MACHINE LEARNING MEETUP

MARK MY MULTIWORDS:

PROCESSING MULTIWORD EXPRESSIONS IN IRISH

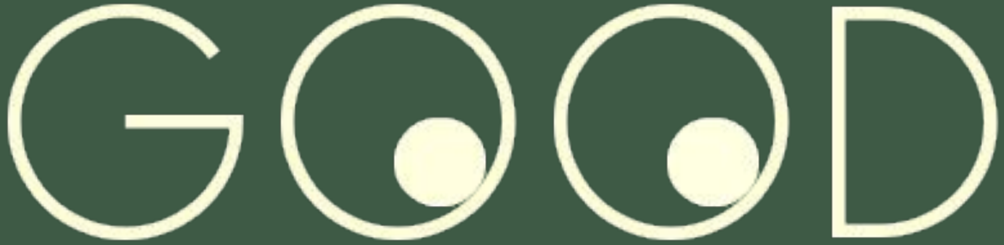
Abigail Walsh

thinking

thinking outside the box

horse chestnut

GOOD



good looking

EDGE

cutting edge

MORE THAN MEETS THE EYE...

DEFINING THE PROBLEM

1

What are Multiword Expressions?

1

What are Multiword Expressions?

- More than one word (multiword)
- Meaning more than sum of the individual words

1

What are Multiword Expressions?

Idioms	More than meets the eye
Phrasal Verbs	*Kick* things *off*
Compound Nouns	Horse chestnut
Light Verbs	Take a turn

UNDERSTANDING THE PROBLEM

2

Why should we care about MWEs in Irish?

2

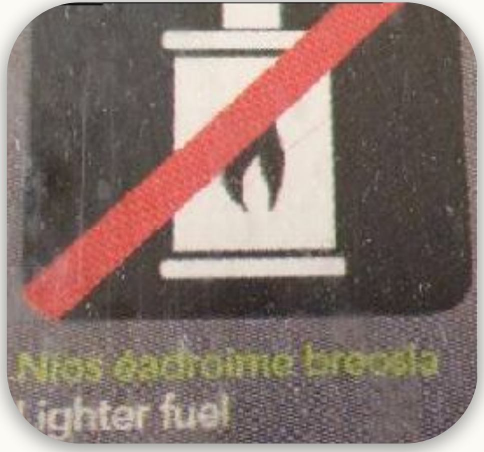
Why should we care about MWEs in Irish?

Downstream Applications

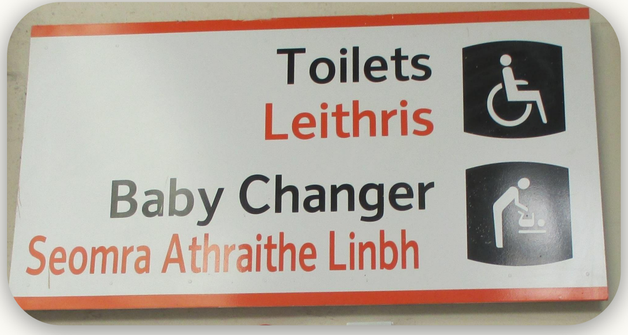
- Machine Translation
- Search Engines
- Grammar Checkers
- Language Learning Apps
- Sentiment Analysis Tools
- ...

Why should we care about MWEs in Irish?

“Níos éadroime breosla”



“Seomra Athraithe Linbh”



ESTIMATED

50%

OF OUR LEXICON ARE
MULTIWORD EXPRESSIONS[1]

