The Multilingual Semantic Annotation System

– also a client GUI and MLCT corpus tool

Scott Piao

UCREL &
School of Computing and Communications
Lancaster University
Lancaster
UK

Email: s.piao@lancaster.ac.uk

Outline of My Talk

- Introduction to the development of UCREL multilingual semantic tagger.
- Main multilingual lexical resources of the semantic tagger.
- Accessing and processing corpus with the semantic tagger using a Graphical Interface (GUI) tool.
- Quick manipulation of the semantically tagged corpus data using the MLCT corpus tool.

Brief History of UCREL Semantic Tagger

- UCREL Semantic tagger (USAS) has been developed at UCREL, Lancaster University over the past two decades (Rayson et al., 2004).
- The semantic tagger has been expanded to annotate English text with a fine-grained semantic categories using a large English thesaurus, leading to the HTST tagger (Samuels Project).
- Initially developed for English, the semantic tagger has been ported for other languages through projects and in-house work, and a Java version was developed for easily handling multilingual data.
- So far, the USAS semantic lexicons that provide knowledge base for the tagger cover 14 languages (including English).
- Based on the lexicons, semantic tagger software have been developed for eight non-English languages.
- Six of them can be accessed via a GUI tool (to be introduced later).
- For further details about USAS, see website http://ucrel.lancs.ac.uk/usas/.

Lancaster University

USAS Semantic Annotation Tagset

--- *22 Major categories and 232 sub-categories*

(http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf)

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			

Course-grained but Generic Semantic Classification

- Based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981), the USAS tagset provides a coarsely-grained lexical semantic classification scheme.
- It is a generic scheme, not constrained to specific domain/s.
- Can be used to analyse high level abstract semantic structures of text, such as key topics of documents.
- Provide extra codes to denote information such as positive/negative, gender etc.
  - Example of tags:
    - ***E4.1+*** and ***E4.1-*** denotes ***happiness*** and ***sadness***;
    - ***S4f*** and ***S4m*** indicate ***female*** and ***male relatives***;
    - Etc.

Main USAS Lexical Resources

- Single word lexicon

  bank NN1 I1/H1 I1.1/I2.1c W3/M4 A9+/H1 O2 M6

- Multi-word expression (MWE) lexicon, including templates.

  giv*_* {R*/Np/PP*} away_* A9- A10+ S4

- For further details, see
  - Rayson, Paul, Dawn Archer, Scott Piao, Tony McEnery (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labeling for NLP tasks, LREC 2004, Lisbon, Portugal, pp. 7-12.
  - Archer, Dawn, Andrew Wilson, Paul Rayson (2002). Introduction to the USAS Category System. URL: http://ucrel.lancs.ac.uk/usas/usas_guide.pdf

Sample of Single Word Lexicon

Manchester	NP1	Z2 Z3
Mancunian	JJ	Z2 Z2/Q3
Mancunian	NN1	Z2/S2mf Z2/Q3
Mandarin-speaking	JJ	Z2/Q3
Mandela	NP1	Z1mf
Mandella	NP1	Z1mf
Manderville	NP1	Z2
Mandeville	NP1	Z2
Mandy	NP1	Z1f
…		
man-to-man	JJ	S5- S1.2.1+ A5.2+ A5.4+
manacles	NN2	O2
manage	VV0	S7.1+ A1.1.1 X9.2+
manageable	JJ	A12+
managed	JJ	S7.1+ A1.1.1 X9.2+
management	NN	S7.1+
management-style	JJ	S7.1+
manager	NN1	S7.1+/S2mf K1/S7.1+/S2mf K5/S7.1+/S2mf
manageress	NN1	S7.1+/S2.1f
manageress	VV0	S7.1+
managerial	JJ	S7.1+

