CYP-LEX:
A novel large-scale lexical database of books for children and young people

Maria Korochkina | **Marco Marelli** | **Marc Brysbaert** | **Kathy Rastle**

Royal Holloway University of London

May 31, 2023

How it all started...

What is morphology and why is it important?

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought

- Most words are built by combining units of meaning (morphemes) which contribute *systematically* to the meaning of the word

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought
- Most words are built by combining units of meaning (morphemes) which contribute *systematically* to the meaning of the word
  - 'un-' + 'clean' → 'unclean' (not clean)

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought

- Most words are built by combining units of meaning (morphemes) which contribute *systematically* to the meaning of the word
  - 'un-' + 'clean' → 'unclean' (not clean)
  - [stem] + '-ify' → 'make [stem]' (e.g., simplify, purify)

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought

- Most words are built by combining units of meaning (morphemes) which contribute *systematically* to the meaning of the word
  - ‘un-’ + ‘clean’ → ‘unclean’ (not clean)
  - [stem] + ‘-ify’ → ‘make [stem]’ (e.g., simplify, purify)

- This combinatorial system allows us to *generalise* and provides a heuristic for vocabulary growth

How it all started...

What is morphology and why is it important?

- Language is a *productive* system in which a *finite* set of elements can be combined to express a *limitless* range of thought

- Most words are built by combining units of meaning (morphemes) which contribute *systematically* to the meaning of the word
  - 'un-' + 'clean' → 'unclean' (not clean)
  - [stem] + '-ify' → 'make [stem]' (e.g., simplify, purify)

- This combinatorial system allows us to *generalise* and provides a heuristic for vocabulary growth
  - 'quick' + '-ify' → 'quickify'

Morphemic knowledge

Not as easy as it seems...

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation
- Their function can & must be inferred from experience with whole words (e.g., solidify, amplify, justify)

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation
- Their function can & must be inferred from experience with whole words (e.g., solidify, amplify, justify)
- However, morphemic regularities are not as regular as one might think...

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation
- Their function can & must be inferred from experience with whole words (e.g., solidify, amplify, justify)

- However, morphemic regularities are not as regular as one might think...
- There is substantial variability in precision & consistency with which a morpheme conveys meaningful information

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation
- Their function can & must be inferred from experience with whole words (e.g., solidify, amplify, justify)

- However, morphemic regularities are not as regular as one might think...
- There is substantial variability in precision & consistency with which a morpheme conveys meaningful information
  - 'artist', 'typist' vs. 'racist', 'sadist'

Morphemic knowledge

Not as easy as it seems...

- Affixes alter words' meanings in highly predictable ways [1, 2] & do not occur in isolation
- Their function can & must be inferred from experience with whole words (e.g., solidify, amplify, justify)

- However, morphemic regularities are not as regular as one might think...
- There is substantial variability in precision & consistency with which a morpheme conveys meaningful information
  - 'artist', 'typist' vs. 'racist', 'sadist'

→ **What** do we need to learn and **how**?

The perks of reading

...or why we need a new corpus

The perks of reading

…or why we need a new corpus

- Most new words children aged 8+ acquire are encountered through reading [3]

The perks of reading

…or why we need a new corpus

- Most new words children aged 8+ acquire are encountered through reading [3]
- In adults, morphological knowledge is modulated by reading experience [4]

The perks of reading

...or why we need a new corpus

- Most new words children aged 8+ acquire are encountered through reading [3]
- In adults, morphological knowledge is modulated by reading experience [4]
- In English, morphological information is far more salient in spelling than it is in spoken language (e.g., ‘sign’ vs. ‘signature’; ‘active’ vs. ‘activity’) [5]

The perks of reading

...or why we need a new corpus

- Most new words children aged 8+ acquire are encountered through reading [3]
- In adults, morphological knowledge is modulated by reading experience [4]
- In English, morphological information is far more salient in spelling than it is in spoken language (e.g., ‘sign’ vs. ‘signature’; ‘active’ vs. ‘activity’) [5]
- Yet, available corpora are too small, target narrow age cohorts, comprise outdated reading materials, or have restricted access

The perks of reading

...or why we need a new corpus

- Most new words children aged 8+ acquire are encountered through reading [3]
- In adults, morphological knowledge is modulated by reading experience [4]
- In English, morphological information is far more salient in spelling than it is in spoken language (e.g., 'sign' vs. 'signature'; 'active' vs. 'activity') [5]
- Yet, available corpora are too small, target narrow age cohorts, comprise outdated reading materials, or have restricted access

↓

We need a large-scale publicly available corpus of books that children and young people read!

