

# Tapadóir: Statistical Machine Translation for Irish

Meghan Dowling

ADAPT Centre, School of Computing, Dublin City University, Ireland

29th April 2015



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

# Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion

# Outline

## Motivation

## The Tapadóir Project

## Results

## Ongoing and Future Work

## Conclusion

# Motivation for implementing a SMT system

- ▶ Current demand for translated content exceeds the capacity of the DAHG's translation team
- ▶ Previously:
  - ▶ SDL Trados (TM)
  - ▶ Not suitable for previously unseen text
  - ▶ If low % TM match, must be translated from scratch
- ▶ Clear need for something which would speed up the translator's work, while also being domain-specific

# Motivation for implementing a SMT system

- ▶ Current demand for translated content exceeds the capacity of the DAHG's translation team
- ▶ Previously:
  - ▶ SDL Trados (TM)
  - ▶ Not suitable for previously unseen text
  - ▶ If low % TM match, must be translated from scratch
- ▶ Clear need for something which would speed up the translator's work, while also being domain-specific

## Motivation for implementing a SMT system

- ▶ Current demand for translated content exceeds the capacity of the DAHG's translation team
- ▶ Previously:
  - ▶ SDL Trados (TM)
  - ▶ Not suitable for previously unseen text
  - ▶ If low % TM match, must be translated from scratch
- ▶ Clear need for something which would speed up the translator's work, while also being domain-specific

# Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion

# The Tapadóir Project

## Main goals:

- ▶ Build a SMT engine tailored to the translation needs of the DAHG
  - ▶ Domain-specific
  - ▶ User-friendly and non-disruptive to the usual workflow
- ▶ Gather and curate appropriate data
- ▶ Explore algorithmic parameters to best fit the language(s) and use case



# The Tapadóir Project

## Main goals:

- ▶ Build a SMT engine tailored to the translation needs of the DAHG
  - ▶ Domain-specific
  - ▶ User-friendly and non-disruptive to the usual workflow
- ▶ Gather and curate appropriate data
- ▶ Explore algorithmic parameters to best fit the language(s) and use case

# The Tapadóir Project

## Main goals:

- ▶ Build a SMT engine tailored to the translation needs of the DAHG
  - ▶ Domain-specific
  - ▶ User-friendly and non-disruptive to the usual workflow
- ▶ Gather and curate appropriate data
- ▶ Explore algorithmic parameters to best fit the language(s) and use case

## Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate

- ▶ reports, staff notices, communications, annual reports, etc
- ▶ a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement

## Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate

- ▶ reports, staff notices, communications, annual reports, etc
- ▶ a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement

## Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate

- ▶ reports, staff notices, communications, annual reports, etc
- ▶ a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement

# Acquiring Data

- ▶ Previous publicly available corpora:
  - ▶ **Parallel English–Irish corpus of legal texts (Paradocs)**<sup>1</sup>:  
English-Irish corpus of legal texts
  - ▶ **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**<sup>2</sup>:  
Bilingual corpus crawled from the web
- ▶ Corpora gathered by Tapadóir:
  - ▶ **General domain data**: Data crawled using the ILSP web-crawler<sup>3</sup>
  - ▶ **Domain-specific data**: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

---

<sup>1</sup>[www.gaois.ie](http://www.gaois.ie)

<sup>2</sup><http://borel.slu.edu/corpas/>

<sup>3</sup><http://nlp.ilsp.gr/redmine/projects/ilsp-fc>

# Acquiring Data

- ▶ Previous publicly available corpora:
  - ▶ **Parallel English–Irish corpus of legal texts (Paradocs)**<sup>1</sup>:  
English-Irish corpus of legal texts
  - ▶ **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**<sup>2</sup>:  
Bilingual corpus crawled from the web
- ▶ Corpora gathered by Tapadóir:
  - ▶ **General domain data**: Data crawled using the ILSP web-crawler<sup>3</sup>
  - ▶ **Domain-specific data**: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

---

<sup>1</sup>[www.gaois.ie](http://www.gaois.ie)

<sup>2</sup><http://borel.slu.edu/corpas/>

<sup>3</sup><http://nlp.ilsp.gr/redmine/projects/ilsp-fc>

# Acquiring Data

- ▶ Previous publicly available corpora:
  - ▶ **Parallel English–Irish corpus of legal texts (Paradocs)**<sup>1</sup>:  
English-Irish corpus of legal texts
  - ▶ **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**<sup>2</sup>:  
Bilingual corpus crawled from the web
- ▶ Corpora gathered by Tapadóir:
  - ▶ **General domain data**: Data crawled using the ILSP web-crawler<sup>3</sup>
  - ▶ **Domain-specific data**: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

---

<sup>1</sup>[www.gaois.ie](http://www.gaois.ie)

