Engaging Content
Engaging People

Giving Out or Happy Out? Processing Multiword Expressions in Irish

Abigail Walsh, ADAPT Center, Dublin City University
Supervised by Teresa Lynn & Jennifer Foster

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.

About Me

- PhD student at ADAPT Center, Dublin City University
- Research: Automatic processing of Irish MWEs for NLP
  - Computational approach linguistically informed
- Research visit to work with Carlos Ramisch
  - Potential multilingual approaches to the problem

1. Defining the Problem

MWEs and the Irish Language

GIVING OUT ??

HAPPY OUT ??

Multiword Expressions in NLP

- Sag et al. 2002 -- a well-known pain in the neck for NLP!
- Poses challenge in many NLP tasks
  - Machine Translation, Parsing, Text Summarisation, Question Answering, etc.
- Varying linguistic properties and levels of idiosyncrasy
- Multiple definitions
  - (Constant et al. 2017)
- We use the following definition: *a string of two or more tokens, which form a unit at a semantic, syntactic or lexical level*

What are Multiword Expressions?

Dia duit

Serves you right

Make a decision

Beat around the bush

Craic agus ceol

Cúpla focail

Nil aon tinteán mar do thinteán fein

Status quo

Salt and pepper

Put up with it

Céad míle fáilte

Irish Language Overview

- Celtic language (Indo-European)
- Verb-Subject-Object word order
- Inflected language
  - Nouns (Case, Gender, Number)
  - Adjectives (Case, Gender, Number)
  - Prepositions (Gender, Person, Number)
  - Verbs (Tense, Aspect, Mood, Person, Number)
- No indefinite article
- Two verbs ‘to be’: copula and substantive verb
- Idiomatic constructions e.g. *tá ocras orm* (hunger is on me) `I am hungry’

Types of MWEs in Irish

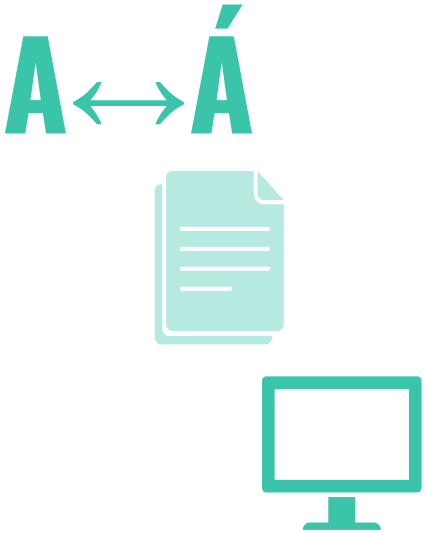
- Verb-Particle Constructions
  - Ó Domhnalláin and Ó Baoill (1975)
  - Foclóir Briathra Gaeilge (Irish Valency Dictionary), Wigger (2008)
- Light Verb Constructions
  - Stenson (1981)
  - Bayda (2014)
  - Bloch-Trojnar (2009, 2010)
- Idiomatic constructions with “be”
  - Stenson (1981)
- Idioms
  - Ní Loingsigh (2016, 2018, 2021)
- Fixed Expressions
  - Uí Dhonnchadha (2008)

2. Understanding the Problem

Why does it matter?

Downstream Applications

- Machine Translation
- Search Engines
- Grammar Checkers
- Language Learning Apps
- ...



Downstream Applications

***Níos éadroime breosla* =**
less heavy fuel??

***Seomra Athraithe Linbh* =**
Baby Exchanging Room??

ESTIMATED

50%

OF OUR LEXICON ARE
MULTIWORD EXPRESSIONS[1]

