Language development in children's writing from six to sixteen

Phil Durrant

University of Exeter

one luge time ago there was a king colld king james the first and the cathlixs did not like him. and there was a bad man called Guy Fawkes he wantied to bow the houses of Parliament he wantid to cill the king to as well as the cathlixs he had 36 barols of gunpowder and he hid it. Robert Catesby sent a leter to the king.

Dear Sir, I am writing to express my views on the article you recently printed, detailing a scheme by the Divert Trust to help difficult students. At first I was unsure if this scheme could ever work, and was indignant, like so many others, that many good students remained unrewarded. However, after researching this scheme I have come to realise that it is rather a brilliant idea. Research shows that around 88% of schools admit to not being able to cope with difficult students.

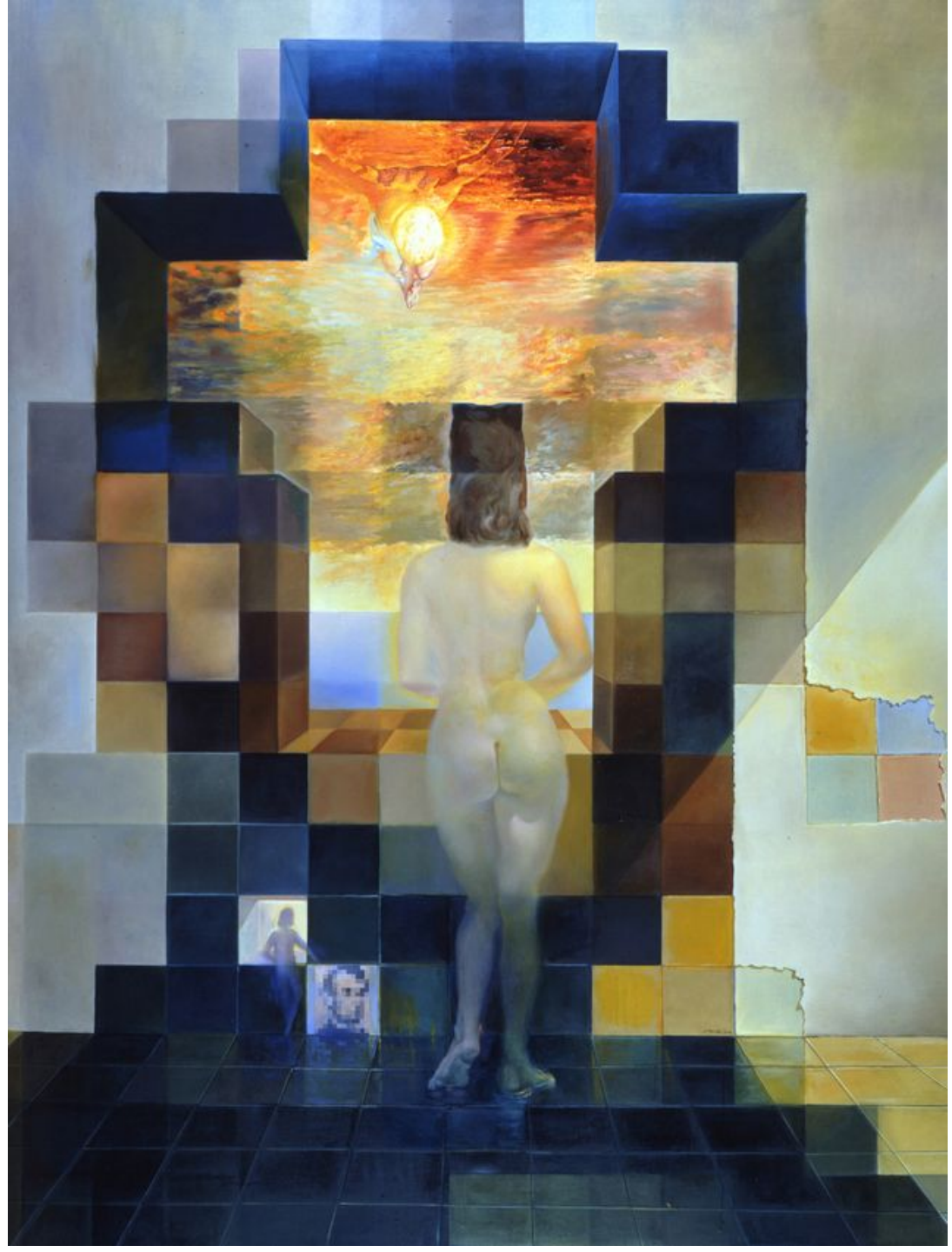
The Growth in Grammar Project

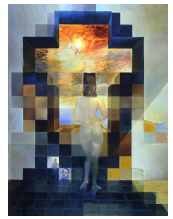
Overview

- Creating a corpus of educationally authentic writing from children in schools across England at Key Stages 1-4
- To be analysed for changes in lexis, phraseology and syntax.
- Corpus to be made publicly available late 2018 .

Methodological foundations: Quantitative text analysis

- Enables us to study large numbers of texts.
- Which enables robust generalizations
- And enables the emergence of patterns which are not obvious in smaller samples…





But...

- Requires transformation of texts to electronic format, so loss of some original features.
- Analysis is limited to features that can be reliably counted.
- Features are decontextualized.

Creating the corpus

- Texts collected from volunteering schools/children
- Classified: Literary vs. Non-literary
- Transcribed/anonymized/normalized

Our corpus

		English	Humanities	Science	Total
Year 2	Literary	258	0	5	263
	Non-literary	277	96	2	375
Year 4	Literary	23	0	0	23
	Non-literary	2	22	2	26
Year 6	Literary	293	0	0	293
	Non-literary	298	106	171	575
Year 9	Literary	220	0	0	220
	Non-literary	305	113	166	584
Year 11	Literary	66	0	0	66
	Non-literary	367	49	58	474
Total		**2110**	**386**	**404**	**2899**

Our corpus

	Schools	Writers	Titles
Year 2	6	160	77
Year 4	2	10	24
Year 6	7	185	78
Year 9	12	457	86
Year 11	9	171	90
Total	**24**	**983**	**351**

Our corpus

Stage	Gender: Female	FSM/PP	EAL
Primary	53.0%	21.7%	21.9%
Secondary	60.0%	22.3%	3.7%

Preparing for vocabulary analysis: CLAWS tagging

Word	POS
Dear	JJ
Editor	NP1
,	,
I	PPIS1
am	VBM
writing	VVG
to	TO
express	VVI
my	APPGE
opposition	NN1

Preparing for analysis: Stanford NLP

Sentence number	Word	POS	Dep. on	Dep.
1	Dear	NNP	2	compound
2	Editor	NNP	6	nsubj
3	,	,	6	punct
4	I	PRP	6	nsubj
5	am	VB	6	aux
6	writing	VB	0	ROOT
7	to	TO	8	mark
8	express	VB	6	xcomp
9	my	PRP$	10	nmod:poss
10	opposition	NN	8	dobj

