The potential role of data-driven content analysis

Andrew Salway

Uni Research, Bergen

Data-driven science

- Nothing new, cf. Kepler, Darwin
- Automatically detecting patterning in vast data sets has become a norm in various scientific fields, e.g. astronomy, genomics, neuroscience
- It doesn't make sense to think of "data-driven" and "hypothesis-driven" as mutually exclusive

Content Analysis

- Content analysis = counting instances of linguistic forms that have meaning with respect to a conceptual framework and non-textual phenomena
- Four main steps:
  1. Select non-textual phenomena to investigate
  2. Determine an appropriate conceptual framework
  3. Establish a mapping between concepts and linguistic forms that can be counted
  4. Identify significant statistical results in the frequency counts

Data-driven content analysis

- As textual material becomes more diverse and bigger then data-driven becomes more relevant:
  - researchers cannot assume that they know the material, i.e. they can make fewer assumptions in steps 1-3 (especially 3)
  - opportunity/need to challenge existing theory and conceptual frameworks and coding schemes
- A partial solution? Automatic data-driven techniques incorporated into "discovery tools" to:
  - provide manageable views of large text corpora
  - elucidate interesting aspects of the content
  - stimulate new hypotheses
  - challenge/confirm existing conceptual frameworks
  - inform the development of coding schemes

What is required of techniques for data-driven content analysis?

- They should elucidate interesting characteristics of the content
- They should be well understood and reliable
- They should not rely on prior linguistic resources such as lexicons and grammars, for both practical and methodological reasons:
  - the diversity of material means techniques should be portable across domains, text types and languages, without the cost of generating resources each time
  - prior linguistic resources introduce biases: better to minimise assumptions about the domain and the language used

Current techniques

- Techniques for unsupervised clustering and scaling mostly meet these requirements but are limited by treating texts as bags of words:
  - most meaning is lost
  - can can only compare text-level features
- Corpus linguistics has established techniques for exploring corpora in a data-driven manner – frequency lists, keyword lists, n-grams, collocations, concordances:
  - Useful for an overview of frequent content, and some information about word sequences and co-occurrences
  - However, still quite a shallow view of language, and these techniques generate a lot of data to inspect
- Language visualization can help to understand word co-occurrence, but it relies on text analysis to provide a manageable view

An example of data-driven content analysis

- The material was a corpus of blogs related to climate change:
  - about 3000 blogs, 1.4m blog posts, 400m words
  - focused on 330,000 sentences containing either "climate change" or "global warming"
- Interesting and challenging for content analysis:
  - climate change is a complex and contested issue
  - diverse sub-topics, perspectives and opinions
  - polarized (sceptics / acceptors)
  - framed in different ways, e.g. science, politics, national / local issues

Fløttum, K., Gjerstad, Ø., Gjesdal, A.M., Koteyko, N. and Salway, A. (2014). Representations of the future in English language blogs on climate change. *Global Environmental Change* 29, 213–222.

Salway, A., Fløttum, K. and Elgesem, D. (2015). Representations of the future in "accepting" and "sceptical" climate change blogs. To appear: *Procs. Corpus Linguistics 2015*, Lancaster University.

1. Select non-textual phenomenon

- It was decided in advance that the focus would be on how people think about climate change and the future: this was motivated by a review of the literature on climate change communication.

2. Select conceptual framework

- We did not make reference to any previously existing conceptual frameworks relating to how climate change is thought about.
- Rather, the framework was developed inductively based on data-driven content analysis.

3. Establish mapping between concepts and linguistic forms

- We first identified frequent linguistic forms that could be related to representations of the future,
  - Frequency lists
  - Word clusters
  - Sorted concordances
- Then, with some close reading, these forms were interpreted to propose nine categories of meaning representations: "(1) sustainability, (2) value-laden positive, (3) value-laden negative, …."

