How to read 100,000+ concordance lines?

Andrew Salway, Uni Research
Samia Touileb, University of Bergen

Motivation

- Humanities and social science researchers need to read concordance lines as part of…
  - corpus-based discourse analysis
  - corpus-based lexicography
  - corpus stylistics
  - etc.
- As corpora continue to grow, it becomes impossible to read even a reasonable sample of all concordance lines for a single word of interest

Problem Definition

- The solution should:
  - generate a convenient overview of salient lexico-grammatical patterning in the co-texts of a word of interest, i.e. it should elucidate the common uses/meanings of the word
  - be portable across languages and domains
  - be as unbiased as possible, i.e. minimise assumptions about the language used in the corpus
- cf. discovery tools for data-driven science

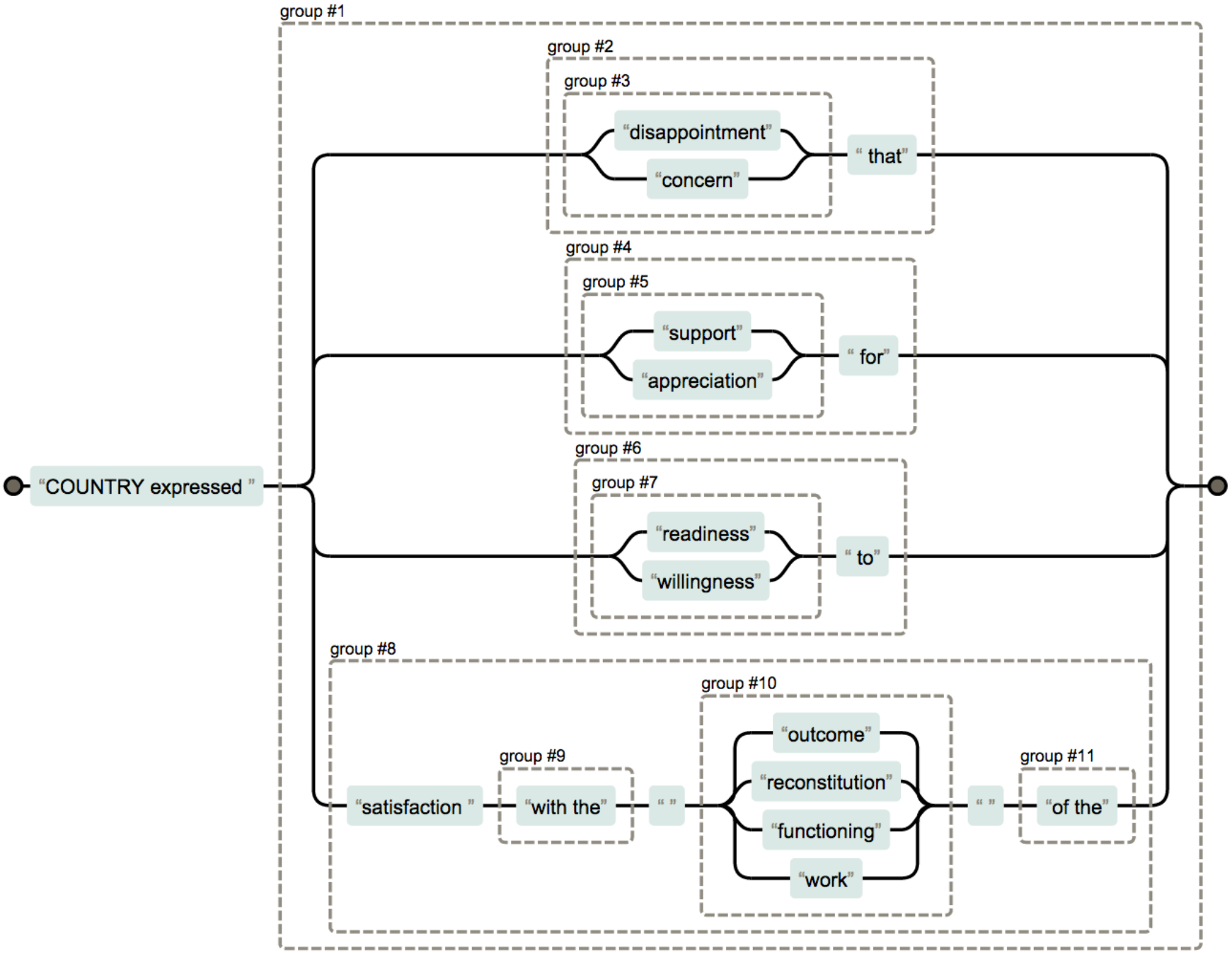
A brief review of text mining and corpus linguistic techniques

- *n-grams, word clusters*: identify frequent word sequences containing the word of interest
- *collocations, topic modelling*: identify words that tend to co-occur with the word of interest in some window
- *semantic classes*: identify other words that are used in similar ways, based on distributional information
- *Sketch Engine*: generates rich summaries of a word's usage, but requires the prior defintion of grammar patterns

We see an opportunity to complement these with a solution that combines sequential and paradigmatic information, without specifying a grammar.

