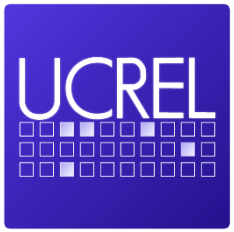


Lancaster University

We don't just count words, you know!
Or ...
Computational language analysis research @ Lancaster

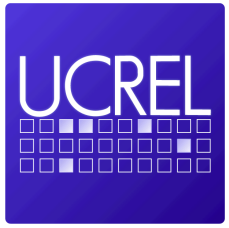
Dr Paul Rayson (@perayson) Director of UCREL research centre, School of Computing and Communications, Lancaster, UK.
Dr Alistair Baron (@barona), Lecturer, UCREL, SCC.

Motivation

Lancaster University

The availability and rapidly expanding nature of language corpora continues to outstrip our abilities as researchers carry out qualitative analysis on a large scale. Hence, new techniques and methods are required in order to support quantitative analysis while facilitating deep analysis in order to understand meaning and other important features of texts.

UCREL Research Centre



Lancaster University

- Incorporates postgraduate students, researchers, staff in *School of Computing and Communications* and *Linguistics and English Language Department*
- Major themes: Corpus Linguistics and Natural Language Processing
- Applications in Social Sciences, Humanities, Health & Medicine and other disciplines where large quantities of text can be studied
- Key outputs and methodological contributions: compiled and annotated corpora, software for corpus annotation and retrieval

Related centres and institutes

- ESRC Centre for Corpus Approaches to Social Science (CASS)
  - http://cass.lancs.ac.uk/
  - BNC2014, Climate Change discourse, Hate Speech
- Lancaster Digital Humanities
  - http://wp.lancs.ac.uk/dighum/
  - Spatial Humanities
- Data Science Institute (DSI)
  - http://www.lancaster.ac.uk/dsi/
- Security Centre
  - http://www.lancaster.ac.uk/security-lancaster/
  - Native Language Influence Detection (NLID), Language of extremism and counter-extremism, Author profiling, Centre for Research and Evidence on Security Threats (CREST)

Our Methodologies

- Natural Language Processing / Text Mining
  - Annotation: Part-of-speech tagging, Semantic Tagging, Word sense disambiguation
  - Variant spelling detection
  - Sentiment analysis
  - Text reuse
  - Named entity recognition
  - Language profiling
  - Unstructured text extraction
  - Dependency parsing
  - Information fusion
  - Network analysis
  - Topic modelling
  - Vector-based and NN approaches
  - Distributed / parallel processing
  - Machine Learning
  - GIS

- Corpus Linguistics
  - Frequency lists
  - Concordances
  - Keywords
  - N-grams / Clusters
  - Collocations

Spatial Humanities:

texts, geographic information systems and places

Lancaster University

European Research Council

erc

- Five year project (2012-16)
- Funded by European Research Council under starting researcher grant to Prof Ian Gregory (agreement number 283850)
- Building on technical expertise in Digital Humanities, Corpus and Computational Linguistics and Historical Geographical Information Systems (HGIS)
- Themes
  - Methodologies (GIS & text analysis)
  - Analysing qualitative sources (Lake District literature in 18th and 19th centuries; Histpop and Bopcris datasets; BL 19th century newspapers and books)
  - Developing the skills bases (training events and workshops)

SAMUELS project

http://www.gla.ac.uk/samuels/

Arts & Humanities Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Lancaster University

SAMUELS

- SAMUELS: Semantic Annotation and Mark-Up for Enhancing Lexical Searches
  - funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1)
  - January 2014 to March 2015
- Aims
  - delivered a system for automatically annotating words in texts with their precise meanings, disambiguating between possible meanings of the same word
  - provided for each word in a text the Historical Thesaurus of English reference code for that concept.
- Extremely large corpora:
  - Early English Books Online (EEBO) Text Creation Partnership (TCP) consisting of over 53,830 books published between 1473 and 1700 (1.27 billion words; Phase 2 November 2014 release)
  - Two hundred years of UK Parliamentary Hansard consisting of over 7 million files (~2 billion words)

Though I **speake** with the tongues of men & of Angels, and **haue** not charity, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **remooue mountaines**, and **haue** no **charitie**, I am nothing...

