Quantifying Lexical Usage and Subjectivity in the CLAEVIPS Project

Diana McCarthy‡
in collaboration with
Kate Wild‡, Andrew Church† and Jacquelin Burgess†

‡ Lexical Computing Ltd and † NEA

5th November 2011

Outline

1. Related Work
2. Scope
3. Methodology
4. Results
5. Replicability, Future Ideas, Discussion

Background

- Study Commissioned by National Ecosystem Assessment (NEA)
- how are ecosystems and the natural environment discussed in the public sphere?
- what are the keys terms?
- are these used in subjective or emotive texts?
- are subjective uses in positive or negative contexts?
- are there differences in usages in different genres (e.g. newspapers, blogs, NGOs, governmental organisations, academic texts)

Macmillan Blog (July 4th) Michael Rundell

http://www.macmillandictionaryblog.com/you-say-global-warming-i-say-climate-change

But in the last two years, things have changed dramatically: climate change overtook global warming in 2010, and the data for 2011 year shows that it is now four times more frequent

. . .

Luddites, tree huggers, or 'beardy environmentalists' vs deniers

. . .

'carbon trading' compared to medieval indulgences (Martin Palmer)

Existing Research on Environmental Language:

Critical discourse analysis (CDA) political analysis, social context, small datasets and qualitative:

- [Goatly, 1996, Schleppegrell, 1997] agency (passive and nominalised forms to avoid ascribing agency)
- [Kuha, 2009] *global warming* in US newspapers - climate change as certain or not
- [Carvalho and Burgess, 2005] political orientations of broadsheet newspapers 1985 to 2003, different framing of *climate change*
- [Alexander, 2009] analysis of small texts on environment, no attempt to establish a norm

Environmental Language and Corpus Linguistics

- [Nerlich and Koteyko, 2009] compounds with *carbon* in blogs and newspapers
- [Grundmann and Krishnamurthy, 2010] compare references to *climate change* and *global warming* in English, French and German

Environmental Language and Corpus Linguistics

- [Nerlich and Koteyko, 2009] compounds with *carbon* in blogs and newspapers
- [Grundmann and Krishnamurthy, 2010] compare references to *climate change* and *global warming* in English, French and German

[Baker, 2011] (Refuges and asylum seekers):

- Combination of CDA and Corpus linguistics
- ensures data to support analysis and reduce researcher bias

CLAEVIPS: A Corpus Linguistics Analysis of Ecosytems Vocabulary in the Public Sphere

- large scale corpus analysis
- broad range of vocabulary
- modest budget (duration 3 months, part-time)
- look for collocates and patterns, dominant discourses and then examine underlying texts (CDA)
- reference corpus plus three custom-built specialised corpora
- inter-analyst reliability

CLAEVIPS: Resources

- 136 words and phrases concerning the ecosystem (45 from NEA)
- Sketch Engine
- WebBootCat
- 4 corpora:
  - UKWaC [Ferraresi et al., 2008]
  - 3 specialised corpora

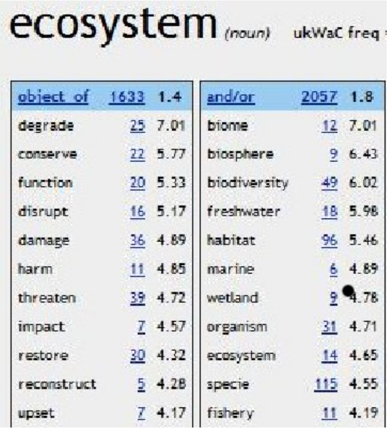
CLAEVIPS: Corpora

- ukWaC [Ferraresi et al., 2008] 1.5 billion word corpus from internet domains ending '.uk'
- three specialised corpora harvest from the web. Web pages contain at least:
  - three types from a set of seed words, and
  - at least three occurrences of a subset of whitelist words
- the three corpora (each approx 1.5 million words)
  1. academic (ac.uk)
  2. government (.gov.uk)
  3. public (news, NGO, blogs)

CLAEVIPS: Methodology

- examine salient collocates using ‘word sketch’ (words), and contrasted in the 3 specialised corpora
- examine 100 random citations from UKWaC:
  - subjective/objective
  - positive / negative / neutral
  - other ...
- (phrases) find collocates in above citations and contrast to 50 random from specialised corpora
- some words selected for additional examination using thesaurus and sketch diff