[1]Sag et al. Multiword Expressions: A Pain in the Neck for NLP

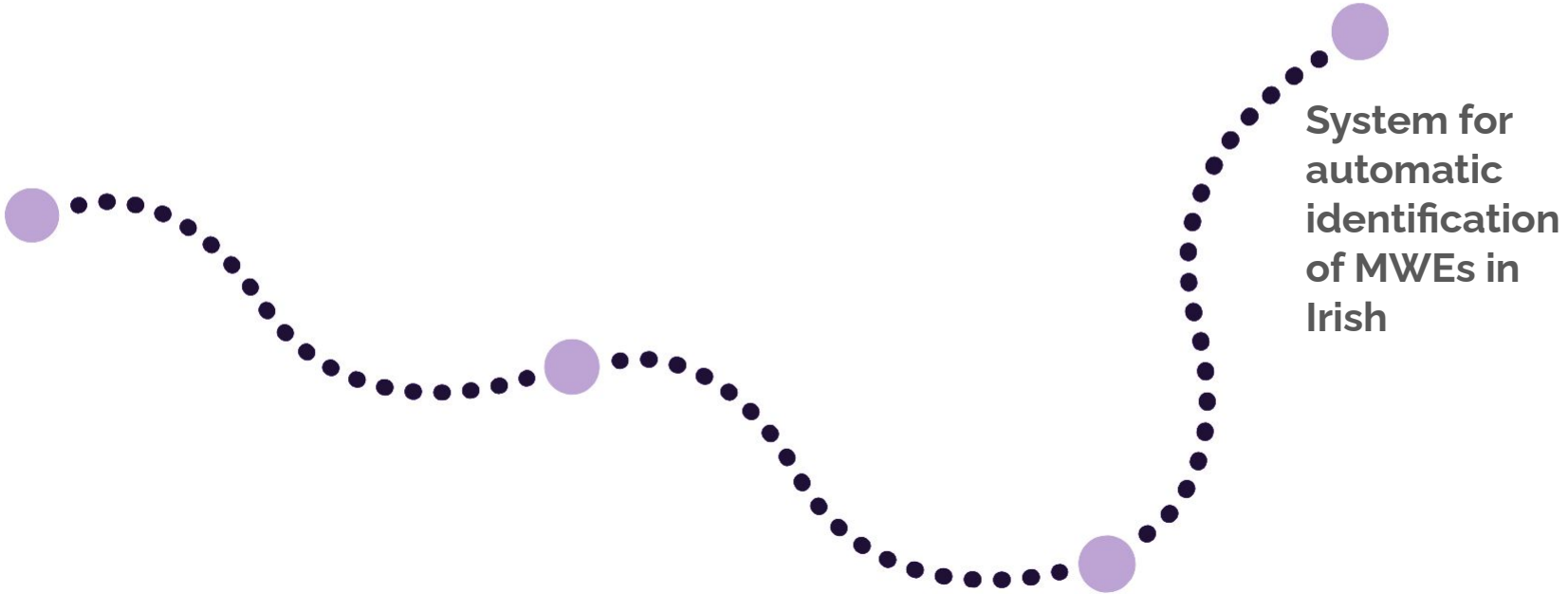
3. Solving the Problem

How to process MWEs in Irish

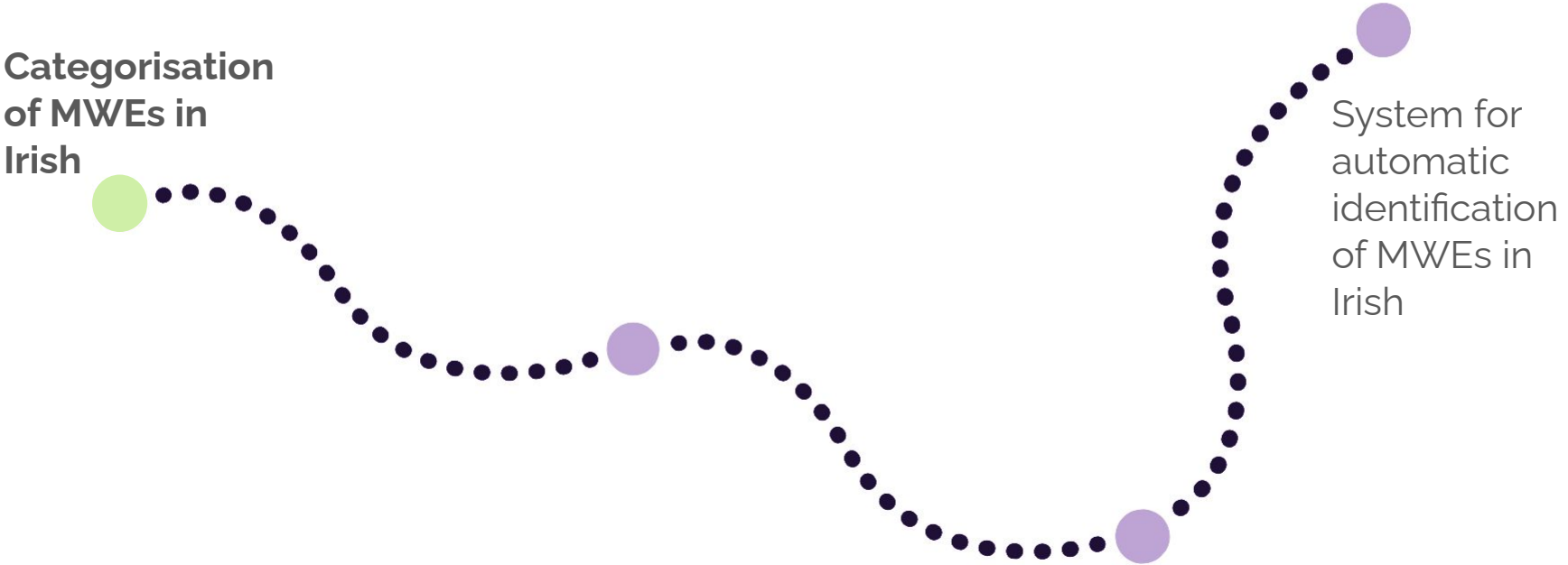
Challenges in handling MWEs

- Discontinuity
  - look the top secret information up
- Ambiguities
  - take the cake
- Productivity
  - Make a decision, point, statement, etc.
- Level of flexibility
  - “Ad hoc” vs “Spilling all the beans”

Road Map



Road Map


Categorisation of MWEs in Irish
System for automatic identification of MWEs in Irish

Approaches to Categorisation

- Levels of idiomaticity
  - Lexicalised phrases vs. institutionalised phrases
- Morphosyntactic information
  - Parts of speech
  - Valency
- Flexibility / difficulty to parse
  - Syntactically fixed vs. syntactically flexible
  - Discontiguity
  - Ambiguity
  - etc.

PARSEME Identification of verbal MWEs

- PARSEME Shared Task on Automatic Identification of verbal MWEs
  - PARSEME project funded as a COST action
  - Intended to increase and enhance ICT support of the European languages
  - Shared Task in verbal MWE identification came from this initiative
  - Edition 1.0, 1.1 and 1.2 took place in 2017, 2018 and 2020 respectively
  - Cross-lingual annotation guidelines provided basis for categorisation of Irish MWEs

PARSEME Identification of verbal MWEs