Preparing for analysis: tagging/parsing

Sentence number	Word	POS	Dep. on	Dep.
1	Dear	adj	2	pre_mod
2	Editor	noun_com	0	voc
3	,			
4	I	pro	6	subj
5	am			
6	writing			
7	to	conj_sub	8	
8	express	verb_lex_act	6	obj
9	my	det	10	
10	opposition	noun_com	8	dobj

Reliability of automated parsing

Feature		IRR
Noun Phrases	per text	.99
	direct dependents per	.81
	words per	.78
Relative clauses	per text	.86
	direct dependents per	.62
	words per	.52
Adjective complement clauses	per text	.38
	direct dependents per	.11
	words per	.04

Vocabulary Analysis

Lexical sophistication

“selection of low-frequency words that are appropriate to the topic and style of the writing, rather than just general, everyday vocabulary”
(Read, 2000:200)

Part 1: Word frequency

Step 1: Get frequencies from a reference corpus

- Corpus of Contemporary American (COCA): 450 million words of English from fiction, newspaper, magazines, academic texts and recorded speech.

A	B	C	D	E	F	G	H	I
WORD	**LEMMA**	**POS**	**TOTAL**	**SPOKEN**	**FICTION**	**MAGAZINE**	**NEWS**	**ACADEMIC**
the	the	at	54124.71	46393.26	53301.68	53775.83	53613.78	63981.74
and	and	cc	26636.86	26089.72	25756.04	26458.18	24577.22	30346.6
of	of	ii	25782.79	21502.94	19640.22	25872.17	23814.03	38260.97
a	a	at1	22240.78	21403.59	22960.81	24395.42	23734.44	18637.48
in	in	ii	17306.2	15433.91	13194.97	17503.23	18490.76	21952.5
to	to	to	15672.75	18518.54	14851.19	15252.34	15091.21	14527.89
to	to	ii	9597.23	8584.14	9325.53	9766.66	9376.73	10973.72
is	be	vbz	9125.93	12536.61	5146.32	9019.08	8862.46	9875.33
that	that	cs	8445.32	11721.96	5711.03	7831.18	7256.74	9563.24
for	for	ii	8157.1	7459.64	6309.93	8729.55	9218.25	9053.34

Step 2: tokenize text and retrieve reference frequencies for each word

Dear Editor, I am writing to express my opposition to the article regarding "teenage tearaways" which was recently published in your newspaper.

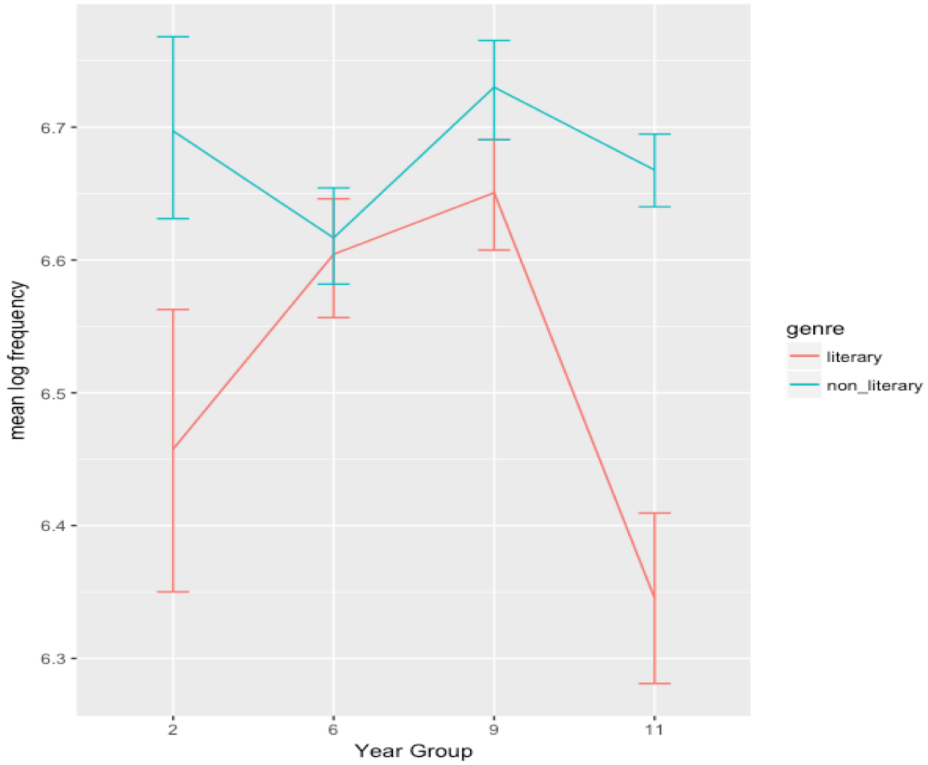
	words	lemmas	pos_gig	counts	
1	words	lemmas	pos_gig	counts	
2	dear	dear	adj	21.02	
3	editor	NA	noun_prop	NA	
4	,	NA	,	NA	
5	i	i	pro	10500.52	
6	am	be	verb	241.88	
7	writing	write	verb	36.32	
8	to	to	conj_sub	16171.84	
9	express	express	verb	25.58	
10	my	my	det	2384.02	
11	opposition	opposition	noun_com	49.95	
12	to	to	prep	9902.85	
13	the	the	det	55848.28	
14	article	article	noun_com	107.00	
15	regarding	regarding	prep	44.59	
16	<quotemark>	NA	<quotemark>	NA	
17	teenage	teenage	adj	15.34	
18	tearaways	NA	noun_com	NA	
19	<quotemark>	NA	<quotemark>	NA	

