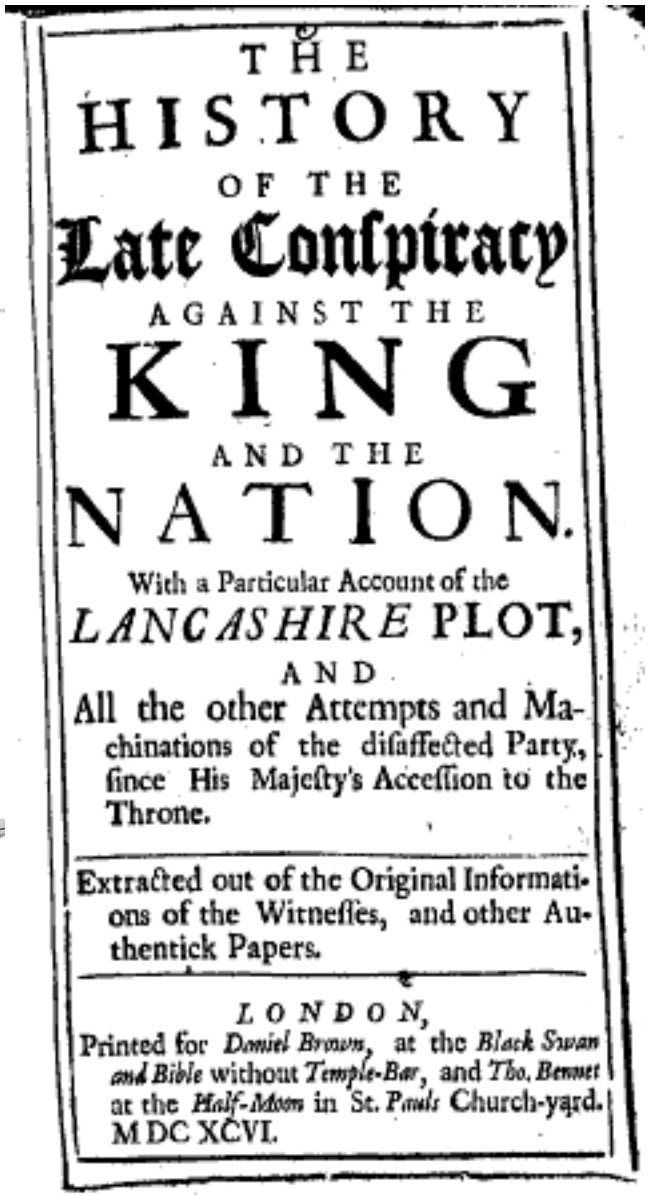
Adjusting a semantic taxonomy and annotation tool for historical corpora

Dr Paul Rayson @perayson
Director of UCREL research centre, School of Computing and Communications, Lancaster, UK

Joint work with Alistair Baron, Scott Piao, and Steve Wattam at Lancaster University, Dawn Archer (MMU) plus others from the Universities of Glasgow and Huddersfield, and OUP.

Slides at http://ucrel.lancs.ac.uk/paul/



THE

HISTORY

OF THE

Late Conſpiracy

AGAINST THE

KING

AND THE

NATION.

With a Particular Account of the

LANCASHIRE PLOT,

AND

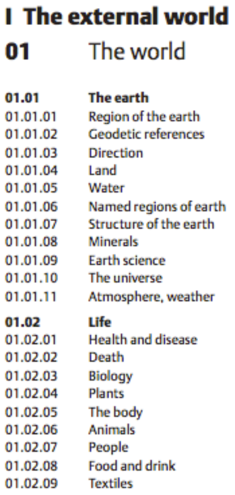
All the other Attempts and Machinations of the diſaffected Party, ſince His Majeſty's Acceſſion to the Throne.

Extracted out of the Original Informations of the Witneſſes, and other Authentick Papers.

LONDON,
Printed for *Daniel Brown*, at the *Black Swan and Bible* without *Temple-Bar*, and *Tho. Bennet* at the *Half-Moon* in St. *Pauls* Church-yard.
MDCXCVI.

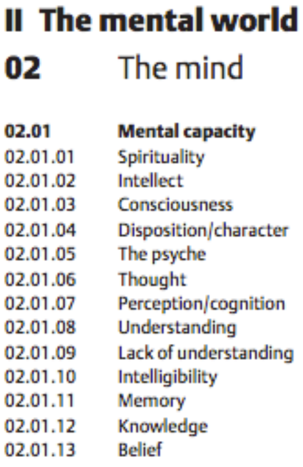
Though I **speake** with the tongues of men & of Angels, and **haue** not charity, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **remooue mountaines**,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

(Authorised Version of the Bible, 1611)



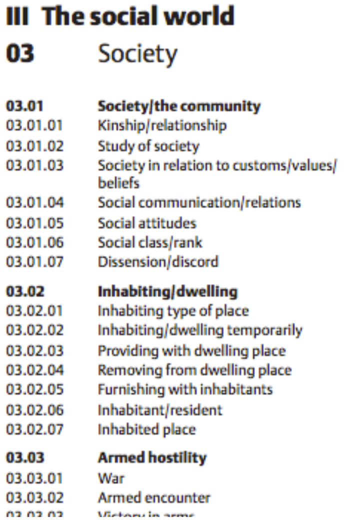
I The external world

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



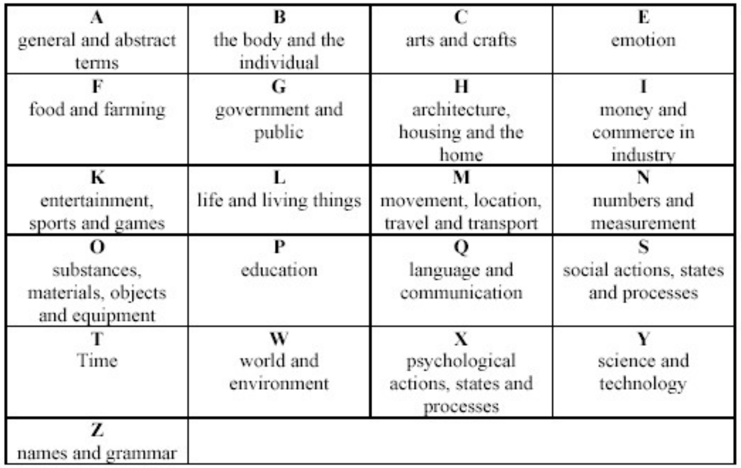
II The mental world

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
03.03.03	Victory in arms



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			

SAMUELS project

http://www.gla.ac.uk/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.
- Project team:
  - Lancaster: Alistair Baron, Scott Piao, Steve Wattam
  - University of Glasgow (lead institution), Lancaster University, University of Huddersfield, University of Central Lancashire, University of Strathclyde, Oxford University Press
  - international partners: Brigham Young University (Utah), Åbo Akademi University (Finland), and the University of Oulu (Finland).