(Authorised Version of the Bible, 1611)

I The external world

01 The world

01.01	**The earth**
01.01.01	Region of the earth
01.01.02	Geodetic references
01.01.03	Direction
01.01.04	Land
01.01.05	Water
01.01.06	Named regions of earth
01.01.07	Structure of the earth
01.01.08	Minerals
01.01.09	Earth science
01.01.10	The universe
01.01.11	Atmosphere, weather
01.02	**Life**
01.02.01	Health and disease
01.02.02	Death
01.02.03	Biology
01.02.04	Plants
01.02.05	The body
01.02.06	Animals
01.02.07	People
01.02.08	Food and drink
01.02.09	Textiles

II The mental world

02 The mind

02.01	**Mental capacity**
02.01.01	Spirituality
02.01.02	Intellect
02.01.03	Consciousness
02.01.04	Disposition/character
02.01.05	The psyche
02.01.06	Thought
02.01.07	Perception/cognition
02.01.08	Understanding
02.01.09	Lack of understanding
02.01.10	Intelligibility
02.01.11	Memory
02.01.12	Knowledge
02.01.13	Belief

III The social world

03 Society

03.01	**Society/the community**
03.01.01	Kinship/relationship
03.01.02	Study of society
03.01.03	Society in relation to customs/values/ beliefs
03.01.04	Social communication/relations
03.01.05	Social attitudes
03.01.06	Social class/rank
03.01.07	Dissension/discord
03.02	**Inhabiting/dwelling**
03.02.01	Inhabiting type of place
03.02.02	Inhabiting/dwelling temporarily
03.02.03	Providing with dwelling place
03.02.04	Removing from dwelling place
03.02.05	Furnishing with inhabitants
03.02.06	Inhabitant/resident
03.02.07	Inhabited place
03.03	**Armed hostility**
03.03.01	War
03.03.02	Armed encounter

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Software Architecture for Mental Health Self Management (SAMS)

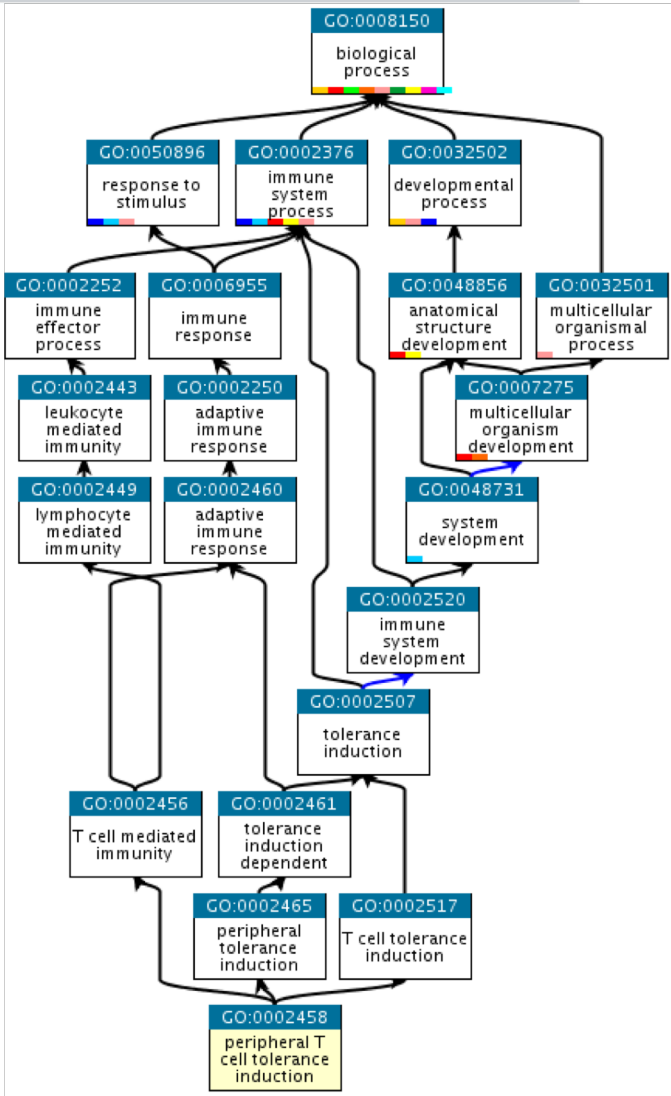
- EPSRC working together project EP/K015796/1
- Alistair Sutcliffe, Paul Rayson, Chris Bull and Pete Sawyer @ Lancaster
- Early diagnosis of dementia by monitoring interaction with a computer to monitor mental health
- circa 900K people in UK alone, only 44% receive a diagnosis
- combining data and text analysis
- http://ucrel.lancs.ac.uk/sams/

Metaphor in end-of-life care (MELC) project

- The way in which the experience of end of life (care) is talked about can shed light on people's views, needs, challenges, and emotions, as well as identify areas with a potential for increased anxiety and/or misunderstanding.
- This project's aim overall was to investigate how members of different stakeholders groups (patients, unpaid family carers and healthcare professionals) use metaphor to talk about their experiences, attitudes and expectations of end of life care (e.g. palliative treatment; preparations for dying).
- Jane Demmen, Andrew Hardie, Veronika Koller, Sheila Payne, Paul Rayson and Elena Semino, (Lancaster University) and Zsófia Demjén (Open University)
- 18 months funding from ESRC ES/J007927/1
- http://ucrel.lancs.ac.uk/melc/

