Corpus Research on the Development of Children's School Writing

Phil Durrant

University of Exeter

UNIVERSITY OF EXETER

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

That morning a meerkat mob was snoring there heds of Suddenly a snaka slithed into the brow. The snaka saw the baby. The baby ran to the mum, the mum froo the snak. Next a jackal ran in to Sunny. The jackal sede can I be your frend? And they play together tag. Sunny a vitid him for tea time the end.

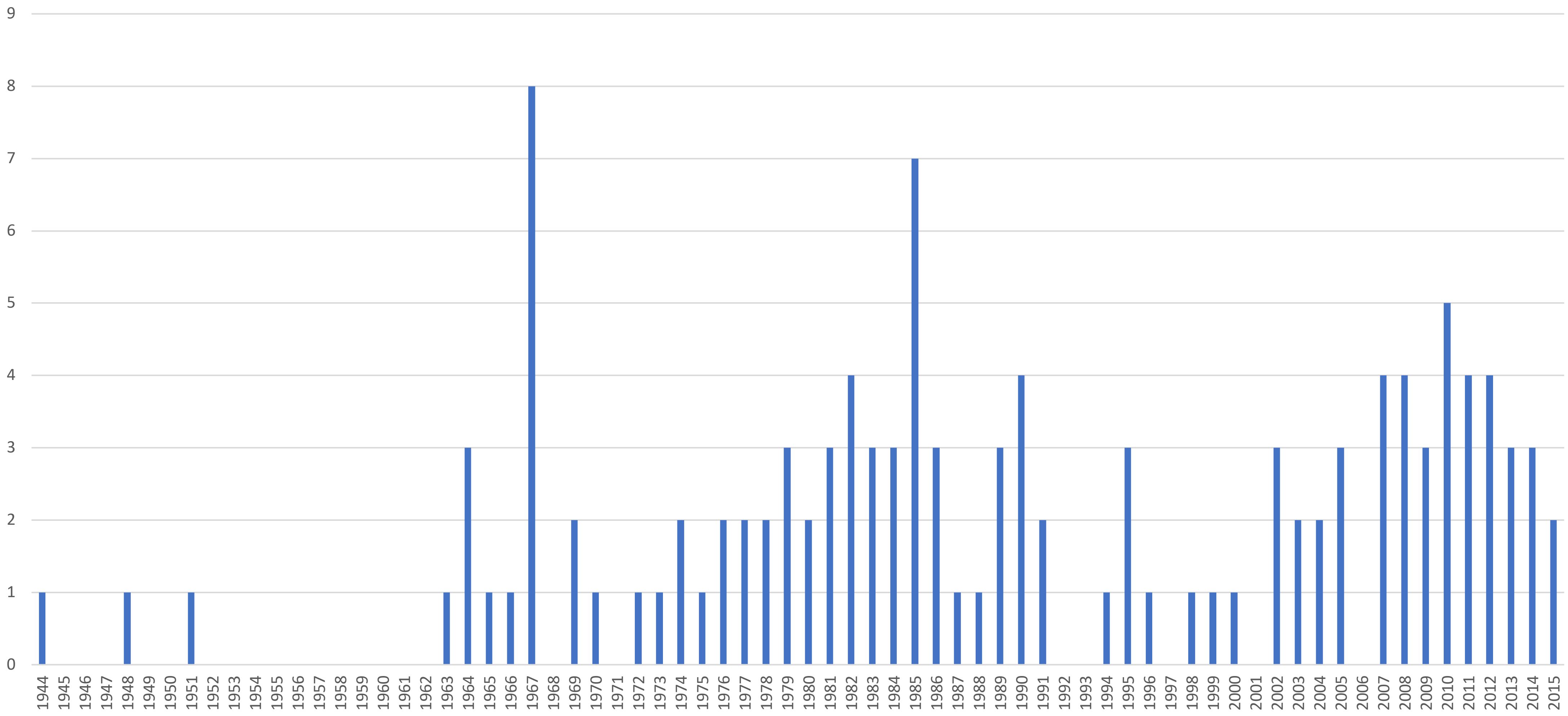
one luge time ago there was a king colld king james the first and the cathlixs did not like him. and there was a bad man called Guy Fawkes he wantied to bow the houses of Parliament he wantid to cill the king to as well as the cathlixs he had 36 barols of gunpowder and he hid it. Robert Catesby sent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

Historical Overview

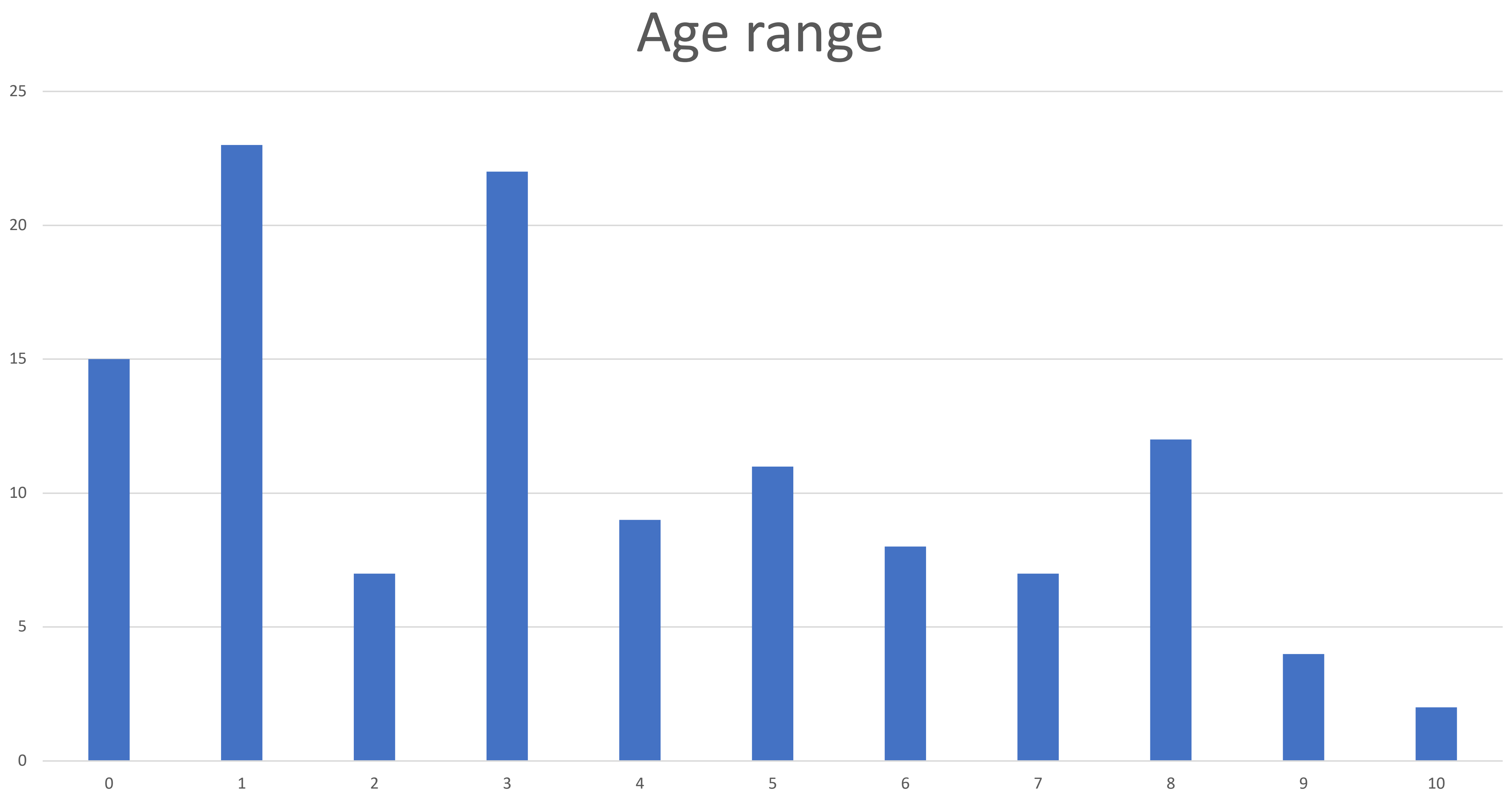
Quantitative comparisons of linguistic features in mainstream children’s writing

Studies per year: 1944-2015
(Total 120 studies)

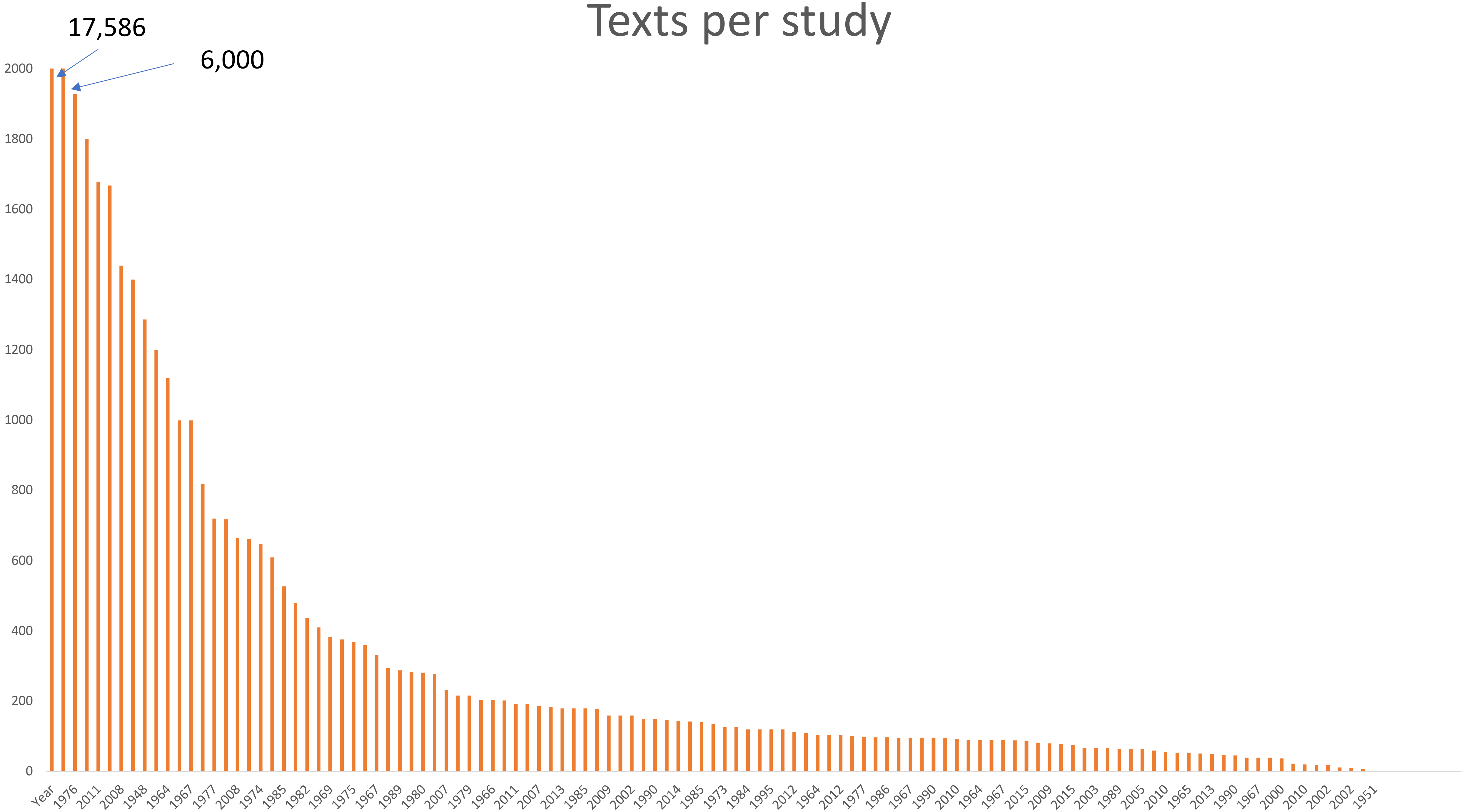


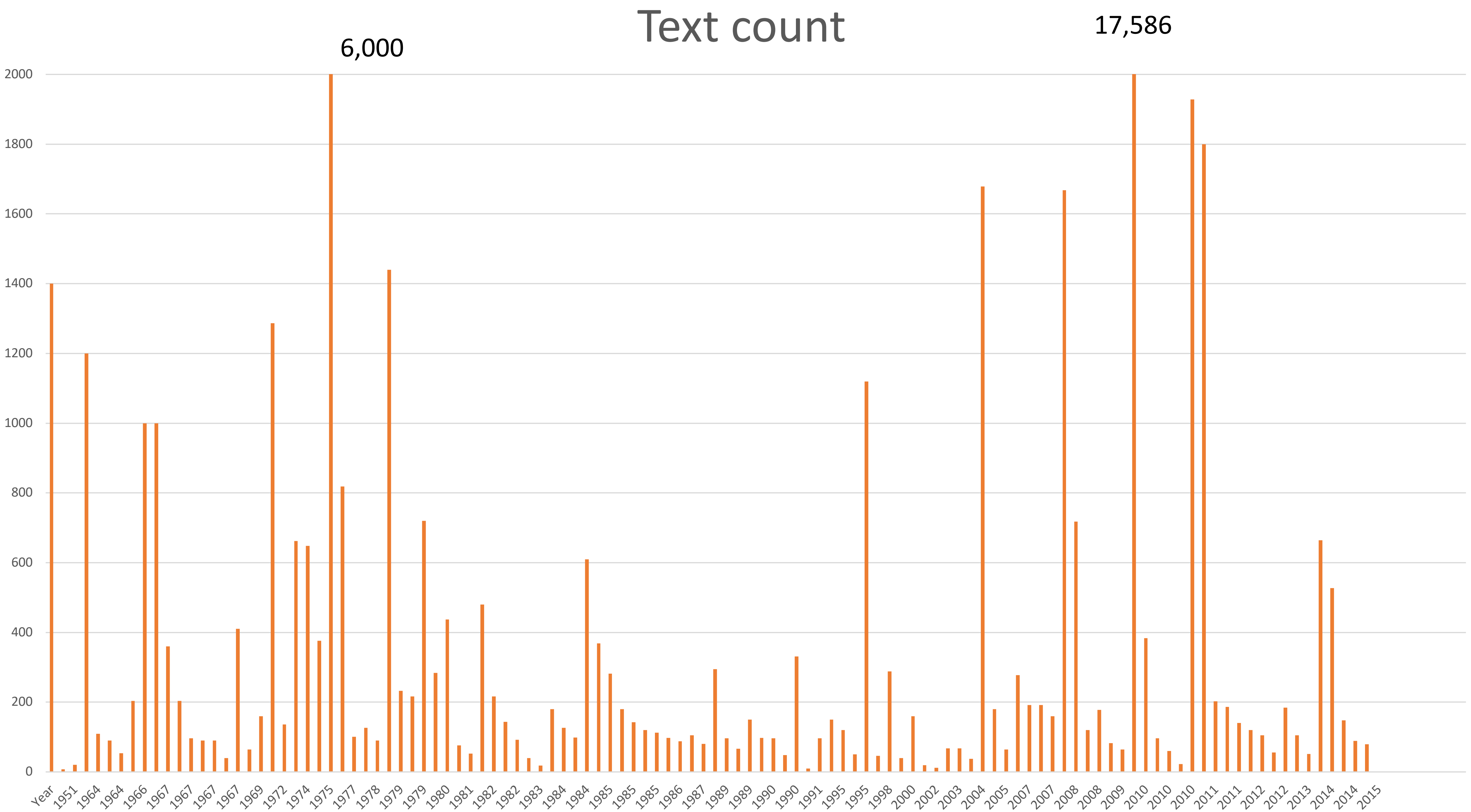
Provenance

Country	Number of Studies
USA	77
UK	22
Canada	16
Australia	2
USA, UK & New Zealand	2
USA & New Zealand	1


