UCREL NLP Summer School 2024

Session 3: Semantic and sentiment analysis

Paul Rayson & Daisy Lal

Types of ambiguity

- Pragmatic
  - Meaning in context
- Semantic
  - Meaning of the sentence
- Anaphoric
  - Referential
  - What/who does each pronoun refer to?
- Syntactic
  - Structure/grammar
  - How do the words fit together in the sentence?
- Lexical ambiguity
  - Word level
  - Word can have more than one meaning

Lexical ambiguity

- Question:
  - How many senses does *spring* have?
- Answer:
  - A: 3
  - B: 4
  - C: 5
  - D: 31

to arrange the escape of (someone) from prison. **17** (*intr*) *Archaic or poetic.* (of daylight or dawn) to begin to appear. ◆ *n* **18** the act or an instance of springing. **19** a leap, jump, or bound. **20a** the quality of resilience; elasticity. **20b** (*as modifier*): *spring steel.* **21** the act or an instance of moving rapidly back from a position of tension. **22a** a natural outflow of ground water, as forming the source of a stream. **22b** (*as modifier*): *spring water.* **23a** a device, such as a coil or strip of steel, that stores potential energy when it is compressed, stretched, or bent and releases it when the restraining force is removed. **23b** (*as modifier*): *a spring mattress.* **24** a structural defect such as a warp or bend. **25a** (*sometimes cap.*) the season of the year between winter and summer, astronomically from the March equinox to the June solstice in the N hemisphere and from the September equinox to the December solstice in the S hemisphere. **25b** (*as*

etc., running down the inside of the handle of a cricket bat, hockey stick, etc. **29** Also called: **spring line.** *Nautical.* a mooring line, usually one of a pair that cross amidships. **30** a flock of teal. **31** *Architect.* another name for **springing.** [Old English *springan;* related to Old Norse *springa,* Old High German *springan,* Sanskrit *sprhayati* he desires, Old Slavonic *pragu* grasshopper] ▸ **'springless** *adj* ▸ **'spring,like** *adj*

Definition of Collins Englis (21st Century

spring balance *or esp. U.S.* **spring scale** *n* a device in which an object to be weighed is attached to the end of a helical spring, the extension of which indicates the weight of the object on a calibrated scale.

Spring (https://dictionary.cambridge.org)

- *spring* was found in the Cambridge Advanced Learner's Dictionary at the entries listed below.
  - spring (MOVE QUICKLY)
  - spring (APPEAR SUDDENLY)
  - spring (SEASON)
  - spring (CURVED METAL)
  - spring (WATER)
  - box spring
  - spring chicken
  - spring-clean
  - spring greens
  - spring onion
  - spring roll
  - spring from sth
  - spring sth on sb
  - be full of the joys of spring
  - spring to life
  - spring to mind
  - a spring in your step

Word Sense Disambiguation (WSD): overview

- Defined as “the process of automatically deciding the senses of words in context”
- An important area of research for 40+ years (and still not solved, or even started for many of the world’s languages)
- Many applications further along the NLP pipeline and to other real world tasks e.g.
  - machine translation: French translation of ‘free’ in the context of ‘free prisoner’ is *libre* and ‘free gift’ is *gratuit*
  - Information retrieval: finding the correct *jaguar, pandas*, or *apple*, etc

Word Sense Disambiguation (WSD): definitions

- **Homonymy**: distinct meanings with no obvious connection, e.g.
  - *Bank*: ‘edge of a river’ and ‘financial institution’
- **Polysemy**: meanings are related in some way, e.g.
  - *Leak*: ‘let liquid escape’ and ‘give away information’
  - *Mouse*: ‘small furry animal with a tail’ and ‘computer equipment’

Word Sense Disambiguation (WSD): approaches and issues

- **Knowledge-based**: relies on large manually created knowledge sources e.g. machine readable dictionaries and WordNet, dictionary definition overlap, LDOCE
  - Only as good as the knowledge base
  - Most frequent sense is a very strong baseline
  - Can lexicographers agree on the number / grouping of senses?
- **Supervised machine learning**: use of Wikipedia links or other linked open data source
  - A variety of approaches: decision lists, naïve bayes, maximum entropy, k-nearest neighbour using contextual word embeddings, support vector machines
  - Time consuming to create training data
- **Unsupervised machine learning**: clustering by context similarity
- **Domain specific approaches**: one sense per discourse, e.g.
  - *Bat*: ‘sports equipment’ or ‘nocturnal mammal’
  - *Party*: ‘politics’ or ‘entertainment’