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

SOLVING THE PROBLEM

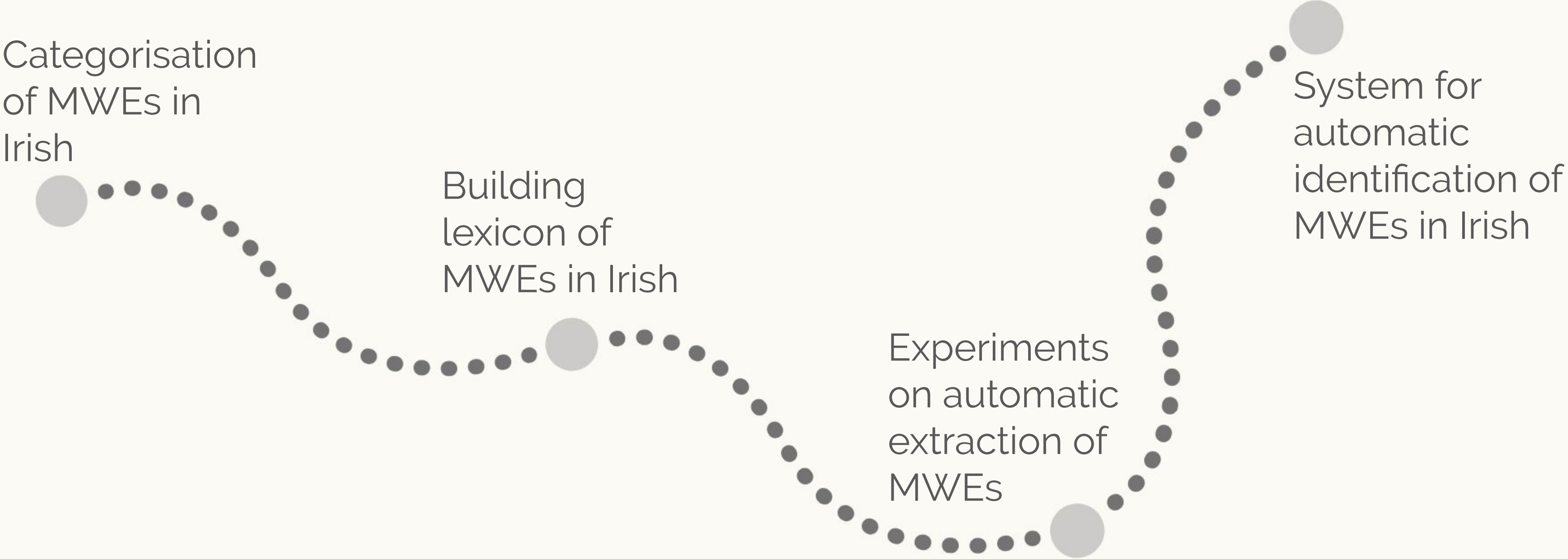
3

How can we process MWEs in Irish?

