

The Australian Qualitative Archive

A Node of the Australian Social Sciences Data Archive





Online Archiving and Mining of Qualitative Data

Andrew E. Smith

**Key Centre for Human Factors & Applied Cognitive Psychology
The University of Queensland
asmith@humanfactors.uq.edu.au**



A Question

Consider the value of these fragments of qualitative data from our history:

The Code of Hammurabi, Homer, Aristotle's Library, The Rosetta Stone, Plato, The Dead Sea Scrolls, Ancient Cave Paintings, Works of the Gawain Poet, ...

Our understanding of the contexts of these fragments is at best incomplete.

Should we not, in our turn, archive as much descriptive information as we can?

The Opportunity

- Large amounts of unstructured data are available for researchers to explore and use to test their hypotheses.
- Longitudinal and cross cultural qualitative data may exist which can inform research questions.
- Rich qualitative data, which contains more information on its context, survives better over time and space.
- The pace of social change and disruption requires urgent social research with broad scope.



Some Problems

How do researchers find and interrogate larger quantities of unstructured data?

How do they reinstate as much of the original meaning as possible?

How do researchers find and interrogate larger quantities of unstructured data?

- **Search** is handicapped by changing word usage – with **time** and **context**.
- Only fragments of documents may be relevant – **granularity**.
- **Static classification** can be **unreliable**, and **irrelevant** to current needs. Classification decisions are subject to contextual influences (Nisbett & Wilson, Tversky & Kahneman)



How do researchers find and interrogate larger quantities of unstructured data?

- Use **Bootstrapped Concept Learning** to map researcher's selected search terms onto more relevant terms in target data collection.
- Resulting concept **thesaurus** used to classify and retrieve related text data fragments.

How do researchers reinstate as much of the original meaning as possible?

- **Archived Metadata** – background and method description; time and place data. Vital for image and video data.
- **Cognitive Mapping of Data and Metadata** – to provide **awareness** to researcher of contextual meaning of the data set.



Using Leximancer Text Analysis to help address these issues

- Leximancer performs unsupervised or semi-supervised **concept discovery** from text.
- Bootstraps thesaurus classifiers from **seed terms**.
- Classifies and indexes **small text segments**, such as individual survey responses.
- Generates a **concept map** of the material.

Functional Design of AQUA Archiving

Gateway – search engine to retrieve **data bundles** using metadata, plus authentication and access control.

Retrieved data bundles can be:

- **downloaded,**
- **browsed,** or
- **analysed online.**

Functional Design of AQUA Archiving

Data Bundle – Qualitative **data** plus supporting metadata describing **context** and **method**.

Expressed using QDIF (Qualitative Data Interchange Format).

➤ **original data** plus collection background and method, along with an unsupervised **Leximancer bundle** for exploration support,

OR

➤ **secondary analysis** – from CAQDAS software.

Functional Design of AQUA Archiving

Metadata - contextualisation

- Assigned at deposition.
- Title, Author, Date, Method, Background,...
- Topical metadata extracted using Leximancer and stored as concept co-occurrence matrix.



Concept Matrix extracted from Enron Email Data



	A	E	F	G	H	I	J	K	L	M
1	Entity	Enron	power	California	market	should	time	energy	electricity	information g
2	Enron	221719	11111	7189	9973	10820	11981	13826	7086	5541
3	power	11111	96591	21900	13661	4250	4684	10199	19440	1208
4	California	7189	21900	81953	11166	3700	3985	11560	16620	900
5	market	9973	13661	11166	76968	5298	5588	5970	9833	1943
6	should	10820	4250	3700	5298	143538	9293	2144	2380	6169
7	time	11981	4684	3985	5588	9293	144796	2679	2887	2632
8	energy	13826	10199	11560	5970	2144	2679	54623	6681	1407
9	electricity	7086	19440	16620	9833	2380	2887	6681	62612	614
10	information	5541	1208	900	1943	6169	2632	1407	614	65587
11	gas	6052	6554	3963	4561	2317	2069	3699	5365	500
12	state	4454	21995	14045	4829	2345	2361	7267	12116	541
13	message	5435	51	80	40	2873	425	59	9	10597
14	e-mail	6683	115	95	172	899	928	143	75	5667
15	natural	3593	4192	2485	2798	1159	1158	2681	3385	276
16	price	3906	6169	5998	7297	2558	2504	2998	4860	669
17	business	8706	1718	1011	2131	1866	1993	2369	1380	2484
18	year	6081	2863	2841	2189	1440	2159	2924	1962	470
19	million	8371	2703	2241	1144	728	804	2531	1458	256
20	Houston	10574	1552	1121	1168	1850	5265	2636	1208	1121
21	review	5969	194	110	219	902	745	135	72	3430
22	receive	5110	155	587	105	485	202	157	105	1818