Corpus development

...or how, in the blink of an eye, a year was gone...

Corpus development

...or how, in the blink of an eye, a year was gone...

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

Corpus development

...or how, in the blink of an eye, a year was gone...

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

$\Downarrow$

1,200 popular fiction & non-fiction e-books
400 books per age band (7-9, 10-12, 13+)

Corpus development

...or how, in the blink of an eye, a year was gone...

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

$\Downarrow$

1,200 popular fiction & non-fiction e-books
400 books per age band (7-9, 10-12, 13+)

$\Downarrow$

Semi-automatic conversion, cleaning, & pre-processing

Corpus development

...or how, in the blink of an eye, a year was gone...

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

⇓

1,200 popular fiction & non-fiction e-books
400 books per age band (7-9, 10-12, 13+)

⇓

Semi-automatic conversion, cleaning, & pre-processing

⇓

Tokenisation, lemmatisation, Part-of-Speech tagging
71,296,249 tokens & 323,670 types

Corpus development

...or how, in the blink of an eye, a year was gone...

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

$\Downarrow$

1,200 popular fiction & non-fiction e-books
400 books per age band (7-9, 10-12, 13+)

$\Downarrow$

Semi-automatic conversion, cleaning, & pre-processing

$\Downarrow$

Tokenisation, lemmatisation, Part-of-Speech tagging
71,296,249 tokens & 323,670 types

$\Downarrow$

Invalid token removal

Corpus development

...or how, in the blink of an eye, a year was gone...

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

⇓

1,200 popular fiction & non-fiction e-books
400 books per age band (7-9, 10-12, 13+)

⇓

Semi-automatic conversion, cleaning, & pre-processing

⇓

Tokenisation, lemmatisation, Part-of-Speech tagging
71,296,249 tokens & 323,670 types

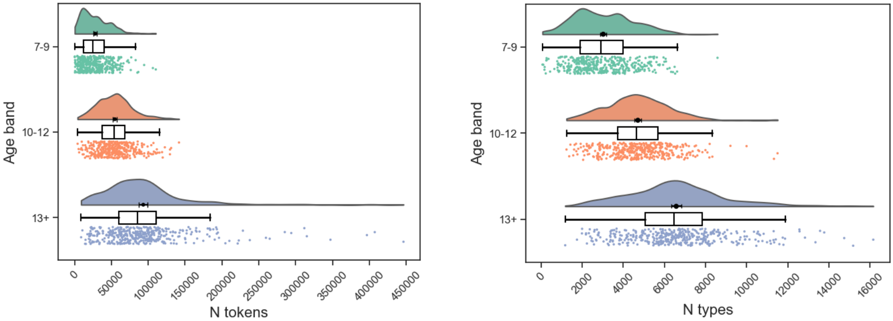
⇓

Invalid token removal

⇓

CYP-LEX: Children and Young People's Books Lexicon
70,287,217 tokens & 105,694 types

CYP-LEX size



	CYP-LEX 7–9	**CYP-LEX 10–12**	**CYP-LEX 13+**
N tokens	11,162,653	21,837,794	37,286,770
μ (σ) tokens	27,906.63 (19,212.43)	54,594.48 (24,011.91)	93,216.92 (57,718.38)
N types	52,851	70,945	90,980
μ (σ) types	3,027.81 (1,452.05)	4,712.81 (1,550.43)	6,446.57 (2,365.59)

Corpus validation

CYP-LEX vs. other corpora

	Cbeebies 0–6 years $N = 27{,}236$		**CPWD** 5–9 years $N = 12{,}452$		**CBBC** 6–12 years N = *58,691*		**SUBTLEX-UK** Adults $N = 160{,}024$	
	% shared	r	% shared	r	% shared	r	% shared	r
CYP-LEX 7–9 $N = 52{,}851$	39%	.67	19%	.71	70%	.77	91%	.72
CYP-LEX 10–12 $N = 70{,}945$	30%	.63	14%	.68	58%	.75	86%	.76
CYP-LEX 13+ $N = 90{,}980$	24%	.58	11%	.62	48%	.72	79%	.76

