Engaging Content
Engaging People

An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media

A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

The challenges of working with low-resourced languages in NLP: An Irish story

Teresa Lynn, Dublin City University

HOST INSTITUTION
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

HOST INSTITUTION
DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

PARTNER INSTITUTIONS
TU Dublin
UCD Dublin
CIT Cork Institute of Technology
Institiúid Teicneolaíochta Chorcaí
AIT Connect & Discover
Maynooth University National University of Ireland Maynooth

FUNDED BY:
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund

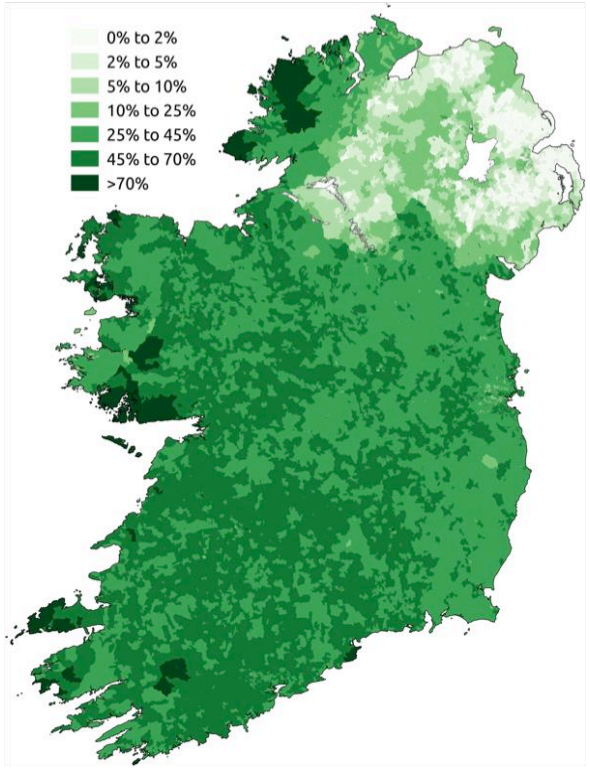
Talk Outline

- **Irish Language Technology - Overview**
- GaelTech Project
- Parsing Irish Tweets
- gaBERT
- European Language Equality Project

The Irish Language

- First official language
- National language

- Census (2016): pop. 4,761,865
- Ability to speak: 1,761,420
- Daily Usage: 73,803
- Colonialism



Engaging Content
Engaging People

Word Order

English: `I saw the boy'

Irish: *Chonaic mé an buachaill*

Gloss: Saw I the boy

Morphology/ Inflection

LENITION

*sa **c**heantar* ‘in the area’

*airgead a **th**uillfeadh sé*

‘money he would earn’

*a **dh**eartháir* ‘his brother’

VOWEL HARMONY

Caith**im** `I spend’

Cas**aim** `I turn’

Rith**finn** `I would run’

D’íos**fainn** `I would eat’

ECLIPSIS

Tír na **n**Óg ‘Land of the Youth’

i **m**Béarla ‘in English’

ar an **m**bord ‘on the table’

Engaging Content
Engaging People

There is no “yes” or “no”

Do you **approve** of the proposal to amend the constitution?

vs

Are you consen**ting** to the proposal to amend the constitution?

A Low-resourced Language

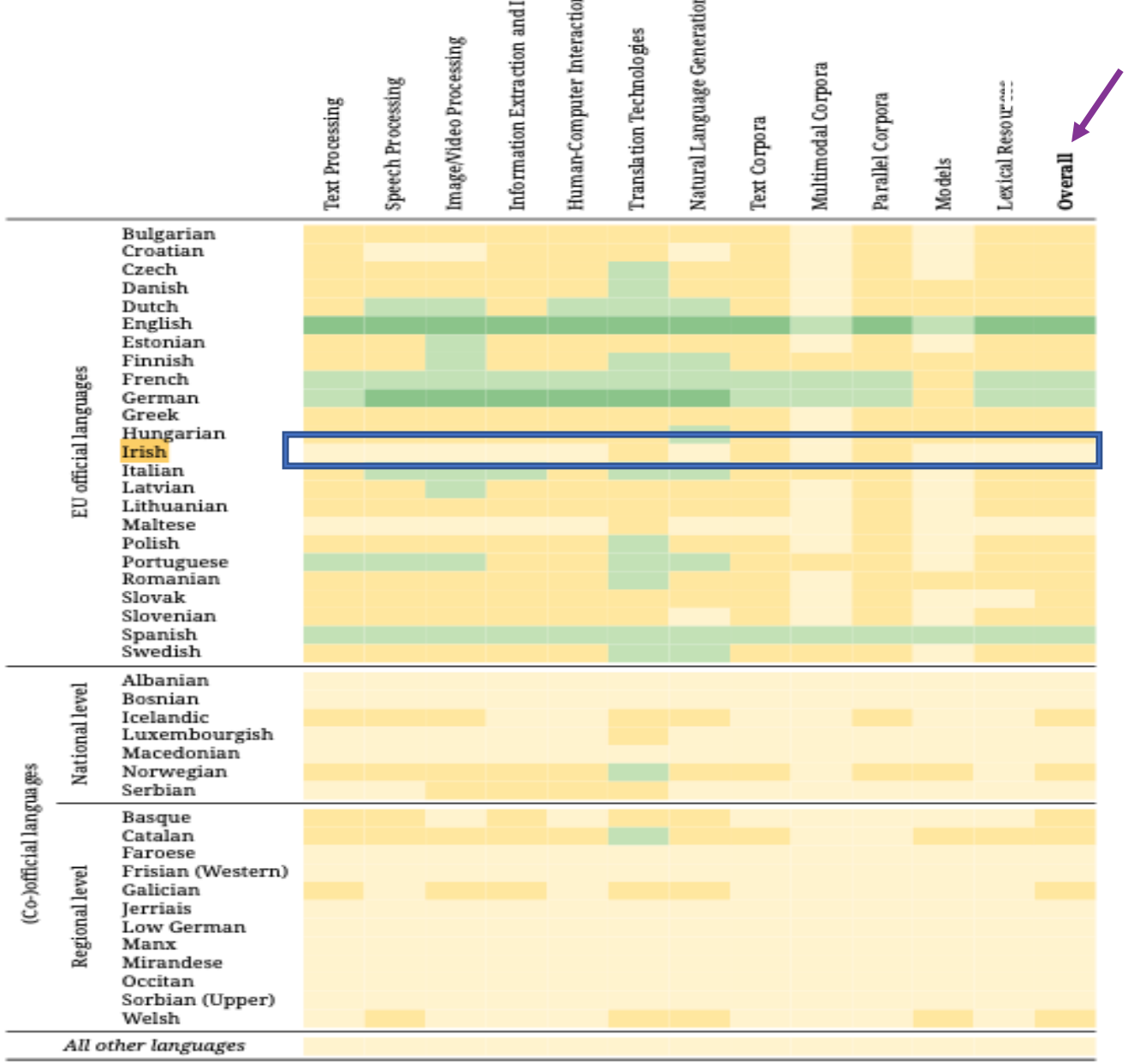


Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

EUROPEAN LANGUAGE EQUALITY

Teresa Lynn. 2022. Report on the Irish language. Technical Report D1.20, European Language Equality. https://european-language-equality. eu/deliverables/.

Irish – A Low-resourced Language

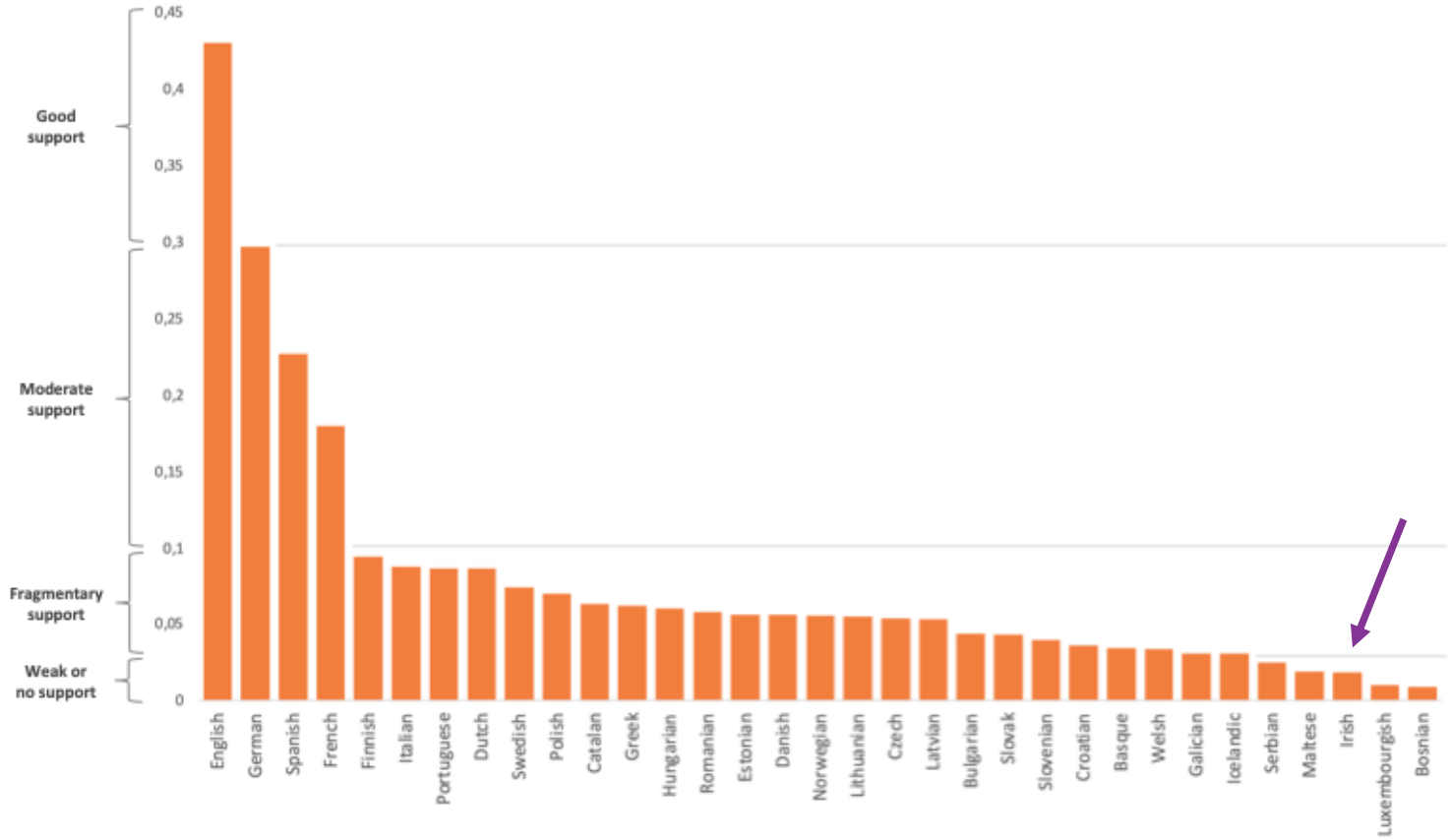


Figure 1: Overall state of technology support for selected European languages (2022)

Teresa Lynn. 2022. Report on the Irish language. Technical Report D1.20, European Language Equality. https://european-language-equality. eu/deliverables/.

Risk of Digital Extinction

"The Printing Press resulted in the extinction of many Regional and Minority Languages"

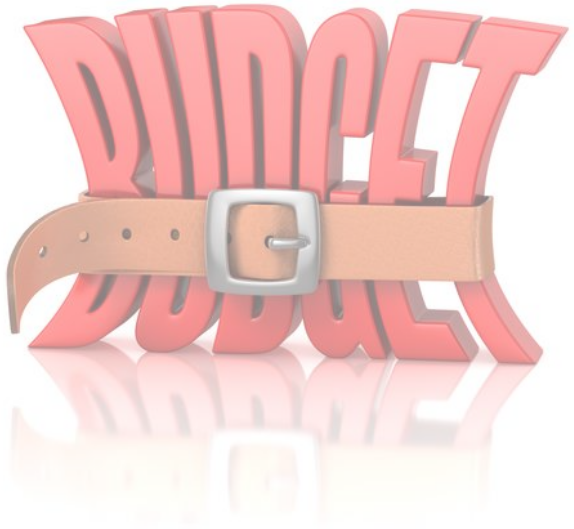
Will technology have the same impact on Irish?