3. Establish mapping between concepts and linguistic forms

- Frequency lists: we inspected the 1500 most frequent words and identified 11 that could be part of future representations, e.g. “future’, “risks”, “opportunities”; the 11 selected words had 30,000 instances in total
- Word clusters and sorted concordances: to give a more condensed view of the co-texts around the identified words → 42 patterns

Part of a sorted concordance

46	, men, farmers and pastoralists can have	a bright	future and never again suffer from famine hopefully th
47	industrialist oleg deripaska said he saw	a bright	future for nuclear development "because only nuclear c
48	energy that will allow mankind to have	a brighter	future, and this needs to happen now the only bothersor
49	erved climate changes do not portend	a calamitous	future, global warming alarmism is invading nearly eve
50	ot day proof of global warming and	a catastrophic	future during their grandparents' early lives (and som
51	hat the world is on the brink of	a "catastrophic"	future of killer heatwaves, floods and droughts unless
52	ot day proof of global warming and	a catastrophic	future during times of natural global warming, elevate
53	nly seem to have a strong faith in	a catastrophic	future has global warming really stopped? has global w
54	e legislation an important step towards	a cleaner	future for australia but said much more needed to be d
55	e change, guilt, love of nature, wanting	a decent	future for your children as an illustration of the li
56	climate change, we might fail to create	a decent	future - we're pretty close to the edge now and there
57	limate change under control and preserve	a decent	future for our grandchildren unless we leave most of t
58	ostlethwaite as an old man living in	a devastated	future earth, watching archive film of the planet and
59	at it is how we focus collectively on	a different	future, and in focusing on it, make it happen "a power
60	te chapter of climate wars described	a different	future scenario, exploring how climate change would af
61	ng itself to the inevitability of	a discontinuous	future, with our institutions and life support systems
62	ing, and if they're right, the state has	a dismal	future if nothing is done to stop it a group of envir
63	ta and projecting perceived trends into	a distant	future that is difficult to grasp so much of the publi
64	ta and projecting perceived trends into	a distant	future that is difficult to grasp so much water is ext
65	d one to think so climate change is not	a distant	future climate change is not a forever problem climate
66	e " and far from being a threat only in	a distant	future, "climate change is happening now " and if ther
67	das of governments and it is not just	a distant	"future" climate change that threatens us and it is no
68	e change and climate model projections of	a drier	future across the south-east * iucn press release, dec
69	argue that it won't be a crisis in	a foreseeable	future either neither howard nor rudd have committed
70	and when the main actor in that movie is	a former	future president, the rules of the game suddenly under
71	baird, wikimedia commons) re-imagining	a global	future through dialogue and action tippingpointaustral
72	walk out of any presentation that showed	a gloomy	future; how people in her church would immediately dis
73	lm the age of stupid projects forward to	a gloomy	future climate change is a real phenomenon climate cha

Three of the induced patterns

Pattern	Unique fillers	Total instances	Number of instances for the five most frequent fillers
a|an WORD future	97	239	sustainable (34); low-carbon (19); better (15); uncertain (12); greener (7)
risk(s)|danger(s)| threat(s) facing WORD	30	142	the (43); our (26); humanity (25); mankind (10); humankind (5)
opportunity(ies) to WORD	325	843	make (39); address (18); put (16); build (16); take (15)

4. Identify significant statistical results in the frequencies of linguistic forms

- The established mapping between the categories and linguistic forms (i.e. pattern-filler combinations) facilitates quantitative analyses and identification of samples for close reading.
- For example, to test the hypotheses that: "accepting" climate change blogs would be more concerned with the future than "sceptical" blogs.

Critique

- Data-driven techniques provided a manageable view of a large text corpus, and, in concert with manual interpretation of the results and close reading of samples, they assisted in developing a conceptual framework and a mapping between concepts and countable linguistic forms.
- It should be noted that, in this example, the mapping between categories and their textual realisations is not comprehensive – rather, as the result of a frequency-led analysis, we expect that it captures the most common textual realisations.
- Furthermore, we cannot guarantee that every instance of a certain linguistic form is being used to convey the same meaning – we assume that most of them are, based on the close reading of some examples.
- The method that identified salient patterns and provided a condensed view of their co-texts relied on manual analysis of lists of word clusters and sorted concordances, which was somewhat ad hoc and time consuming.