Big Data Challenges

- Big 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)
- Big taxonomies:
  - Historical Thesaurus of English (developed at the University of Glasgow) and the Oxford English Dictionary to help us improve methods for the automatic semantic analysis of historical texts.
  - The Historical Thesaurus contains 793,742 word forms arranged into 225,131 semantic categories.

Big Data Challenges

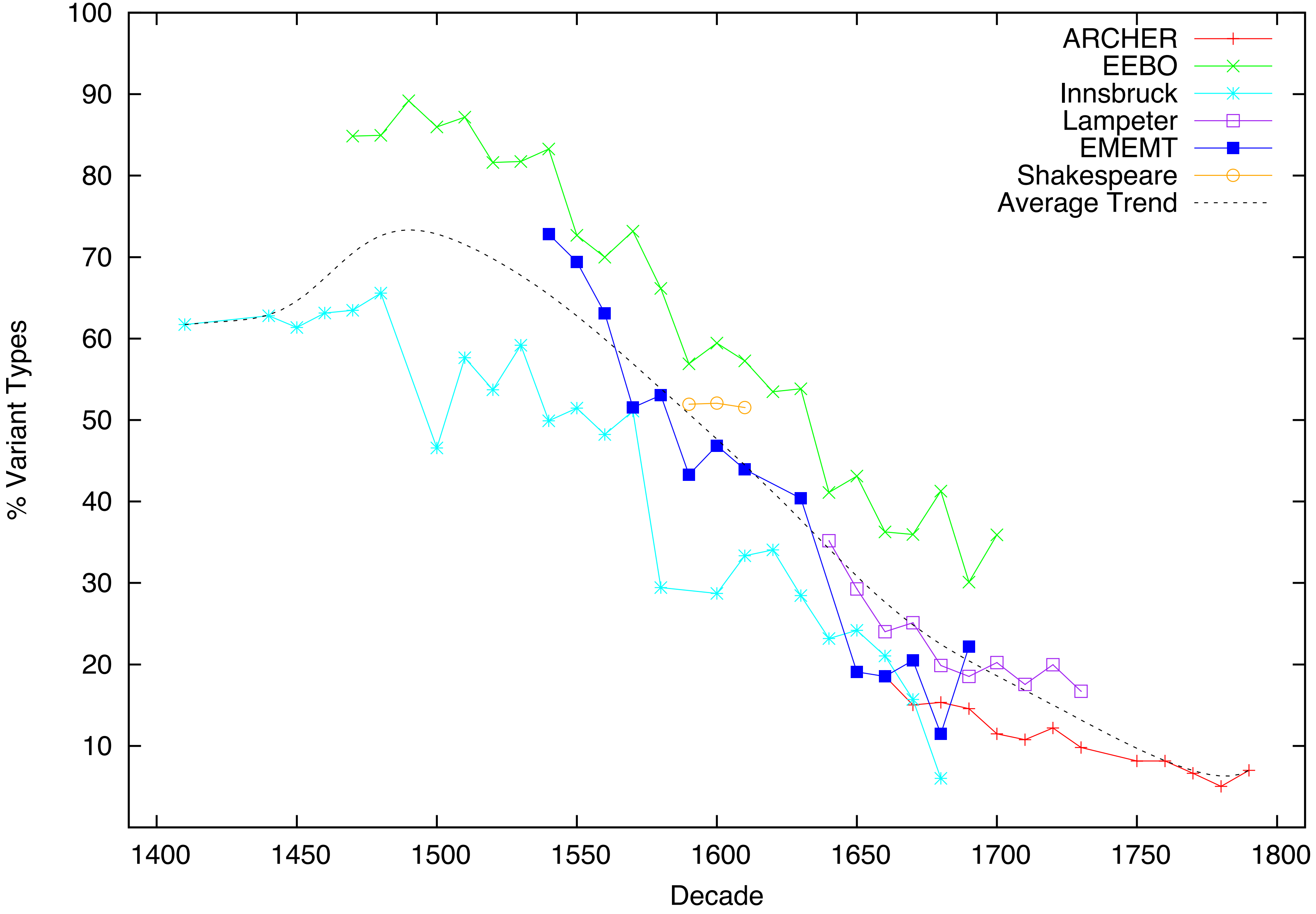
- The combination of scale (and historical nature) of the corpora and the taxonomy pose significant computational challenges for existing retrieval methods (Wmatrix) and annotation software (USAS)

- Our solutions
  - Variant Spelling methods
  - Improved semantic disambiguation techniques (Historical Thesaurus Semantic Tagger – HTST)
  - Use of big data methods e.g. cluster and cloud computing

