# PROFILING MEDICAL JOURNAL ARTICLES USING A GENE ONTOLOGY SEMANTIC TAGGER

MAHMOUD EL-HAJ

PAUL RAYSON

SCOTT PIAO

JO KNIGHT

SHERYL PRENTICE

NATHAN RUTHERFORD

# ORIGIN AND OUTCOMES

- Currently funded through a Wellcome Trust Seed award
- Collaboration with UCREL through DSI
- Previous work presented at International Genetic Epidemiology Society (2017), LREC (2018)
- Explosion of literature in human medical genetics (e.g. genome wide association studies: 5 in 1995, 141 in 2005, 3,633 in 2015)
- Goal is to utilise text mining and corpus linguistic methods to feed into a tool that will assist the generation of new hypotheses from this rapidly growing set of data

# DATASET

- Medical journal abstracts from PubMed
- English articles discussing human genetics studies in psychiatry and immune related disorders.

# DATASET

| Corpus | #Articles | #Words | Keywords |
|---|---|---|---|
| Immune | 21.5K | 4.8M | (geneti* OR gene OR genot*) AND (immunol* OR immunog* OR immune) |
| Psychiatric | 15.2K | 2.8M | (geneti* OR gene OR genot*) AND (psychi) |
| Reference | 296.5K | 79.0M | (geneti* OR gene OR genot*) |
| Total | 333.2K | 86.7M | |

# DATA EXTRACTION

- Search PubMed website directly
- Saved results to large XML file
- Built a Java Suite for parsing PubMed XML file format.
- Java suite extracts abstracts, titles, authors, pub-date, DOI …etc.
- Code freely available on github:

  https://github.com/drelhaj/BioTextMining

# FINE-GRAINED MEDICAL TERMS

- Next step…tagging
- Currently used methods not suitable for specialist terms, e.g. cytokines, lymphocyte mediated immunity
- Extra level of annotation required for tagging
- The Gene Ontology Consortium's[1] OBO Basic Gene Ontology (go-basic.obo) categories[2].
- What is GO (Gene Ontology)? Provides consistent descriptions of gene products across databases.
- Focussed on tracing ancestors and children for each entry in the ontology

[1] http://geneontology.org/

[2] http://purl.obolibrary.org/obo/go/go-basic.obo

# GENE ONTOLOGY SEMANTIC TAGGER

- Corpora uploaded to Wmatrix
- POS tagged using CLAWS.
- Semantically tagged using USAS
- Counted frequencies
- Compared sub-corpora using methods from Corpus Linguistics.

# PARSING OBO

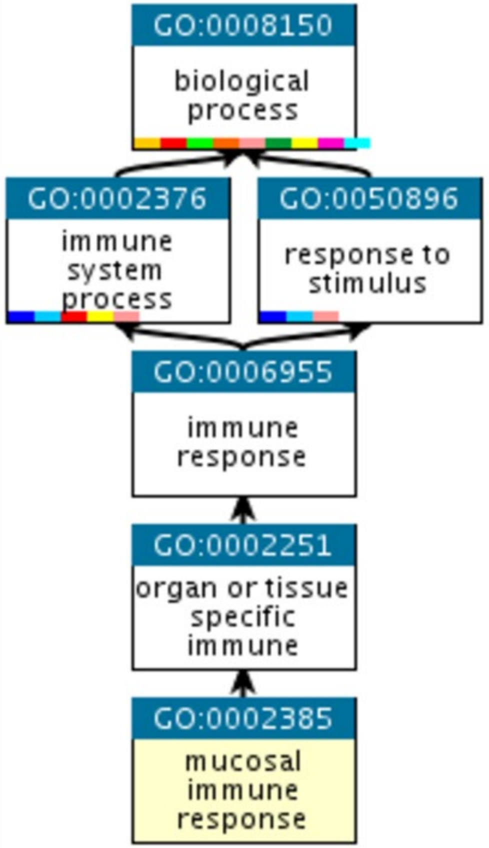
- we created Java code that combines the use of publicly available OBO library[1]
- with Java Directed Graph (Digraphs)
- to trace the paths from a node child to the root.
- The code used Breadth First and Depth First algorithms to quickly and accurately extract the paths.

