

# Vocabulary and morphological complexity in books popular with British children

Maria Korochkina<sup>1</sup> | Marco Marelli<sup>2</sup> | Marc Brysbaert<sup>3</sup> | Kathy Rastle<sup>1</sup>

<sup>1</sup>Department of Psychology, Royal Holloway University of London, UK

<sup>2</sup>Department of Psychology, University of Milano-Bicocca, Italy

<sup>3</sup>Department of Experimental Psychology, Ghent University, Belgium



FRiLL meeting  
University of Reading  
11 December 2023



# The CYP-LEX project

- Large body of scientific knowledge on

- Large body of scientific knowledge on
  - ▶ how children learn to read & how they should be taught [1]

- Large body of scientific knowledge on
  - ▶ how children learn to read & how they should be taught [1]
  - ▶ the prerequisites for becoming an expert reader [2–6]

- Large body of scientific knowledge on
  - ▶ how children learn to read & how they should be taught [1]
  - ▶ the prerequisites for becoming an expert reader [2–6]
- The speed with which children gain reading expertise depends on the *nature of language* they are exposed to

- Large body of scientific knowledge on
  - ▶ how children learn to read & how they should be taught [1]
  - ▶ the prerequisites for becoming an expert reader [2–6]
- The speed with which children gain reading expertise depends on the *nature of language* they are exposed to
- Yet, 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

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+

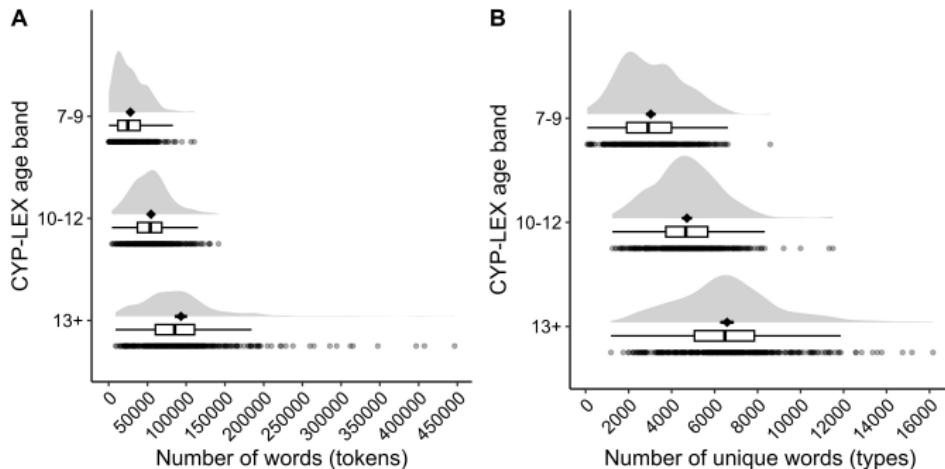


# The CYP-LEX corpus

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 $N$ ( $\sigma$ ) 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 $N$ ( $\sigma$ ) unique words per book	3,028 (1,452)	4,713 (1,550)	6,447 (2,366)

Children may encounter many unfamiliar words in books

# Children may encounter many unfamiliar words in books

*Percentage of CYP-LEX words that children DO NOT encounter on TV*

	Cbeebies 0–6 years	CBBC 6–12 years	SUBTLEX-UK adults
7–9 age band	40%	30%	
10–12 age band		60%	14%
13+ age band		48%	21%

# Children may encounter many unfamiliar words in books

*Percentage of CYP-LEX words that children DO NOT encounter on TV*

	Cbeebies 0–6 years	CBBC 6–12 years	SUBTLEX-UK adults
7–9 age band	40%	30%	
10–12 age band		60%	14%
13+ age band		48%	21%



From the earliest years of independent reading, children may be encountering a large proportion of words in books that are not in their spoken vocabulary

# Few words used in books are encountered frequently

# Few words used in books are encountered frequently

## A 7-9 age band

around asked  
looked your think  
going now over very  
been this all they from here how  
could for is he but me an am before  
go like had out at be by where  
us have in **ofato** she their  
see my her  
them his **i the** if not can here  
did that him right  
get are on up down  
when who do with you said so just too  
some more would what were no off  
eventime back into or  
know little head  
after