Local grammar induction

(Salway and Touileb 2014)

- Aim: to generate an overview of the content that preserves more linguistic structure – and hence meaning – than is possible with bag of words approaches
- Main points:
  - Highlights distinctive patterning in large unannotated text corpora
  - Automatically induces frequent local grammatical structures: these characterise what is typically written about key domain terms, and may reflect salient information structures
  - Does not rely on linguistic resources
  - Does rely on repeating patterns in the texts, i.e. constrained domain + stylized language → more structure is induced
  - Like other unsupervised techniques, the output can be very sensitive to small changes in input and parameters
  - Work in progress: we don't fully understand what it captures and what it misses; it is computationally intensive

Example output

```
((carbon|(greenhouse gas)|co2) emissions)

((anthropogenic|manmade|(man made))
global_warming)

((source|emitter|emitters|producers) of
greenhouse_gases)
```

Example output

```
((to (combat|minimize|tackle)) climate change)

(((due to)|(caused by)) ((climate change)|(global warming)))
```

Example output

```
((of global warming) (was|are|is))

(in (order|(the (atmosphere|recessions))))
```

Example output

```
(((greenhouse gases)|emissions|gases|(carbon
emissions)|pollution) blamed ((for|to)
global_warming))

((would|should|to|must) (control|reduce|regulate|
regulating|release) greenhouse_gases)

(((((global|some|sophisticated|complex|the)
climate models)|climate models) (project|suggest|
predict)) that)
```

Using induced structures to highlight interesting content? (Touileb and Salway 2014)

- Patterns that are unusually frequent in a particular blog can give insights into its content, for example:

(the (causes|effects) | (consequences|impacts) ((of|for) climate change))): 460 – *the impacts of climate change (224), the effects of climate change (203) , the consequences of climate change (29), the causes of climate change (4)*

((developing|poor) countries): 1172 – *developing countries(1061), poor countries (111)*

((to (combat|minimize|tackle)) climate change): 130 – *to tackle climate change (72), to combat climate change (57), to minimize climate change (1)*

((you|we) (can|should)): 590 – *we can (302), you can (196), we should (84), you should (8)*

Using induced structures for infromation extraction (Salway, Touileb and Tvinnereim 2014)

```
(COUNTRY ((supported|opposed) by) COUNTRY)
```

This induced pattern was used to extract data about country relations.

