

The Children and Young People's Books Lexicon (CYP-LEX): How does book language change as children transition into and through adolescence?

Maria Korochkina¹ | Marco Marelli² | Marc Brysbaert³ | Kathy Rastle¹

¹Department of Psychology, Royal Holloway University of London, UK

²Department of Psychology, University of Milano-Bicocca, Italy

³Department of Experimental Psychology, Ghent University, Belgium



EPS meeting
Swansea University

July 12, 2023

The CYP-LEX project

Why do we need a children's books corpus?

The CYP-LEX project

Why do we need a children's books corpus?

- Hard to overestimate the importance of literacy for an individual's prosperity

The CYP-LEX project

Why do we need a children's books corpus?

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on
 - ▶ how children learn to read and how they can best be taught [1]

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on
 - ▶ how children learn to read and how they can best be taught [1]
 - ▶ the prerequisites for becoming a skilled reader [2–6]

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on
 - ▶ how children learn to read and how they can best be taught [1]
 - ▶ the prerequisites for becoming a skilled reader [2–6]
 - e.g., amount of print exposure & of high-quality lexical representations, morphological knowledge

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on
 - ▶ how children learn to read and how they can best be taught [1]
 - ▶ the prerequisites for becoming a skilled reader [2–6]
 - e.g., amount of print exposure & of high-quality lexical representations, morphological knowledge
- Crucially, the speed with which children gain reading expertise depends on the *nature of language* they are exposed to

- Hard to overestimate the importance of literacy for an individual's prosperity
- Large body of scientific knowledge on
 - ▶ how children learn to read and how they can best be taught [1]
 - ▶ the prerequisites for becoming a skilled reader [2–6]
 - e.g., amount of print exposure & of high-quality lexical representations, morphological knowledge
- Crucially, the speed with which children gain reading expertise depends on the *nature of language* they are exposed to
- However, presently, we know very little about *what* children and young people are reading

Corpus development

National reading surveys, publisher data, & book sales statistics from Amazon UK, BookTrust, Goodreads, LoveReading4Kids, etc.

National reading surveys, publisher data, & book sales statistics
from Amazon UK, BookTrust, Goodreads, LoveReading4Kids, etc.



1,200 popular fiction & non-fiction e-books
400 books per age band

Corpus development

National reading surveys, publisher data, & book sales statistics from Amazon UK, BookTrust, Goodreads, LoveReading4Kids, etc.



1,200 popular fiction & non-fiction e-books
400 books per age band

7–9



10–12



13+



Corpus development

National reading surveys, publisher data, & book sales statistics from Amazon UK, BookTrust, Goodreads, LoveReading4Kids, etc.



1,200 popular fiction & non-fiction e-books
400 books per age band

7–9



10–12



13+



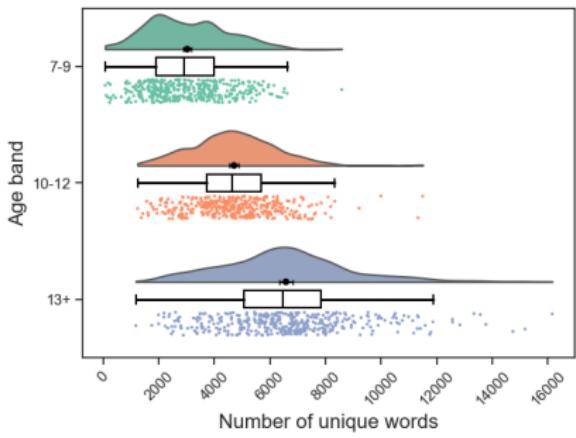
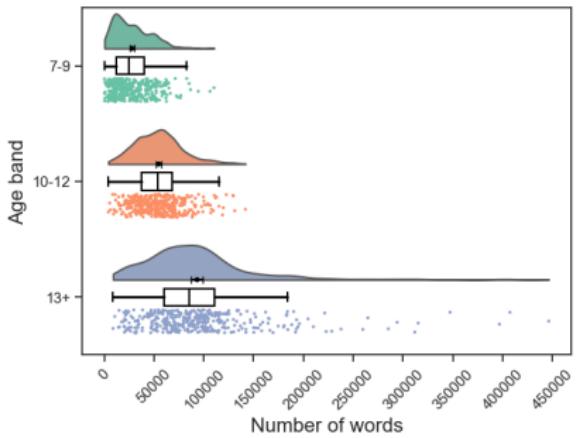
Cleaning, tokenisation, lemmatisation, PoS-tagging...