[1] https://github.com/sugang/bioparser

[2] http://purl.obolibrary.org/obo/go/go-basic.obo

# OBO GRAPH SAMPLE

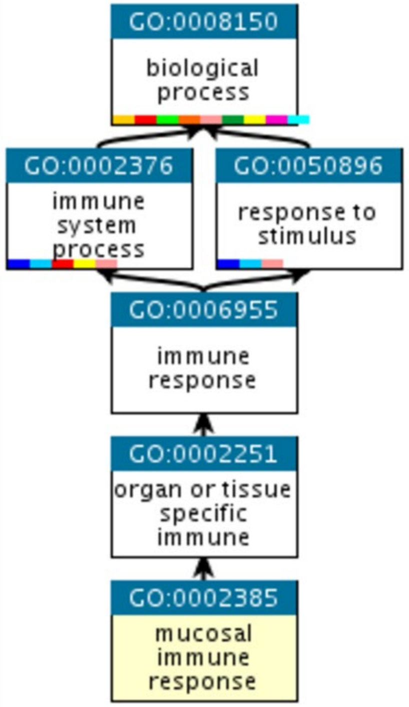
- Our code allowed us to generate a USAS tagger dictionary file
- where each entry in the OBO ontology is tagged with the GO IDs shown in its path.
- In the figure we can see two paths from the child node towards the ``biological process" root.



# DICTIONARY CREATION

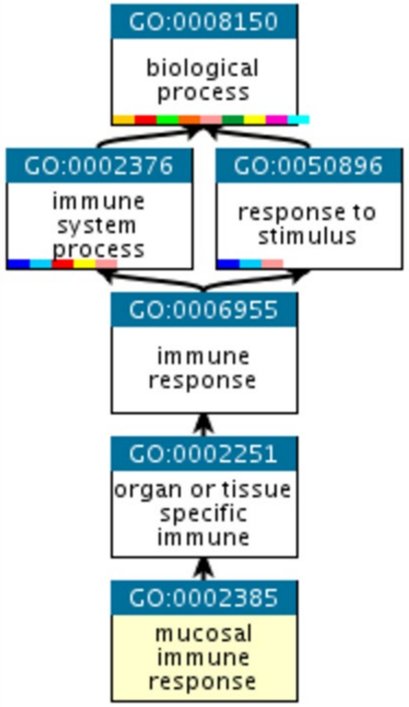
The dictionary creation process works as follows:

1. Is child node single word or multi-word expression.
2. get number of paths towards the root.
3. get each path's GoID entries (child node's ancestors)
4. include the level of each ancestor by adding that to the end of each entry (e.g. .1 to refer to the first parent (GOO:0002251).
5. Check if path passes through an ``immune system process" (i.e. GoID: 0002376).
6. If so we add .I to the end of the GoID tag to refer to immune entry, otherwise we add .N referring to a non-immune entry.



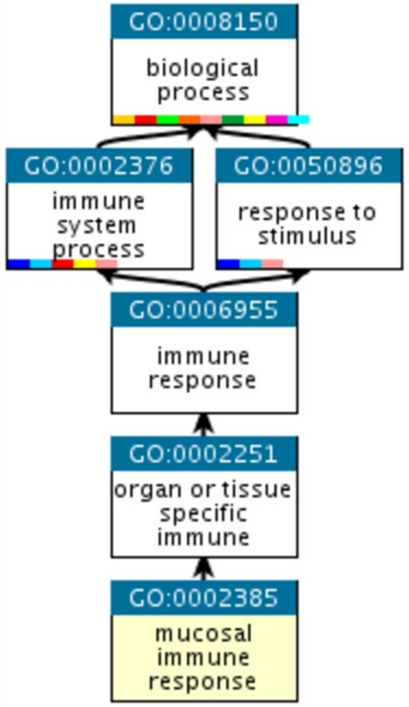
# TAGGING EXAMPLE

- Following the steps in previous slide, the child node GO:0002385 is multi-word expression entry with following semantic dictionary tags:
- {GO:0008150.4.I, GO:0002376.3.I, GO:0050896.3.N, GO:0006955.2.I, GO:0002385.0.I, GO:0002251.1.N, GO:0006955.2.N, GO:0002385.0.N, GO:0002251.1.I, GO:0008150.4.N}.