Challenges in Automatic Identification of Irish MWEs

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

Road Map



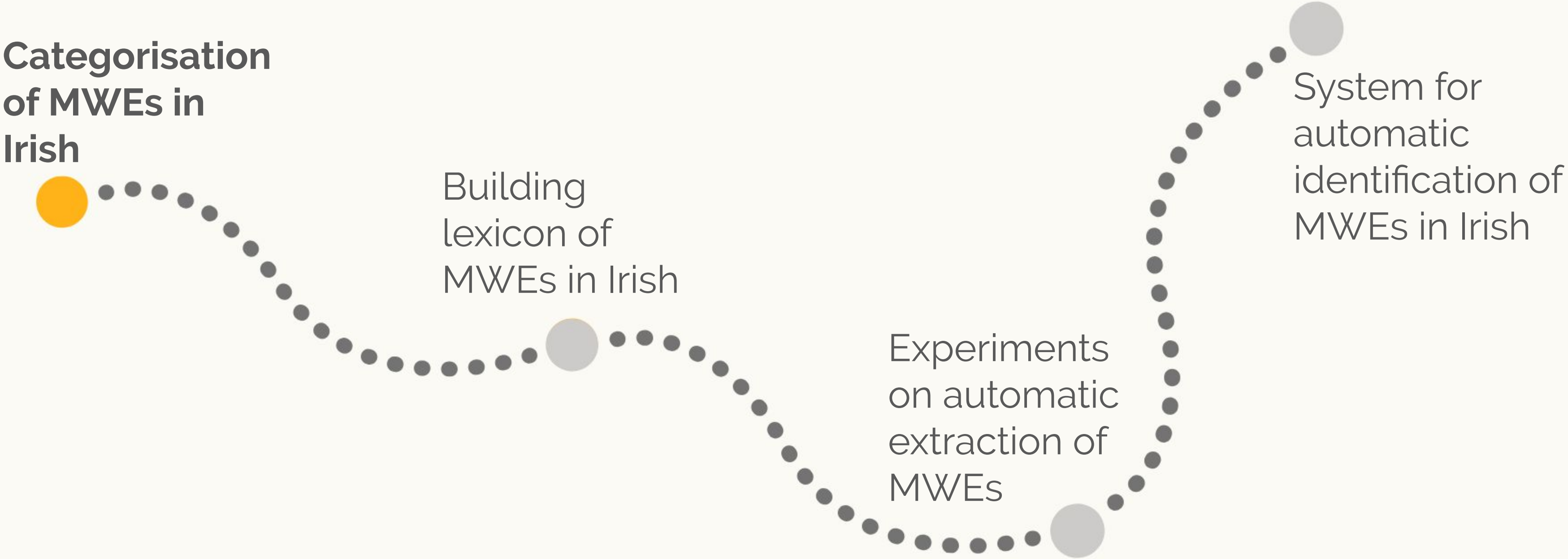
Categorisation of MWEs in Irish

Building lexicon of MWEs in Irish

Experiments on automatic extraction of MWEs

System for automatic identification of MWEs in Irish

Road Map



Categorisation of MWEs in Irish

Building lexicon of MWEs in Irish

Experiments on automatic extraction of MWEs

System for automatic identification of MWEs in Irish

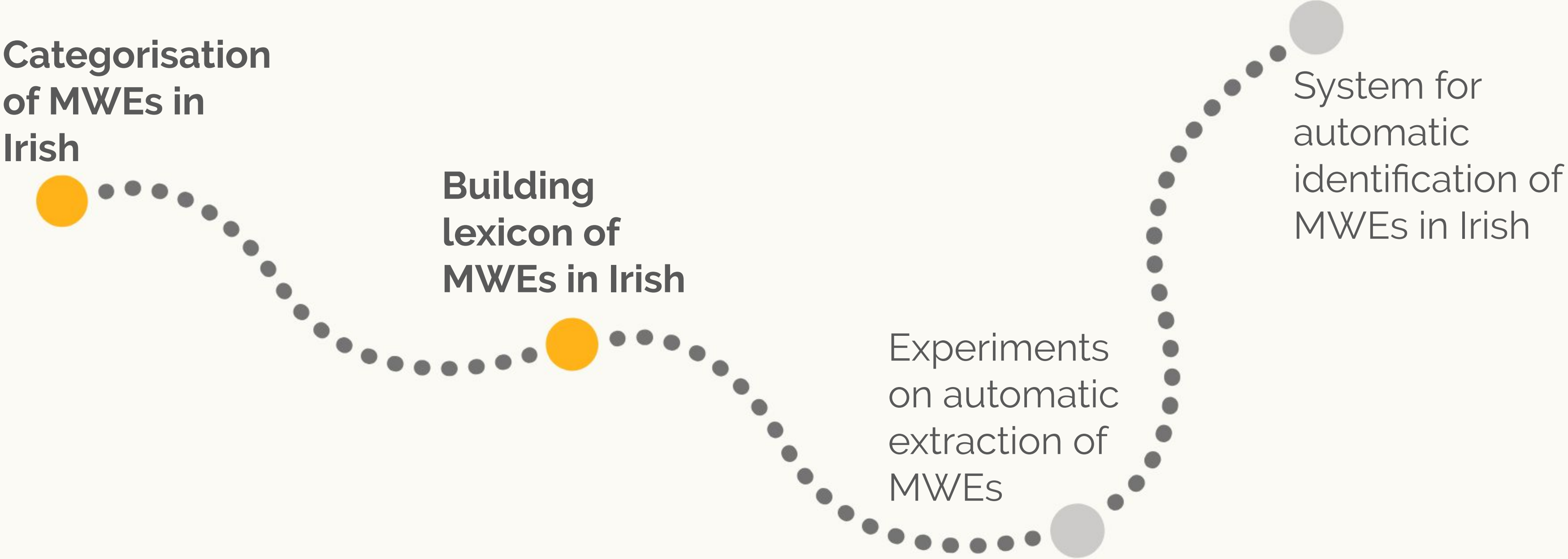
Categories of MWEs in Irish

Idiom	Gearraíonn beirt bóthar *‘Two shorten the road’*
Copular Construction	Is maith liom *‘I like’*
Verb Particle Construction (VPCs)	Tabhair amach *‘Give out’*
Inherently Adpositional Verbs (IAVs)	Abair le *‘Say to’*
Light Verb Constructions (LVCs)	Déan dearmad *‘Forget’*
Compound Nouns	Madra rua *‘fox’*
Compound Prepositions	In aice *‘beside’*

PARSEME Classification of Verbal MWEs

- EU Project: COST Action
- Shared Task 1.1: Identification of verbal MWEs across 19 languages
- Annotation guidelines for six broad categories of MWEs
- Four categories appropriate for Irish (LVCs, IAVs, VPCs, Idioms)