Functional Design of AQuA Archiving

Leximancer Data Bundle

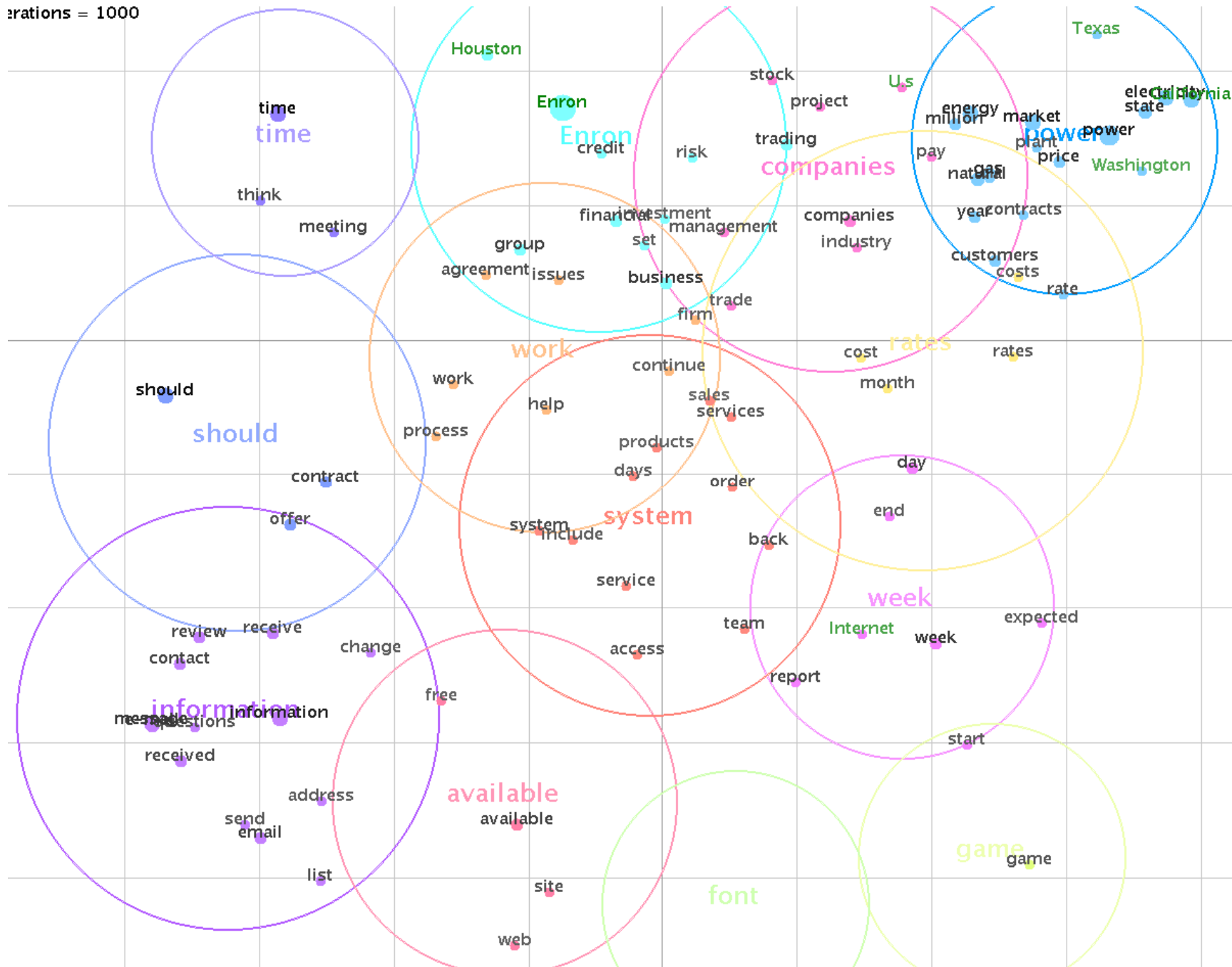
- Added to **original data** bundle at deposition.
- Unsupervised **topical and relational map**.
- Allows researcher to browse data bundle for relevant data.



