Networks of Texts and People

Andrew Salway
Uni Research, Bergen

Networks of Texts and People (NTAP)

- New methods to analyse the distribution, flow and development of statements in online social networks
  - e.g. climate change discourse in the blogosphere
  - to contribute to social science research into framing, information diffusion and polarization
- Text analysis, network analysis and visualization
  - text analysis: corpus linguistics, text mining and IE
  - identify salient linguistic constructions / information structures ➔ analyse their occurrence over the blogosphere and over time

Networks of Texts and People (NTAP)

- Dept. of Information Science and Media Studies, UiB and Computational Language Unit, Uni Research
  - Nick Diakopoulos
  - Dag Elgesem
  - Knut Hofland
  - Andrew Salway
  - Lubos Steskal
  - Samia Touileb
- Funded by Research Council of Norway, 2012-2016.
- www.ntap.no

Climate Change Corpus

January 1, 2012 to March 1, 2012

8,415 authors | **112,510** posts | **8,123** blogs

(enter search terms here) Go

1/12 2/12 3/12

Statements

Sort by: frequency

Climate change

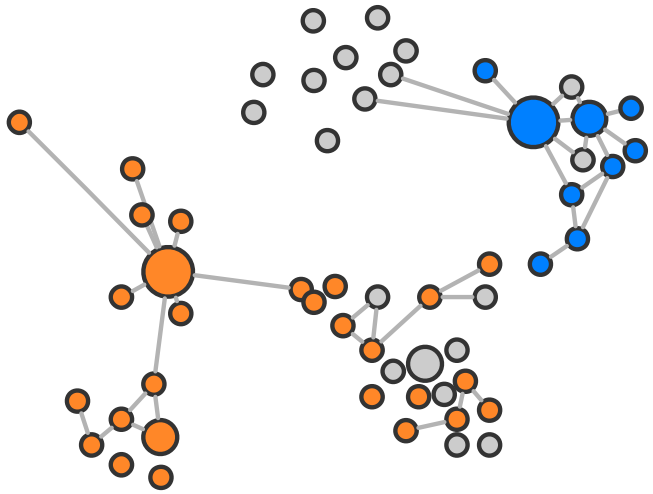
- is_caused_by, “natural causes" — 35 — 40%
- is_caused_by, “the excessive amount of carbon emissions poured into the atmosphere" — 32 — 75%
- is, “changing the arctic" — 25 — 85%
- is, “based on fraudulent science" — 16 — 12%
- will be, “arrested by peak oil" — 5 — 60%

Global warming

- is_caused_by, “burning fossil fuels like coal" — 45 — 92%

Network

View Options



NTAP blog corpus

- English-language blogs that mention broad climate change issues across science, politics, environment, etc.
- Method:
  - Handpicked 5 seed blogs - well connected in the blogosphere, representing different viewpoints in the climate debate
  - Harvested all posts from the seed blogs and extracted key terms to be used to determine topical relevance in the crawl:
    - frequent words typical of the domain: e.g. *climate, global, carbon, emissions, temperature, sea, solar, greenhouse …*
    - n-grams (2 <= n <=5) containing words, e.g. *climate change, climate science, carbon dioxide, emissions trading, sea level …*
  - Breadth first crawl from the seed blogs:
    - harvest blog if English-language and has a key term on homepage
    - follow links from the homepage of each blog
    - limited to WordPress and Blogspot blogs