Road Map

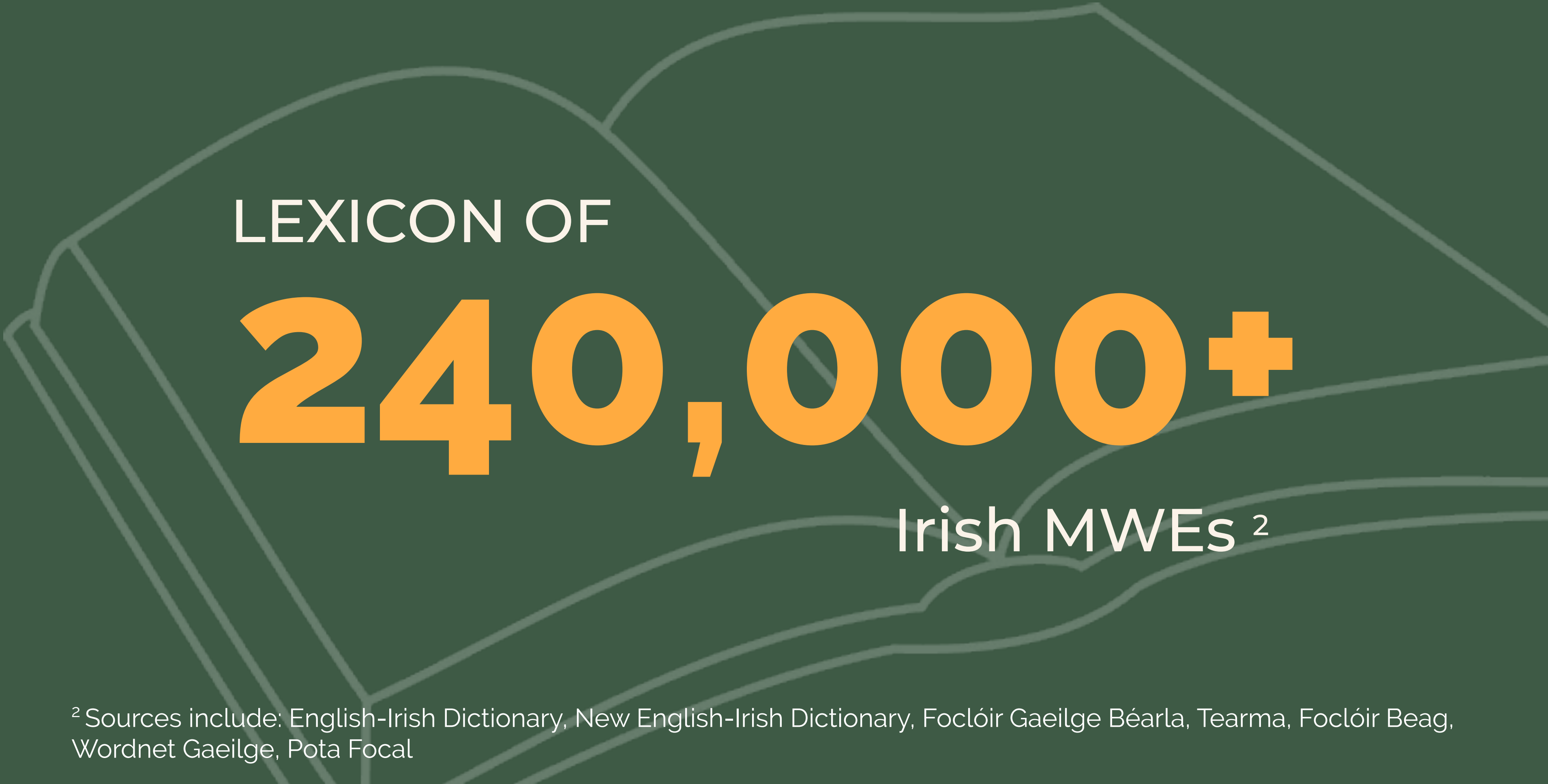


Categorisation of MWEs in Irish

Building lexicon of MWEs in Irish

Experiments on automatic extraction of MWEs

System for automatic identification of MWEs in Irish