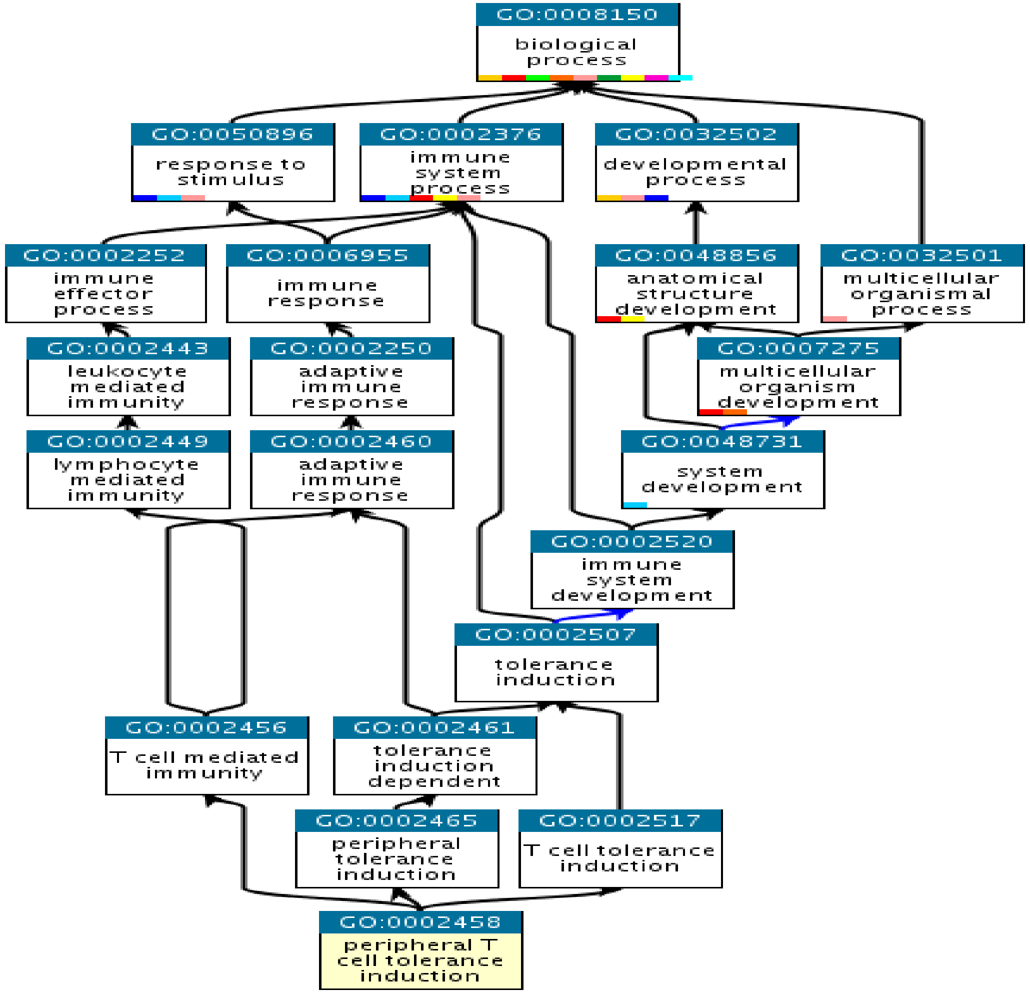
# TAGGING EXAMPLE

- Tags such as GO:0006955 ends with .2 suffix referring to level two (counting from level zero).
- and will appear twice;
  - once as an immune entry with a .I suffix (GO:0006955.2.I)
  - and another as a non-immune entry with a .N suffix (GO:0006955.2.N).