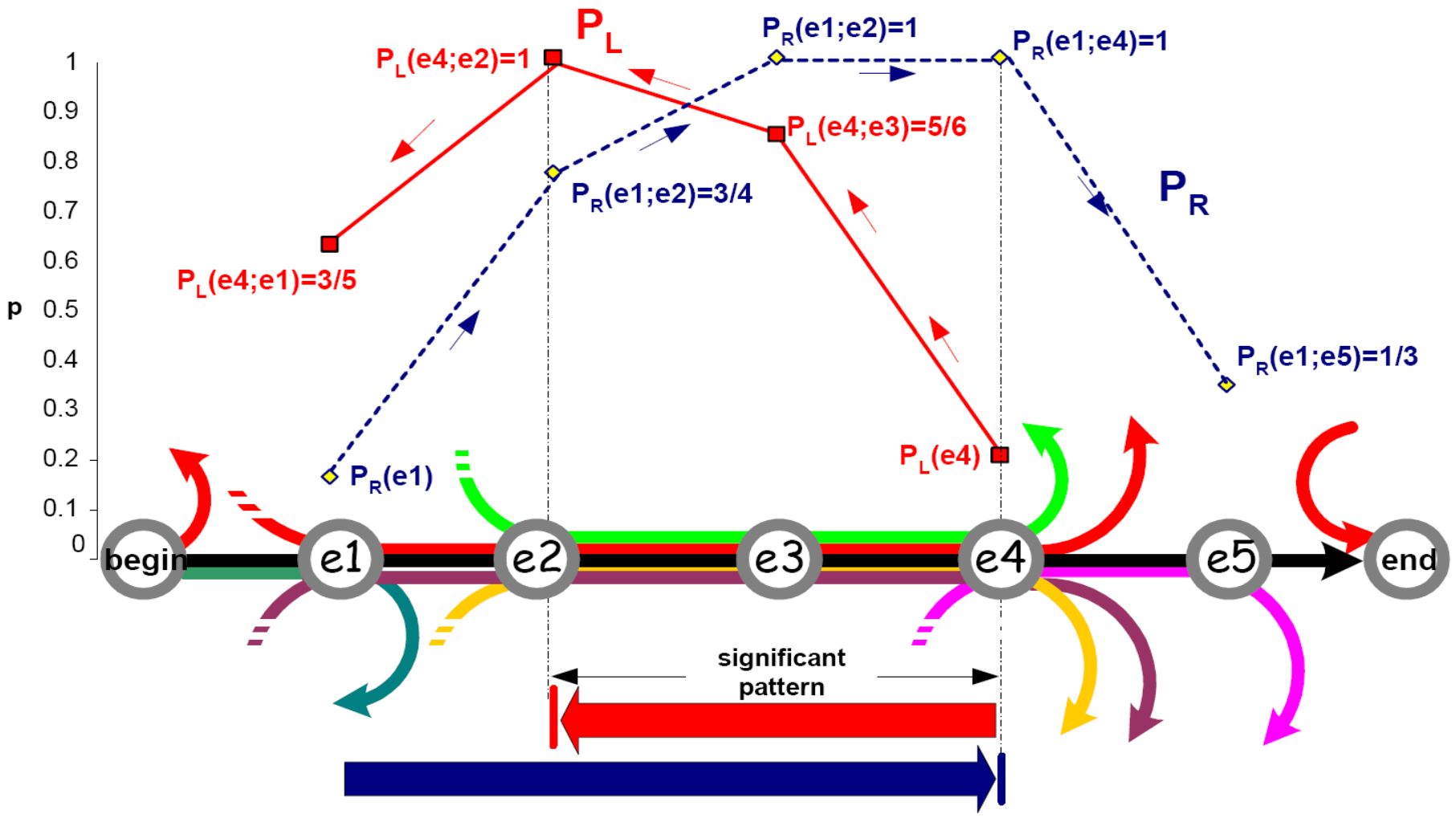
Proposed solution: local grammar induction