Biomedical Text Mining

- Purpose and Motivation
  - Explosion of academic literature leading to stove-piping.
  - Researchers are no longer able to maintain knowledge of related areas.
  - Derive and compare corpora from bodies of genetics literature.
  - Develop techniques to provide new clues to disease aetiology.
- Apply text mining to all open access papers in PubMed
- Annotation with new Gene Ontology Semantic Tagger (GOST)
- FHM & SCC: Jo Knight, Scott Piao, Mahmoud El-Haj, Sheryl Prentice, Nathan Rutherford
- http://wp.lancs.ac.uk/btm/



Corporate Financial Information Environment (CFIE) projects x 3

- ESRC, ICAEW, FRC, FCA funded three projects
- Martin Walker, Manchester Business School
- Steven Young, Lancaster University Management School
- Paul Rayson, Lancaster University School of Computing & Communications
- Mahmoud El-Haj, Lancaster University School of Computing & Communications
- Vasiliki Athanasakou, London School of Economics
- Thomas Schleicher, Manchester Business School

- Project seeks to analyse UK financial narratives, their association with financial statement information, and their informativeness for investors

- Automated, large sample analysis of UK annual report narratives represents a cornerstone of the project
- Develop software for general use by academics
- http://ucrel.lancs.ac.uk/cfie/

Measuring online user sentiment around branding of products

- LU-Sunway University small grants (2014-15 and 2016-17)
- Fine grained sentiment analysis
- English and Malay language data
- Malay semantic analysis system

Cardiff University / Prifysgol Caerdydd

Swansea University / Prifysgol Abertawe

Lancaster University

Prifysgol Bangor University

Arts & Humanities Research Council

E·S·R·C Economic & Social Research Council

- Creation of a major language resource for Welsh speakers, Welsh learners, Welsh language researchers, and indeed anyone who is interested in the Welsh language. Multiple language samples will be gathered from real-life communication, and presented in a searchable online 'corpus', allowing users to explore Welsh as it is actually used.
- £1.8M funding (2016-19)
- http://www.corcencc.org/

Llywodraeth Cymru Welsh Government

LLGC NLW

saysomethinginWelsh

wjec cbac

ylLolfa www.ylolfa.com cyhoeddwyr ac argraffwyr

S4C

BBC cymru wales

Cynulliad Cenedlaethol Cymru / National Assembly for Wales

GPC

PhD projects

Lancaster University

- Ed Dearden: *Alternative Fakes: The Roles of Belief and Deceptive Intent in the Language of False Information*
- Rob Larson: Weak Signals as Predictors of Sophisticated Social Engineering Attacks
- Andrew Moore: *Target dependent sentiment analysis*
- Jawad Shafi: *Urdu semantic analysis* (with Adeel Nawab, Comsats)
- John Vidler: *GPU and graph pipeline architectures for NLP* (with Andrew Scott, SCC)
- Matt Coole: *Next generation database architectures for CL & NLP* (with John Mariani)
- Muhammad Sharjeel: *Cross lingual text reuse* (with Adeel Nawab, Comsats)
- Alex Reinhold*: Geospatial Innovation in the Digital Humanities* (with Ian Gregory in History)
- Henry Moss: *Improving cross validation methodology* (with David Leslie of M&S)
- Glorianna Jagfield: *Talking about personal recovery in bipolar disorder*, (with Steven Jones and Fiona Lobban of the Spectrum Centre)
- Lama Al-Sudias: *Using Twitter to support Digital Surveillance of Infectious Diseases in the Arab subcontinent*

My own research

- Applying Natural Language Processing (NLP) and Data Science techniques to (primarily) security and cybercrime issues.
- Particularly analysing language style for insights into authors and their intentions.
- Mostly concerned with online text (CMC), e.g. emails, chat, SMS, social media, etc., including the "noise" it contains.
- Who is behind online personas, deception, and influence.

Previous / current projects

Lancaster University

- Uncovering masquerading behaviour, especially fake online profiles used for grooming.
- Social engineering attack surfaces, and analysing social engineering vulnerabilities.
- Disinformation: “fake news”, conspiracies, “flat-earthers”, anti-vax.
- Native Language Influence Detection.

Research direction

Lancaster University

- Explainability of NLP / Machine Learning systems – not the what, but the why.
- “Human-in-the-loop” systems for decision support, e.g. for assisting with investigations, disaster recovery, and crisis management.
- Privacy-preserving.
- Change over time.