- Annotation Guidelines developed through annotation for each edition
  - 27 languages in total
- Tests for 6 broad categories (8 fine-grained categories) of verbal MWEs
  - Light Verb Constructions (**full; causative**)
  - Verbal Idioms
  - Inherently Reflexive Verbs
  - Verb Particle Constructions (**full; semi**)
  - Inherently Adpositional Verbs
  - Multi-Verb Constructions

Annotation guidelines

PARSEME shared task on automatic identification of verbal MWEs - edition 1.1 (2018)

Structural tests (S)

Structural tests are quite simple preliminary tests that help determining the syntactic structure of the VMWE. This is required in order to point at the right category-specific identification tests. In practice, annotators will rarely need them since they will already have an intuition about the VMWE candidate category when they identify it.

Test S.1 (prev. 6) - [HEAD] - Syntactic head

Does the candidate contain a unique verb functioning as the functional syntactic head of the whole?

↳ **NO ⇒** Apply the VID-specific tests

- *(EN) to* ***pretty-print*** → there is an unusual case of an adjective modifying a verb
  to ***drink and drive*** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination

↳ **YES ⇒** continue to the next test

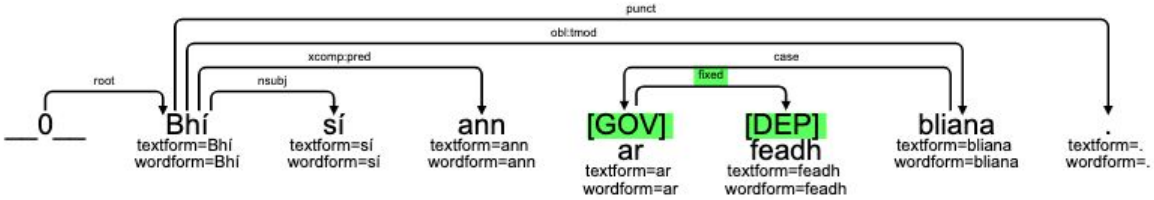
Universal Dependencies Classification of MWEs

