



# NLP for low-resourced languages

Teresa Lynn, PhD

*Research Fellow*

*ADAPT Centre*

*Dublin City University*

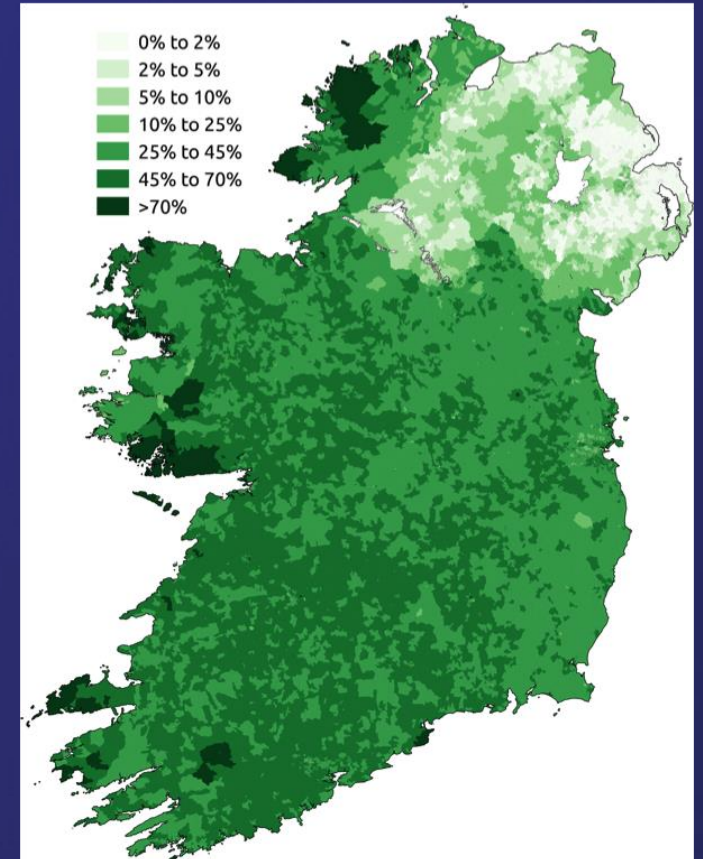


# AI Challenges for Low-resourced Languages

- Overview of The Irish Language
- NLP with few resources
- Addressing the Lack of Irish Data
- The Future?

# Irish language - status

- First Official Language
- National Language
- Census (2016): Pop. 4,761,865
- Ability to speak: 1,761,420
- Daily usage: 73,803



# EU Language status



- Official EU Language
- Minority Language (low-resourced)
- Derogation on official translations (until 2021)

# Morphology/ Inflection



## LENITION

sa **che**antar 'in the area'

airgead a **thu**illfeadh sé 'money he would earn'

a **dh**eartháir 'his brother'



## ECLIPSIS

Tír na **nÓ**g 'Land of the Youth'

i **m**Béarla 'in English'

ar an **m**bord 'on the table'



## VOWEL HARMONY

Caithim 'I spend'

Casaim 'I turn'

Rith**finn** 'I would run'

D'íos**fainn** 'I would eat'

# Inflected Prepositions



**le – with**  
*liom* `with **me**`  
*leat* `with **you**`



**ag – at**  
*agam* `at **me**`  
*agat* `at **you**`



**fai – 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**`

# Word Order



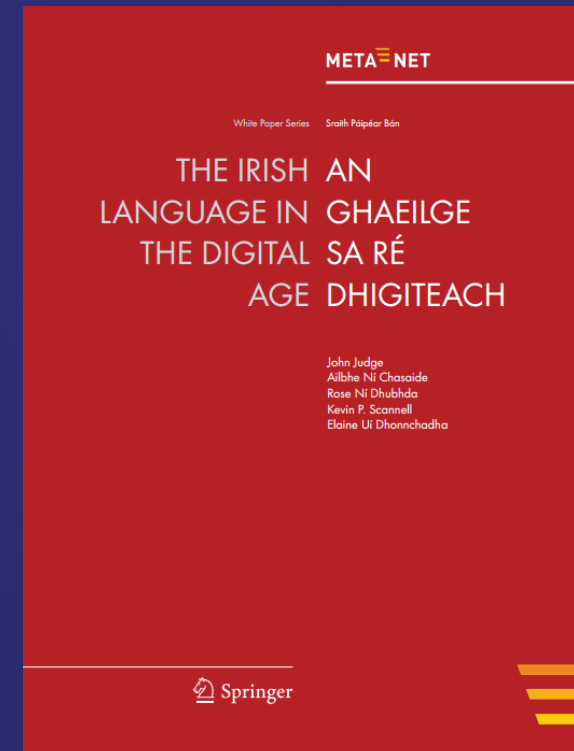
**English:** 'I saw the boy'