Corpus validation

CYP-LEX vs. other corpora

	Cbeebies 0–6 years $N = 27{,}236$		**CPWD** 5–9 years $N = 12{,}452$		**CBBC** 6–12 years $N = \mathit{58{,}691}$		**SUBTLEX-UK** Adults $N = 160{,}024$	
	% shared	r	% shared	r	% shared	r	% shared	r
CYP-LEX 7–9 $N = 52{,}851$	39%	.67	19%	.71	70%	.77	91%	.72
CYP-LEX 10–12 $N = 70{,}945$	30%	.63	14%	.68	58%	.75	86%	.76
CYP-LEX 13+ $N = 90{,}980$	24%	.58	11%	.62	48%	.72	79%	.76

Corpus validation

CYP-LEX vs. other corpora

	Cbeebies		**CPWD**		**CBBC**		**SUBTLEX-UK**	
	0–6 years		5–9 years		6–12 years		Adults	
	$N = 27{,}236$		$N = 12{,}452$		$N = \mathit{58{,}691}$		$N = 160{,}024$	
	% shared	r	% shared	r	% shared	r	% shared	r
CYP-LEX 7–9 $N = 52{,}851$	39%	.67	19%	.71	70%	.77	91%	.72
CYP-LEX 10–12 $N = 70{,}945$	30%	.63	14%	.68	58%	.75	86%	.76
CYP-LEX 13+ $N = 90{,}980$	24%	.58	11%	.62	48%	.72	79%	.76

Corpus validation

CYP-LEX vs. other corpora

	Cbeebies		**CPWD**		**CBBC**		**SUBTLEX-UK**	
	0–6 years		5–9 years		6–12 years		Adults	
	$N = 27{,}236$		$N = 12{,}452$		$N = \mathit{58{,}691}$		$N = 160{,}024$	
	% shared	r	% shared	r	% shared	r	% shared	r
CYP-LEX 7–9 $N = 52{,}851$	39%	.67	19%	.71	70%	.77	91%	.72
CYP-LEX 10–12 $N = 70{,}945$	30%	.63	14%	.68	58%	.75	86%	.76
CYP-LEX 13+ $N = 90{,}980$	24%	.58	11%	.62	48%	.72	79%	.76

CYP-LEX 10–12 vs. CYP-LEX 7–9

25,627 unshared words

CYP-LEX 10–12 vs. CYP-LEX 7–9

25,627 unshared words

Raw frequency ≤ 3
73% ($N = 18{,}646$) of unshared words

CYP-LEX 10–12 vs. CYP-LEX 7–9

25,627 unshared words

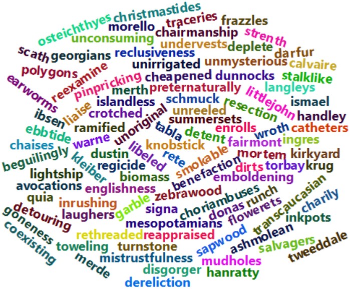
Raw frequency ≤ 3
73% ($N = 18{,}646$) of unshared words

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

CYP-LEX 10–12 vs. CYP-LEX 7–9

25,627 unshared words

Raw frequency ≤ 3
73% ($N = 18{,}646$) of unshared words



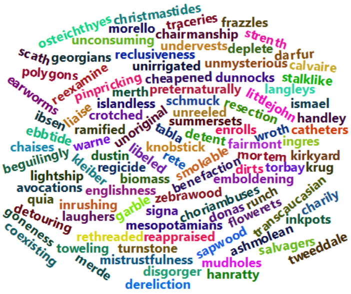
Raw frequency > 100
$< 1\%$ ($N = 249$) of unshared words