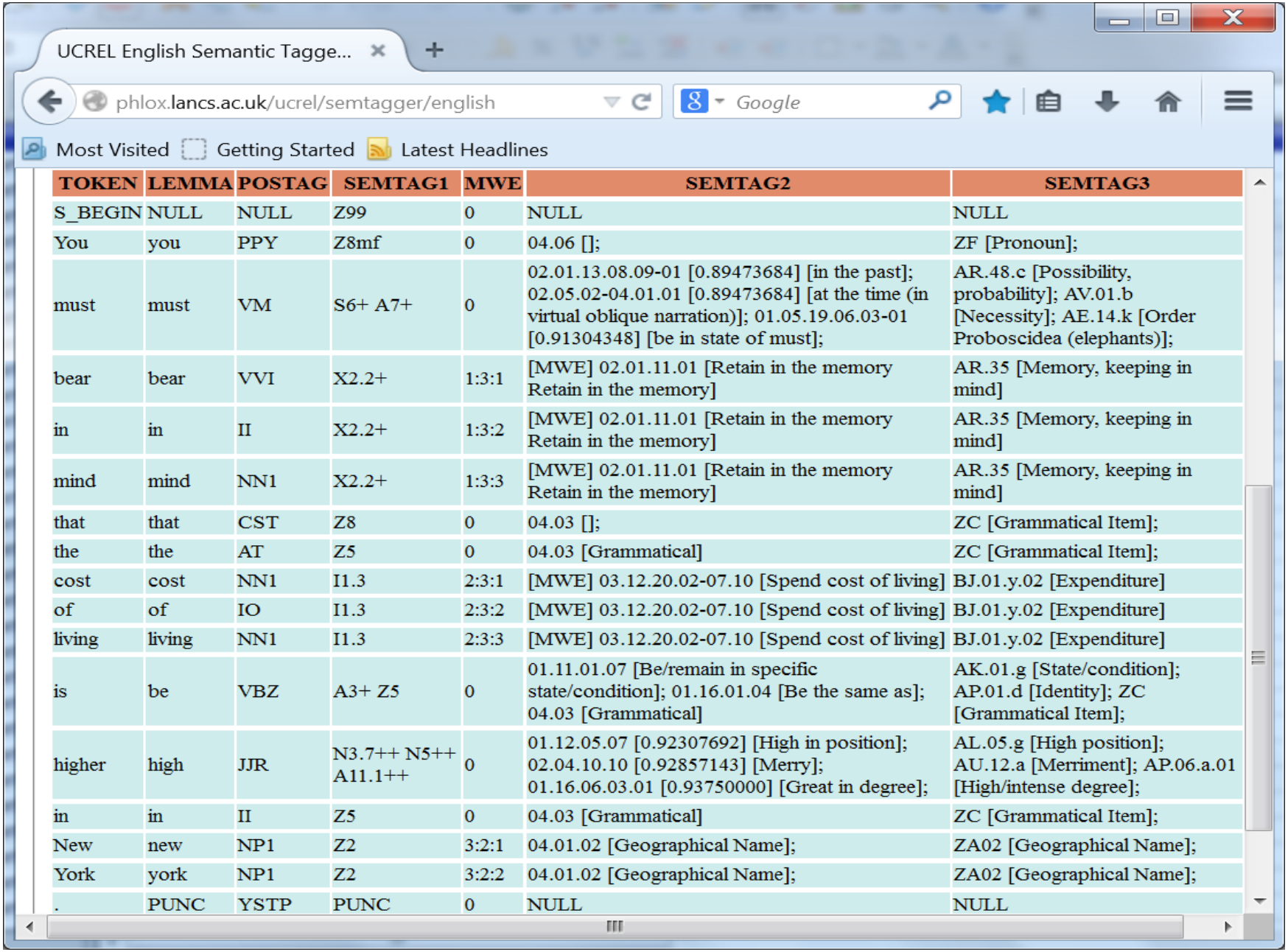
Sample of Multi-Word Expression (MWE) Lexicon

at_II the_AT very_RG least_DAT	A13.7
at_II the_AT very_RG minimum_*	A13.7
at_II the_AT {J*/UH} offset_NN1	T2+
at_II the_AT {J*} forefront_NN1 of_IO	A11.1+
at_II the_AT {J*} mercy_NN1 of_IO	S7.1-
at_II the_AT {J*} moment_NN1	T1.1.2
at_II the_AT {J*} outset_NN1	T2+

HTST Tagger, An Extension of English Semantic Tagger

- In the Samuels Project, the USAS was extended to tag English text in a highly fine-grained semantic classification scheme based on a English Historical Thesaurus, named HTST.
- For details of the thesaurus, see websites
  - http://historicalthesaurus.arts.gla.ac.uk/
  - http://public.oed.com/historical-thesaurus-of-the-oed/
- HTST employs 225,131 semantic categories, which are mapped to about 4,000 broader semantic categories for practical applications.