Machine Learning

"Data-driven": learning patterns in language

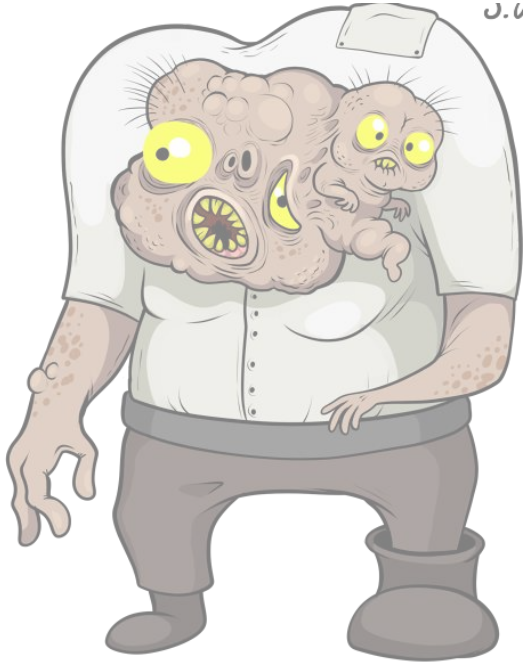
Latest AI techniques @ ADAPT

What does low-resourced mean?



FUNDING

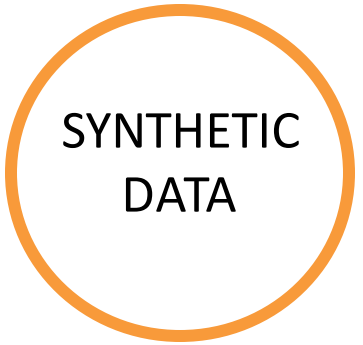
SKILL SHORTAGE



MORPHOLOGY

NUMBER OF SPEAKERS

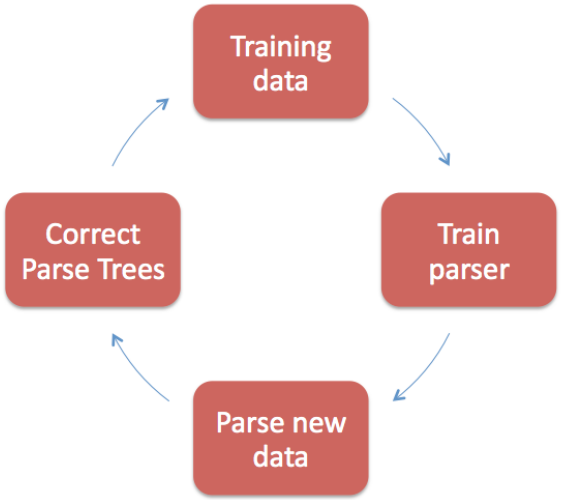
Addressing the lack of data



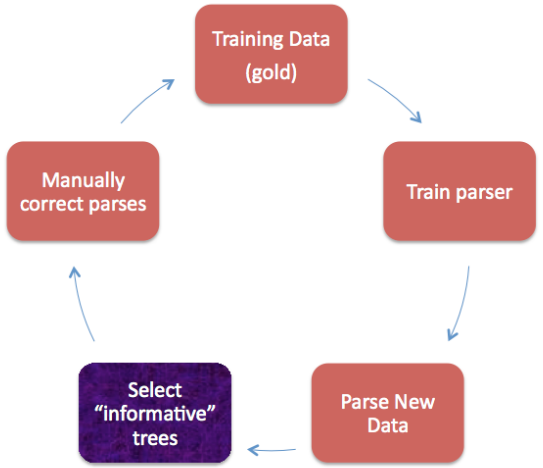
Bootstrapping treebank development

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

PASSIVE LEARNING



ACTIVE LEARNING



Leverage expertise in international projects

Universal Dependencies (treebanks)

European Language Resource Coordination
Connecting Europe Facility

MT data collection

PARSEME

Multiword Expressions

Recent progress on resource development

- **Machine translation**:
  - Involvement in the European Language Resource Coordination (ELRC)
  - CEF-funded data-collection projects (ELRI, PRINCIPLE)
  - Open Data Directive
  - Science Foundation Ireland – funding 2 PhDs in Irish MT at Dublin City University
  - Irish now a working EU language (eTranslation more important than ever)
- **Speech Technologies**
  - Steps towards ASR (data collection) at Trinity College Dublin
  - Integration of TTS into CALL and accessibility aids (e.g. screen-reader)
- New **national corpus** under development (with NLP in mind!)
- **GaelTech** project data-driven POS-tagging, parsing, treebanks, MWE-processing, gaBERT

Talk Outline

- Irish Language Technology - Overview
- **GaelTech Project**
- Parsing Irish Tweets
- gaBERT
- European Language Equality Project

2017-2023 @ €735,000

An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media

3 Research Strands:

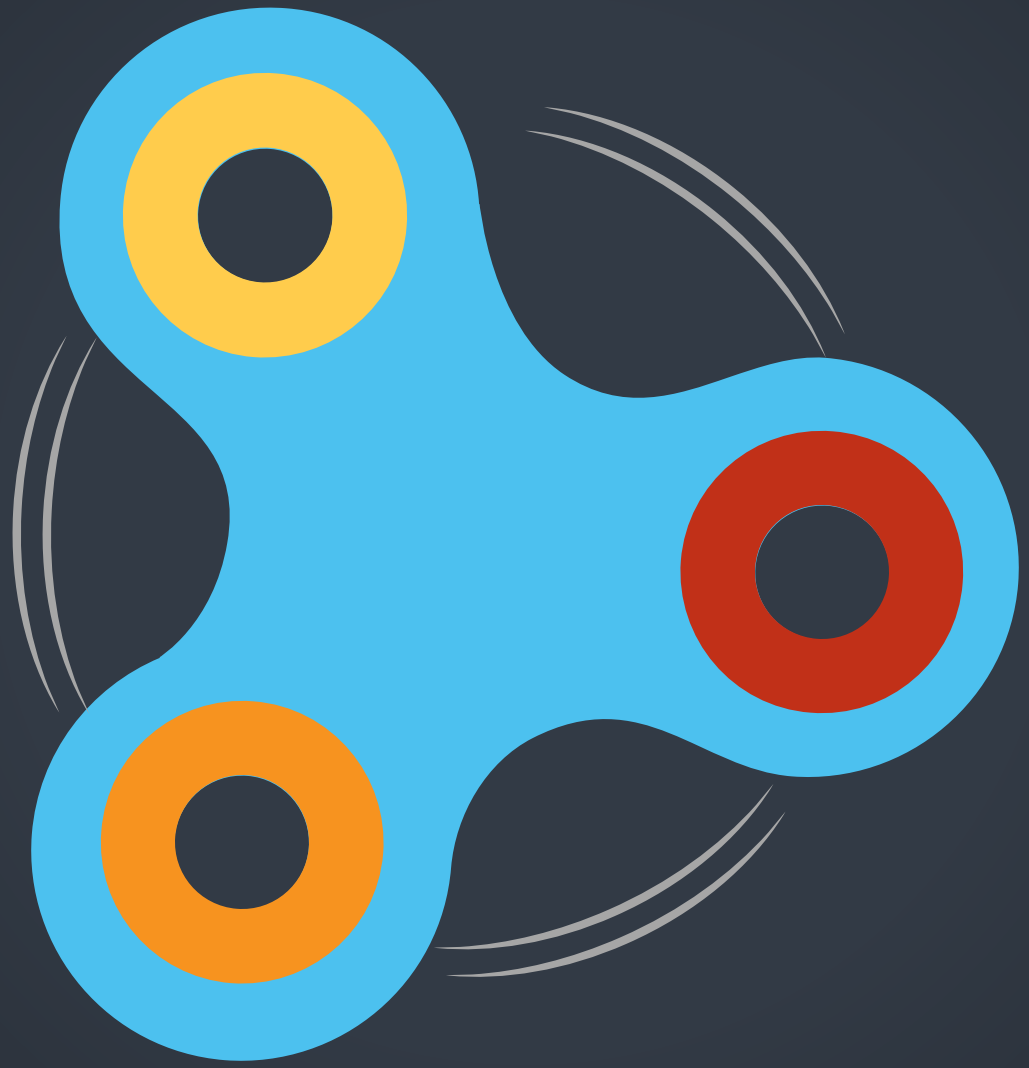
- Treebank Development (Parsing)
- Automatic Processing of Irish Multiword Expressions
- NLP for Irish User-Generated Content

GaelTech

GaelTech Project Design

Syntactic Parsing

Treebank Development
POS-tagging & morphological information
Syntax information
Multiword Expressions (MWEs)
Annotation guidelines
Parsing models
gaBERT language model



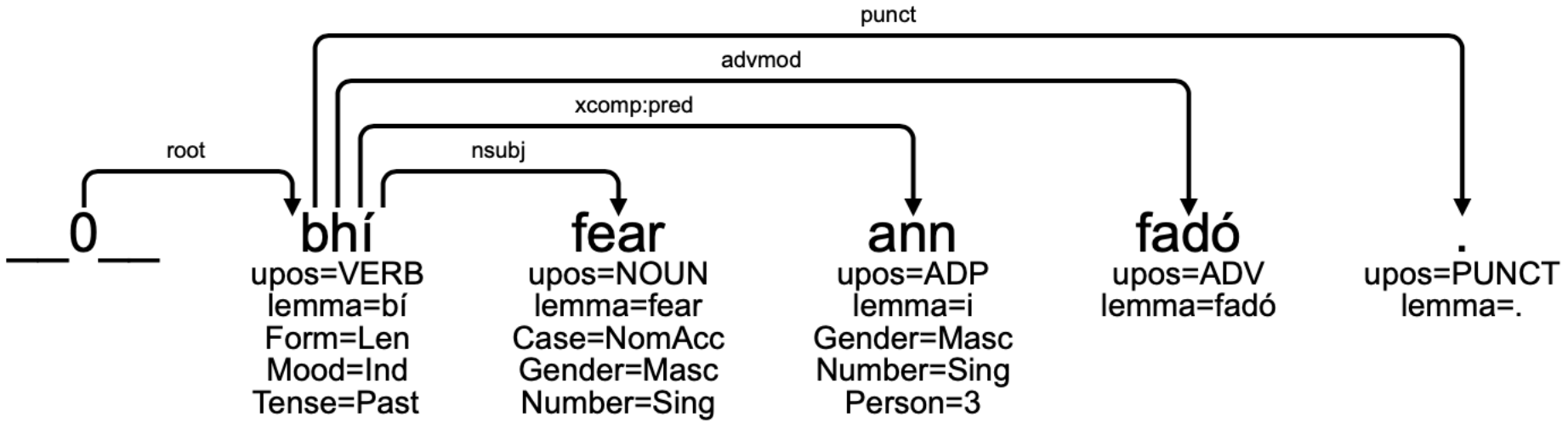
MWE Processing

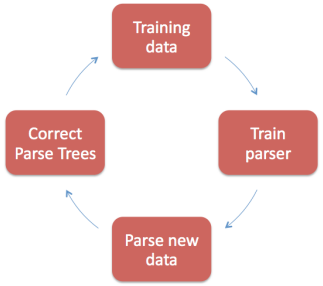
Fundamental Irish MWE research
MWE lexicon
Annotated corpora (test data)
Machine Translation evaluation
Automatic MWE identifier

NLP for UGC

Data collection
Treebank development
POS-tagging, morphological information
Syntax information
Annotation Guidelines
POS-tagging models
Parsing models
Code-switching processing

1. Treebank Expansion and Parser Improvement





Named Entities

github.com/UniversalDependencies/UD_Irish-IDT
match.grew.fr

2. Irish Multiword Expressions – Abigail Walsh

Idiom	*Gearraíonn beirt bóthar* 'Two shorten the road'
Copular Construction	*Is maith liom* 'I like'
Verb Particle Construction	*Tabhair amach* 'Give out'
Inherently Adpositional Verbs	*Éirigh le* 'Get on with'
Light Verb Constructions	*Déan dearmad* 'Forget'
Compound Nouns	*Madra rua* 'fox'
Compound Prepositions	*In aice* 'beside'

2. Irish Multiword Expressions

Engaging Content
Engaging People

Perspective
Sentence
page: 1
Selector
Automatic (deepest)

Legend • Entity
(Hide)
- VID
- LVC.full
- {optional} IAV
- LVC.cause
- IRV
- VPC.full
- VPC.semi

Ar leath i ndiaidh a seacht ar maidin a caitheadh anuas an bhonnóg.

bhí fear ann fadó.