Unsupervised Leximancer Map of Enron Email Data



Iterations = 1000



Functional Design of AQuA Archiving

Online Analysis in Leximancer

- **New map** of original text data.
- Further work on **existing secondary analysis**.
- Category analysis.
- **Customisation** of concepts.
- **Profiling** to explore framing of selected topics.

Leximancer Thesaurus

Primary Classes
Converged in **6** iterations.

[Go to next concept set.](#)

- illnesses

Secondary Classes
Converged in **8** iterations.

[Go to previous concept set.](#)
[Go to next concept set.](#)

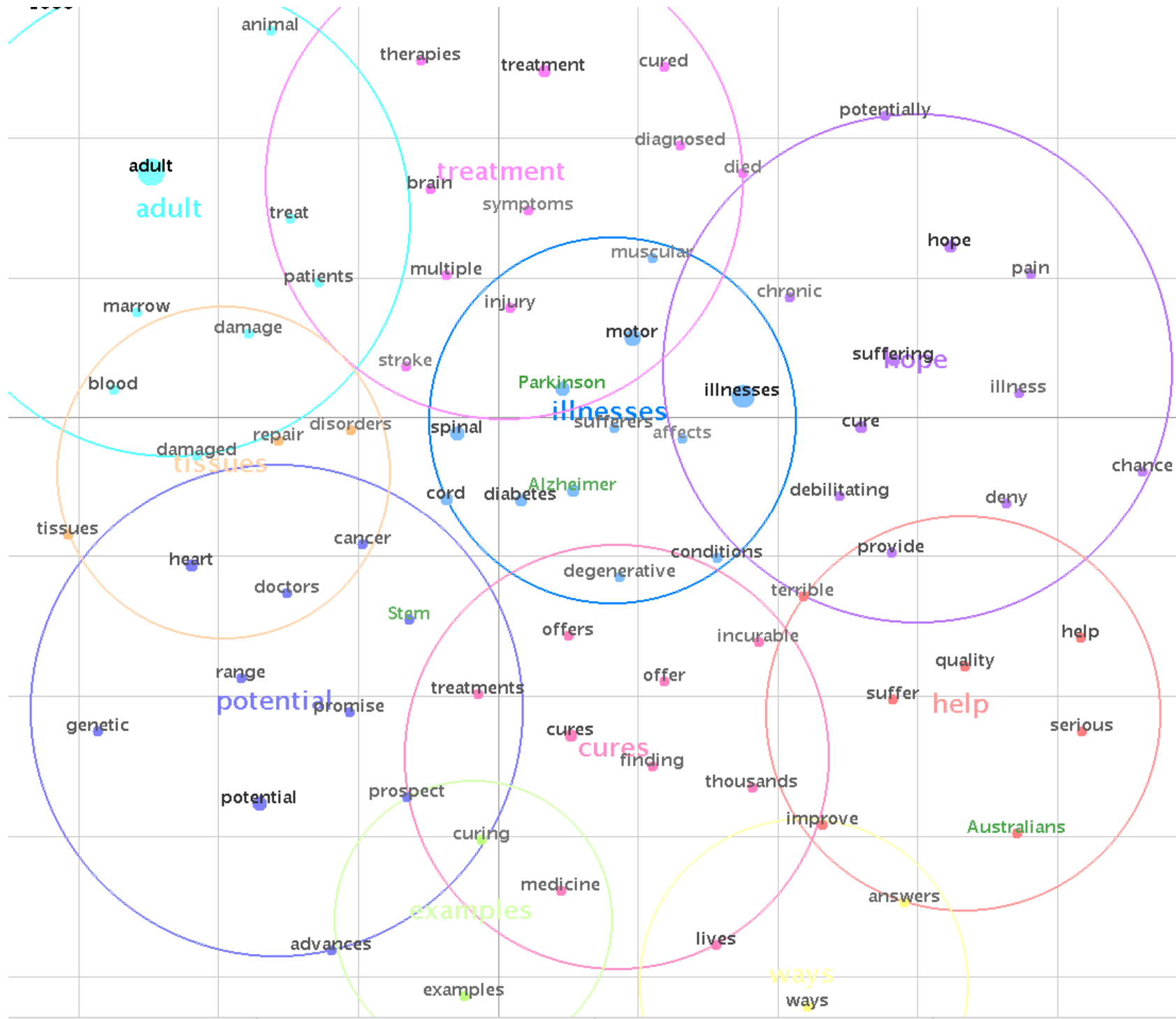
- adult
- advances
- affects
- Alzheimer
- animal
- answers
- Australians
- blood
- brain
- cancer
- chance
- chronic
- conditions
- cord