The CYP-LEX corpus

70,287,217 tokens & 105,694 types

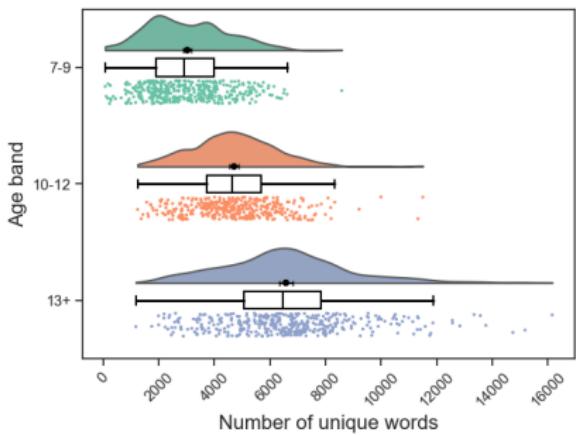
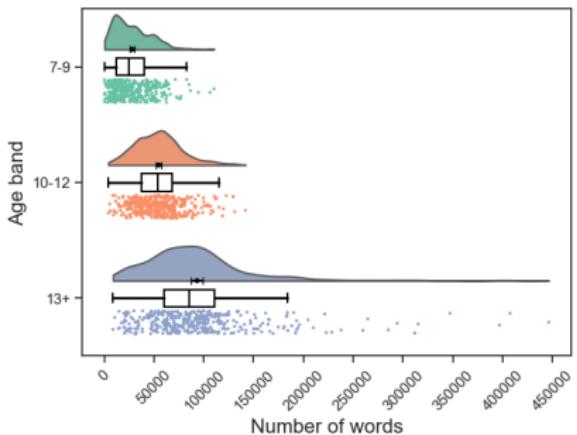
The CYP-LEX corpus

70,287,217 tokens & 105,694 types



The CYP-LEX corpus

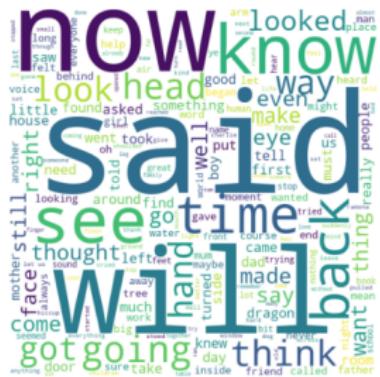
70,287,217 tokens & 105,694 types



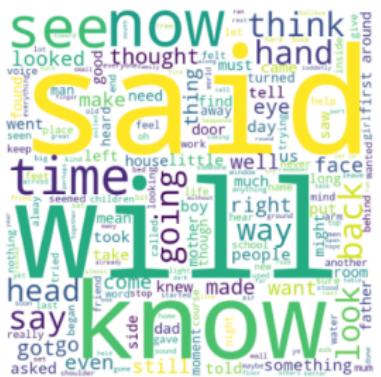
	7-9	10-12	13+
N words	11,162,653	21,837,794	37,286,770
Average N (σ) words per book	27,907 (19,212)	54,594 (24,012)	93,217 (57,718)
N unique words	52,851	70,945	90,980
Average N (σ) unique words per book	3,028 (1,452)	4,713 (1,550)	6,447 (2,366)

Some words occur very widely...

7-9



10-12



13+



...and amount to half of the corpus

...and amount to half of the corpus

" " " "
The to too.
" "
" !"
The is a with !"
This be on the !"
now, was if there was he could do?
Could be intc a ? There was only way
to out.
The n't know he in and had, in ,
that very was " , !".
So, the in his head, he out the of the .
" "
then in the that would his .
The her head and looked at the as if he was .
by this, The .
" "
, and she her and her . Then a her
" "
A out, again the with .
" with that !"
" you'll be it to the !"
"Or for the !"
" "
could , but he was it was more .
" "
This time the . in!
" "

...and amount to half of the corpus

" " " "
The to too.
" "
" , !"
"The is a with !"
"This be on the !"
now, was if there was he could do?
Could be intc a ? There was only way
to out. The n't know he in and had, in .
that very was " , !".
So, the in his head, he out the of the .
" "
then in the that would his .
The her head and looked at the as if he was .
by this, The .
" "
, and she her head to the . Then a her .
" and she her and her .
A out, again the with .
" with that !" a from .
" you'll be it to the !"
"Or for the !"
" "
" could , but he was it was more .
" "
This time the in!
" "