'Cad chuige nach n-abrófá féin staic d'amhrán dúinn, is gur tú an fear ceoil is fearr ar an mbaile?

Uaisle saibhre de chuid na Róimhe a bhí ina magistri tráth a bhíodh ann.

Agus nuair a mhínigh mé dó seo go mba i dtaobh poist a bhí mé ag caint leis siúd, dúirt sé lena mheangadh milis nach raibh aon fholúntas ann.

AN FILE i gcead do George Mackay Brown.

An rud a theastaigh ó Bhreandán agus uaim féin ná post buan ollscoile a chur ar fáil, agus faraor, de bharr easpa airgeadais, níor éirigh leis an bplean ag an am.

Ar ith sí an dinnéar?

Tá ceist eile ag an Scéaltacht.

Is fíor gur láidre Fianna Fáil ag cosaint na neodra
ceisteanna a chuireann a cheannairí faoin bpolasa

Sliogáin Eile Faoin am seo tá sé tugtha faoi deara agat, is dócha, gur sliogéisc cuid mhaith de na hainmhithe beaga a fheiceann tú ar an gcladach.

Le bliain nó dhó anuas amháin a thosaigh sé ag tabhairt na difríochta faoi deara.

ART: Céard tá i gceist agat?

Bhí gá leis an gcogadh sin cé gur trua i gcónaí nuair nach mbíonn aon bhealach eile ann chun saoirse a bhaint amach seachas an

Annotation Editor • Word/Token
ga_ud-idt_100_test.text.1.s.18.w.15

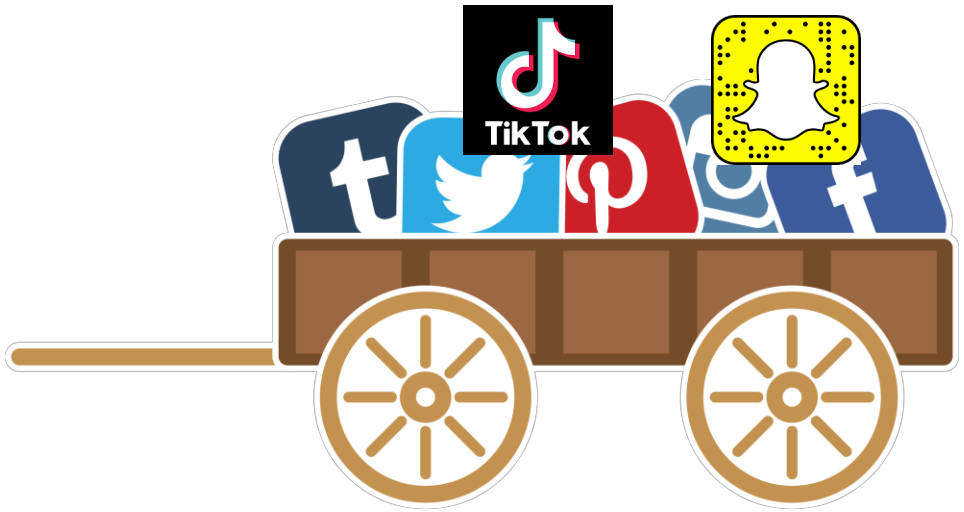
Entity
https://github.com/proycon/parseval/support/raw/master/parseme-mwe-alllanguages2018.foliaset.xml

chur ar fáil
LVC.cause
Select span>
D N
confidence: (not set)

Queue for later submission
Repeat this annotation for the next target
Open console window after submission

Ok

Irish User-generated Content – Lauren Cassidy



Talk Outline

- Irish Language Technology - Overview
- GaelTech Project
- **Parsing Irish Tweets**
- gaBERT
- European Language Equality Project

TwittIrish

- A Universal Dependencies treebank of Irish tweets
- Why are we interested in annotating tweets?
  - As a way of capturing the kind of informal Irish that is used on social media
    (Provides a basis for linguistic and sociolinguistic studies)
  - Standard parsers have been shown to perform poorly on noisy UGC
    - (e.g. Foster et al., 2011, Seddah et al., 2012)
  - Part of our ongoing effort to develop Irish language technology to help address the risk of digital extinction

Where do the tweets come from?

- The Indigenous Tweets Project INDIGENOUS TWEETS.COM

- 700 (out of 1,500) tweets sampled from 2006-2014
  - Manually POS-tagged (Lynn and Scannell (2015)) and mapped to UD

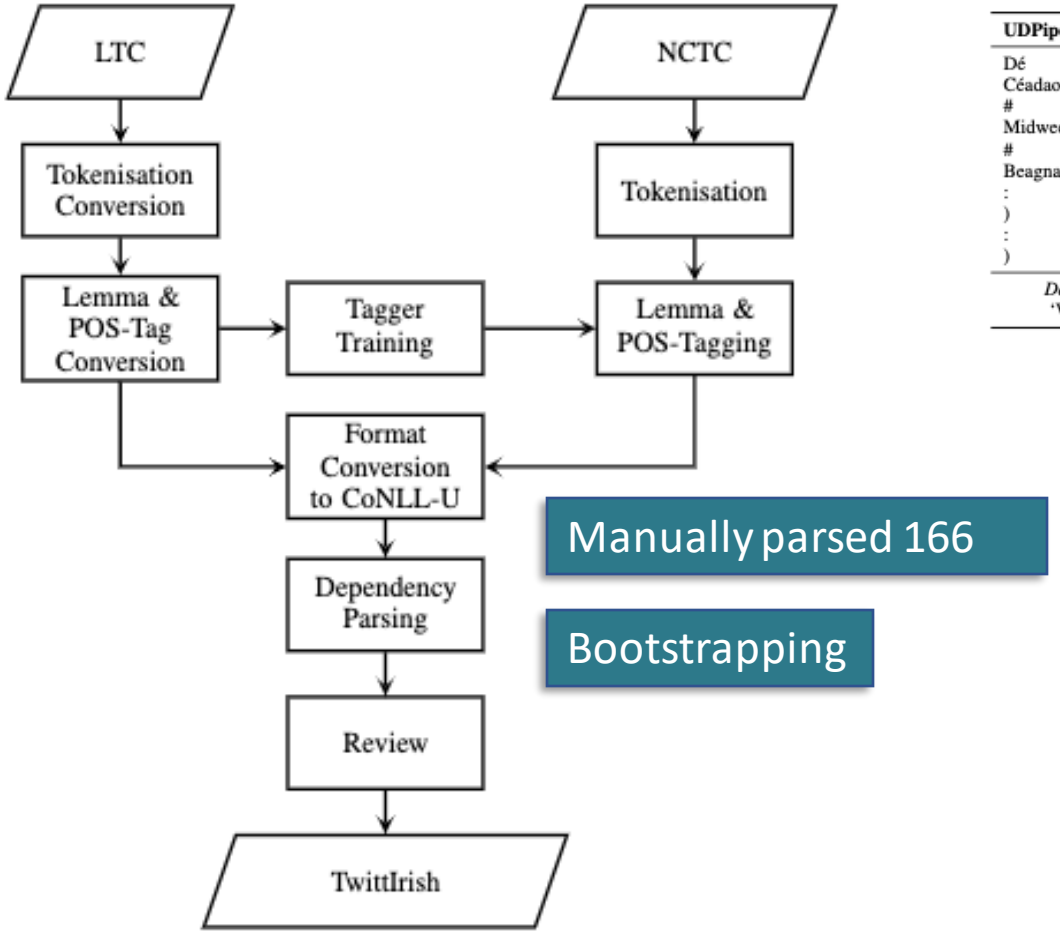
- A further 166 recent tweets sampled from 2010-2019
  - Automatically lemmatised & POS-tagged with Morfette

- **Result:** TwittIrish **test set** released in UD v2.8

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland.

Building the TwittIrish treebank

MWE: *in_aice* → *in aice*
Symbols: *5pm* → *5 pm*
10% → *10 %*



Manually parsed 166

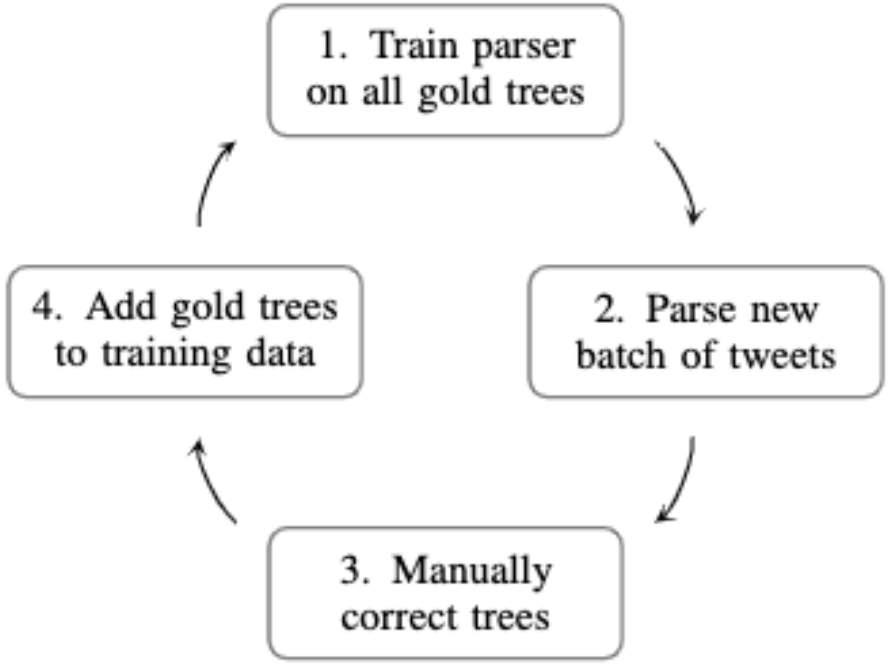
Bootstrapping

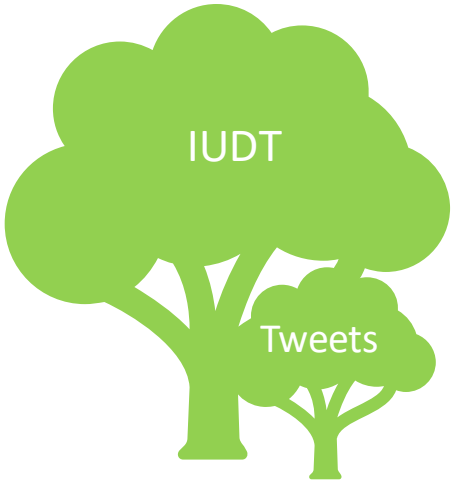
UDPipe (IUDT)	NLTK Tweettokenizer
Dé	Dé
Céadaoin	Céadaoin
#	#Midweek
Midweek	
#	#Beagnachann
Beagnachann	
:	:)
)	
:	:)
)	

Dé Céadaoin #Midweek #Beagnachann :) :)
'Wednesday #Midweek #Almostthere :) :)'

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland.

Bootstrapping approach to annotation



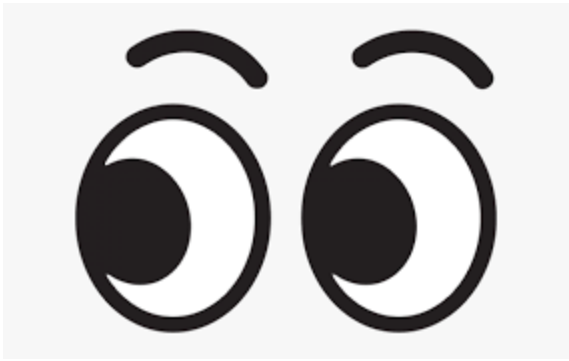


Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland.

What is in the Data?

Examined deviation from standard text along the following lines:

1. Orthographic
2. Lexical
3. Syntactic



Concurrent with UD-UGC study

2019-2021
Extensive study by several UGC treebankers representing various treebank language groups:

- Irish
- French
- English
- Italian
- Turkish
- German

Refer to:

Sanguinetti, M., Bosco, C., Cassidy, L. *et al.* Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Lang Resources & Evaluation* (2022).