**Irish:** *Chonaic mé an buachaill*

**Gloss:** Saw I the boy

# Irish language technology

- META-NET white paper series (Judge et al., 2012)
- EU-led survey
- 31 EU languages
- Language resources and technologies





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, <b>Irish</b> , 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, <b>Irish</b> , 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, <b>Irish</b> , 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, <b>Irish</b> , Latvian, Lithuanian, Maltese, Welsh



# Risk of Digital Extinction



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

Will technology have the same impact on Irish?

# Risk of Digital Extinction

Need to ensure **continuing** language usage through technology

- Edutainment packages/ CALL
- Multi-platform Word processing tools
- Automated translation
- Search engines
- Games
- Social media/ Online data mining
- Text Generation (e.g. weather reports)
- Automatic subtitling
- ...



# Why do we need NLP?

○ TEXT  
SUMMARISATION

○ SENTIMENT  
ANALYSIS

○ INFORMATION  
RETRIEVAL

○ GRAMMAR CHECKING

○ RECOMMENDER SYSTEMS


○ TEXT MINING

○ MACHINE  
TRANSLATION

○ QUESTION-ANSWERING  
SYSTEMS

○ LANGUAGE LEARNING APPS

○ VIDEO SUMMARISATION

- 
- Overview of The Irish Language
  - **NLP with few resources**
  - Addressing the Lack of Irish Data
  - The Future?

# Why is NLP a hard task?

- One word/sentence may have many meanings
- Many ways of saying the same thing
- Meaning depends on context
- Literal and figurative language (metaphor)
- Language and culture  
(different ways of conceptualising the same thing)

# Ambiguous Headlines



## Syntactic Ambiguity

EYE DROPS OFF SHELF

SQUAD HELPS DOG BITE VICTIM

ENRAGED COW INJURES FARMER WITH AXE

STOLEN PAINTING FOUND BY TREE



## Semantic Ambiguity

PANDA MATING FAILS; VETERINARIAN TAKES OVER

SAFETY EXPERTS SAY SCHOOL BUS PASSENGERS SHOULD BE BELTED

POLICE BEGIN CAMPAIGN TO RUN DOWN JAYWALKERS

# What does a machine know about language?

character	binary code
1	0011 0001
2	0011 0010
3	0011 0011
4	0011 0100
...	...
A	0100 0001
B	0100 0010
C	0100 0011
D	0100 0100
...	...
a	0110 0001
b	0110 0010
c	0110 0011
d	0110 0100



# What does a machine know about language?

Sentence = a string/sequence of characters:

*“The man saw the boy with the telescope”*



# SYNTACTIC PARSING 101

Who is doing what? Who has the telescope?



## Part of Speech Tagging

*"The man saw the boy with the telescope"*  
DET NOUN VERB DET NOUN PREP DET NOUN

# “Traditional” Parsing

S → NP VP

S → NP VP PP

NP → Noun | Pronoun

VP → Verb NP | Verb PP

PP → Preposition Noun

Noun → ‘ice-cream’ | ‘summer’