## B 10-12 age band

little off our  
again your now of looked  
about anonly  
know were out there us very  
from had for have like more  
this in **initiato** his is or but no get  
over me with **the** she then  
time ambe he was all by away  
go up not would we their  
are not going even  
too her what on and of said just  
think will at if you that did can  
could here into they so as him back  
way when them see down who before  
through right to other

## C 13+ age band

head still  
eyes over down which  
say get when like then know  
now out said have so or our  
way into with this **YOU** this **at**  
from here no as **ofato** i not what think  
who him were that **the** in at up more  
go they is it me by  
time we my and **he** she am some where  
through there on them  
just be **call** **was** her are too  
us did do but would back  
going your will see  
been could can see  
before their even right other  
than again

# Few words used in books are encountered frequently

A 7-9 age band

around asked  
looked your think  
going now from there how  
been this all they me an am before  
away could for is he but out be by where  
go like had in of ato she their  
us have in of ato see my her  
them his i the if not can him right here  
did that get are on and it was will as will right  
when who do with you said so we just too  
some more would what were no off  
eventime then back into or  
know little head after

B 10-12 age band

little now of our  
again your look about anonly  
know were out there us very  
from had for have like more  
this in it i a to she then  
over me with in it i a to she then  
ambe he was all by would  
time up go not was we their  
are too her what on and of said just going even  
think will at if you that did been  
could here into they so as him back  
way when them see down who before  
through right to other

C 13+ age band

head still  
eyes over down which  
say get when like then know  
now out said have so or our  
way into with this you this  
from here no as of a to i if not what think  
who him were that the in at up more  
go they is it me by  
time we my and he she am  
there just be o all was her are too  
us did do but would back  
going your will but could can see  
been before their even right other  
than again

- The top-100 words amount to half of each age band

# Few words used in books are encountered frequently

A 7-9 age band

around asked  
looked your think  
going now from there how  
been this all they me an am before  
away could for is he but we are  
go like had out at be by where  
us have in of ato she their  
see my her them his i the if not can  
them that was will him right  
did get are on and it was will  
through when do with you said so just too  
some who would what were no off  
more then back into or  
eventime about know little  
head after

B 10-12 age band

little of our  
again your now of looked  
about anonly  
know were out there us very  
from had for have like more  
this in it i a to she then  
over me with in it i a to she then  
time ambe he was all by but no get  
go up not was would we their  
are not what on and of was going even away  
too her what on and of was said just been  
think will at if you that did been  
could here into they so as him back  
way when them see some  
through right to other

C 13+ age band

head still  
eyes over down which  
say get when like then know  
now out said have so or our  
way into with this you this  
from here no as o f a t o i if not what think  
who him were that the in at up more  
go they is it me by  
time we my and he she am  
there on them some where  
through just be call was her are too  
us did do but would back  
going your will but could can see  
been before their even right other  
before than again

- The top-100 words amount to half of each age band
- Max 11% of words in each age band with  $fpmw > 10$

# Few words used in books are encountered frequently

A 7-9 age band

around asked  
looked your think  
going now over very  
from there how  
been this all they  
me an am before  
could for she but  
she out be by where  
go like had  
us have in of ato  
see my her  
them his i the if not can  
did that was will him right  
get are on and it was  
when up down  
some who do with you said so just too  
more would then what were no off  
eventime back into or  
know little head  
after

B 10-12 age band

little of our  
again your now of looked  
about anonly  
were out there us very  
from had for have like more  
his is or but no get  
how this in with in ita to  
over me do she then  
ambe he was all by would  
time up go not we their  
are not what on and of was  
too her what on and of was  
think will at if you that did been  
could here into they as him back  
way when them see some  
through right to other

C 13+ age band

head still  
eyes over down which  
say get when like then know  
now out said have so or our  
way into with this you this  
from here no as of a to i if not what think  
who him were that the in at up more  
go they is it me by  
time we my and he she am  
there on them some where  
through just be call was her are too  
us did do but would back  
going your will but could can see  
been before their even right other  
than again

- The top-100 words amount to half of each age band
- Max 11% of words in each age band with  $fpmw > 10$