Orthographic variation

2.5% of tokens exhibit orthographic differences to standard text:

Variation type	Example	Correct form	Gloss
Diacritic drop	***Leacht*** *faoi stair Príosún Dún Dealgain*	***Léacht*** *faoi stair Príosún Dún Dealgain*	`Liquid* (Lecture) about the history of Dundalk Prison'
Abbreviation	*sa bhaile an* ***tseacht*** *seo*	*sa bhaile an* ***tseachtain*** *seo*	`at home this week'
Lengthening	*Tá siad go léir* ***buuuuuuî***	*Tá siad go léir* ***buî***	`they are all yellow'

Orthographic variation

Variation type	Example	Gloss
Casing	*ach in aon áit **AR DOMHAIN***	*`but anywhere ON EARTH'*
Punctuation	*sin a dhóthain-)*	*`that's enough :-)'*
Transliteration	*Fair **plé** daoibh* ♥	`Fair ***play*** to you' ♥
Spelling Variation	***O'B**ama*	Obama

Lexical Variation

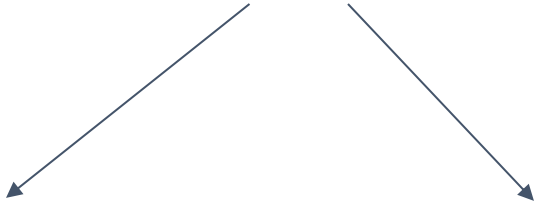
Only 38% of unique lemmata overlap with IUDT

Variation type	Example	Standard form	Gloss
Dialect	**fé**	faoi	`under’
Initialism	**BÁC**	Baile Átha Cliath	`Dublin’
Pictogram	**<3** mór	**grá** mór	`lots of **love**’
Truncation	Thart fa ‘53 **nó….**	Thart fa ‘53 **nóiméad**	`Over 53 **min**…. (minutes)’
Hypercorrection	i ndiaidh **concise** mór	i ndiaidh **coicíse** mór	`After a big **fortnight**’

Code-switching

- 75% of tweets are completely in Irish, 25% are bi- or multi-lingual.
- 67% of tokens in Irish, 5% are in English

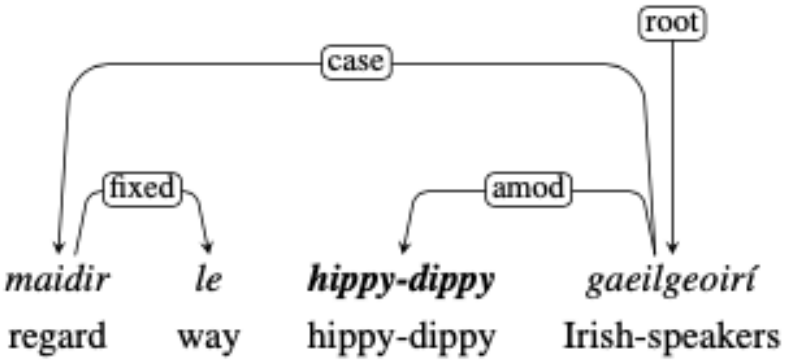
Remaining are neither or both:



Punctuation, meta language tags, words in other languages

Intraword code-switching or nonce borrowing (e.g. *happenáil* `happening')

Code-switching – syntactic variation



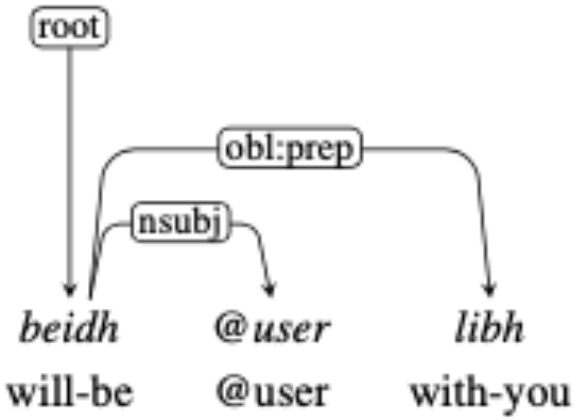
`As for hippy-dippy Irish-speakers…'

Intra-sentential CS

Inter-sentential CS

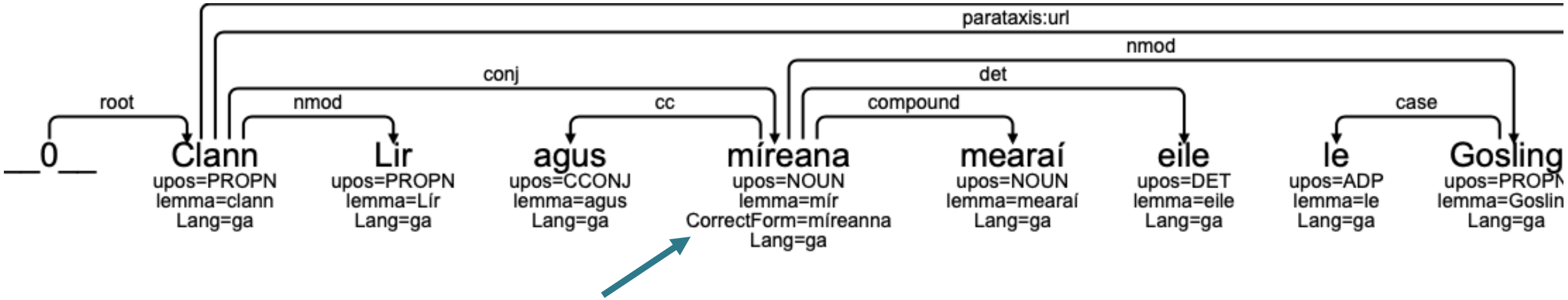
Syntactic Variation

- Metatags



`@user will be with you’

How to account for typos, non-standard forms, etc.?



MISC column

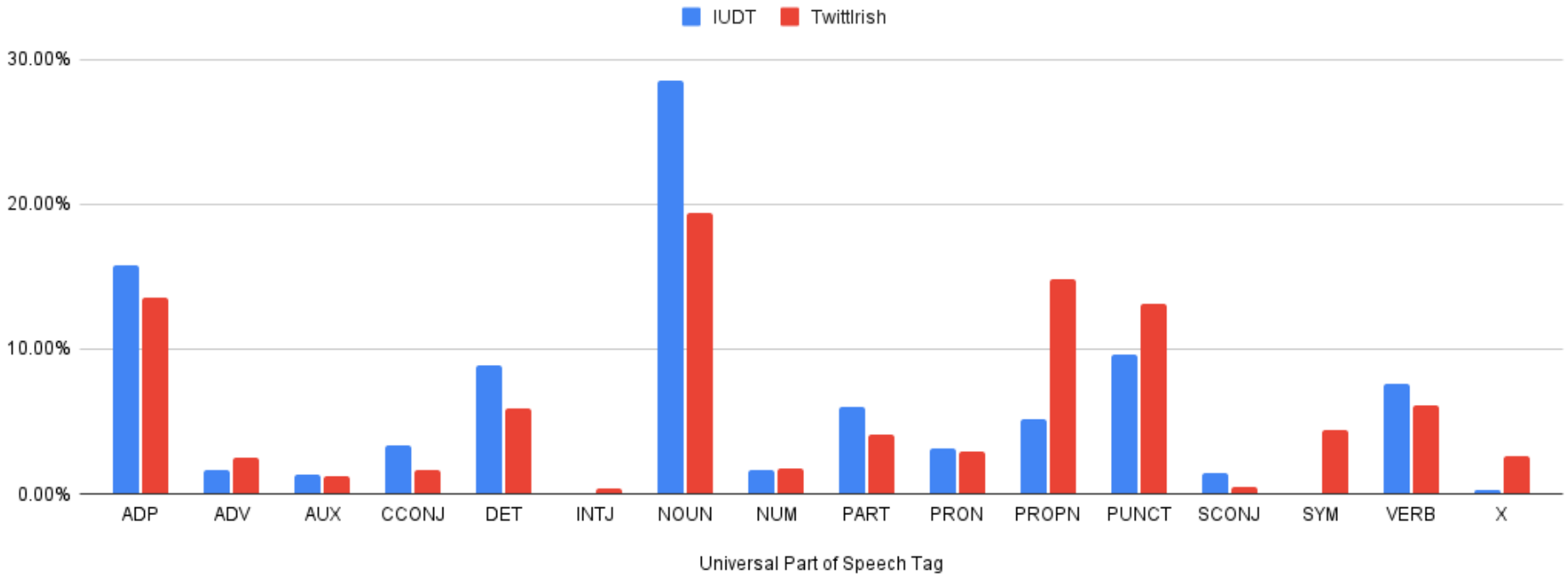
NonCan=X
AutoC: autocorrection **CharOm**: character omission, **Cont**: contraction, **Neo**: neologism, **OS**: over-splitting, **Phon**: phonetization, **PunctVar**: punctuation variation, **SpellVar**: spelling variation, **Stretch**: graphemic stretching, **Transl**: transliteration, **Trunc**: truncation.

CorrectForm=X, **FullForm**=X, **CorrectSpaceAfter**=**Yes** (cases of non-canonical language, abbreviations and incorrectly merged words respectively)

Lang=X (code-switching)

Refer to:
Sanguinetti, M., Bosco, C., Cassidy, L. *et al.*
Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Lang Resources & Evaluation* (2022).

Differences in UPOS distribution: TwittIrish vs IUDT



Similar distribution to English (Liu et al., 2018) and Italian (Sanguinetti et al., 2018)

- Symbols, punctuation, and proper nouns are more frequent in tweets
- Nouns, determiners, and prepositions are more frequent in standard text

Parsing Experiments

Experimental setup:

Training	Test
IUDT (train) 4005 trees	**IUDT (test)** 454 trees
	TwittIrish 866 trees

Parsers

- UDPipe (v1) – transition-based parser

Straka et al. (2016).
UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.

- AllenNLP – graph-based parser
  (Biaffine dependency parser with BiLTSM encoder)

Gardner et al., (2018).
AllenNLP: A Deep Semantic Natural Language Processing Platform.

Results

System	LAS IUDT	LAS TwittIrish
UDPipe v1	70.58	47.33
AllenNLP	71.56	48.73

- Significant **drop in performance** when tested on **tweets**
- Drop consistent for both parsers

What about adding contextualised word embeddings?

System	LAS IUDT	LAS TwittIrish
UDPipe v1	70.58	47.33
AllenNLP	71.56	48.73
AllenNLP + gaBERT	**84.25**	**59.34**

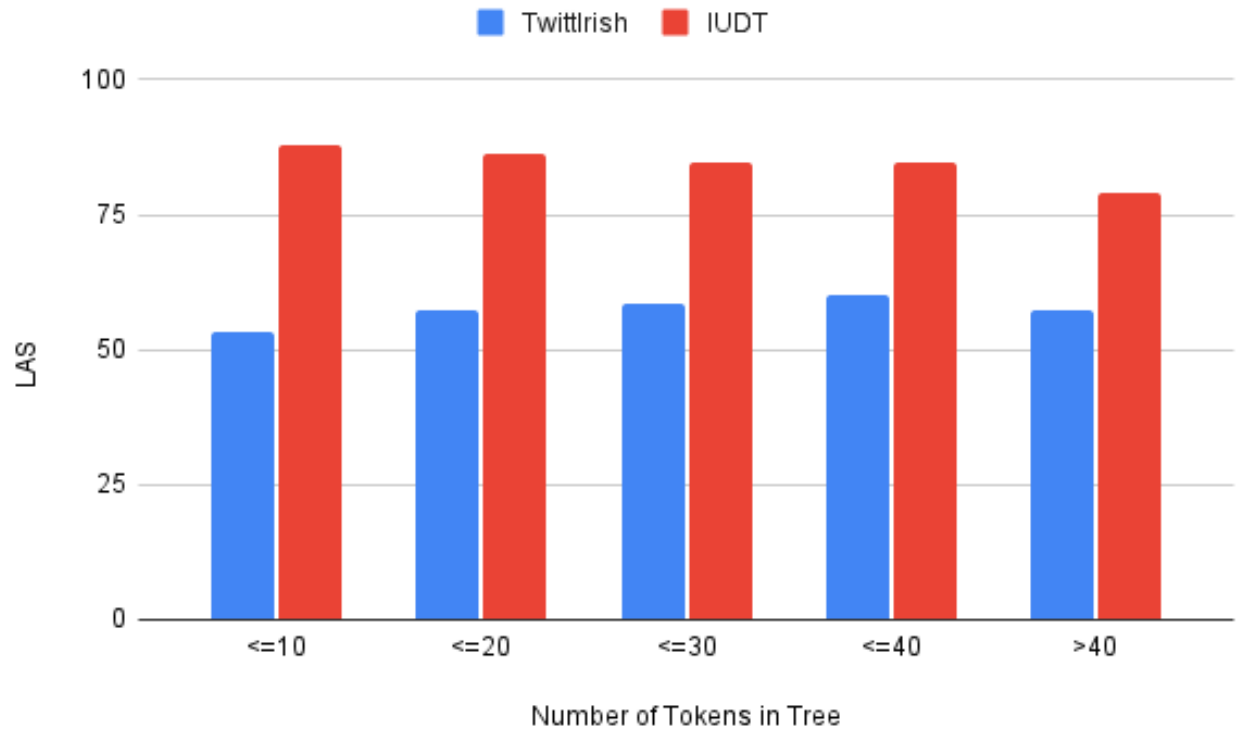
- A big improvement for both IUDT and TwittIrish (10+ points)
- BUT the gap between the two remains