Step 3

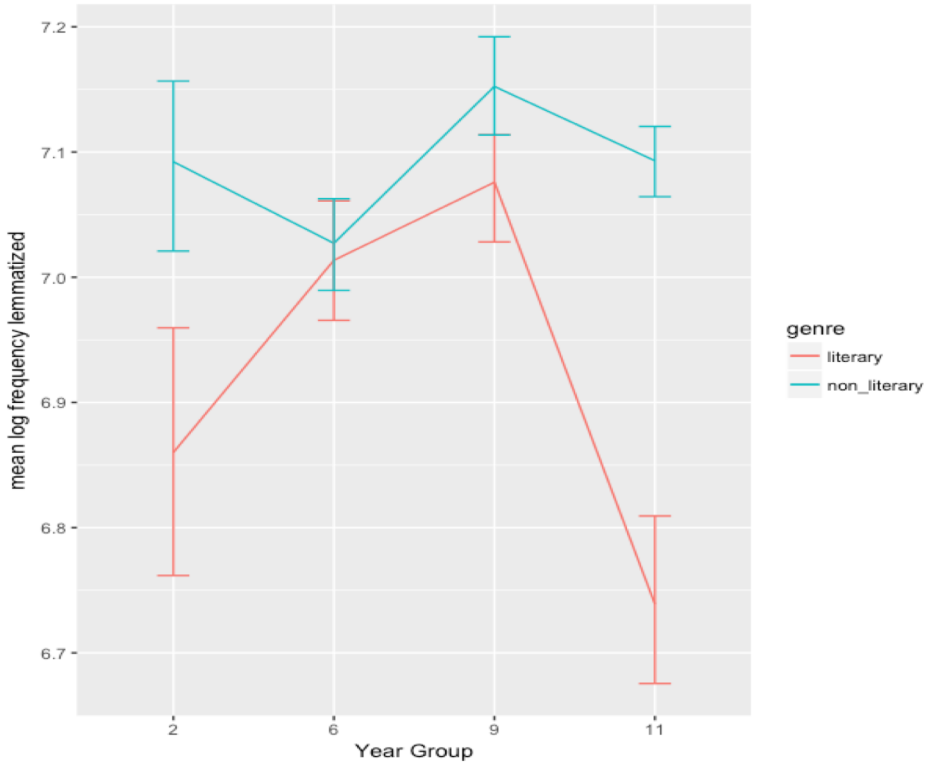
- For each text:
  - Find mean log frequency of all lexical tokens (noun/verb/adj/adv)
  - Repeat for:
    - Lemmatized tokens
    - Word types
    - Lemmatized word types