illnesses:

- diseases -> 7.9612
- [[parkinson]] -> 7.058
- [[alzheimer]] -> 6.4293
- illnesses -> 6.1579
- neurone -> 6.1095
- degenerative -> 5.5687
- disease -> 5.4958
- sufferers -> 5.4805
- died -> 4.8485
- stroke -> 4.8485
- symptoms -> 4.6475
- muscular -> 4.6475
- dystrophy -> 4.5276
- affects -> 4.5276
- [[tom]] -> 4.3896
- debilitating -> 4.3584
- breast -> 4.2266
- brains -> 4.2266
- dependent -> 4.2266
- sickle -> 4.2266
- prevention -> 4.2266
- helping -> 4.2266
- motor -> 4.1711
- afflicted -> 4.0277
- diabetics -> 4.0277
- fatal -> 4.0277
- fibrosis -> 4.0277
- sourced -> 4.0277
- anaemia -> 4.0277
- neurones -> 4.0277
- life-threatening -> 4.0277
- [[rett]] -> 4.0277
- cystic -> 4.0277
- dopamine -> 4.0277
- plague -> 3.7717



Profile, or Framing, of Concept - **Illnesses**





Future Plans

Scale up for operational use:

- Upgrade hardware and hosting.
- Test usability.
- Revise metadata schema.
- Establish and document procedure for deposition, access control, and management.



Acknowledgement

This research is supported by Australian Research Council grant LE0560677: Australian Social Science Data Archive: Facility Enhancement and Network Development.