Age range
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10

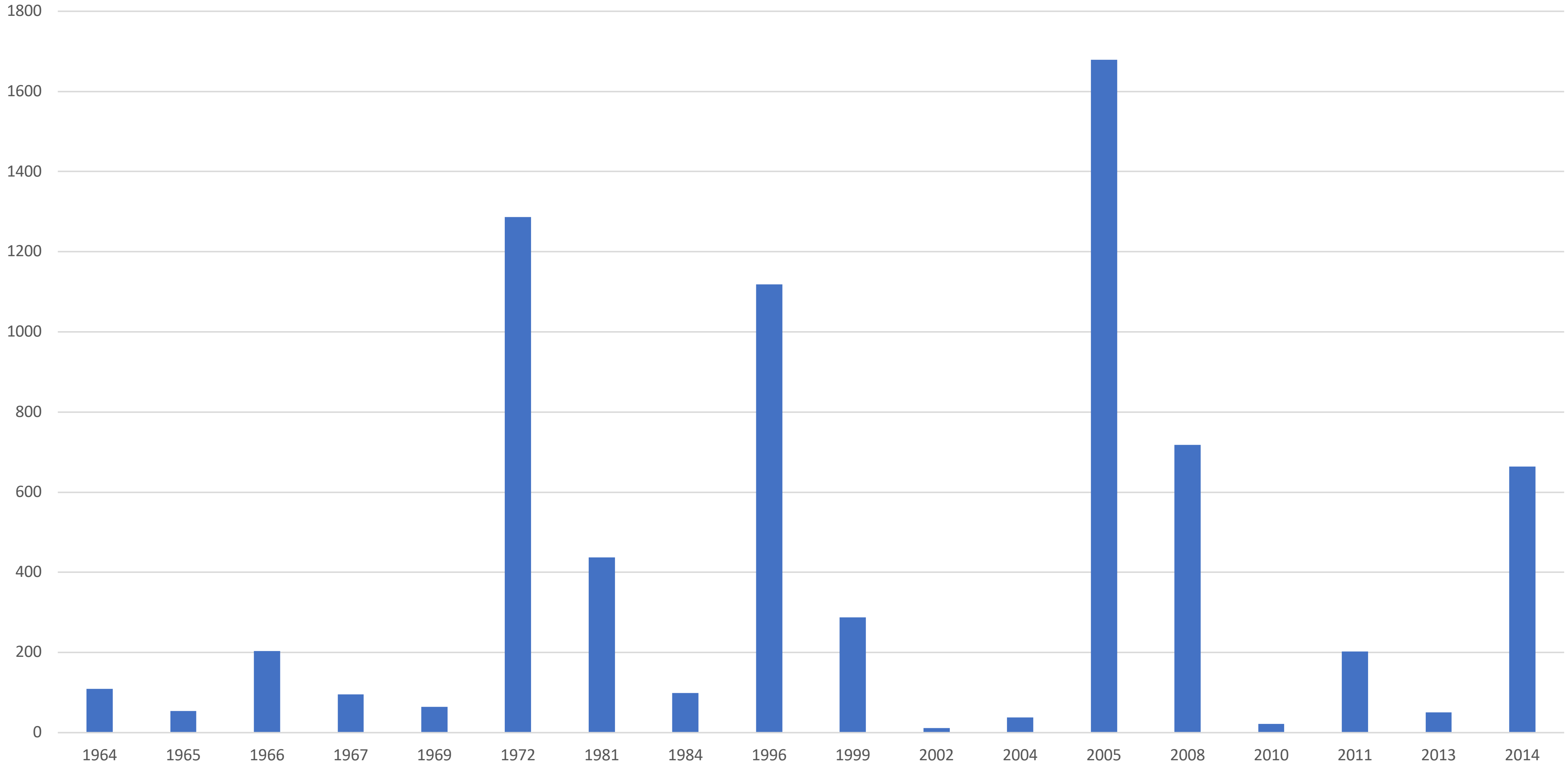
Texts per study



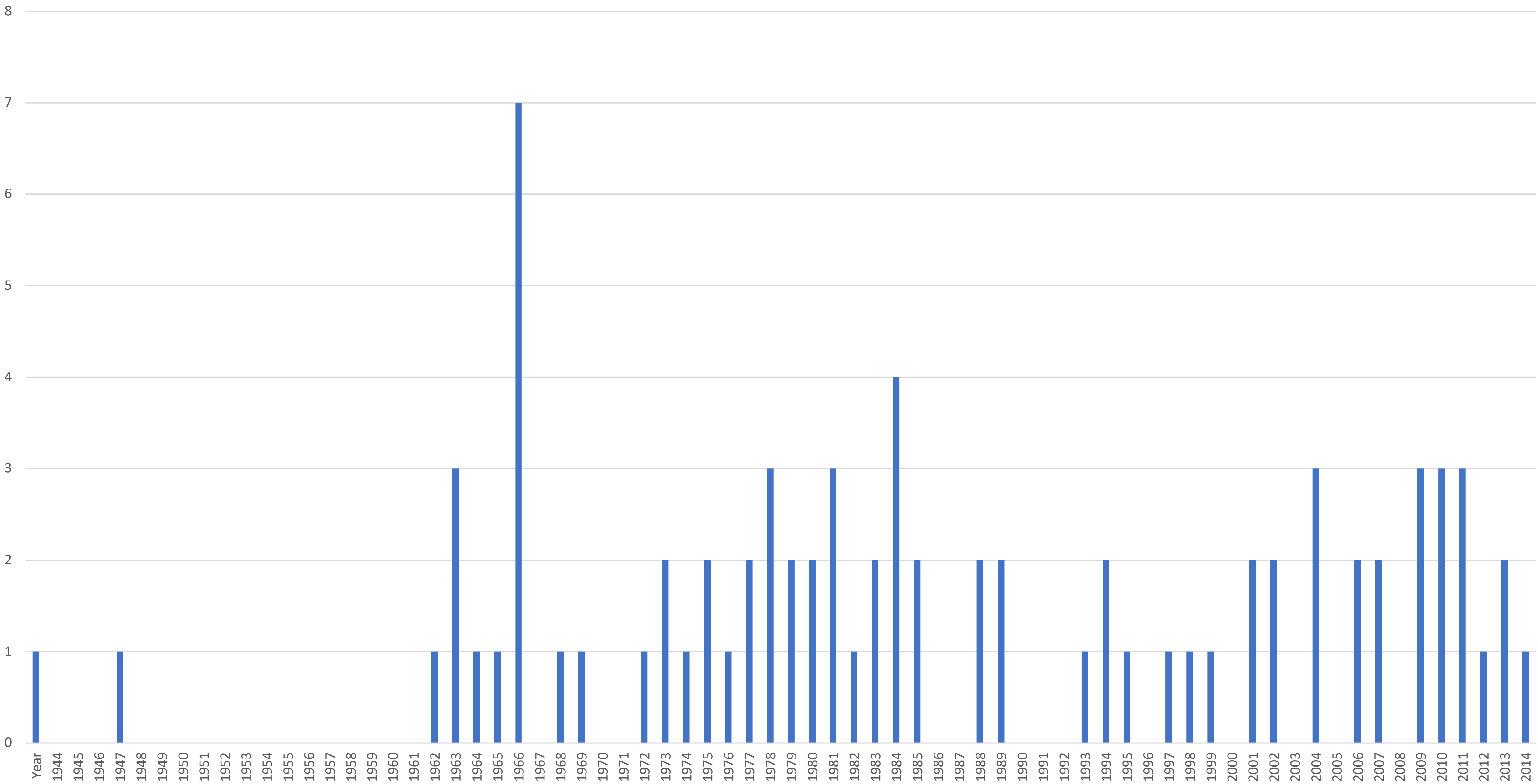


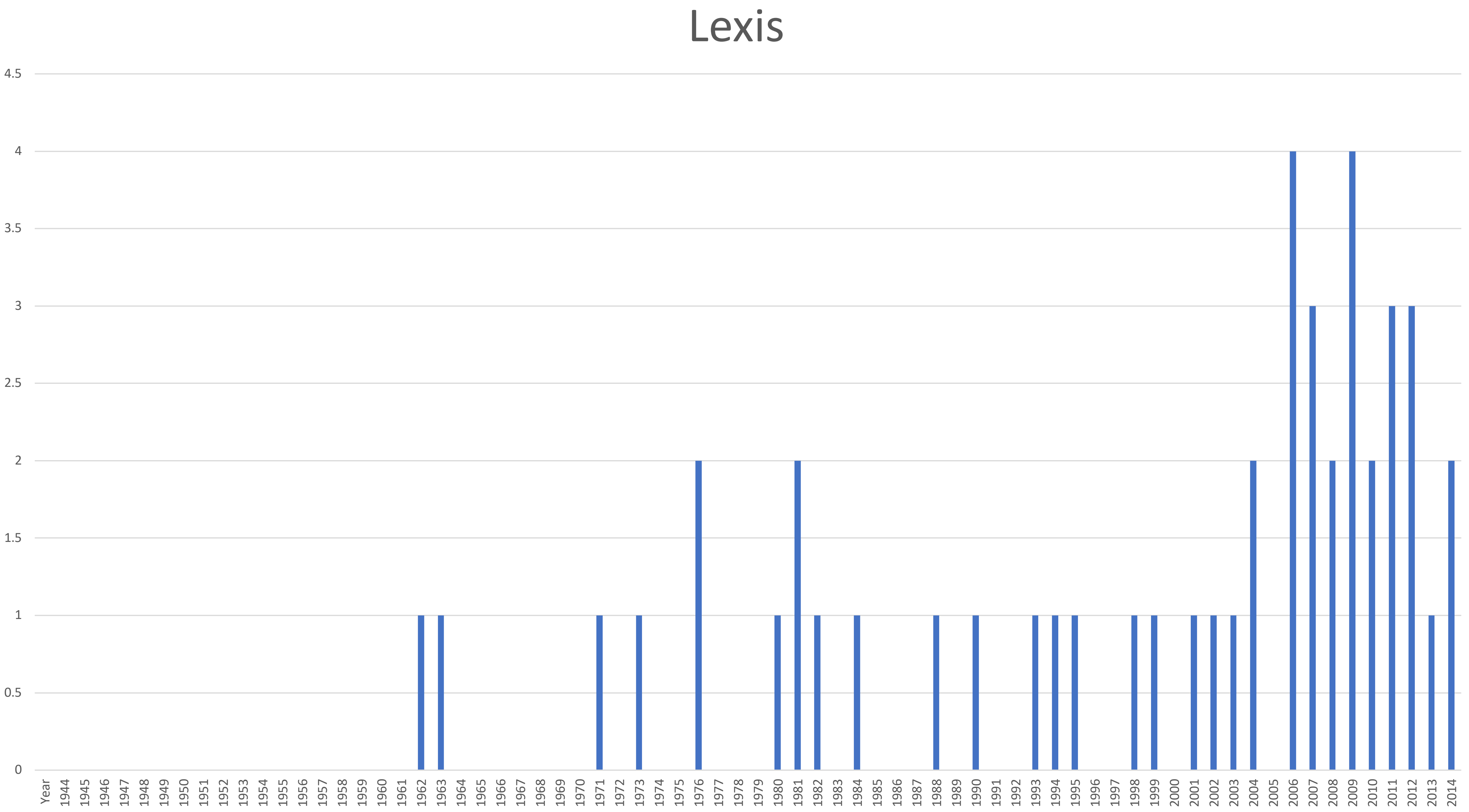
Text count
6,000
17,586
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Year
1951
1964
1964
1966
1967
1967
1967
1967
1969
1972
1974
1975
1977
1978
1979
1979
1980
1981
1982
1982
1983
1984
1984
1985
1985
1985
1986
1987
1989
1989
1990
1990
1991
1995
1995
1998
2000
2002
2003
2004
2005
2007
2007
2008
2008
2009
2010
2010
2010
2011
2011
2012
2012
2013
2014
2014
2015

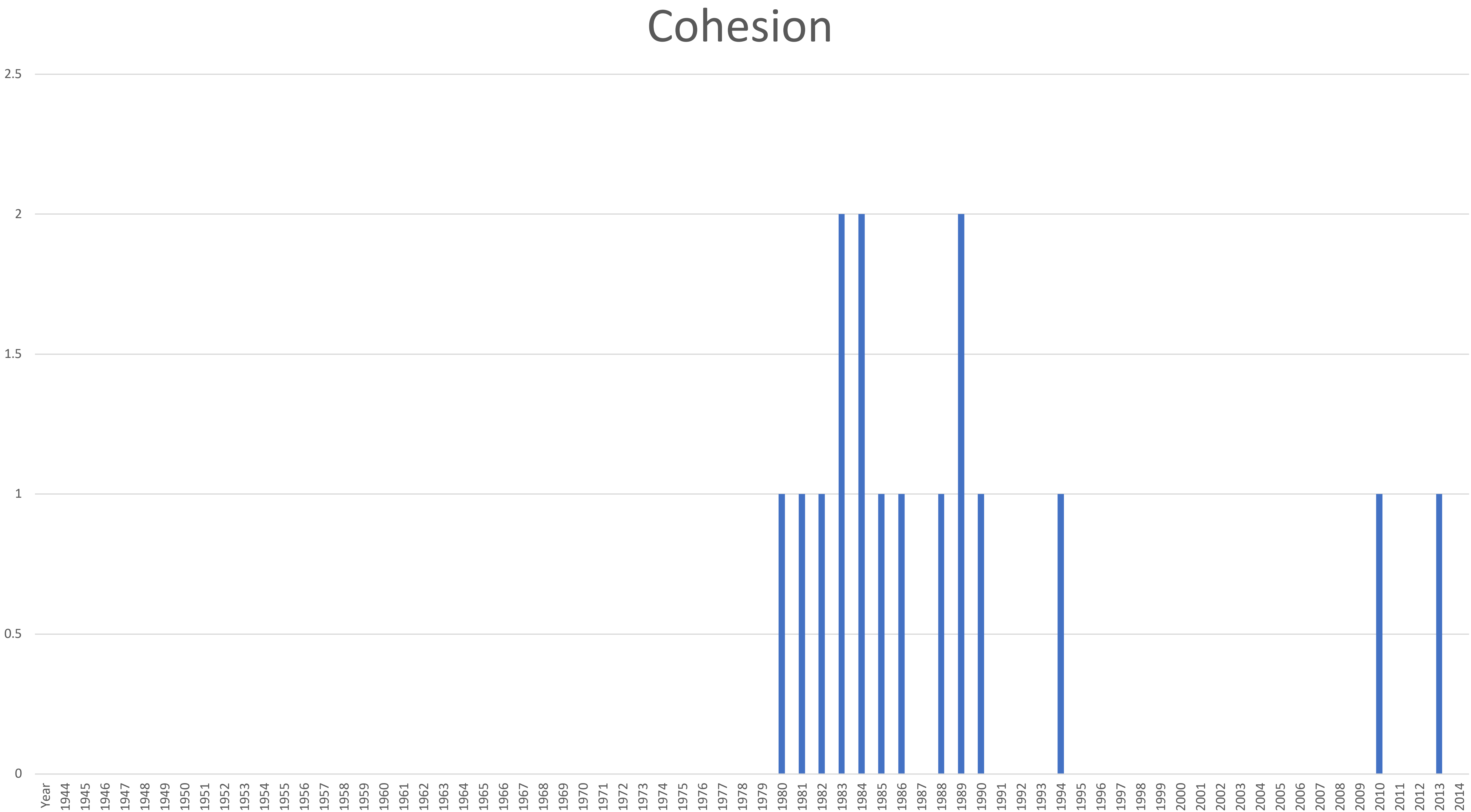
Authentic Texts



Syntax


8
7
6
5
4
3
2
1
0
Year
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014


Lexis
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Year
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014


Cohesion
2.5
2
1.5
1
0.5
0
Year
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

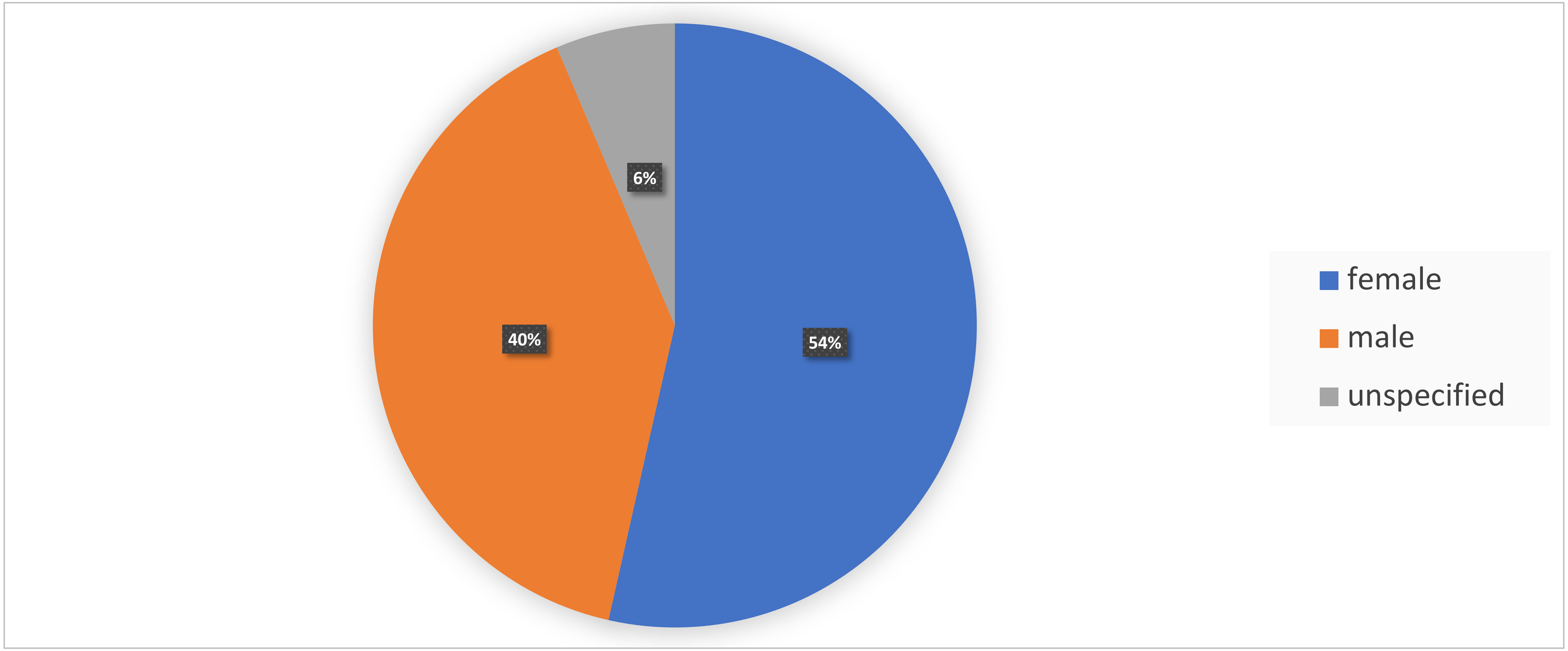
The Growth in Grammar Project

- Creating a corpus of educationally authentic writing from children in schools across England at Key Stages 1-4
- To be analysed for changes at the levels of lexis, phrase and clause.
  - NB: Analysis of structures used, not of accuracy.
- Corpus to be made publicly available from August 2018