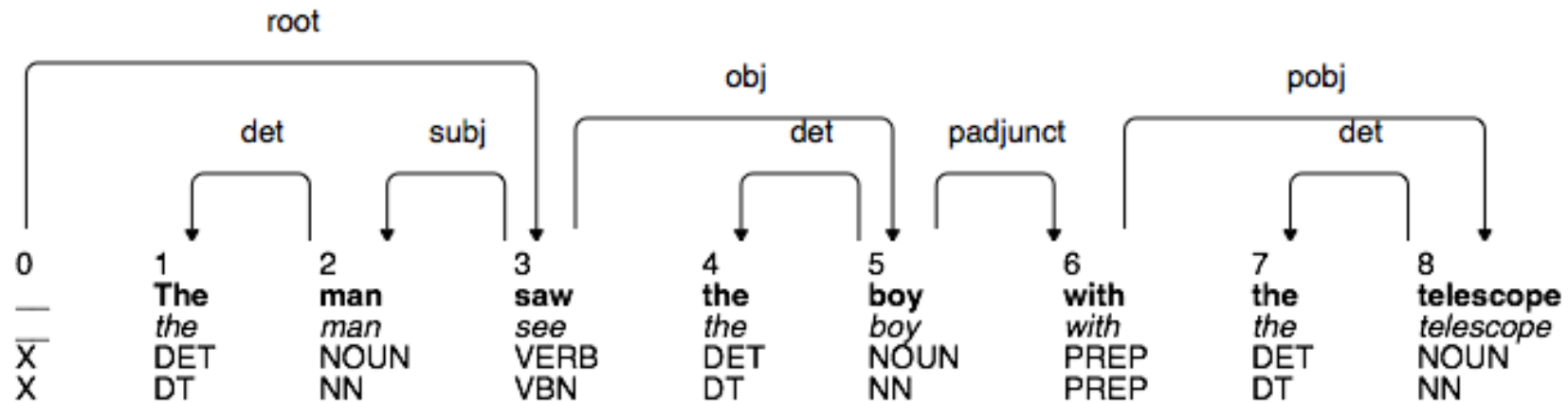
Pronoun → ‘I’

Verb → ‘like’

Preposition → ‘in’

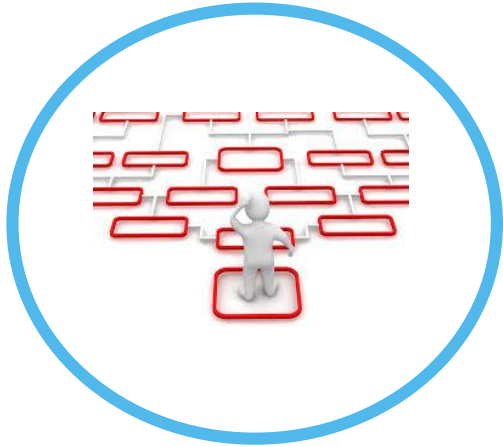


# STATISTICAL PARSING

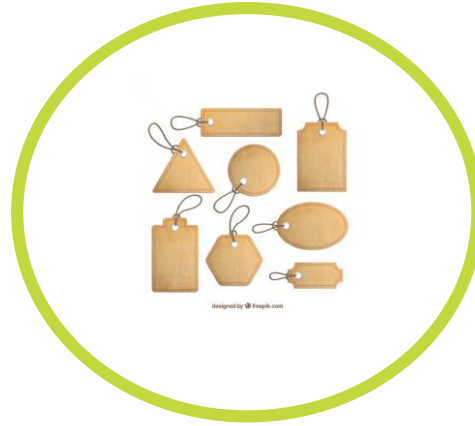


# Machine Learning in NLP

(data driven approaches)



STRUCTURED  
DATA



LABELLED  
DATA



RELIABLE  
DATA

# Machine Learning – data sparsity

what's the  
opposite of  
sparse?



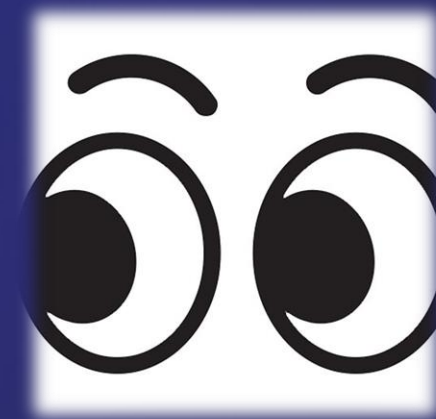
plentiful, abundant, sufficient,  
lush, fat, dense, frequent,  
adequate, enough, full





