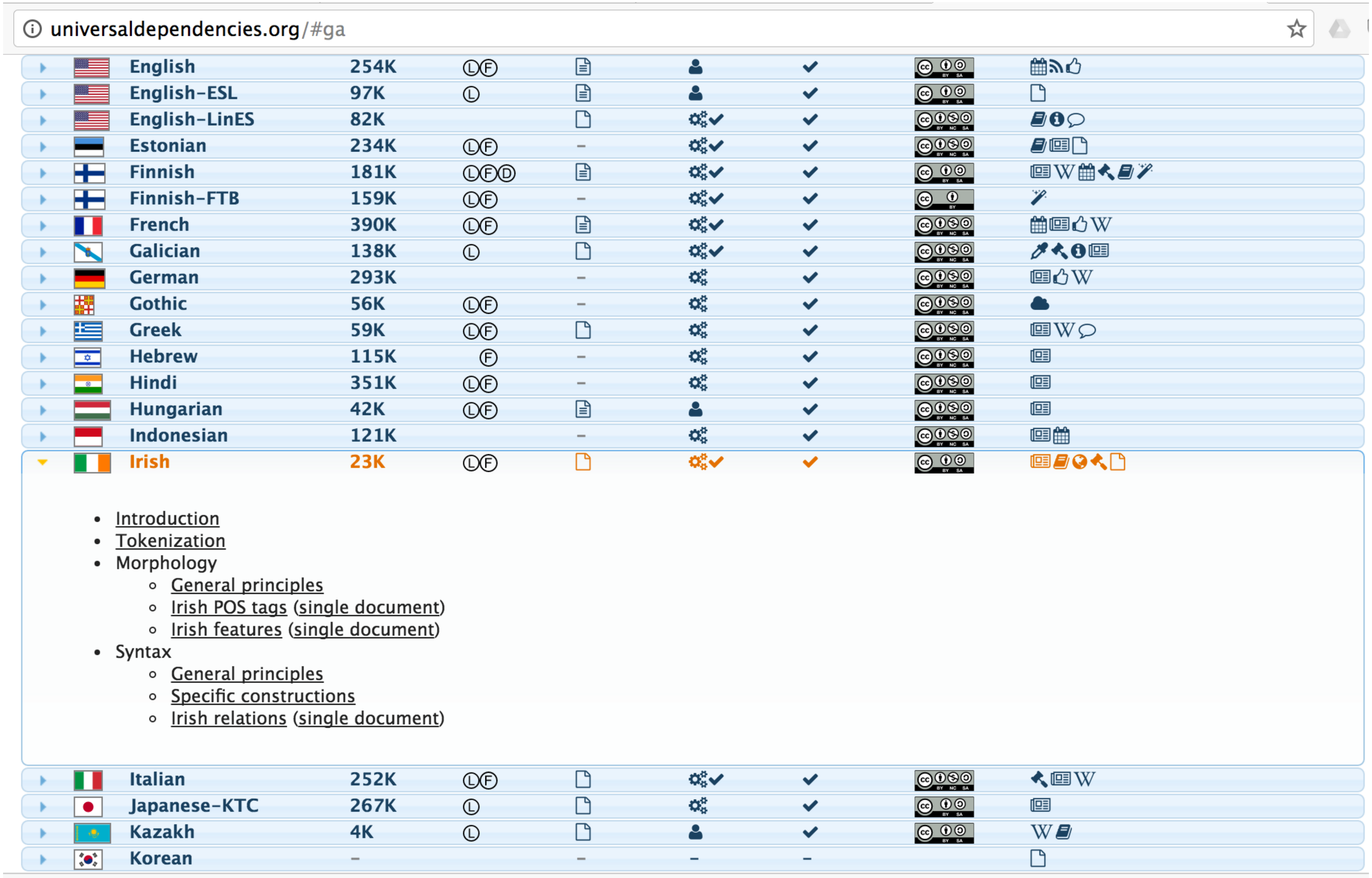
Release of annotation guidelines (v1): October 2014

- **10** treebanks: January 2015 🇮🇪
- **18** treebanks: May 2015
- **37** treebanks: November 2015
- **54** treebanks: May 2016
- **64** treebanks: November 2016

Release of annotation guidelines (v2): December 2016

- **70** treebanks (**50** languages) : March 2017
- **102** treebank (**60** languages) : November 2017
- **122** treebanks (**71** languages) : July 2018

universaldependencies.org



universaldependencies.org/#ga

English	254K	ⓁⒻ	🗎	👤	✓	CC BY SA	📅 📡 👍
English-ESL	97K	Ⓛ	🗎	👤	✓	CC BY SA	🗎
English-LinES	82K		🗎	⚙✓	✓	CC BY NC SA	📖 ℹ 💬
Estonian	234K	ⓁⒻ	–	⚙✓	✓	CC BY NC SA	📖 📰 🗎
Finnish	181K	ⓁⒻⒹ	🗎	⚙✓	✓	CC BY SA	📰 W 📅 🔨 📖 🪄
Finnish-FTB	159K	ⓁⒻ	–	⚙✓	✓	CC BY	🪄
French	390K	ⓁⒻ	🗎	⚙✓	✓	CC BY NC SA	📅 📰 👍 W
Galician	138K	Ⓛ	🗎	⚙✓	✓	CC BY NC SA	💉 🔨 ℹ 📰
German	293K		–	⚙	✓	CC BY NC SA	📰 👍 W
Gothic	56K	ⓁⒻ	–	⚙	✓	CC BY NC SA	☁
Greek	59K	ⓁⒻ	🗎	⚙	✓	CC BY NC SA	📰 W 💬
Hebrew	115K	Ⓕ	–	⚙	✓	CC BY NC SA	📰
Hindi	351K	ⓁⒻ	–	⚙	✓	CC BY NC SA	📰
Hungarian	42K	ⓁⒻ	🗎	👤	✓	CC BY NC SA	📰
Indonesian	121K		–	⚙	✓	CC BY NC SA	📰 📅
Irish	23K	ⓁⒻ	🗎	⚙✓	✓	CC BY SA	📰 📖 🌐 🔨 🗎

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

Italian	252K	ⓁⒻ	🗎	⚙✓	✓	CC BY NC SA	🔨 📰 W
Japanese-KTC	267K	Ⓛ	🗎	⚙	✓	CC BY SA	📰
Kazakh	4K	Ⓛ	🗎	👤	✓	CC BY SA	W 📖
Korean	–		–	–	–		🗎

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

TEDx

Conclusion

All this can lead to:

Understanding of **need** for language technology

#GRMA

Go raibh maith agaibh

Thank you (pl)