Word Sense Disambiguation (WSD): evaluation

- SensEval and SemEval evaluation exercises
  - https://web.eecs.umich.edu/~mihalcea/senseval/
  - https://siglex.org/
  - https://semeval.github.io/
- Lexical sample task, training data provided, untagged data to be tagged by participants systems

Further reading

- Eneko Agirre and Mark Stevenson (2023) Word sense disambiguation. In Ruslan Mitkov (ed.) The Oxford Handbook of Computational Linguistics, second edition.
- Speech and Language Processing (3rd edition 2023) chapter 23: https://web.stanford.edu/~jurafsky/slp3/23.pdf

Motivation for semantic analysis

- Understanding the meaning of words and texts
- Various types in the Natural Language Processing literature
  - Named entity recognition
  - Semantic role labelling
  - Word sense disambiguation
  - Summarisation
  - Sentiment analysis
  - Content analysis
  - Semantic field tagging

UCREL Semantic Analysis System (USAS)

What is Semantic Tagging?

Lancaster University

- Semantic field annotation has applications for conceptual or topic tagging:
- *Last_T1.1.1 year_T1.1.1* was_A3+ the_Z5 UK_Z2 's_Z5 second_N4 warmest_O4.6+++ *on_A11.2+ record_A11.2+* ,_PUNC *according_Z5 to_Z5* provisional_T1.3- data_X2.2 from_Z5 the_Z5 Met_S3.1 Office_I2.1/H1c ._PUNC This_Z8 *puts_X2.2-* it_Z8 just_A14 *behind_X2.2-* 2022_N1 ,_PUNC which_Z8 recorded_Q1.2 an_Z5 average_A6.2+ temperature_O4.6 of_Z5 only_A14 0.06C_Z99 higher_N3.7++ ._PUNC

- A3+ = being; A6.2 = comparing; A11.2 = importance; A14 = exclusivisers; H1 = architecture, buildings; I2.1 = business; N1 = numbers; N3.7 = measurement; N4 = linear order; O4.6 = temperature; Q1.2 = documents, writing; S3.1 = relationship; T1.1.1 = Time past; T1.3 = time period; X2.2 = knowledge; Z2 = geographical names; Z5 = grammatical bin; Z8 = pronouns etc; Z99 = unmatched

Multiword expressions: plain sailing?

- Phrasal verbs
  - *Stubbed out*
- Noun phrases
  - *Riding boots*
  - *Pony nuts*
- Proper names
  - *United States of America*
- *Named entities*
  - *23rd November 1963*
  - *British Broadcasting Corporation*
- Multiword prepositions
  - *In terms of*
  - *As soon as*
- Idiomatic expressions
  - *Spill the beans*
  - *A pain in the neck*

UCREL Semantic Analysis System (USAS)