NTAP blog corpus

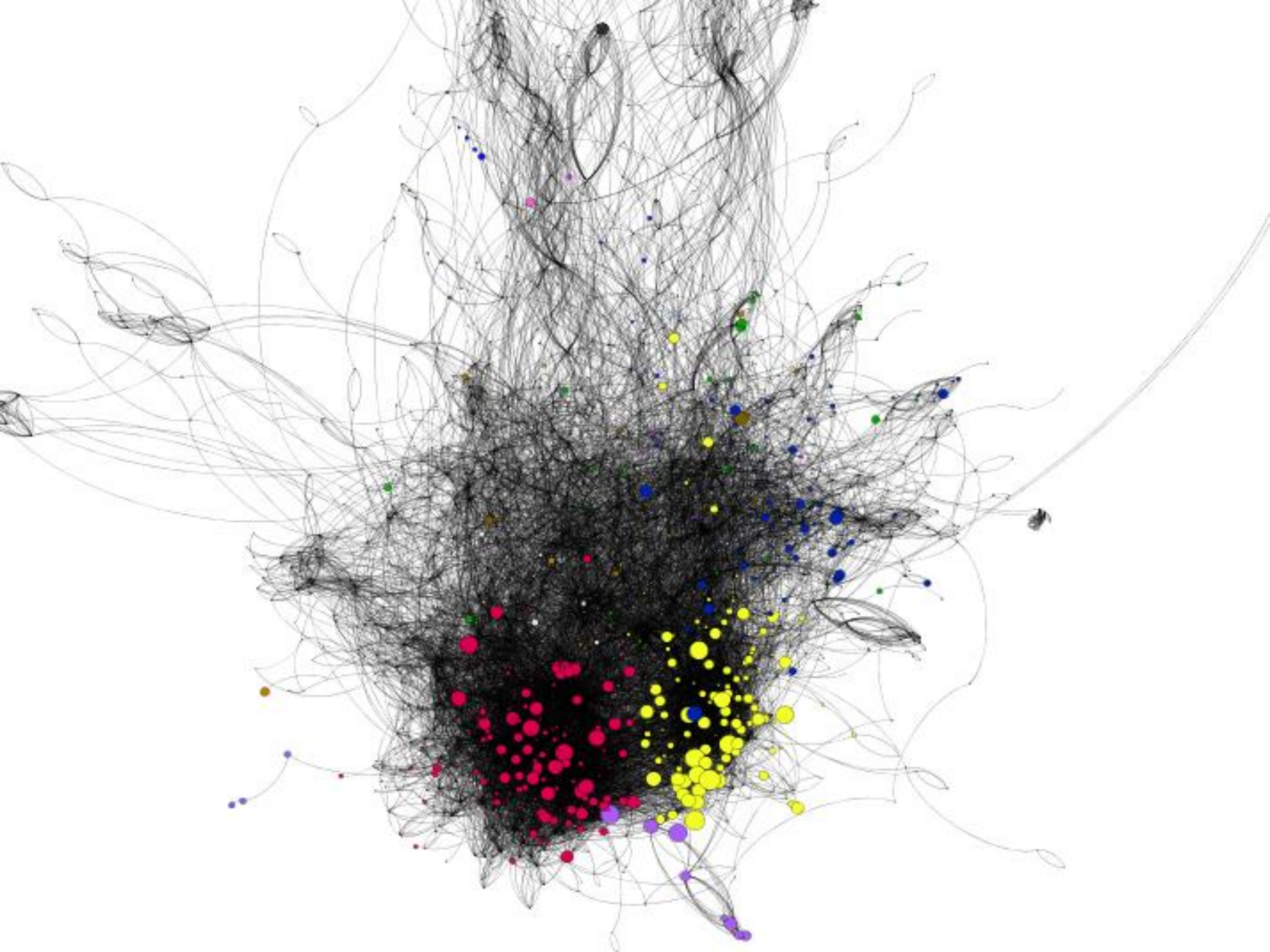
- Crawl carried out June-September 2012
- Harvested the complete content of about 3,000 English-language blogs, about 1.5m blog posts
- Text content of each post extracted using the Alchemy API and stored in a MongoDB database, with date
- Hyperlinks extracted links and stored these in a Neo4j graph database.