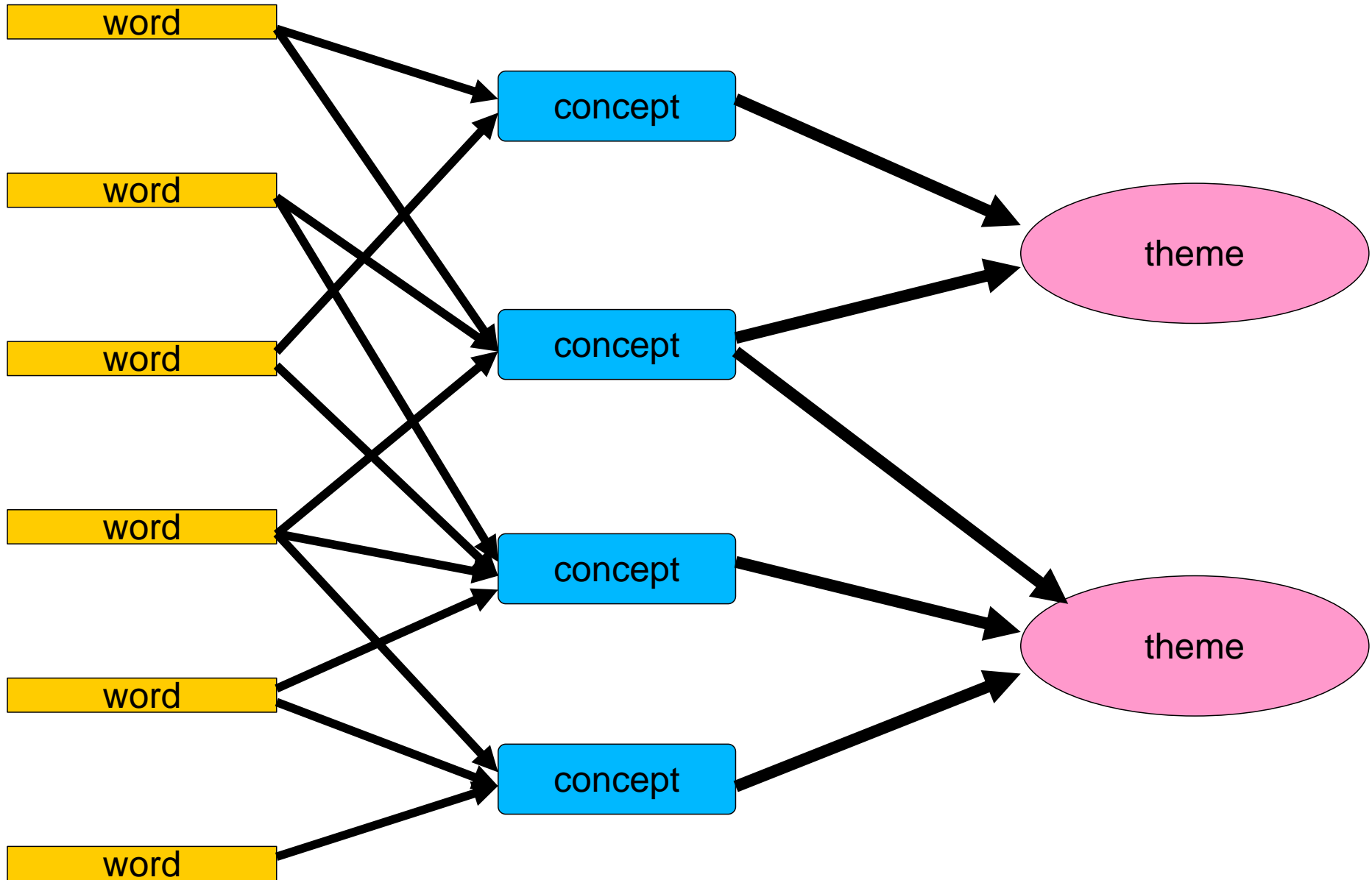


Leximancer

**A Method for Analysing Large
Quantities of Text**

Aims of the System

- To **automatically** extract the principal **concepts** from a body of text.
- To allow the user to encode further **custom concepts**.
- To automatically code text segments.
- To quantify and explore the resulting concept space.