"HA! HA! HA!"
The stern-faced crowd began to chuckle too.
"HO! HO! HO!"
"Well played, boy!"
"The child is a marvel with animals!"
"This pair should be on the stage!"
Feeling ten-foot tall now, Eric was wondering if there was something else he could do? Could these raspberries be blown into something resembling a tune? There was only one way to find out.
The boy didn't know many songs. One he often sang in school assembly and had, in fact, sung that very morning was "Rule, Britannia!".
So, rejoining the tune in his head, he began raspberrying out the notes of the chorus.
"PFFFFT! PFT! PFT! PFT!"
Eric then fell silent in the hope that Gertrude would follow his lead.
The gorilla tilted her head and looked at the boy as if he was barmy.
Undeterred by this, Eric persisted. The boy repeated himself.
"PFFFFT! PFT! PFT! PFT!"
Gertrude tilted her head to the other side. Then a mischievous thought flashed across her eyes, and she pursed her lips together and pushed her tongue forward.
"PFFFFFFF! PFT!"
A long, low raspberry came out, once again covering the boy with gorilla sputtle.
"Good luck with that one, lad!" snorted a voice from behind.
"Next you'll be teaching it to play the piano!"
"Or dance for the Royal Ballet!"
"HA! HA! HA!"
Eric could sense people ebbing away, but he was sure it was worth one more try.
"PFFFFT! PFT! PFT! PFT!"
This time the most wondrous thing happened. Gertrude joined in!
"PFFFFT! PFT! PFT! PFT!"

...and amount to half of the corpus

" " " "
The to too.
" "
" , !"
"The is a with !"
"This be on the !"
now, was if there was he could do?
Could be intc a ? There was only way
to out. The n't know he in and had, in .
that very was " , !".
So, the in his head, he out the of the .
" "
then in the that would his .
The her head and looked at the as if he was .
by this, The .
" "
, and she her head to the . Then a her .
" and she her and her .
A out, again the with .
" with that !" a from .
" you'll be it to the !"
"Or for the !"
" "
could , but he was it was more .
" "
This time the . in!

"HA! HA! HA!"
The stern-faced crowd began to chuckle too.
"HO! HO! HO!"
"Well played, boy!"
"The child is a marvel with animals!"
"This pair should be on the stage!"
Feeling ten-foot tall now, Eric was wondering if there was something else he could do? Could these raspberries be blown into something resembling a tune? There was only one way to find out.
The boy didn't know many songs. One he often sang in school assembly and had, in fact, sung that very morning was "Rule, Britannia!".
So, rejoining the tune in his head, he began raspberrying out the notes of the chorus.
"PFFT! PFT! PFT! PFT!"
Eric then fell silent in the hope that Gertrude would follow his lead.
The gorilla tilted her head and looked at the boy as if he was barmy.
Undeterred by this, Eric persisted. The boy repeated himself.
"PFFT! PFT! PFT! PFT!"
Gertrude tilted her head to the other side. Then a mischievous thought flashed across her eyes, and she pursed her lips together and pushed her tongue forward.
"PFFFFFFFFFFF!"
A long, low raspberry came out, once again covering the boy with gorilla sputtle.
"Good luck with that one, lad!" snorted a voice from behind.
"Next you'll be teaching it to play the piano!"
"Or dance for the Royal Ballet!"
"HA! HA! HA!"
Eric could sense people ebbing away, but he was sure it was worth one more try.
"PFFT! PFT! PFT! PFT!"
This time the most wondrous thing happened. Gertrude joined in!
"PFFT! PFT! PFT! PFT!"

...but it's the other, less common, words that make up the stories!

...yet, many of the less common words may be unfamiliar

Percentage of CYP-LEX words that children encounter on TV

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	39%	70%	91%
10–12 age band <i>N</i> = 70,945	30%	58%	86%
13+ age band <i>N</i> = 90,980	24%	48%	79%

...yet, many of the less common words may be unfamiliar

Percentage of CYP-LEX words that children encounter on TV

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	39%	70%	91%
10–12 age band <i>N</i> = 70,945	30%	58%	86%
13+ age band <i>N</i> = 90,980	24%	48%	79%

...yet, many of the less common words may be unfamiliar

Percentage of CYP-LEX words that children encounter on TV

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	39%	70%	91%
10–12 age band <i>N</i> = 70,945	30%	58%	86%
13+ age band <i>N</i> = 90,980	24%	48%	79%

Book language more complex than that on TV?