Network analysis

- Used a community detection algorithm based on modularity maximization
  - blogs grouped to maximize inter-group hyperlinking (Louvain method, implemented in the Gephi tool)
- This suggested 11 major communities in the corpus, accounting for about 60% of all blogs
- A selection of blogs were inspected to manually code each community as “accepting”, “skeptical” or “neutral” regarding anthropogenic global warming

Topic modelling

- Latent Dirichlet Allocation (LDA) used to identify topics within the corpus, using MALLET tool
- Two (out of 20) topics related strongly to climate change:
  - "climate change science": *climate, warming, global, change, ice, data, temperature, years, science, scientists, carbon, sea, earth, year, ocean, time, temperatures, scientific, research*
  - "climate change politics": *climate, change, countries, world, environmental, international, development, global, emissions, carbon, india, environment, people, government, nations, policy, china, issues, sustainable*
- Other topics: "energy", "wildlife", "legal", "education", "economic policy", "transportation", "American politics", "storms and floods", "farming", "health", "new age", and some noise (incoherence topics and non-English words)



Sub-corpora for two communities

- Our network analysis, topic modelling and manual coding suggested two communities concerned with climate change science, one broadly “accepting”, one “skeptical”

➔ sub-corpora

- “accepting” 204 blogs, 69k posts, 27m words
- “skeptical” 417 blogs, 290k posts, 127m words

Sub-corpora for two communities

- Little difference between the 300 most frequent words (and the 300 top keywords) in the two sub-corpora
  - perhaps not surprising given role of topic modelling in selecting the sub-corpora, and a large volume of peripheral material?

- But, some lower frequency words, and word clusters, seem to be preferred by one community or the other…

Sub-corpora for two communities

	"accepting" (204 blogs, 69k posts)		**"skeptical" (417 blogs, 290k posts)**	
	no. of blogs	no. of posts	no. of blogs	no. of posts
acidification	64	1782	75	412
coral	70	939	117	701
ocean	122	4627	206	6550
species	112	3455	196	5319
tax	100	2339	213	17,180
Gore	84	845	200	10,669

Sub-corpora for two communities

	"accepting" (204 blogs, 69k posts)		**"skeptical" (417 blogs, 290k posts)**	
	no. of blogs	no. of posts	no. of blogs	no. of posts
climate science	99	2115	155	3551
anthropogenic c c	55	349	79	360
human-caused c c	37	124	57	201
human-induced c c	33	144	57	273
man-made c c	31	73	97	566
climate change denial	38	165	41	109
climate alarmist	6	6	47	216

Sub-corpora for two communities

Mentions of causes of climate change

	"skeptical" (35k instances of climate change)	"accepting" (22k instances of climate change)
cause [verb forms] climate change	147	49
cause(s) of climate change	117	34
contribute(s|d) to climate change	68	34
affect [verb forms] climate change	18	7
lead to [verb forms] climate change	6	5
result in [verb forms] climate change	3	2
TOTAL	**359**	**131**