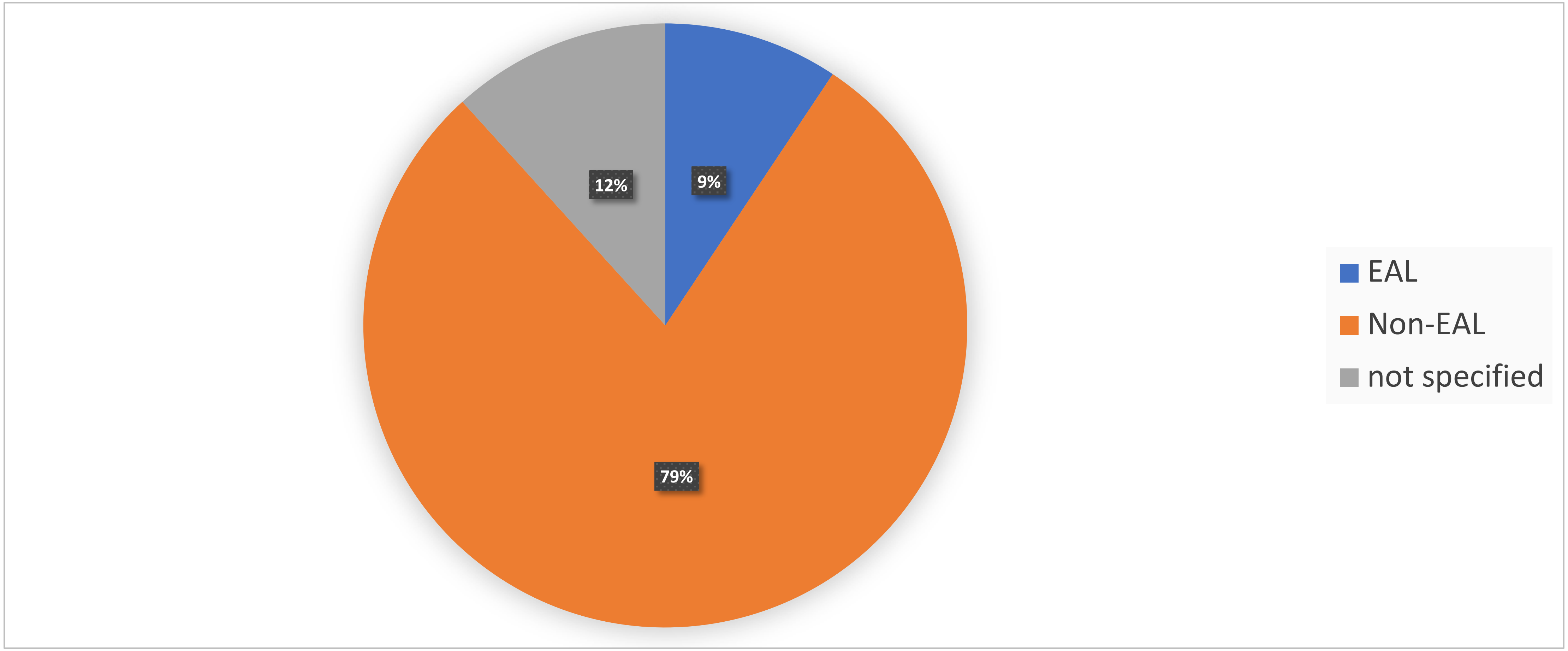
Corpus to date

	Total texts	Total schools	Total children	Discipline		
				English	Hums	Science
Year 2	543	5	138	452	84	7
Year 4	49	2	10	25	22	2
Year 6	868	7	185	548	149	171
Year 9	761	12	457	483	112	166
Year 11	316	9	165	166	49	54
All years	2,537	23	955	1,674	416	400

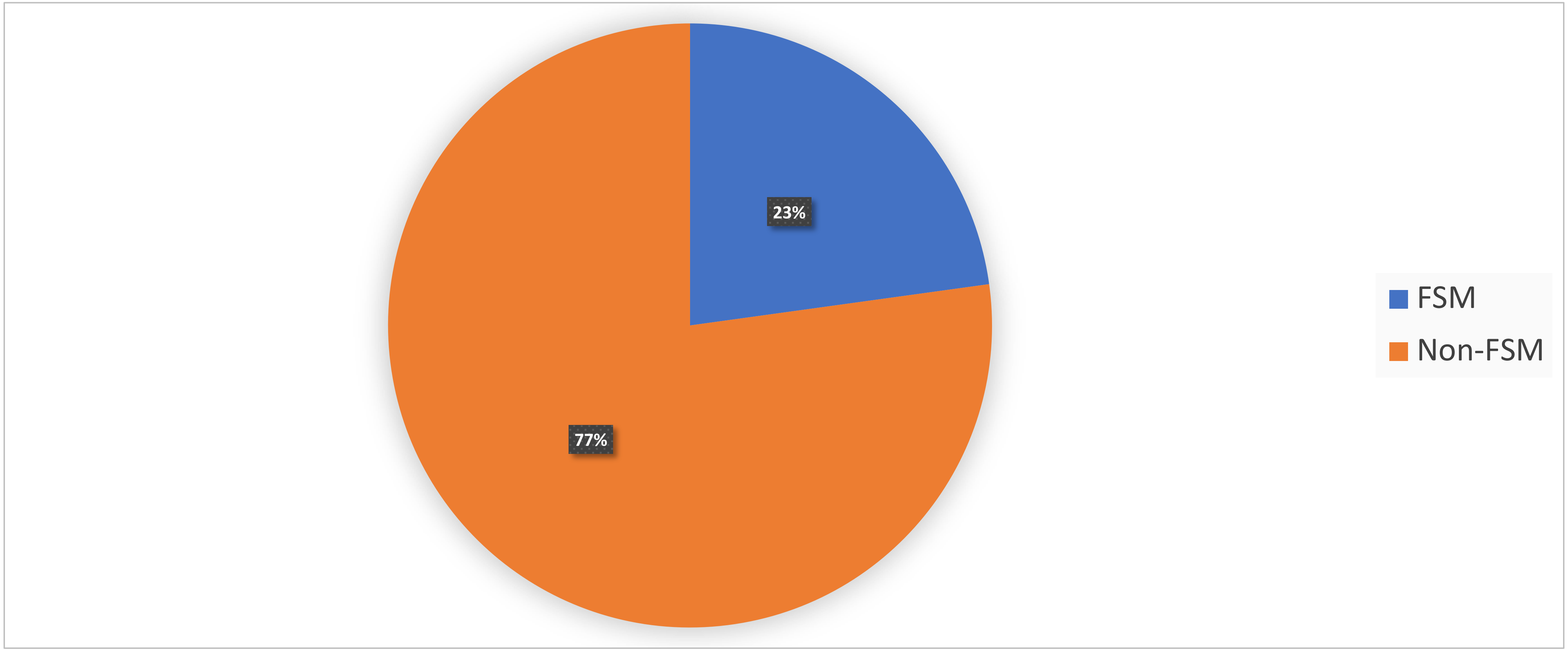
Corpus to date: Gender



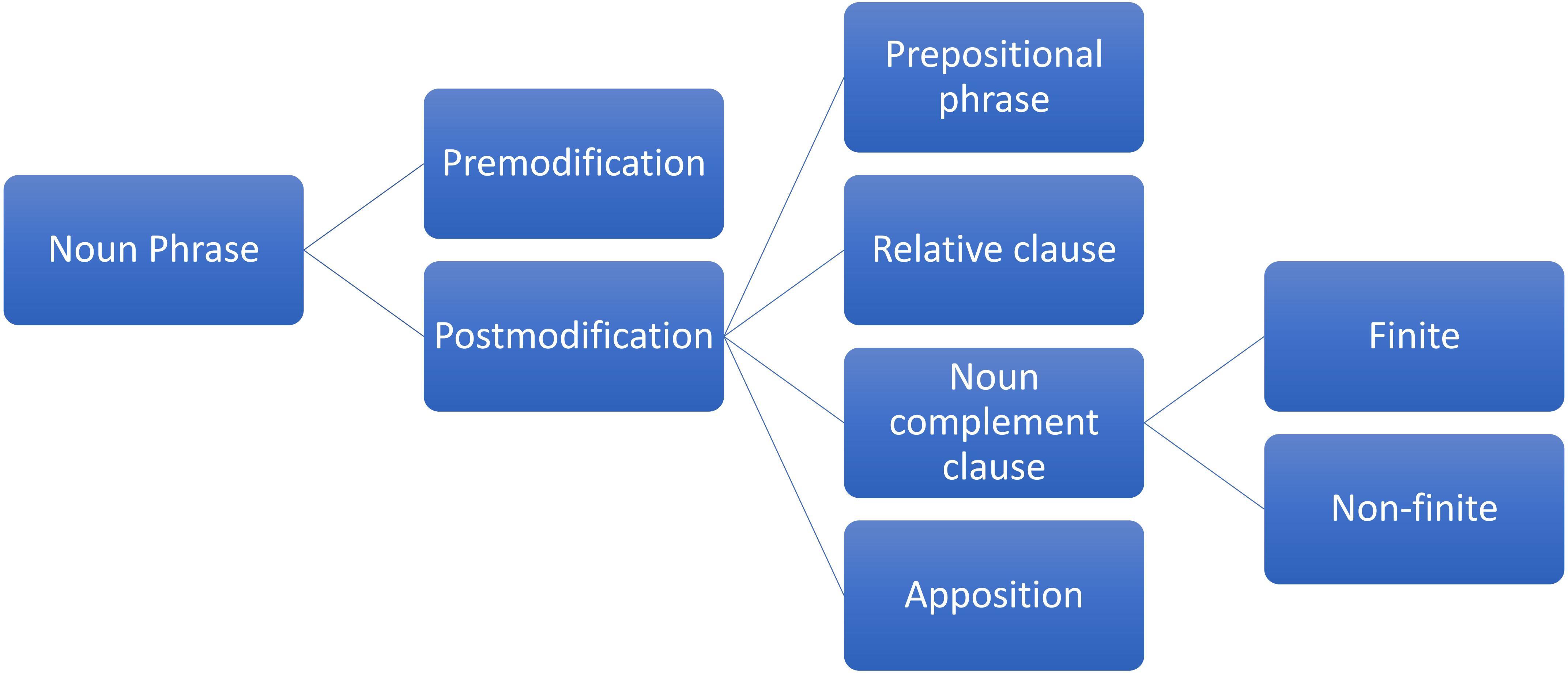
Corpus to date:
English as an Additional Language