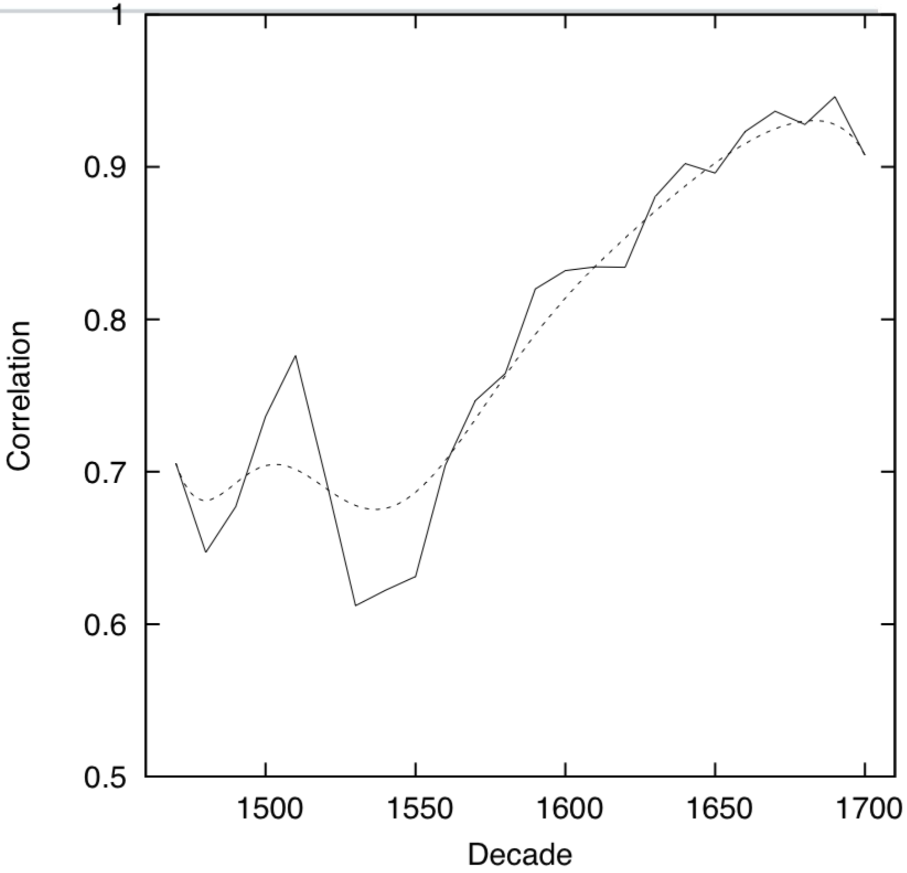
- Addition or removal of ‘e’, e.g. *aske, workes, dos*
- Doubling and singling of letters, e.g. *smels, heere, leggs*
- Interchanged letters: { u , v }, { j , i }, { ie , y }, { vv , w }, e.g. *haue, vnder, maiestie, vvas*
- Usage of apostrophe, e.g. *vow’d, ‘em*
- Spellings which are variable still today, e.g. *centre/center, -or/-our, -ise/-ize*
- Fused forms, e.g. *t’is, t’was, o’th*
- Archaic *–(e)th* and *–(e)st* endings, e.g. *hath, doth, seemeth, shouldst*
- Archaic forms, e.g. *betwixt, howbeit*
- Phonetic spellings, e.g. *publiquely, blew (blue)*
- + any combination of the above and other irregular spellings, e.g. *ligge (Jig), diuell (devil), shak’d (shook)*

The extent of spelling variation in EmodE corpora

- And its effect on corpus methods such as keywords
  - Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. In Anglistik: International Journal of English Studies, 20 (1), pp. 41-67.


ARCHER
EEBO
Innsbruck
Lampeter
EMEMT
Shakespeare
Average Trend
% Variant Types
100
90
80
70
60
50
40
30
20
10
1400
1450
1500
1550
1600
1650
1700
1750
1800
Decade

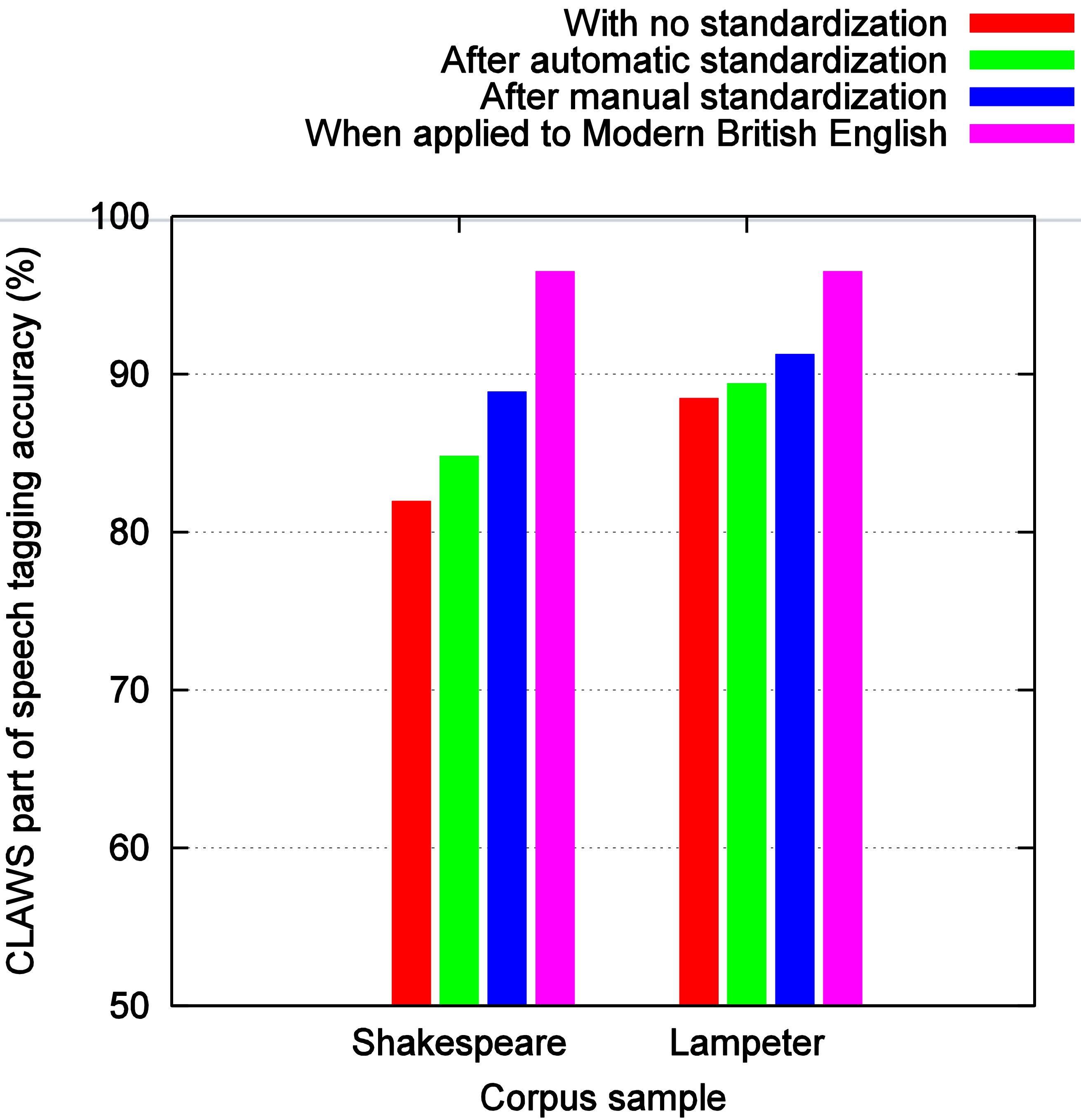
- Searching for words can be problematic: *would, wolde, woolde, wuld, wulde, wud, wald, vvould, vvold*, etc.
- Frequencies split by multiple spellings.
- Knock-on effect on key words (Baron *et al.*, 2009), key word clusters (Palander-Collin & Hakala, 2011) and collocates.