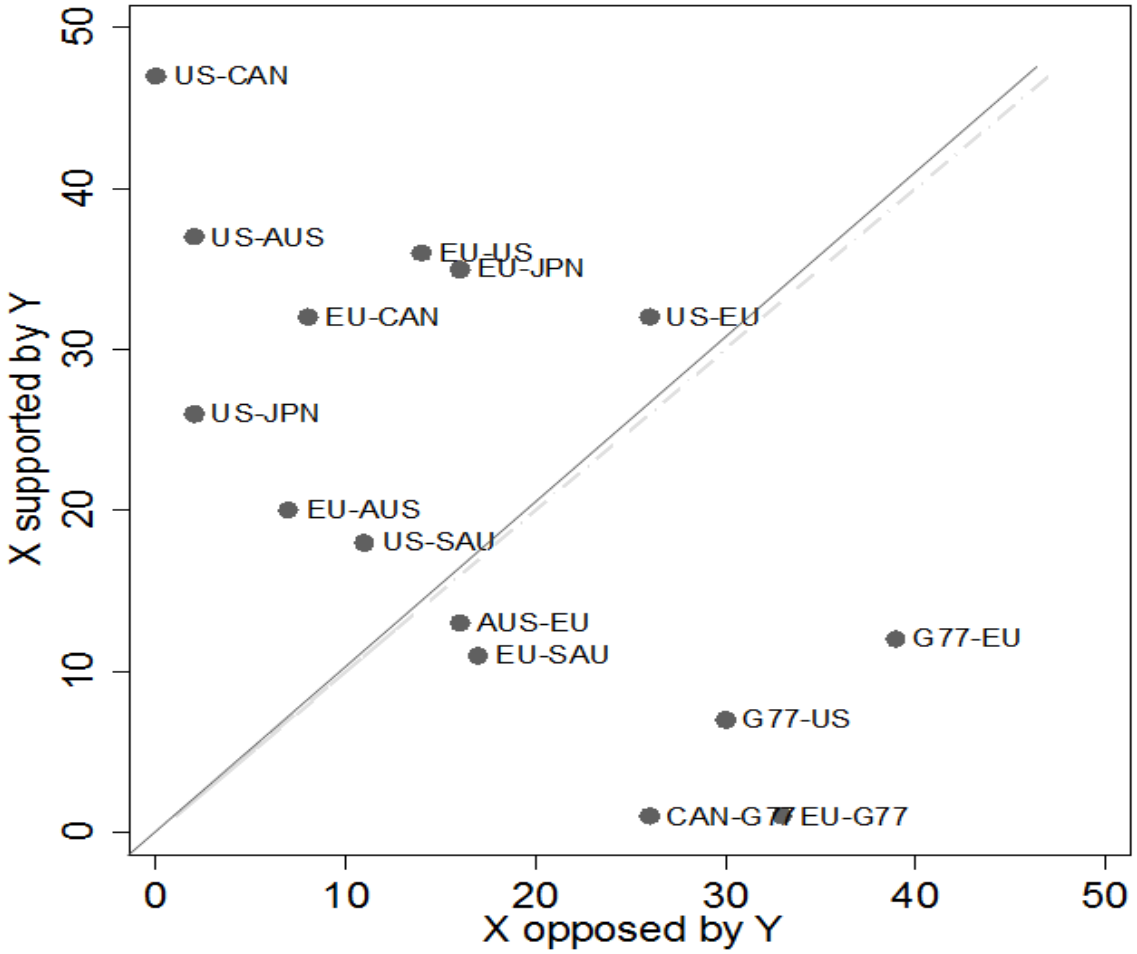
```
(COUNTRY (said|noted|recommended|explained|
responded|stressed|questioned|addressed|reiterated|
reported|urged|amended|invited…))

(COUNTRY ((clarified|urged|reported) that)

(COUNTRY ((presented|demanded|outlined|favored (the|
a))
```