Tokens

Non-lemmatized

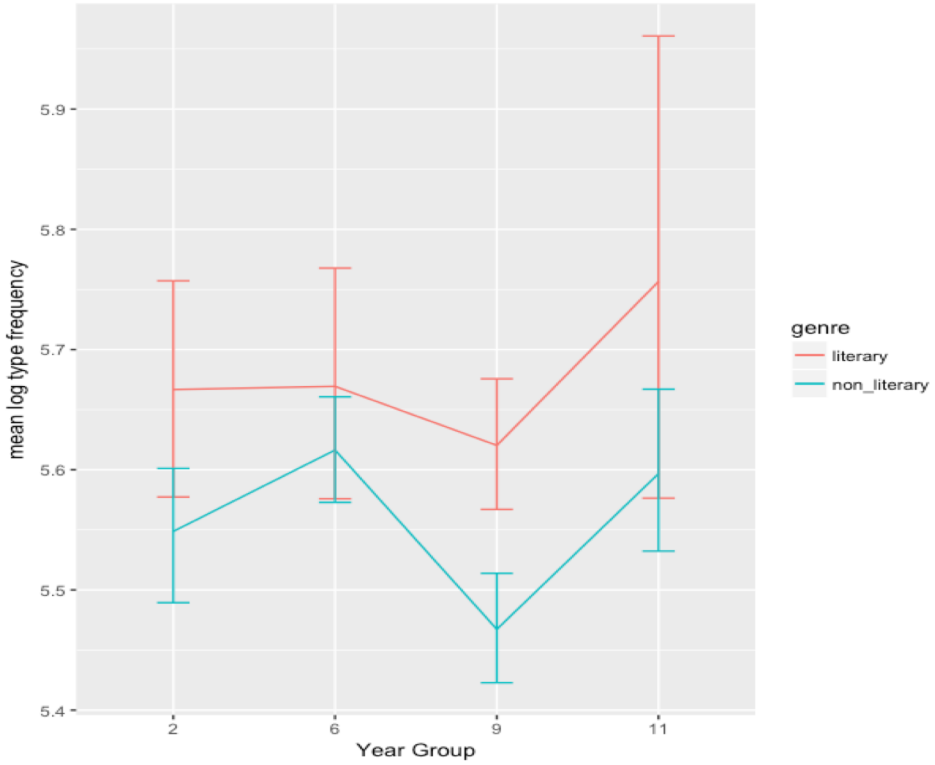


Lemmatized

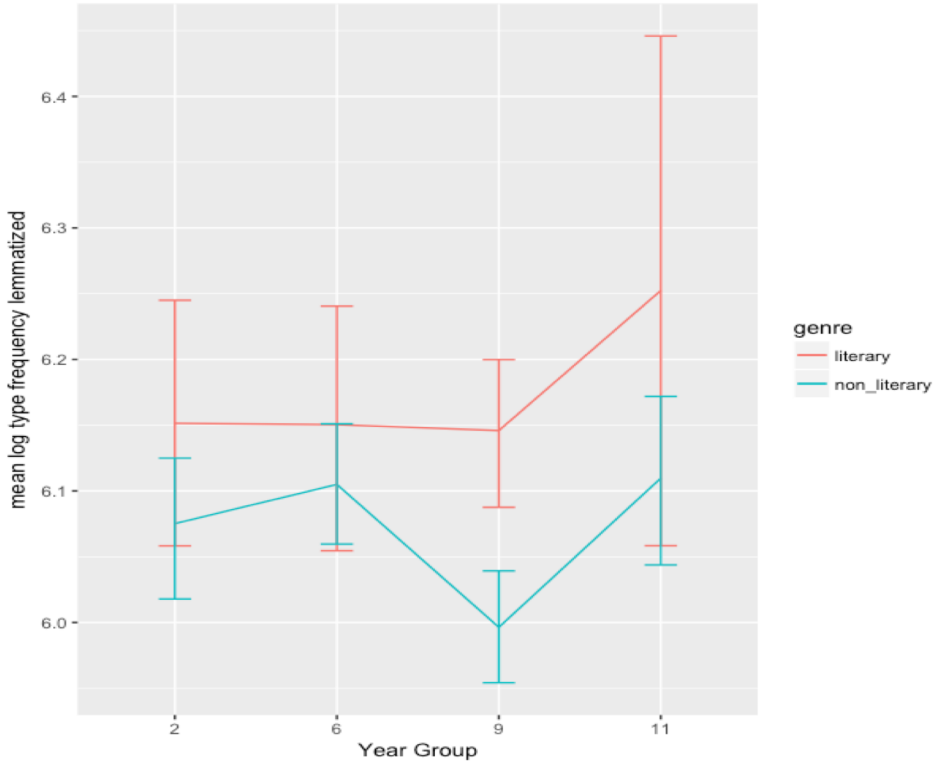


Types

Non-lemmatized



Lemmatized

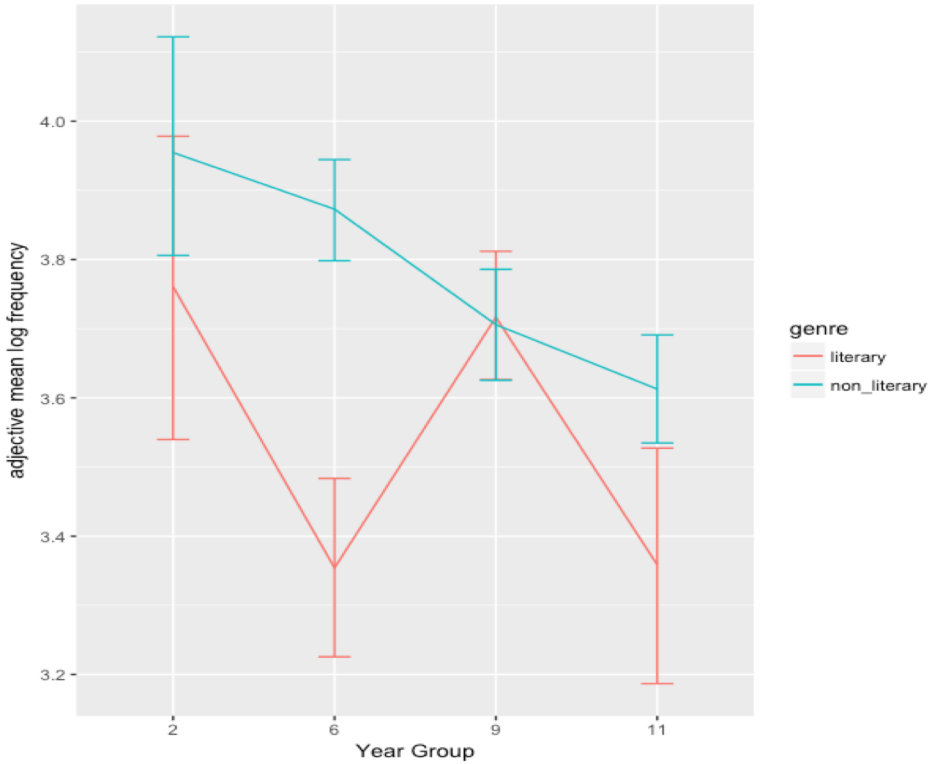


Conclusions 1

Overall word frequencies do not vary across genres or year groups

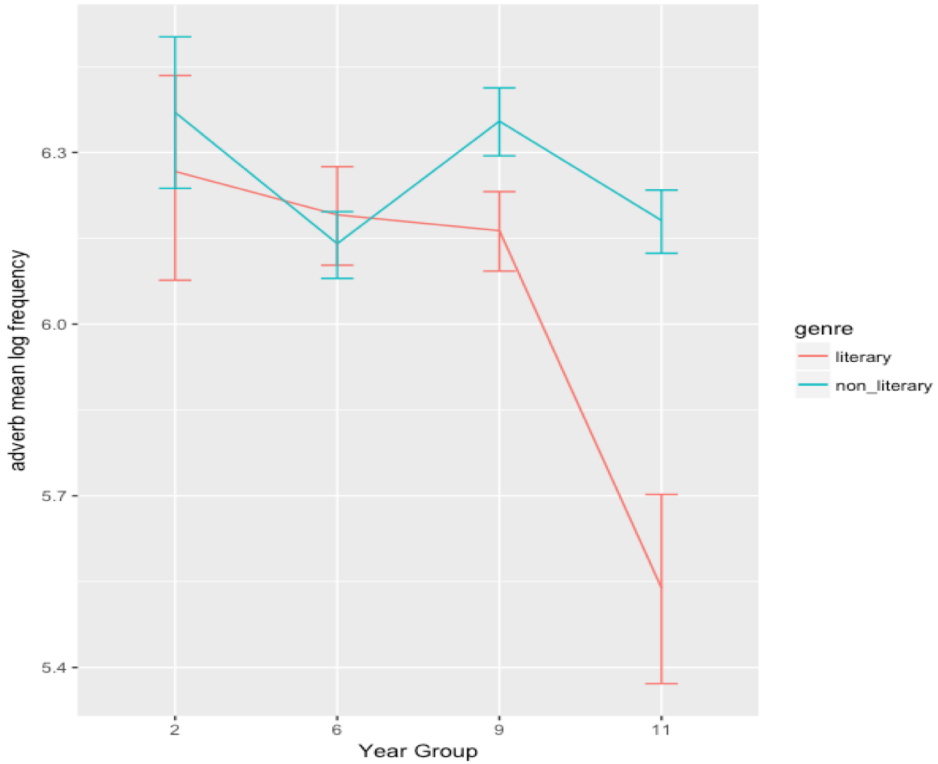
Non-lemmatized tokens by POS

Adjectives



Year: $t(1738)=-4.51$, $p<.0001$
Genre: $t(1738)=3.34$, $p<.001$*

Adverbs

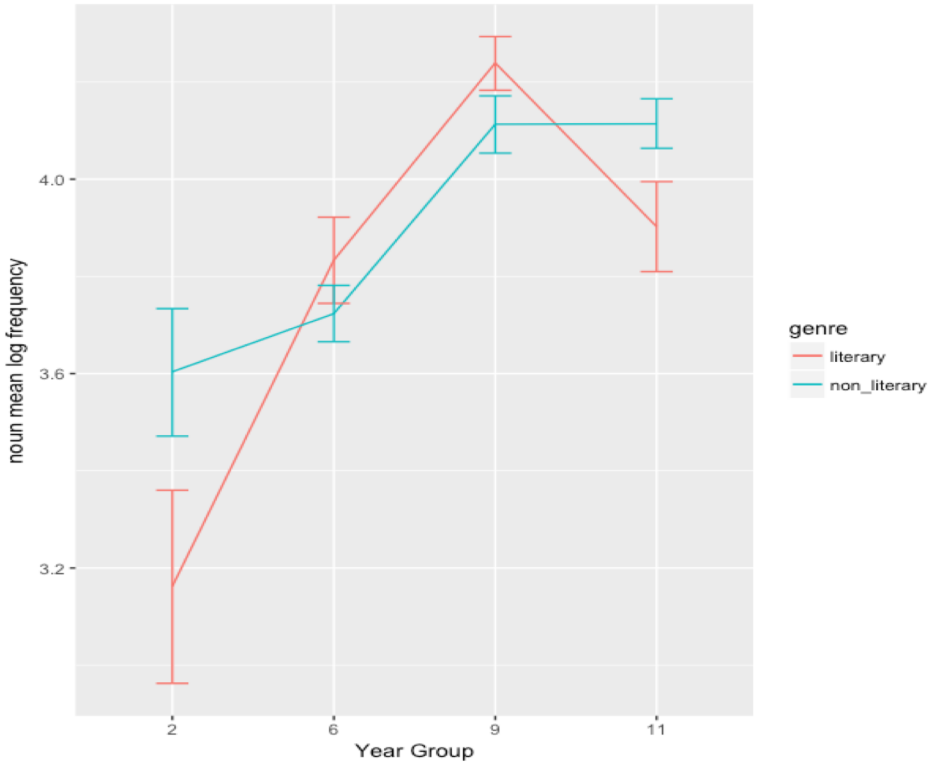


Year: $t(1743)=-2.45$, $p<.05$
Genre: $t(1743)=3.93$, $p<.0001$*

* Random intercepts for text topic

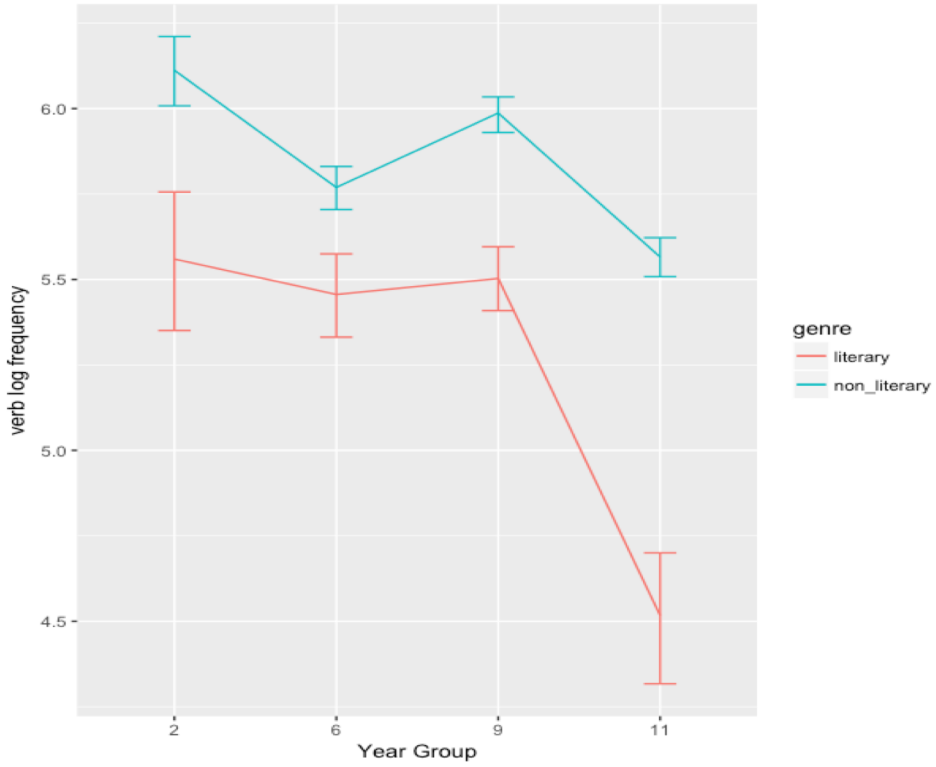
Non-lemmatized tokens by POS

Nouns



Year: $t(1765)=3.40$, $p<.0001$*

Verbs



Year: $t(1764)=-5.49$, $p<.0001$
Genre: $t(1764)=6.76$, $p<.0001$*

* Random intercepts for text topic

Typical Year 2 texts: use of nouns

Non-literary

*I am writing to you about the **sea** **turtle** because they are not safe. **People** like the **fishermen** throw **nets** in the **sea** and if **turtles** get stuck in them they will ILLEGIBLE_TEXT die. Also that can happen to other **animals** in the **sea** and at the **beach**. **People** throw **pollution** in the **sea** but they are hunted for their **shells** and they're killed just for their **shells**. They've been alive since the **dinosaurs** were alive and **hunters** can kill them…*

Literary

*One **day** on a stormy, wet, cold, **morning** Rosie saw her first red **fairy**. She looked up and saw a magical **fairy**. She let the red **fairy** in the **house**. After **playtime** the more Rosie looked the more **book** **fairies** she saw in the **sky** but no one else noticed them. Rosie went into the Institutionname **class room** and in the **trays** she found a **map** to **fairyland**.*

Typical Year 11 texts: use of nouns

Non-literary

*At the **start** of the **play** when the **inspector** is hinting at the **accusation** that the **family** is responsible for the **death** of Eva Green. Sheila immediately questions the **inspector**. Saying "you talk as of we are responsible". This comes across childish and it's as if it's almost impossible that they are related to the **incident**. Before that when Gerald is proposing to her, a big **moment** in anyone's **life**, she doesn't seem to take it very seriously.*

Literary