Analysis

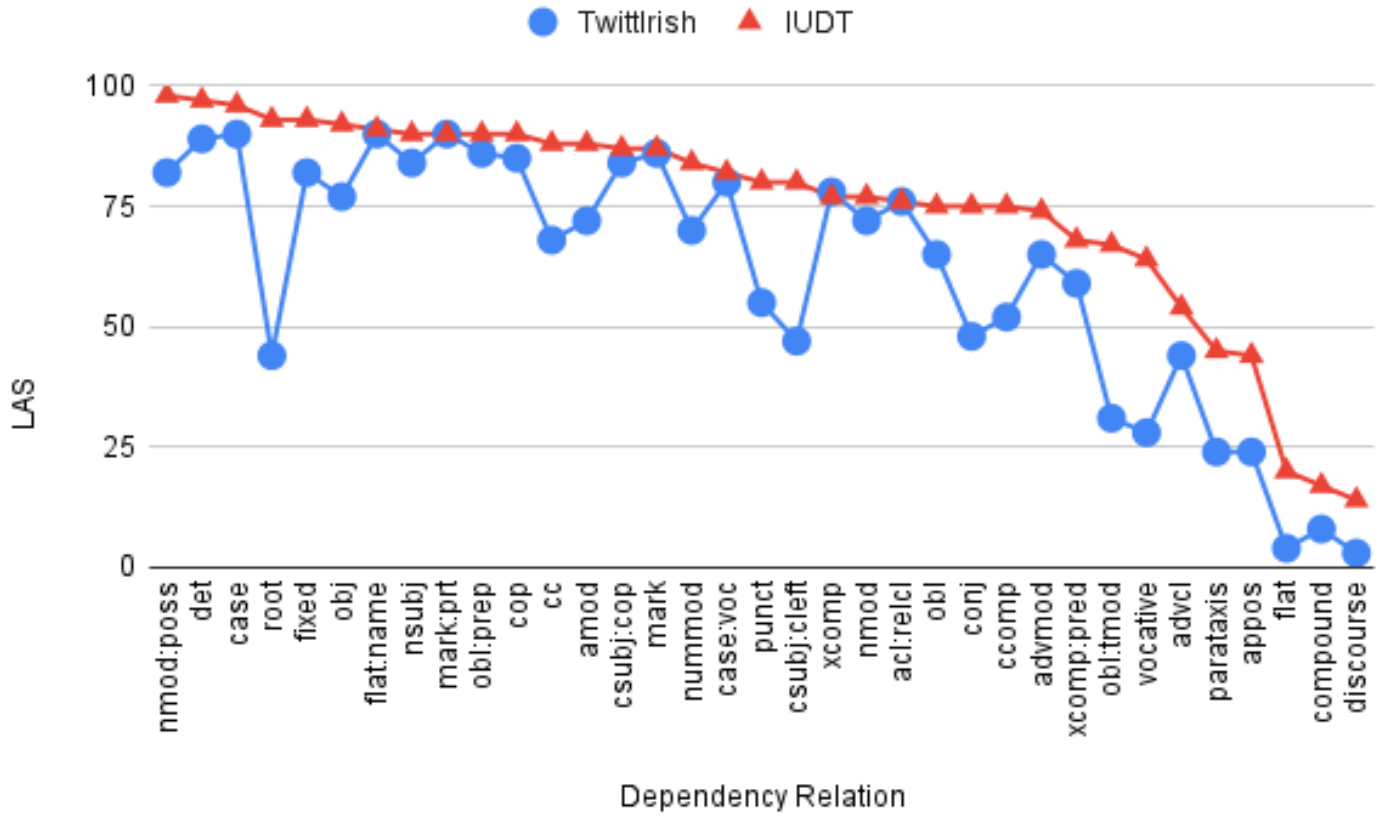
- Analysis based on the best model (AllenNLP parser with gaBERT embeddings)

  … using Dependable (Choi et al., 2015).

LAS by length



LAS by deprel



Error Analysis

Examine two extremes:

- LAS between 0 and 5
- LAS between 95 and 100

Phenomenon	Easiest Tweets	Hardest Tweets
Emoji	0	15
English Token	1	9
Username	3	10
Ellipsis	2	5
Hashtag	1	3
RT	0	3
URL	0	3
Spelling variation	2	2

Talk Outline

- Irish Language Technology - Overview
- GaelTech Project
- Parsing Irish Tweets
- **gaBERT**
- European Language Equality Project

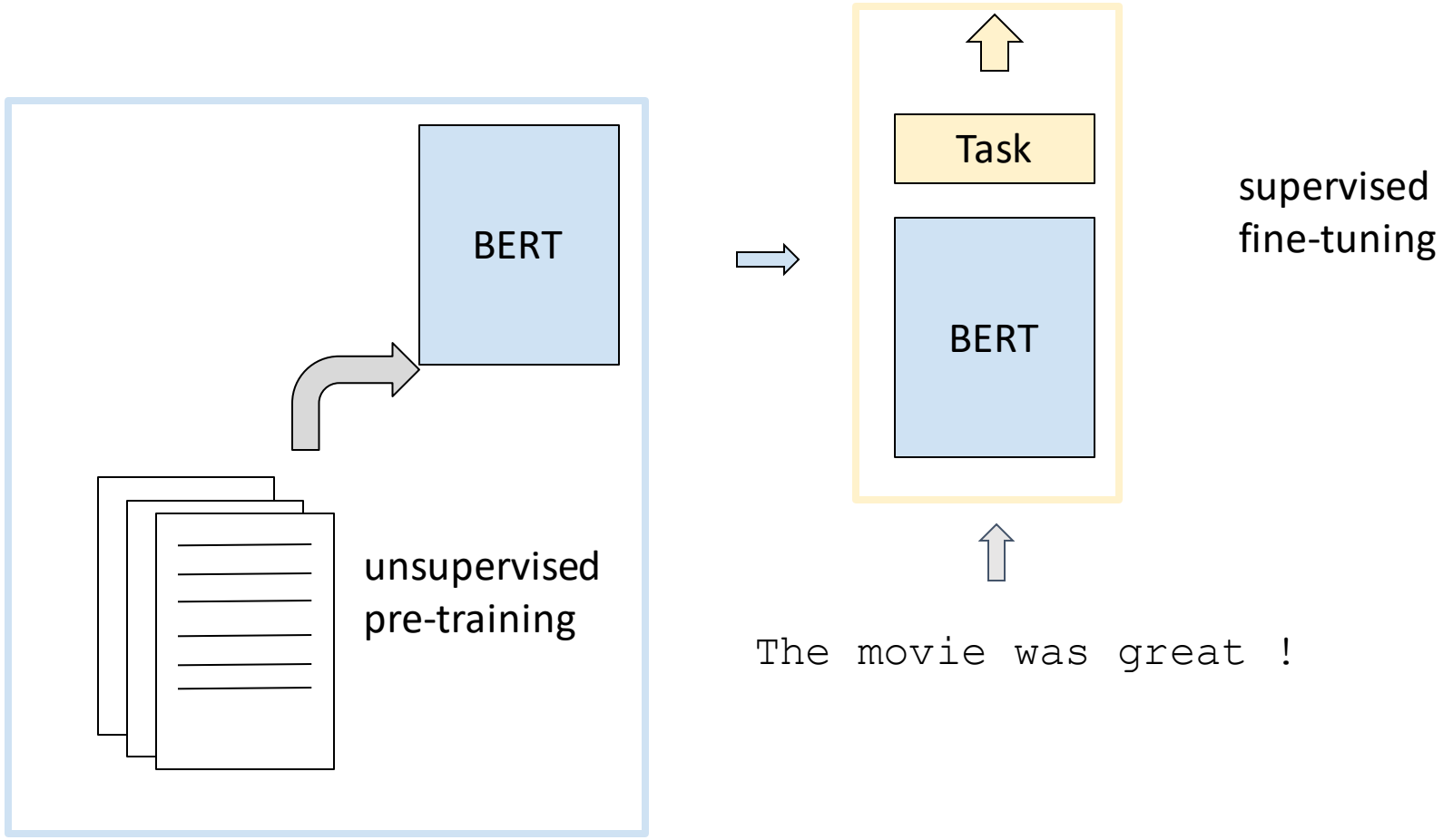
gaBERT

- A monolingual BERT model for Irish
  - **Language Models** (LMs) are used in NLP to learn representations of words

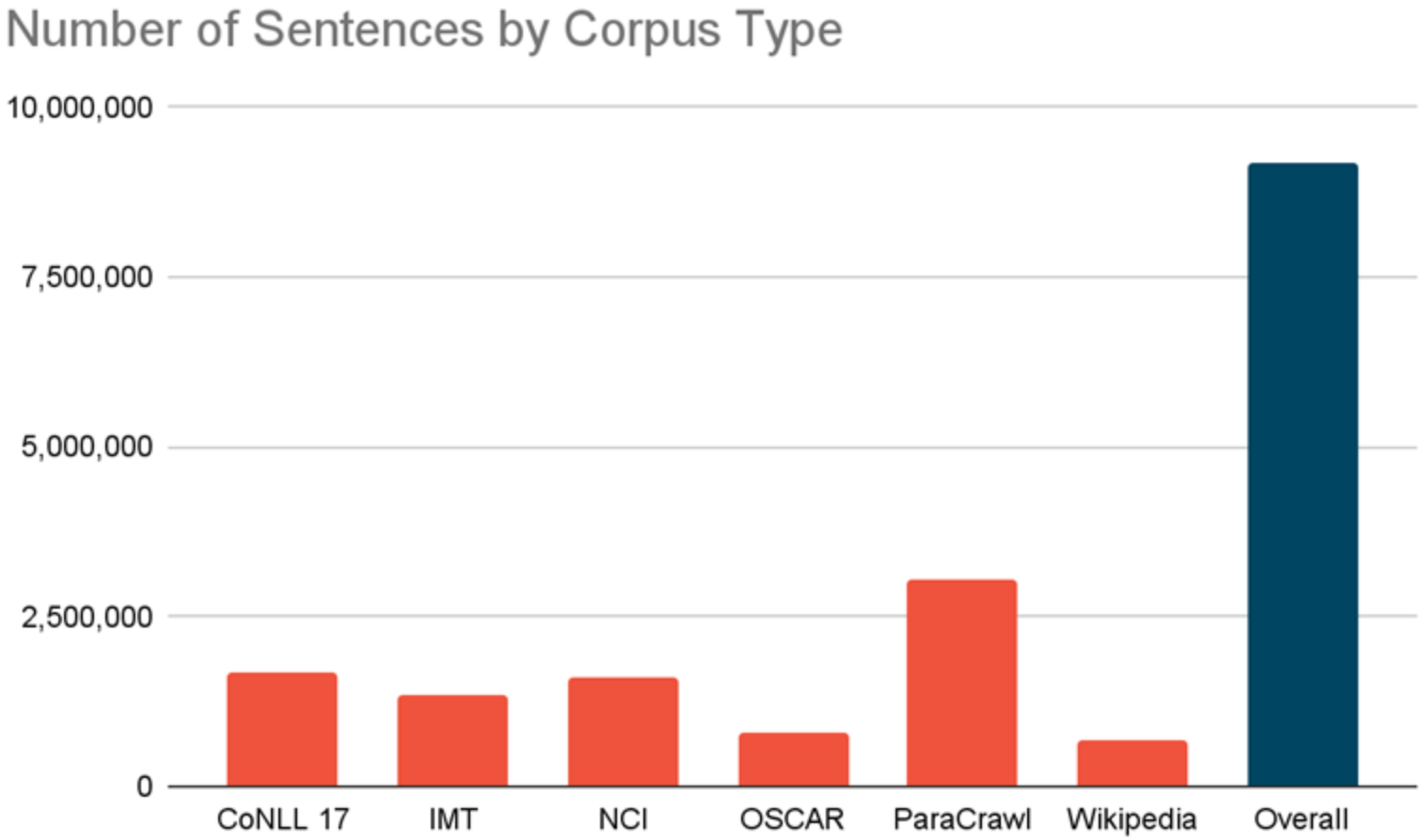
James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair and Jennifer Foster, **gaBERT — an Irish Language Model**, *In Proceedings of the 13th Language Resources and Evaluation Conference,* Marseille, France, June 2022