Corpus to date: Socio-Economic Status



Ongoing:
Coding Grammatical Features



Noun Phrase

Premodification

Postmodification

Prepositional phrase

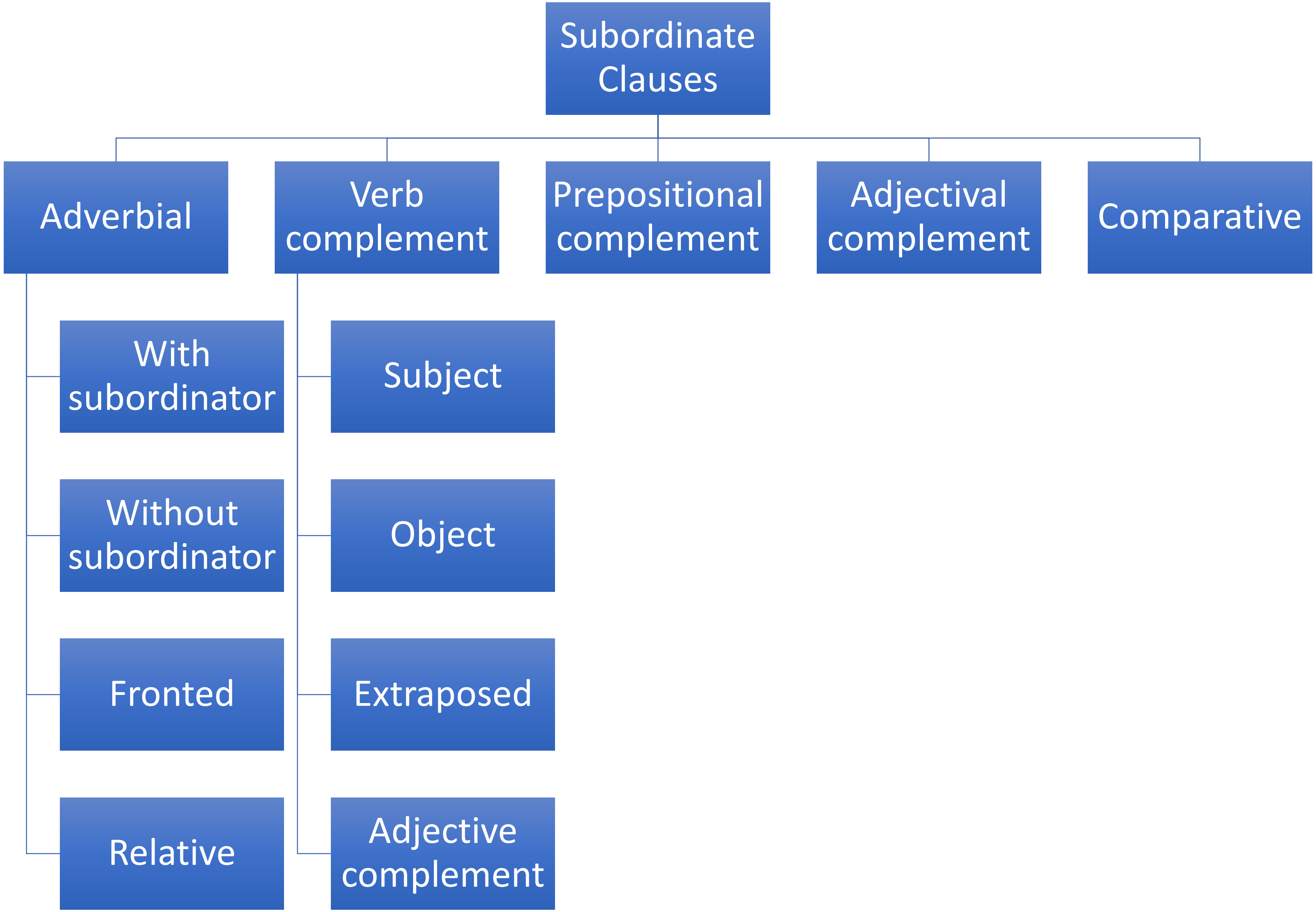
Relative clause

Noun complement clause

Apposition

Finite

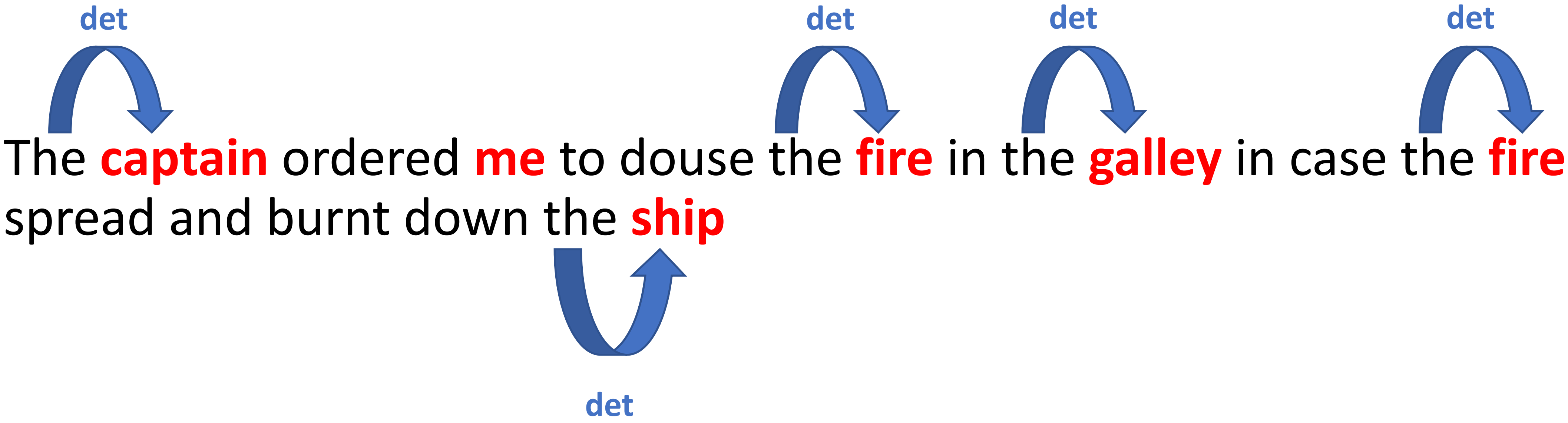
Non-finite



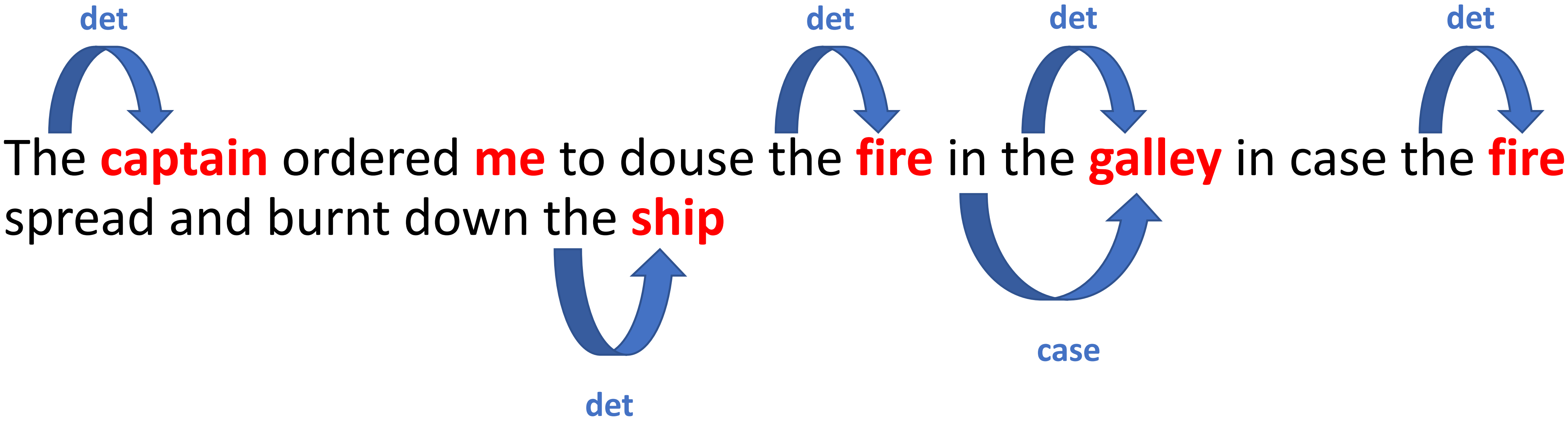
- Subordinate Clauses
  - Adverbial
    - With subordinator
    - Without subordinator
    - Fronted
    - Relative
  - Verb complement
    - Subject
    - Object
    - Extraposed
    - Adjective complement
  - Prepositional complement
  - Adjectival complement
  - Comparative

The captain ordered me to douse the fire in the galley in case the fire spread and burnt down the ship

The **captain** ordered **me** to douse the **fire** in the **galley** in case the **fire** spread and burnt down the **ship**



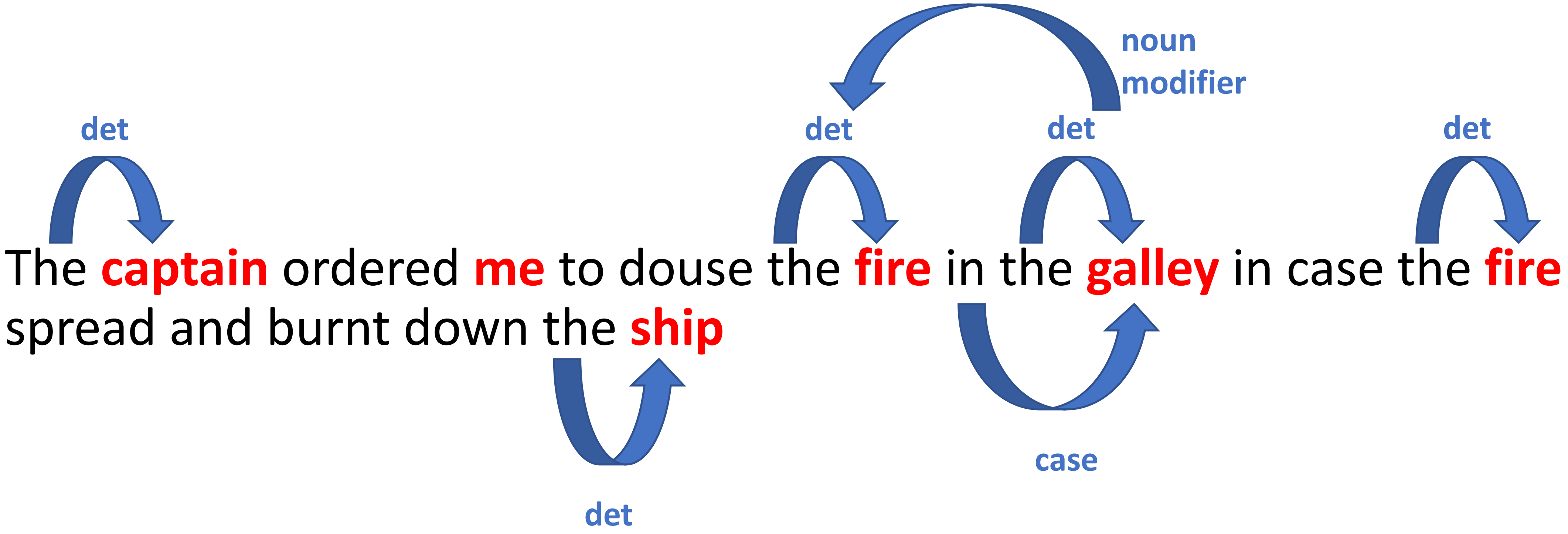
The **captain** ordered **me** to douse the **fire** in the **galley** in case the **fire** spread and burnt down the **ship**



det det det det

The **captain** ordered **me** to douse the **fire** in the **galley** in case the **fire** spread and burnt down the **ship**

det case



The **captain** ordered **me** to douse the **fire** in the **galley** in case the **fire** spread and burnt down the **ship**

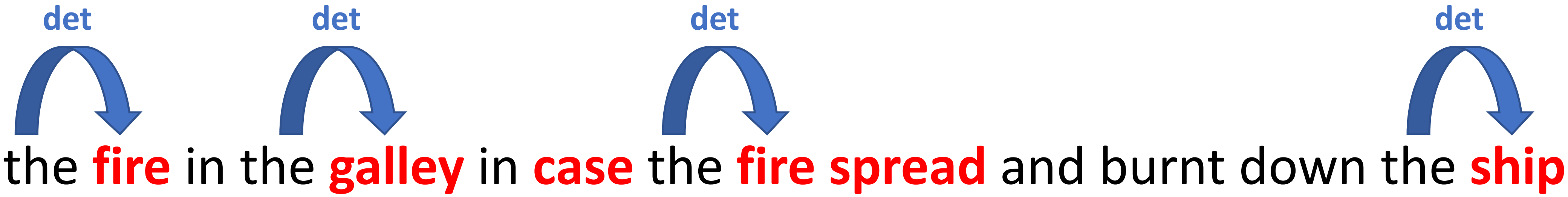
- *the **captain***
- ***me***
- *the **fire** in the galley*
  - *in the **galley***
- *the **fire***
- *the **ship***

The captain ordered me to douse the fire in the galley in case the fire spread and burnt down the ship

The **captain** ordered **me** to douse the **fire** in the **galley** in case the **fire** spread and burnt down the **ship**

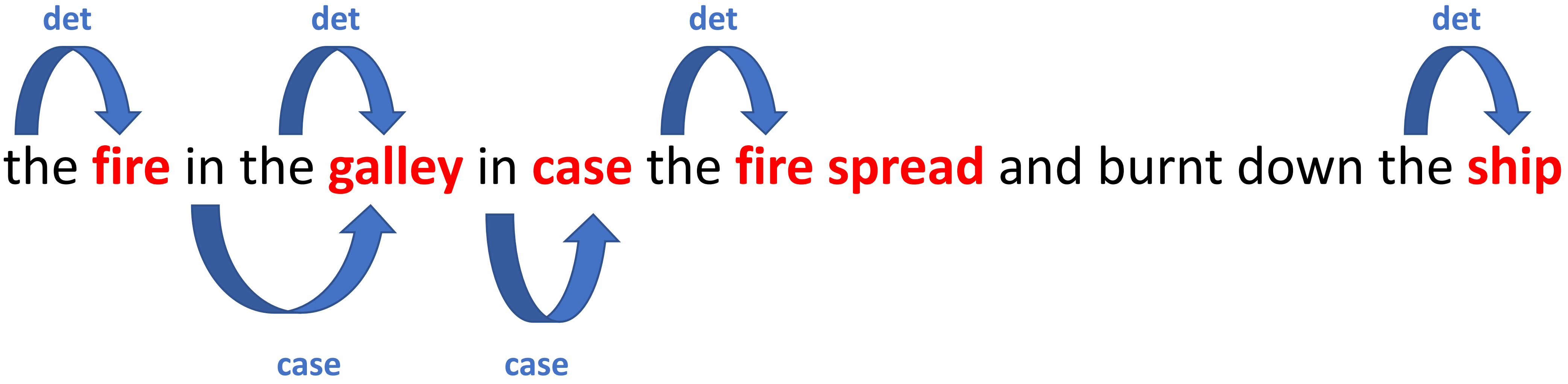
The **captain** ordered **me** to douse the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

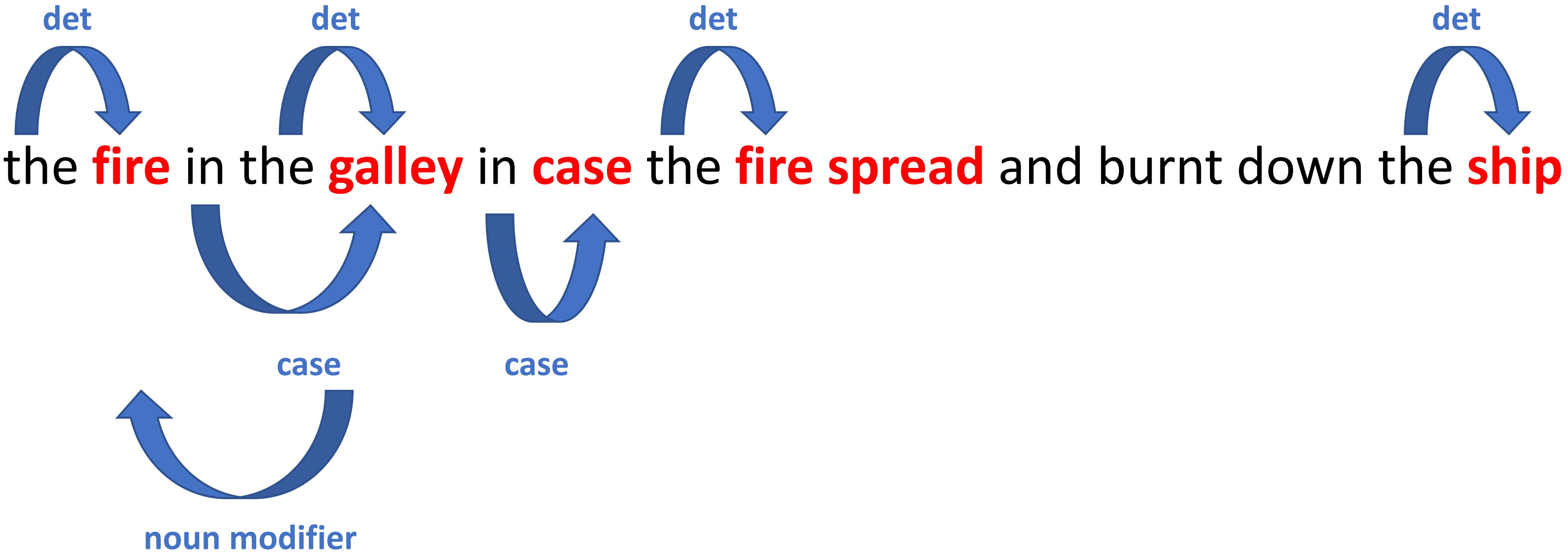


det det det det

the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**



the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

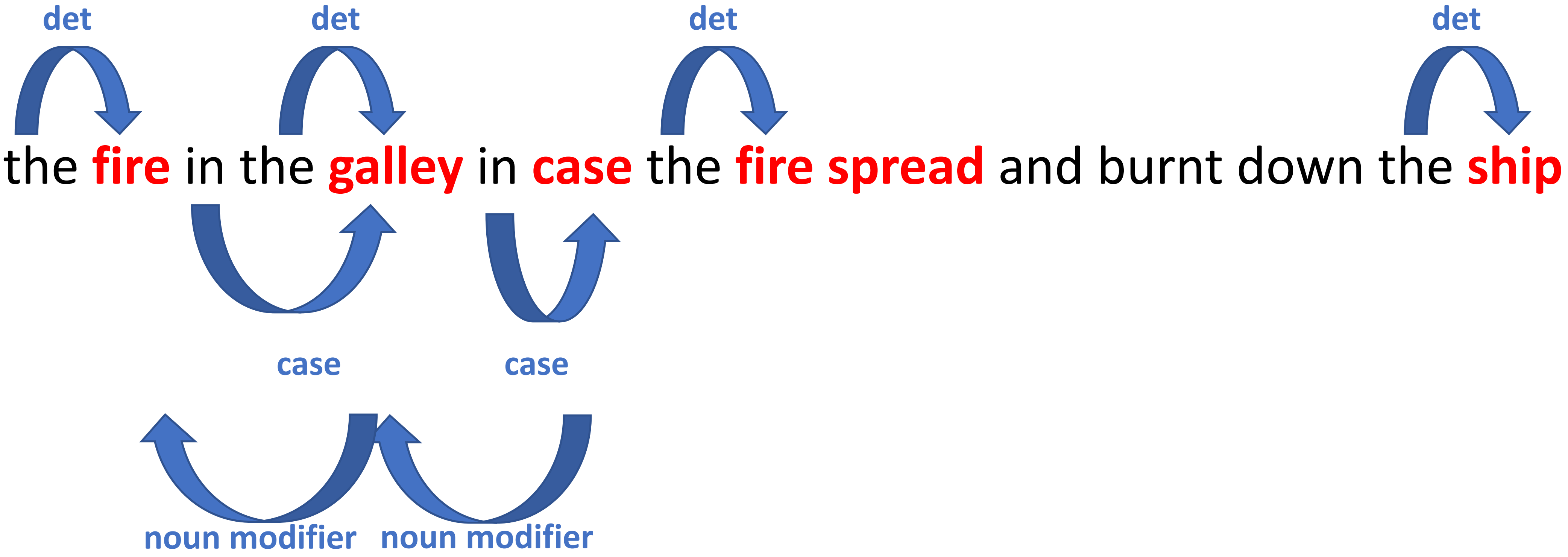


det det det det

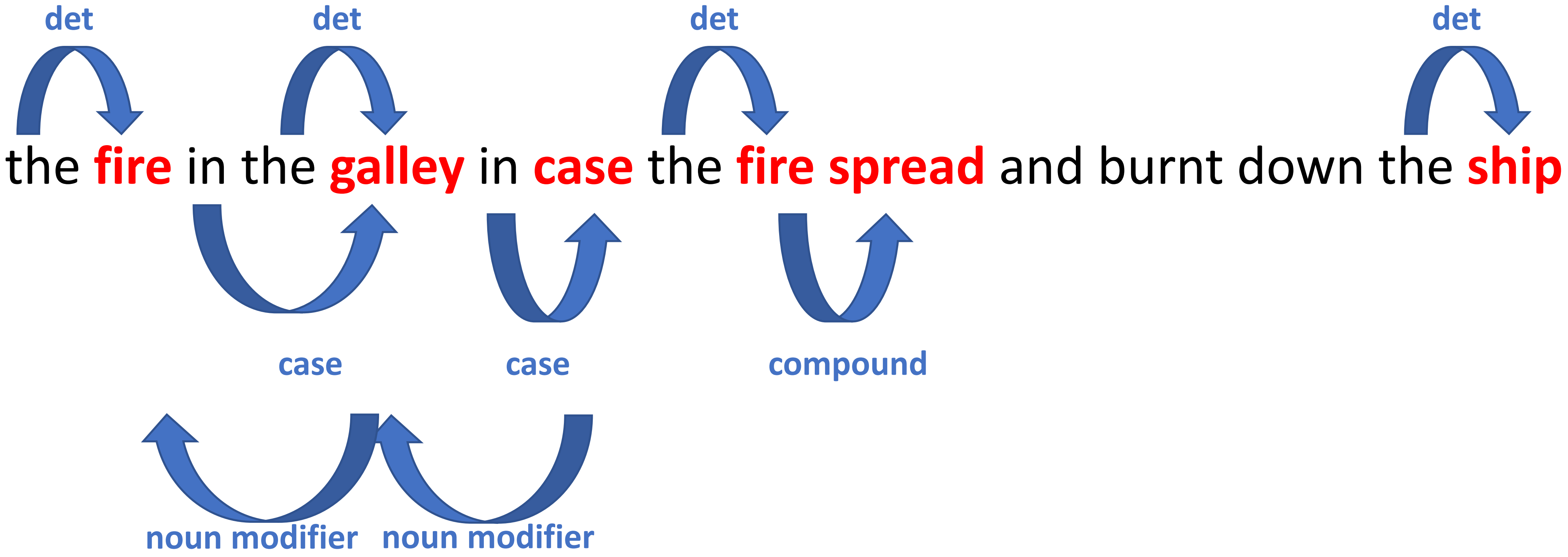
the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