CYP-LEX 10–12 vs. CYP-LEX 7–9

25,627 unshared words

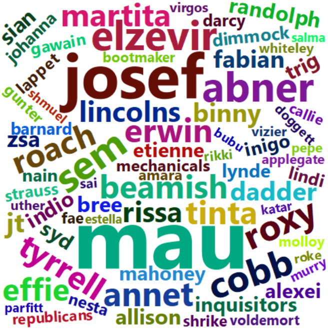
Raw frequency ≤ 3

73% ($N = 18{,}646$) of unshared words



Raw frequency > 100

$<$ 1% ($N = 249$) of unshared words



CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3

74% ($N = 22{,}855$) of unshared words

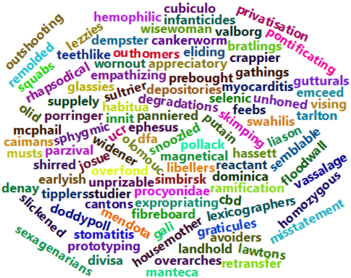
cubiculo
hemophilic
infanticides
privatisation
outshooting
lezzies
wisewoman
valborg
pontificating
dempster
cankerworm
remolded
teethlike
outhomers
eliding
bratlings
squabs
wornout
appreciatory
crappier
rhapsodical
empathizing
prebought
gathings
glassies
sultrier
depositories
myocarditis
gutturals
emceed
supplely
habitua
degradations
selenic
unhoned
vising
olid
porringer
innit
panniered
feebs
tarlton
putain
skimping
swahilis
mcphail
ephesus
caimans
sphygmic
ucr
liason
dfa
snoozled
semblable
musts
parzival
widener
olomouc
pollack
hassett
floodwall
shirred
josue
magnetical
overfond
libellers
reactant
vassalage
earlyish
simbirsk
dominica
denay
unprizable
ramification
homozygous
tipplersstudier
procyonidae
cantons
expropriating
cbd
slickened
doddypoll
fibreboard
lexicographers
misstatement
mendota
gali
stomatitis
graticules
sexagenarians
housemother
avoiders
prototyping
landhold
lawtons
divisa
overarches
retransfer
manteca

CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3
74% ($N = 22{,}855$) of unshared words

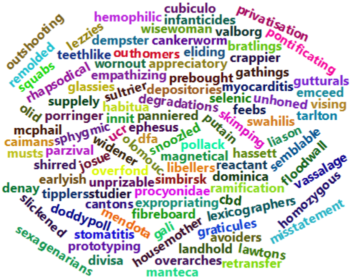
Raw frequency > 100
1% ($N = 326$) of unshared words



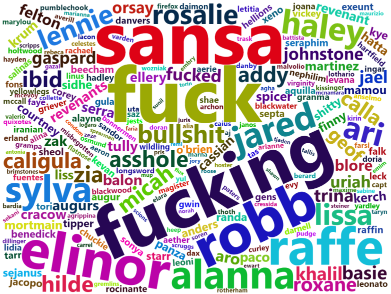
CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3
74% ($N = 22{,}855$) of unshared words



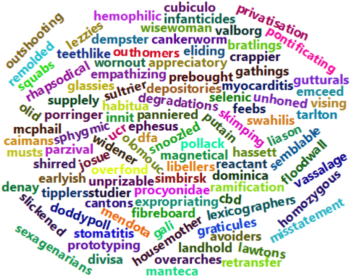
Raw frequency > 100
1% ($N = 326$) of unshared words



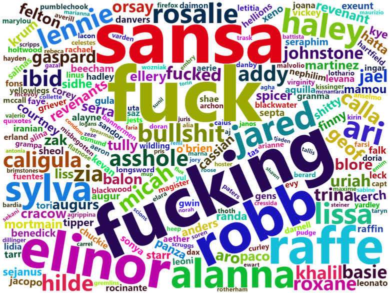
CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3
74% ($N = 22{,}855$) of unshared words



Raw frequency > 100
1% ($N = 326$) of unshared words

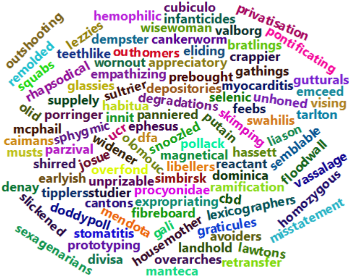


On that note...