The need for normalisation …

- Automatic semantic analysis of EmodE corpora
  - Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.
- Automatic POS tagging of historical corpora
  - Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In proceedings of Corpus Linguistics 2007, July 27-30, University of Birmingham, UK.
- Corpus annotation in general
  - Rayson, P. (2007) Travelling through time with corpus annotation software. PALC2007 keynote talk.

Lancaster University

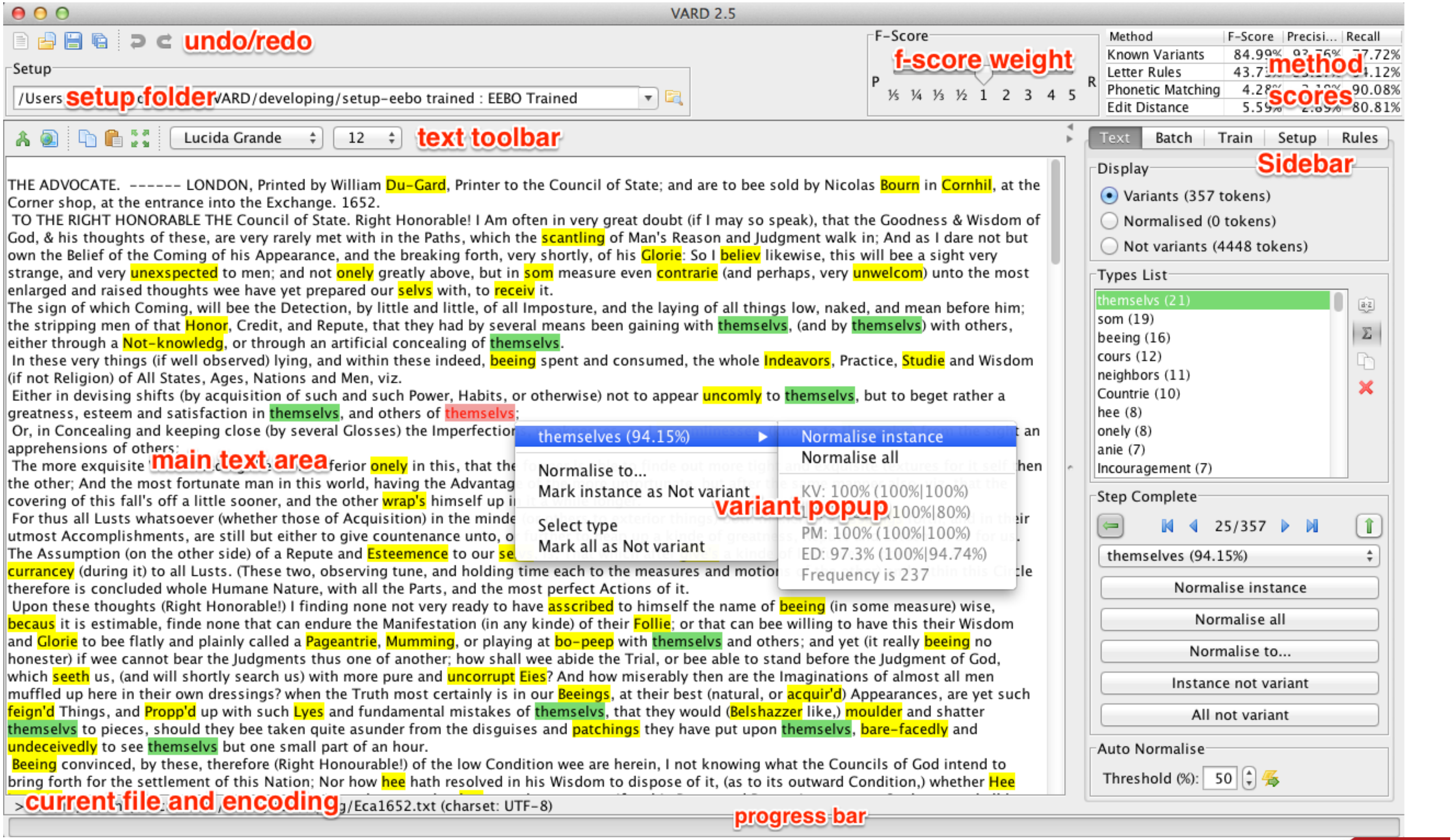


Development of VARD …

- Use of existing spell checking techniques
  - Rayson, P., Archer, D., Smith, N., (2005), VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. In Proceedings of Corpus Linguistics 2005, Birmingham University, July 14-17
- Hybrid methods
  - Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. In proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, 22nd May 2008.

VARD (VARiant Detector)

http://ucrel.lancs.ac.uk/vard/



Setup: /Users[…]VARD/developing/setup-eebo trained : EEBO Trained

F-Score: P ⅕ ¼ ⅓ ½ 1 2 3 4 5 R

Method	F-Score	Precisi...	Recall
Known Variants	84.99%	93.76%	77.72%
Letter Rules	43.7[illegible]	[illegible]	[illegible]4.12%
Phonetic Matching	4.28%	[illegible]	90.08%
Edit Distance	5.59%	2.89%	80.81%

