UNIVERSITY OF BIRMINGHAM

Closed-class keywords and corpus-driven discourse analysis

Nicholas Groom
Centre for English Language Studies

Discourses

- **Sets of meanings and values which are associated with particular communities or institutions, and which are produced and reproduced through characteristic and often highly conventionalised linguistic choices (e.g. Carter 1995; Stubbs 1996, 1997, 2001; Partington 2003; Baker and McEnery 2005; Teubert 2005; Baker 2006; Mautner 2007).**

Corpora and discourses

- **Corpora allow for more robust statements to be made about discourses than can be made through the analysis of single texts**
- **Corpus analysis allows researcher to identify meanings and values in the language of many individuals in a discourse community**
  - Dominant and contested meanings
  - Synchronic and diachronic perspectives

UNIVERSITY OF BIRMINGHAM

Keywords, corpora and discourses

- **Keywords analysis:**
  - Delegates task of identifying objects of analysis to computer
  - Not based on any particular theory of language
  - Identifies items that would be difficult or impossible for human analyst to observe or predict
  - Tends to generate very long lists of candidate items for analysis

UNIVERSITY OF BIRMINGHAM

Selecting keywords for discourse analysis

- **Strategy 1: Topslicing**
  - –Words at top of list have highest sig. scores
  - –Ad hoc solution
  - –What lurks beneath?

UNIVERSITY OF BIRMINGHAM

Selecting keywords for discourse analysis

- **Strategy 2: Pruning**
  - 3 kinds of keyword (Scott 1999; Scott & Tribble 2006):
    - proper nouns
    - other open-class ‘lexical’ words
    - closed-class ‘grammatical’ words
  - Default strategy: discard or gloss over proper nouns and closed class words, and focus on remaining open-class items

UNIVERSITY OF BIRMINGHAM

Selecting keywords for discourse analysis

- **Strategy 2: Pruning**
  - Rationale:
    - proper nouns only reveal *dramatis personae* and other obvious content features
    - closed-class words only tell us about the **style** of a text or corpus of texts
    - open-class words indicate the **aboutness** of a text or corpus, and are thus "generally those which are most interesting to analyse" (Baker 2006: 127).

UNIVERSITY OF BIRMINGHAM

Selecting keywords for discourse analysis

- **Strategy 2: Pruning**
  - Alternative strategy (Gledhill 2000; Groom 2007, forthcoming):
    - discard all open-class items and focus exclusively on remaining closed-class words
  - Validity?
  - Viability?
  - Desirability?

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as valid objects of analysis

- **How meaningful are open-class words?**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as valid objects of analysis

- **How meaningful are open-class words?**
  - ***gift*** **(n)**

Closed-class keywords as valid objects of analysis

- **How meaningful are open-class words?**

You do have a **gift** for understatement, Jill.	
England gave Romania a **gift** of a goal.	
I wish I could [help you], but it's not in my **gift**.	

Closed-class keywords as valid objects of analysis

- **How meaningful are open-class words?**

You do have a **gift** for understatement, Jill.	talent
England gave Romania a **gift** of a goal.	unmissable opportunity
I wish I could [help you], but it's not in my **gift**.	power/authority

Closed-class keywords as valid objects of analysis

- **How meaningful are open-class words?**

HAVE + a + **gift** + for + N.	talent
a + **gift** + of + a + N	unmissable opportunity
N/PRON + BE + (not) + in + poss. det. + **gift**.	power/authority

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as valid objects of analysis

- **How meaningless are closed-class words?**

a + gift + **for** + N.	talent
a + gift + **of** + **a** + N	unmissable opportunity
in + **poss. det.** + gift.	power/authority

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as valid objects of analysis

- **Closed-class words act as meaning classifiers (cf. pattern grammar, construction grammar)**

a + gift + **for** + N.	talent
a + gift + **of** + **a** + N	unmissable opportunity
in + **poss. det.** + gift.	power/authority

