

Using salient grammatical words to identify and analyse multi- word units

Nicholas Groom

Department of English

University of Birmingham

Using salient grammatical words
to identify and analyse
phraseology

Nicholas Groom

Department of English

University of Birmingham

Phraseology

- Frequently-occurring sequences in natural language data
- “meaning is attached to frequently-occurring sequences rather than to their constituent lexical or grammatical items” (Hunston 2003: 32).

My research

An analysis of **phraseological variation**
in **written academic English**
across **two disciplines**
and **two genres**

My corpora

History research articles

(*HistArt*): 3,148,745 tokens; 239 texts

History book reviews

(*HistRev*): 3,200,810 tokens; 4,013 texts

Literary Critical research articles

(*LitArt*): 4,057,104 tokens; 575 texts

Literary Critical book reviews

(*LitRev*): 1,011,238 tokens; 685 texts

General principles

- *Corpus-driven* analysis (Tognini-Bonelli 2001)
- Minimal imposition of *a priori* categories on the data (Sinclair 1991, 2004)
- No tagging or parsing
- Use of concordancer to identify latent patterning in local environment of node words

Methodological problem 1

- How to retrieve frequently-occurring sequences in a manner consistent with corpus-driven principles?
- Directly, via lexical bundles, n-grams, chains, etc.
- Indirectly, via Keywords (Scott 1997)

Methodological problem 2

- Keywords procedure yields A LOT of data.
How to select?
- Gledhill's solution (2000a, 2000b): only look at grammatical words (e.g. *the*, *of*, *we*, *and*).

Results of Keywords analysis

- 49 salient grammatical words identified across the four corpora
- 19 found in all four corpora
- 9 found in three corpora
- 15 found in two corpora
- 6 found in one corpus only

Words salient in two corpora: discipline-salient

	<i>LitArt</i>	<i>LitRev</i>	<i>HistArt</i>	<i>HistRev</i>
1. AGAINST			713	672
2. DURING			807	997
3. HIMSELF	769	582		
4. HIS	8,040	7,060		
5. ONE'S	120	99		
6. THEIR			3,541	4,119
7. THOUGH	502	694		
8. UPON	652	568		
9. WHICH	4,438	5,024		

Discipline-salient words across the four corpora (per million words).

Words salient in two corpora: genre-salient

	<i>HistRev</i>	<i>LitRev</i>	<i>HistArt</i>	<i>LitArt</i>
1. HOW	1,460	1,176		
2. LESS	588	567		
3. MORE	2,786	2,734		
4. MOST	1,671	1,651		
5. MUCH	1,242	1,138		
6. THAN	1,922	1,930		

Genre-salient grammatical words across the four corpora (per million words)

Words salient in one corpus

	<i>HistRev</i>	<i>LitArt</i>	<i>LitRev</i>	<i>HistArt</i>
1. ALTHOUGH	970			
2. HER		3,914		
3. HERSELF		260		
4. MANY	1,350			
5. SEVERAL	338			
6. THOSE	1,051			

Grammatical words salient in one corpus (per million words).

Words salient in three corpora

	<i>HistRev</i>	<i>LitRev</i>	<i>LitArt</i>	<i>HistArt</i>
1. AN	4,081	4,773	4,087	
2. BY	6,159		5,918	5,986
3. DESPITE	362	234		210
4. THAT	11,980	13,199	11,920	
5. THIS	5,803	5,906	5,218	
6. THROUGH	1,001	1,064	1,039	
7. WHEREIN	10	12	28	
8. WHILE	907	937	791	
9. WHOM	194	216	263	

Grammatical words salient in three corpora (per million words).

Words salient in all 4 corpora

	<i>HistArt</i>	<i>LitArt</i>	<i>HistRev</i>	<i>LitRev</i>
1. AND	27,961	28,679	34,757	32,378
2. BETWEEN	1,508	1,391	1,698	1,833
3. BEYOND	171	222	252	290
4. BOTH	940	1,224	1,260	1,410
5. ITS	2,011	2,253	2,832	3,096
6. OF	43,413	43,799	47,087	49,533
7. SUCH	1,388	1,391	1,663	1,927
8. THESE	1,616	1,521	1,972	2,085
9. THROUGHOUT	203	203	280	259
10. WITHIN	638	570	699	713
1. AMONG	568	342	752	371
2. IN	25,473	23,954	24,357	23,618
3. THE	71,193	67,770	73,839	66,589
4. THEMSELVES	388	294	424	345
1. AS	7,774	10,573	9,598	12,082
2. ITSELF	317	599	347	718
3. NEITHER	128	162	141	162
4. WHOSE	309	446	390	459
1. NOR	224	313	253	237

Grammatical words salient in all four corpora (per million words).

Case study: *against vs. upon*

	<i>LitArt</i>	<i>LitRev</i>	<i>HistArt</i>	<i>HistRev</i>
AGAINST			713	672
UPON	652	568		

Against in History: the usual suspects

N *against* n

*Venice took no part in the war **against** the Normans*

(>400 p.m.w.)