Lucida Grande 12

THE ADVOCATE. ------ LONDON, Printed by William Du-Gard, Printer to the Council of State; and are to bee sold by Nicolas Bourn in Cornhil, at the Corner shop, at the entrance into the Exchange. 1652.
TO THE RIGHT HONORABLE THE Council of State. Right Honorable! I Am often in very great doubt (if I may so speak), that the Goodness & Wisdom of God, & his thoughts of these, are very rarely met with in the Paths, which the scantling of Man's Reason and Judgment walk in; And as I dare not but own the Belief of the Coming of his Appearance, and the breaking forth, very shortly, of his Glorie: So I believ likewise, this will bee a sight very strange, and very unexspected to men; and not onely greatly above, but in som measure even contrarie (and perhaps, very unwelcom) unto the most enlarged and raised thoughts wee have yet prepared our selvs with, to receiv it.
The sign of which Coming, will bee the Detection, by little and little, of all Imposture, and the laying of all things low, naked, and mean before him; the stripping men of that Honor, Credit, and Repute, that they had by several means been gaining with themselvs, (and by themselvs) with others, either through a Not-knowledg, or through an artificial concealing of themselvs.
In these very things (if well observed) lying, and within these indeed, beeing spent and consumed, the whole Indeavors, Practice, Studie and Wisdom (if not Religion) of All States, Ages, Nations and Men, viz.
Either in devising shifts (by acquisition of such and such Power, Habits, or otherwise) not to appear uncomly to themselvs, but to beget rather a greatness, esteem and satisfaction in themselvs, and others of themselvs;
Or, in Concealing and keeping close (by several Glosses) the Imperfectio[…] an apprehensions of others;
The more exquisite […]ferior onely in this, that the former is able to finde out more tight and exquisite textures for it self then the other; And the most fortunate man in this world, having the Advantag[e of the more unfortunate, but after the same manner also, viz. that the] covering of this fall's off a little sooner, and the other wrap's himself up i[n it a while longer.]
For thus all Lusts whatsoever (whether those of Acquisition) in the minde [(or others to exterior things)] […]eir utmost Accomplishments, are still but either to give countenance unto, o[r further to heap up a kinde of greatness,] […] for us.
The Assumption (on the other side) of a Repute and Esteemence to our se[lvs, is, that which after] […] a kinde of […]
currancey (during it) to all Lusts. (These two, observing tune, and holding time each to the measures and motio[ns of the other) within this Ci]rcle therefore is concluded whole Humane Nature, with all the Parts, and the most perfect Actions of it.
Upon these thoughts (Right Honorable!) I finding none not very ready to have asscribed to himself the name of beeing (in some measure) wise, becaus it is estimable, finde none that can endure the Manifestation (in any kinde) of their Follie; or that can bee willing to have this their Wisdom and Glorie to bee flatly and plainly called a Pageantrie, Mumming, or playing at bo-peep with themselvs and others; and yet (it really beeing no honester) if wee cannot bear the Judgments thus one of another; how shall wee abide the Trial, or bee able to stand before the Judgment of God, which seeth us, (and will shortly search us) with more pure and uncorrupt Eies? And how miserably then are the Imaginations of almost all men muffled up here in their own dressings? when the Truth most certainly is in our Beeings, at their best (natural, or acquir'd) Appearances, are yet such feign'd Things, and Propp'd up with such Lyes and fundamental mistakes of themselvs, that they would (Belshazzer like,) moulder and shatter themselvs to pieces, should they bee taken quite asunder from the disguises and patchings they have put upon themselvs, bare-facedly and undeceivedly to see themselvs but one small part of an hour.
Beeing convinced, by these, therefore (Right Honourable!) of the low Condition wee are herein, I not knowing what the Councils of God intend to bring forth for the settlement of this Nation; Nor how hee hath resolved in his Wisdom to dispose of it, (as to its outward Condition,) whether Hee

KV: 100% (100%|100%)
[illegible] 100%|80%)
PM: 100% (100%|100%)
ED: 97.3% (100%|94.74%)
Frequency is 237

>[…]g/Eca1652.txt (charset: UTF-8)

Display

○ Variants (357 tokens)
○ Normalised (0 tokens)
○ Not variants (4448 tokens)

Types List

themselvs (21)
som (19)
beeing (16)
cours (12)
neighbors (11)
Countrie (10)
hee (8)
onely (8)
anie (7)
Incouragement (7)

Step Complete

25/357

themselves (94.15%)

Auto Normalise

Threshold (%): 50

- Freely available for academic use: http://ucrel.lancs.ac.uk/vard/
- Designed to assist researchers in standardising spelling variation in historical corpora both manually and automatically.
- Uses methods from modern spellchecking to find spelling variants and offer/select appropriate modern equivalents.
- The original spelling is always retained in the text with an xml tag surrounding the replacement.
  - <normalised orig=”charitie">charity</normalised>
- Allows for the use of standard corpus linguistics tools without any modification.
- Used to normalise released historical (and other) corpora, e.g. EMEMT (Lehto *et al.*, 2010) and CEEC (Palander-Collin & Hakala, 2011).

VARDing guidelines

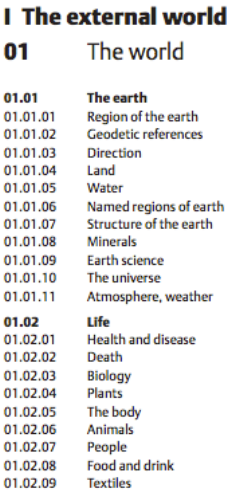
- Dawn Archer, Merja Kytö, Alistair Baron, Paul Rayson (2014) Normalising the Corpus of English Dialogues (1560-1760) using VARD2: Decisions and Justifications. Presented at the ICAME 2014 conference, University of Nottingham, UK, 30 April – 4 May 2014.
- Dawn Archer, Merja Kytö, Alistair Baron, Paul Rayson (2015). Guidelines for normalising Early Modern English corpora: decisions and justifications. ICAME Journal, Volume 39, May 2015. DOI: 10.2478/icame-2015-0001

VARDing EEBO

- *£7k funding from JISC, September 2014*
- uVARD crowdsourcing server prototype created by Charlie Revett (July-August 2014)
- VARDsourcing data preparation by Mahmoud El-Haj (Feb-Mar 2015)
- VARDsourcing server development by Andrew Moore (2016-17)
- EEBO corpus (Phase 1 texts) split into 10 x 25 year periods x 8 blocks (2,000 words); estimating 2 hours per 1,000 words; total ~160K words
- Training of participants via gold standard
- Evaluation of inter-rater reliability via VARD API
- Timescale: call for participants and training of VARD subsequently

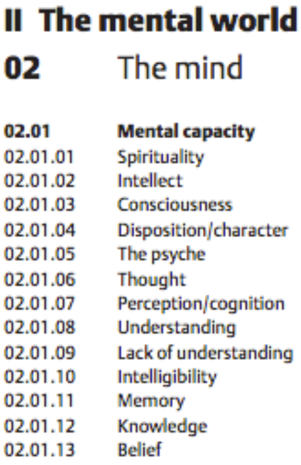
Though I **speake** with the tongues of men & of Angels, and **haue** not charity, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **remooue mountaines**,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