HTST Sample Output



TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	NULL	NULL
You	you	PPY	Z8mf	0	04.06 [];	ZF [Pronoun];
must	must	VM	S6+ A7+	0	02.01.13.08.09-01 [0.89473684] [in the past]; 02.05.02-04.01.01 [0.89473684] [at the time (in virtual oblique narration)]; 01.05.19.06.03-01 [0.91304348] [be in state of must];	AR.48.c [Possibility, probability]; AV.01.b [Necessity]; AE.14.k [Order Proboscidea (elephants)];
bear	bear	VVI	X2.2+	1:3:1	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
in	in	II	X2.2+	1:3:2	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
mind	mind	NN1	X2.2+	1:3:3	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
that	that	CST	Z8	0	04.03 [];	ZC [Grammatical Item];
the	the	AT	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
cost	cost	NN1	I1.3	2:3:1	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
of	of	IO	I1.3	2:3:2	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
living	living	NN1	I1.3	2:3:3	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
is	be	VBZ	A3+ Z5	0	01.11.01.07 [Be/remain in specific state/condition]; 01.16.01.04 [Be the same as]; 04.03 [Grammatical]	AK.01.g [State/condition]; AP.01.d [Identity]; ZC [Grammatical Item];
higher	high	JJR	N3.7++ N5++ A11.1++	0	01.12.05.07 [0.92307692] [High in position]; 02.04.10.10 [0.92857143] [Merry]; 01.16.06.03.01 [0.93750000] [Great in degree];	AL.05.g [High position]; AU.12.a [Merriment]; AP.06.a.01 [High/intense degree];
in	in	II	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
New	new	NP1	Z2	3:2:1	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
York	york	NP1	Z2	3:2:2	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
.	PUNC	YSTP	PUNC	0	NULL	NULL

HTST is beyond scope of this talk. If interested, see paper:

Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron, Paul Rayson (2015). *Metaphor, Popular Science and Semantic Tagging: Distant reading with the Historical Thesaurus of English*. Digital Scholarship in the Humanities, Oxford University Press, UK.

Multilingality of Semantic Tagging

- Multilinguality is an important aspect of corpus linguistics and natural language processing, and so to semantic analysis.
- Would be nice to create an ecosystem for multilingual semantic tagging and analysis under the same semantic classification framework.
- The USAS multilingual semantic tagger can help to build such a system.
- After fourteen years' of progress, the current USAS lexicons cover Italian, Portuguese, Chinese, Spanish, Arabic, Russian, French, Czech, Finnish, Dutch, Malaysian, Welsh, Urdu besides English. Available at *https://github.com/UCREL/Multilingual-USAS/*
- Based on the lexicons, semantic tagging software have been developed for Italian, Portuguese, Chinese, Spanish, French, Russian, Finnish, Dutch, and a prototype for Welsh.
- Semantic taggers are in different stages of development for different languages, hence they provide various lexical coverages and accuracies.

Multilingual Semantic Lexicon Construction

- A critical part of multilingual semantic tagger development is to construct semantic lexicons for the languages.
- Various approaches have been employed so far:
  - Automatically translating the core English semantic lexicon using bilingual dictionaries and other publicly available lexicons.
  - Using crowd-sourcing methods to clean and expand the automatically generated lexicons.
  - Where possible, using bilingual parallel corpora to align words across languages, thereby allowing the application of above two methods.
  - Using machine translation tools to directly translate existing lexicons into new languages.
  - Manually cleaning and curating the lexicons whenever possible.
  - *There should be more good methods … that we can try.*

Statistics of Semantic Lexicons for 13 Languages

Language	Single Word Entries	MWE Entries	Tagger developed?
Arabic	31,154	0	N
Chinese	64,541	19,048	Y
Czech	28,161	0	N
Dutch	4,220	0	Y
Finnish	46,225	4,422	*Y
French	2,754	0	Y
Italian	13,098	5,622	Y
Malay	64,863	0	N
Portuguese	13,499	1,781	Y
Russian	17,443	713	*Y
Spanish	3,665	0	Y
Urdu	1,765	235	N
Welsh	174,000	0	N