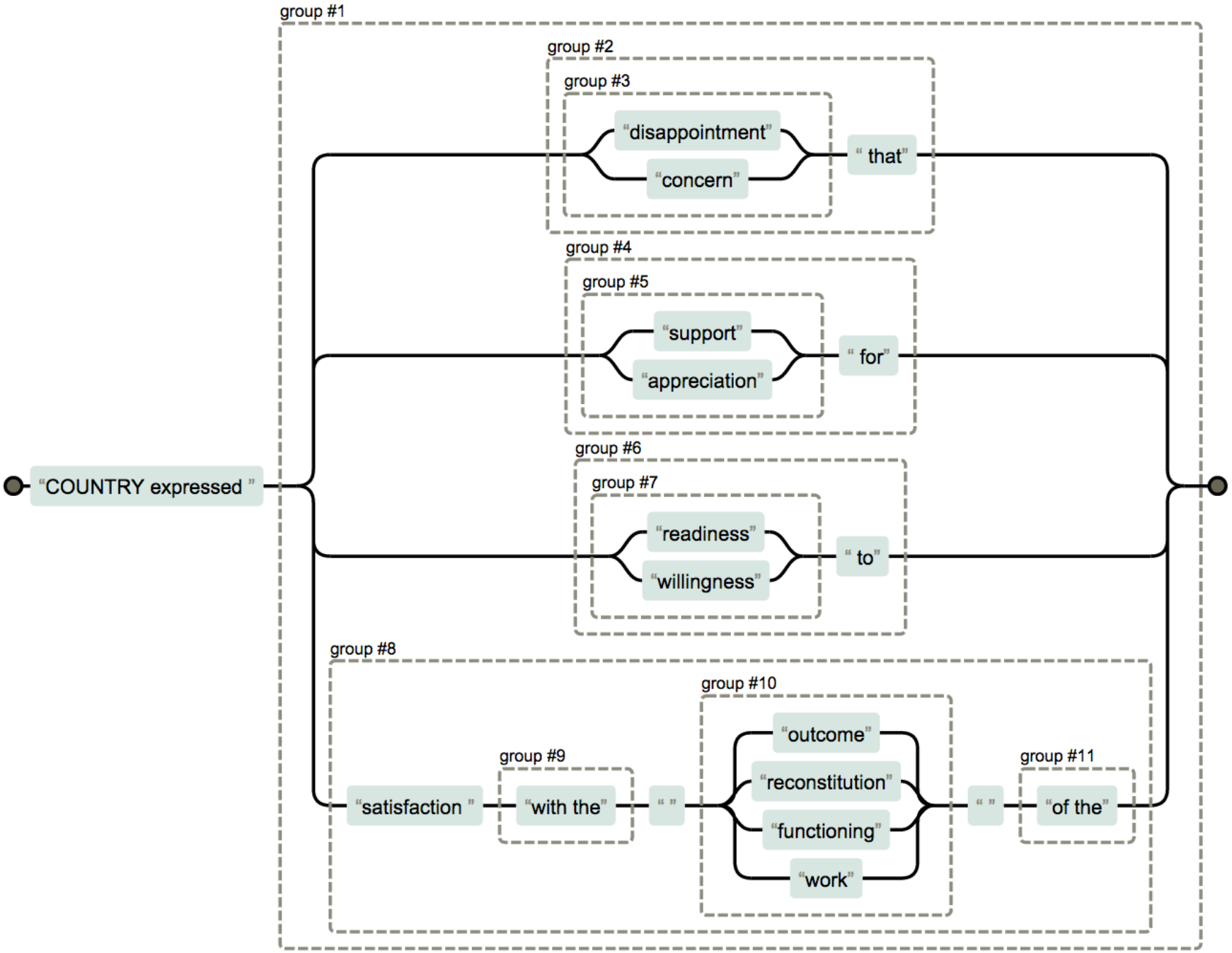
These patterns were used to extract statements relating to countries' positions. The statements were grouped by country and scaled.

Dyads of support and opposition



Scale of climate change statements

Austria (-2.38), Belgium, Germany, the UK, Switzerland, the US, Canada, Australia, Norway, France, Russia, New Zealand, Japan (-.62) ,Papua New Guinea (-.26), Tuvalu, Peru, Mexico, Brazil, Argentina, Malaysia, South Korea, Colombia, Saudi Arabia, Chile, Kuwait, Nigeria, Grenada, Uganda, Bangladesh, China, Egypt, the Philippines, South Africa, Indonesia, Venezuela, Iran, Bolivia, Barbados, India, Algeria (1.44)


group #1
group #2
group #3
"disappointment"
"concern"
" that"
group #4
group #5
"support"
"appreciation"
" for"
group #6
group #7
"readiness"
"willingness"
" to"
"COUNTRY expressed "
group #8
group #10
"outcome"
"reconstitution"
"functioning"
"work"
group #9
"satisfaction "
"with the"
" "
group #11
" "
"of the"

Summary

- Bigger and more diverse material →

  need data-driven techniques that highlight unusual textual patterning, i.e. interesting content

- Some useful data-driven techniques already, from text mining and corpus linguistics, but these do not capture much linguistic structure and meaning →

  "local grammar induction" may complement these