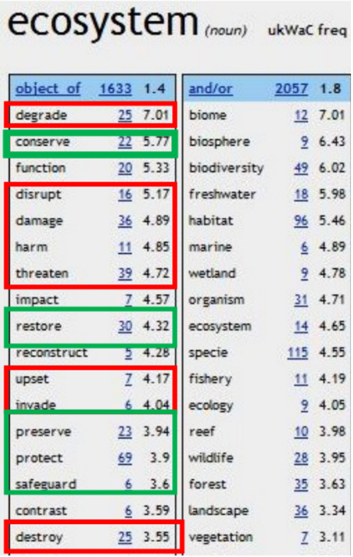
Word Sketch for *ecosystem* in ukWaC



ecosystem (noun) ukWaC freq

object_of	1633	1.4	and/or	2057	1.8
degrade	25	7.01	biome	12	7.01
conserve	22	5.77	biosphere	9	6.43
function	20	5.33	biodiversity	49	6.02
disrupt	16	5.17	freshwater	18	5.98
damage	36	4.89	habitat	96	5.46
harm	11	4.85	marine	6	4.89
threaten	39	4.72	wetland	9	4.78
impact	7	4.57	organism	31	4.71
restore	30	4.32	ecosystem	14	4.65
reconstruct	5	4.28	specie	115	4.55
upset	7	4.17	fishery	11	4.19

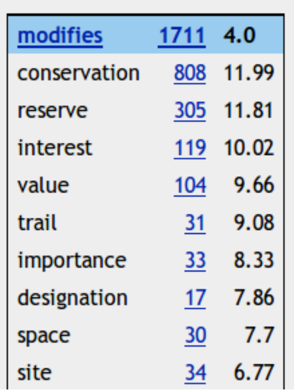
Word Sketch for *ecosystem* in ukWaC



ecosystem (noun) ukWaC freq

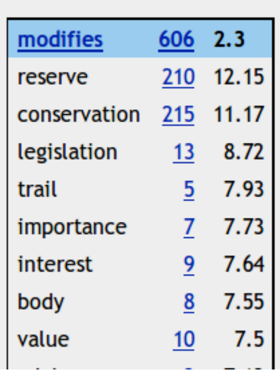
object of	1633	1.4	and/or	2057	1.8
degrade	25	7.01	biome	12	7.01
conserve	22	5.77	biosphere	9	6.43
function	20	5.33	biodiversity	49	6.02
disrupt	16	5.17	freshwater	18	5.98
damage	36	4.89	habitat	96	5.46
harm	11	4.85	marine	6	4.89
threaten	39	4.72	wetland	9	4.78
impact	7	4.57	organism	31	4.71
restore	30	4.32	ecosystem	14	4.65
reconstruct	5	4.28	specie	115	4.55
upset	7	4.17	fishery	11	4.19
invade	6	4.04	ecology	9	4.05
preserve	23	3.94	reef	10	3.98
protect	69	3.9	wildlife	28	3.95
safeguard	6	3.6	forest	35	3.63
contrast	6	3.59	landscape	36	3.34
destroy	25	3.55	vegetation	7	3.11

Word Sketch for *nature* in Specialised Corpora



modifies	1711	4.0
conservation	808	11.99
reserve	305	11.81
interest	119	10.02
value	104	9.66
trail	31	9.08
importance	33	8.33
designation	17	7.86
space	30	7.7
site	34	6.77

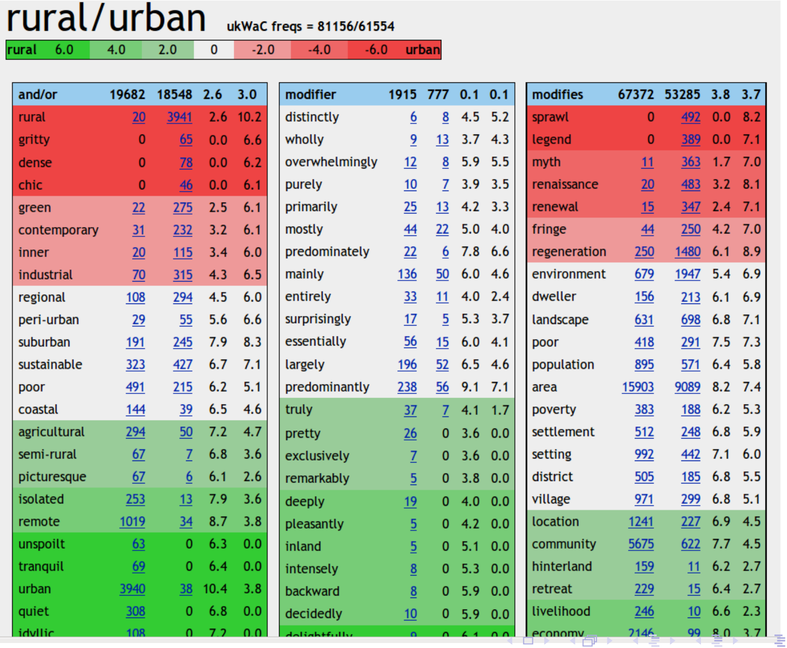
Government



modifies	606	2.3
reserve	210	12.15
conservation	215	11.17
legislation	13	8.72
trail	5	7.93
importance	7	7.73
interest	9	7.64
body	8	7.55
value	10	7.5