(Authorised Version of the Bible, 1611)



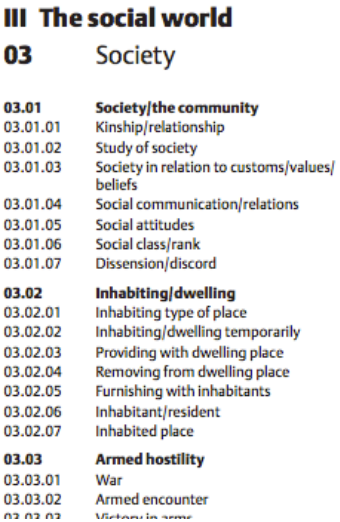
I The external world

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



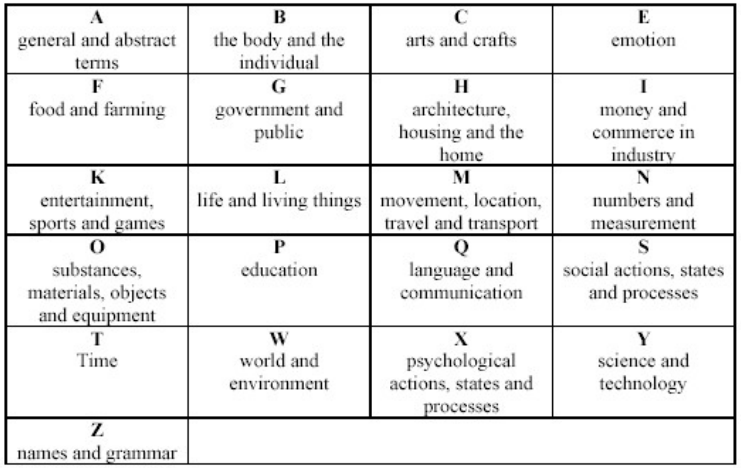
II The mental world

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			

USAS (Modern English) semantic tagger

- Full text tagging, not just selected words (c.f. Diction, LIWC, RID)
- Tagging the coarse-grained sense in context, not just the word
- Not task specific categories
- Flexible category set with hierarchical structure
- Words and multi-word expressions (MWE) e.g. phrasal verbs (stubbed out), noun phrases (riding boots), proper names (United States of America), true idioms (living the life of Riley)

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			

Lexical resources

- Lexicon of 56,316 items
  - presentation NN1 Q2.2 A8 S1.1.1 K4
- MWE list of 18,971 items
  - travel_NN1 card*_NN* M3/Q1.2
- A small wildcard lexicon
  - *kg NNU N3.5
- Unknown words using WordNet synonym lookup

Disambiguation methods (1)

- 1. POS tag
  - *spring* noun [season sense] [coil sense]
  - *spring* verb [jump sense]
- 2. General likelihood ranking for single-word and MWE tags
  - *green* referring to [colour] is generally more frequent than *green* meaning [inexperienced]
- 3. Overlapping MWE resolution
  - Heuristics applied: semantic MWEs override single word tagging, length and span of MWE also significant

Disambiguation methods (2)

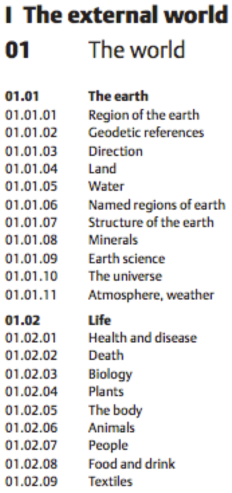
- 4. Domain of discourse
  - adjective *battered*
    - [Violence] (e.g. battered person)
    - [Judgement of Appearance] (e.g. battered car)
    - [Food] (e.g. battered cod)
- 5. Text-based disambiguation
  - one sense per text
- 6. Template rules
  - *Auxiliary verbs (be/do/have)*
  - *account* of NP [narrative]
  - balance of xxx *account* [financial]

Disambiguation methods (3)

- 7. Local probabilistic
  - *account* occurring in the company of *financial, bank, overdrawn, money*
  - surrounding words, POS tags or semantic fields
  - span of words
  - co-occurrence measures rather than HMM

Evaluation (modern data)

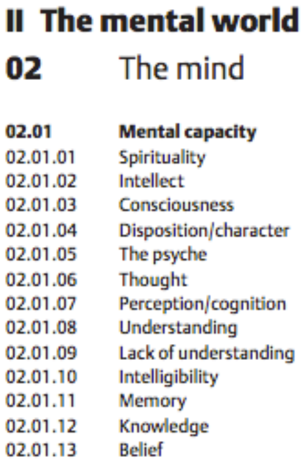
- Hand tagged test corpus of 124,839 words
- Error rate of 8.95%
- Ambiguity ratio 47.73%
- Reduced to 17.06% by disambiguation
- Not all ambiguity is resolved, but 1st choice tag selection gives 91% accuracy.



I The external world

01 The world

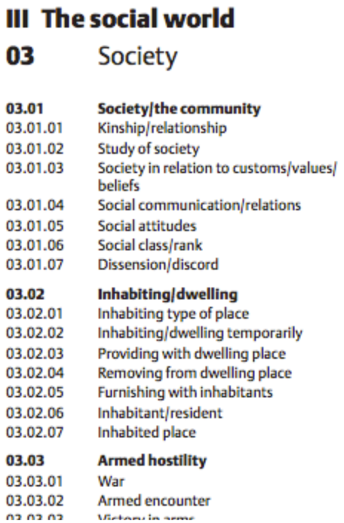
01.01	**The earth**
01.01.01	Region of the earth
01.01.02	Geodetic references
01.01.03	Direction
01.01.04	Land
01.01.05	Water
01.01.06	Named regions of earth
01.01.07	Structure of the earth
01.01.08	Minerals
01.01.09	Earth science
01.01.10	The universe
01.01.11	Atmosphere, weather
01.02	**Life**
01.02.01	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