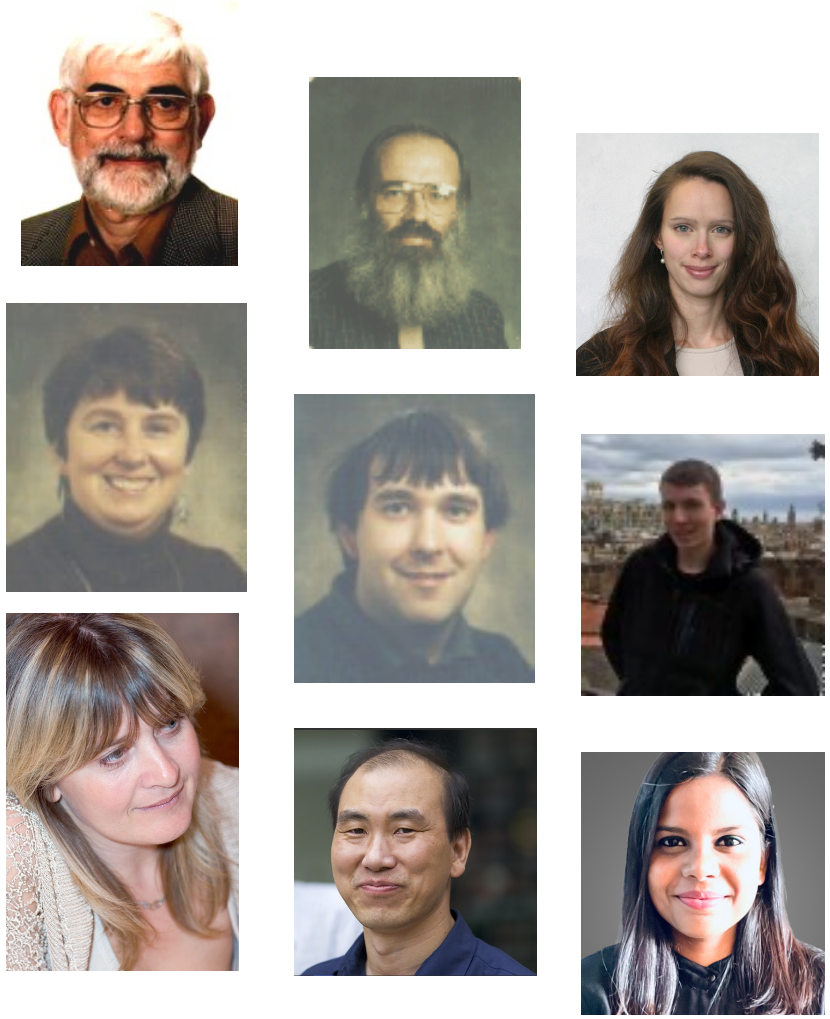
Lancaster University

- 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)

- https://ucrel.lancs.ac.uk/usas/
- Lexicons freely available:
  - https://github.com/UCREL/Multilingual-USAS

The work of many hands …

- Joint research with
  - Geoffrey Leech
  - Roger Garside
  - Jenny Thomas
  - Andrew Wilson
  - Dawn Archer
  - Scott Piao
  - Sheryl Prentice
  - Andrew Moore
  - Daisy Lal



Semantic fields

- AKA concepts, semantic domains
- ‘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

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 for English

- 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

English 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

English 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]

Evaluation (English 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.

Applications of semantic analysis 100+ papers listed at https://ucrel.lancs.ac.uk/wmatrix/

Lancaster University

- Analysis of market research interview transcripts
- Intelligent dictionaries
- Assistance for human translators
- Software Engineering domain understanding
- Language profiling for online child protection
- Actionability
- Corpus stylistics
- Prediction of real-world events from social media
- Metaphor and end-of-life care
- Pattern analysis of the language of psychopaths
- Political discourse analysis
- Describing the language of extremism and counter-extremism
- UK General Election Manifestos (Rayson 2008)

BRITISH NATIONAL CORPUS

ParlaMint

BNC BRITISH NATIONAL CORPUS 2014

CorCenCC Corpws Cenedlaethol Cymraeg Cyfoes National Corpus of Contemporary Welsh

Metaphor, cancer and end of life care (MELC)

Lancaster University

- Analysis of metaphorical language used to talk about cancer, dying and death: people 'fight' their cancer, 'win' or 'lose' their 'battle' against it, hope for a positive end to their cancer 'journey', and so on.
- 1.5M word corpus of interviews and online forum posts from patients, carers and healthcare professionals
- Methods: Manual analysis (MIP) and Wmatrix (Semantic analysis & concordancing)
- http://wp.lancs.ac.uk/melc/

G3 Warfare (e.g. *fight* as a verb, *battle*)
A1.1.1 General actions, making (e.g. *blast*, *confront*)
A1.1.2 Damaging and destroying (e.g. *destroy*, *shatter*)
E3– Violent/angry (e.g. *hit*, *attack*)
S8+ Helping (e.g. *defend*, *protect*)
S8– Hindering (e.g. *fight* as a noun)
X8+ Trying hard (e.g. *struggle*)