Engaging Content  
Engaging People

# Data Envy



**ISN'T IT WELL FOR YE?**





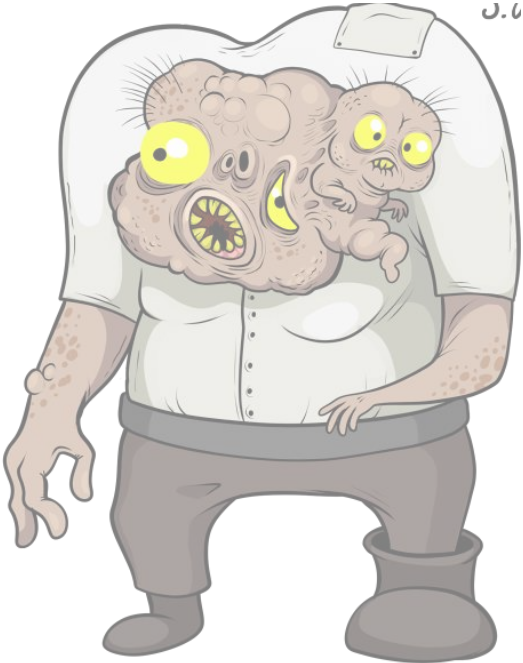
# Irish Data Sparsity



FUNDING




SKILL  
SHORTAGE



MORPHOLOGY



NUMBER OF  
SPEAKERS

- 
- Overview of The Irish Language
  - NLP with few resources
  - **Addressing the Lack of Irish Data**
  - The Future?

# Addressing the lack of data



CROSS-  
LINGUAL  
TRANSFER



BOOT-  
STRAPPING



SYNTHETIC  
DATA



TRAIN  
MORE  
EXPERTS

# CROSS-LINGUAL TRANSFER

- Using data from one language to help build a system for another

UNIVERSAL  
DEPENDENCIES



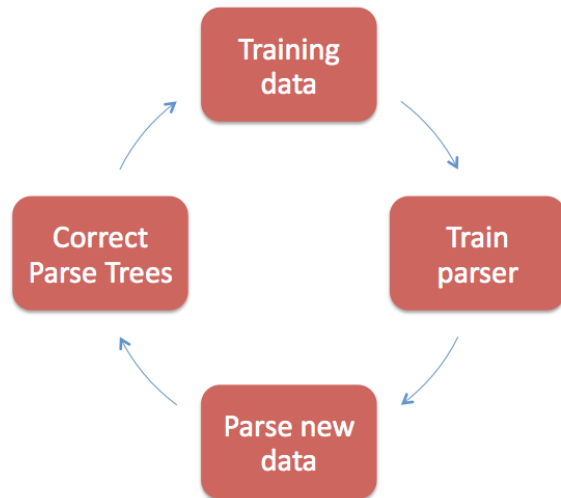
MULTI-WORD EXPRESSIONS

P A R S  M E

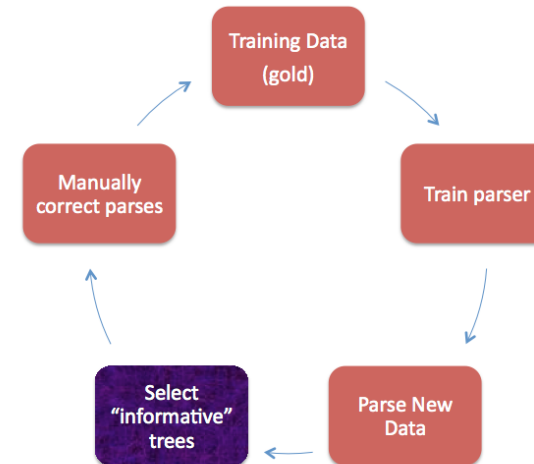
# BOOTSTRAPPING

- Using limited data to train a sub-standard system to help further annotations (human correction rather than annotate from scratch)