- Universal Dependencies framework
  - Develop cross-linguistically consistent treebank annotation for languages
  - Version 2.8 (released 15th May 2020) includes 202 treebanks in 114 languages
- Irish UD Treebank
  - Irish UD Treebank (IUDT): UD v.1 release in 2015
  - Small treebank (4910 trees in v2.8 release)
  - In-depth analysis of MWEs in recent editions
    - McGuinness et al. (2020)

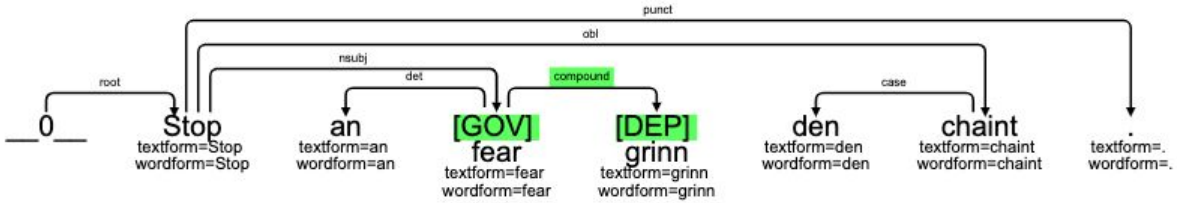
Universal Dependencies Classification of MWEs

- Fixed expressions
  - Used for fixed grammaticized expressions that behave like function words or short adverbials
- Compound expressions
  - Used for any kind of X^0 compounding (e.g. noun compounds, adjective compounds), as well as particle verbs, serial verbs, etc.
- Flat constructions
  - Used for exocentric (headless) semi-fixed MWEs like names and dates

Universal Dependencies Classification of MWEs



Bhí sí ann ***ar*** ***feadh*** *bliana.*
Was she there **on** **duration** year
She was there for a year.



Stop an ***fear*** ***grinn*** *den chaint.*
Stopped the **man** **of-humour** off-the talk
The clown stopped talking.

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	
*Named Entities	
*Fixed Expressions	
Institutionalised Phrases	
†Light Verb Constructions	
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	
*Fixed Expressions	
Institutionalised Phrases	
†Light Verb Constructions	
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	
Institutionalised Phrases	
†Light Verb Constructions	
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	
†Light Verb Constructions	
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	***déan obair*** ‘do work’
†Verb Particle Constructions	
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	***déan obair*** ‘do work’
†Verb Particle Constructions	***leag amach*** ‘lay out/arrange’
†Inherently Adpositional Verbs	
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	***déan obair*** ‘do work’
†Verb Particle Constructions	***leag amach*** ‘lay out/arrange’
†Inherently Adpositional Verbs	***bain le*** (take with) ‘associate’
†Verbal Idioms	
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	***déan obair*** ‘do work’
†Verb Particle Constructions	***leag amach*** ‘lay out/arrange’
†Inherently Adpositional Verbs	***bain le*** (take with) ‘associate’
†Verbal Idioms	***an lá atá inniu (ann)*** (the day that is today) ‘nowadays’
Copular Constructions	

MWE Categories for Irish

Category Labels	Example
*Nominal Compounds	***mac tíre*** (son of the land) ‘wolf’
*Named Entities	***Baile Átha Cliath*** ‘Dublin’
*Fixed Expressions	***in aghaidh*** ‘against’
Institutionalised Phrases	***domhan uile*** ‘whole world’
†Light Verb Constructions	***déan obair*** ‘do work’
†Verb Particle Constructions	***leag amach*** ‘lay out/arrange’
†Inherently Adpositional Verbs	***bain le*** (take with) ‘associate’
†Verbal Idioms	***an lá atá inniu (ann)*** (the day that is today) ‘nowadays’
Copular Constructions	***is maith liom*** (is good with me) ‘I like’

*UD inspired
†PARSEME inspired

Contentious Questions

Annotate constructions with selected prepositions as LVC + IAV?

