Modern and Historical Aspects of the UCREL Semantic Analysis System

Paul Rayson, Scott Piao, Dawn Archer

UCREL

Department of Computing,

Department of Linguistics and English Language

Lancaster University, UK.

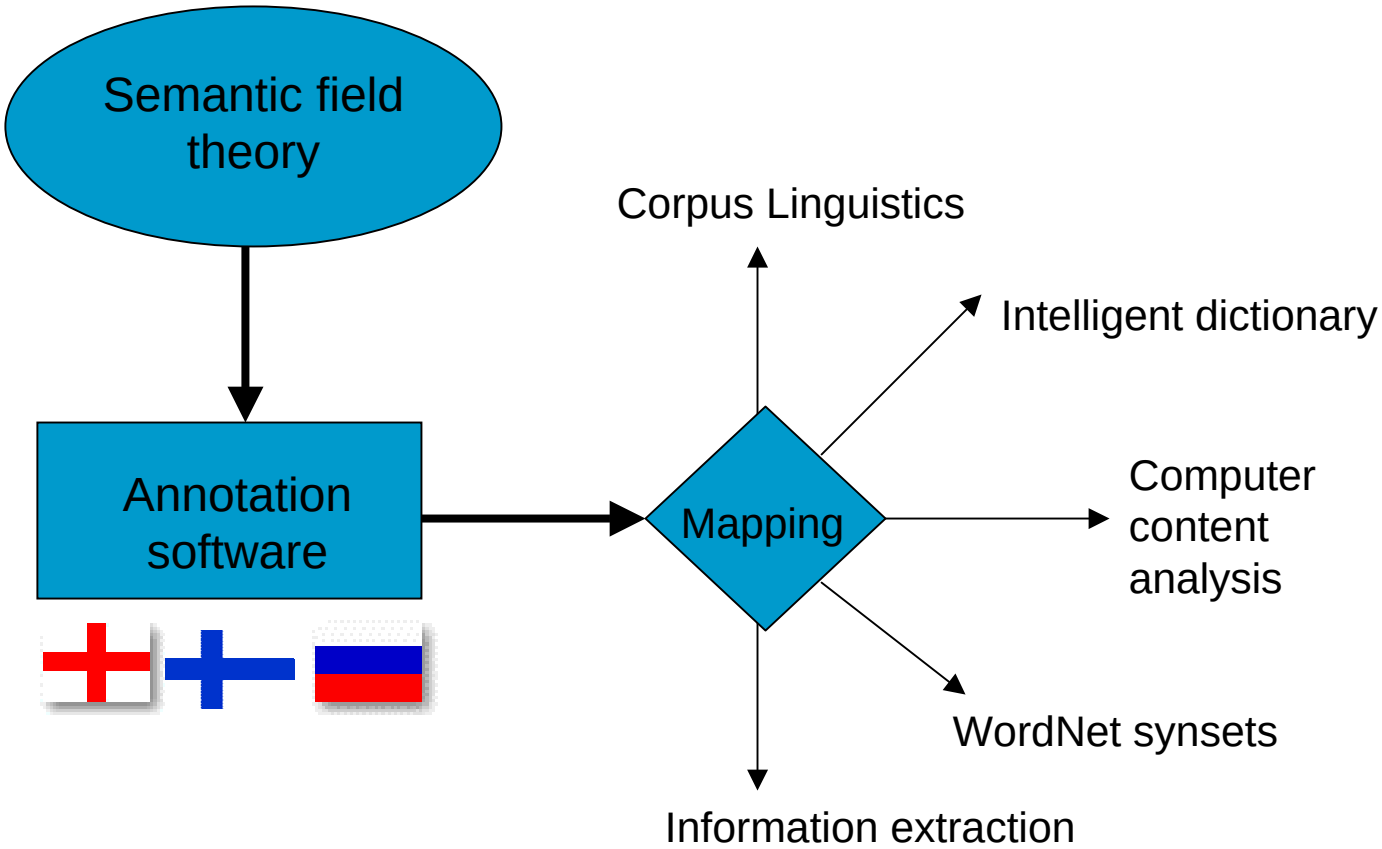
Outline

- The purpose of this talk is to introduce the UCREL Semantic Analysis System (henceforth USAS), an automated software tool for the semantic field annotation of running texts. The talk will consist of three parts:
  - How and why USAS was developed.
  - Its recent application as part of the Benedict project, including its porting to the Finnish language.
  - The modifications we have made to the system so that it can cope with historical texts ranging from the 1600s onwards.
- For more information, see: www.comp.lancs.ac.uk/ucrel/usas/
- Acknowledgements:
  - Original PIs were Roger Garside, Geoffrey Leech, Jenny Thomas
  - Currently funded EU IST Benedict project IST-2001-34237

How and why USAS was developed

Paul’s bit

Overview


Semantic field theory
Annotation software
Mapping
Corpus Linguistics
Intelligent dictionary
Computer content analysis
WordNet synsets
Information extraction

Application contexts

- Semantic field analysis
- Content analysis
  - Conceptual analysis: USAS, Louw/Nida categories in OpenText.org
  - General category: General Inquirer, Minnesota Contextual Content Analysis
  - Specialised content analysis: RID, Diction
- Market research interview transcript analysis
- Word sense disambiguation: Senseval
- Information extraction / text mining
- Electronic dictionaries

Information extraction

- Requirements reverse engineering to support business process change (Revere)
- Reducing rework through decision management (Tracker)

Links to Lexicography

- The New Intelligent Dictionary (Benedict)
- Providing an interactive user-specified access interface, tailoring the dictionary information supply according to user specifications, incorporating multi-layered entry structure with new information categories and links to corpus data and syntactically- and semantically-based corpus search tools in the dictionary data base.