Lexical Coverage Evaluation on Running Text

No	Language	Blogs (million words)	News (million words)	Average	Tagger or Lexicon only?
1	Finnish	95.98	95.89	95.93	Tagger
2	Italian	91.14	89.34	90.24	Tagger
3	Czech	87.95	86.05	86.99	Tagger
4	Russian	84.93	86.66	85.79	Tagger
5	Chinese	82.98	79.36	81.17	Tagger
6	Portuguese (EU)	76.79	77.47	77.13	Tagger
7	Portuguese (BR)	76.11	77.75	76.93	Tagger
8	Dutch	61.55	59.87	60.71	Tagger
9	Spanish (EU)	57.81	55.73	56.77	Tagger
10	Spanish (SA)	57.20	56.11	56.65	Tagger
11	Arabic	86.43	91.33	88.88	Lexicon only
12	Urdu	86.26	84.21	85.24	Lexicon only
13	Malay	53.83	54.91	54.37	Lexicon only

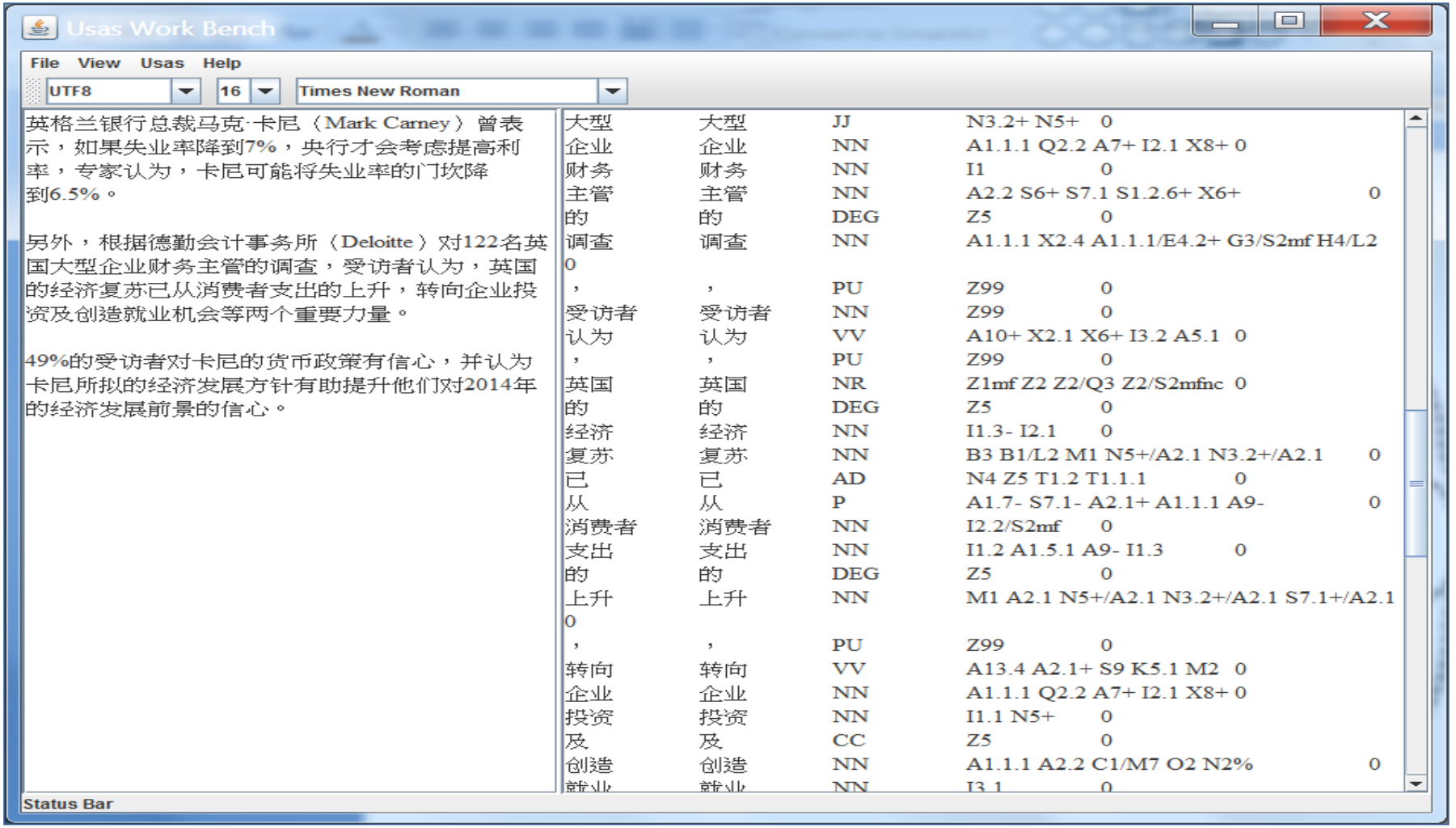
Current and Future Research

- Welsh – current focus
  - UCREL is involved in the CorCenCC Project (The National Corpus of Contemporary Welsh), in which UCREL team is developing a Welsh semantic tagger, in collaboration with Welsh Universities.
  - An initial Welsh semantic lexicon has been constructed, currently containing over 174,000 Welsh words.