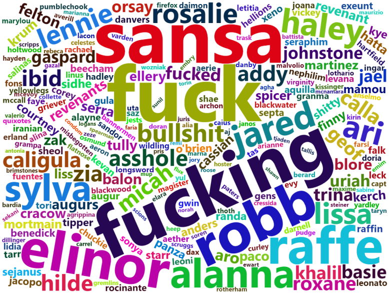
CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3
74% ($N = 22{,}855$) of unshared words



Raw frequency > 100
1% ($N = 326$) of unshared words



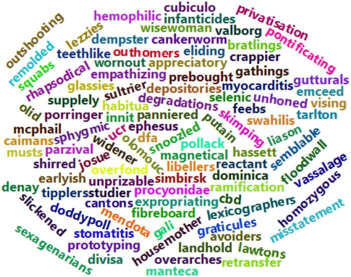
On that note...

- CYP-LEX 13+ vs. CBBC: ‘shit’ (*Zipf* = 4.92, $N = 3{,}077$)

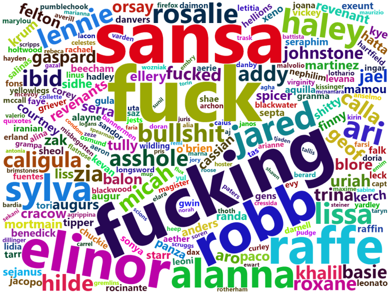
CYP-LEX 13+ vs. CYP-LEX 10–12

31,025 unshared words

Raw frequency ≤ 3
74% ($N = 22{,}855$) of unshared words



Raw frequency > 100
1% ($N = 326$) of unshared words

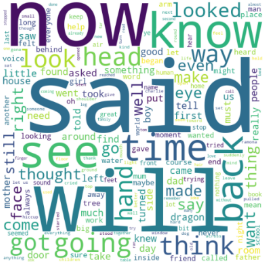


On that note...

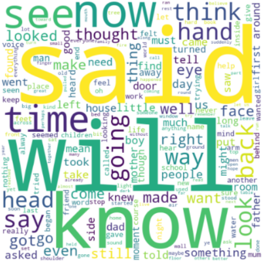
- CYP-LEX 13+ vs. CBBC: ‘shit’ (*Zipf* $= 4.92$, $N = 3{,}077$)
- CYP-LEX 13+ vs. CPWD: ‘hell’ (*Zipf* $= 5.21$, $N = 6{,}103$)

Wait a minute... aren't they identical?

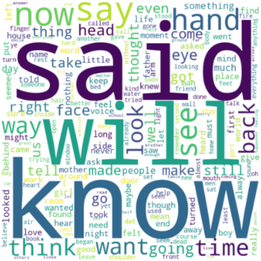
7–9



10–12

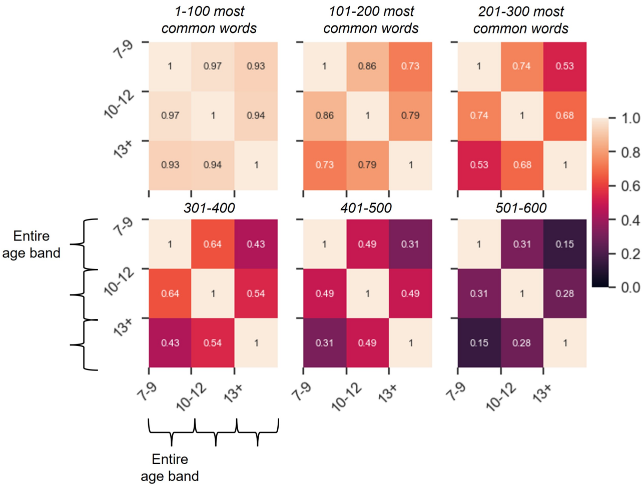


13+



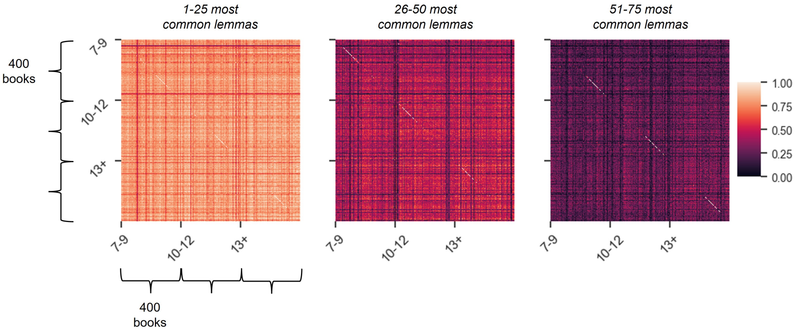
Semantic similarity across the age bands