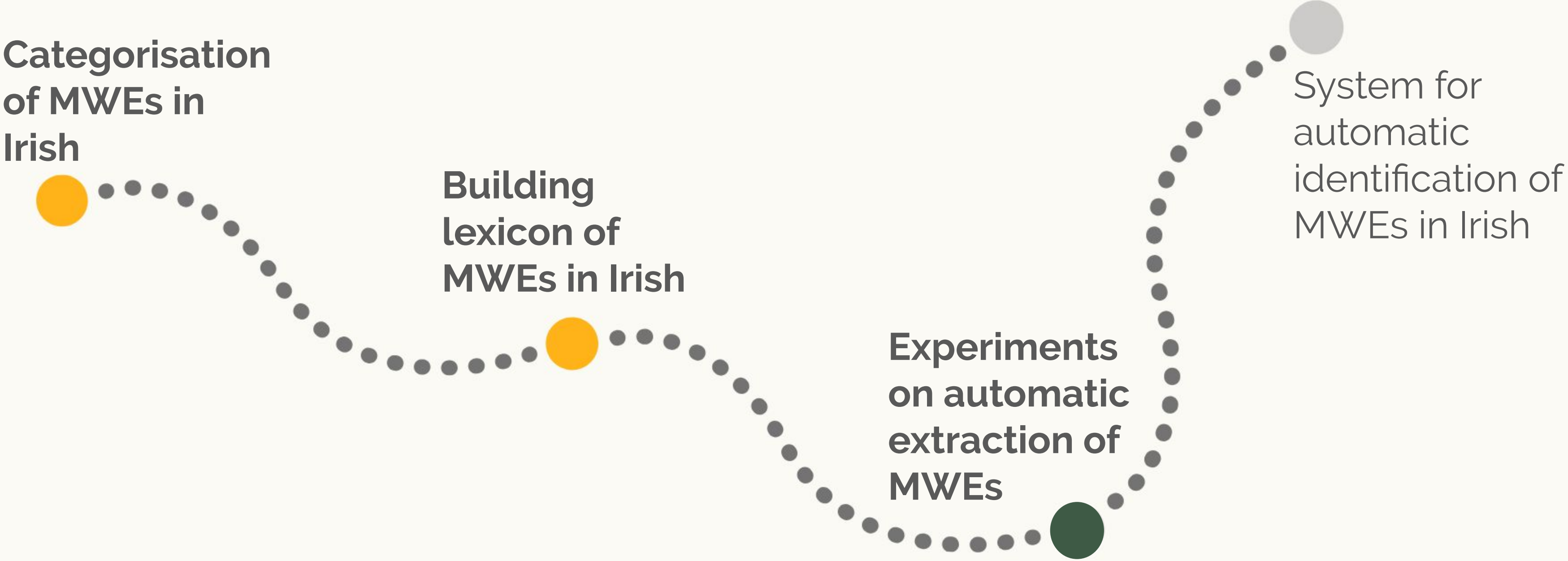
LEXICON OF

240,000+

Irish MWEs [2]

[2] Sources include: English-Irish Dictionary, New English-Irish Dictionary, Foclóir Gaeilge Béarla, Tearma, Foclóir Beag, Wordnet Gaeilge, Pota Focal