Bain trial ***as*** ‘try’ -- construction does not occur without preposition

Light verb + Fixed Adverbial, Prepositional or Complement constructions

Cuir *rud* ***ar fáil*** (Put a thing available) ‘make available’ -- annotated as LVC

Substantive ‘be’ combines with preposition to create idiomatic constructions

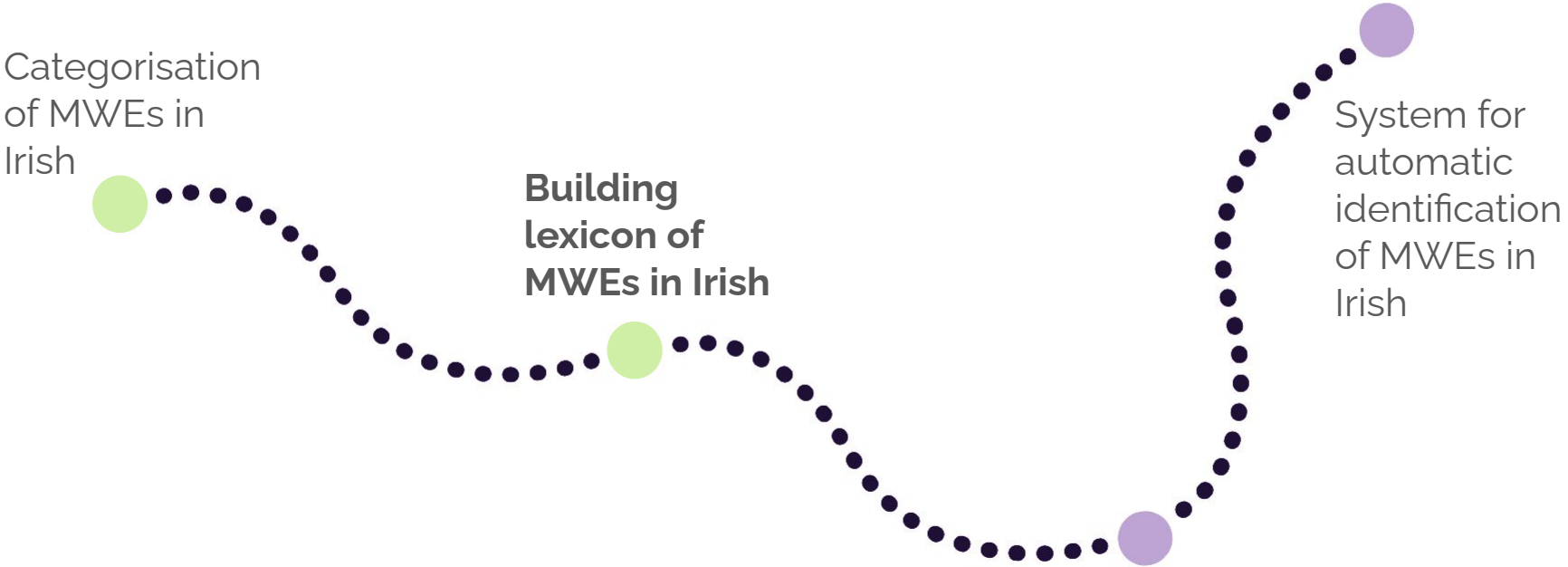
Bí ag (be at) ‘have’ -- annotated as IAV

In Irish, reflexive pronoun is formed through the combination of *féin* + personal pronoun

iompair mé ‘I carry’ vs. *iompair mé féin* (carry I self) ‘I behave myself’

This type of construction appears to be rare in Irish -- not annotated in large part

Road Map


Categorisation of MWEs in Irish
Building lexicon of MWEs in Irish
System for automatic identification of MWEs in Irish

Extracting and Cleaning

Resources Used

- NEID and EID
- Foclóir Gaeilge Béarla
- Tearma database
- An Foclóir Beag
- Irish Wordnet
- Pota Focal Gluais Tí
- Peadar Ó Laoghaire Idiom Collection

Method

1. **Extract** from lexical resources
2. **Clean** redundant/duplicate entries, noise, etc.
3. **Define categorisation scheme** for MWE types
4. **Enhance** lexicon
5. **Share**

Current Progress

- **210,000+** MWE entries extracted
- Noisy/redundant entries **cleaned or removed**
- **Automatic annotation** of MWE categories
- **Licence** for releasing portions of the corpus seeking approval