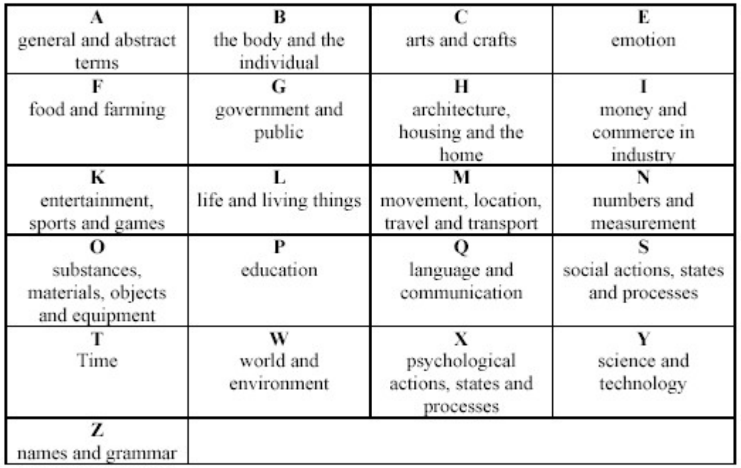
02.01	**Mental capacity**
02.01.01	Spirituality
02.01.02	Intellect
02.01.03	Consciousness
02.01.04	Disposition/character
02.01.05	The psyche
02.01.06	Thought
02.01.07	Perception/cognition
02.01.08	Understanding
02.01.09	Lack of understanding
02.01.10	Intelligibility
02.01.11	Memory
02.01.12	Knowledge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



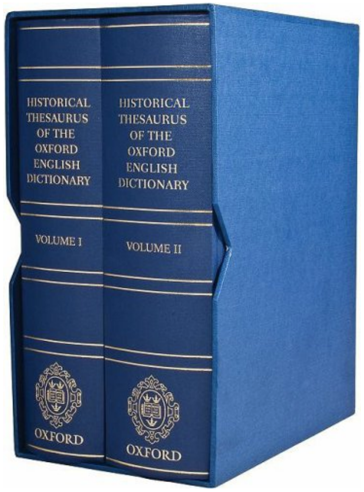
III The social world

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
03.03.03	Victory in arms



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			

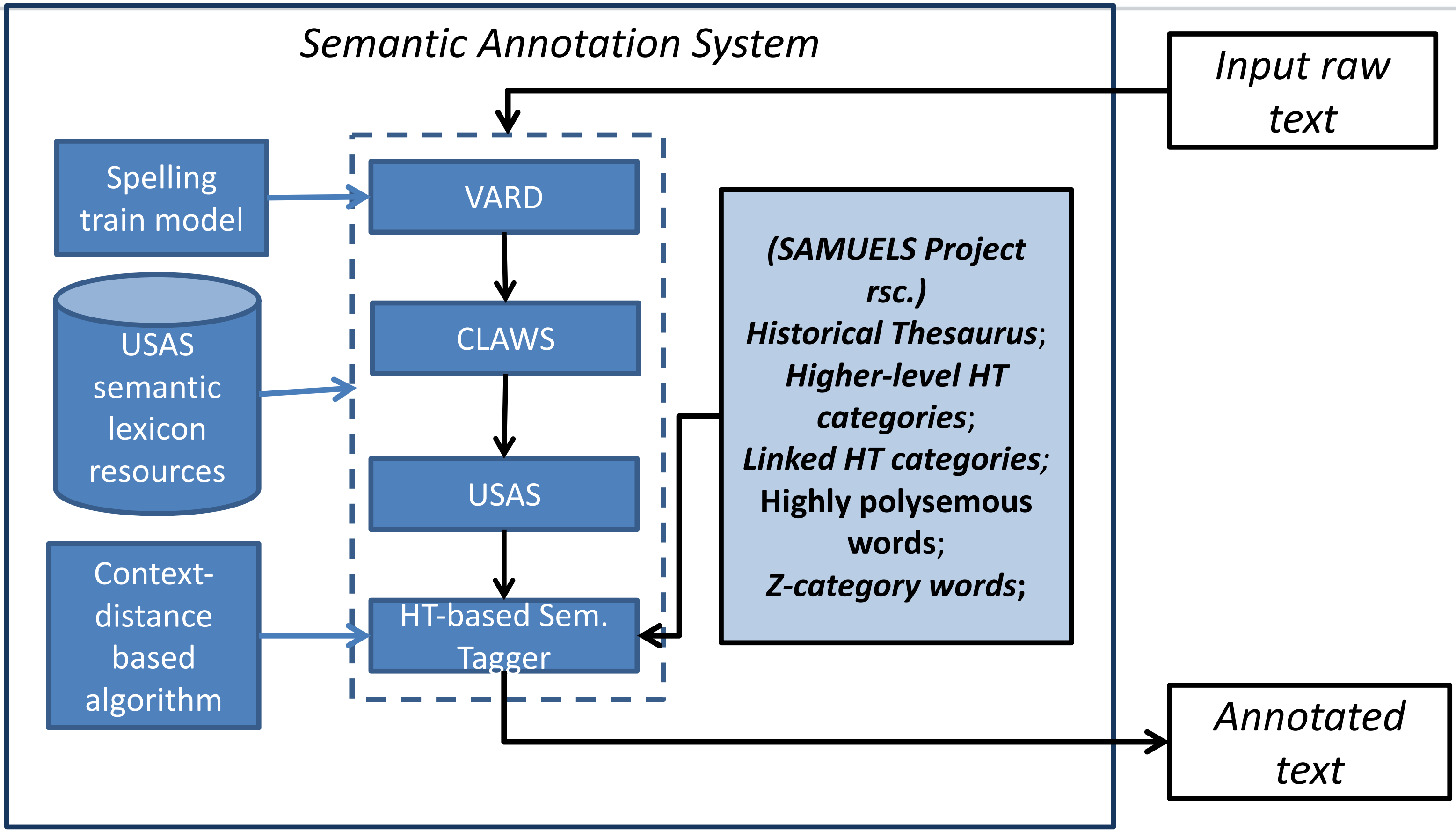


HISTORICAL THESAURUS OF THE OXFORD ENGLISH DICTIONARY
VOLUME I
OXFORD
HISTORICAL THESAURUS OF THE OXFORD ENGLISH DICTIONARY
VOLUME II
OXFORD

Historical Thesaurus of English (Samuels, Kay, Alexander et al)