  - In an initial evaluation, our current Welsh wordlist has reached over 97% lexical coverage – the wordlist includes raw Welsh words extracted from corpus resources
  - Work is under way to classify more Welsh words into USAS semantic categories.
  - Initial version of Welsh semantic tagger is under development.
- Works under way or plan:
  - Swedish, Norwegian, possibly Greek later.

Accessing the Multilingual Semantic Taggers

- The semantic taggers are built as web services.
- Three ways to access the tools:
  - Webpage interfaces for a simple trial, available at URL: http://ucrel.lancs.ac.uk/usas/
  - For processing larger corpus data in multiple files, a GUI tool is available for six languages, as shown in next slide.
  - Tool developers can access the service using web service API (beyond scope of this talk).

Desktop Graphical User Interface (GUI)



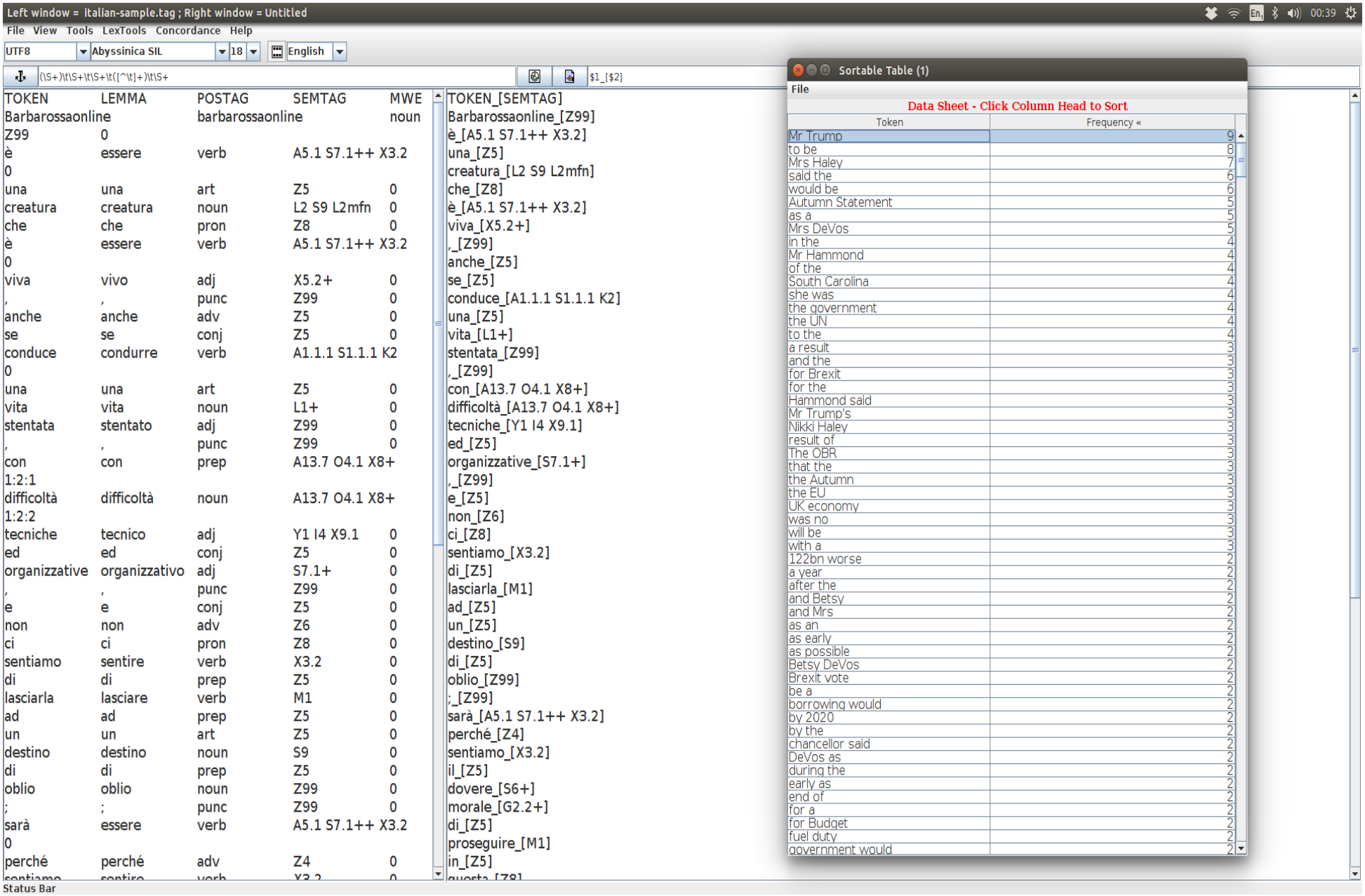
How to get it and run it?