Language Models

Pre-train, then fine-tune



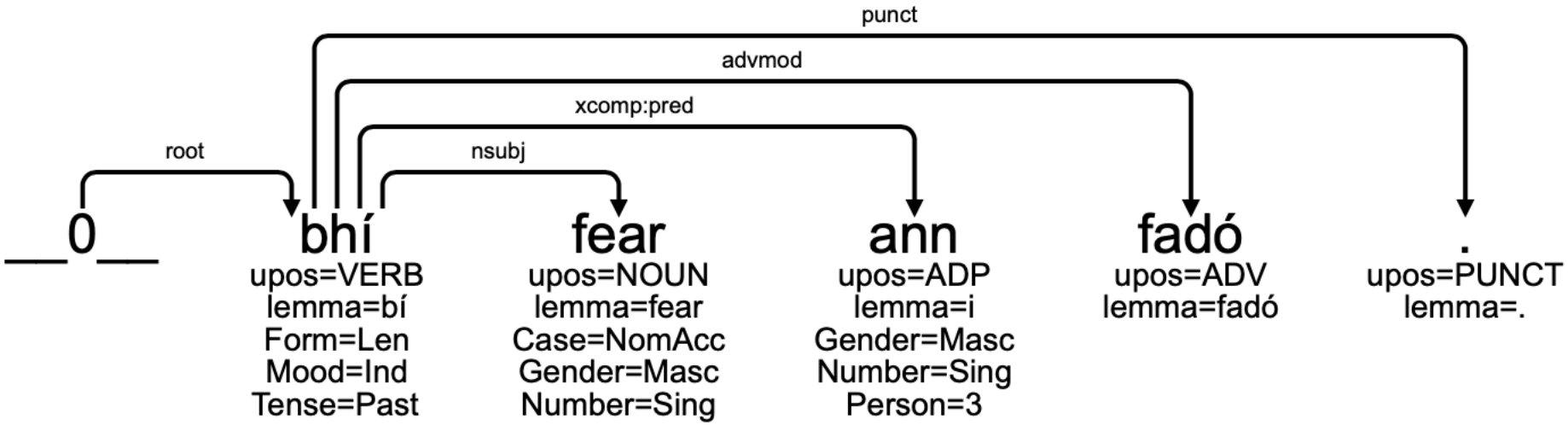
Pre-training Data



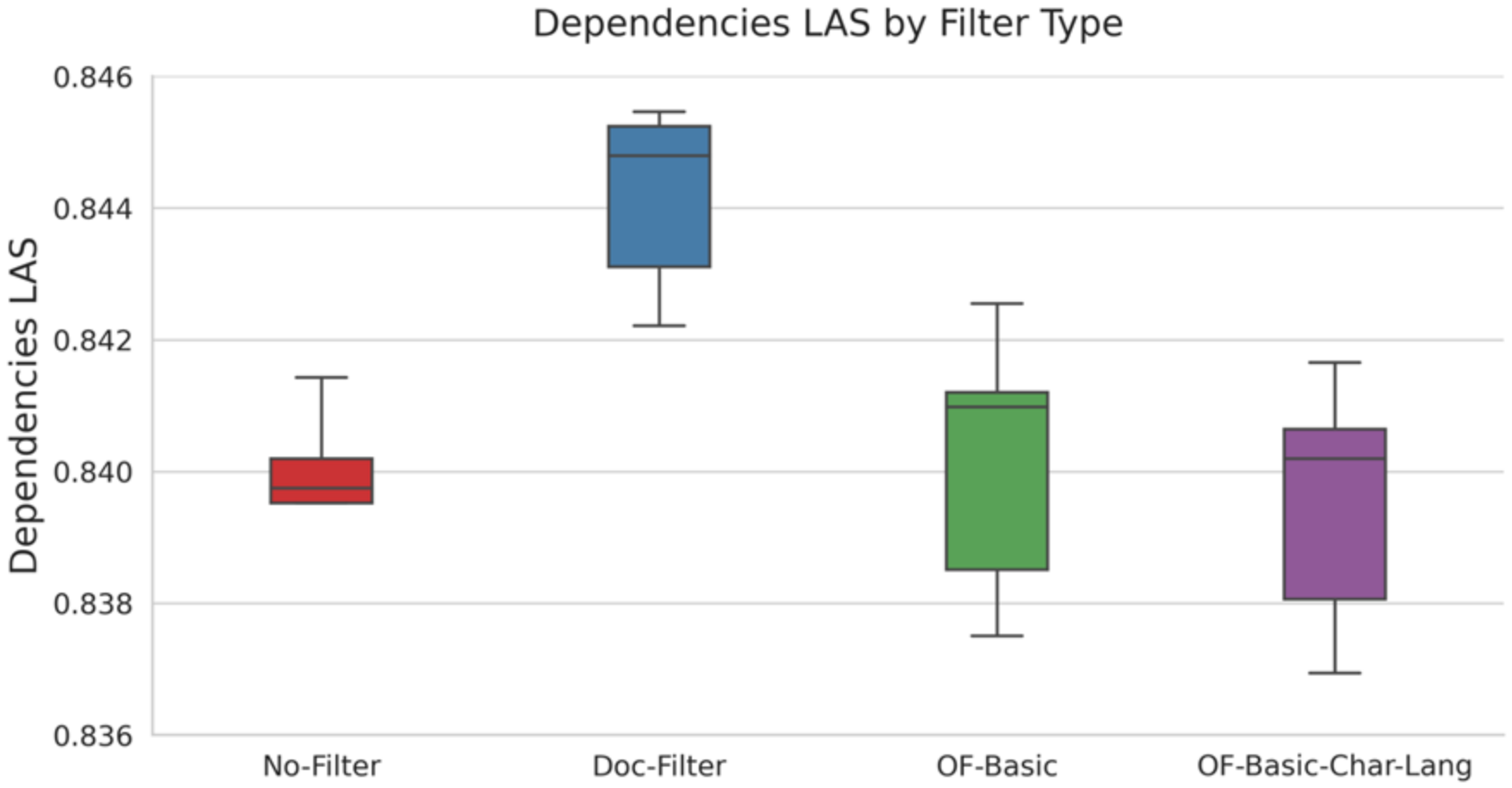
Collection of publicly available and internally collected corpora

Experiments

- Dependency Parsing
- Used Irish Universal Dependencies Treebank (v2.8)

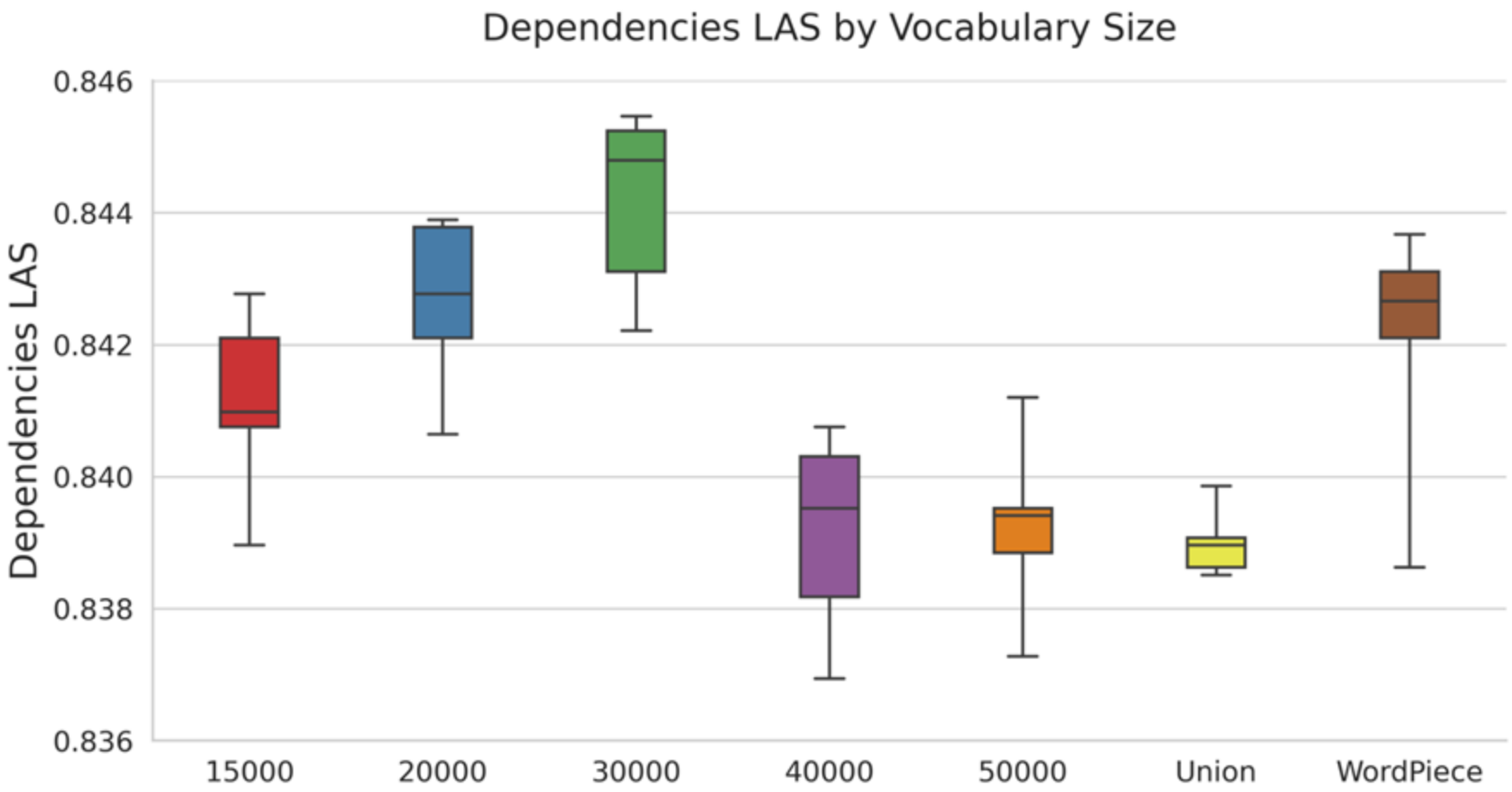


Experiment 1: Corpus Filtering



- Document-level filtering performs best

Experiment 2: Vocabulary Size



– Vocabulary size of 30K works best

Parsing Results

- ~10 point increase with word embeddings!
- gaBERT model outperforms other multilingual and monolingual baselines

		LAS	
Model	UD	Dev	Test
BiLSTM	2.8	73.4	71.4
mBERT	2.8	81.8	80.3
WikiBERT	2.8	81.9	80.4
mBERT-cp	2.8	84.3	82.3
gaBERT	2.8	**85.6**	**84.0**

Interim Summary

- Irish UGC is noisy! We need UGC-specific tools.
- **UGC** corpora might seem like a "nice-to-have", but is extremely **valuable**
  - Gives a voice to minority group
  - Provides sociolinguistic research resource
  - Supports effort to tackle digital extinction
- **Leveraging** existing **resources** is really important for low-resourced languages
- It's possible to widen NLP R&D to those who don't speak a language
  - **Collaboration**
- **gaBERT** is a game-changer for Irish NLP!!