Categorisation Scheme

Category	Example in Irish
Nominal MWEs	*Madra crainn* 'Squirrel'
Compound Prepositions	*Tar éis* 'After'
Verb Particle Constructions	*Tabhair amach* 'Give out'
Adpositional Verbs	*Buail le* 'Meet'
Light Verb Constructions	*Déan dearmad* 'Forget'
Copular Constructions	*Is léir* 'Clearly'
Idioms	*Giorraíonn beirt bóthar* 'Two shorten the road'
Institutionalised Phrases	*Aire agus forcamás* 'Care and attention'

Categorisation Scheme

Category	Example in Irish
Nominal MWEs	*Madra crainn* 'Squirrel'
Compound Prepositions	*Tar éis* 'After'
Verb Particle Constructions	*Tabhair amach* 'Give out'
Adpositional Verbs	*Buail le* 'Meet'
Light Verb Constructions	*Déan dearmad* 'Forget'
Copular Constructions	*Is léir* 'Clearly'
Idioms	*Giorraíonn beirt bóthar* 'Two shorten the road'
Institutionalised Phrases	*Aire agus forcamás* 'Care and attention'
Named Entities	

Structure

GA-Head	GA	POS	EN	Source	ID
cú	cú allta	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	muc mhara	s.	porpoise	eid	141920eid
min	min sáibh	noun	UNK	lsg	140450lsg
min	min sáibh	UNK	sawdust	tearma	140451tearma

Structure

GA-Head	GA	POS	EN	Source	ID
cú	cú allta	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	muc mhara	s.	porpoise	eid	141920eid
min	min sáibh	noun	UNK	lsg	140450lsg
min	min sáibh	UNK	sawdust	tearma	140451tearma

Structure

GA-Head	GA	POS	EN	Source	ID
cú	cú allta	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	muc mhara	s.	porpoise	eid	141920eid
min	min sáibh	noun	UNK	lsg	140450lsg
min	min sáibh	UNK	sawdust	tearma	140451tearma