case case

noun modifier



the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

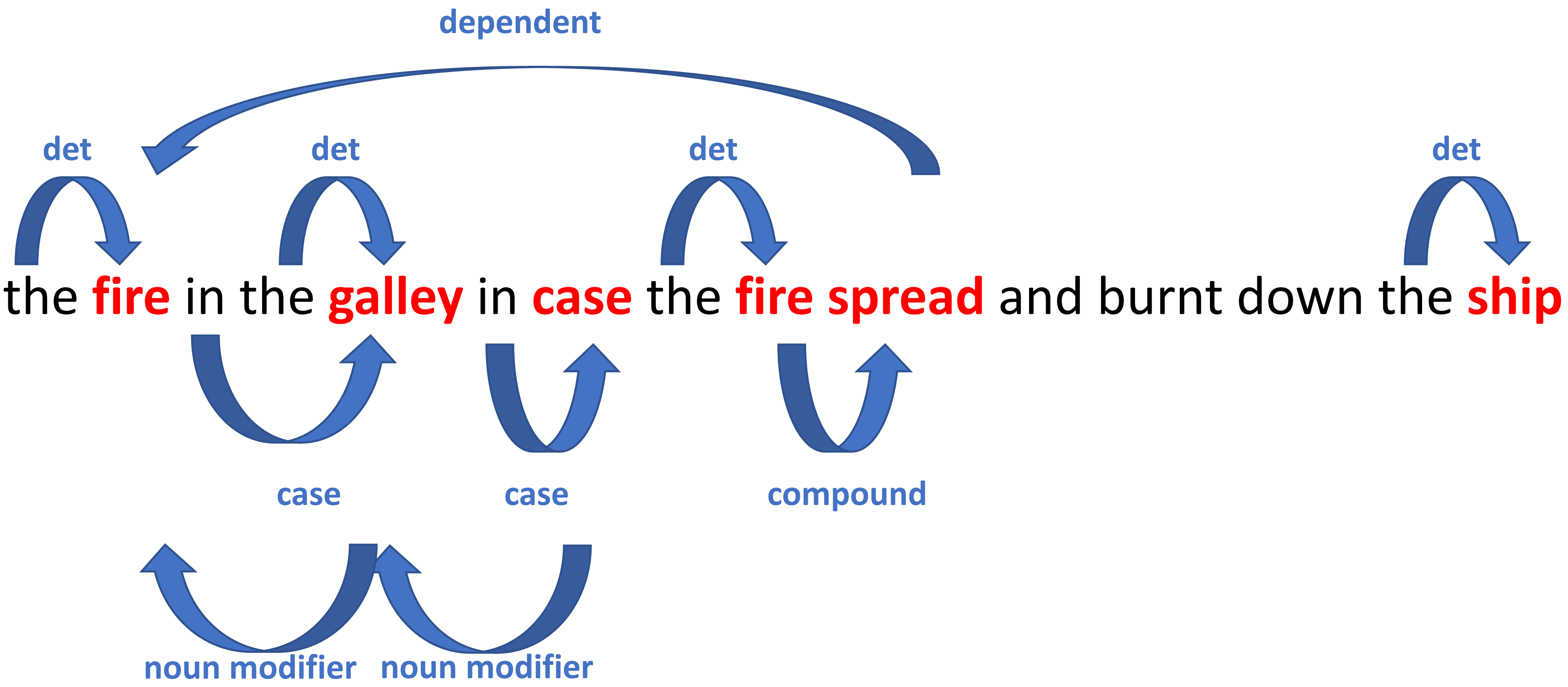


det det det det

the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

case case compound

noun modifier noun modifier



dependent

det det det det

the **fire** in the **galley** in **case** the **fire spread** and burnt down the **ship**

case case compound

noun modifier noun modifier

- *the **captain***
- ***me***
- *the **fire** in the galley*
    - *in the **galley***
- *the **fire***
- *the **ship***

- *the **captain***
- ***me***
- *the **fire** in the galley*
    - *in the **galley***
- *the **fire***
- *the **ship***

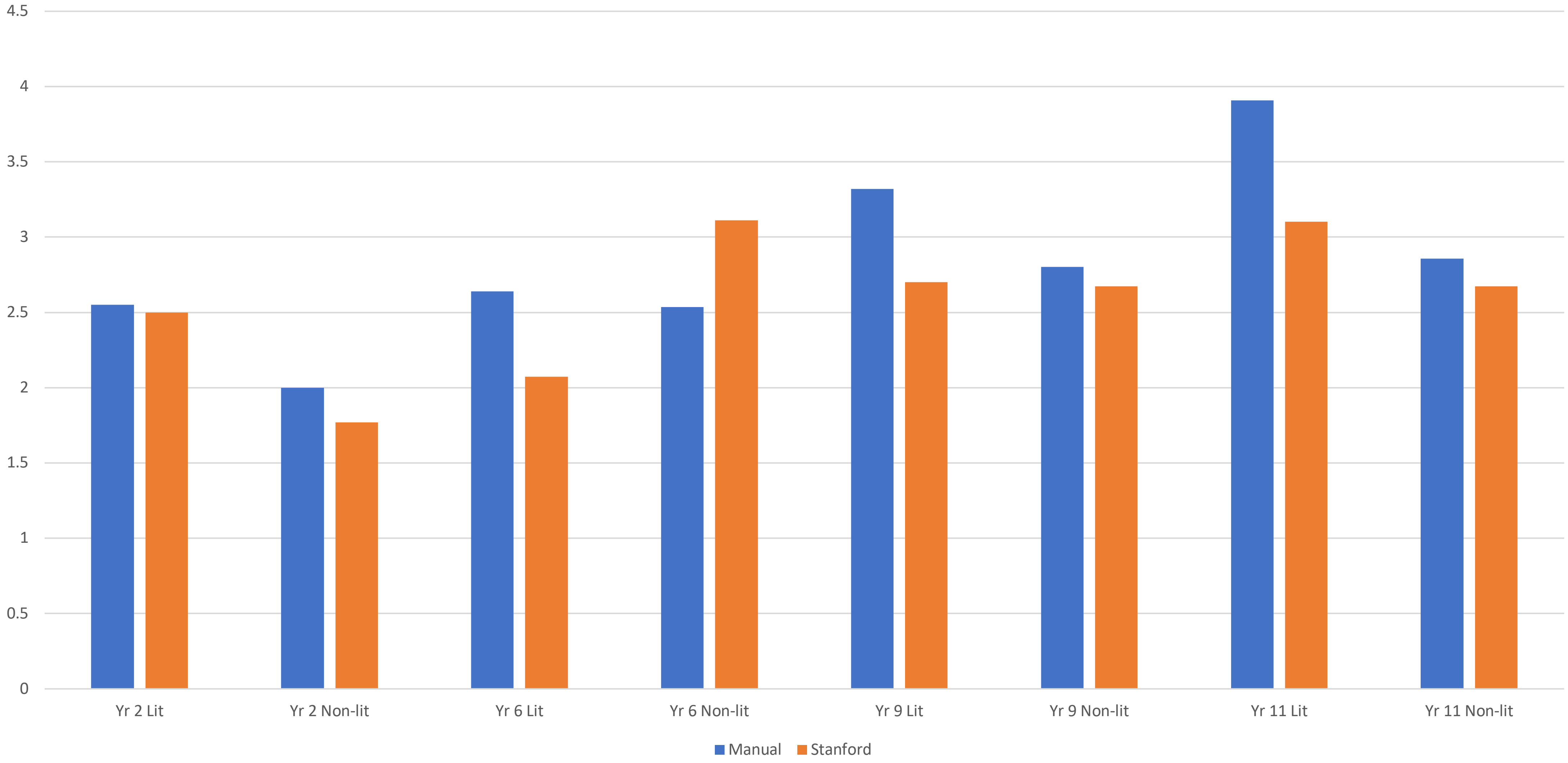
- *the **captain***
- ***me***
- *the **fire** in the galley in case the fire spread*
    - *in the **galley** in case*
        - *in **case***
    - *the fire **spread***
        - *the **fire***
- *the **ship***

- 1 x 5-word NP
- 1 x 3-word NP
- 3 x 2-word NP
- 1 x 1-word NP

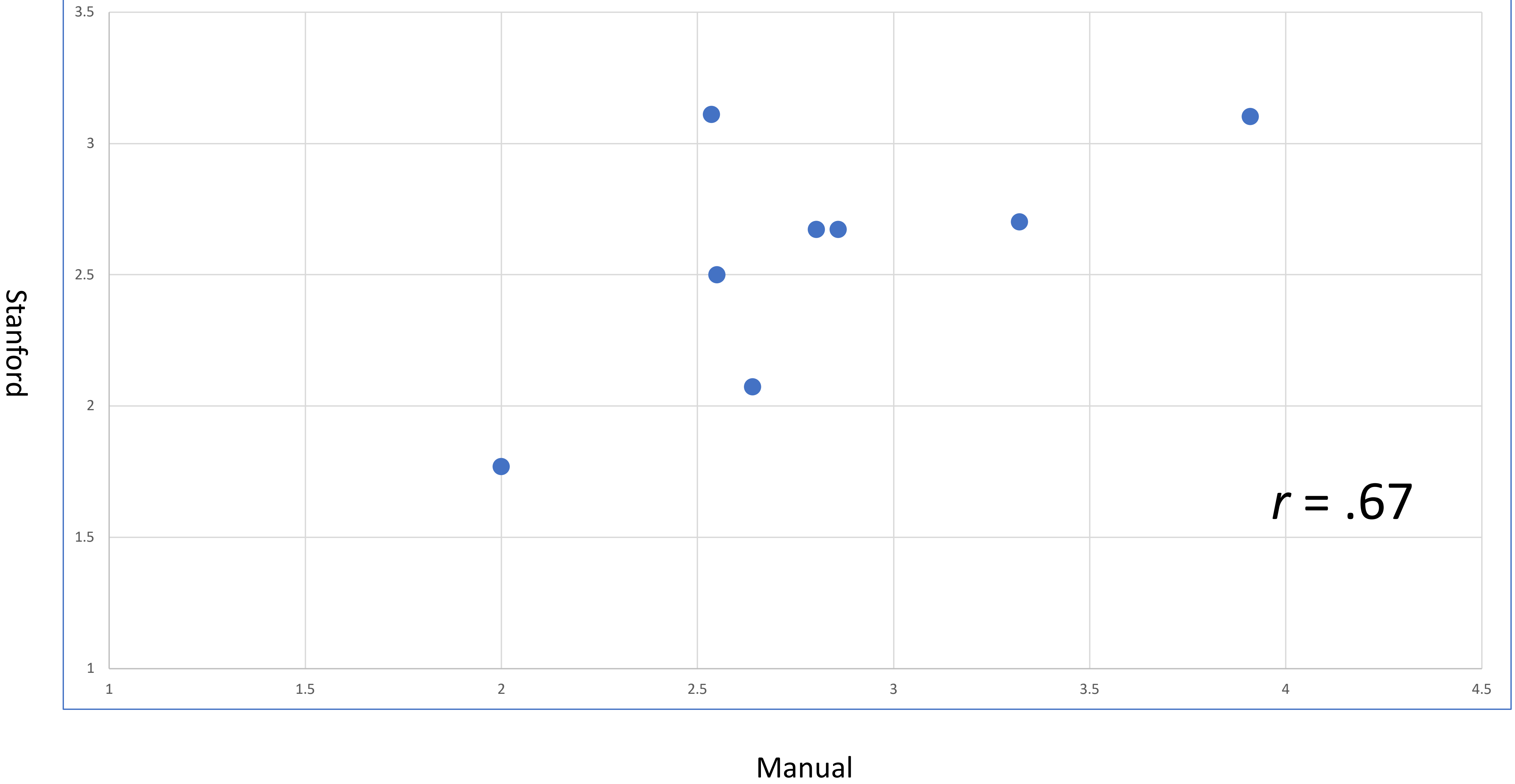
- Total NPs: 6
- Mean NP length: 2.5 words

- 1 x 10-word NP
- 1 x 5-word NP
- 1 x 3-word NP
- 4 x 2-word NP
- 1 x 1-word NP
- Total NPs: 8
- Mean NP length= 3.4 words