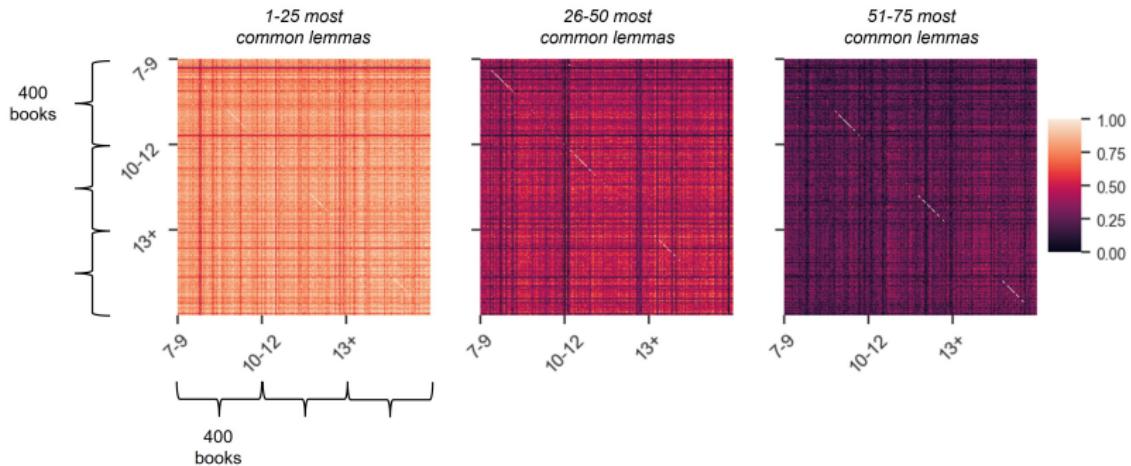
Only a small subset of words may be learned to the degree that they are recognised automatically and effortlessly

# Most words are infrequent and used in few books

75 most common *lemmas* in sets of 25

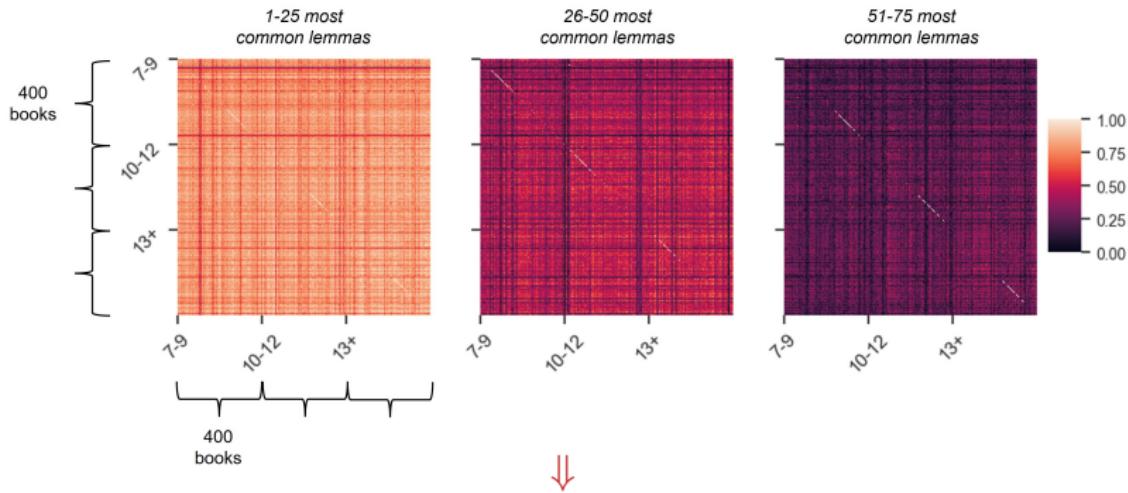
# Most words are infrequent and used in few books