<sup>2</sup><http://borel.slu.edu/corpas/>

<sup>3</sup><http://nlp.ilsp.gr/redmine/projects/ilsp-fc>



## Increased bilingual corpora

Table of Irish-English corpora gathered

<b>Corpus</b>	<b>Size(lines)</b>	<b>Size(words)</b>
Paradocs	89,000	1,526,498
CCGB	6,000	113,889
DAHG (baseline)	29,000	67,418
DAHG (Additional)	13,500	68,691
Crawled (cleaned)	10,000	183,999
Crawled (uncleaned)	55,000	1,062,942
DCEP & DGT-TM	29,000	439,262

**Table 1** : Current data sets collected for Irish↔English MT. Word counts given for the English files only.

## Developing a test set

- ▶ 1500 sentences of gold-standard TM data provided by DAHG
- ▶ Domain-specific
- ▶ Representative of the type of text Tapadóir will need to translate
- ▶ Used to evaluate using automatic metrics

# SMT Engine

- ▶ Built using Moses 2.0 (Baseline) and Moses 3.0 (Current)
- ▶ 6-gram language model
- ▶ Trained using combinations of previously mentioned corpora
- ▶ KenLM

# System Setting Modifications

## ▶ **Tuning**

- ▶ Performed on a held-out section of 3000 domain-specific sentence pairs

## ▶ **Hierarchical-based model**

- ▶ Reordering table changed from phrase-based to hierarchical
- ▶ Better able to handle larger ordering differences – treats adjacent pairs as one unit

# System Setting Modifications

## ▶ **Tuning**

- ▶ Performed on a held-out section of 3000 domain-specific sentence pairs

## ▶ **Hierarchical-based model**

- ▶ Reordering table changed from phrase-based to hierarchical
- ▶ Better able to handle larger ordering differences – treats adjacent pairs as one unit

# Post-Processing Module

## Automated Post-Editing

- ▶ corrects common MT mistakes
- ▶ uses Irish surface orthography rather than deeper morphological analysis
- ▶ aims to improve grammar and readability
- ▶ *ag mé\** → *agam*  
*ag an baile\** → *ag an mbaile*

# Outline

Motivation

The Tapadóir Project

**Results**

Ongoing and Future Work

Conclusion

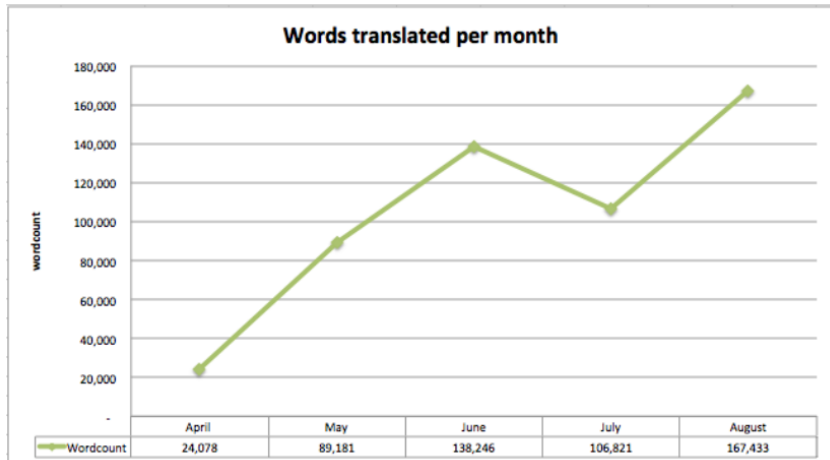
# System Results

	<b>BLEU</b>
Google	33.91
Tapadóir (Baseline)	39.44
<b>Tapadóir (Current)</b>	<b>43.19</b>

**Table 2 :** Comparison of results for Google Translate, our baseline system and our current system



# Increased MT usage



# Integration into DAHG translation workflow

The screenshot displays the DAHG translation workflow interface. The main window is titled "Editor" and shows a bilingual document with source text in Irish and target text in English. The source text includes: "Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: Gnáthscríbhín an Aonaid Iconic Plugin - Transl... 12/8/2014 5:38:56 PM TSR/vrabbhatagha". The target text includes: "the young performers, who range in age from 10 to 15, have been drawn from across the island of Ireland, and are all exceptional and award-winning artists and performers. Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: "This promises to be an extremely exciting and uplifting event; the ideal start to our year of commemorations in 2016. The Peace Proms will be a wonderful occasion, bringing together talented and committed young musicians from across the island of Ireland. I would like to thank all of the young performers who will be taking part; through their music and song they will set the perfect tone as kick off what is set to be a very special year."

The interface includes several toolbars and panels:

- Project Settings:** Configuration, Clipboard, Batch Tasks, File Actions, Formatting, QuickInsert, Concordance Search, Translation Memory, Terminology, and Segment Actions.
- Term Recognition:** A list of terms including "speaking", "speaking announcement", "announcement", "peace", "pron", "pron", "programmable read-only memory", "PRCM", and "Minister Humphreys".
- Translation Memory Table:**

Source Text	Match	Target Text
Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said:	71%	Agus í ag labhairt ag seoladh an leabhair inniu, dúirt an tAire Mhícheál Ó Súilleabháin:
Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said:	AT	Ag labhairt di nuair a fógraíodh an peace Proms, dúirt an tAire Heather Mhícheál Ó Súilleabháin:
- Translation Details Panel:** Shows "Status: Draft", "Origin: Interactive", and "Score: 0%".
- Bottom Status Bar:** Displays "All segments INS 53.25% 32.06% 14.65% Chars: 0 0/2630" and a language indicator for Irish/English.