# COMPLEX EXAMPLE

- Dictionary creation can be complex
  - Overlapping hierarchies
  - Levels that can be skipped



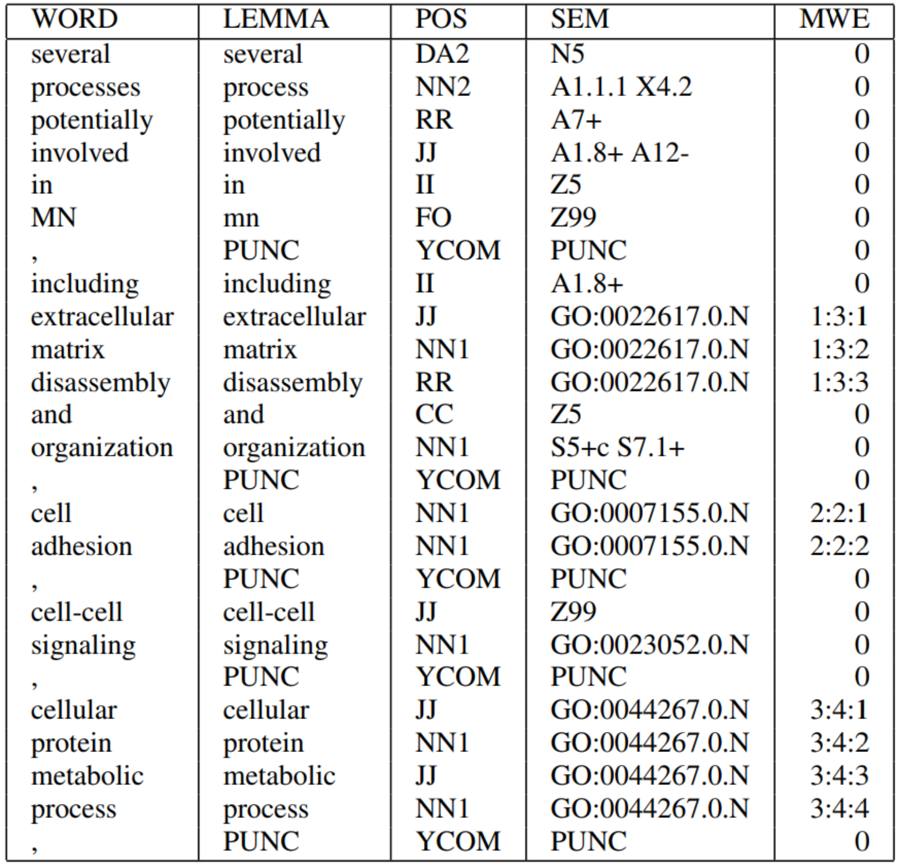
# GOST

- The resultant GO term and ID map collection from the process described above contains:
  - 433 single word bioterms
  - and 44,180 multiword bioterms
- merged into the Lancaster UCREL Semantic lexicons to create a new version of the Lancaster USAS semantic annotation system named:

“GOST” (Gene Ontology Semantic Tagger)