- Make sure your PC has Java Runtime Environment (JRE) installed – download from url:
  http://www.oracle.com/technetwork/java/javase/downloads/index-jsp-138363.html.
- Download the file “sem-tagger-gui.tar.gz” from url:
  http://ucrel.lancs.ac.uk/usas/gui/
- Unzip it somewhere on your PC.
- Go into the tool folder, click on file “run_semtagger-gui.bat” in Windows, or in Linux/Unix type
  >run_semtagger-gui.sh [RETURN]
- The interface starts up.

MLCT (Multi-Lingual Corpus Toolkit)

- After tagging corpus using the semantic tagger GUI, you often want to process the data further for research.
- A light-weight corpus tool, MLCT, can be used together with the semantic tagger GUI.
- It provides numerous functions for manipulating corpus data, including
  - Search, replace and re-format text (using regular expressions)
  - Extract word frequency list, n-grams and collocations
  - Extract concordance lists
  - Many more useful small useful functionalities.
- Not everything is fully automatic, needs users' involvement, like writing regular expression languages, but you can do creative and complex work with your own data.
- For processing moderate-sized corpus data, not for a large-scale corpus processing.
- Reference paper:
  - Piao, Scott, Andrew Wilson and Tony McEnery (2002). A Multilingual Corpus Toolkit, AAACL-2002, Indianapolis, Indiana, USA.

