Tapadóir: Statistical Machine Translation for Irish

Meghan Dowling

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

29th April 2015

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

Outline

Outline

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

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)**[1]**:** English-Irish corpus of legal texts
  - **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**[2]**:** Bilingual corpus crawled from the web
- Corpora gathered by Tapadóir:
  - **General domain data:** Data crawled using the ILSP web-crawler[3]
  - **Domain-specific data:** Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

[1] www.gaois.ie
[2] http://borel.slu.edu/corpas/
[3] http://nlp.ilsp.gr/redmine/projects/ilsp-fc

Acquiring Data

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

[1] www.gaois.ie
[2] http://borel.slu.edu/corpas/
[3] http://nlp.ilsp.gr/redmine/projects/ilsp-fc

Acquiring Data

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

[1] www.gaois.ie
[2] http://borel.slu.edu/corpas/
[3] http://nlp.ilsp.gr/redmine/projects/ilsp-fc

Increased bilingual corpora

Table of Irish-English corpora gathered

Corpus	**Size(lines)**	**Size(words)**
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*

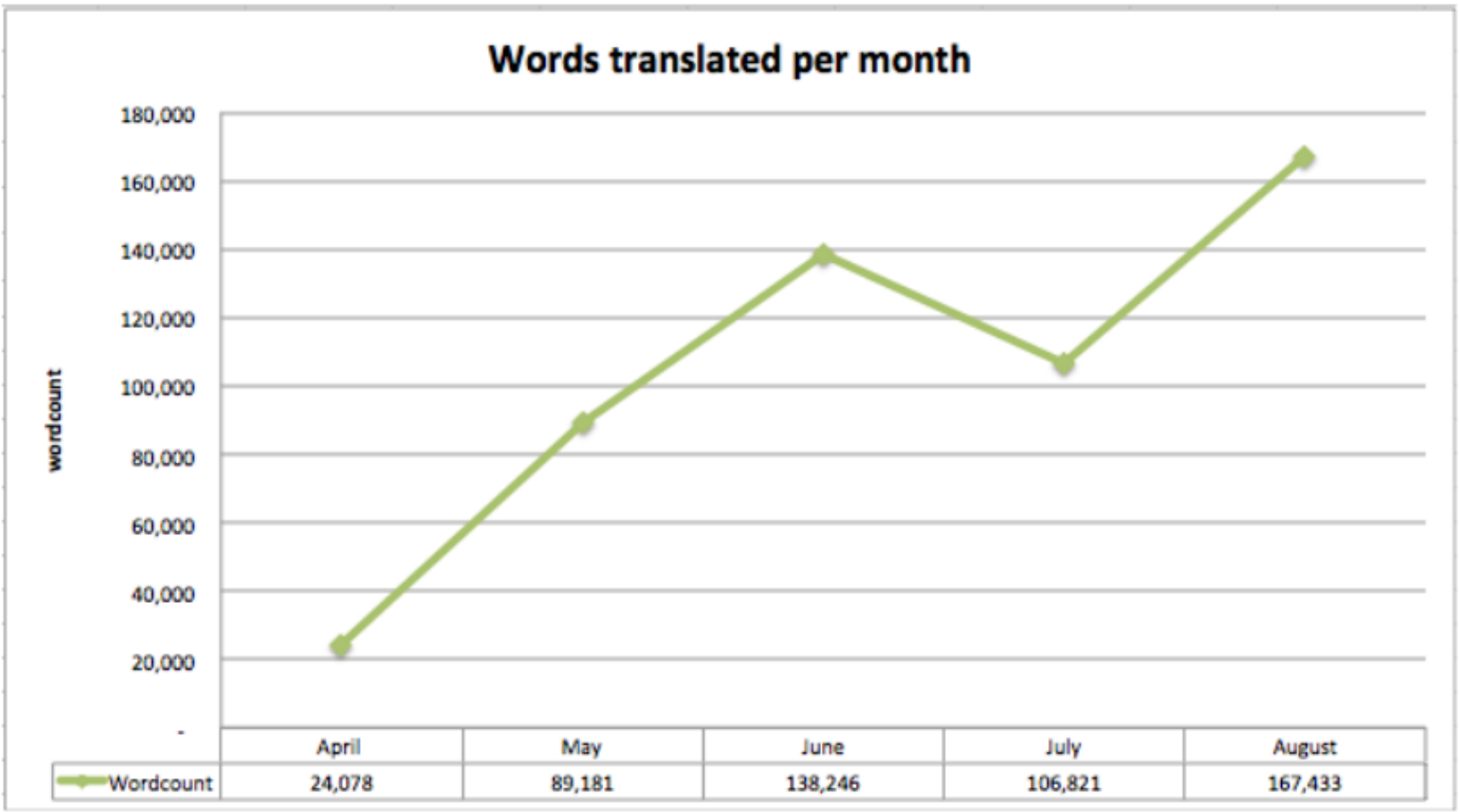
Outline

System Results

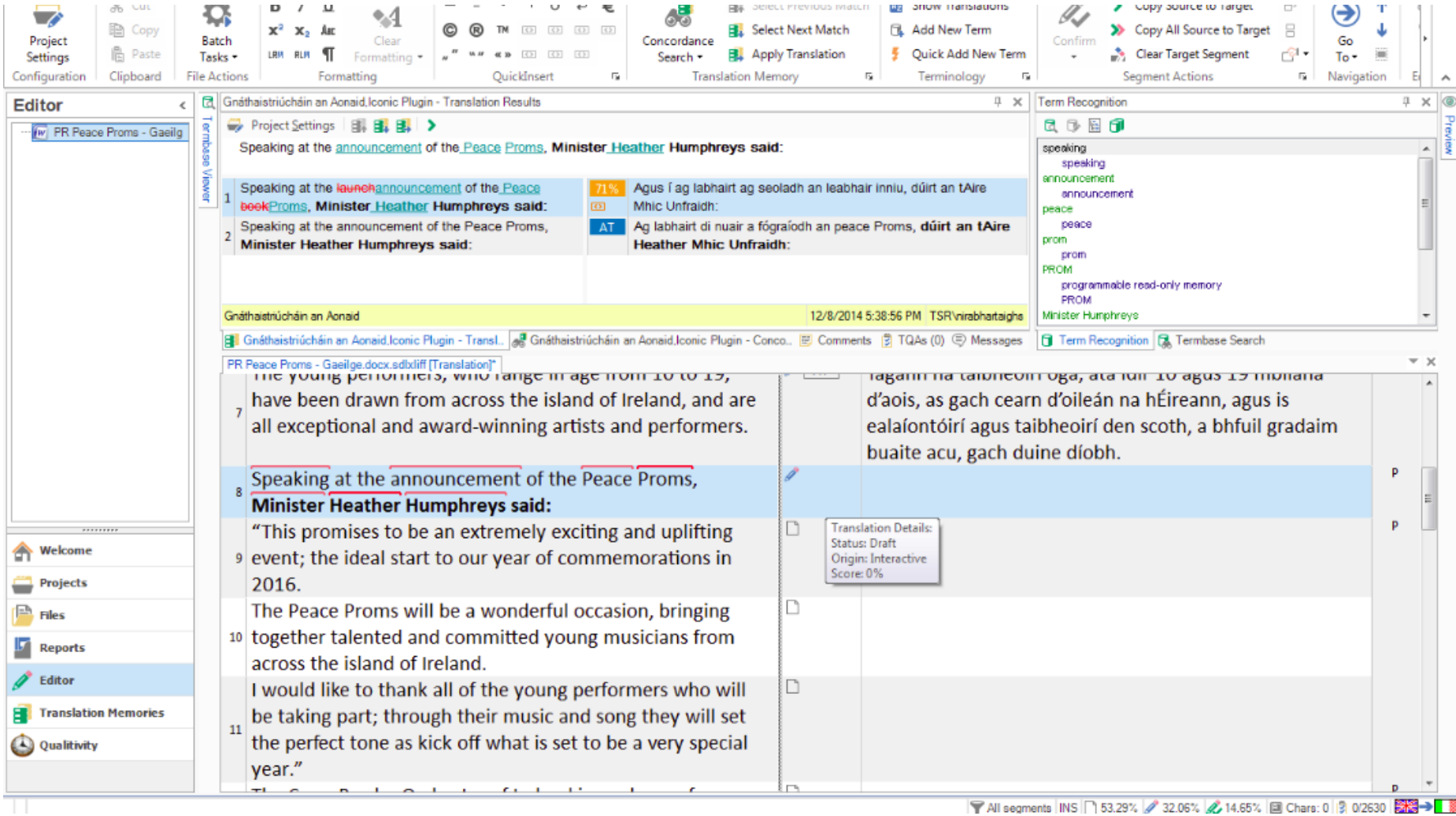
	BLEU
Google	33.91
Tapadóir (Baseline)	39.44
Tapadóir (Current)	43.19