600 most common words in sets of 100

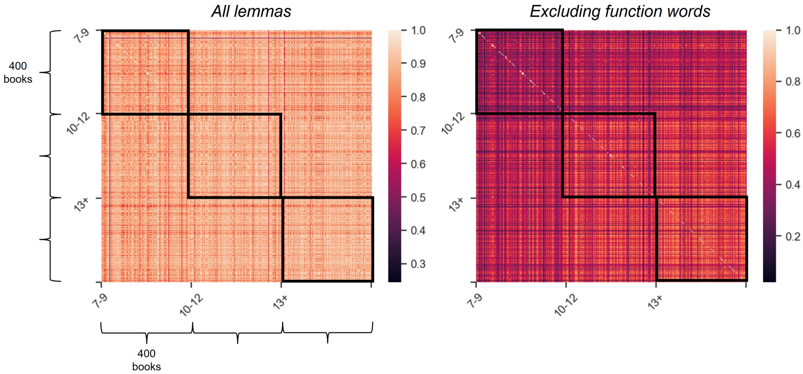


Semantic similarity across the books

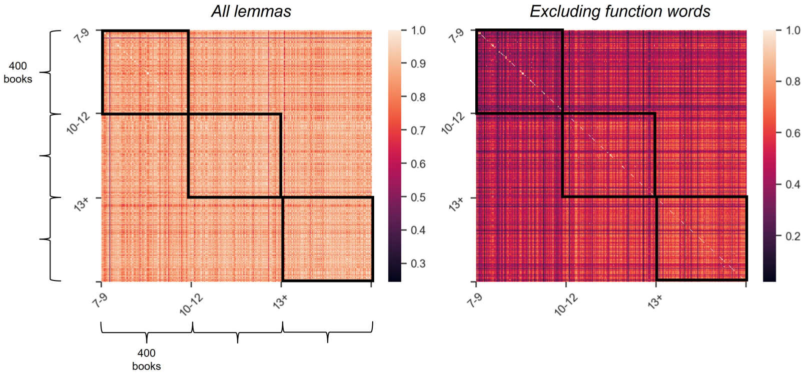
75 most common *lemmas* in sets of 25



Semantic similarity within the age bands



Semantic similarity within the age bands



↓

The ‘older’ the children, the more similar the books?..

Semantic similarity within the age bands

Type-token ratio

Semantic similarity within the age bands

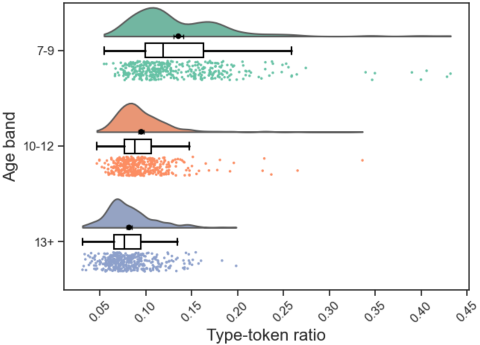
Type-token ratio

- Large corpora typically have higher TTR: the longer the sample, the higher the probability of encountering a new word

Semantic similarity within the age bands

Type-token ratio

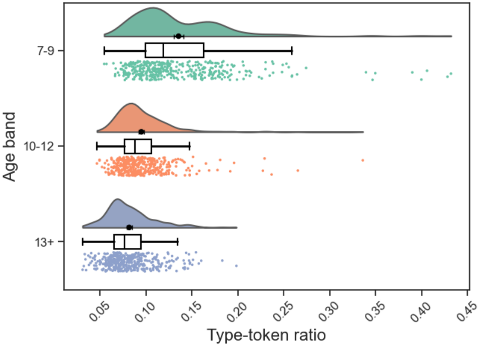
- Large corpora typically have higher TTR: the longer the sample, the higher the probability of encountering a new word



Semantic similarity within the age bands

Type-token ratio

- Large corpora typically have higher TTR: the longer the sample, the higher the probability of encountering a new word
- In CYP-LEX, this relationship is inverse!