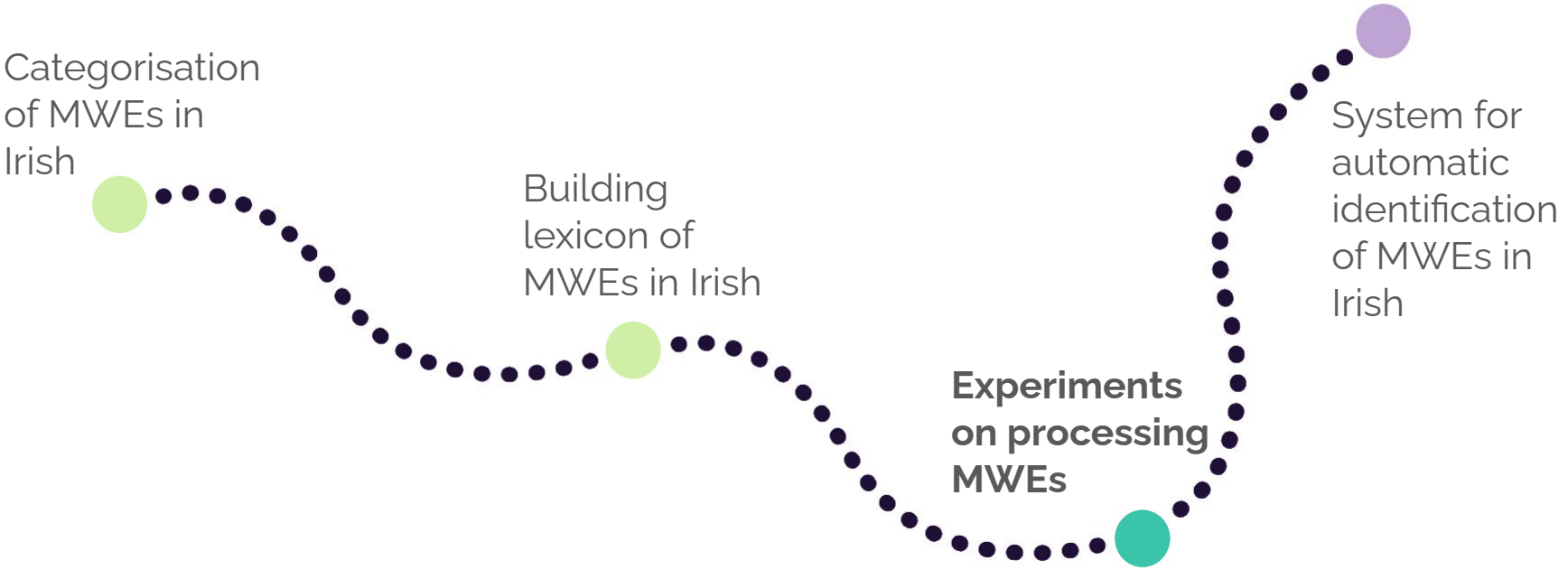
Structure

GA-Head	GA	POS	EN	Source	ID
cú	cú allta	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	muc mhara	s.	porpoise	eid	141920eid
min	min sáibh	noun	**UNK**	lsg	140450lsg
min	min sáibh	**UNK**	sawdust	tearma	140451tearma

Structure

GA-Head	GA	POS	EN	Source	ID
cú	cú allta	m. gs. pl. gs. & gpl. f	wolf	x_fgb	2388x_fgb
muc	muc mhara	s.	porpoise	eid	141920eid
min	min sáibh	noun	UNK	lsg	140450lsg
min	min sáibh	UNK	sawdust	tearma	140451tearma

Road Map

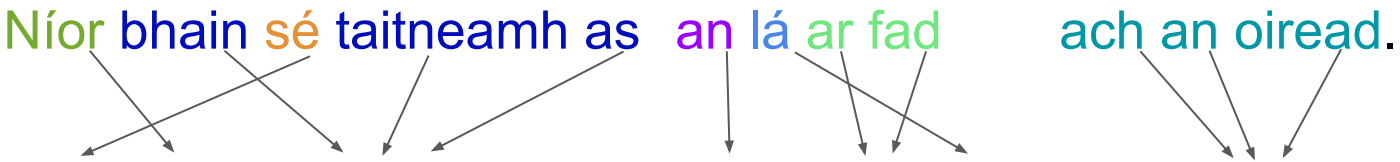

Categorisation of MWEs in Irish
Building lexicon of MWEs in Irish
Experiments on processing MWEs
System for automatic identification of MWEs in Irish

PMI Scores and Word Alignments

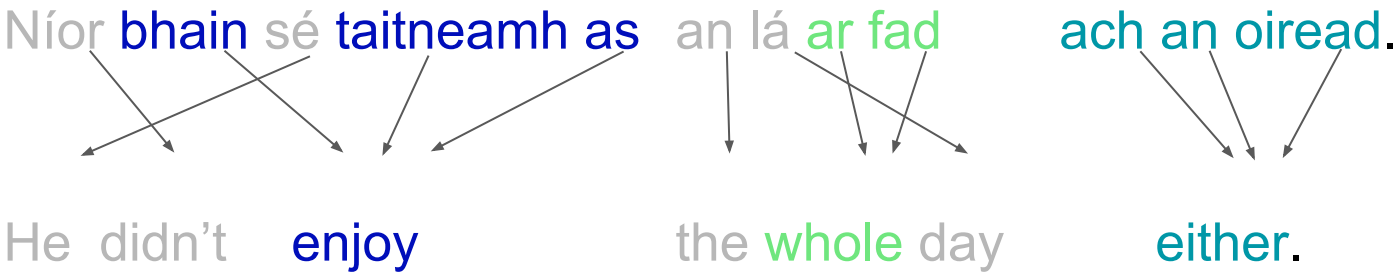
Method (*Tsvetkov and Wintner, 2010*)

1. Align two parallel corpora
2. Extract all one to many or many to many alignments (potential MWEs)
3. Calculate PMI score of bigrams in extracted phrases, using large monolingual corpus
4. Accept bigrams above certain threshold as MWEs

Extracting MWEs from misalignments



Extracting MWEs from misalignments



Extracting MWEs from misalignments

Extracted

bhain taitneamh as

ar fad

ach an oiread

Candidates filtered using PMI scores generated from monolingual corpus

PMI Scores and Word Alignments

Results

- PMI scores revealed some common collocations
- Word alignments were poor: word order?
- Repeat experiment, focus on better word alignments

MWEs in MT: Experiment

- Tag MWEs from GA and EN lexicons in parallel data
  - Ilfhocail + English MWE lexical lists
- Train an NMT system with MWEs as features

```
Bain|MWE úsáid|MWE as|MWE an|NONE mhodh|NONE HTML|NONE .|NONE

Use HTML mode .
```