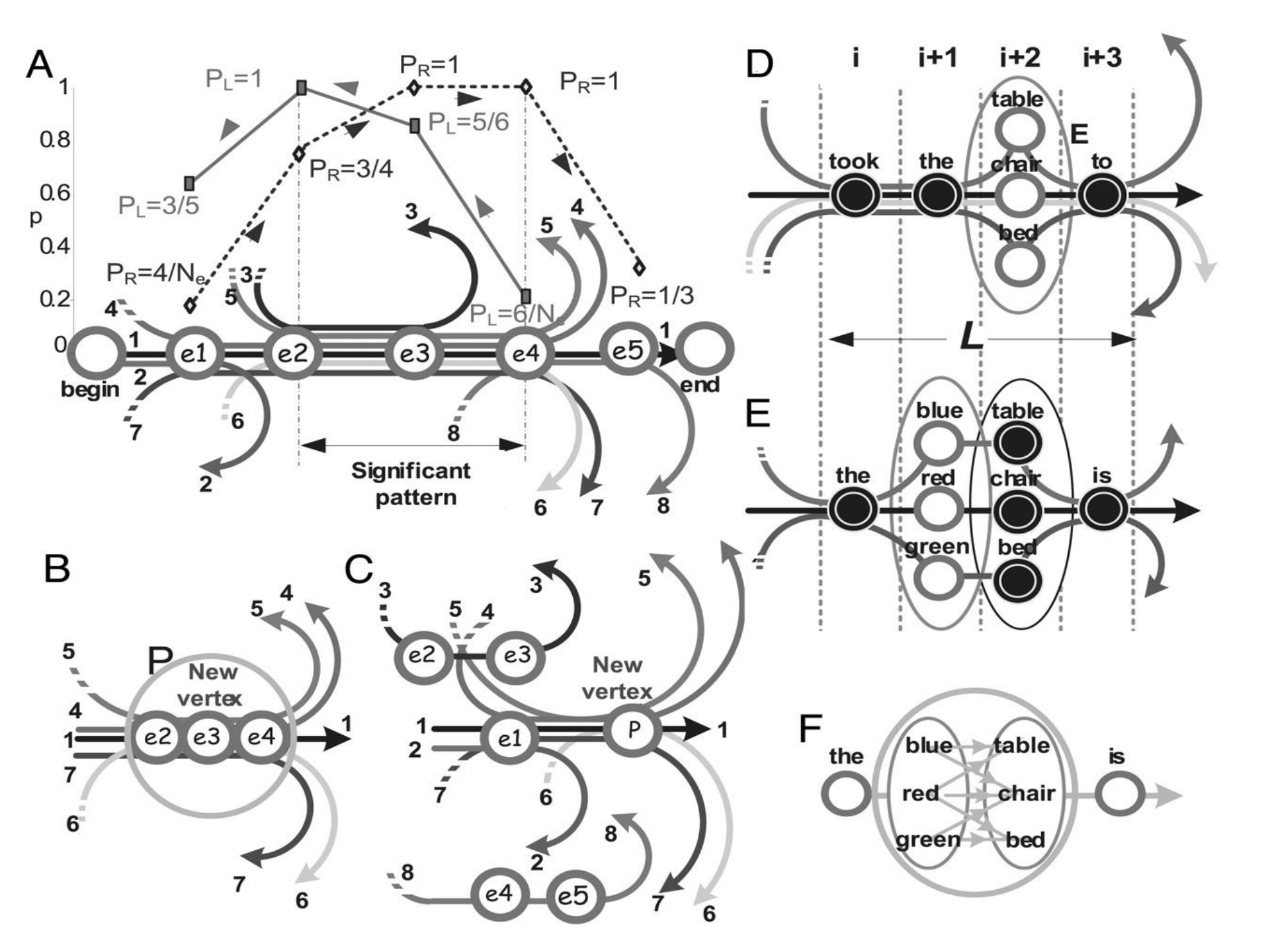
- Modify a grammar induction algorithm to induce local grammar fragments from an unannotated corpus
- Background concepts/theory:
  - Grammar induction, e.g. ADIOS (Solan et al. 2005)
  - Distributional analysis can identify linguistic units and structures (Harris 1954); induced linguistic structures reflect important information structures, especially in sublanguages (Harris 1988)
  - Local grammar (no general word classes) provides a better fit with usage (Gross 1997); cf. pattern grammar, construction grammar


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"

ADIOS: automatic distillation of structure (Solan et al. 2005)