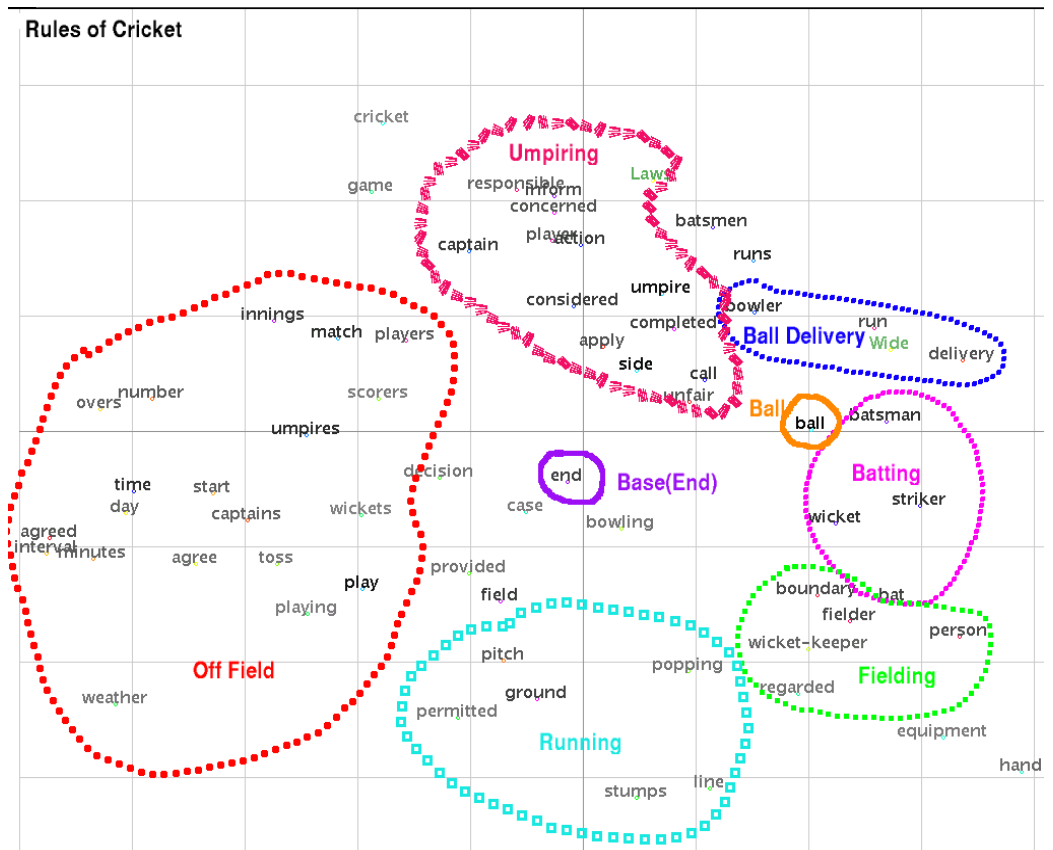
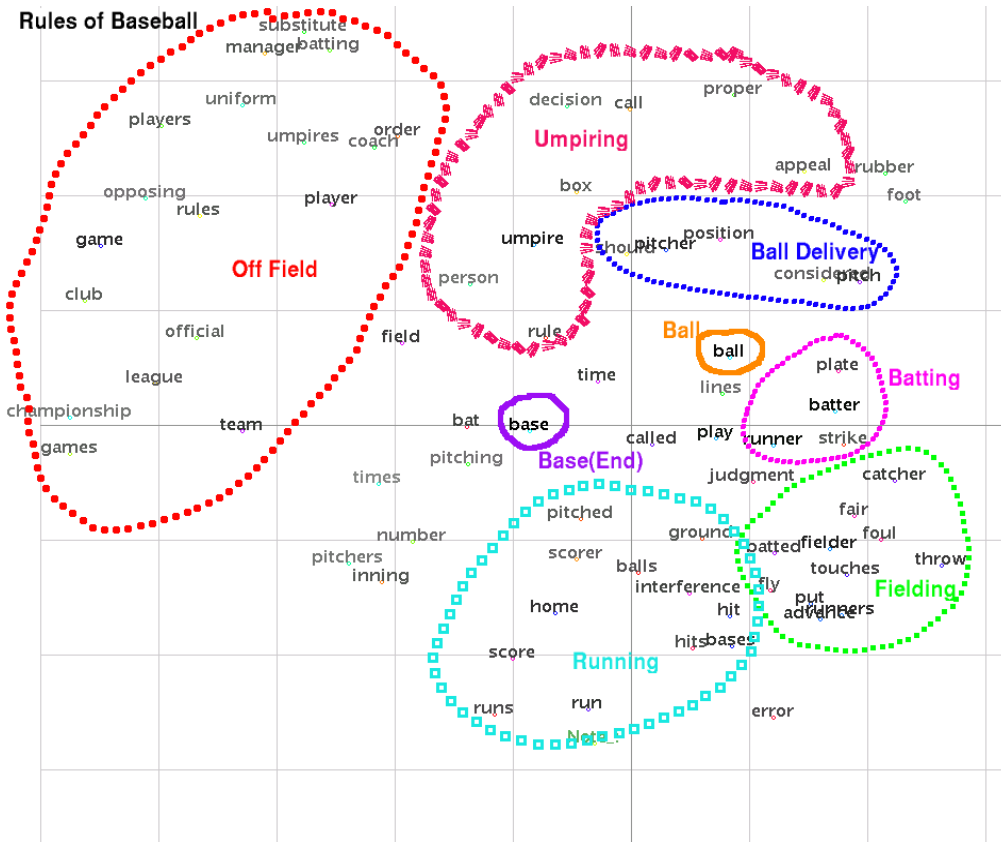