The task we set ourselves

- Full text tagging, not just selected words
- Tagging the sense in context, not just the word
- Not task specific categories
- Tag set should make sense (psycho)- linguistically
- Flexible category set with hierarchical structure
- Words and multi-word expressions e.g. phrasal verbs (*stubbed out*), noun phrases (*riding boots*), proper names (*United States of America*), true idioms (*living the life of Riley*)

Semantic fields

- AKA conceptual field, a semantic domain, a lexical field, or a lexical domain
- ‘groups together word senses that are related by virtue of their being connected at some level of generality with the same mental concept’
- Not only *synonymy and antonymy* but also *hypernymy and hyponymy*
- E.g. EDUCATION: academic, coaching, coursework, deputy head, exams, PhD, playschool, revision notes, studious, swot, viva

The UCREL Semantic Analysis System

- Hierarchy of 21 major discourse fields expanding into 232 category labels:

Table 1 : The top level of the USAS system

A: General & Abstract Terms	**B:** The Body & the Individual	**C:** Arts & Crafts	**E:** Emotional Actions, States & Processes
F: Food & Farming	**G:** Government & the Public Domain	**H:** Architecture, Building Houses & the Home	**I:** Money & Commerce in Industry
K: Entertainment, Sports & Games	**L:** Life & Living Things	**M:** Movement, Location, Travel & Transport	**N:** Numbers & Measurement
O: Substances, Materials, Objects & Equipment	**P:** Education	**Q:** Linguistic Actions, States & Processes	**S:** Social Actions, States & Processes
T: Time	**W:** The World & Our Environment	**X:** Psychological Actions, States & Processes	**Y:** Science & Technology
Z: Names & Grammatical Words			

Lexical resources

- Lexicon of 44,668 items
  - workshop NN1 I4/H1 P1
- MWE list of 18,553 items
  - travel_NN1 card*_NN* M3/Q1.2
- A small wildcard lexicon
  - *kg NNU N3.5
- A small context rule set of 350 items
  - VB*[Z5] (R*n) (XX) (R*n) V*G*
- Unknown words using WordNet synonym lookup

Disambiguation methods (1)

- 1. POS tag
  - *spring* temporal noun [season sense]
  - *spring* common noun [coil sense] [water source 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 wife)
    - [Judgement of Appearance] (e.g. battered car)
    - [Food] (e.g. battered cod)
- 5. Text-based disambiguation
  - one sense per text
- 6. Context rules
  - *Auxiliary verbs (be/do/have)*
  - *account* of NP [narrative]
  - balance of xxx *account* [financial]

Disambiguation methods (3)

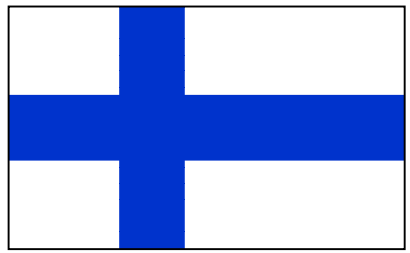
- 7. Local probabilistic
  - *account* occurring in the company of *financial, bank, overdrawn, money*
  - surrounding words, POS tags or semantic fields
  - span of words
  - co-occurrence measures rather than HMM

Its recent application as part of the Benedict project, including its porting to the Finnish language

Scott’s bit

USAS in the Benedict Project

- Restructuring and updating of USAS
- Integrating USAS into MT System
- Porting USAS to Finnish language
- Future possibilities



The Benedict Project

- EC Project (ref. IST-2001-34237)
- **Aim**: develop a new methodology and a prototype of intelligent machine dictionary.
- **Project partners**: Lancaster University, HarperCollins Publisher, Kielikone Ltd., University of Tampere, Gummerus Kustannus Ltd and Nokia
- **Websites**:
  http://www.comp.lancs.ac.uk/ucrel/projects.html#benedict
  http://mot.kielikone.fi/benedict/

Restructuring and Updating USAS (1)

- Why the trouble?
  - The original tagger is designed in procedural approach, difficult to interact with/integrate into other program/systems.
  - The original tagger is C-coded and difficult to be ported across different platforms (from UNIX to Windows in our case).
  - Difficult to port it to other languages.

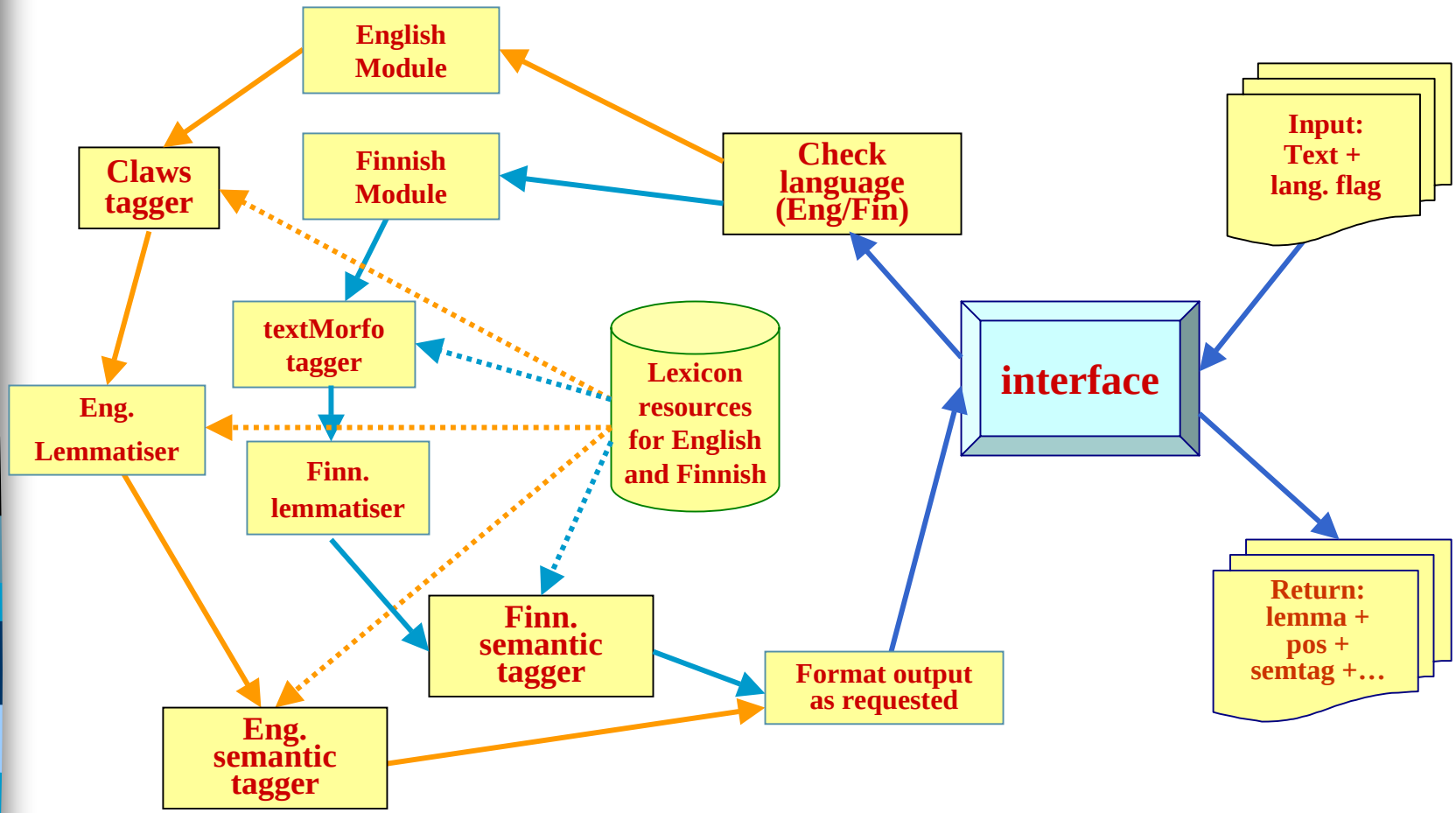
Restructuring and Updating USAS (2)