Word frequency correlations for shared words

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	.67	.77	.72
10–12 age band <i>N</i> = 70,945	.63	.75	.76
13+ age band <i>N</i> = 90,980	.58	.72	.76

Book language more complex than that on TV?

Word frequency correlations for shared words

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	.67	.77	.72
10–12 age band <i>N</i> = 70,945	.63	.75	.76
13+ age band <i>N</i> = 90,980	.58	.72	.76

Book language more complex than that on TV?

Word frequency correlations for shared words

	Cbeebies 0–6 years <i>N</i> = 27,236	CBBC 6–12 years <i>N</i> = 58,691	SUBTLEX-UK adults <i>N</i> = 160,024
7–9 age band <i>N</i> = 52,851	.67	.77	.72
10–12 age band <i>N</i> = 70,945	.63	.75	.76
13+ age band <i>N</i> = 90,980	.58	.72	.76

Many new words in each band

25,627 new words in 10–12 compared to 7–9

Many new words in each band

25,627 new words in 10–12 compared to 7–9

73% encountered \leq 3 times

Many new words in each band

25,627 new words in 10–12 compared to 7–9

73% encountered \leq 3 times

osteichthyes christmastides frazzles
unconsuming traceries strength
morello chairmanship darfur
scath georgians reclusiveness deplete
unirrigated unmysterious calvaire
polygons unredeemable preternaturally stalklike
earworms reeks pinpricking cheapened dunnocks langleys
ibsen laise islandless schmuck littlejohn handley
ebbtide crotched unreel summersets ismael
chaises warne unoriginal tabula resection
beguilingly kleiber ramified enrols wroth catheters
lightship dustin knobstick fairmont ingres
avocations libeled retre
quia inrushing biomass smokable mortem kirkyard
laughers garble zebrawood emboldening
detouring mesopotamians choriambs donas runch
coexisting toweling turnstone sapwood transcaucasian
merde rethreaded reappraised flowerets inkpots
mistrustfulness disgorger mudholes salvagers
dereliction hantray tweeddale

Many new words in each band

25,627 new words in 10–12 compared to 7–9

73% encountered \leq 3 times

1% encountered \geq 100 times

A circular arrangement of words from the 10-12 band, color-coded by frequency. The words are:

- osteichthyes
- christmastides
- frazzles
- morello
- traceries
- unconsuming
- chairmanship
- strength
- undervests
- deplete
- scath
- georgians
- reclusiveness
- darfur
- unirrigated
- unmysterious
- calvaire
- polygons
- rexamine
- pinpricking
- cheapened
- dunnocks
- stalklike
- merth
- preternaturally
- littlejohn
- langleys
- earworms
- ibsen
- liaise
- islandless
- schnuck
- resection
- ismael
- crotched
- ramified
- unreel
- summersets
- enrolls
- wroth
- catheeters
- chaises
- ebbtide
- warne
- unoriginal
- tabla
- detent
- fairmont
- ingres
- beguilingly
- dustin
- libeled
- knobstick
- mortem
- kirkyard
- lightship
- kleider
- regicide
- rete
- smokable
- fairmont
- dirts
- torbay
- krug
- avocations
- englishness
- benefaction
- zebrawood
- emboldening
- mesopotamians
- donas
- runch
- charily
- quia
- inrushing
- laughers
- garble
- choriambuses
- flowerets
- transcaucasian
- detouring
- goneness
- toweling
- turnstone
- reappraised
- sawwood
- ashmolean
- inkpots
- coexisting
- merde
- mistrustfulness
- disgorger
- mudholes
- salvagers
- tweeddale
- dereliction