Evaluation of Unsupervised Semantic Mapping of Natural Language with Leximancer Concept Mapping

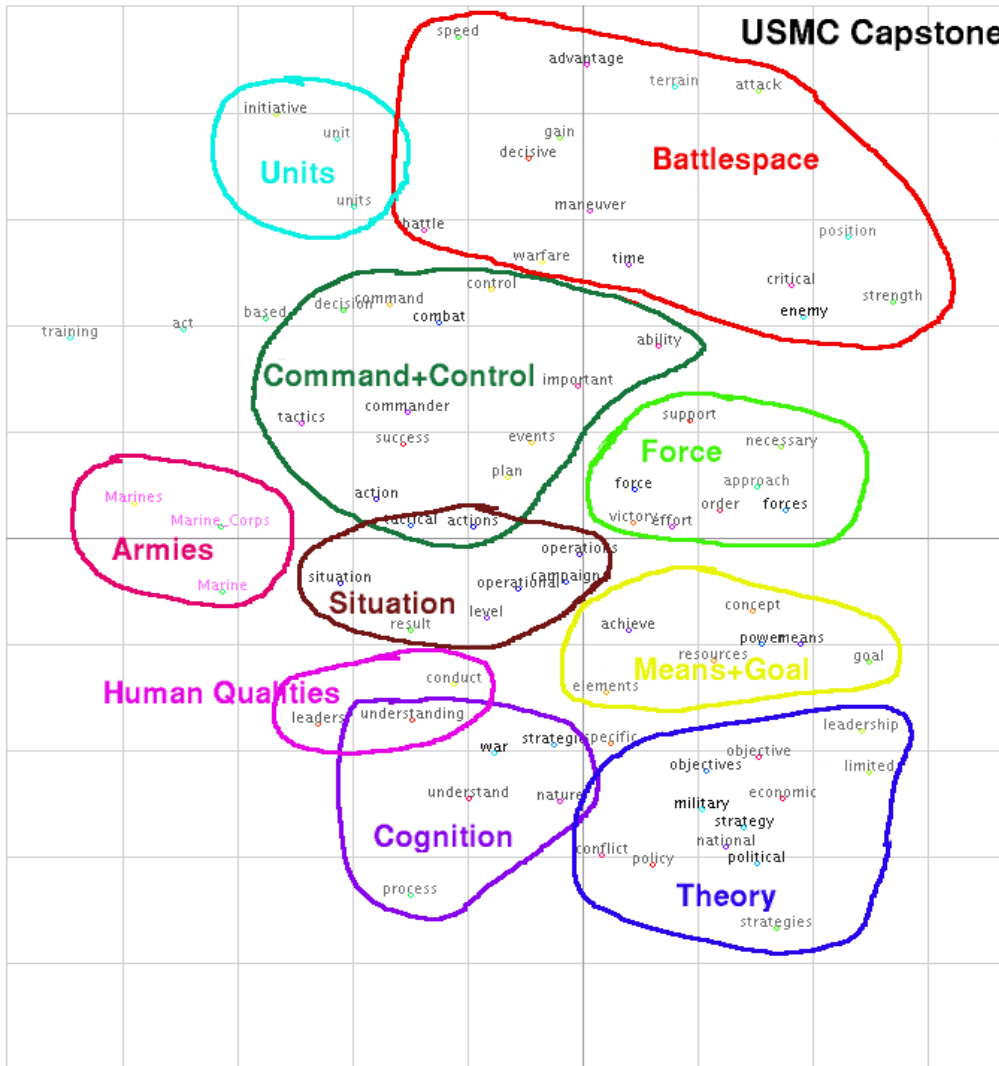
Andrew Smith & Mike Humphreys
Behavior Research Methods
(to appear)

Validation - Rules of Baseball and Cricket



Validation - Military Doctrine, Old and New

USMC Capstone Doctrine



Clausewitz - "On War"