*The monotonous, shrill **screech** of the **alarm clock** brought me to my **senses**, as I wearily stumbled out of my **bed** and into the **bathroom**. A **shroud** of **darkness** lingered outside, accompanied by the persistent **patter** of **rain**. As I looked through the **window**, dark **clouds** slowly circled around, menacing and patient. I caught the 7:21 **train** on platform 3, like I do every **day** and the familiar **scent** was oddly welcoming.*

Typical Year 2 texts: use of verbs

Non-literary

*Dear Romeo. I **am going** to **drink** a special medicine but I **will sleep** for 2 days. Here **is** the plan. Oh Romeo **did**n't **leave** any poison for me. Then we **can get married** . Juliet **loves** Romeo and **wants** to **marry** him too so they **decided** not to **fight**. But Juliet's family **thinks** that's Juliet **is** dead. Romeo **heard** that Juliet **is** dead. Romeo **is** so upset.*

Literary

*The creatures of the glittery blue sea **loved** their home. The glittery blue sea **was covered** with fishes, corrals, sea fish, starfish and seaweed. The corrals **were** as purple as a flower. Everybody **lived** happily together. That **was** until the bad mermaid Emilia **arrived**. She **travelled** on her own.*

Typical Year 11 texts: use of verbs

Non-literary

*"Good evening ladies and gentlemen", he **announced**. "We'**re** tonight's entertainment". Carelessly, he **shoved** food into his mouth. "Where'**s** Harvey Dent" he **shouted**, spitting shrimp everywhere. **Using** his arrogant personality, the Joker **tries** to **intimidate** his audience. **Grabbing** a civilian's face because he **was**n't **threatened** by him.*

Literary

*I **repeated** my investigation, because it **helps** me to **compare** my results with other results, also to **see** if I **had** any anomalies in my answer, so I **would get** a more accurate result. This **meant** I could **work** out the mean by **adding** all my results together **dividing** by the number there **is**, **ignoring** all my anomalous results **creating** an accurate result.*

Conclusions 2

- Frequency of all POS except nouns:
  - decreases with age
  - Is lower in literary than in non-literary texts
- Frequency of nouns:
  - increases with age
  - does not vary across genres
- Analysis of types does not show variation across ages or genres