Many new words in each band

25,627 new words in 10–12 compared to 7–9

73% encountered \leq 3 times

1% encountered \geq 100 times

osteichthyes christmastides
morello traceries frazzles
unconsuming chairmanship strength
scath georgians reclusiveness undervests deplete
polygons unirrigated unmysterious darfur
earworms reexamine pinpricking cheapened dunnocks stalklike
ibsen merth preternaturally langley
ebbtide liaise islandless schmuck littlejohn ismael
chaises warne crotched summersets handley
beguilingly kleiber ramified tabla rete enrolls wroth catheters
lightship regicide biomass smokable mortem kirkyard
avocations englishness zebrawood emboldening
quia inrushing garble signa choriambs donas runch
detouring laughers mesopotamians flowersn sapwood transcaucasian
coexisting toweling merde rethreaded reappraised mudholes inkpots
merde mistrustfulness disgorger hanratty salvagers tweeddale

sian johanna gawain elzevir bootmaker jappet gunther shalmuel lincolns binny erwin etienne amara rikki inigo lynde lindi barnard zsa nain strauss saj uthere bree fae estella beamish daddar rissa tinta katar jt indio tyrell syd tyrrell nestha parfitt republicans mau mahoney annet allison cobb inquisitors alexei virgos darcy dimmock whiteley fabian trig callie doggett vizier pepe applegate strauss salma whiteley randolph salma

Many new words in each band

31,025 new words in 13+ compared to 10–12

Many new words in each band

31,025 new words in 13+ compared to 10–12

74% encountered \leq 3 times

Many new words in each band

31,025 new words in 13+ compared to 10–12

74% encountered \leq 3 times

outshooting hemophilic cubiculo privatisation
remodelled lezzies infanticides valborg pontificating
squabs teethlike Dempster cankerworm bratlings
rhapsodical glasses outhers eliding crappier
supinely suitor empathizing prebought gathings
olid porringer habitus depositories myocarditis gutturalis
mcpail phymic init panniered putain emceed
caimans sphymic ucr ephesus skimping swahilis
musts parzial widener snoozed pollack liaison
shirred jousc overford oblong, magnetical hassett
earlyish unprizable simbirk dominica sellable
deny tipplers studious procyonidean ramification
slickened doddypoll cantons expropriating cbd
sexagenarians stomatitis mendoza fibreboard
prototyping divisa galí graticles floodwall
housemother manteca avoiders vassalage
overarches landhold lawtons homozgous
retransfer

Many new words in each band

31,025 new words in 13+ compared to 10–12

74% encountered \leq 3 times

1% encountered \geq 100 times

outshooting hemophilic cubiculo privatisation
remoided lezzies infanticides valborg pontificating
squats teethlike Dempster cankerworm bratlings
rhapsodical glasses othumers eliding crappier
supinely sultrier prebought gathings
olid porringer habitus depositories myocarditis gutturalis
mcpheil sphymic init panniered putain emceed
caimans phrygian ucr ephesus skimping swahilis
musts parzial widener dfa snoozed pollack liaison
shirred jouse overfond oblong libellers reactant sellable
earlyish unprizable simborsk magnetical hassett
tipplers studious procyonidean dominica ramification
deny doddypoll cantons expropriating cbd floodwall
slickened stomatis mendoza feebroad vassalage
sexagenarians prototyping galis graticles homozygous
divisa housemother overarches avoiders ni statement
manteca landhold lawtons retransfer

Many new words in each band

31,025 new words in 13+ compared to 10–12

74% encountered \leq 3 times



1% encountered \geq 100 times

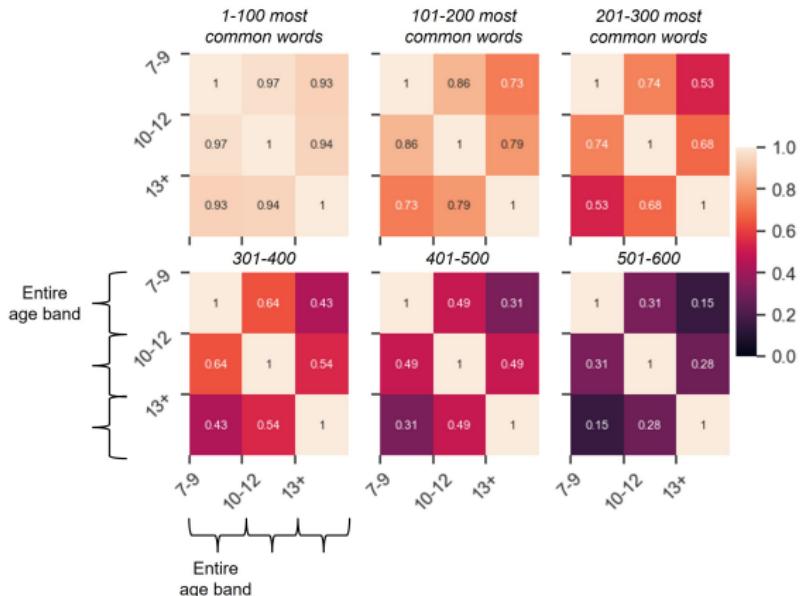


Do the three age bands use the same vocabulary?