- What has been done?
  - The original package has been re-structured into an Object-oriented (OO) architecture.
  - Semantic tagger has been re-coded in Java; a Windows version of the CLAWS tagger (C-code) is integrated using JNI.
  - It Interacts with other Benedict programs through an interface.
  - Improved maintainability – each component can be updated/modified without affecting other components.
  - The Unicode-based text processing – may be not so useful for English, but can be critical for some other languages.

Porting USAS to Finnish Language

- Building Finnish semantic tagger (FST) is one of the aims of the Benedict Project.
- FST is based on the architecture of the English tagger with only minor modification (mainly for different POS tagger).
- A Finnish morphological analyser *textMorfo*, provided by Kielikone Oy, is used in the place of CLAWS tagger.
- A beta version of Finnish semantic lexical resources are built and provided by Finnish project partners.
- Finnish semantic tagger has become a module of the USAS package.

Structure of USAS (v2)



Evaluation of English Semantic Tagger

- The English semantic tagger has been evaluated on various test data.
- Difficulty experienced: Very few manually semantically tagged data are available.
- Our evaluation:
  - Hand tagged test corpus
  - MWE (Multiword Expression) recognition
  - Lexical coverage

Evaluation on Hand-tagged Corpus (1)

- Test corpus size = 124,839 words
- Error rate = 8.95%
- Ambiguity ratio
  - *Before disambiguation:*

    Ambiguity ratio = 47.73%
  - *After disambiguation*:

    Ambiguity ratio = 17.06%

Evaluation on Hand-tagged Corpus (2)

Breakdown of tagging methods

Tagging method	Freq. %
Lexicon	63.68
Lexicon after suffix strip	3.41
Lexicon on lemma	0.03
Auto-tag rule	0.39
Domain of discourse	7.67
Auxiliary verb	6.76
Context rules	0.83
Lexicon ignoring POS	0.92
Lexicon stem ignoring POS	0.07
WordNet	0.05
Wildcard MWE	0.54
Normal MWE	11.60
MWE and domain of discourse	4.06

Evaluation on Hand-tagged Corpus (3)

Breakdown of Errors by POS

POS tag	% Error relative to testbed	% Error relative to tag
A:Article	0.21	2.47
C:Conjunction	0.05	0.60
D:Determiner	0.21	4.69
E:Existential there	0.01	1.22
F:Formula Foreign word	0.00	0.31
G:Genitive	0.01	6.62
I:Preposition	0.36	4.16
J:Adjective	0.87	17.65
M:number	0.29	23.93
N:noun	2.62	16.29
P:pronoun	0.06	0.51
R:adverb	1.08	13.47
T:to	0.11	7.52
U:interjection	0.02	0.94
V:verb	3.03	13.21
X:not	0.01	1.25
Z:letter	0.00	2.67
TOTAL	8.95	

Evaluation on MWE Recognition

MWE length	Total MWEs	Accepted MWEs	Precision
2	3,378	3,105	91.92%
3	700	575	82.14%
4	95	91	95.44%
5	18	17	94.44%
6	4	4	100.00%
Total	4,195	3,792	90.39%

For further details, see: **Scott Songlin Piao, Paul Rayson, Dawn Archer and Tony McEnery (2005). Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE Extraction. Computer Speech and Language. (to appear).**

Evaluation on Lexical Coverage (1)

Modern English Corpora

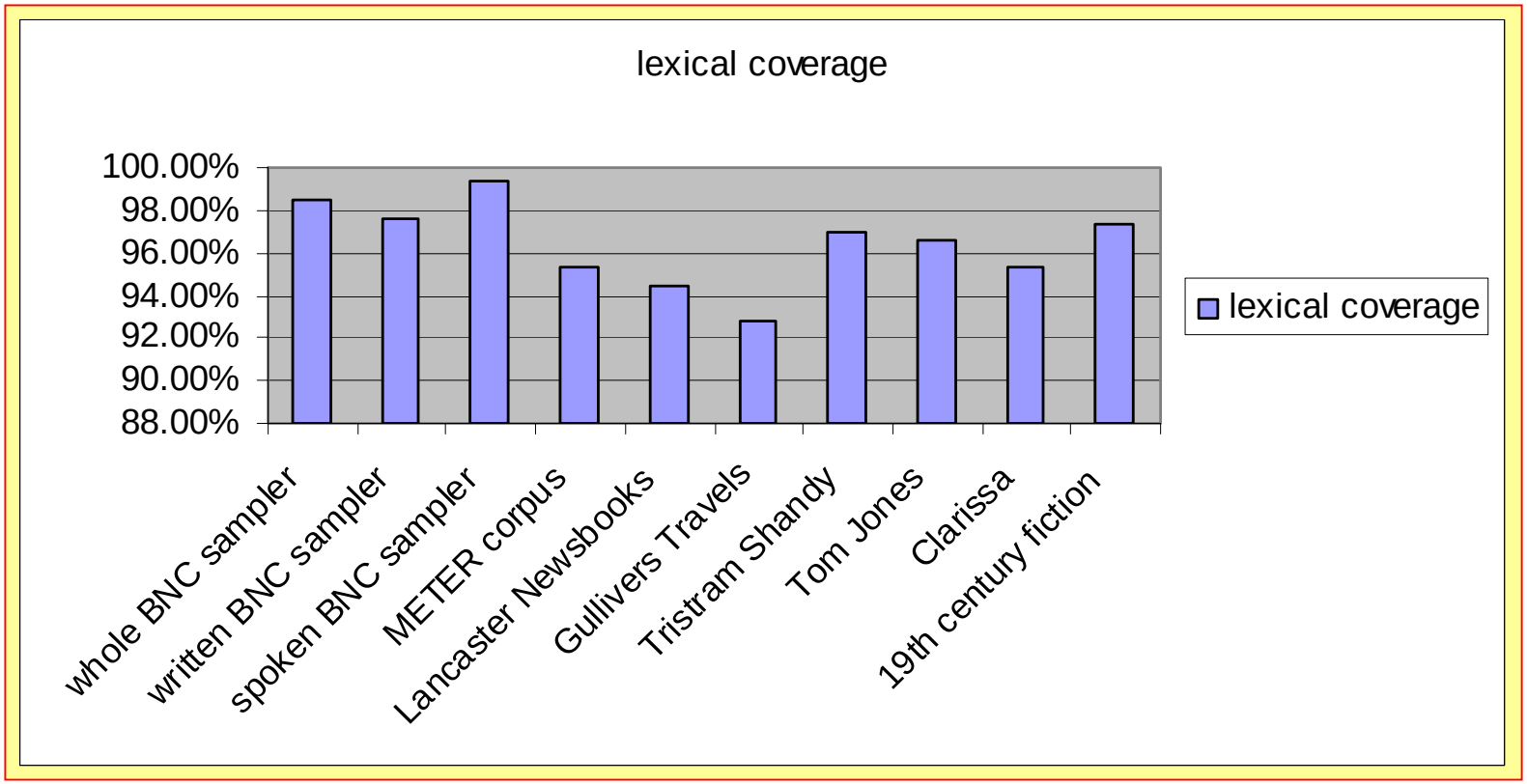
Test Corpus	Total Tokens	Unmatched Tokens	Lexical Coverage
Whole BNC Sampler	1,956,171	29,517	98.49%
BNC Samp. Written sect.	970,532	23,407	97.59%
BNC Samp. Spoken sect.	985,639	6,110	99.39%
METER Corpus	241,311	11,143	95.38%