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

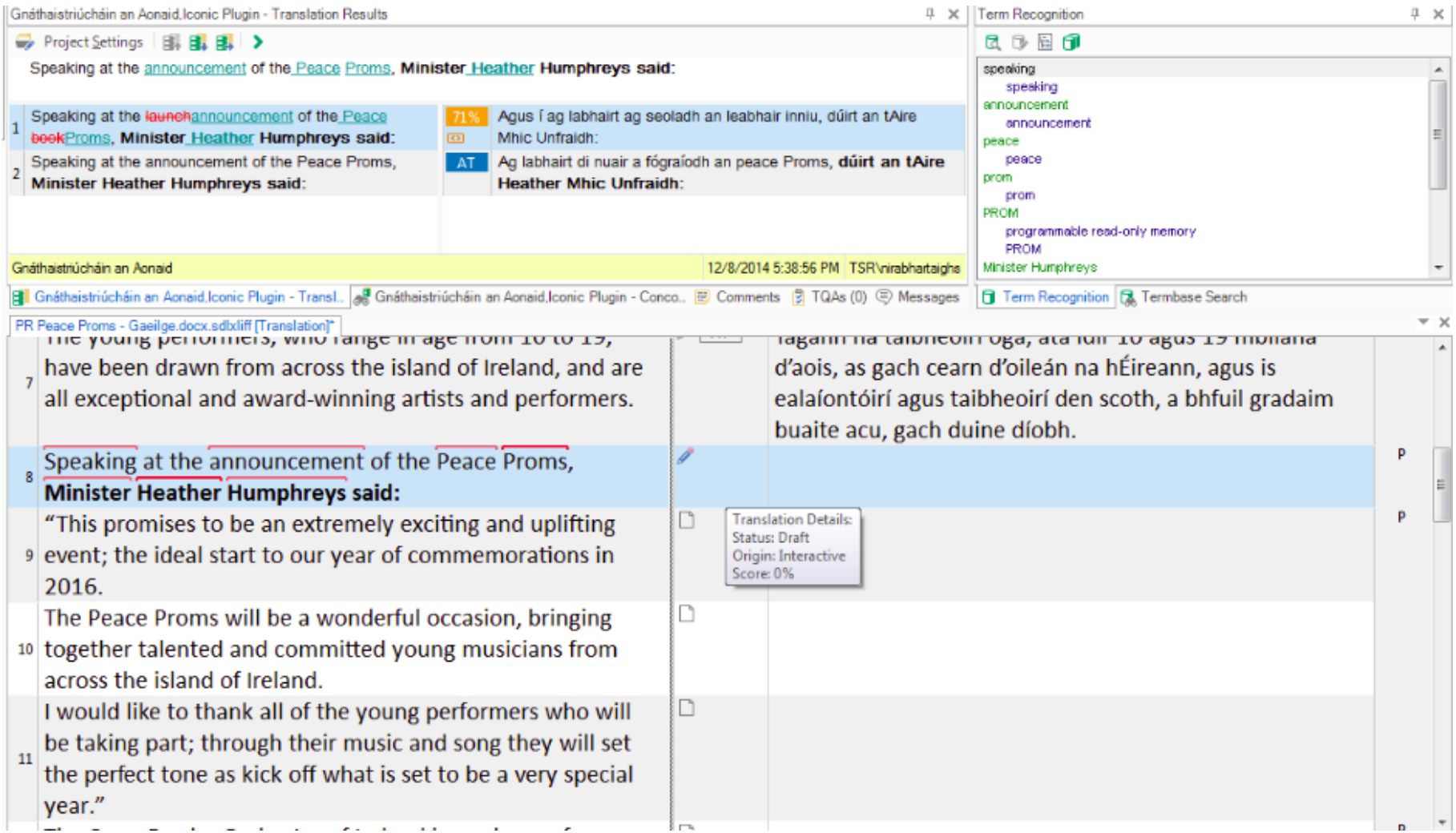
Increased MT usage



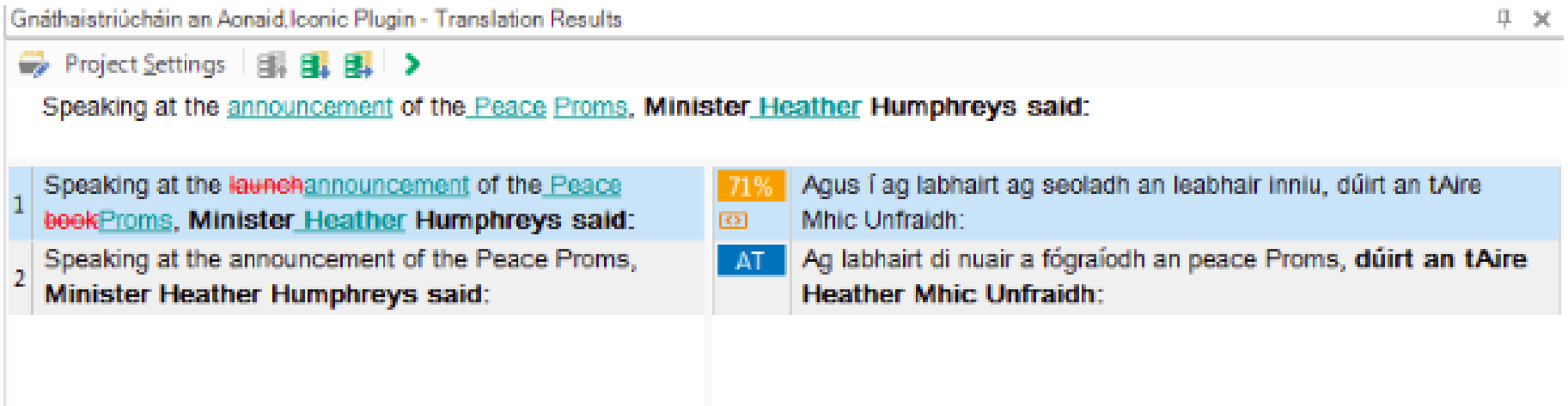
Integration into DAHG translation workflow



Integration into DAHG translation workflow



Integration into DAHG translation workflow



Typical MT errors

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

Outline

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

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!)