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as valid objects of analysis

- **Meaning typically resides in *sequences* of words, not individual word forms**
- **Both open and closed-class words contribute to the meaning of an idiomatic sequence/ pattern/construction/ ...**
- **... so closed-class words are just as valid as open-class words are as starting points for semantically-oriented analysis**
- **Whither the style vs. aboutness distinction?**

Closed-class keywords as viable objects of analysis

- of
- very significant keyword in 3.2m-word corpus of history journal articles (HistArt) and 4m-word corpus of literary criticism journal articles (LitArt)
- ‘Style’: nominalisation

UNIVERSITY OF BIRMINGHAM

Sequences	Examples	Sample 1 (%)	Sample 2 (%)	Sample 3 (%)
n *of* n	it could not be truly democratic if it presupposed the continuation **of** colonial control.	90	92	87
prep n	they were eventually forced to seek help from local people, particularly from farmers in the vicinity of the camps.	3	5	9
adj *of* n	most of us are aware of rules of evidence	1	2	2
v n *out of* n	the laws acted to take weapons out of men's hands	2	1	0
v *of* n	Bainville … wrote of this recurrent dilemma of French foreign policy	1	0	2
fixed phrase	they facilitated the entry of respectable women into what one turn-of-the-century writer termed the "Night Side of London"	1	0	0
***the* adj-superl *of* pl-n**	the dubbing of a knight is the most familiar of the new ceremonies	1	0	0
v n *of* n	Chulaki accused them both of cosmopolitanism	1	0	0

Closed-class keywords as viable objects of analysis

- of
- very significant keyword in 3.2m-word corpus of history journal articles (HistArt) and 4m-word corpus of literary criticism journal articles (LitArt)
- ‘Style’: nominalisation
- ‘Aboutness’ …?

UNIVERSITY OF BIRMINGHAM

Semantic sequence	Example	Sample 1 (%)	Sample 2 (%)	Sample 3 (%)	Average
PROCESS + *of* + OBJECT	control of female sexuality	11	17	15	14.333
PROPERTY + *of* + PHENOMENON	the basic tenets of Marxist theory	14	13	16	14.333
CONCEPTUALISATION + *of* + PHENOMENON	a tidal wave of conservative loyalism	7	15	15	12.333
QUANTITY + *of* + PHENOMENON	Hundreds of thousands of shares	7	8	7	7.333
PROCESS + *of* + ACTOR	the death of Henry II	6	7	6	6.333
AUTHORITY + *of* + DOMAIN	the third Duke of Northumberland	9	4	4	5.666
PART + *of* + WHOLE	the east wall of that tiny church	3	5	8	5.333
QUALITY + *of* + PHENOMENON	the insignificance of the English	3	7	5	5.000
TEXT + *of* + CONTENT	Journal of Imperial and Commonwealth History	9	2	2	4.333
GROUP + *of* + MEMBERS	a transatlantic 'community of ideals, interests and purposes'	4	2	3	3.000

Closed-class keywords as viable objects of analysis

- **‘Aboutness’:**
  - **CONCEPTUALISATION + *of* + PHENOMENON:**
    LitArt>HistArt
- **Agentive processes (e.g. PROCESS + *of* + OBJECT):**
  HistArt>LitArt

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Tractable lists of items for analysis**
  - **Gledhill (2000): 39 CCKWs**
  - **Groom (2007): 26 CCKWs**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Tractable lists of items for analysis**
  - **Gledhill (2000): 39 CCKWs**
  - **Groom (2007): 26 CCKWs**
- **CCKWs tend to be scattered throughout KW lists (when ranked by keyness) - alternative to random sampling**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Coverage (1)**
  - **The 26 CCKWs obtained for HistArt alone constitute 20.28% of the whole corpus**
  - ***"the majority of text is made of the occurrence of common words in common patterns"* (Sinclair 1991: 108)**
  - **So it is arguably preferable to select the commonest of these common words for analysis.**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**
  - **concordance sample of *of*:**