Evaluation on Lexical Coverage (2)

Historical English Corpora

Test Corpus	Total Tokens	Unmatched Tokens	Lexical Coverage
Lancaster Newsbooks	61,065	3,418	94.40%
Gulliver's Travels	194,987	14,117	92.76%
Tristram Shandy	108,137	3,235	97.01%
Tom Jones	352,942	11,944	96.62%
Clarissa	887,276	40,988	95.38%
19th century fiction	5,000,000	135,661	97.29%

Evaluation on Lexical Coverage (3)



***For further details, see*: Scott S. L. Piao, Paul Rayson, Dawn Archer and Tony McEnery (2004) Evaluating lexical resources for a semantic tagger. In Proceedings of LREC-04. Lisbon, Portugal. pp. 499-502.**

Development of an historical tagger

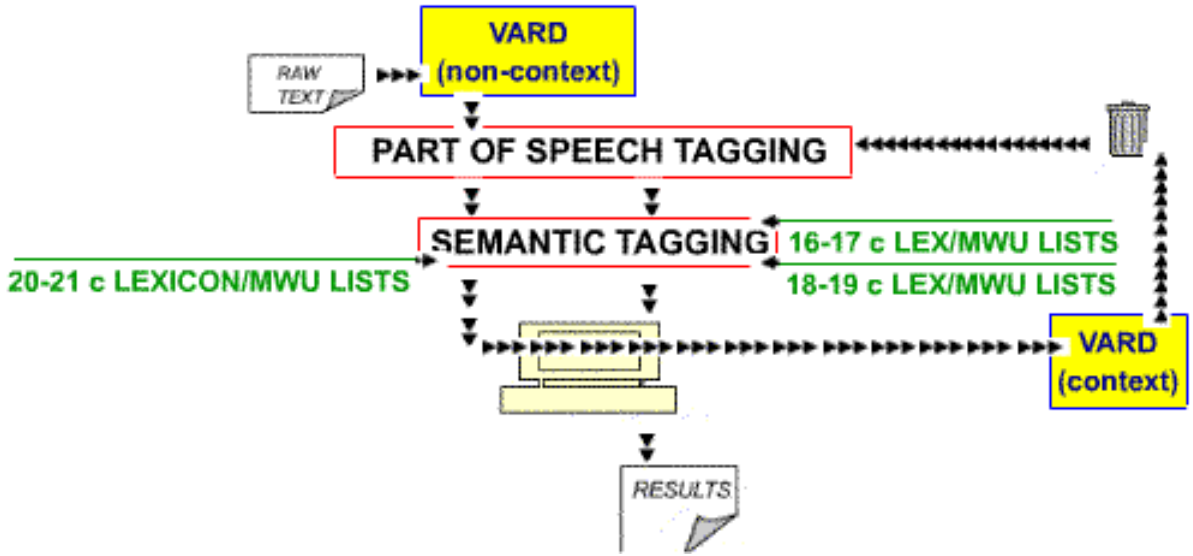
Dawn’s bit

Problems and solutions

- Variant spellings
  e.g. *bee, doe*
  -(e)th, -(e)st,
  (e)s genetive
- Change in meaning

- VARD
  - Non-contextualised (search and replace script)
  - Contextualised (context rules)
- Different lexicons for different periods
  - 16th and 17th C
  - 18th and 19th C
  - 20 and 21st C

Historical Tagger



Resources for 16-17th lexicon:

Nameless Shakespeare, Lampeter Corpus (1640-1740), Lancaster Newsbooks (1653/54), Corpus of English Dialogues (1560-1760)

Resources for 18-19th lexicon:

18th and 19th fiction from ProQuest (EEBO texts)

Example VARDED text: Midsummer Night's Dream

Theseus. Now <reg o="faire"> fair Hippolita, our <reg o="nuptiall“> nuptial <reg o="houre"> hour <reg o="Drawes“> Draws on apace: <reg o="foure"> four happy <reg o="daies"> days bring in Another Moon: but oh, me <reg o="thinkes"> thinks, how slow This old Moon wanes; She lingers my desires Like to a Step-dame, or a Dowager, Long withering out a <reg o="yong"> young <reg m="mans"> man's <reg o="reuennew"> revenue.

Hip. <reg o="Foure“> Four <reg o="daies"> days <reg o="wil"> will quickly steep <reg o=“thselues”> themselves in nights; <reg o="foure"> Four nights <reg o="wil"> will quickly <reg o="dreame"> dream away the time: And then the <reg o="Moone"> Moon, like to a <reg o="siluer“> silver bow, Now bent in <reg o= "heauen“> heaven, <reg o="shal"> shall behold the night Of our solemnities.