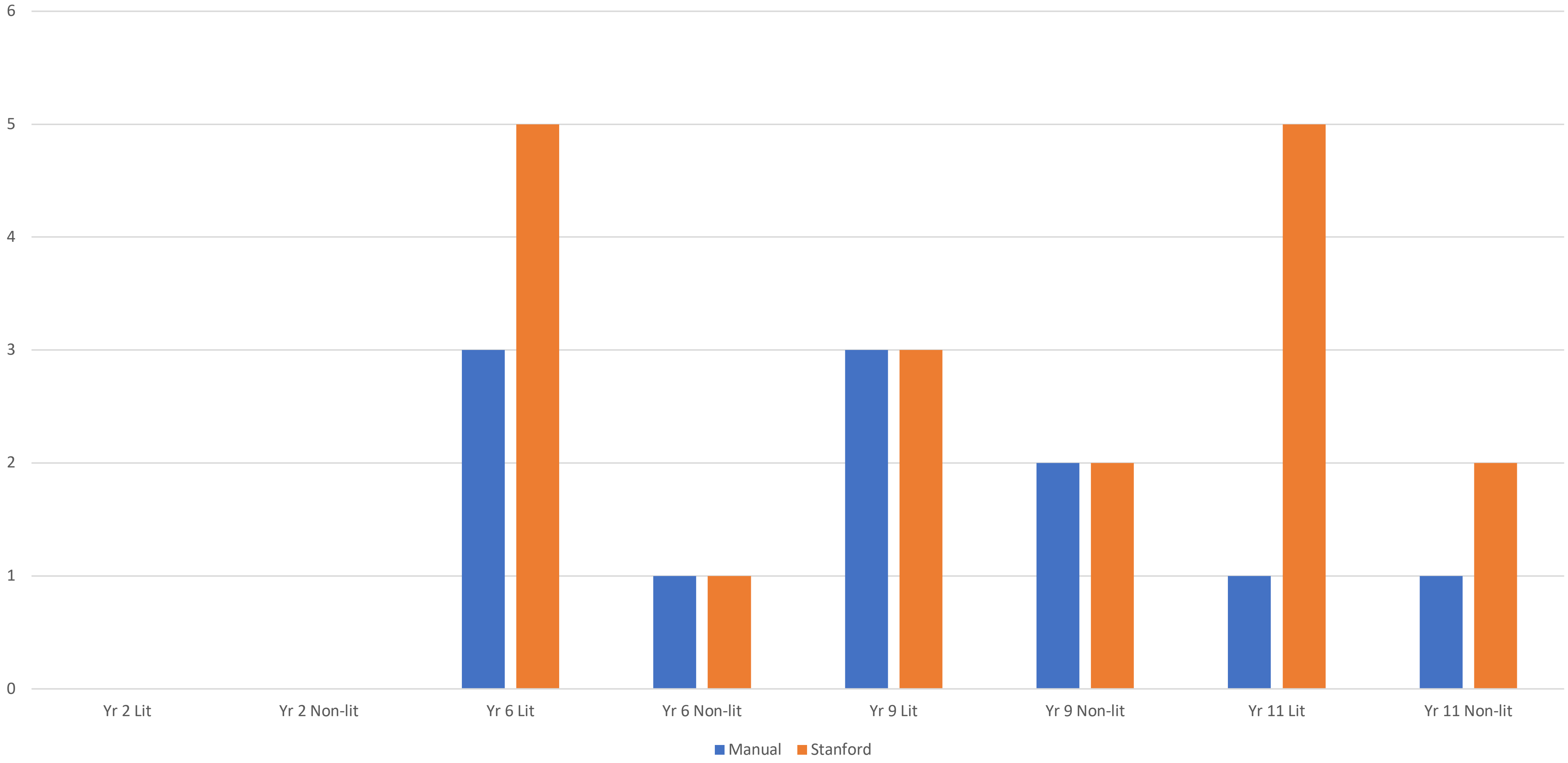
Mean Words per NP



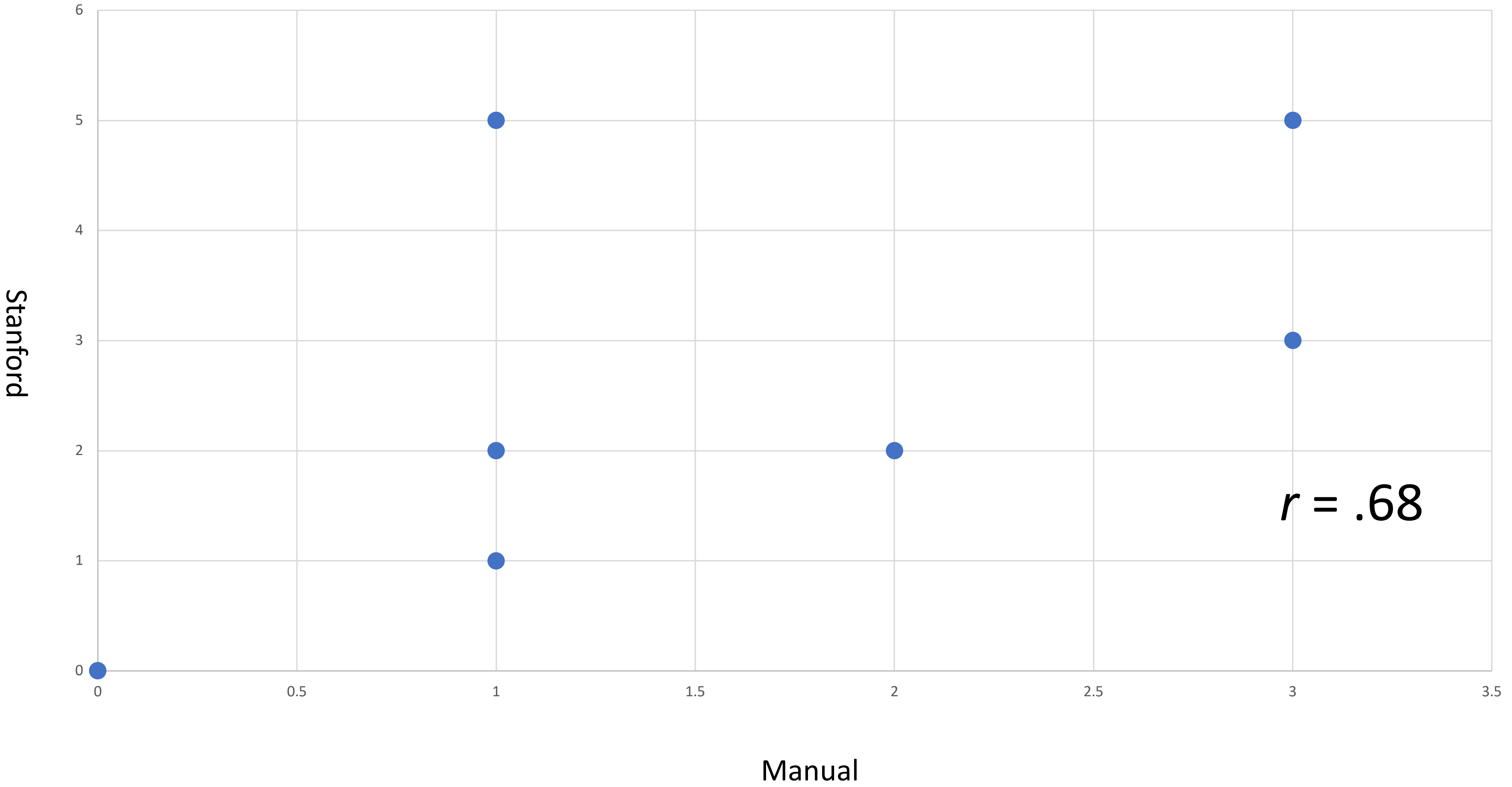
Mean Words per NP



Relative Clauses per Text

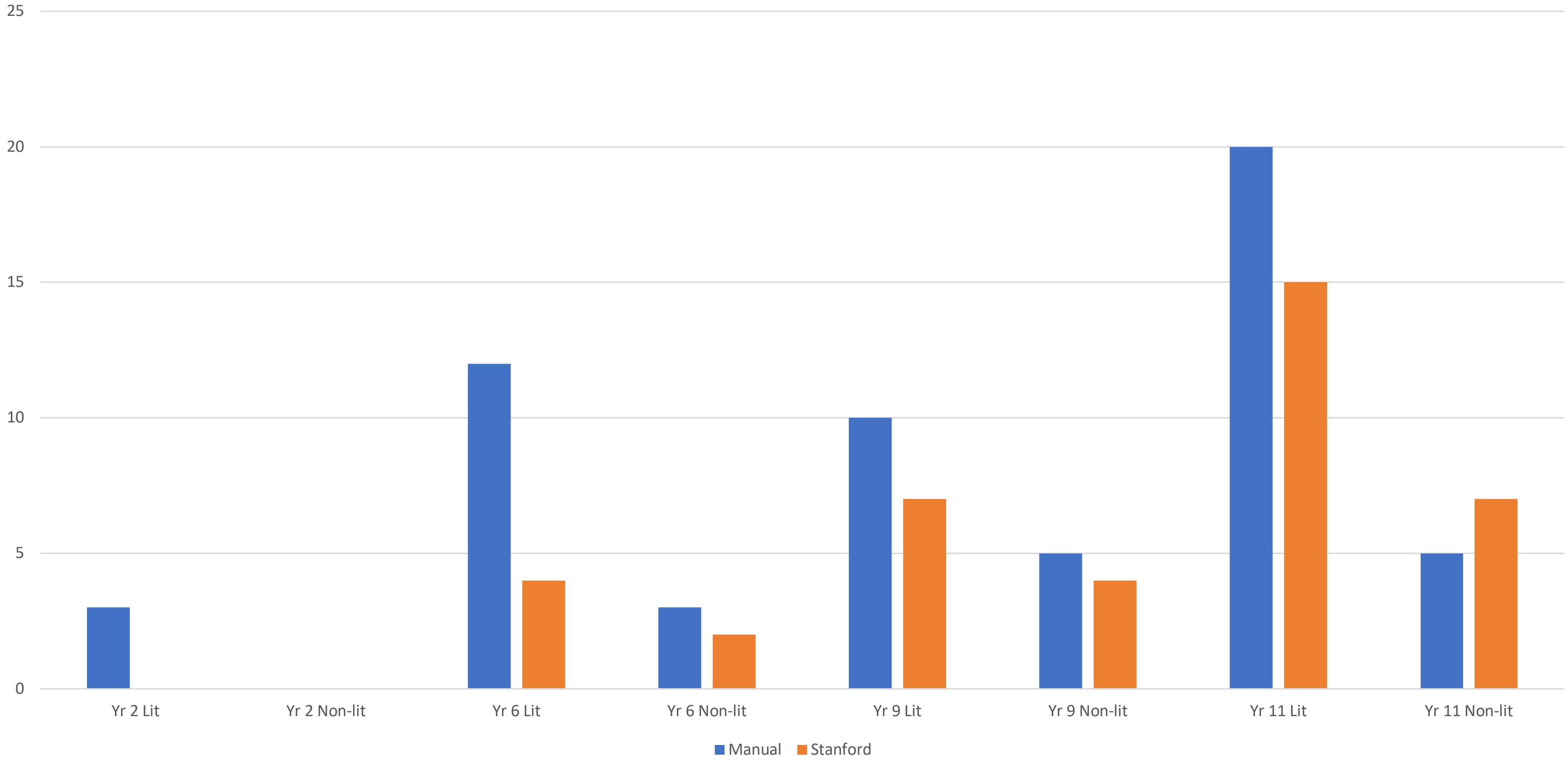


Relative Clauses per Text

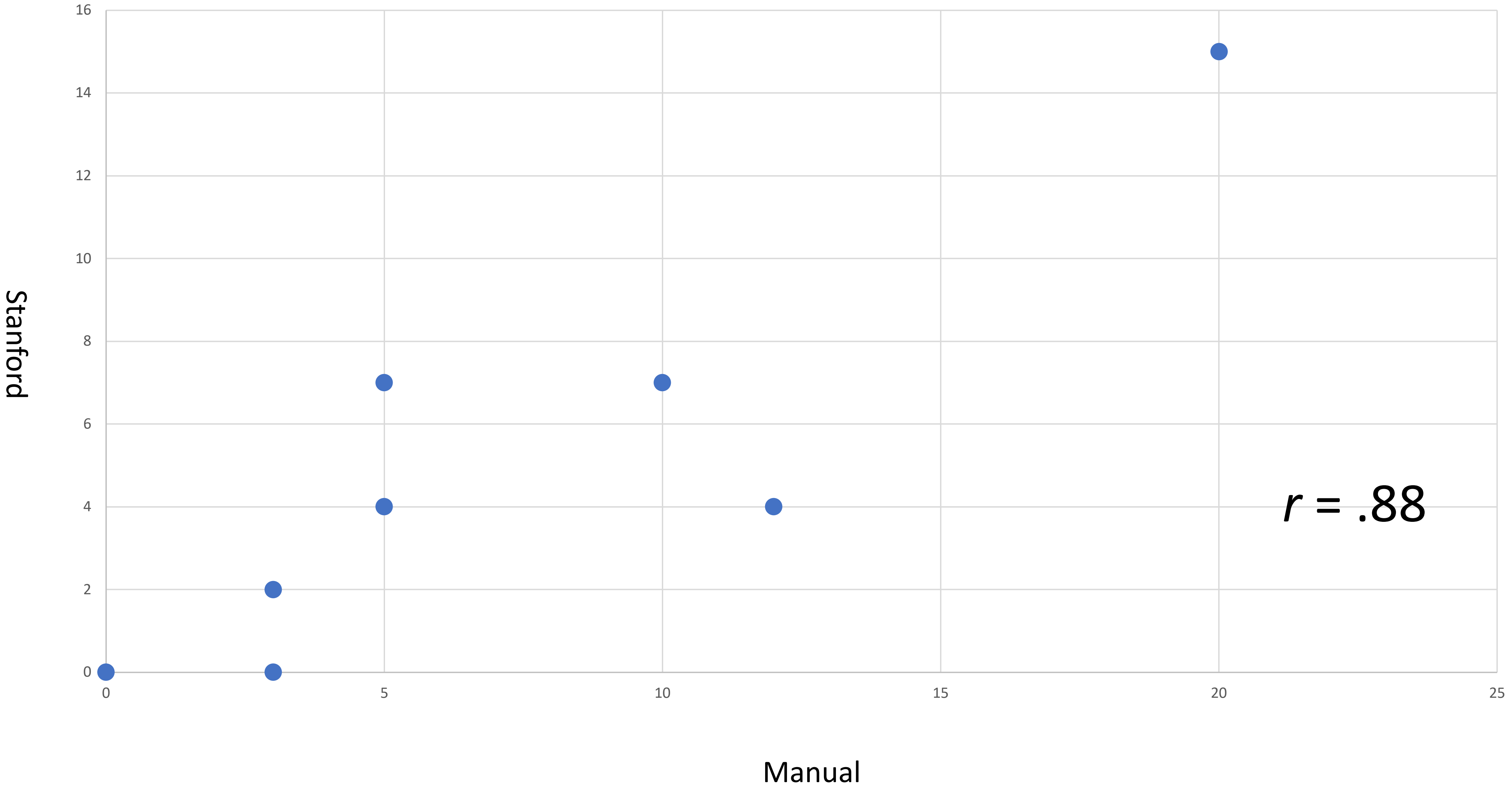


r = .68

Adverbial clauses per text

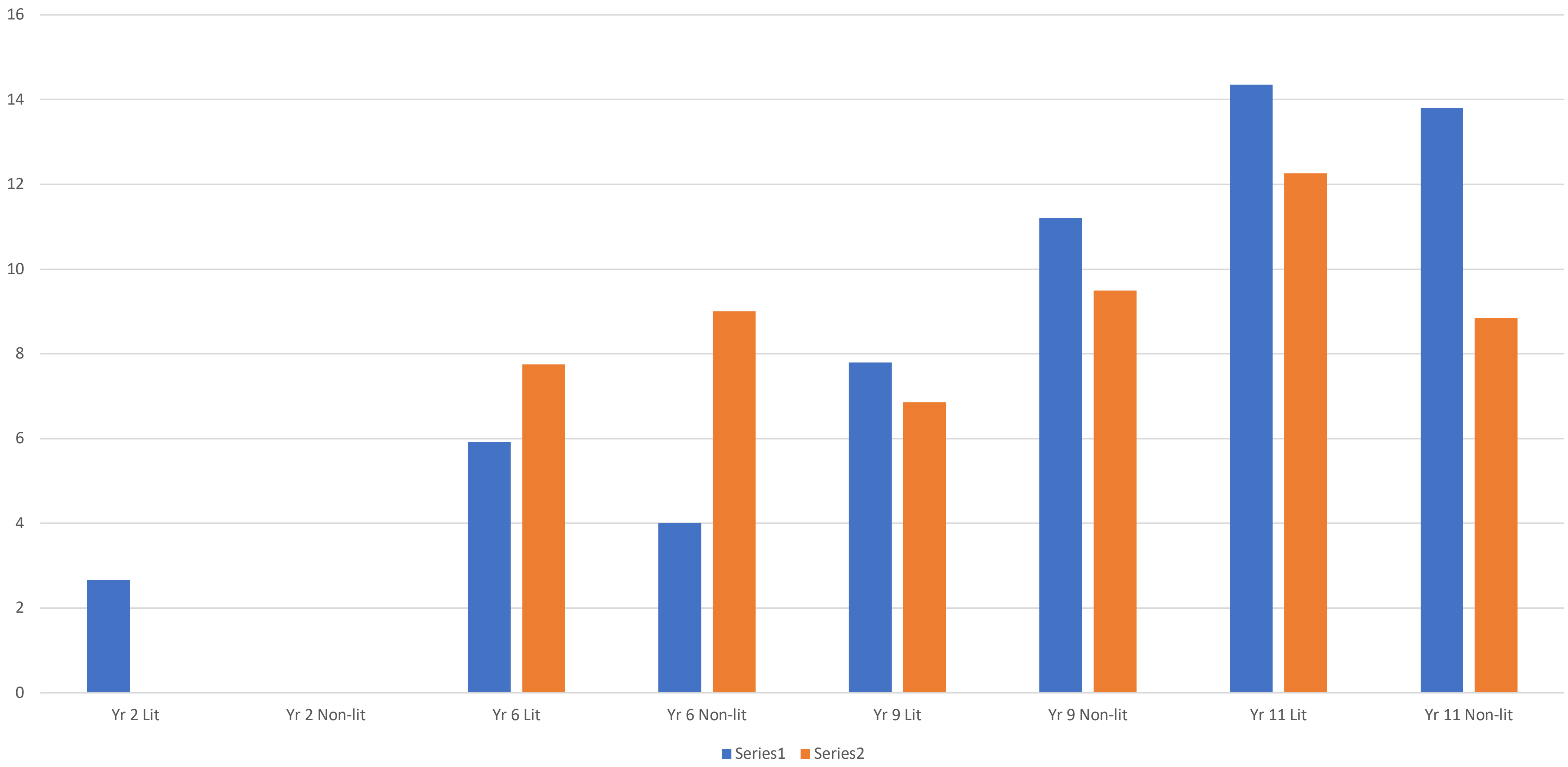


Adverbial clauses per text

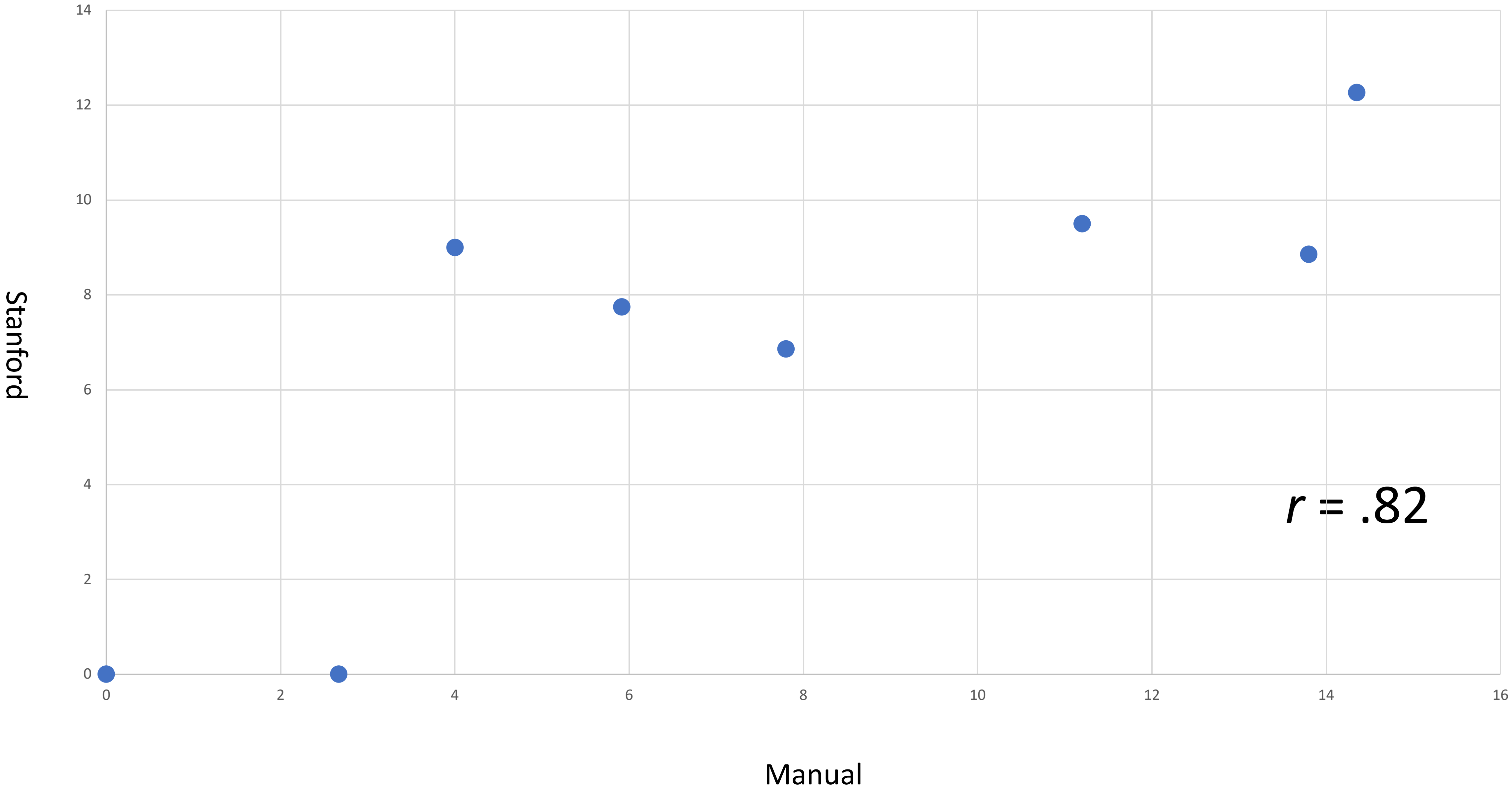


r = .88

Mean Words per Adverbial Clause



Mean Words per Adverbial Clause



Some Initial Findings: Vocabulary development

Literature on word frequency and age

Effect	Measure	Source	Ages
Decreases with age	% words < 1/100K	Finn 1977	9-10 < 16-17
	% words not on high-frequency list	Olinghouse & Leaird 2009	7-8 > 9-10
	% 1K words	Sun, Zhang & Scardamalia 2010	8 > 10
	% 2K words	Sun, Zhang & Scardamalia 2010	8 < 10
	% off-list words	Sun, Zhang & Scardamalia 2010	8 < 10
Ambiguous age effects	% words not on high-frequency list	Lawton 1963	12 > 14 for working-class children, not middle class
	Mean frequency from reference corpus	Crossley et al 2011	14-15 < 18-19 14-15=16-17 16-17=18-19
No age effects	P-Lex	Malvern et al 2004	7 = 11 = 14

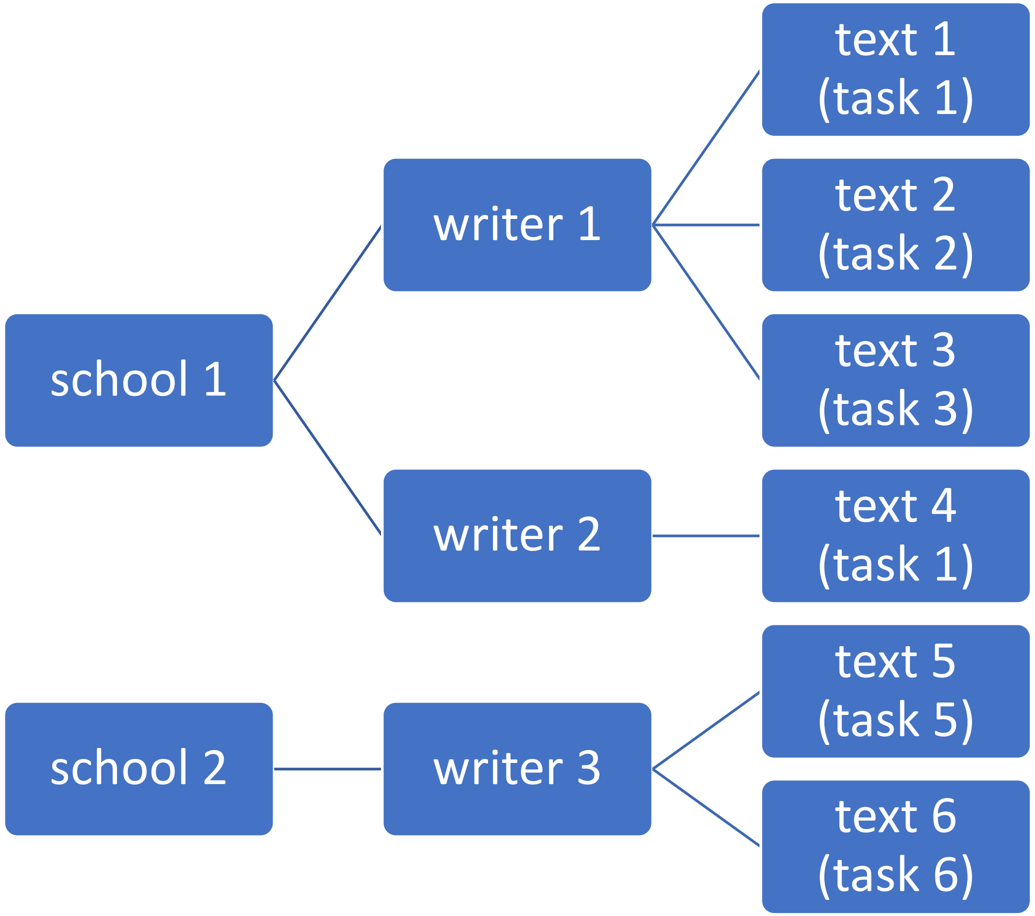
Sample for the current study

- Years 2, 6 and 9 only
- English/Humanities classes only
- Exclude texts with > 100 illegible words per 1,000
- Exclude poems
- Exclude samples more than 1SD from mean word length
- Spelling errors corrected
- Randomly select texts to give equal numbers in each year group

Study Corpus

	Schools	Writers	Texts	Text Length				Genre		
				Mean	Median	Min	Max	Story	Exposition	Persuasion
Year 2	3	78	219	66.6	62	27	131	116	99	4
Year 6	4	90	219	284.2	261	120	521	114	82	23
Year 9	6	189	219	343.3	330	181	560	130	59	30

Corpus structure



school 1