## PASSIVE LEARNING

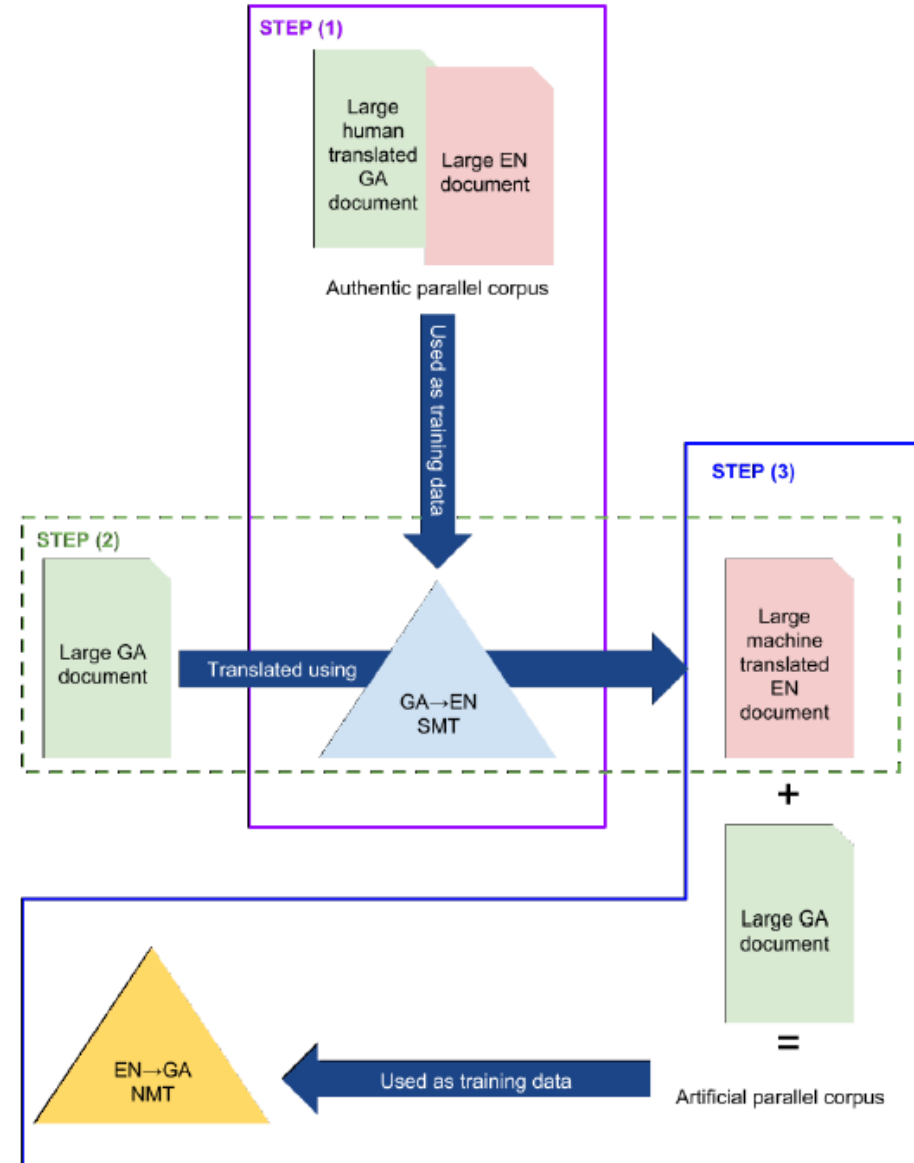


## ACTIVE LEARNING



# SYNTHETIC DATA

e.g. Back Translation for Machine Translation



# On that MT note.....



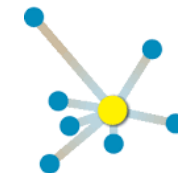
- Tapadóir SMT system (BLEU 46)
- SMT vs NMT (NMT BLEU 40)
- Domain-tuning, linguistic features (hybrid)
- Increasing data collection (European Language Resource Coordination)




An Roinn  
Cultúir, Oidhreachta agus Gaeltachta  

---

Department of  
Culture, Heritage and the Gaeltacht



European Language  
Resource Coordination  
*Connecting Europe Facility*

- 
- Overview of The Irish Language
  - NLP with few resources
  - Addressing the Lack of Irish Data
  - **The Future?**



# Digital Strategy for the Irish Language 2019

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)

# TRAINING MORE EXPERTS

- Machine Translation
- Irish Twitter Analysis
- Processing Irish Multiword Expressions
- Irish Syntactic Parsing



An Roinn  
Cultúir, Oidhreachta agus Gaeltachta  

---

Department of  
Culture, Heritage and the Gaeltacht



**Engaging Content**  
Engaging People

# Go Raibh Maith Agaibh

#GRMA

[teresa.lynn@adaptcentre.ie](mailto:teresa.lynn@adaptcentre.ie)



@cigilt