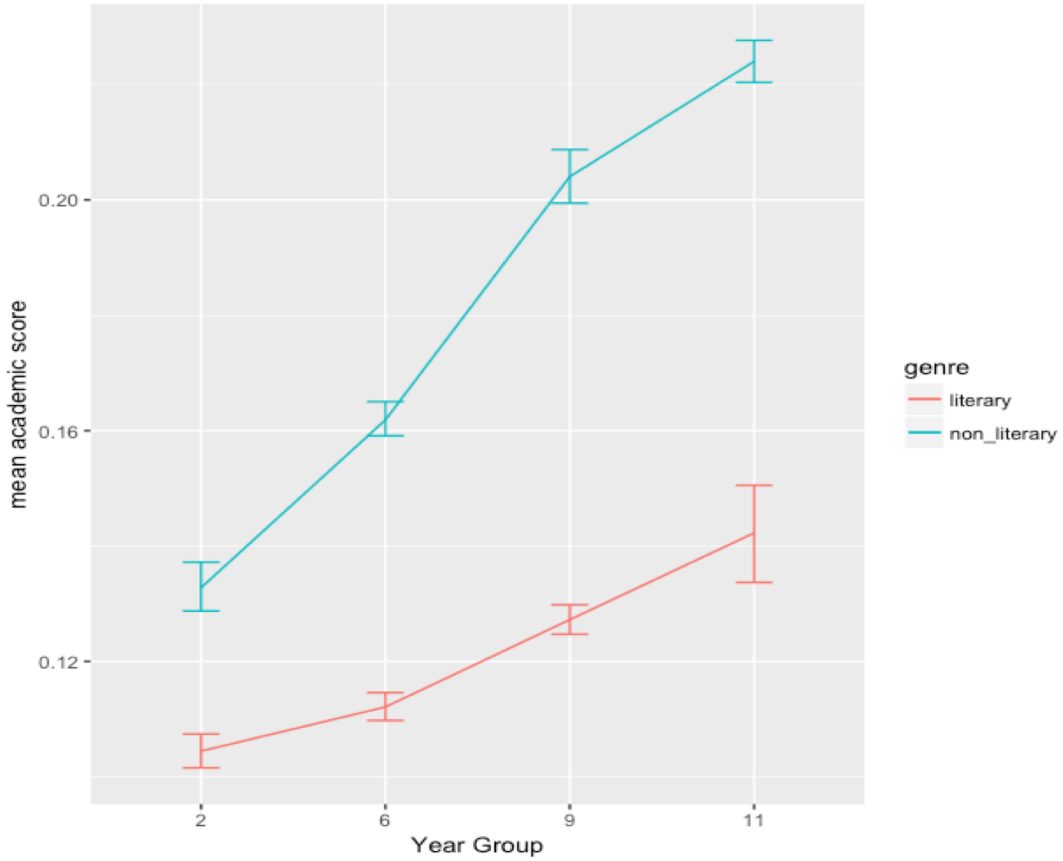
Part 2: Appropriateness

Quantifying appropriateness

A	B	C	D	E	F	G	H	I
WORD	**LEMMA**	**POS**	**TOTAL**	**SPOKEN**	**FICTION**	**MAGAZINE**	**NEWS**	**ACADEMIC**
the	the	at	54124.71	46393.26	53301.68	53775.83	53613.78	63981.74
and	and	cc	26636.86	26089.72	25756.04	26458.18	24577.22	30346.6
of	of	ii	25782.79	21502.94	19640.22	25872.17	23814.03	38260.97
a	a	at1	22240.78	21403.59	22960.81	24395.42	23734.44	18637.48
in	in	ii	17306.2	15433.91	13194.97	17503.23	18490.76	21952.5
to	to	to	15672.75	18518.54	14851.19	15252.34	15091.21	14527.89
to	to	ii	9597.23	8584.14	9325.53	9766.66	9376.73	10973.72
is	be	vbz	9125.93	12536.61	5146.32	9019.08	8862.46	9875.33
that	that	cs	8445.32	11721.96	5711.03	7831.18	7256.74	9563.24
for	for	ii	8157.1	7459.64	6309.93	8729.55	9218.25	9053.34

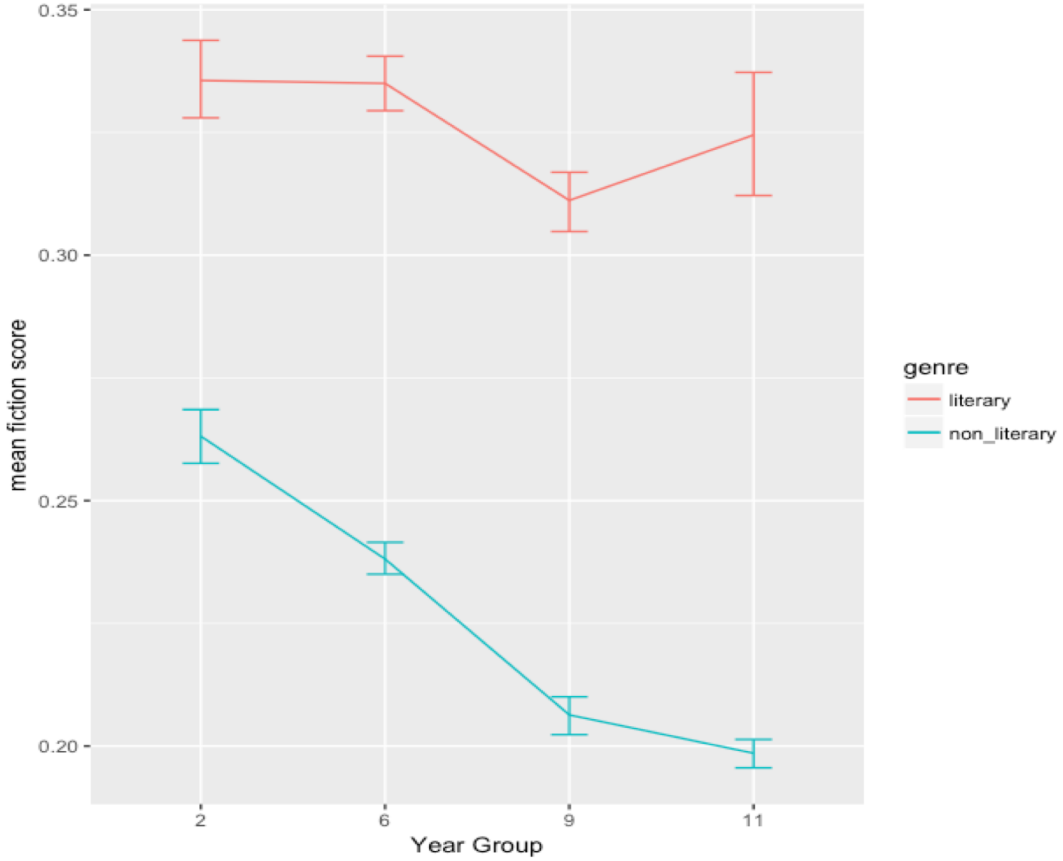
Word Form	POS	Academic	Fiction	Magazine	News	Spoken
the	article	.24	.20	.20	.20	.17
and	conjunction	.23	.19	.20	.18	.20
shuddered	verb	.01	.90	.06	.02	.01
tunelessly	adverb	.03	.90	.08	.00	.00
metacognitive	adverb	1.00	.00	.00	.00	.00
reflectivity	noun	.98	.00	.02	.00	.00