- Comprehensive analysis of English as found in the 2nd edition of the OED
- 793,742 word forms arranged into 225,131 semantic categories
- The HT semantic categories are mapped to 4,028 thematic-level categories.
- three primary divisions are
  - I The External World
  - II The Mental World
  - III The Social World
- each category is given a nested reference code such as "01.02.08.02.02.06.01 n" for the category *Whisky*

Architecture of Annotation system



TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	04.10 [Unrecognised];	04.10 [Unrecognised];
The	the	AT	Z5	0	04.03 [null];	04.03 [Grammatical Word];
cat	cat	NN1	L2 M3	0	03.10.12.02.12.01-08 [0.94736842] [.of cat]; 01.02.04.13.09.02.12-01 [1.00000000] [.types of]; 01.02.06.16.07.04-09 [1.00000000] [.member of family Pimelodidae/common cat-fish];	Y12a07a [Skin with hair attached/fur]; B20t [Particular food plant/product]; B22j [Fish];
sat	sit	VVD	M8 C1 P1 G1.1 G2.1 M6 A9+	0	[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];	B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];
on	on	II	Z5	0	[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];	B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];
the	the	AT	Z5	0	04.03 [null];	04.03 [Grammatical Word];
mat	mat	NN1	H5 O2	0	03.02.07.03.09.14-03 [0.93750000] [.mat]; 03.11.04.13.16.15-14 [0.93750000] [.mat]; 03.02.07.03.09.10.01-02 [0.94444444] [.table mat];	Q06f05m [Floor-covering]; Z08v11 [Bowls/bowling]; Q06f05i [Household linen];
.	PUNC	YSTP	PUNC	0	NULL	NULL [];
S_END	NULL	NULL	Z99	0	04.10 [Unrecognised];	04.10 [Unrecognised];

HTST current disambiguation methods (1)

- Disambiguate words and MWEs that have multiple HT categories
  - Filter by POS.
  - For each candidate category, extract all possible parent categories and collect headings (simple definition) of them, including current heading. Words in the headings form a feature set $\boldsymbol{HW_i} = \{h_1, h_2, \ldots, h_m\}$.
  - Collect up to five content words from each side of the key word/MWE. Together with the target word/mwe w_t, they form a context feature set $\boldsymbol{CW}=\{w_t, w_1, w_2, \ldots, w_n\}$.
  - Measure Jaccard Distance between ***CW*** and each $\boldsymbol{HW_i}$, and select the candidate categories (up to three) that have close distances to the context.

HTST current disambiguation methods (2)

- Time filtering
  - Filter word senses whose usage appear outside a given time window in the HT thesaurus.
  - Users can set upper and lower time boundaries (in years) to increase the relevance of the HT categories to the given time.
    - E.g. if a text was published in 1800, using the time filter, ignore the word senses which appear after that era.
  - Particularly useful for tagging historical data.

Further disambiguation methods

- Detecting linked HT categories in context to determine the core senses;
- Apply co-occurrence based statistical training model based on HT-OED sense mapping, OED example sentences (50.2M tokens) and sense definitions (14.5M tokens).
  - At word level: based on co-occurrence between HT category and context words
  - At semantic level: Based on co-occurrence between HT category and USAS tags.
- Core HT category detection based on density of polysemy;
- Core HT category detection based on OED sense ordering;
- Improve VARD with OED spelling variants data linked to headwords & dates.

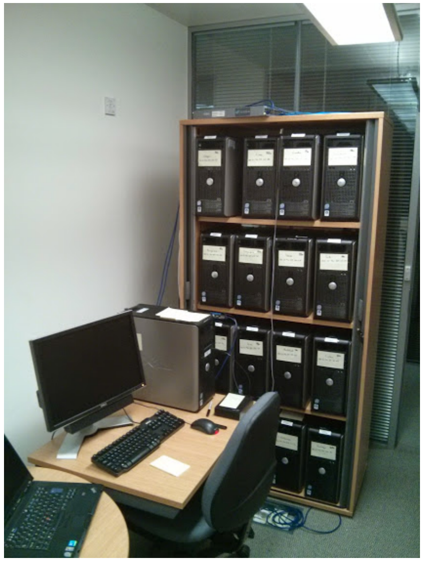
Evaluation

- Ten texts were selected from different genres (e.g. spoken and written).
- Publication time spans from 1820 to 2014.
- Each text contains about 1,000 words.
- Evaluated for both HT sense codes and thematic sense codes.
- Examined the impact of the time filter.
- Evaluation criterion: If top three of the candidate tags suggested by the system contain the correct tag(s), it is considered to be correct annotation.
  - *In our evaluation, we see maximum 84.4% for the HT codes and 86.2% for the thematic codes.*

Further reading ...

- Piao, SS, Dallachy, F, Baron, A, Demmen, JE, Wattam, S, Durkin, P, McCracken, J, Rayson, PE & Alexander, M 2017, 'A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation' Computer Speech and Language, vol 46, pp. 113-135. DOI: 10.1016/j.csl.2017.04.010

Cluster & cloud computing



- MapReduce (Hadoop) framework
- Hansard corpus processing
  - 2.2 billion words
  - 32.7GB of data including mark-up
  - 7.5 million files
  - 3 days to complete versus 98 days on one PC (HPC-USAS)
  - 6 days to complete on our hand-made cluster (HTST)

In summary …

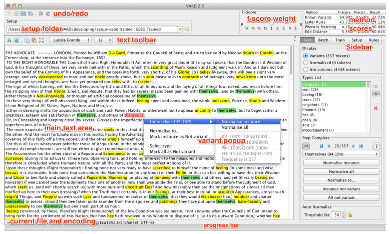
- In order to adapt our modern semantic tagger you need:
  - Variant Spelling methods
  - Historically sensitive semantic taxonomy
  - Improved semantic disambiguation techniques (Historical Thesaurus Semantic Tagger – HTST)
  - Use of big data methods e.g. cluster and cloud computing
- Ongoing and future work
  - Visualisations / GIS
  - Multilingual semantic tagger for 12+ languages

Thanks for listening!

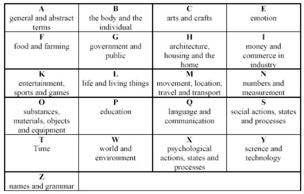
- p.rayson@lancaster.ac.uk
- @perayson

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

- http://ucrel.lancs.ac.uk/vard/



- http://ucrel.lancs.ac.uk/usas/
- http://phlox.lancs.ac.uk/ucrel/semtagger/english



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			