Road Map



Categorisation of MWEs in Irish

Building lexicon of MWEs in Irish

Experiments on automatic extraction of 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

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

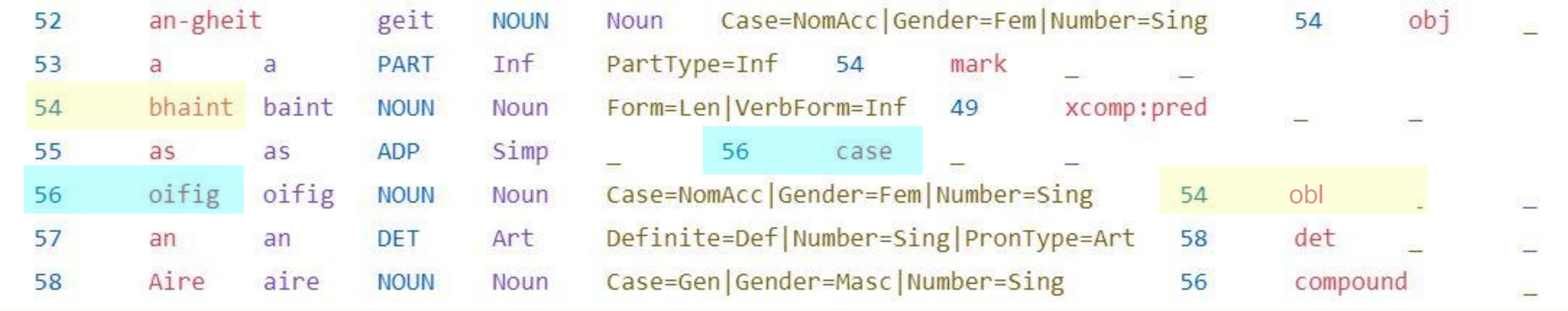
Universal Dependency Relations

- MWEs are labelled in UD as fixed, flat and compound
    - Fixed and compound relations allow for certain types of Irish MWEs
- Extraction of constructions using UD information
    - Verb-Particle Constructions, Compound Nouns, Compound Prepositions, Light-verb Constructions?

3

How can we process MWEs in Irish?

Universal Dependency Relations



```
52    an-gheit        geit    NOUN    Noun    Case=NomAcc|Gender=Fem|Number=Sing    54    obj    _
53    a       a       PART    Inf     PartType=Inf    54      mark    _       _
54    bhaint  baint   NOUN    Noun    Form=Len|VerbForm=Inf   49      xcomp:pred      _       _
55    as      as      ADP     Simp    _       56      case    _       _
56    oifig   oifig   NOUN    Noun    Case=NomAcc|Gender=Fem|Number=Sing      54      obl     _       _
57    an      an      DET     Art     Definite=Def|Number=Sing|PronType=Art   58      det     _       _
58    Aire    aire    NOUN    Noun    Case=Gen|Gender=Masc|Number=Sing        56      compound        _
```

MWEs in Machine Translation for Irish

- Encoding MWEs in Neural EN↔GA Machine Translation
- Two experiments:
    - Encoding uncategorised fixed MWEs (large lexicon)
    - Encoding four categories of semi-fixed MWEs (small lexicon)
        - Test different domains for different categories of MWEs
- Collecting MWEs for labelling dataset

Road Map

Categorisation of MWEs in Irish

Building lexicon of MWEs in Irish

Experiments on automatic extraction of MWEs

System for automatic identification of MWEs in Irish

System for Automatic Identification of MWEs in Irish

- Information used for MWE identification
  - Statistical (association measures)
  - Linguistic analysis (POS, lemmas)
    - VPCs captured with linguistic analysis
    - NNs, Compound Prepositions using statistical
    - IAVs, LVCs using both
- How to capture idiomaticity?
  - Idioms, copular constructions, LVCs

System for Automatic Identification of MWEs in Irish

- Features for identification come from this information
  - POS, PMI scores, etc.
- Compare traditional ML methods using feature engineering, and neural methods using pre-trained word embeddings
- Combine best of both worlds

Go raibh maith agaibh