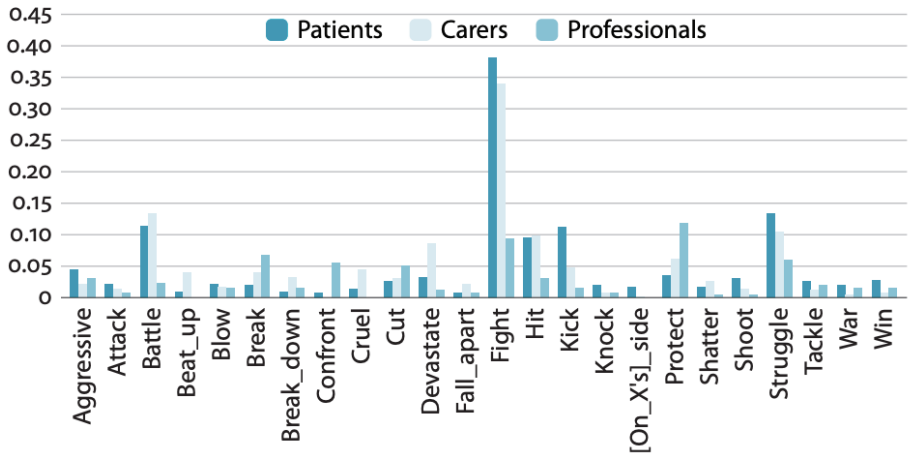
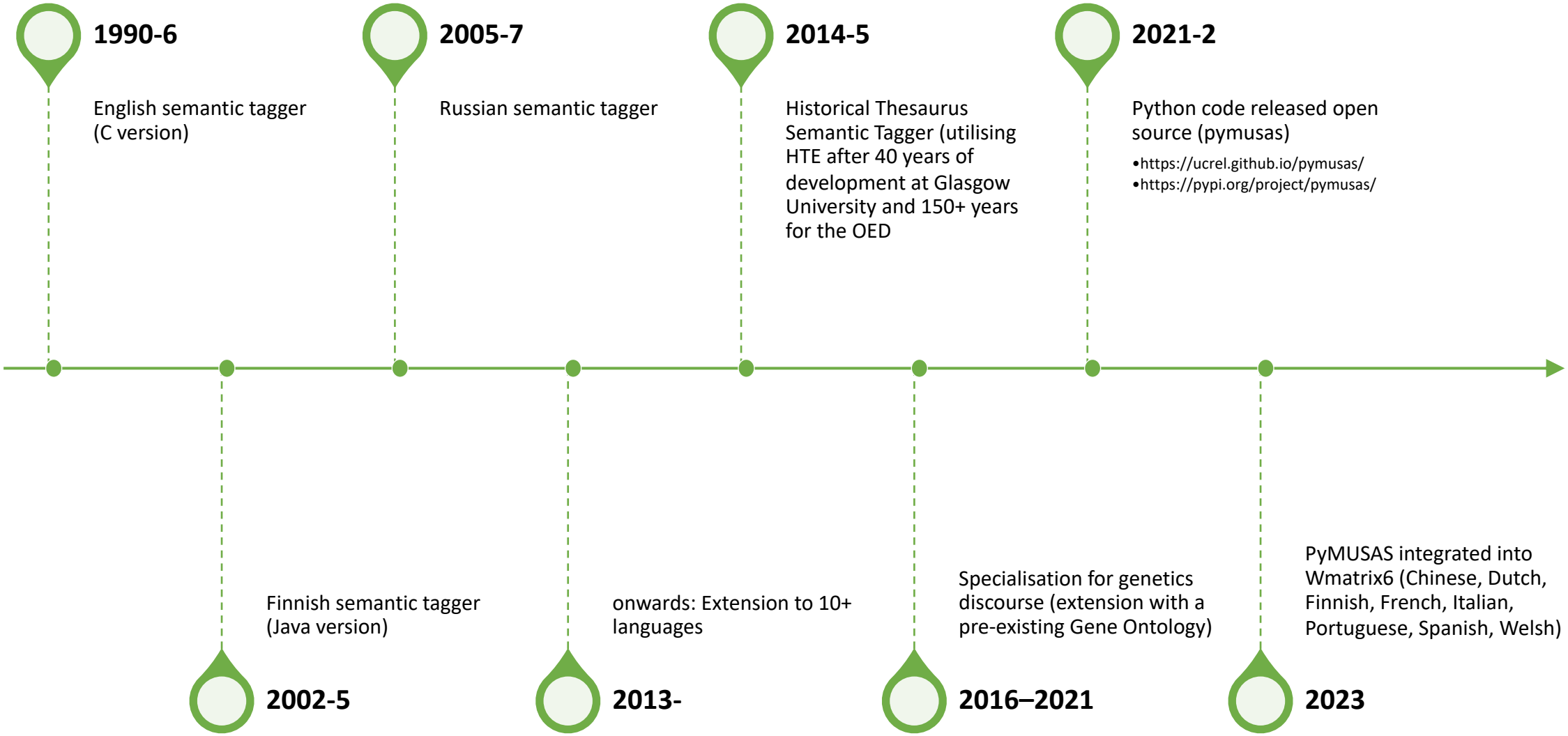


Figure 3. Relative use of most frequent Violence metaphors by each stakeholder group (per 1,000 tokens): Online forum posts


1990-6
English semantic tagger (C version)
2002-5
Finnish semantic tagger (Java version)
2005-7
Russian semantic tagger
2013-
onwards: Extension to 10+ languages
2014-5
Historical Thesaurus Semantic Tagger (utilising HTE after 40 years of development at Glasgow University and 150+ years for the OED
2016–2021
Specialisation for genetics discourse (extension with a pre-existing Gene Ontology)
2021-2
Python code released open source (pymusas)
•https://ucrel.github.io/pymusas/
•https://pypi.org/project/pymusas/
2023
PyMUSAS integrated into Wmatrix6 (Chinese, Dutch, Finnish, French, Italian, Portuguese, Spanish, Welsh)