Talk Outline

- Irish Language Technology - Overview
- GaelTech Project
- Parsing Irish Tweets
- gaBERT
- **European Language Equality Project**

Photo credit: Teresa Lynn

"Language Equality" EP Resolution (2018)

EP Resolution "Language equality in the digital age"
P8_TA(2018)0332 – partially based on the STOA study

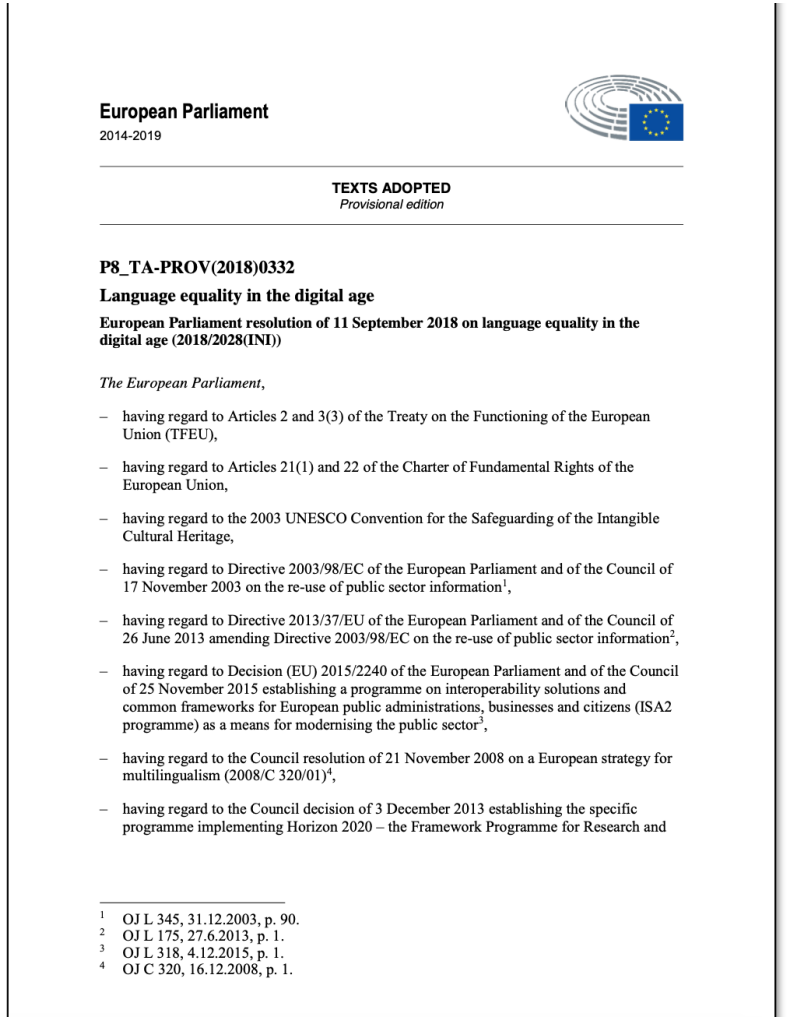
Voting (11 Sept. 2018): **592 yes** – 45 no

Selected Recommendations addressed by ELE:

25. Establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands

29. Create a European LT platform for sharing of services

27. Europe has to secure its leadership in language-centric AI



European Parliament
2014-2019

TEXTS ADOPTED
Provisional edition

P8_TA-PROV(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

The European Parliament,

– having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),

– having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

– having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,

– having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information[1],

– having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information[2],

– having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector[3],

– having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)[4],

– having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and

[1] OJ L 345, 31.12.2003, p. 90.
[2] OJ L 175, 27.6.2013, p. 1.
[3] OJ L 318, 4.12.2015, p. 1.
[4] OJ C 320, 16.12.2008, p. 1.

European Language Equality Project

EUROPEAN LANGUAGE EQUALITY

EUROPEAN LANGUAGE GRID

Consortium: 52 partners from all over Europe

Coordinator: ADAPT Centre (Dublin City University)

Co-Coordinator: DFKI

Objective: *development of a strategic research, innovation and implementation agenda to achieve digital language equality in Europe by 2030*

Runtime: 18 months – ELE and ELG are both finishing up in June 2022

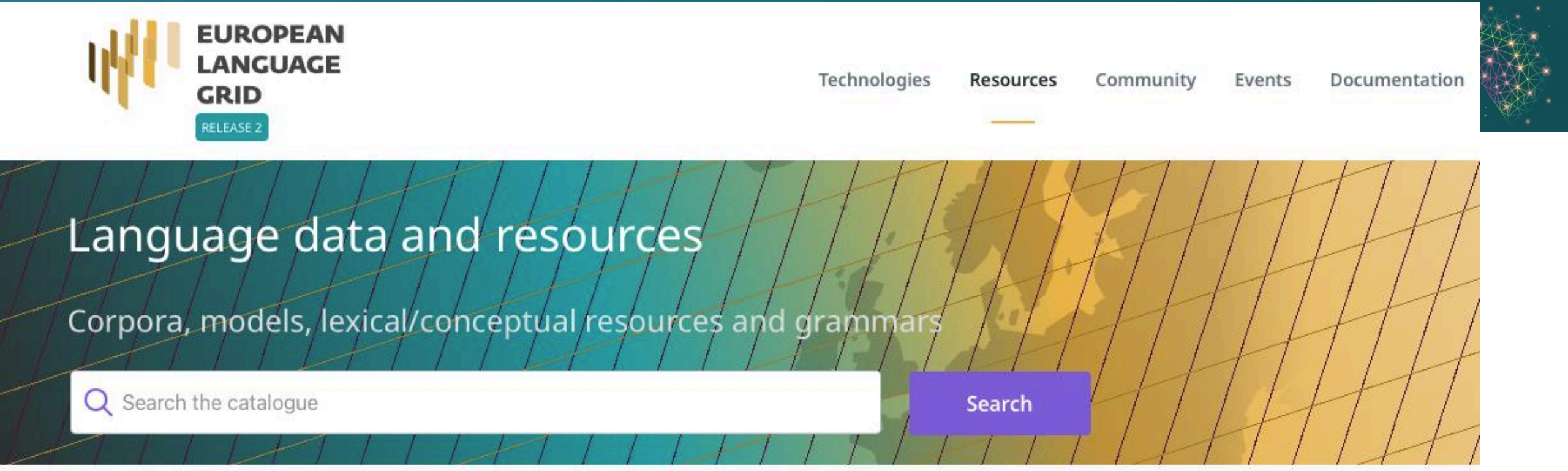
Started on **1 January 2021**

http://www.european-language-equality.eu

European Language Equality Project

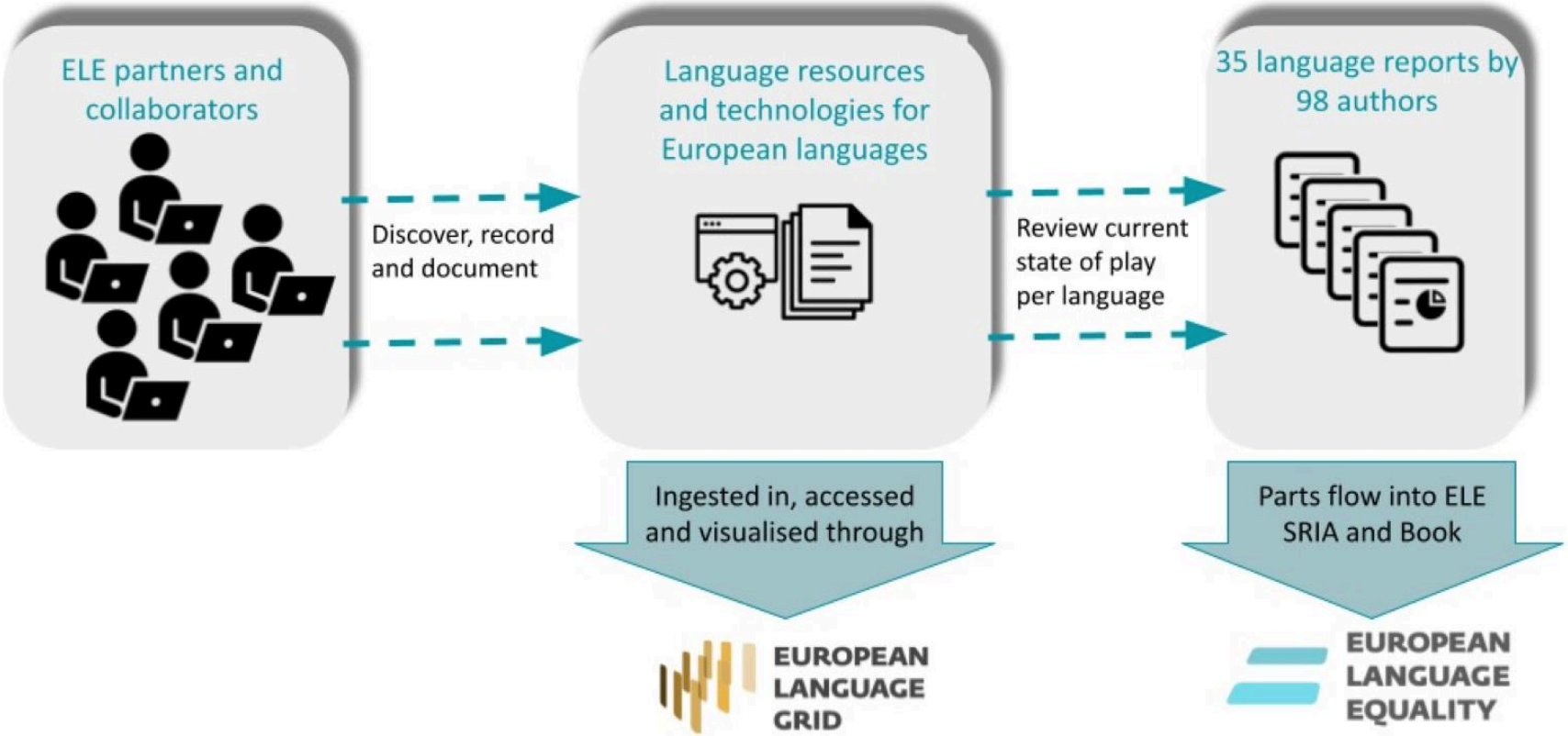
- Main result: **Strategic agenda and roadmap**
- Research partners produce **updates** of the **32 META-NET White Papers**