# Integration into DAHG translation workflow

Project Settings | >

Speaking at the [announcement](#) of the [Peace Proms](#), **Minister Heather Humphreys** said:

1	Speaking at the <a href="#">leathorannouncement</a> of the <a href="#">Peace bookProms</a> , <b>Minister Heather Humphreys</b> said:	71% 	Agus í ag labhairt ag seoladh an leabhair inniu, dúirt an tAire Mhíic Unfraidh:
2	Speaking at the announcement of the Peace Proms, <b>Minister Heather Humphreys</b> said:		Ag labhairt di nuair a fógraíodh an peace Proms, <b>dúirt an tAire Heather Mhíic Unfraidh</b> :

Gnáthstrúichín an Aonaid 12/8/2014 5:38:56 PM TSR/viabhartaghe

Gnáthstrúichín an Aonaid.Icnic Plugin - Transl... Gnáthstrúichín an Aonaid.Icnic Plugin - Conco... Comments TQAs (0) Messages Term Recognition Termbase Search

PR Peace Proms - Gaeilge.docx.edbdfi! [Translation]

7	the young performers, who range in age from 10 to 15, have been drawn from across the island of Ireland, and are all exceptional and award-winning artists and performers.	ragairín na taibheoirí óga, atá tuim 10 agus 15 bliana d'aois, as gach cearn d'oileán na hÉireann, agus is ealaíontóirí agus taibheoirí den scoth, a bhfuil gradaim buaite acu, gach duine díobh.	
8	Speaking at the announcement of the Peace Proms, <b>Minister Heather Humphreys</b> said:		P
9	"This promises to be an extremely exciting and uplifting event; the ideal start to our year of commemorations in 2016.		P
10	The Peace Proms will be a wonderful occasion, bringing together talented and committed young musicians from across the island of Ireland.		
11	I would like to thank all of the young performers who will be taking part; through their music and song they will set the perfect tone as kick off what is set to be a very special year."		

Translation Details:  
Status: Draft  
Origin: Interactive  
Score: 0%

# Integration into DAHG translation workflow

Gnáthaisríúcháin an Aonaid, Iconic Plugin - Translation Results 🔍 ✕

Project Settings | 📄 📄 📄 >

Speaking at the [announcement](#) of the [Peace Proms](#), [Minister Heather Humphreys](#) said:

1	Speaking at the <a href="#">launch announcement</a> of the <a href="#">Peace book Proms</a> , <a href="#">Minister Heather Humphreys</a> said:	71% 📄	Agus í ag labhairt ag seoladh an leabhair inniu, dúirt an tAire Mhic Unfraidh:
2	Speaking at the announcement of the Peace Proms, <a href="#">Minister Heather Humphreys</a> said:	AT	Ag labhairt di nuair a fógraíodh an peace Proms, dúirt an tAire Heather Mhic Unfraidh:

## Typical MT errors

- ▶ Incorrect case used
- ▶ Inaccurate reordering - especially with longer sentences
- ▶ Copula constructions

# Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion

# Ongoing and Future Work

- ▶ Gather additional general-domain parallel data

- ▶ **Source-side re-ordering**

To address divergent word order (SVO–VSO)

**SRC:** The timeframe **can** be extended

**RO:** **can** The timeframe be extended

**REF:** 'Is féidir an tráthchlár a shíneadh'

- ▶ **Factored Models**

Using a POS-tagger (Uí Dhonnchadha), build a language model which contains grammatical information

## Ongoing and Future Work

- ▶ Gather additional general-domain parallel data

- ▶ **Source-side re-ordering**

To address divergent word order (SVO–VSO)

**SRC:** The timeframe **can** be extended

**RO:** **can** The timeframe be extended

**REF:** '**Is féidir** an tráthchlár a shíneadh'

- ▶ **Factored Models**

Using a POS-tagger (Uí Dhonnchadha), build a language model which contains grammatical information



## Ongoing and Future Work

- ▶ Gather additional general-domain parallel data

- ▶ **Source-side re-ordering**

To address divergent word order (SVO–VSO)

**SRC:** The timeframe **can** be extended

**RO:** **can** The timeframe be extended

**REF:** 'Is féidir an tráthchlár a shíneadh'

- ▶ **Factored Models**

Using a POS-tagger (Uí Dhonnchadha), build a language model which contains grammatical information

# Ongoing and Future Work

- ▶ DAHG vision for a Shared Translation Service
- ▶ Interest from DGT
- ▶ ELRC helping data gathering cause
  - ▶ DAHG helping to identify and gather public admin data from around the country

# Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion

# Conclusion

- ▶ Clear need for further research in English-Irish MT
- ▶ Clear need for new methods in overcoming the morphology problem/ lack of data problem

# Conclusion

Go Raibh Maith Agaibh!  
(Thank You!)