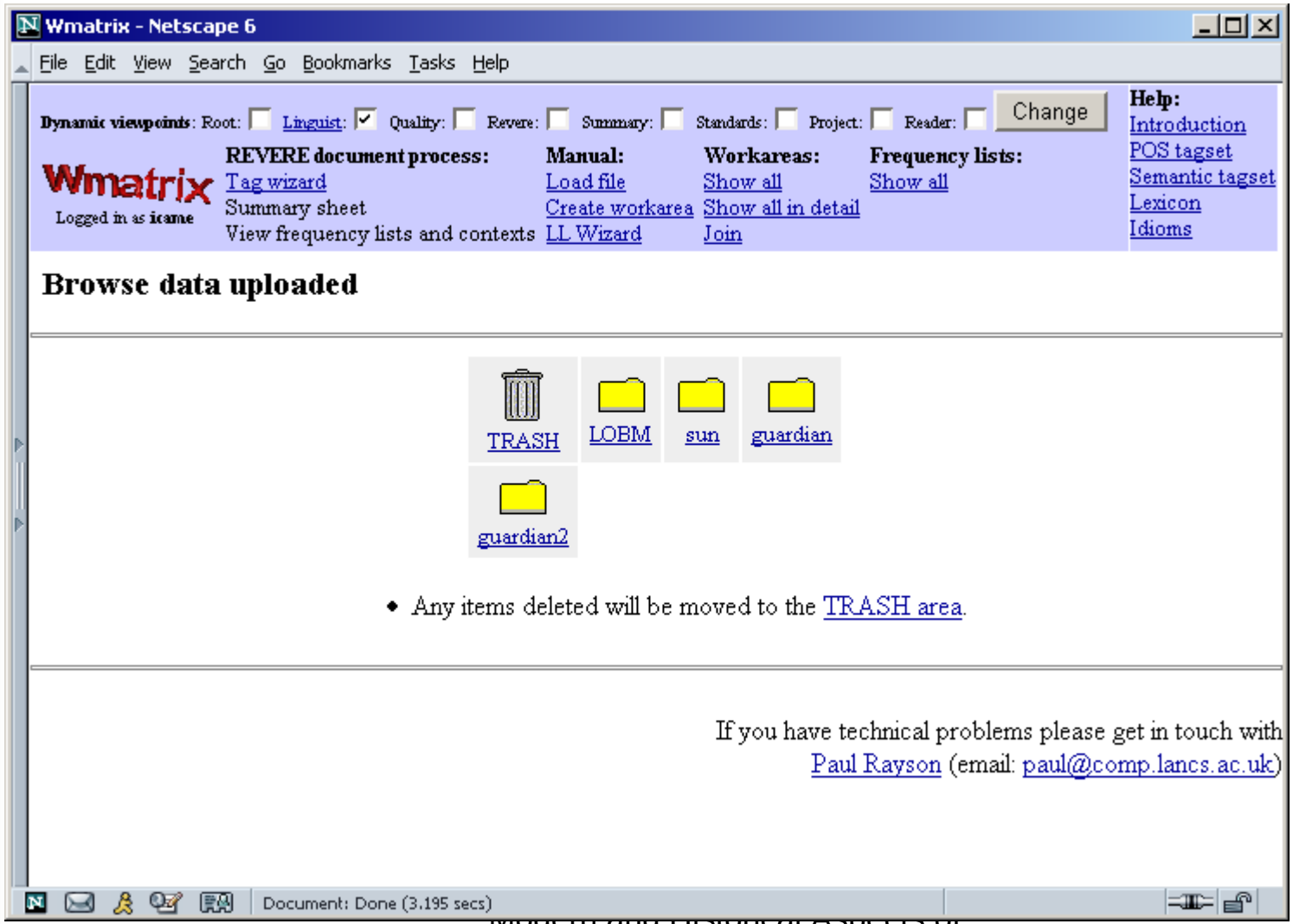
Questions?