Public

Sketch Difference: Rural vs Urban



rural/urban ukWaC freqs = 81156/61554

rural	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	urban

and/or	19682	18548	2.6	3.0
rural	20	3941	2.6	10.2
gritty	0	65	0.0	6.6
dense	0	78	0.0	6.2
chic	0	46	0.0	6.1
green	22	275	2.5	6.1
contemporary	31	232	3.2	6.1
inner	20	115	3.4	6.0
industrial	70	315	4.3	6.5
regional	108	294	4.5	6.0
peri-urban	29	55	5.6	6.6
suburban	191	245	7.9	8.3
sustainable	323	427	6.7	7.1
poor	491	215	6.2	5.1
coastal	144	39	6.5	4.6
agricultural	294	50	7.2	4.7
semi-rural	67	7	6.8	3.6
picturesque	67	6	6.1	2.6
isolated	253	13	7.9	3.6
remote	1019	34	8.7	3.8
unspoilt	63	0	6.3	0.0
tranquil	69	0	6.4	0.0
urban	3940	38	10.4	3.8
quiet	308	0	6.8	0.0
idyllic	108	0	7.2	0.0

modifier	1915	777	0.1	0.1
distinctly	6	8	4.5	5.2
wholly	9	13	3.7	4.3
overwhelmingly	12	8	5.9	5.5
purely	10	7	3.9	3.5
primarily	25	13	4.2	3.3
mostly	44	22	5.0	4.0
predominately	22	6	7.8	6.6
mainly	136	50	6.0	4.6
entirely	33	11	4.0	2.4
surprisingly	17	5	5.3	3.7
essentially	56	15	6.0	4.1
largely	196	52	6.5	4.6
predominantly	238	56	9.1	7.1
truly	37	7	4.1	1.7
pretty	26	0	3.6	0.0
exclusively	7	0	3.6	0.0
remarkably	5	0	3.8	0.0
deeply	19	0	4.0	0.0
pleasantly	5	0	4.2	0.0
inland	5	0	5.1	0.0
intensely	8	0	5.3	0.0
backward	8	0	5.9	0.0
decidedly	10	0	5.9	0.0
delightfully	9	0	6.1	0.0

modifies	67372	53285	3.8	3.7
sprawl	0	492	0.0	8.2
legend	0	389	0.0	7.1
myth	11	363	1.7	7.0
renaissance	20	483	3.2	8.1
renewal	15	347	2.4	7.1
fringe	44	250	4.2	7.0
regeneration	250	1480	6.1	8.9
environment	679	1947	5.4	6.9
dweller	156	213	6.1	6.9
landscape	631	698	6.8	7.1
poor	418	291	7.5	7.3
population	895	571	6.4	5.8
area	15903	9089	8.2	7.4
poverty	383	188	6.2	5.3
settlement	512	248	6.8	5.9
setting	992	442	7.1	6.0
district	505	185	6.8	5.5
village	971	299	6.8	5.1
location	1241	227	6.9	4.5
community	5675	622	7.7	4.5
hinterland	159	11	6.2	2.7
retreat	229	15	6.4	2.7
livelihood	246	10	6.6	2.3
economy	2146	99	8.0	3.7

CLAEVIPS: (some) Findings

- words not widely understood e.g. *biotype*, *natural capital*
- differences in specialised corpora e.g. public interest in rainforest and global warming
- promotional use of nature in advertising *'eco'*
- nature as a commodity (esp government corpus)
- in ukWaC and public corpus: evidence of scepticism regarding empty use of words *sustainable* and claims on climate change
- relationship between humans and nature
- fear of open spaces
- avoid reference to agency with words such as *pollute*, see also [Schleppegrell, 1997]

Replicability . . .

- different data, ideally if available
- annotators
- methodology
- other works:
  - [Marchi and Taylor, 2009] 2 researchers - journalists talk about selves and profession - same corpus and methodology; convergent, dissonant and complementary findings
  - [Baker, 2011] 5 researchers - foreign doctors in British Press - different methods, 4% findings made by all, 65% from one person. Overall feel similar

Ours small study: 3 months - part-time

- divided work
- methodology, worked out together after start point and few iterations before commencing
- same methodology, same data, 3 lemmas - same findings broadly