600 most common words in sets of 100

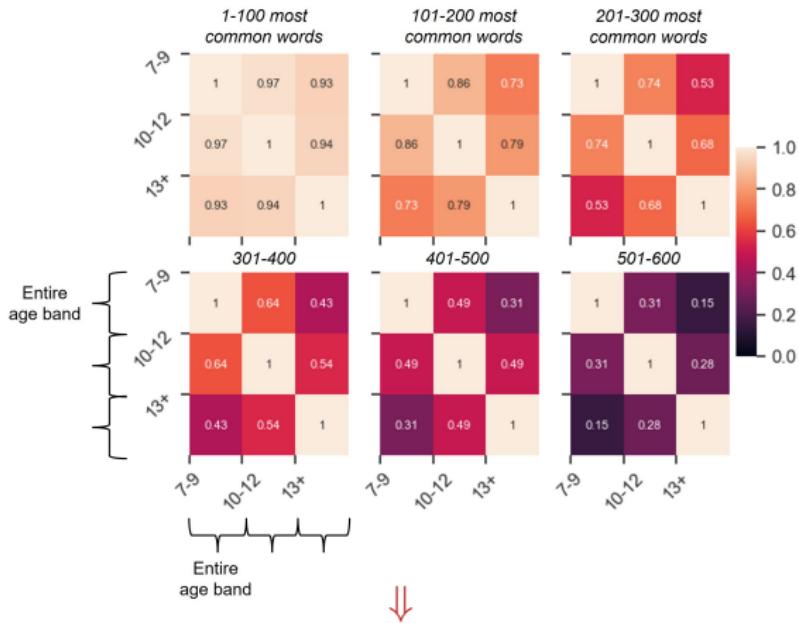
Do the three age bands use the same vocabulary?

600 most common words in sets of 100



Do the three age bands use the same vocabulary?

600 most common words in sets of 100



Yes, but only in terms of the 200–300 most common words

Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

A **lemma** is the unmarked form of a set of inflected word forms

Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

A **lemma** is the unmarked form of a set of inflected word forms

- go, goes, going, went, gone → go

Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

A **lemma** is the unmarked form of a set of inflected word forms

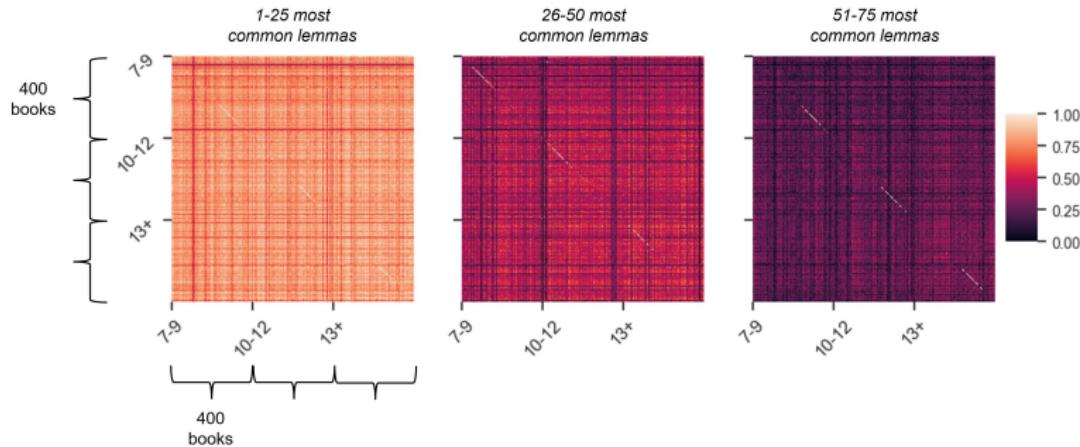
- go, goes, going, went, gone → go
- he, his, him → he

Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

A **lemma** is the unmarked form of a set of inflected word forms

- go, goes, going, went, gone → go
- he, his, him → he

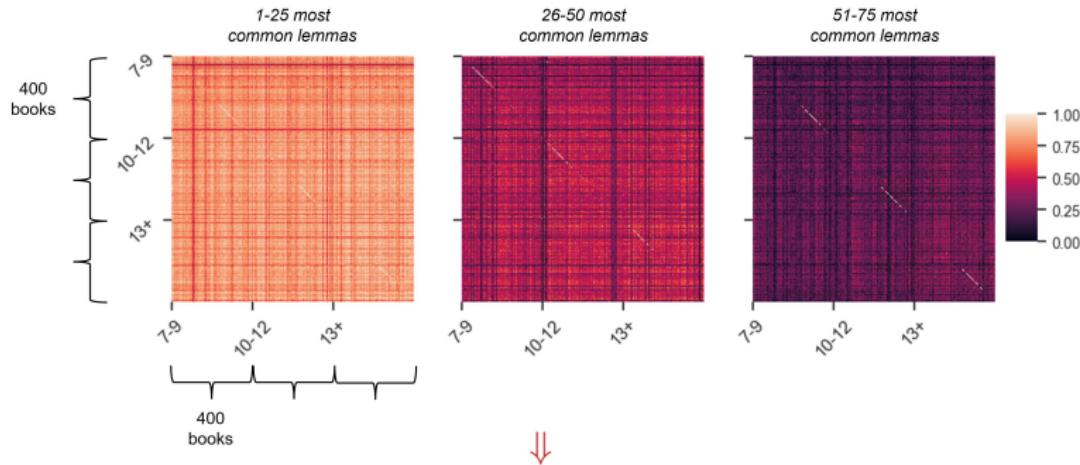


Do individual books use similar vocabulary?

75 most common *lemmas* in sets of 25

A **lemma** is the unmarked form of a set of inflected word forms

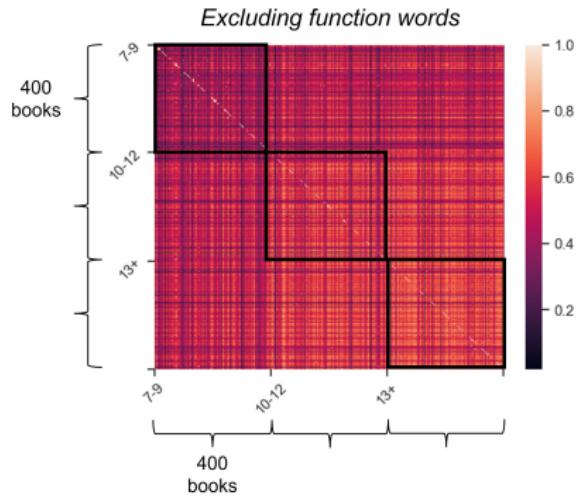
- go, goes, going, went, gone → go
- he, his, him → he



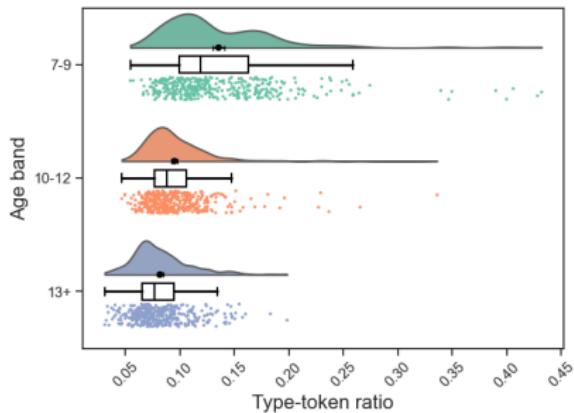
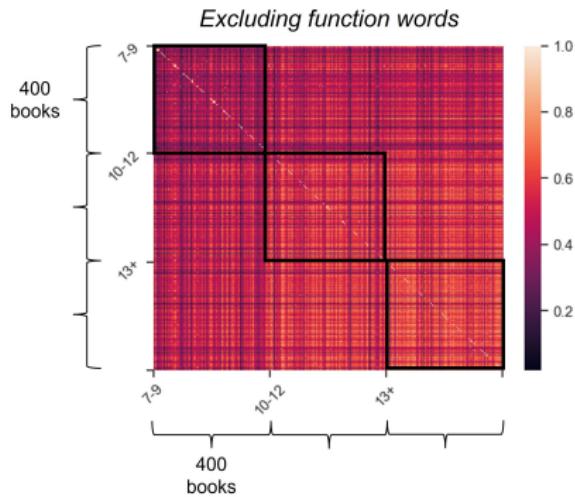
Similar in terms of their most frequent lemmas but rapidly diverge