Further information at
http://www.comp.lancs.ac.uk/ucrel/usas/

{p.rayson, s.piao. d.archer}@lancs.ac.uk

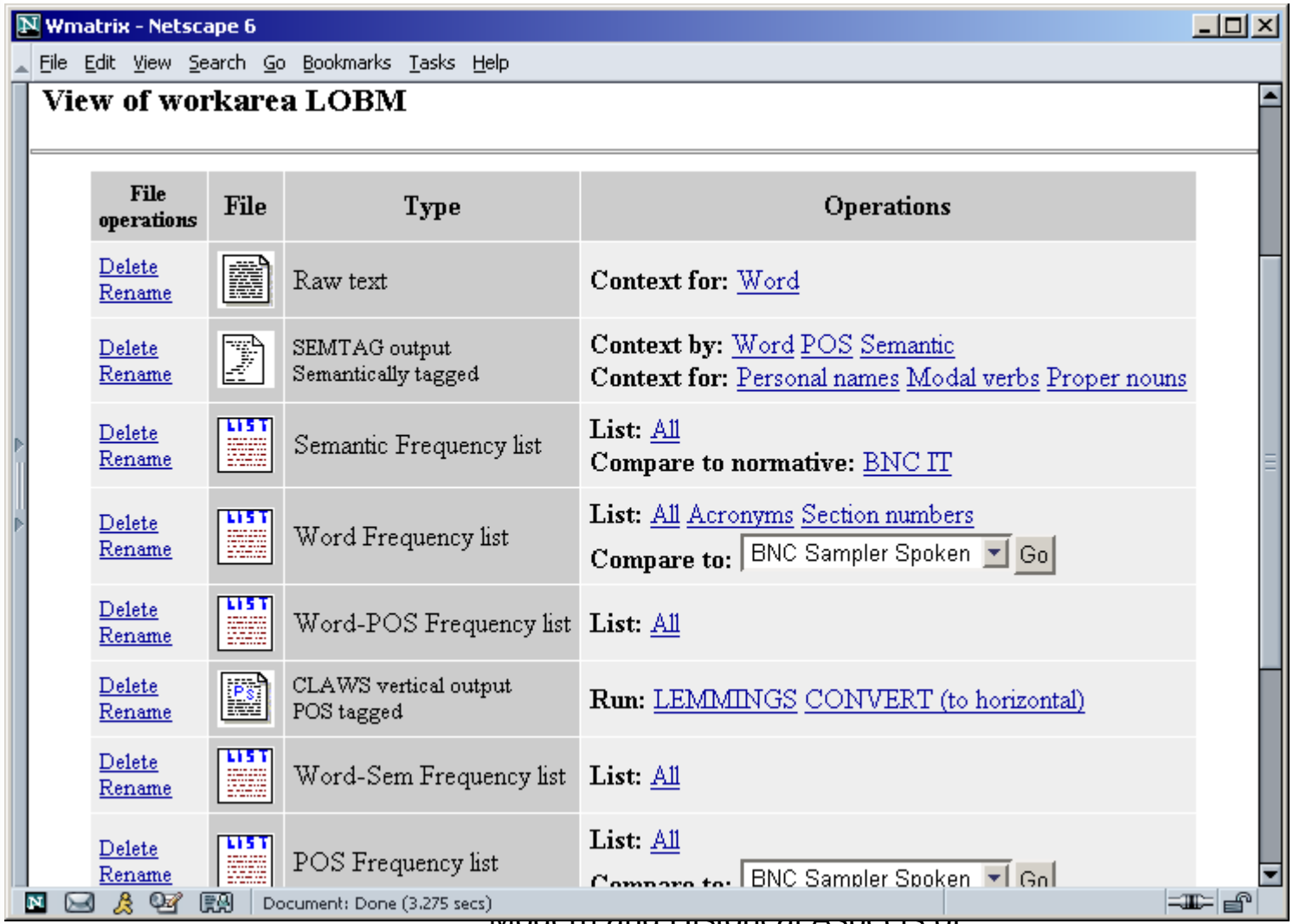
Appendix

Wmatrix work areas

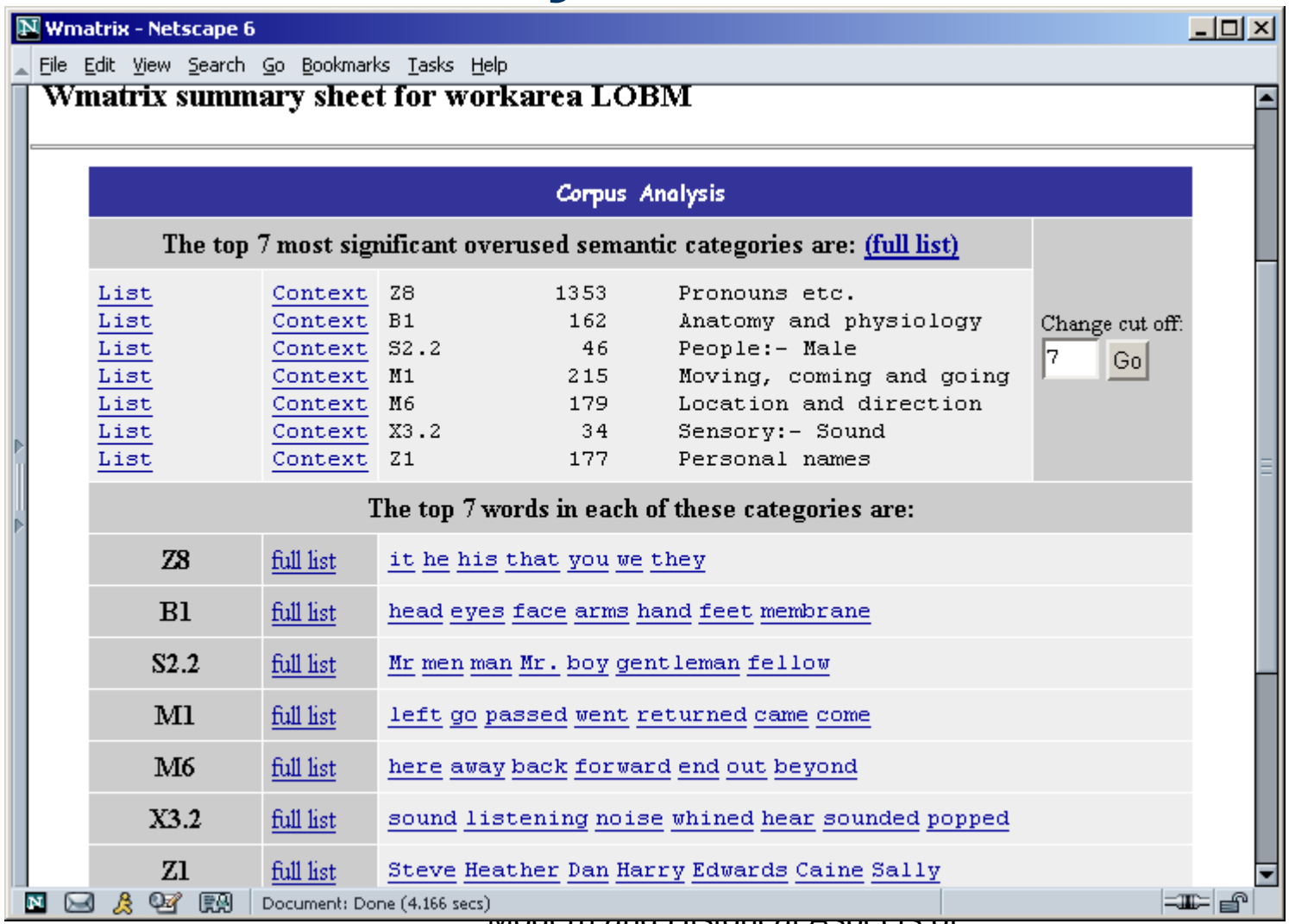


Wmatrix tag wizard

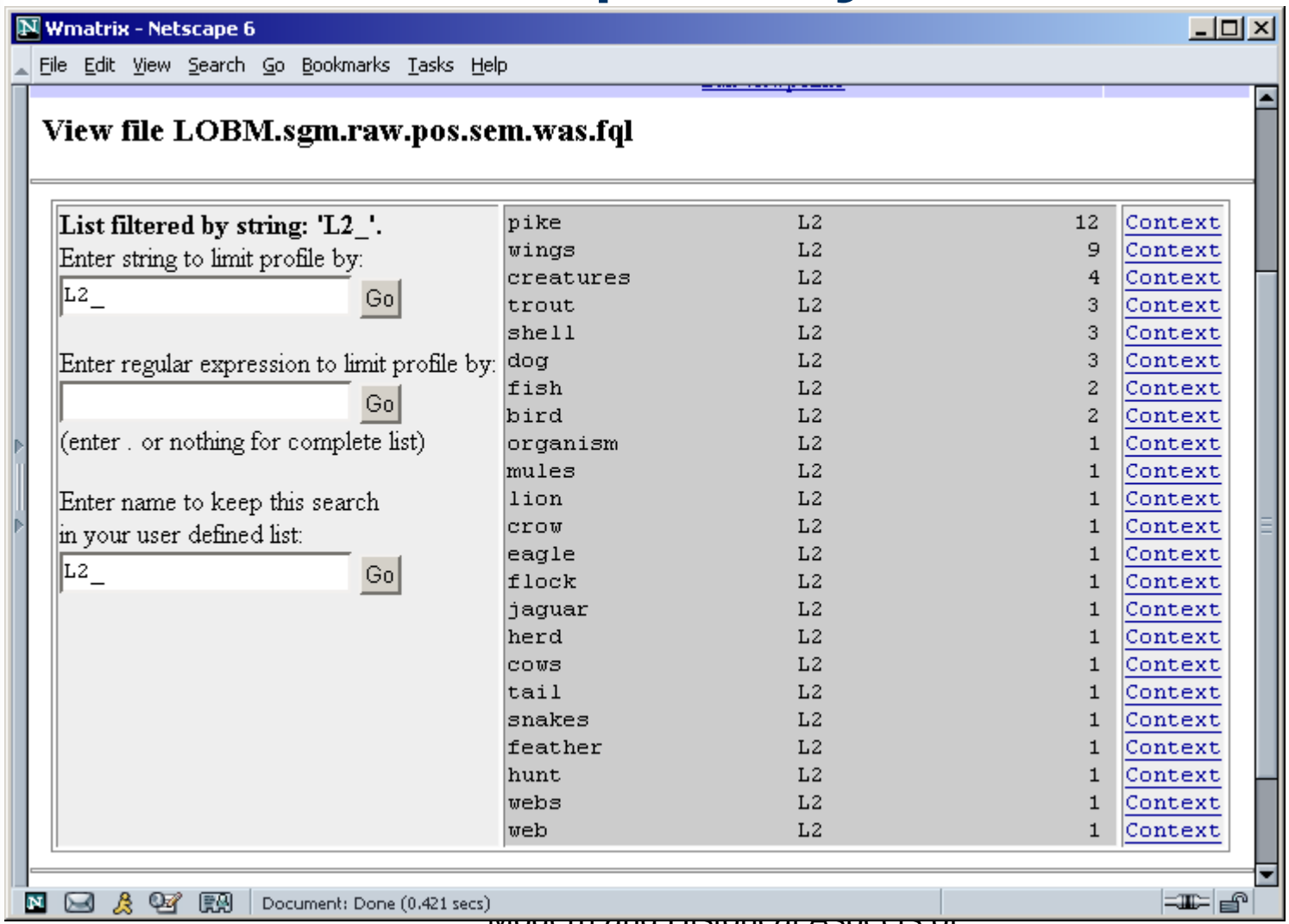
Wmatrix views



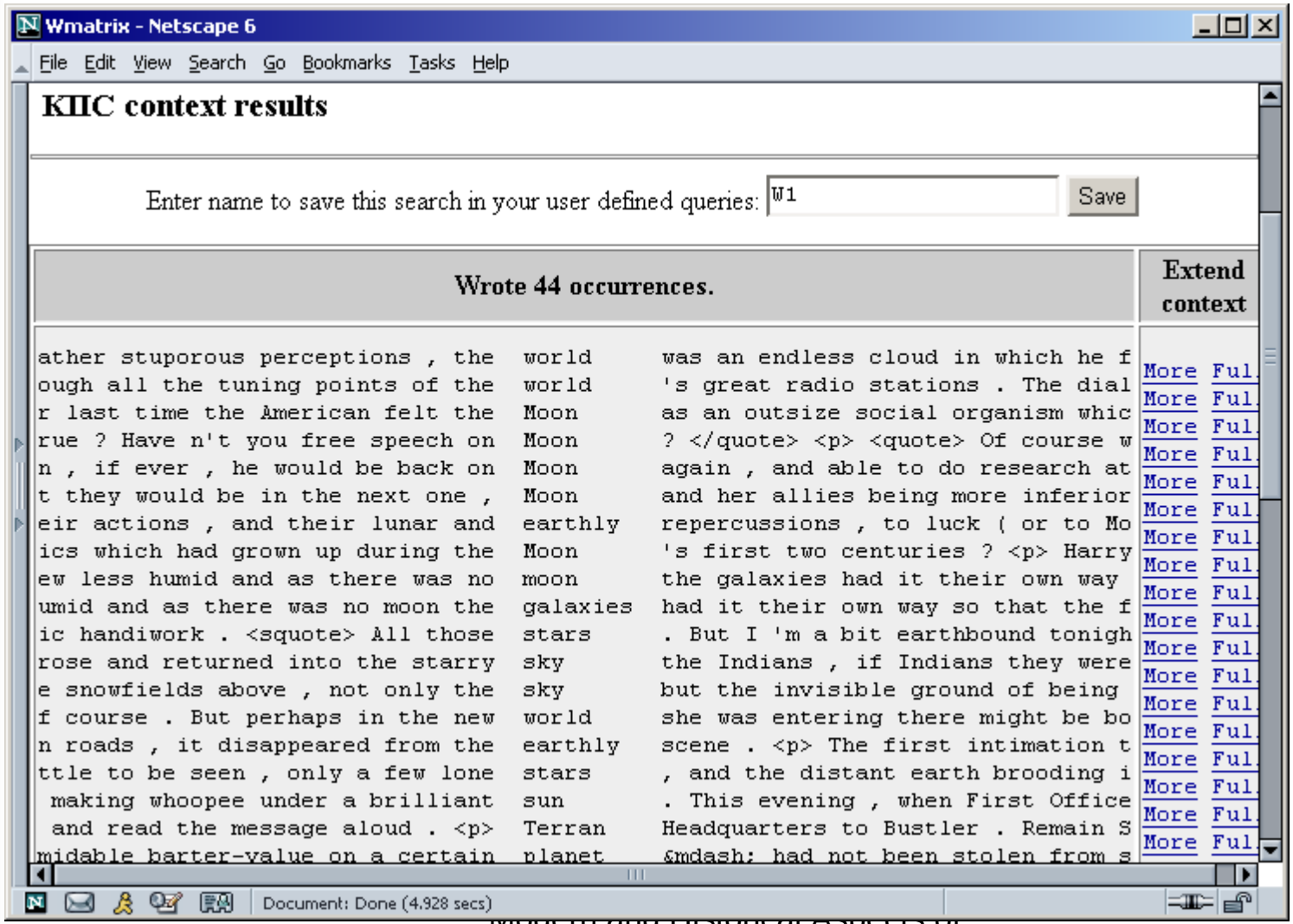
Wmatrix key items



Wmatrix frequency lists



Wmatrix KIIC



A1	GENERAL AND ABSTRACT TERMS
A1.1.1	General actions, making etc.
A1.1.2	Damaging and destroying
A1.2	Suitability
A1.3	Caution
A1.4	Chance, luck
A1.5	Use
A1.5.1	Using
A1.5.2	Usefulness
A1.6	Physical/mental
A1.7	Constraint
A1.8	Inclusion/Exclusion
A1.9	Avoiding
A2	Affect
A2.1	Affect:- Modify, change
A2.2	Affect:- Cause/Connected
A3	Being
A4	Classification
A4.1	Generally kinds, groups, examples
A4.2	Particular/general; detail
A5	Evaluation
A5.1	Evaluation:- Good/bad
A5.2	Evaluation:- True/false
A5.3	Evaluation:- Accuracy
A5.4	Evaluation:- Authenticity
A6	Comparing
A6.1	Comparing:- Similar/different
A6.2	Comparing:- Usual/unusual
A6.3	Comparing:- Variety
A7	Definite (+ modals)
A8	Seem
A9	Getting and giving; possession
A10	Open/closed; Hiding/Hidden; Finding; Showing
A11	Importance
A11.1	Importance: Important
A11.2	Importance: Noticeability
A12	Easy/difficult
A13	Degree
A13.1	Degree: Non-specific
A13.2	Degree: Maximizers
A13.3	Degree: Boosters
A13.4	Degree: Approximators
A13.5	Degree: Compromisers
A13.6	Degree: Diminishers
A13.7	Degree: Minimizers
A14	Exclusivizers/particularizers
A15	Safety/Danger
B1	Anatomy and physiology
B2	Health and disease
B3	medicines and medical treatment
B4	Cleaning and personal care
B5	Clothes and personal belongings
C1	Arts and crafts
E1	EMOTIONAL ACTIONS, STATES AND PROCESSES General
E2	Liking
E3	Calm/Violent/Angry
E4	Happy/sad
E4.1	Happy/sad: Happy
E4.2	Happy/sad: Contentment
E5	Fear/bravery/shock
E6	Worry, concern, confident
F1	Food
F2	Drinks
F3	Cigarettes and drugs
F4	Farming & Horticulture
G1	Government, Politics and elections