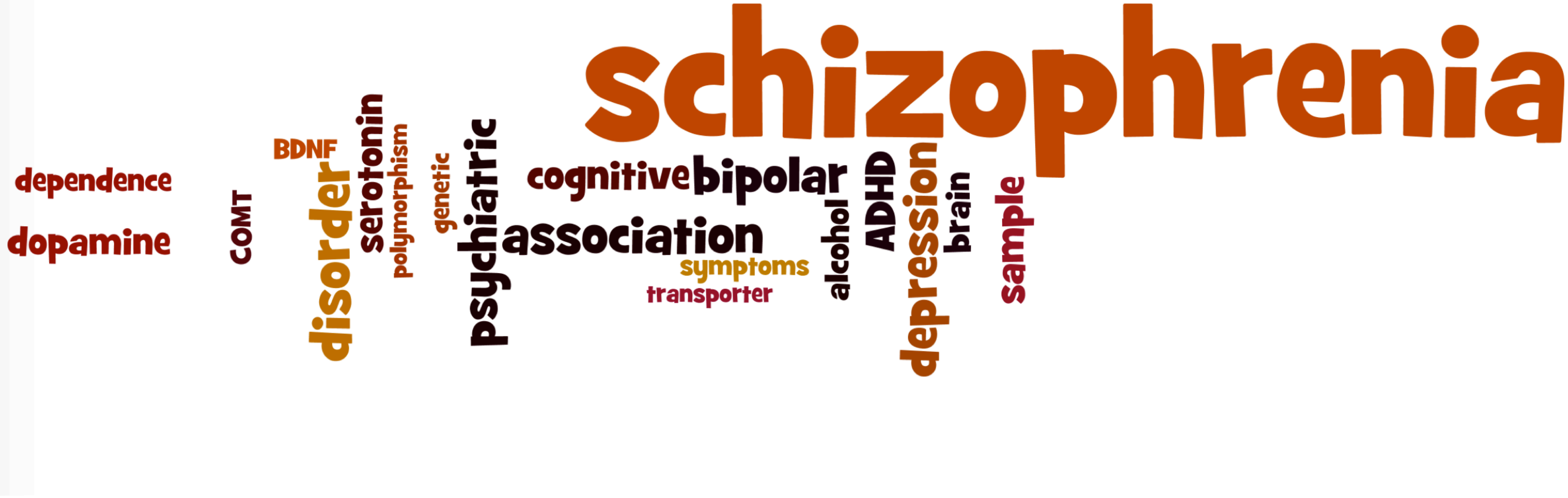
# USING THE GOST

- Using the GOST, we have tagged 237,615 PubMed abstracts in our corpus.
- This corpus provides a valuable new resource for mining Biomedical and health information from the Biomedical literature.
- The table shows a sample from a tagged abstract, where the part-of-speech tags are from CLAWS C7 tagset
- the generic semantic tags are from the USAS tagset,
- and the MWE tags encode multiword term information including sequential number, term length and location of each word in the given term.



| WORD | LEMMA | POS | SEM | MWE |
|---|---|---|---|---|
| several | several | DA2 | N5 | 0 |
| processes | process | NN2 | A1.1.1 X4.2 | 0 |
| potentially | potentially | RR | A7+ | 0 |
| involved | involved | JJ | A1.8+ A12- | 0 |
| in | in | II | Z5 | 0 |
| MN | mn | FO | Z99 | 0 |
| , | PUNC | YCOM | PUNC | 0 |
| including | including | II | A1.8+ | 0 |
| extracellular | extracellular | JJ | GO:0022617.0.N | 1:3:1 |
| matrix | matrix | NN1 | GO:0022617.0.N | 1:3:2 |
| disassembly | disassembly | RR | GO:0022617.0.N | 1:3:3 |
| and | and | CC | Z5 | 0 |
| organization | organization | NN1 | S5+c S7.1+ | 0 |
| , | PUNC | YCOM | PUNC | 0 |
| cell | cell | NN1 | GO:0007155.0.N | 2:2:1 |
| adhesion | adhesion | NN1 | GO:0007155.0.N | 2:2:2 |
| , | PUNC | YCOM | PUNC | 0 |
| cell-cell | cell-cell | JJ | Z99 | 0 |
| signaling | signaling | NN1 | GO:0023052.0.N | 0 |
| , | PUNC | YCOM | PUNC | 0 |
| cellular | cellular | JJ | GO:0044267.0.N | 3:4:1 |
| protein | protein | NN1 | GO:0044267.0.N | 3:4:2 |
| metabolic | metabolic | JJ | GO:0044267.0.N | 3:4:3 |
| process | process | NN1 | GO:0044267.0.N | 3:4:4 |
| , | PUNC | YCOM | PUNC | 0 |

# RESULTS – WORD COMPARISON



# RESULTS - NEW GOST ANNOTATED CORPORA

| GOID | NAME | IMMUNE | % | PSYCH | % | O/U | Keyness |
|---|---|---|---|---|---|---|---|
| GO:0005623 | cell | 33346 | 7.31 | 1524 | 1.02 | + | 10696.95 |
| GO:0005575 | Cellular Component | 34577 | 7.58 | 1808 | 1.20 | + | 10332.02 |
| GO:0007610 | behavior | 199 | 0.04 | 2095 | 1.40 | - | 4611.01 |
| GO:0032501 | multicellular organismal process | 616 | 0.13 | 2364 | 1.57 | - | 3915.62 |
| GO:0008150 | Biological Process | 7253 | 1.59 | 88 | 0.06 | + | 3416.63 |
| GO:0006955 | immune response | 6992 | 1.53 | 84 | 0.06 | + | 3298.74 |
| GO:0050877 | neurological system process | 426 | 0.09 | 1756 | 1.17 | - | 2991.92 |
| GO:0050896 | response to stimulus | 7034 | 1.54 | 192 | 0.13 | + | 2764.12 |
| GO:0002376 | immune system process | 2958 | 0.65 | 28 | 0.02 | + | 1443.03 |
| GO:0050890 | cognition | 10 | 0.00 | 536 | 0.36 | - | 1402.85 |