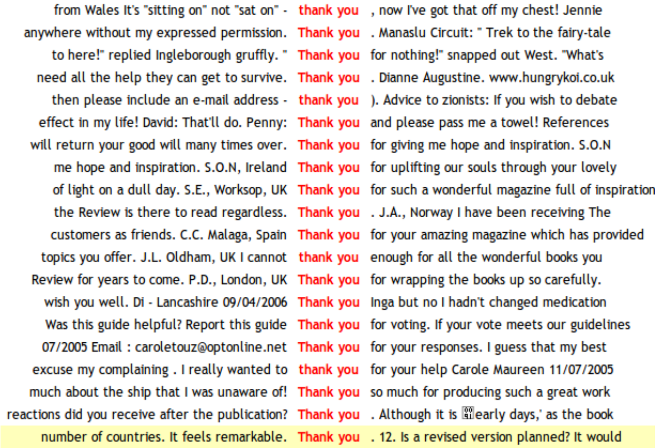
allotment

Feature	**Researcher 1**	**Researcher 2**
collocates UKWaC	neutral: *gardening, holder, plot, gardener* negative: *derelict, disused, unused, overgrown*	relating to gardens: *gardening, gardener, garden* ownership: *smallholding, rent* negative: *derelict, disused, overgrown*
positive/negative	neut or pos: uses and benefits	
freq diffs	freq in govt, rare acad	
other differences	-	acad: lack of allotments
other findings:	pos use of neg collocates	

Further ideas

- data - repeating process with WBC
- intra-annotator agreement, repeat the process 6 months later
- diff methodology, same data, same general findings?

Finally



```
         from Wales It's "sitting on" not "sat on" -  thank you  , now I've got that off my chest! Jennie
      anywhere without my expressed permission.  Thank you  . Manaslu Circuit: " Trek to the fairy-tale
          to here!" replied Ingleborough gruffly. "  Thank you  for nothing!" snapped out West. "What's
         need all the help they can get to survive.  Thank you  . Dianne Augustine. www.hungrykoi.co.uk
           then please include an e-mail address -  thank you  ). Advice to zionists: If you wish to debate
        effect in my life! David: That'll do. Penny:  Thank you  and please pass me a towel! References
      will return your good will many times over.  Thank you  for giving me hope and inspiration. S.O.N
          me hope and inspiration. S.O.N, Ireland  Thank you  for uplifting our souls through your lovely
        of light on a dull day. S.E., Worksop, UK  Thank you  for such a wonderful magazine full of inspiration
         the Review is there to read regardless.  Thank you  . J.A., Norway I have been receiving The
          customers as friends. C.C. Malaga, Spain  Thank you  for your amazing magazine which has provided
       topics you offer. J.L. Oldham, UK I cannot  thank you  enough for all the wonderful books you
       Review for years to come. P.D., London, UK  Thank you  for wrapping the books up so carefully.
          wish you well. Di - Lancashire 09/04/2006  Thank you  Inga but no I hadn't changed medication
         Was this guide helpful? Report this guide  Thank you  for voting. If your vote meets our guidelines
        07/2005 Email : caroletouz@optonline.net  Thank you  for your responses. I guess that my best
       excuse my complaining . I really wanted to  thank you  for your help Carole Maureen 11/07/2005
       much about the ship that I was unaware of!  Thank you  so much for producing such a great work
reactions did you receive after the publication?  Thank you  . Although it is □early days,' as the book
         number of countries. It feels remarkable.  Thank you  . 12. Is a revised version planned? It would
```

Alexander, R. J. (2009).
Framing Discourses on the Environment: A Critical Discourse Approach.
Routledge, New York.

Baker, P. (2011).
Discourse, news representations and corpus linguistics.
In *Plenary paper presented at Corpus Linguistics 2011*, Birmingham, UK.

Carvalho, A. and Burgess, J. (2005).
Cultural circuits of climate change in uk broadsheet newspapers,1985-2003.
Risk Analysis, 25(6):14571469.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008).
Introducing and evaluating ukwac, a very large web-derived corpus of english.
In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Goatly, A. (1996).
Green grammar and grammatical metaphor, or language and myth of power, or metaphors we die by.
Journal of Pragmatics, 25(4):537560.

Grundmann, R. and Krishnamurthy, R. (2010).
The discourse of climate change: A corpus-based approach.
Critical Approaches to Discourse Analysis across Disciplines, 4(2):125–146.

Kuha, M. (2009).
Uncertainty about causes and effects of global warming in u.s. news coverage before and after bali.
Language and Ecology, 4(2):¡http://www.ecoling.net/journal.html¿.

Marchi, A. and Taylor, C. (2009).
If on a winters night two researcher . . . a challenge to assumptions of soundness of interpretation.
Critical Approaches to Discourse Analysis across Disciplines, 3(1):1–20.

Nerlich, B. and Koteyko, N. (2009).
Compounds, creativity and complexity in climate change communication: The case of 'carbon indulgences'.
Global Environmental Change, 19:345353.

Schleppegrell, M. (1997).
Agency in environmental education.
Linguistics and Education, 9:49–67.