MWEs in MT: Experiment

- Trained MT models using four datasets
  - baseline, MWE-fixed, MWE-semi-fixed, and MWE-joined
  - RNN, BRNN and Transformer models
- Created two small parallel datasets manually tagged on the source side with MWE data (described in resources)
- Evaluation shows some small increases in BLEU scores
- Manual inspection reveals differences between systems
  - No discernible pattern in choice of translation

MWEs in MT: Experiment

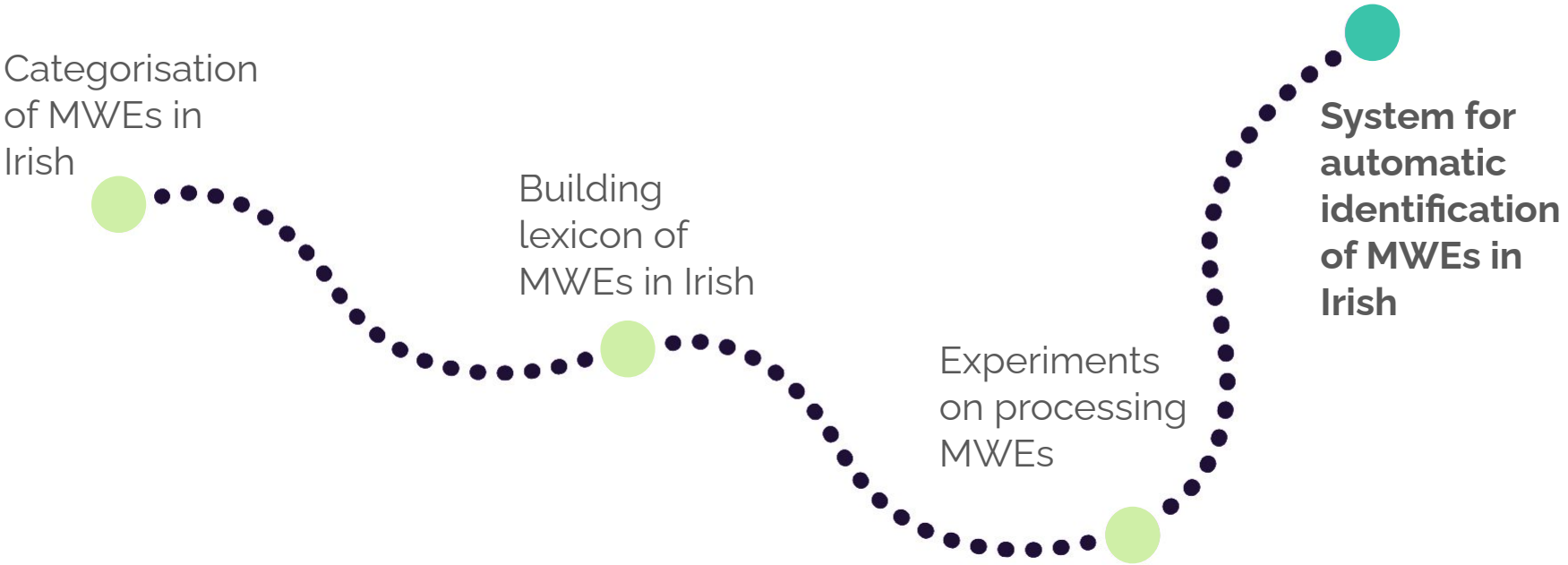
BLEU scores for RNN model

	Baseline dataset	Fixed MWEs annotated	Flexible MWEs annotated
EN-GA	47.0	**47.8**	47.6
GA-EN	53.0	53.0	**53.6**

BLEU scores for Transformer model

	Baseline dataset	Fixed MWEs annotated	Flexible MWEs annotated
EN-GA	56.8	56.6	**56.9**
GA-EN	**62.5**	62.3	62.1

Road Map


Categorisation of MWEs in Irish
Building lexicon of MWEs in Irish
Experiments on processing MWEs
System for automatic identification of MWEs in Irish

Current Project!

GIVING OUT

COMPLAINING

HAPPY OUT

CONTENT

Merci pour votre attention!

Go raibh maith agaibh

abigail.walsh@adaptcentre.ie