Intensity of vocabulary in the 7–9 age band

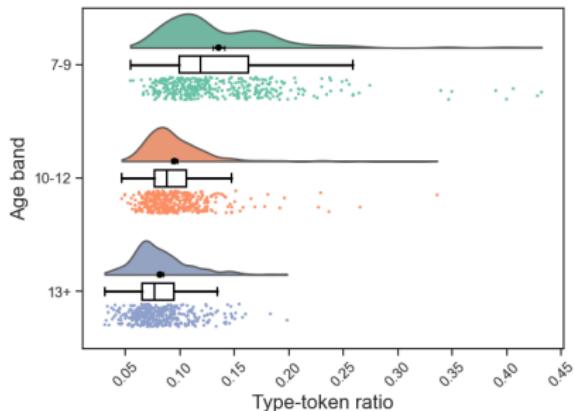
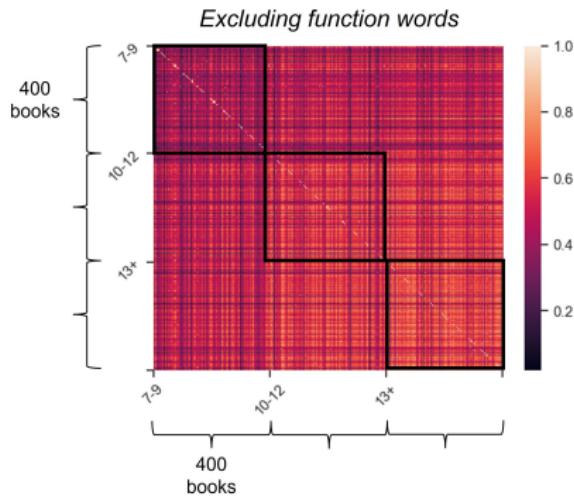
Intensity of vocabulary in the 7–9 age band



Intensity of vocabulary in the 7–9 age band



Intensity of vocabulary in the 7–9 age band



Books in the 7–9 age band are less similar to one another than those in the other age bands

Conclusions

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge
 - It is crucial to develop reading skills and motivation

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge
 - It is crucial to develop reading skills and motivation
- Vast number of (new) & morphologically complex words in children's books

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge
 - It is crucial to develop reading skills and motivation
- Vast number of (new) & morphologically complex words in children's books
 - Importance of tools to decode these words, e.g., morphological knowledge

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge
 - It is crucial to develop reading skills and motivation
- Vast number of (new) & morphologically complex words in children's books
 - Importance of tools to decode these words, e.g., morphological knowledge
- Beyond function words, children's books have low similarity to one another

- Initial experience of independent reading (7–9 age band) is intense & vitally important for building word knowledge
 - It is crucial to develop reading skills and motivation
- Vast number of (new) & morphologically complex words in children's books
 - Importance of tools to decode these words, e.g., morphological knowledge
- Beyond function words, children's books have low similarity to one another
 - Reading widely is key

Thank you!

- [1] A. Castles, K. Rastle, and K. Nation, “Ending the Reading Wars: Reading acquisition from novice to expert,” *Psychological Science*, vol. 19, no. 1, pp. 5–51, 2018. DOI: <https://doi.org/10.1177/1529100618772271>.
- [2] A. Castles, C. Davis, P. Cavalot, and K. Forster, “Tracking the acquisition of orthographic skills in developing readers: Masked priming effects,” *Journal of Experimental Child Psychology*, vol. 97, pp. 165–182, 2007. DOI: <https://doi.org/10.1016/j.jecp.2007.01.006>.
- [3] S. E. Mol and A. G. Bus, “To read or not to read: A metaanalysis of print exposure from infancy to early adulthood,” *Psychological Bulletin*, vol. 137, pp. 267–296, 2017. DOI: <https://doi.org/10.1037/a0021890>.

- [4] K. Nation, "Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill," *npj Science of Learning*, vol. 2, pp. 1–4, 2017. DOI: <https://doi.org/10.1038/s41539-017-0004-7>.
- [5] K. Rastle, "The place of morphology in learning to read in english," *Cortex*, vol. 116, pp. 45–54, 2019. DOI: <https://doi.org/10.1016/j.cortex.2018.02.008>.
- [6] C. A. Perfetti and L. Hart, "The lexical quality hypothesis," in *Precursors of Functional Literacy*, L. Verhoeven, C. Elbr, and P. Reitsma, Eds., John Benjamins, 2002, pp. 189–212. DOI: <https://doi.org/10.1037/a0021890>.