Sub-corpora for two communities

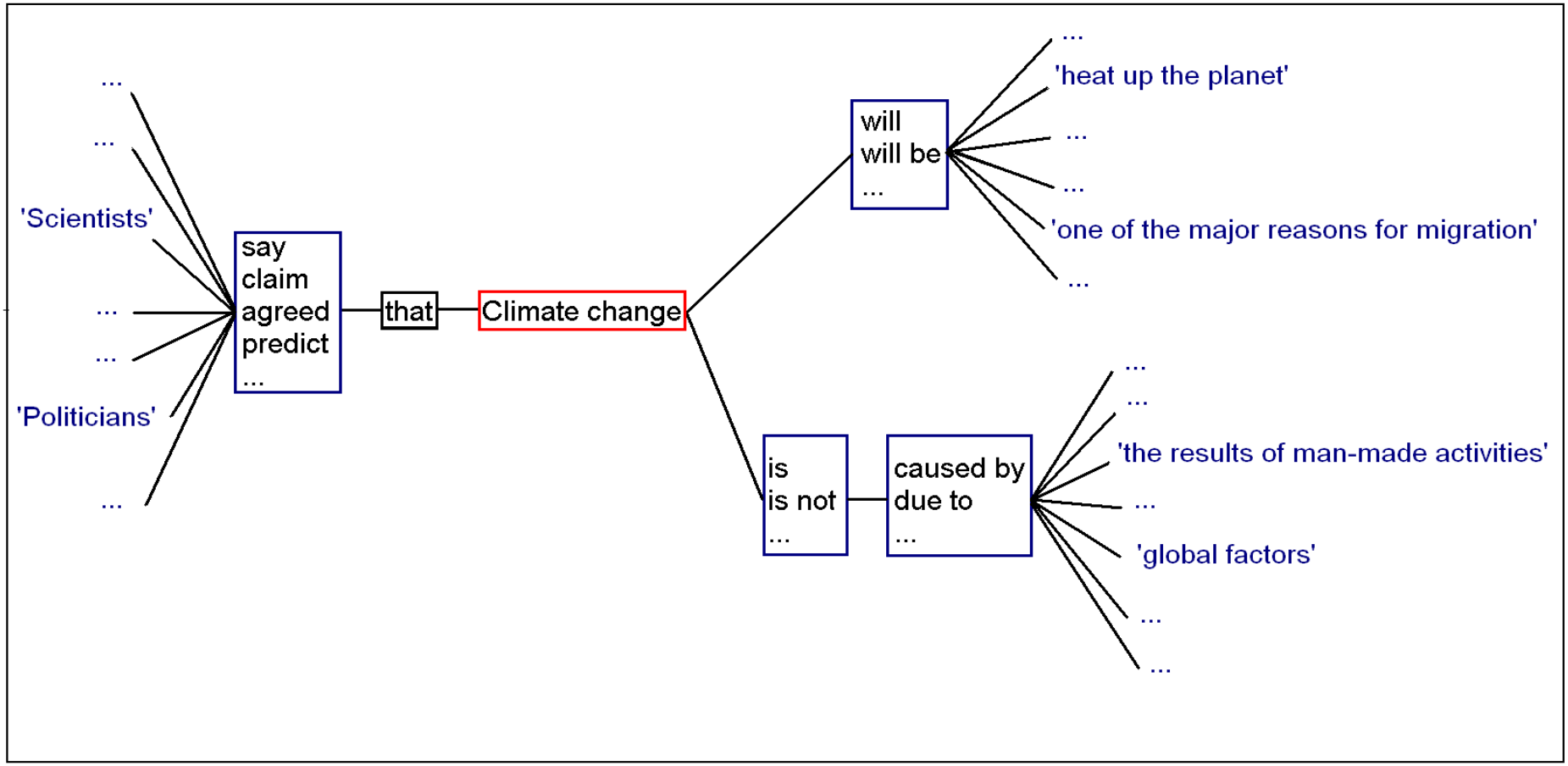
Mentions of effects of climate change

	"skeptical" (35k instances of climate change)	"accepting" (22k instances of climate change)
result|effect(s)|impact(s)|consequence(s) of climate change	1,412	1,034
due to climate change	179	148
climate change cause|affect|lead to|result in|contribute to [verb forms]	68	34
TOTAL	**1,659**	**1,216**