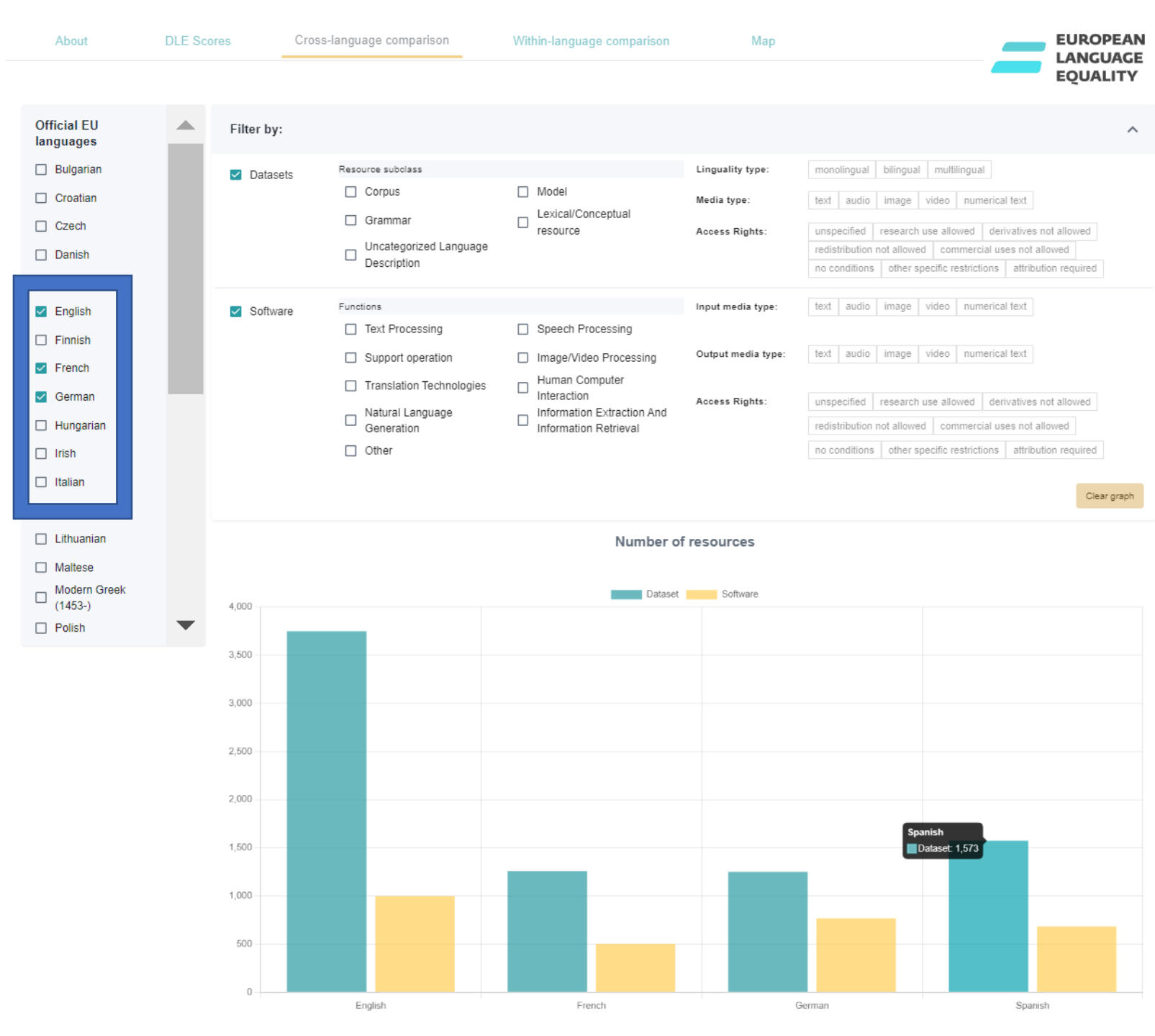
- **Networks** and **initiatives** produce one report on: needs, wishes, demands, visions
- SME partners produce four technical **Deep Dives** for the main technology areas.
- Several additional reports to be produced, primarily by the five core partners.
- **EU citizens' survey** – 20,000+ responses
- **77 languages** taken into account


EUROPEAN LANGUAGE GRID
RELEASE 2
Technologies
Resources
Community
Events
Documentation
Language data and resources
Corpora, models, lexical/conceptual resources and grammars
Search the catalogue
Search
Corpora
Models
Lexical/Conceptual resources

Populating the ELG through “language informants”

An evidence-based investigation





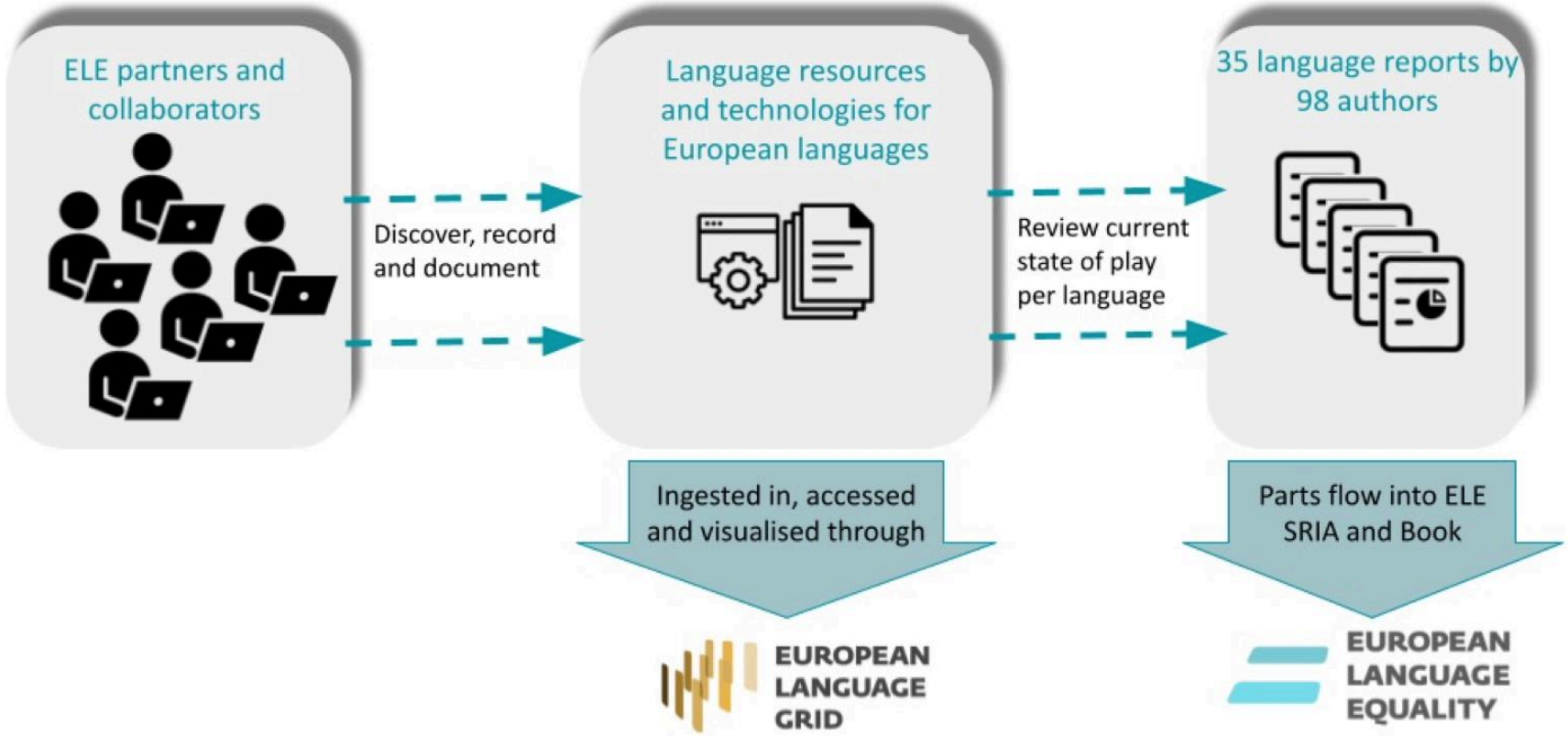
European Language Grid “Dashboard“

Cross-language comparison

11,500+ records

EU Language Reports – technical readiness

An evidence-based investigation



ELE – Irish Report

EUROPEAN LANGUAGE EQUALITY

D1.20

Report on the Irish Language

https://european-language-equality.eu/deliverables/

Gaps and shortcomings

- Still **no National Digital Plan** or Strategy for Irish
  - Ireland's AI Strategy is focused on English LT R&D

- Still no:
  - **Automatic Speech Recognition**
  - **Automatic Subtitling**
  - **Question-Answering Systems**
  - **Named Entity Recognition**
  - **Information Retrieval**
  - **Information Extraction**
  - **Virtual Agents**
  - **Adaptive Learning** (CALL)
  - **NLG**
  - **Semantic Role Labelling**
  - **Linked Data**/ **KGs**

- **Skills gap**
  - Only 1 undergraduate level course in computing and linguistics (with few Irish speaking students)
  - Lack of dedicated funding for (postgraduate) Irish LT research

- Minimal support or investment from **industry** (despite global tech EMEA headquarter location!)

Recommendations for the SRIA

- **Change of focus**
  - Away from current focus on dictionary development and training translators…
- **Untapped potential**
  - Lack of awareness around value of language data
  - Poor language data management practices
- **Need dedicated LT programmes**
  - Upskilling, 3rd level, vocational training, inter-disciplinary studies
- **Long term strategy**
  - Long-awaited National Digital Plan for Irish (2015), need forward thinking
- **Open-source culture**
  - Many corpora are inaccessible due to licensing restrictions
- **Corporate Social Responsibility** – give back to the public

ELE – French Report

EUROPEAN LANGUAGE EQUALITY

D1.14

Report on the French Language

https://european-language-equality.eu/deliverables/

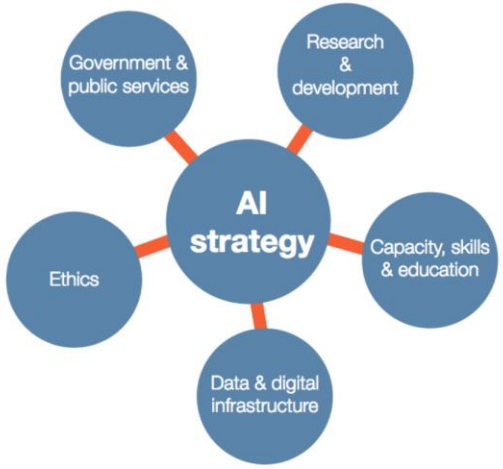
https://hal.archives-ouvertes.fr/hal-03637784v1

ELE – Main Findings

All countries surveyed have a **national AI strategy**

- *Croatia?*

- **No LT-related funding at all**
  - Austria, Serbia
- **Limited LT-related funding**
  - Bulgaria, Czechia, Finland, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Lux., Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Sweden.
- **Funding for LT through AI**
  - Belgium, France, Germany and Malta
- Dedicated **LT programmes**
  - Denmark, Estonia, Iceland and Spain



ELE – Main Findings

- **Number of speakers** determines industry investment (need governments on board)
- **Data-driven paradigm** is leaving languages behind
- Current state-of-the-art LT research and development is based on access to huge, and previously unthinkable, amounts of data and **processing power**.
- Across the EU, there is an **uneven distribution of resources** (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language.
- Expertise hard to find: **lack of interdisciplinary studies**, loss of talent to giant tech companies.
- Awareness raising on **value of data**:
  public sector, publishers, (minority) language communities
- Enterprise and health data: tends to be **locked in regulatory and corporate silos**. By nature is confidential and companies need to respect data protection regulations. Barriers for making data available are high.

ELE – Main Findings

- While **copyright law** is subject to fair-use exceptions in countries such as the US, European law is far less flexible.
- **GDPR**: Since unconstrained, unstructured text often includes personal data: data protection and privacy (DPP) policies **put limits on availability**
- Speech technologies are not accessible nor available to everyone on an equal level, i. e., functions, performance, robustness (**marginalises** language communities, elderly, etc.)
- Translation technology lacks for many language pairs (**restricting access to information**: governments, newswire)
- **NLP Labelled data**: more difficult and expensive to acquire for lesser used languages (crowdsourcing not possible)
- Ethical issues: race/ gender-related **biases** in training data -> resulting models.
- Ethical issues: **Carbon footprint**. Trade-off between an difference in 1 BLEU/WER worth it? Argument for new approaches?

Impact of Digital Inequality

ELE – Impact of Digital Inequality

- The lack of digital support for a language can lead to:
  - **language shift** and eventual language decline, particularly amongst younger generations.
  - A **divide arises in levels of information accessibility** and economic progression across language communities.
- Online data mining is often used by governments and media to gather information on events, political issues and public sentiment. The **voices of many will be left unheard**, unrepresented and unaccounted for.
- **Divide in the levels of education** on offer across language communities - contributing further to **societal inequalities**.

ELE – Impact of Digital Inequality

- Court and criminal justice systems use multimodal content retrieval to find evidence in audio and video content. **Disadvantage for some language communities.**
- Divide in level of **access to healthcare**
- **Marginalised** groups in speech processing (dialects, elderly, disabilities)
- **Selective** customer **support**
- **Fragmented EU market** (businesses at disadvantage)

Ní neart go cur le chéile

– `There is no strength without unity’

Thank you!
#GRMA

Lauren Cassidy, Teresa Lynn, James Barry, Jennifer Foster,
TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish.
In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, May 2022, Dublin, Ireland (to appear)

Sanguinetti, M., Bosco, C., Cassidy, L. *et al.*
Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Lang Resources & Evaluation* (2022).

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J Ó Meachair, and Jennifer Foster. 2021. gaBERT–an Irish language model. *arXiv preprint arXiv:2107.12930*

Engaging Content
Engaging People

www.adaptcentre.ie

FUNDED BY:
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
Science Foundation Ireland sfi For what's next

teresa.lynn@adaptcentre.ie
@cigilt