V *against* n; V n *against* n

*extreme competition ... shaped policies that discriminated **against** blacks.*

*alleged witches and their families also had various strategies that they could employ to defend themselves **against** rumours and formal accusations of witchcraft.*

(c.250 p.m.w.)

Against in History: making connections between events and circumstances

Happenings:

deliberation took place **against** a changing
backdrop of military events

It was **against** this background that abortion
was discussed during the 1930s

What of the normative institutional culture of
charity to the dead, the background **against**
which Stoeckhlin's idiosyncratic views were
drawn?

Perceptions:

Boniface's emphasis on kingship is better understood if viewed **against** the backdrop of the rhetoric of just authority and good rule that surrounded the conflict.

This description should also be seen **against** the backdrop of a new guiding principle for Nordic co-operation, termed 'Nordic usefulness' (nordisk nytte).

Belgium's 'Europeanism' is similarly incomprehensible unless seen **against** the background of its internal dissensions.

Wrightson's book must be seen **against** this pessimistic background.

Upon in Literary Criticism

- *Against* in History: making connections between events and circumstances
- *Upon* in Literary Criticism: making connections between texts, ideas and authors
- **V/ADJ *upon* n**

‘Building’ (usually positive)

Spenser draws upon traditional imagery

*The play was based upon a pamphlet published
in the same year*

*My thinking also builds upon A. Walton Litz's
discussion of Joyce's accretion*

‘Dependency’ (usually negative)

such an interpretation is entirely dependent upon accepting that the household is indeed a private world, securely detached from the public.

it relies rather too heavily upon contemporary critical theory and not enough upon close analysis of the text

The Essay dismisses the works of Godwin and Paley as excessively reliant upon the faculty of 'reason'

Conclusion: why do it this way?

- SGW approach provides analytical depth; complements breadth of coverage provided by n-gram approaches
- Depends on aims of research: are you looking for **phrases**, or **phraseology**?

Conclusion: why do it this way?

deliberation took place against a changing backdrop of military events

It was against this background that abortion was discussed during the 1930s

What of the normative institutional culture of charity to the dead, the background against which Stoeckhlin's idiosyncratic views were drawn?

Words salient in two corpora: discipline-salient

	<i>LitArt</i>	<i>LitRev</i>	<i>HistArt</i>	<i>HistRev</i>
1. AGAINST	NEG 511	NOT 517	713 (+6%)	672
2. DURING	NEG 332	NEG 312	807	997 (+19%)
3. HIMSELF	769 (+24%)	582	NOT 291	NEG 238
4. HIS	8,040 (+12%)	7,060	NEG 3,806	NEG 4,393
5. ONE'S	120 (+17%)	99	NOT 40	NOT 47
6. THEIR	NEG 2,368	NEG 2,320	3,541	4,119 (+14%)
7. THOUGH	502	694 (+28%)	NEG 351	NEG 275
8. UPON	652 (+13%)	568	NOT 279	NOT 227
9. WHICH	4,438	5,024 (+12%)	NEG 3,130	NEG 2,752

Discipline-salient words across the four corpora (per million words).

Words salient in two corpora: genre-salient

	<i>HistRev</i>	<i>LitRev</i>	<i>HistArt</i>	<i>LitArt</i>
1. HOW	1,460 (+19%)	1,176	NEG 554	NEG 721
2. LESS	588 (+4%)	567	NOT 385	NEG 327
3. MORE	2,786 (+2%)	2,734	NEG 1,886	NEG 2,051
4. MOST	1,671 (+1%)	1,651	NOT 1,135	NOT 1,120
5. MUCH	1,242 (+8%)	1,138	NEG 634	NEG 758
6. THAN	1,922	1,930 (+0.5%)	NEG 1,304	NOT 1,491

Genre-salient grammatical words across the four corpora (per million words)

Words salient in one corpus

	<i>HistRev</i>	<i>LitArt</i>	<i>LitRev</i>	<i>HistArt</i>
1. ALTHOUGH	970	NOT 473	NOT 443	NOT 502
2. HER	NEG 1,481	3,914	NEG 2,803	NEG 1,128
3. HERSELF	NEG 62	260	NOT 169	NEG 53
4. MANY	1,350	NEG 701	NOT 1,022	NEG 835
5. SEVERAL	338	NOT 248	NOT 301	NOT 250
6. THOSE	1,051	NEG 771	NOT 907	NOT 908

Grammatical words salient in one corpus

Words salient in three corpora

	<i>HistRev</i>	<i>LitRev</i>	<i>LitArt</i>	<i>HistArt</i>
1. AN	4,081	4,773	4,087	NEG 3,358
2. BY	6,159	NOT 5,944	5,918	5,986
3. DESPITE	362	234	NOT 160	210
4. THAT	11,980	13,199	11,920	NEG 9,236
5. THIS	5,803	5,906	5,218	NEG 4,220
6. THROUGH	1,001	1,064	1,039	NEG 680
7. WHEREIN	10	12	28	NOT 7
8. WHILE	907	937	791	NOT 664
9. WHOM	194	216	263	NOT 167

Grammatical words salient in three corpora