Local grammar induction

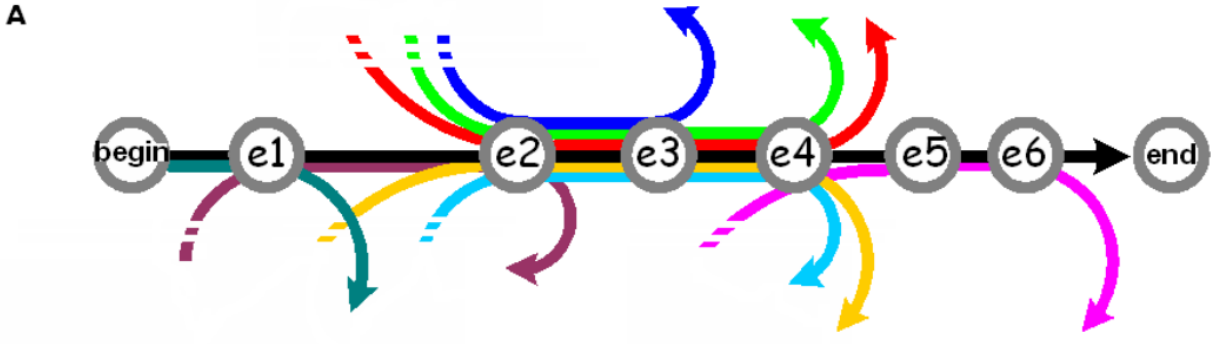
- For exploratory purposes, we need an overview of what is typically written about key terms in the corpus
  - 100,000's lines for "climate change", 10,000's lines for "sea levels"
  - manual analysis of concordance lines is not feasible (?)
- Can we exploit relatively restricted and repetitive language use in the corpus (i.e. same text genre, topic, function) to induce and visualize interesting patterning?
  - cf. collocations, clusters, phrases, constructions, lexical bundles, n-grams, collocation frameworks, formulaic units, multiword expressions, phrase frames
  - cf. The Sketch Engine, The Word Tree

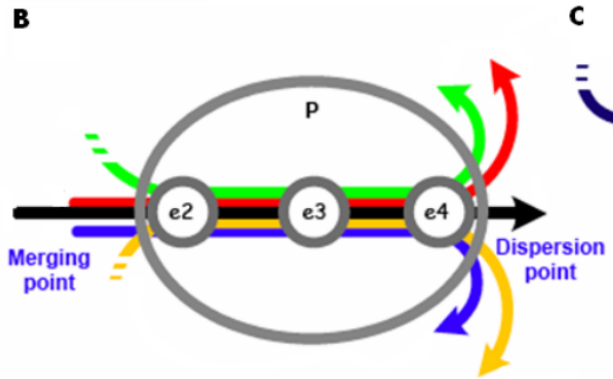
Part of an idealised local grammar for “climate change”

