NLP for Old Irish

Development of Natural Language Processing Techniques and Resources for Old Irish with an Application for the Detection of Authors in the Würzburg Glosses

Candidate: Adrian Doyle
Supervisors: Dr. Clodagh Downey, Dr. John McCrae
Old Irish

• 7th c. – 10th c.

• Surviving Corpora:
  • Würzburg Glosses (mid 8th c.)
  • Milan Glosses (early 9th c.)
  • St. Gall Glosses (mid 9th c.)
Digital Resources Available

- Corpus of Electronic Texts (CELT)
- Parsed Old and Middle Irish Corpus (POMIC)
- Databases:
  - Milan (Griffith. Et al., Vienna)
  - St. Gall (Moran. Et al., NUIG)
- Wurzburg:
  - [www.wuerzburg.ie](http://www.wuerzburg.ie)
The Würzburg Glosses

• 3,511 Interlinear glosses on the text of the Pauline Epistles
• 1,885 glosses with code-switching (Irish – Latin)
• Three distinct scribal hands apparent
• Glosses assumed to have been copied from earlier sources.
• Some glosses from “about or prior to 700 A.D.” ¹ Latest Glosses not later than Milan Glosses (early 9th century)

¹ Thesaurus Palaeohibernicus (1901), Stokes & Strachan (eds.). Volume 1. Dublin Institute for Advanced Studies. p. xxiv.
Research Questions

1. How can a digital corpus be compiled which preserves the orthographic details of the manuscript text?

2. How can tokenisation of the text be standardised so as to account for the inconsistencies of manuscript word-spacing?

3. What is a reliable standard for part-of-speech tagging?
Research Questions

1. How can a digital corpus be compiled which preserves the orthographic details of the manuscript text?

F. 14d, G. 26
• .i. isipersin crist dagníusa sin – (p.596) ¹
• [ie].i.[/ie] isipersin crist dagníusa sin

F. 26d, G. 19
• .i. isómórdethiden file domsa diibsi – (p.670) ¹
• [ie].i.[/ie] isómórdethiden file domsa diibsi

¹ Thesaurus Palaeohibernicus (1901), Stokes & Strachan (eds.). Volume 1. DIAS.
Research Questions

2. How can tokenisation of the text be standardised so as to account for the inconsistencies of manuscript word-spacing?

F. 26d, G. 19
• .i. ismór indethiden file domsa diibsí – (p.670)₁
• Is mór in deithiden file dom-sa diib-sí. – (p.192)²
• .i. is mór in dethiden file dom -sa diib -si

F. 14d, G. 26
• .i. isipersin crist dagníusa sin – (p.596)₁
• .i. Is i persain Chríst da-gnìu-sa sin. – (p.144)²
• .i. is i persin crist d*a*gniu -sa sin

1. Thesaurus Palaeohibernicus (1901), Stokes & Strachan (eds.). Volume 1. DIAS.
Research Questions

3. What is a reliable standard for part-of-speech tagging?

F. 26d, G. 19
• .i. ismór indethiden file domsa diibsi – (p.670)
• .i. is món in dethiden file dom -sa diib -si

F. 14d, G. 26
• .i. isipersin crist dagníusa sin – (p.596)
• .i. is i persin crist d*a*gniu -sa sin

1. Thesaurus Palaeohibernicus (1901), Stokes & Strachan (eds.). Volume 1. DIAS.
Progress to Date

• Digitization of the Würzburg Glosses
  • Tag-set Created
  • Proofing Completed
  • www.wuerzburg.ie Online

• CCURL Workshop at LREC, 12th May 2018

• Tokenisation
  • Rules Set Out
  • Glosses for Gold Standard Collected and Tokenised
Current Work

• Complete Implementation of Tag-Set
• Tokenise the Glosses
  • Get Inter Annotator Agreement
    • 5+ Annotators
  • Create Tokeniser
    • Utilise St. Gall Database
    • Utilise Techniques Shown to Work in Languages where Spacing Is Not a Reliable Basis for Tokenisation

2. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop (2005), Habash, N. and Rambow, O. ACL.
Current Work

- Occurrences of tokens in Glosses
  - dub, dlub
  - cach, clach
  - on, ón
  - act, aclt
  - n, ñ

Each chart represents the occurrence frequency of different tokens across various Glosses.
Future Work

• Create POS Tagger
• POS Tag Corpus
• Implement and Assess Author Recognition Techniques

Go Raibh Maith Agaibh!
Thank You