- An unsupervised algorithm that recursively distils (induces) hierarchical patterns from sequential data
- Each sequence (sentence) is loaded onto a directed pseudograph with one vertex for each vocabulary item, so partially aligned sequences share sub-paths across the graph
- In each iteration:
  - the most significant pattern is identified with a statistical criterion that favours frequent sequences that occur in a variety of contexts
  - equivalence classes are identified within the context of the pattern
  - the new pattern and equivalence classes become vocabulary items in the graph so that they can become part of further patterns and equivalence classes: hence hierarchy


P_L
P_R(e1;e2)=1
P_R(e1;e4)=1
P_L(e4;e2)=1
P_L(e4;e3)=5/6
P_R(e1;e2)=3/4
P_R
P_L(e4;e1)=3/5
P_R(e1;e5)=1/3
P_R(e1)
P_L(e4)
p
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
begin
e1
e2
e3
e4
e5
end
significant pattern


A
p
1
0.8
0.6
0.4
0.2
0
P_L=1
P_R=1
P_R=1
P_L=5/6
P_R=3/4
P_L=3/5
P_R=4/N_e
P_L=6/N_e
P_R=1/3
begin
e1
e2
e3
e4
e5
end
Significant pattern
B
P
New vertex
C
New vertex
D
i
i+1
i+2
i+3
took
the
table
chair
bed
to
E
L
E
the
blue
red
green
table
chair
bed
is
F
the
blue
red
green
table
chair
bed
is

Modifying ADIOS for text mining
(Salway and Touileb 2014)

- Input:
  - instead of whole corpus, present text snippets around a single keyterm, cf. concordance lines; local grammar
  - (optionally, constrain snippets to be within clauses/punctuation)
  - present increasingly large snippets to emphasise the most local patterning
- After each iteration:
  - remove patterns containing large equivalence classes; we assume these are more likely to be semantically nebulous
  - in the input data replace instances of the most frequent patterns with common identifiers so that patterning around them is more explicit in subsequent iterations

Preliminary results

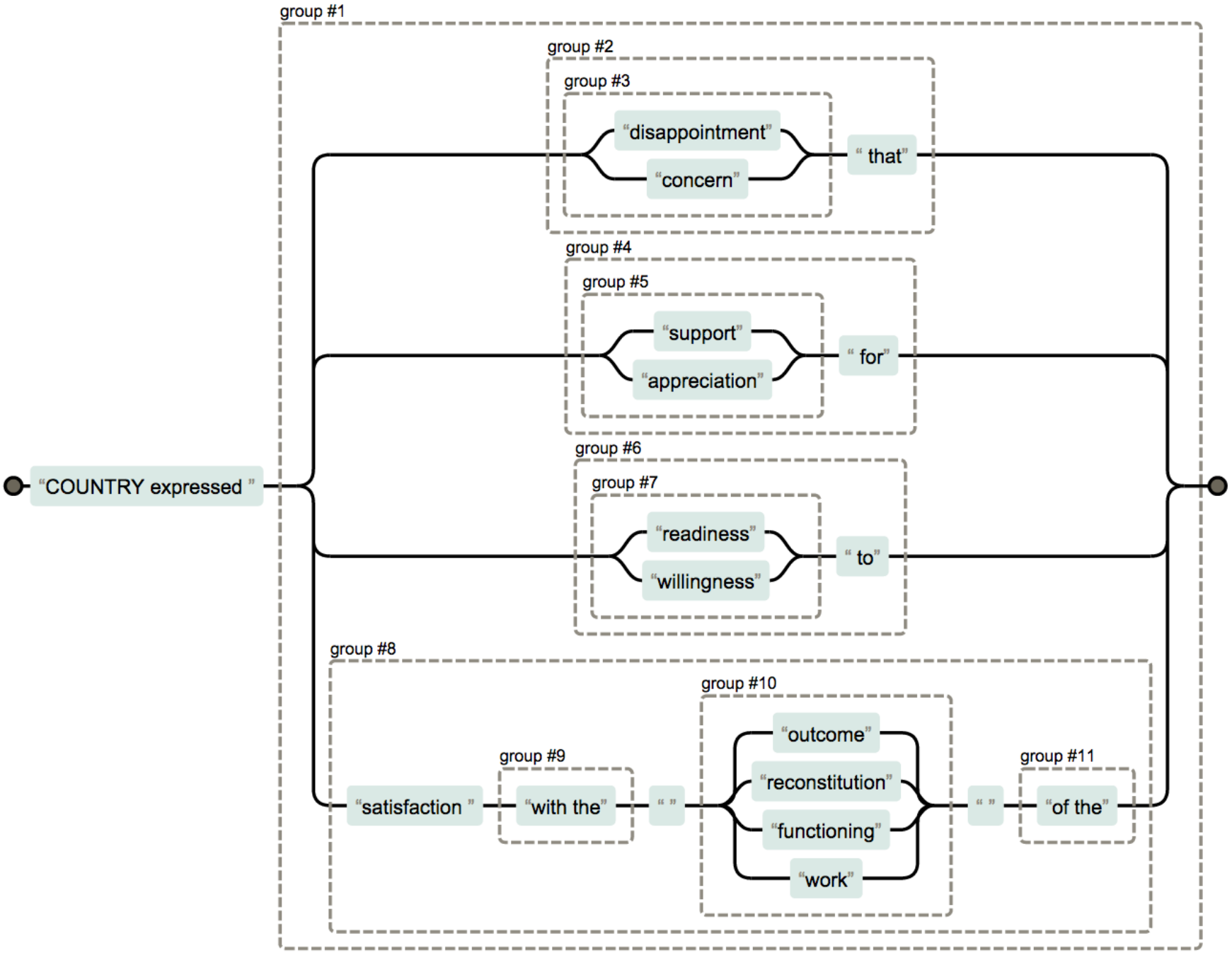
- Climate change blogs (Salway & Touileb 2014; Touileb & Salway 2014)
  - 1.4m English-language blog posts
  - key terms included “climate change” (f ≈ 250,000)