G1.1	Government etc.
G1.2	Politics
G2	Crime, law and order
G2.1	Crime, law and order: Law and order
G2.2	General ethics
G3	Warfare, defence and the army; weapons
H1	Architecture and kinds of houses and buildings
H2	Parts of buildings
H3	Areas around or near houses
H4	Residence
H5	Furniture and household fittings
I1	Money generally
I1.1	Money: Affluence
I1.2	Money: Debts
I1.3	Money: Price
I2	Business
I2.1	Business: Generally
I2.2	Business: Selling
I3	Work and employment
I3.1	Work and employment: Generally
I3.2	Work and employment: Professionalism
I4	Industry
K1	Entertainment generally
K2	Music and related activities
K3	Recorded sound etc.
K4	Drama, the theatre and showbusiness
K5	Sports and games generally
K5.1	Sports
K5.2	Games
K6	Childrens games and toys
L1	Life and living things
L2	Living creatures generally
L3	Plants
M1	Moving, coming and going
M2	Putting, taking, pulling, pushing, transporting &c.
M3	Vehicles and transport on land
M4	Shipping, swimming etc.
M5	Aircraft and flying
M6	Location and direction
M7	Places
M8	Remaining/stationary
N1	Numbers
N2	Mathematics
N3	Measurement
N3.1	Measurement: General
N3.2	Measurement: Size
N3.3	Measurement: Distance
N3.4	Measurement: Volume
N3.5	Measurement: Weight
N3.6	Measurement: Area
N3.7	Measurement: Length & height
N3.8	Measurement: Speed
N4	Linear order
N5	Quantities
N5.1	Entirity; maximum
N5.2	Exceeding; waste
N6	Frequency etc.
O1	Substances and materials generally
O1.1	Substances and materials generally: Solid
O1.2	Substances and materials generally: Liquid
O1.3	Substances and materials generally: Gas
O2	Objects generally
O3	Electricity and electrical equipment
O4	Physical attributes
O4.1	General appearance and physical properties

O4.2	Judgement of appearance (pretty etc.)
O4.3	Colour and colour patterns
O4.4	Shape
O4.5	Texture
O4.6	Temperature
P1	Education in general
Q1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.2	Paper documents and writing
Q1.3	Telecommunications
Q2	Speech acts