writer 1

text 1 (task 1)

text 2 (task 2)

text 3 (task 3)

writer 2

text 4 (task 1)

school 2

writer 3

text 5 (task 5)

text 6 (task 6)

TAALES* Indices

- Frequency
- Ngram frequency/association

*Kyle & Crossley 2014

Frequency: 72 indices

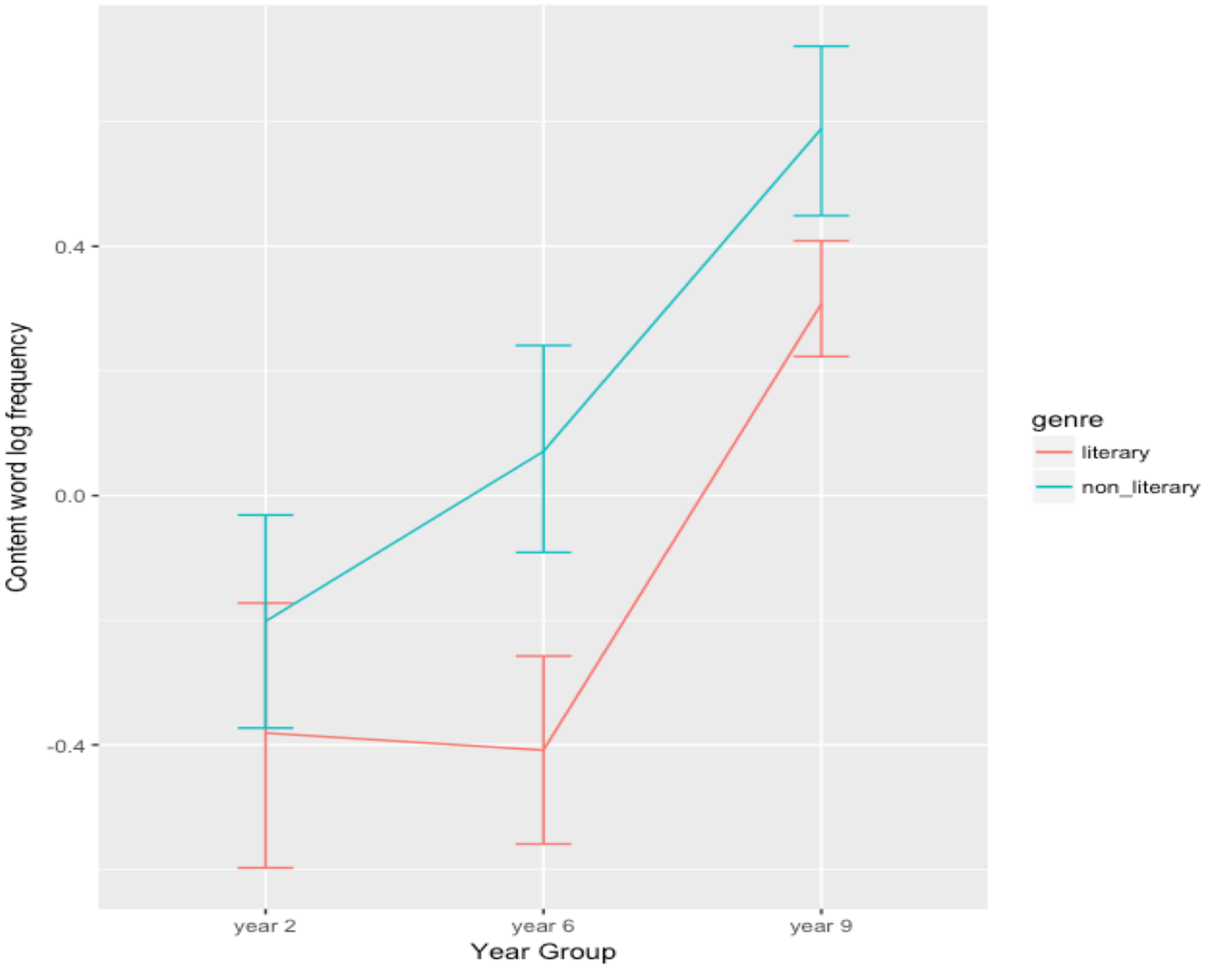
- Range of reference corpora
- Separate indices for:
  - all words vs. content words vs. function words
  - raw frequency vs. log frequency

Combining results from different corpora

Sub-category (1)	Sub-category (2)	Cronbach’s alpha	Deleted
All words	Raw	.99	SUBTLEXus (*r* = .41)
	Log	.98	
Content words	Raw	.98	
	Log	.98	
Function words	Raw	.99	SUBTLEXus (*r* = .41)
	Log	.98	

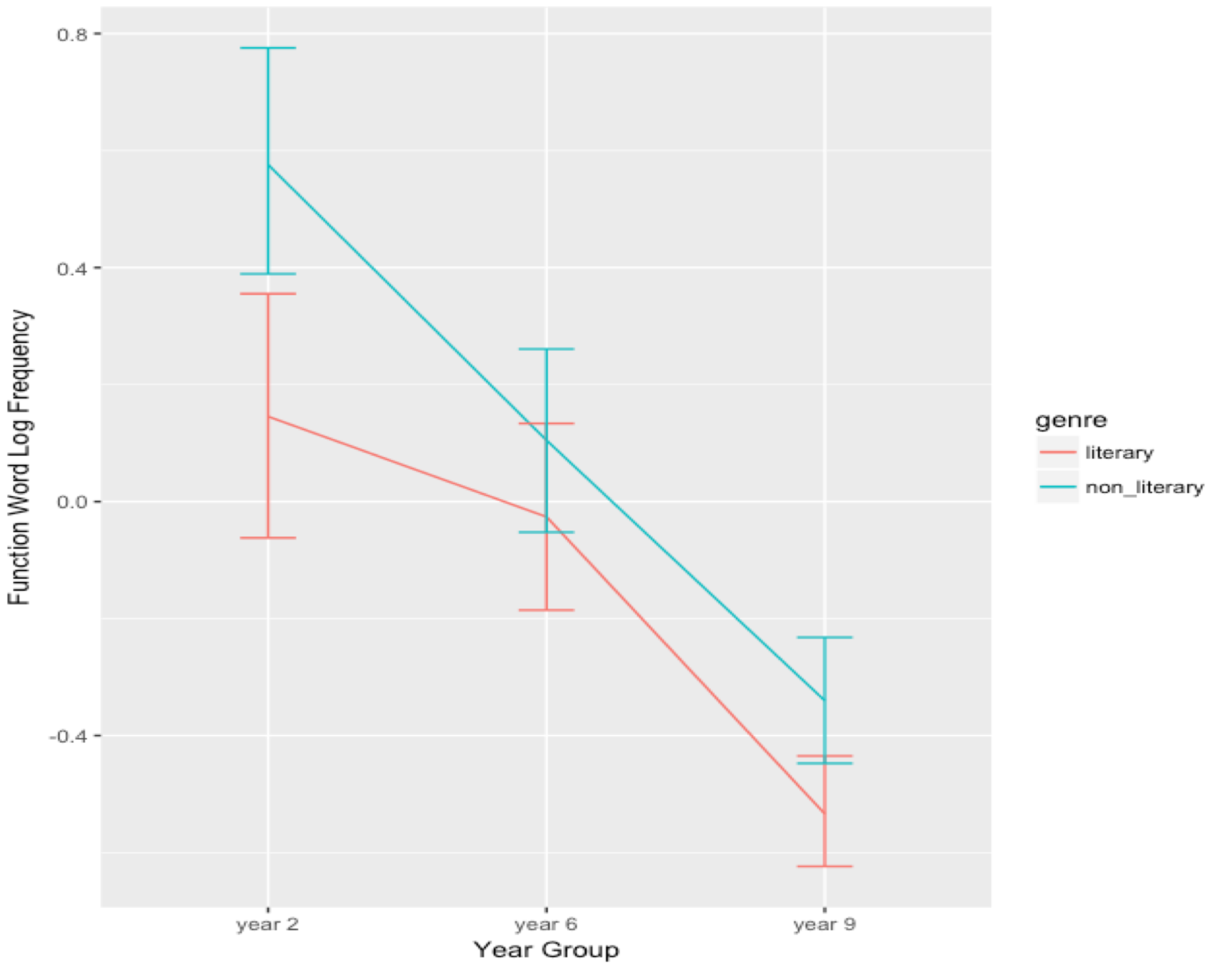
		All words		Content words		Function words	
		Raw	Log	Raw	Log	Raw	Log
All words	Raw	1.00					
	Log	0.18	1.00				
Content words	Raw	-0.15	0.61	1.00			
	Log	-0.27	0.80	0.76	1.00		
Function words	Raw	0.83	-0.17	-0.30	-0.42	1.00	
	Log	0.68	-0.08	-0.29	-0.35	0.86	1.00

Content Word Log Frequency



Age: Year 6: *b*=-.09, *t*(78)=-.49, *p*>.05; **Year 9: *b*=.56, *t*(78)=2.19, *p*<.05**
Genre: non-literary: *b*=.56, *t*(78)=3.05, *p*<.05
(random intercept for task)

Function Word Log Frequency



Age: Year 6: *b*=-.15, *t*(78)=-.84, *p*>.05; **Year 9: *b*=-.72, *t*(78)=3.31, *p*<.05**
Genre: non-literary: *b*=.07, *t*(78)=0.43, *p*>.05
(random intercept for task)

For example

One morning, five meerkats finished eating crunchy scorpians but there preditors like slithering snakes, lions and falcons. But one woke up and went meerkats, meerkats and woke up the others. One stepped on the snake and it went hiss hiss! And the other predators ran away. But a fennec fox was coming to take two pups. The scary fox wishes to eat them even the nice scorpians.