Recipe for creating a tagger in a new language

Lancaster University

1. re-evaluate USAS semantic tagset for new language context
2. find freely available (open source if possible) POS tagger & lemmatiser
3. integrate these into USAS Multilingual software framework (PyMUSAS)
    a. consider whether other new components are needed e.g. tokeniser or compound tool
4. develop single-word semantic lexicon and MWE dictionary
    a. bilingual dictionary
    b. parallel aligned corpus (Moses / Giza)
    c. machine translation / translation memory
    d. crowdsourcing by non-experts
    e. named entity recognition and gazetteers
    f. vector-based approaches
    g. multi-task & deep learning
    h. manual checking and editing by experts
5. extend disambiguation routines
6. release lexicons with CC-BY-NC-SA licence
7. release software as REST API and/or open-source licence

PyMUSAS
https://pypi.org/project/pymusas/

- Open source – Apache License Version 2.0
- Open resources – Creative Commons licence version 4
- Rule based tagger
- Identify and tag Multi Word Expressions (MWE)
- Supports multiple languages through downloadable spaCy pipelines
- Supports Indonesian and Welsh via other POS taggers (TreeTagger for Indonesian and CyTag for Welsh)

Language (BCP 47 language code)	MWE Support	Size
Mandarin Chinese (cmn)	✔	1.28MB
Welsh (cy)	✔	1.09MB
Spanish, Castilian (es)	✔	0.20MB
French (fr)	×	0.08MB
Indonesian (id)	×	0.24MB
Italian (it)	✔	0.50MB
Dutch, Flemish (nl)	×	0.15MB
Portuguese (pt)	✔	0.27MB

Further reading …

- Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.
  - http://www.lancaster.ac.uk/staff/rayson/publications/usas_lrec04ws.pdf

- Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A. and Rayson, P. (2015). Development of the multilingual semantic annotation system. In proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), Denver, Colorado, United States, pp. 1268-1274.
  - http://aclweb.org/anthology/N/N15/N15-1137.pdf

Further reading …

- Piao, Scott Songlin; Rayson, Paul; Archer, Dawn; Bianchi, Francesca; Dayrell Gomes Da Costa, Maria Carmen; El-Haj, Mahmoud; Jiménez, Ricardo-María; Knight, Dawn; Křen, Michal; Lofberg, Laura; Nawab, Rao Muhammad Adeel ; Shafi, Jawad; Teh, Phoey Lee; Mudraya, Olga / Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. 2016. In proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portorož, Slovenia, pp. 2614-2619.
    - http://www.lrec-conf.org/proceedings/lrec2016/pdf/257_Paper.pdf
- El-Haj, M., Rayson, P., Piao, S., & Wattam, S. (2017). Creating and validating multilingual semantic representations for six languages: expert versus non-expert crowds. In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. (pp. 61-71). Association for Computational Linguistics.
    - http://aclweb.org/anthology/W17-1908

Summary so far

- Semantic ambiguity
- The task of Word Sense Disambiguation
  - Definitions
  - Methods
  - Evaluation
  - Resources e.g. WordNet
  - Semantic tagging via USAS
  - Applications

Any final questions?

Task 1: Automatic annotation of sample texts with the USAS online tagger (5 minutes)

Lancaster University

- *Aim: to familiarize yourself with the USAS tagset*
- Copy a sample text e.g. from a news article downloaded from today's BBC News (https://www.bbc.co.uk/news)
- Copy and paste the text into the online USAS tagger (https://ucrel-api.lancaster.ac.uk/usas/tagger.html) in order to semantically tag the text
- View the tagged output and the tag descriptions on the following pages and see if you can understand why certain words and MWEs have been tagged with these USAS labels, and can you spot places where the tags are incorrect?
  - https://ucrel.lancs.ac.uk/usas/
  - https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

Task 2 Install and run the PyMUSAS semantic tagger (10 minutes)

Lancaster University

- *Aim: to create Python code to semantically tag your texts*
- Either, follow the code templates in the how-to guides (https://ucrel.github.io/pymusas/usage/how_to/tag_text) to create your own Python code to semantically tag your BBC News text (pymusas is pip installable: https://pypi.org/project/pymusas/)
- Or, use the existing Python Notebook and run this using Google Colab or on the SCC lab machines locally to semantically tag your texts and compare them:
  - https://github.com/UCREL/pymusas_notebook/blob/main/PyMUSAS.ipynb
- (optional) Pick a new text in a language that you don't understand e.g. Spanish and use the semantic tagger for that language to see if you can understand the main key topics in the text without using Google Translate