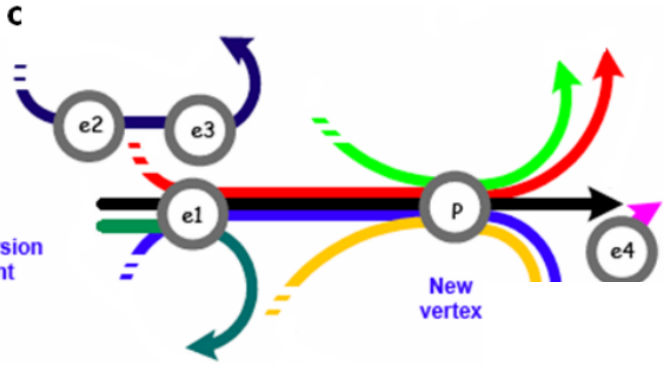


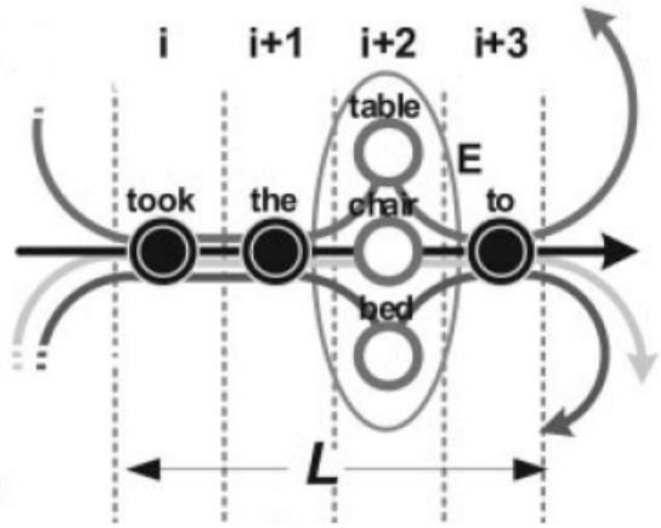
Grammar Induction

- Grammar induction algorithms generate a grammar from an unannotated corpus, i.e. based on formal distributional properties of words(cf. Harris 1954)
- For example, ADIOS (Automatic Distillation of Structure, Solan et al. 2005)
  - Organises all input sequences (sentences) on a graph with a node for each unique vocabulary item
  - In each iteration, simultaneously forms patterns (syntagmatic units) and equivalence classes (paradigmatic units)