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**
  - **concordance sample of *of*:**
- **lexical bundles**

1. debate began in London, **in advance of** the 1907 City Council el
2. men?"37 His defensiveness **on behalf of** the scholastic prerogati
3. nds to loom ever larger **in the eyes of** electors, whether their
4. y, on June 8, 1794. 66 **In the face of** such events, the madonna
5. foreign member states. **As a result of** these rulings, and relat
6. ularly from farmers **in the vicinity of** the camps.31 Although th

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**
  - **concordance sample of *of*:**
- **grammar patterns**

1. tion and most of us are **aware** **of** rules of evidence, which of ne
2. ed, we would **know** very little **of** the context of these oaths of
3. m of money. The vast **majority** **of** cases, however, went before th
4. icular, the frequent **outbreak** **of** disease, which posed an ostens
5. erer's own silence is **typical** **of** a larger phenomenon. Before th
6. brother, he **wrote** repeatedly **of** his bewilderment at the eighte

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**
  - **concordance sample of *of*:**
- **semantic sequences**

CONCEPTUALISATION	*of*	PHENOMENON
the complementary component	**of**	cultural stewardship
masculine images	**of**	cigar smoke and stiff wool
the institution	**of**	Mothers' Day
the problem	**of**	punctuation
a tidal wave	**of**	conservative loyalism

UNIVERSITY OF BIRMINGHAM

Closed-class keywords as preferred objects of analysis

- **Coverage (2)**
  - **Different types of phraseology**
  - **concordance sample of *of*:**
    - **semantic sequences: *"recurring sequences of words and phrases that may be very diverse in form and which are therefore more usefully characterised as sequences of meaning elements rather than as formal sequences"* (Hunston 2008: 271)**

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- *"Hammerfest is a thirty-hour ride by bus from Oslo"*
- *"Ntobeye is a two-hour ride by four wheel drive vehicle from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- *"Hammerfest is a* NUMBER+TIME+JOURNEY *by bus from Oslo"*
- *"Ntobeye is a* NUMBER+TIME+JOURNEY *by four wheel drive vehicle from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- *"Hammerfest is a* NUMBER+TIME+JOURNEY *by bus from Oslo"*
- *"Ntobeye is a* NUMBER+TIME+JOURNEY *by four wheel drive vehicle from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- "DESTINATION *is a* NUMBER+TIME+JOURNEY *by bus from Oslo"*
- " DESTINATION *is a* NUMBER+TIME+JOURNEY *by four wheel drive vehicle from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- " DESTINATION *is a* NUMBER+TIME+JOURNEY *by bus from Oslo"*
- " DESTINATION *is a* NUMBER+TIME+JOURNEY *by four wheel drive vehicle from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- " DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from Oslo"*
- " DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from the vast refugee camp at Ngara"*

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- " DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from Oslo*"
- " DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from the vast refugee camp at Ngara*"

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"
- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- *"Hammerfest is a thirty-hour ride by bus from Oslo"*
- *"Ntobeye is a two-hour ride by four wheel drive vehicle from the vast refugee camp at Ngara"*

Semantic sequences

- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"
- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"

UNIVERSITY OF BIRMINGHAM

Semantic sequences

- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"
- "DESTINATION *is a* NUMBER TIME JOURNEY *by* MODE OF TRANSPORT *from* POINT OF DEPARTURE"

UNIVERSITY OF BIRMINGHAM

Conclusion

- Closed-class keywords are
  - valid
  - viable
  - perhaps even preferable

for corpus-driven analyses of specialized discourses

UNIVERSITY OF BIRMINGHAM

Coda

- **Student essay (2009):**
  - ***"The grammatical words 'of' and 'and' cannot tell us anything at all."***

UNIVERSITY OF BIRMINGHAM