Academic



Year: $t(826)=7.32$, $p<.0001$
Genre: $t(1194)=-3.23$, $p<.0001$
Year x Genre: $t(1194)=10.63$, $p<.0001$*

Fiction



Year: $t(826)=-4.86$, $p<.0001$
Genre: $t(1194)=-11.98$, $p<.0001$
Year x Genre: $t(1194)=-7.19$, $p<.0001$*

* Random intercepts for writer

Conclusions 3

- Children's vocabulary becomes more 'academic-like' and less 'fiction-like' over time.
- Increase in academic words is especially strong in non-literary writing.
- Movement away from fiction words is only found in non-literary writing.
- Appears to show steadily-growing greater register awareness with age

Collocation Analysis

Introduction to collocations

- Pairs of words which frequently co-occur in text, e.g.:
  - *evidence suggests; economic growth; most important*
- Appear to be an indicator of development in L2. e.g.:
  - Bestgen & Granger, 2014; Paquot, 2017)

Quantifying ‘frequent co-occurrence’

- Frequency:
  - *however is; also be; even be*
- Hypothesis-testing measures:
  - *most important; as much; many people; take part*
- Information measures:
  - *pathetic fallacy; evoke pity; draw attention; unconscious mind*
- Directional measures:
  - *best<-fit; local<-resident; grow<-uncontrollably; give<-insight*

Counting collocations

- Span-based approach: e.g. 4-words to left and right
- But:
  - *The old **dream** of wireless communication through space has now been **realized***
  - *She **realizes** that the buzzing sound from her **dream** is still present in her bedroom.*
- Dependency-based approach…

Sentence number	Word	POS	Dep. on	Dep.
1	Dear	NNP	2	compound
2	Editor	NNP	6	nsubj
3	,	,	6	punct
4	I	PRP	6	nsubj
5	am	VB	6	aux
6	writing	VB	0	ROOT
7	to	TO	8	mark
8	express	VB	6	xcomp
9	my	PRP$	10	nmod:poss
10	opposition	NN	8	dobj

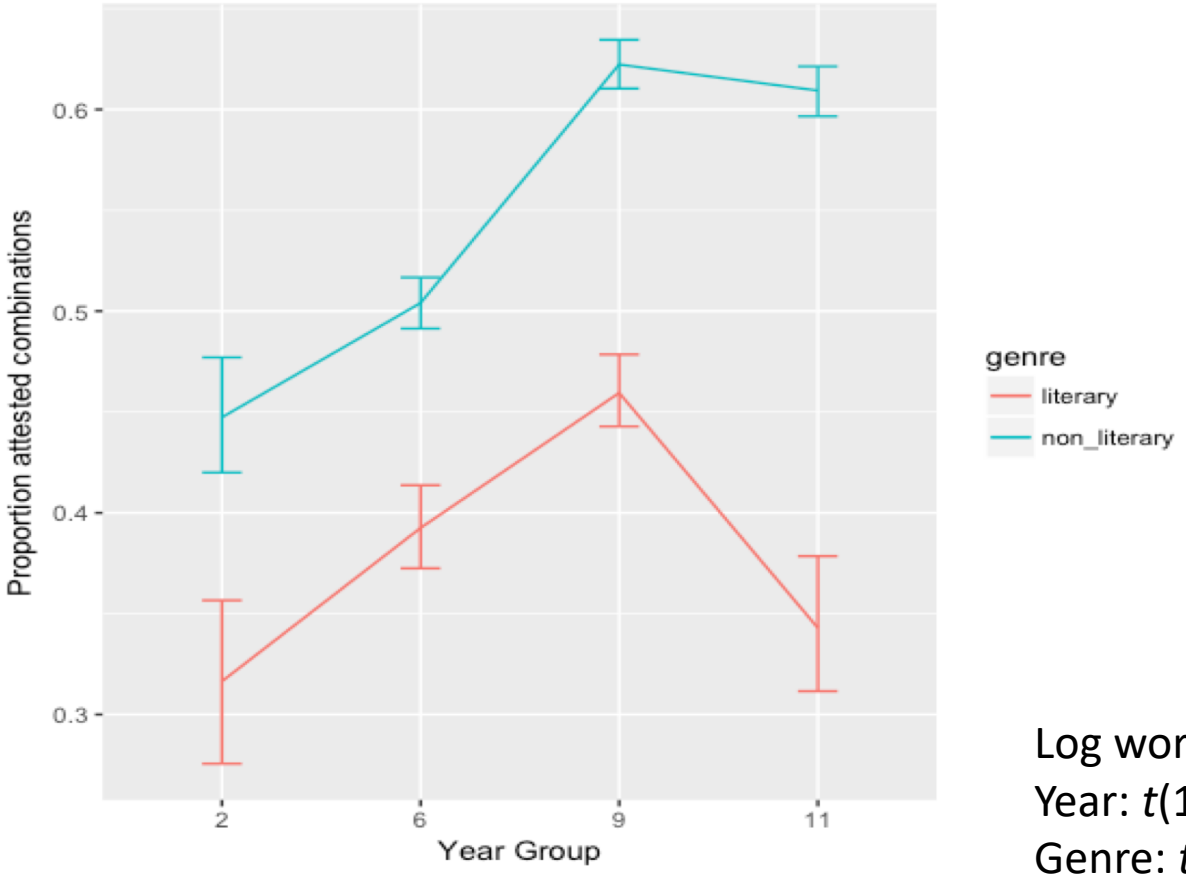
The study

- Focus on combinations of lexical words:
  - Adjective modifying a noun
  - Adverb modifying an adjective
  - Adverb modifying a verb
  - Noun as subject of a verb
  - Noun as object of a verb
- With words lemmatized, e.g.:
  - *argue strongly; argues strongly; arguing strongly* all counted as the same
- Collocation information from BAWE corpus

For example

Item	Frequency	Log Frequency	MI	MI2	t-score	Delta-P (Coll-Node)	Delta-P (Node-Coll)
teenage tearaway	0	0	NA	NA	NA	NA	NA
considerable bias	0	0	NA	NA	NA	NA	NA
own opinion	13	2.56	4.77	8.47	3.47	0.00	0.01
bad behaviour	5	1.61	0.12	2.44	0.17	0.00	0.00