Why reading widely is key

…and why it can be so hard

Why reading widely is key

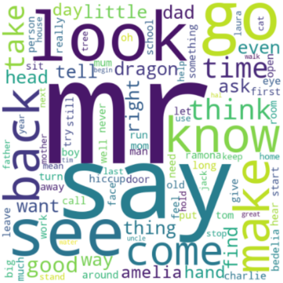
...and why it can be so hard

- 100 most frequent vs. infrequent lemmas across all the books in the 7–9 age band

Why reading widely is key

…and why it can be so hard

- 100 most frequent vs. infrequent lemmas across all the books in the 7–9 age band

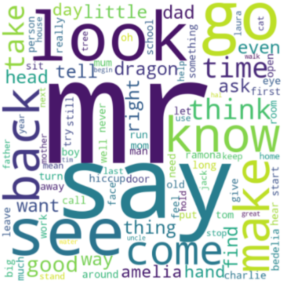


Why reading widely is key

...and why it can be so hard

- 100 most frequent vs. infrequent lemmas across all the books in the 7–9 age band

Most of these occur in all books

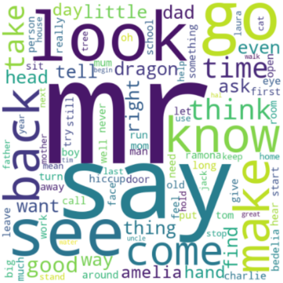


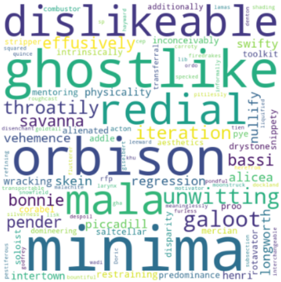
Why reading widely is key

...and why it can be so hard

- 100 most frequent vs. infrequent lemmas across all the books in the 7–9 age band

Most of these occur in all books



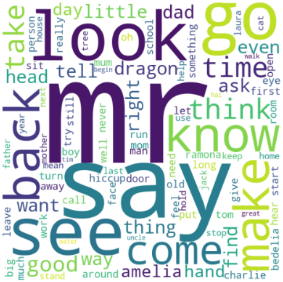


Why reading widely is key

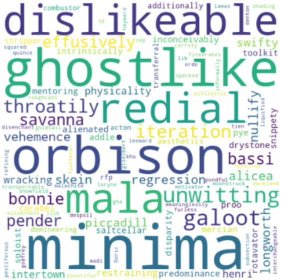
…and why it can be so hard

- 100 most frequent vs. infrequent lemmas across all the books in the 7–9 age band

Most of these occur in all books



These ones occur in 3 books only



Thank you!

References I

[1] D. Plaut and L. Gonnerman. "Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?" In: *Language and Cognitive Processes* 15 (2000), pp. 445–485. DOI: https://doi.org/10.1080/016909600501196G1.

[2] K. Rastle and M. H. Davis. "Morphological decomposition based on the analysis of orthography". In: *Language and Cognitive Processes* 23 (2008), pp. 942–971. DOI: https://doi.org/10.1080/01690960802069730.

[3] W. Nagy and R. Andreson. "How many words are there in printed school English?" In: *Reading Research Quarterly* 19 (1984), pp. 304–330. DOI: https://doi.org/10.2307/747823.

References II

[4] S. Andrews and S. Lo. “Is morphological priming stronger for transparent than opaque words? It depends on individual differences in spelling and vocabulary”. In: *Journal of Memory and Language* 68 (2013), pp. 279–296. DOI: https://doi.org/10.1016/j.jml.2012.12.001.

[5] A. Ulicheva et al. “Skilled readers’ sensitivity to meaningful regularities in English writing”. In: *Cognition* 195 (2020), p. 103810. DOI: https://doi.org/10.1016/j.cognition.2018.09.013.