- Earth Negotiation Bulletin (Salway, Touileb and Tvinnereim 2014)
  - minutes of climate change negotiations
  - key term “COUNTRY” (f ≈ 32,000)

- Charles Dickens, 15 novels (Salway and Mahlberg in preparation)
  - “CHARACTER” (f ≈ 58,000), “said” (f ≈ 26,000), “eyes” (f ≈ 3,600), “hands” (f ≈ 2,500)…

```
(((to (combat|fight))| (to
(battle|slow|minimise|mitigate|tackle)))
climate_change)

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

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

((will|would|to) (push|raise|elevate) (sea_levels
(around|by)))

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

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

(COUNTRY
(said|noted|recommended|explained|responded|stress
ed|questioned|addressed|reiterated|reported|urged|
amended|invited…))
(COUNTRY ((clarified|urged|reported) that)
(COUNTRY ((presented|demanded|outlined|favored
(the|a))

(COUNTRY expressed ((disappointment|concern)
that)|((support|appreciation)
for)|((readiness|willingness) to)|(satisfaction
(with the)
(outcome|reconstitution|functioning|work) (of
the)))
```


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"

```
((black|haggard|blue|dark|small|bright|sparkling|l
ittle) eyes)

(((her
eyes)|((with|casting|fixing|and|keeping|bending)
((my|his) eyes))|((CHARACTER's|their) eyes))
((towards|on|over|upon) the))

((he said)
(bending|shaking|rubbing|raising|letting|wiping)
his)


((inquired|exclaimed|replied|asked|muttered|remark
s|added|adds|continued) CHARACTER)
```

Discussion

+ Results suggest that interesting and meaningful structures can be induced automatically from unannotated corpora

+ Sequential and contextual information is combined

+ No need for language-specific resources; less potential for bias

- Results files are quite noisy; function words

- Languages with free word order(?)

? Need to speed-up, implement on Hadoop ?

? Needs a front-end for interactive visualization

References

Maurice Gross. 1997. The Construction of Local Grammars. In: E. Roche and Y. Schabes (eds.), *Finite-State Language Processing*. The MIT Press, Cambridge MA: 329-354.

Zellig Harris. 1954. Distributional Structure. *Word* 10(2/3):146-162.

Zellig Harris. 1988. *Language and Information*. Columbia University Press, New York.

Sydney Lamb. 1961. On the Mechanization of Syntactic Analysis. *Int. Conf. Machine Translation of Languages and Applied Language Analysis.*

Andrew Salway and Samia Touileb. 2014. Applying Grammar Induction to Text Mining. *Procs. ACL 2014.*

Andrew Salway, Samia Touileb and Endre Tvinnereim. 2014. Inducing Information Structures for Data-driven Text Analysis. *Procs. ACL Workshop on Language Technologies and Computational Social Science*.

Andrew Salway and Michaela Mahlberg. In preparation. Inducing narrative features from literary texts.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. *Procs. of the National Academy of Sciences* 102(33):11629-11634.

Samia Touileb and Andrew Salway. 2014. Constructions: a new unit of analysis for corpus-based discourse analysis. *Procs. of the 28th Pacific Asia Conference on Language, Information and Computation.*