# REMOVING COMMON TERMS

- Previous corpus linguistic research has identified formulaic language
- Created sample corpora of Immune Genetic, Psych Genetic, Immune General and Psych General abstracts
- Corpora were used to create a word family list (using existing lemma list and manual process)
- Word family list used to generate keyness, consistency, lock word, and glossary comparison analyses

# REMOVING COMMON TERMS

- Keyness analysis:
  - identified expected terms when genetics corpora were compared (e.g. disorder names) and when genetics corpora were compared with general corpora (e.g. ‘polymorphism’, ‘allele’)
  - However, ‘identify’, ‘analysis’ (labelling) and ‘we’ overused in genetic literature, many items
- Consistency analysis:
  - ‘be’, ‘have’, ‘use’, ‘study’ stable across corpora, ‘infect’, ‘cell’, ‘response’ and ‘disease’ across the immune corpora, and ‘disorder’ across the psych corpora
  - However, ‘express’ family more characteristic of Immune Genetic corpus, cut-off difficulty
- Lock words:
  - Low variance terms include function words, high variance include ‘disease’, ‘response’, ‘cell’, and ‘disorder’
  - Overall, assigned to appropriate stop lists, but ‘we’ + ‘analysis’ general and ‘express’ genetic, too restrictive

# REMOVING COMMON TERMS

- Glossary comparison: Common terms and MWEs extracted from varying sources to create psych (e.g. ICD-11), immune (e.g. ontology lookup service) and genetics (e.g. NHGRI) glossaries
- Each glossary compared with word list and MWE list created from the sample corpora (at least 3 texts)
- 685 genetic glossary items (overlapped with 174 genetics corpora items), 629 immune glossary items (overlapped with 57 immune corpus items), 274 psych glossary items (overlapped with 42 corpus items)
- Advantage: what we expect to find vs. What actually occurs in the data
- Disadvantage: Misses many common items identified in the CL analysis

- Best approach: combine glossary overlap items with items common to CL analyses to arrive at overall stop lists

# CONCLUSION AND FUTURE WORK

- A method for the creation of a semantic lexicon from an existing Gene Ontology, a Gene Ontology Semantic Tagger (GOST)
- Applied to corpora of scientific papers
- Provided freely available annotated corpora
- Demonstrated the tools extending corpus and computational linguistics, which allows genomics researchers to get sensible answers

# DEMO GOST-BUSTER

- **Website:** http://ucrel-biotm-1.lancs.ac.uk
- **Code:** https://delta.lancs.ac.uk/BioTM/BUSTER

# BUSTER – DATA

- Full text from open access papers
- Queried from PubMed Central - https://www.ncbi.nlm.nih.gov/pmc/
  - “Autoimmune Disease AND GWAS”
- Semantic Tagging provided using USAS Semantic Tagger
- Indexed using LexiDB - corpus database management system
  - Fulfils corpus linguistics retrieval queries on multi-billion-word multiply-annotated corpora

**Website:** http://ucrel-biotm-1.lancs.ac.uk

# BUSTER – WHERE ARE WE NOW?

- Downloading query papers from PMC
- Running these papers through BUSTER NLP pipeline
- Indexing in LexiDB
- Concordance of query search terms

**Website:** http://ucrel-biotm-1.lancs.ac.uk

# BUSTER – WHERE ARE WE GOING?

- Concordance based on meta-data such as GO Tag
- Frequency lists based on sub-corpra in lexiDB
- Key Words based on generated frequency list
- N-Gram analysis
- Collocations
- Regular updates to GOST based on latest version of Gene-Ontology

**Website:** http://ucrel-biotm-1.lancs.ac.uk

# RESOURCES

- The corpora and Java code to parse and annotate the dataset in addition to the ontology lexicon are made publicly available for research purposes.

  https://github.com/drelhaj/BioTextMining

- The Gene Ontology Semantic Tagger will soon be released via the downloadable graphical interface.

  http://ucrel.lancs.ac.uk/usas/gui/

- Project information

  http://wp.lancs.ac.uk/btm/