Scottish Corpus of Texts and Speech

Linguistics

- First corpus to include all major languages of present-day Scotland (eventually including non-indigenous languages such as Chinese and Urdu)
- Phase 1: Concentrating on Scots and Scottish English
- Wide variety of text types (e.g. conversations, letters, articles, non-literary and literary texts)
- Language usage in modern media (emails, faxes, webpages etc.)
- Text, sound and video media
- Orthographic transcriptions available alongside sound files (useful educational resource)
- Problems with variant spellings to overcome
- Free access to all non-profit users
- Anticipated users include academics, those working in education, media, those generally interested in Scottish language and culture
- Problems associated with working with non-standard and relatively unexplored language varieties such as Scots
- Full texts not extracts wherever possible

THE UNIVERSITY OF EDINBURGH

UNIVERSITY of GLASGOW

VIA VERITAS VITA

www.scottishcorpus.ac.uk

The SCOTS project is grant-funded by the Engineering and Physical Sciences Research Council (grant GR/R32772/01). SCOTS is a joint venture by the Department of English Language and STELLA project, University of Glasgow, and the Language Technology Group, University of Edinburgh.

Scottish Corpus of Texts and Speech

Computing

- Administration database containing all author, document, copyright and tracking information (14 tables, 254 fields)
- Administration data held in secure SQL database with admin via MS Access gui front-end
- Publicly accessible document and author data exported as xml for web search system
- Web search system provided by Java servlets accessing xml dataset
- Texts searchable via metadata and lexis (author demographics, language, document type)
- Captioned sound and video streaming
- Development of xml language processing tools to aid searching variant spelling words (possible approaches: fuzzy search techniques or lemmatizing) (e.g. home = hame = haim)
- xml based strict html4 web site providing best in cross-platform support and accessibility

UNIVERSITY of GLASGOW

www.scottishcorpus.ac.uk

The SCOTS project is grant-funded by the Engineering and Physical Sciences Research Council (grant GR/R32772/01). SCOTS is a joint venture by the Department of English Language and STELLA project, University of Glasgow, and the Language Technology Group, University of Edinburgh.