Q2.1	Speech etc:- Communicative
Q2.2	Speech acts
Q3	Language, speech and grammar
Q4	The Media
Q4.1	The Media:- Books
Q4.2	The Media:- Newspapers etc.
Q4.3	The Media:- TV, Radio and Cinema
S1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.2	Reciprocity
S1.1.3	Participation
S1.1.4	Deserve etc.
S1.2	Personality traits
S1.2.1	Approachability and Friendliness
S1.2.2	Avarice
S1.2.3	Egoism
S1.2.4	Politeness
S1.2.5	Toughness; strong/weak
S1.2.6	Sensible
S2	People
S2.1	People:- Female
S2.2	People:- Male
S3	Relationship
S3.1	Relationship: General
S3.2	Relationship: Intimate/sexual
S4	Kin
S5	Groups and affiliation
S6	Obligation and necessity
S7	Power relationship
S7.1	Power, organizing
S7.2	Respect
S7.3	Competition
S7.4	Permission
S8	Helping/hindering
S9	Religion and the supernatural
T1	Time
T1.1	Time: General
T1.1.1	Time: General: Past
T1.1.2	Time: General: Present; simultaneous
T1.1.3	Time: General: Future
T1.2	Time: Momentary
T1.3	Time: Period
T2	Time: Beginning and ending
T3	Time: Old, new and young; age
T4	Time: Early/late
W1	The universe
W2	Light
W3	Geographical terms
W4	Weather
W5	Green issues
X1	PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES
X2	Mental actions and processes
X2.1	Thought, belief
X2.2	Knowledge
X2.3	Learn
X2.4	Investigate, examine, test, search
X2.5	Understand
X2.6	Expect
X3	Sensory
X3.1	Sensory:- Taste
X3.2	Sensory:- Sound
X3.3	Sensory:- Touch
X3.4	Sensory:- Sight
X3.5	Sensory:- Smell
X4	Mental object
X4.1	Mental object:- Conceptual object
X4.2	Mental object:- Means, method
X5	Attention
X5.1	Attention
X5.2	Interest/boredom/excited/energetic
X6	Deciding
X7	Wanting; planning; choosing
X8	Trying
X9	Ability
X9.1	Ability:- Ability, intelligence
X9.2	Ability:- Success and failure
Y1	Science and technology in general
Y2	Information technology and computing
Z0	Unmatched proper noun
Z1	Personal names
Z2	Geographical names
Z3	Other proper names
Z4	Discourse Bin
Z5	Grammatical bin
Z6	Negative
Z7	If
Z8	Pronouns etc.
Z9	Trash can
Z99	Unmatched