MLCT in Work


Left window = italian-sample.tag ; Right window = Untitled
File View Tools LexTools Concordance Help
UTF8
Abyssinica SIL
18
English
(\S+)\t\S+\t\S+\t([^\t]+)\t\S+
$1_[$2]
TOKEN LEMMA POSTAG SEMTAG MWE
Barbarossaonline barbarossaonline noun
Z99 0
è essere verb A5.1 S7.1++ X3.2
0
una una art Z5 0
creatura creatura noun L2 S9 L2mfn 0
che che pron Z8 0
è essere verb A5.1 S7.1++ X3.2
0
viva vivo adj X5.2+ 0
, , punc Z99 0
anche anche adv Z5 0
se se conj Z5 0
conduce condurre verb A1.1.1 S1.1.1 K2
0
una una art Z5 0
vita vita noun L1+ 0
stentata stentato adj Z99 0
, , punc Z99 0
con con prep A13.7 O4.1 X8+
1:2:1
difficoltà difficoltà noun A13.7 O4.1 X8+
1:2:2
tecniche tecnico adj Y1 I4 X9.1 0
ed ed conj Z5 0
organizzative organizzativo adj S7.1+ 0
, , punc Z99 0
e e conj Z5 0
non non adv Z6 0
ci ci pron Z8 0
sentiamo sentire verb X3.2 0
di di prep Z5 0
lasciarla lasciare verb M1 0
ad ad prep Z5 0
un un art Z5 0
destino destino noun S9 0
di di prep Z5 0
oblio oblio noun Z99 0
; ; punc Z99 0
sarà essere verb A5.1 S7.1++ X3.2
0
perché perché adv Z4 0
TOKEN_[SEMTAG]
Barbarossaonline_[Z99]
è_[A5.1 S7.1++ X3.2]
una_[Z5]
creatura_[L2 S9 L2mfn]
che_[Z8]
è_[A5.1 S7.1++ X3.2]
viva_[X5.2+]
,_[Z99]
anche_[Z5]
se_[Z5]
conduce_[A1.1.1 S1.1.1 K2]
una_[Z5]
vita_[L1+]
stentata_[Z99]
,_[Z99]
con_[A13.7 O4.1 X8+]
difficoltà_[A13.7 O4.1 X8+]
tecniche_[Y1 I4 X9.1]
ed_[Z5]
organizzative_[S7.1+]
,_[Z99]
e_[Z5]
non_[Z6]
ci_[Z8]
sentiamo_[X3.2]
di_[Z5]
lasciarla_[M1]
ad_[Z5]
un_[Z5]
destino_[S9]
di_[Z5]
oblio_[Z99]
;_[Z99]
sarà_[A5.1 S7.1++ X3.2]
perché_[Z4]
sentiamo_[X3.2]
il_[Z5]
dovere_[S6+]
morale_[G2.2+]
di_[Z5]
proseguire_[M1]
in_[Z5]
Sortable Table (1)
File
Data Sheet - Click Column Head to Sort
Token
Frequency «
Mr Trump 9
to be 8
Mrs Haley 7
said the 6
would be 6
Autumn Statement 5
as a 5
Mrs DeVos 5
in the 4
Mr Hammond 4
of the 4
South Carolina 4
she was 4
the government 4
the UN 4
to the 4
a result 3
and the 3
for Brexit 3
for the 3
Hammond said 3
Mr Trump's 3
Nikki Haley 3
result of 3
The OBR 3
that the 3
the Autumn 3
the EU 3
UK economy 3
was no 3
will be 3
with a 3
122bn worse 2
a year 2
after the 2
and Betsy 2
and Mrs 2
as an 2
as early 2
as possible 2
Betsy DeVos 2
Brexit vote 2
be a 2
borrowing would 2
by 2020 2
by the 2
chancellor said 2
DeVos as 2
during the 2
early as 2
end of 2
for a 2
for Budget 2
fuel duty 2
government would 2
Status Bar

How to get it and run it?

- Again, make sure your PC has Java JRE installed.
- Download file “mlct_public.zip” from url: *https://sites.google.com/site/scottpiaosite/software/mlct*
- Unzip it somewhere on your PC.
- Go into the tool folder, click on file “run_mlct_public.bat” in Windows, or in Linux/Unix type

  >java -Xmx500m -jar mlct_public.jar [RETURN]

- The MLCT interface starts up.

Summary

- USAS system provides a good corpus tool for multilingual research.
- It will cover more languages and provide better performance.
- The USAS GUI access tool and the MLCT can be combined to help you to work with moderate-sized multilingual corpus data.

Related Papers

- Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron, Paul Rayson (2015). Metaphor, Popular Science and Semantic Tagging: Distant reading with the Historical Thesaurus of English. Digital Scholarship in the Humanities, Oxford University Press, UK.
- McArthur, Tom (1981). Longman Lexicon of Contemporary English. Longman London Quirk R., Greenbaum S., Leech G., Svartvik J. (1985). A Comprehensive Grammar of the English Language. Longman: London.
- Rayson, Paul, Dawn Archer, Scott Piao and Tony McEnery (2004). The UCREL semantic analysis system. In Proceedings of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks, pp. 7-12. Lisbon, Portugal.
- Piao, Scott, Paul Rayson, Dawn Archer, Francesca Bianchi, Carmen Dayrell, Mahmoud El-Haj, Ricardo-María Jiménez, Dawn Knight, Michal Křen, Laura Löfberg, Rao Muhammad Adeel Nawab, Jawad Shafi, Phoey Lee Teh, Olga Mudraya (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. Accepted by The 10th Edition of the Language Resources and Evaluation Conference (LREC2016). To be held during 23-28 May 2016 in Portorož, Slovenia.
- Piao, Scott, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio and Paul Rayson (2015). Development of the Multilingual Semantic Annotation System. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), Denver, Colorado, USA.
- Piao, Scott, Andrew Wilson and Tony McEnery (2002). A Multilingual Corpus Toolkit, AAACL-2002, Indianapolis, Indiana, USA