Findings 1: Proportion of attested combinations



Log word frequency: $t(1740)=6.39$, $p<.001$
Year: $t(1740)=-4.8$, $p<.0001$
Genre: $t(1740)=8.29$, $p<.0001$*

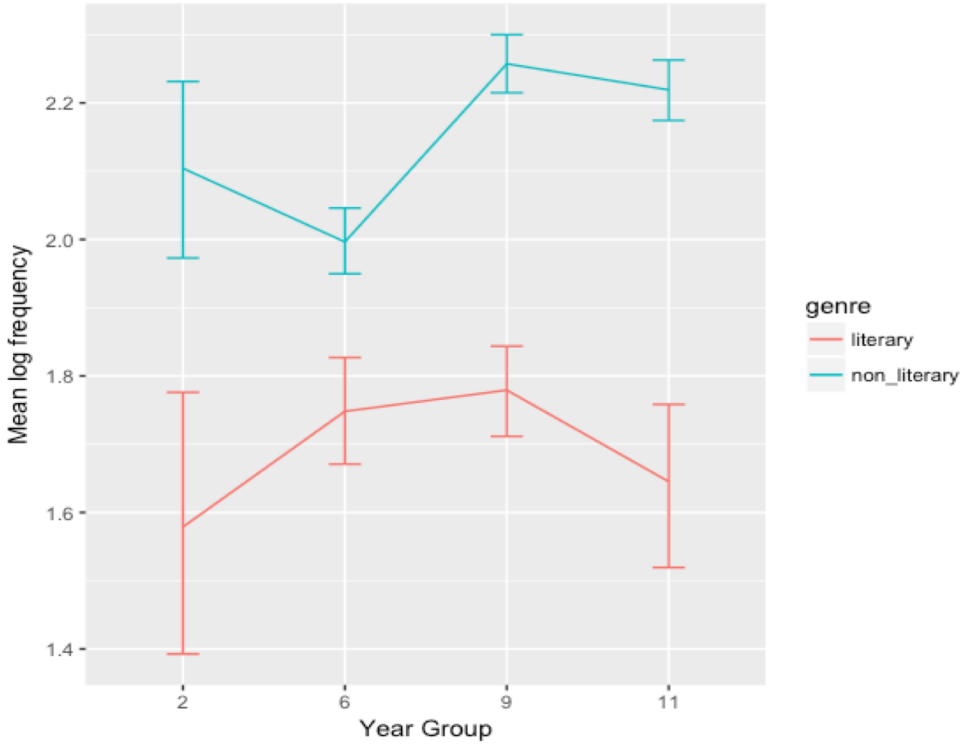
* Random intercepts for title

Comparing quantitative measures

	Log Frequency	MI	MI2	t-score	Delta-P (Coll-Node)	Delta-P (Node-Coll)
Log Frequency	1.00					
MI	.11	1.00				
MI2	.56	.86	1.00			
t-score	.60	.74	.89	1.00		
Delta-P (Coll-Node)	.30	.81	.81	.79	1.00	
Delta-P (Node-Coll)	.31	.82	.81	.81	.55	1.00

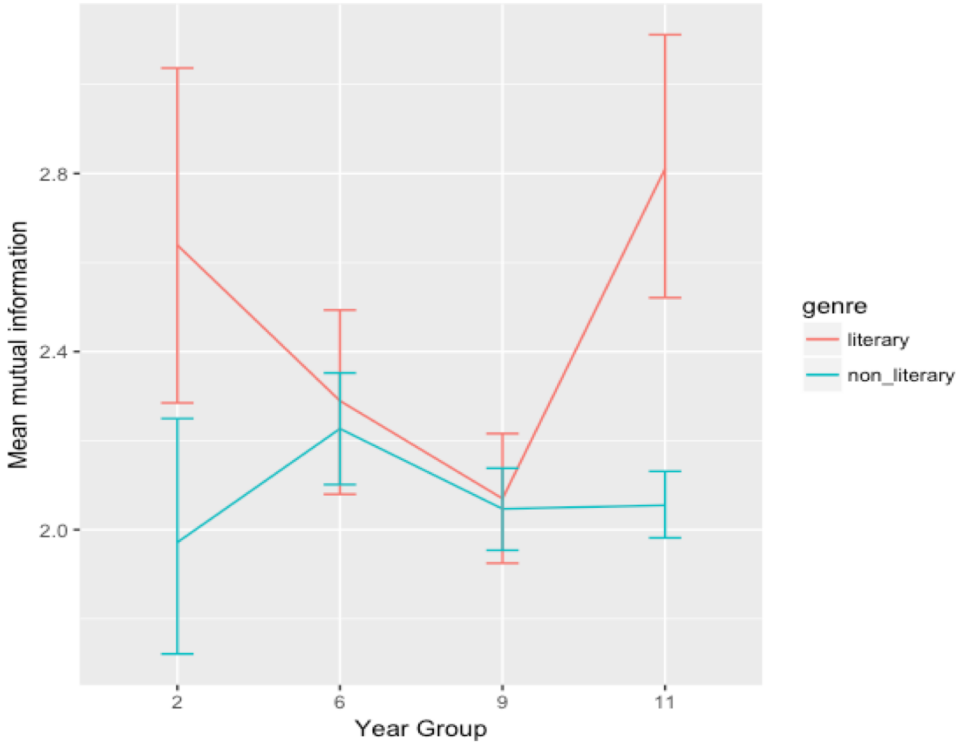
Findings 2: log frequency & MI

Log frequency



Log word frequency: $t(1740)=23.72$, $p<.001$
Year: $t(1740)=2.63$, $p<.01$
Genre: $t(1740)=4.63$, $p<.0001$*

MI



Log word frequency: $t(1740)=-13.97$, $p<.0001$
Year: $t(1740)=.04$, $p>.05$
Genre: $t(1740)=-1.00$, $p>.05$*

* Random intercepts for title

Conclusions 4

- Use of ‘academic’ collocations distinguishes non-literary from literary writing.
- As children mature:
    - Overall use of such collocations increases
    - Differentiation between genres becomes stronger
- No evidence that association measures are developmentally relevant

Summary

- GiG corpus of children’s writing will be available late 2018

Summary

- Development in lexical sophistication:
  - Mean log frequency of words varies across year groups, but:
    - Development differs across different parts-of-speech
    - Developments is seen to token counts only: repetition is influencing results
  - Children's vocabulary becomes more 'academic-like' and less 'fiction-like' over time, but:
    - Increase in academic words is especially strong in non-literary writing
    - Movement away from fiction words is only found in non-literary writing

Summary

- Development in phraseology:
  - ‘Academic’ collocations distinguish children’s genres
  - Children use more academic collocations as they progress through school
  - Children’s use of collocations becomes more genre-sensitive.
  - Association measures do not appear to be developmentally relevant

Thank you!

p.l.durrant@exeter.ac.uk

phildurrant.net

@growinggrammar