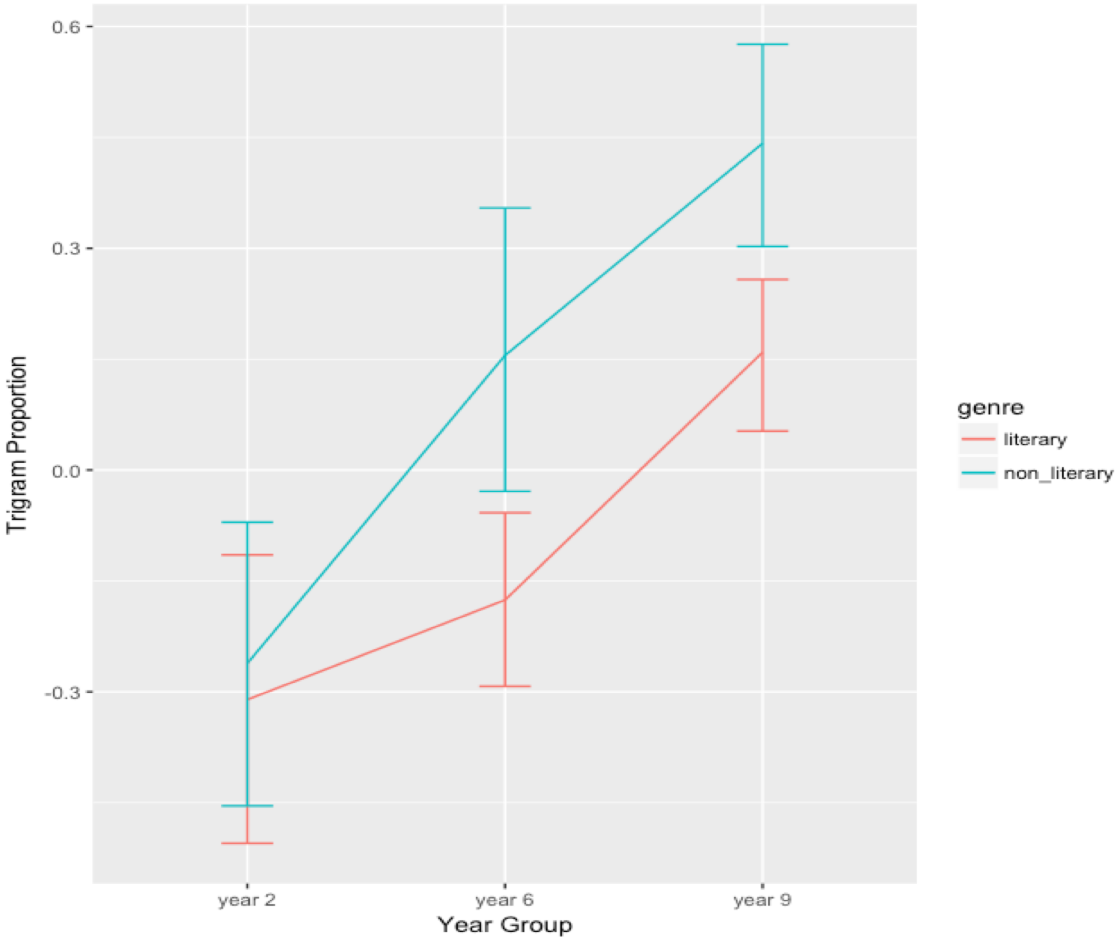
Ngram measures: Proportion & Frequency

- Bigram & Trigram
- 7 x reference corpora; Association: 5 x reference corpora
- Proportion: 10K; 20K; 30K…100K)

Combining results from different corpora

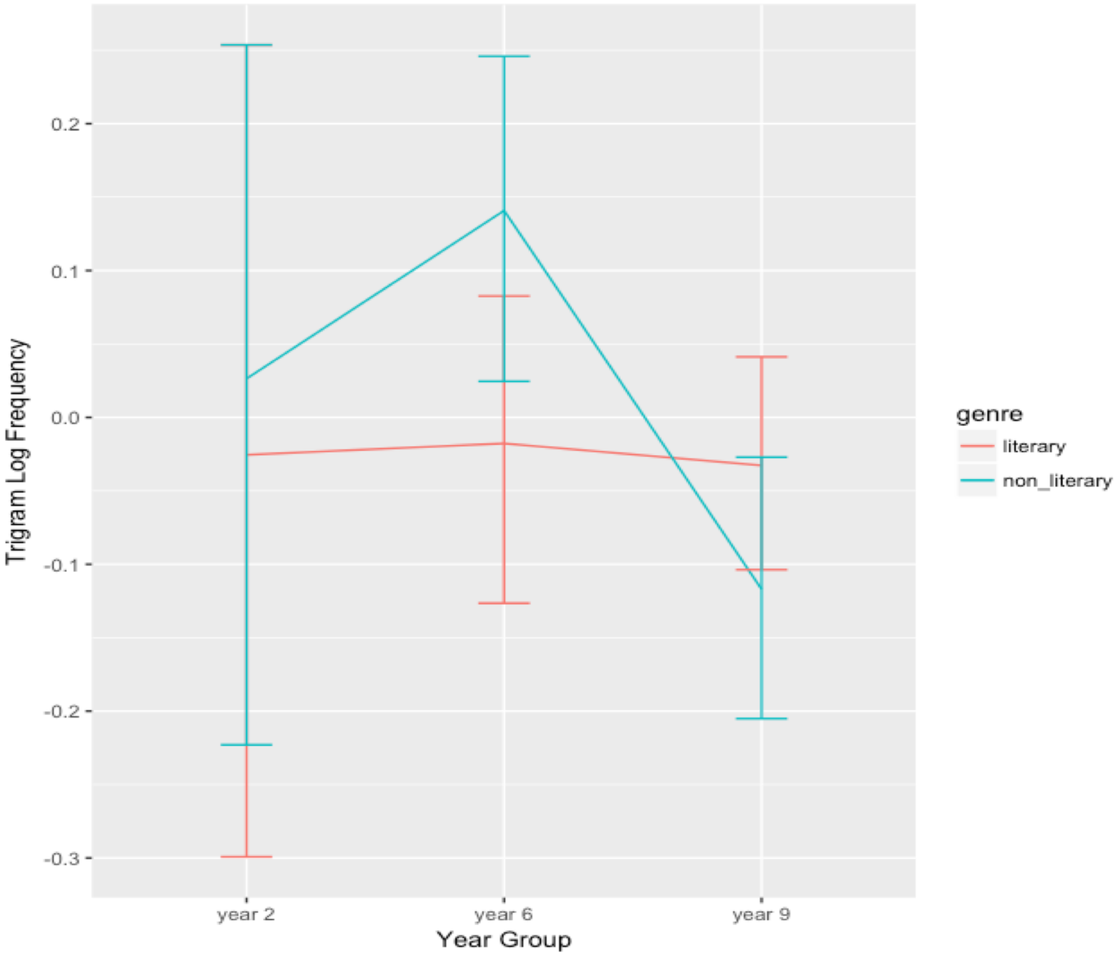
Category	Cronbach’s alpha	Deleted
Trigram Proportion (combines all frequency bands)	1	
Trigram Log Frequency	.93	BNC Spoken: .14 BNC Written: .20 COCA Academic: .65

Trigram Proportion



Age: Year 6: *b*=.13, *t*(78)=.76*, p*>.05; **Year 9: *b*=.62, *t*(78)=2.06*, p*<.05**
Genre: non-literary: *b*=.19, *t*(78)=1.27*, p*>.05
(random intercept for task)

Trigram Log Frequency



Age: Year 6: *b*=.13, *t*(78)=.1.05 *p*>.05; Year 9: *b*=.15, *t*(78)=-.07*, p*>.05
Genre: non-literary: *b*=.03, *t*(78)=.28*, p*>.05
(random intercept for task)

Trigram measures: Association

- Frequent n-grams tend to be combinations of frequent words:
    - *to be a*
    - *he didn’t*
    - *back to the*
    - *it was the*
- Often use measures of *association*: is the combination frequent in relation to the frequency of the words?

Trigram measures: Association

- *t-score:* Is the combination more frequent than chance would predict, given the frequency of the component words?
- *mutual information (MI):* how strongly do the component words predict each other?
- *delta-P:* conditional probability from one component to another

Trigram measures: Association

- Range of reference corpora
- MI, MI2; t-score; Delta-P; Collexeme
- Trigram segmentation
  - Trigram 1 (*double* – *espresso please*)
  - Trigram 2 (*double espresso* – *please*)

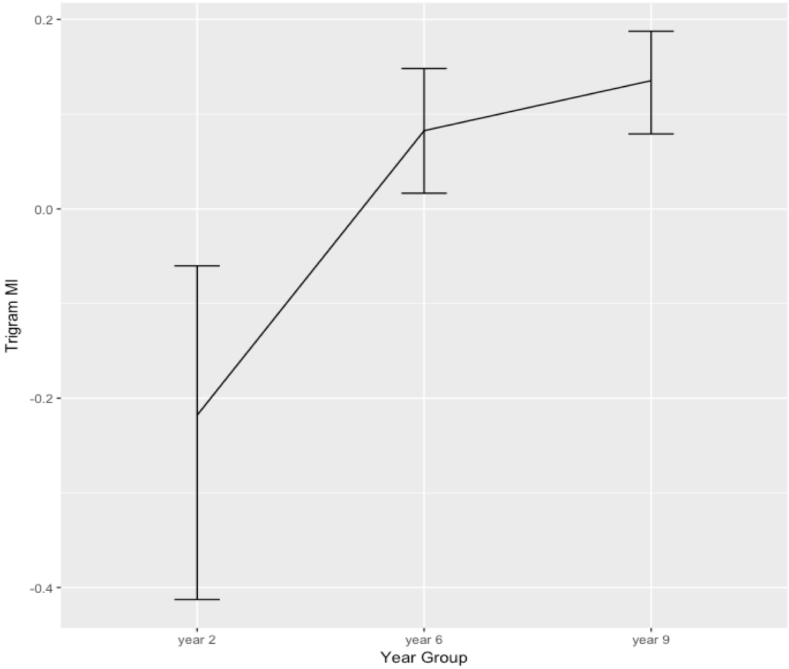
Combining results from different corpora

Category	Cronbach’s alpha	Deleted
MI	.94	COCA Academic Tri 1: .58 COCA Academic Tri 2: .63
MI2	.97	
t-score	.96	COCA Academic Tri 1: .61 COCA Academic Tri 2: .66
Delta-P	.91	COCA Academic Tri 1: .63 COCA Academic Tri 2: .68
Collexeme	.97	COCA Academic Tri 1: .69 COCA Academic Tri 2: .63

	MI	MI2	t-score	Delta-P	Collexeme
MI	1.00				
MI2	0.99	1.00			
t-score	0.66	0.66	1.00		
Delta-P	0.42	0.44	0.42	1.00	
Collexeme	0.45	0.46	0.89	0.40	1.00

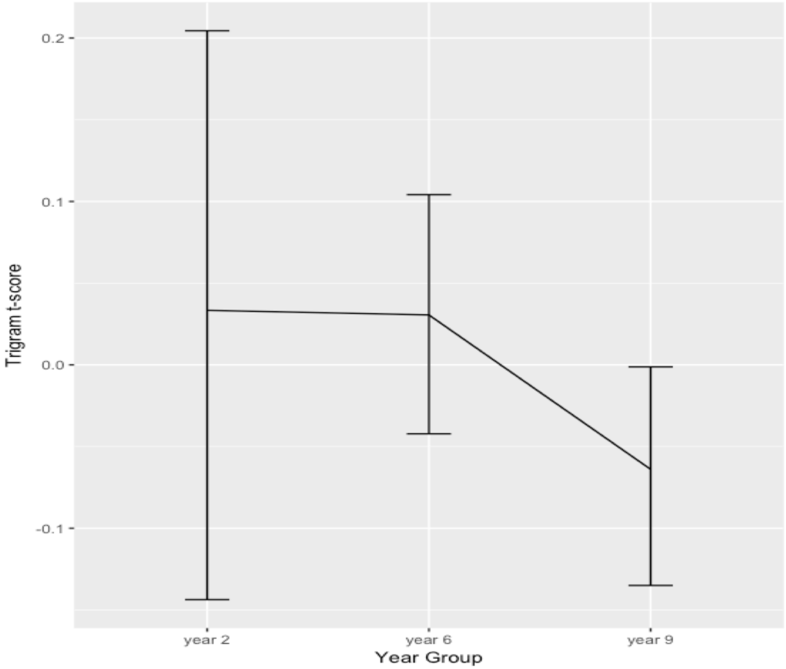
Trigram associations

Mutual Information



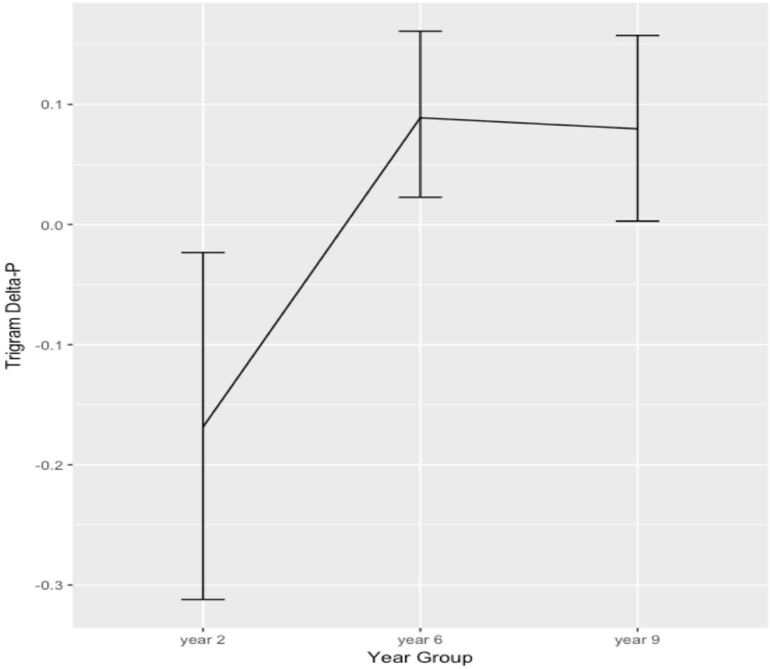
Age: Year 6: *b*=.35, *t*(78)=3.01, *p*<.005;
Year 9: *b*=.38, *t*(78)=2.76, *p*<.01
Genre: non-literary: *b*=.02, *t*(78)=.17, *p*>.05
(random intercept for task)

t-score



Age: Year 6: *b*=.03, *t*(78)=.28, *p*>.05;
Year 9: *b*=-.1, *t*(78)=-.80, *p*>.05
Genre: non-literary: *b*=.00, *t*(78)=-.03, *p*>.05
(random intercept for task)

Delta-p



Age: Year 6: *b*=.40, *t*(78)=3.35, *p*<.005;
Year 9: *b*=-.32, *t*(78)=2.35, *p*<.05
Genre: non-literary: *b*=-.14, *t*(78)=-1§.39, *p*>.05
(random intercept for task)

Conclusions - methodological

- Counts from different reference corpora mostly consistent
- Log frequencies enable patterns to emerge more clearly

Conclusions: word frequency

- Mean content word frequency increases with age
- Mean function word frequency decreases with age

Conclusions: trigrams

- Percentage of ngrams attested in corpora increases with age
- MI & DP of attested trigrams increase with age

So…

- Corpus research under-exploited in study of L1 English writing development.
- Our corpus to be completed early 2018
- Online late 2018
- Full analyses of vocabulary, NP-expansion, subordination soon.
- And the book of the lit. review…
- Future prospects:
    - Historical oral/written corpus
    - Studies of attainment

Keep in touch!

- Twitter: @growing grammar
- Facebook: facebook.com/growthingrammar
- Email: p.l.durrant@exeter.ac.uk

References

- Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282-311.
- Department of Education (2014). *The national curriculum in England: Framework document. December 2014.*
- Finn, P. J. (1977). Computer-aided description of mature word choices in writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, Measuring, Judging* (pp. 69-89). Urbana, Illinois: National Council of Teachers of English.
- Malvern, D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development*. Basingstoke: Palgrave Macmillan.
- Olinghouse, N., G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrarive writing quality in second- and fourth-grade students. *Reading and Writing, 22*, 545-565.
- Sun, Y., Zhang, J., & Scardamalia, M. (2010). Knowledge building and vocabulary growth over two years, Grades 3 and 4. *Instructional Science, 38*, 147-171.