A
begin
e1
e2
e3
e4
e5
e6
end


B
P
e2
e3
e4
Merging point
Dispersion point


C
e2
e3
e1
P
e4
New vertex

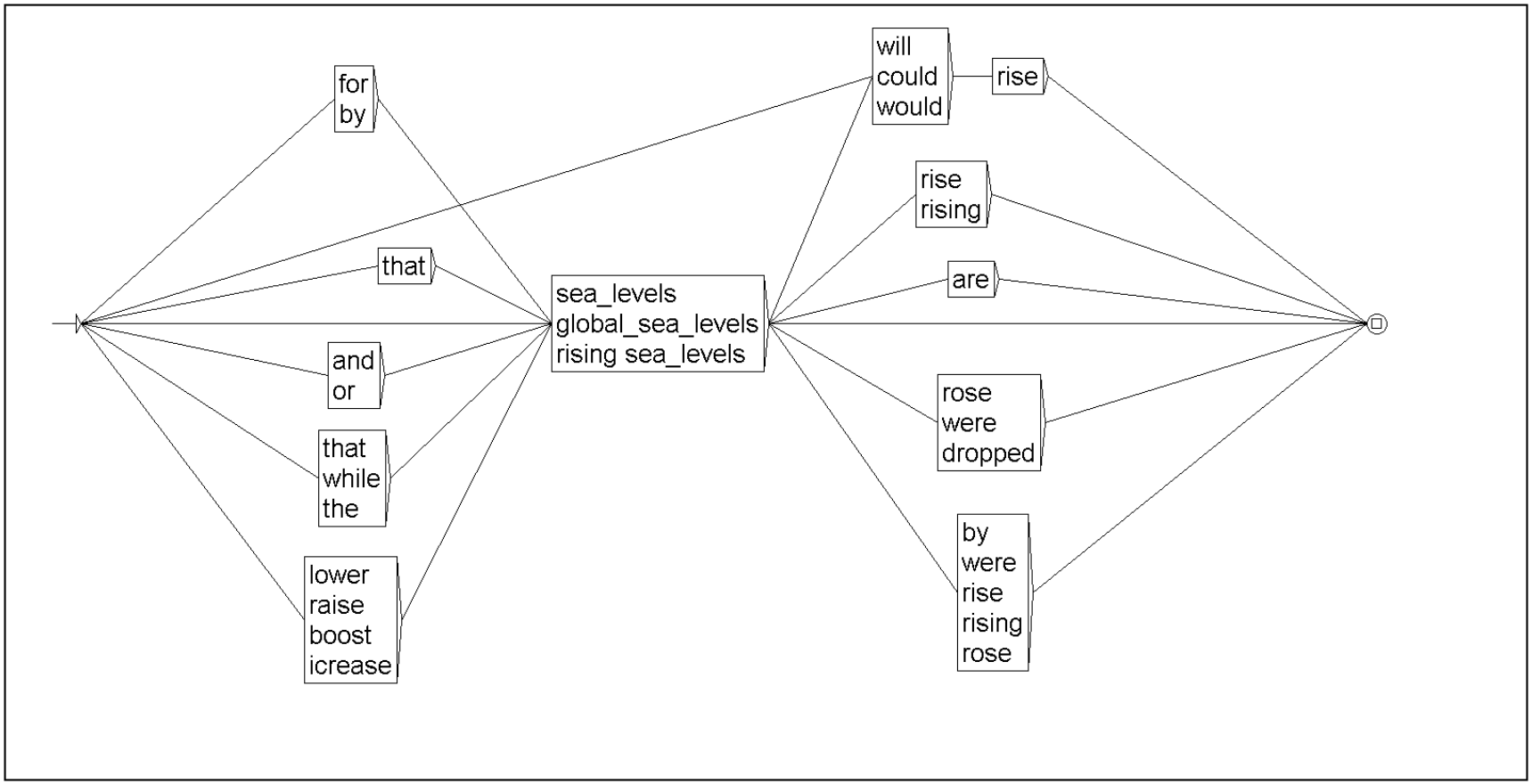

i
i+1
i+2
i+3
table
E
took
the
chair
to
bed
L

Adapting ADIOS
(PhD work of Samia Touileb)

- Input is a set of sentences that contain the same key term
- The sentences are turned into sets of snippets, of varying sizes, around the key term
- Sets of snippets are presented to algorithm in size order, e.g. N iterations for the smallest snippets, then N for the next size, and so on
  - to focus the algorithm on patterns local to the key term
- After each iteration the 'best' patterns and equivalence classes are selected and inserted into the input file
  - to make further patterning more explicit

P_0 (of climate_change)
P_1 (of the)
P_2 (to climate_change)
P_3 (on climate_change)
P_4 (climate_change is)
P_5 ((to|with) the)
P_6 (the (affects|effect|effects|impact|impacts) P_0)
P_7 ((while|and|induced) climate_change)
P_8 (on the)
P_9 (the (dangers|science|problem|risks) P_0)
P_10 (to (meet|tackle|take|redirect))
P_11 (to (escape|address) climate_change)
P_12 (the (psychological|inevitable|worst|visible|negative) effects P_0)
P_13 (to (prevent|combat) climate_change)
P_14 ((threat|threats|risk|risks|challenges) posed by)
P_15 (in (attempts|order) to)
P_16 (the (details|issue) P_0)
P_17 (climate_change (conference|summit) in)
P_18 (climate_change will (have|reduce|increase) the)
P_19 ((scientific|academic) literature)

A step towards a local grammar fragment for “sea levels”



Closing Remarks

- Network analysis, e.g. community detection, may offer a fruitful way to add a dimension to the analysis of social media corpora; perhaps need to fuse network analysis and text analysis to detect communities
- Early results from local grammar induction show promise, i.e. as a way to elucidate interesting patterning in large sets of concordance lines

Ongoing / future work

- Norwegian and French blog corpora
- Comparison of sub-corpora, and temporal comparisons, with regards to the framing of key terms
  - How to determine significant differences between sub-corpora? (NB. skewed distribution across blogs, variable amounts of peripheral material due to permissive crawl)
- Analyses of causality, modality, proposed climate solutions
- Development of local grammar induction
  - Optimising parameters
  - How to create local grammar fragments from P's and E's
  - Evaluation

A few patterns and equivalence classes for “sea levels” snippets

```
P_0     P2339 (E_0,E_1) 360
E_0     E2340 {and,or}
E_1     E2334 {P_1,sea_levels}
P_1     P2316 (rising,sea_levels) 1535
P_2     P2323 (E_2,rise) 282
E_2     E2324 {will,could,would}
P_3     P2324 (E_3,are) 253
E_3     E2325 {global_sea_levels,sea_levels,P_1}
P_4     P2329 (sea_levels,P_2) 234
P_5     P2344 (E_4,sea_levels) 219
E_4     E2343 {that,while,the}
P_6     P2358 (E_5,E_6) 198
E_5     E2359 {lower,raise,boost,increase}
E_6     E2360 {global_sea_levels,sea_levels}
.
.
.
```