75 most common *lemmas* in sets of 25



# Most words are infrequent and used in few books

75 most common *lemmas* in sets of 25



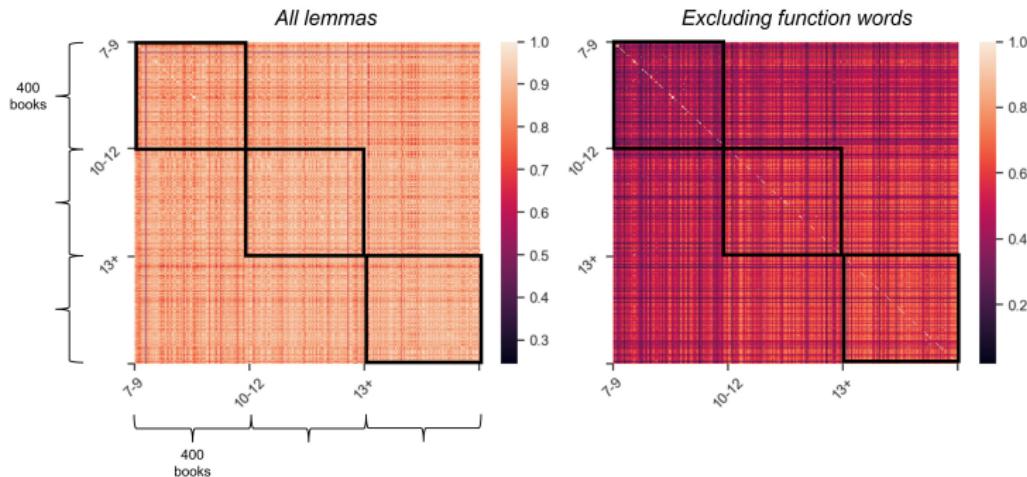
Books similar regarding most common lemmas but rapidly diverge

# Most words are infrequent and used in few books

All lemmas & all lemmas excluding function words

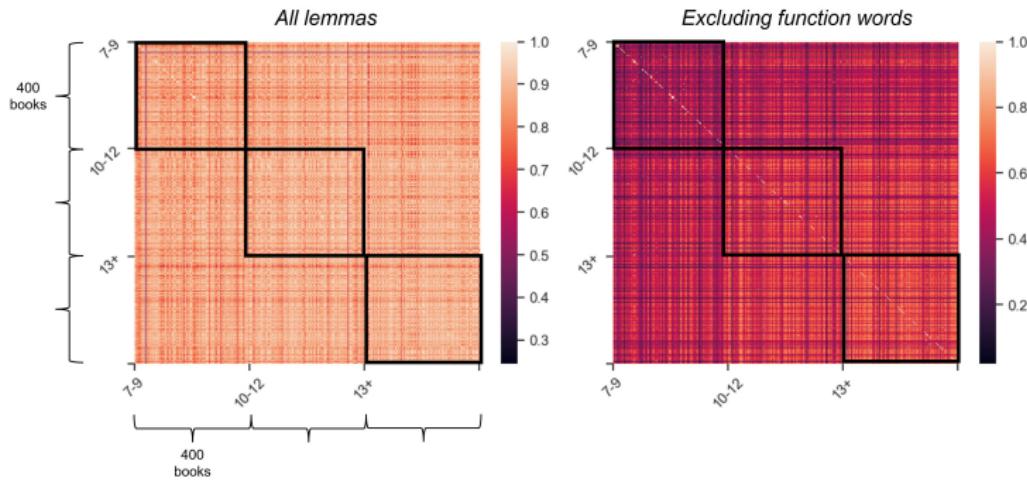
# Most words are infrequent and used in few books

All lemmas & all lemmas excluding function words



# Most words are infrequent and used in few books

All lemmas & all lemmas excluding function words



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

# Many new words in each age band

About 25,000-31,000 words

# Many new words in each age band

About 25,000-31,000 words

10-12 vs. 7-9:

73% encountered  $\leq$  3 times

# Many new words in each age band

About 25,000-31,000 words

10-12 vs. 7-9:  
73% encountered  $\leq$  3 times

A dense cluster of colorful words from the CYP-LEX corpus, primarily in green, blue, and purple, illustrating the vocabulary used by children aged 10-12. The words include: osteichthyes, christmastides, traceries, frazzles, morello, chairmanship, strength, unconsuming, undervests, deplete, scath, georgians, reclusiveness, darfur, polygons, reexamine, pinpricking, unirrigated, unmysterious, calvaire, earworms, islandless, schmuck, stalklike, ibsen, liaison, mirth, preternaturally, langley, crotched, unreled, resection, ismael, ramified, unoriginal, summersets, handley, ebbtide, warne, tabla, detent, wroth, catheters, chaises, Dustin, knobstick, fairmont, ingres, beguilingly, kleiner, libeled, rete, mortem, kirkyard, lightship, regicide, biomass, smokable, dirts, torbay, krug, avocations, englishness, benefaction, emboldening, quia, englischness, garble, zebrawood, mesopotamians, donas runch, inrushng, laughers, signa, choriambs, uses, turnstone, sapwood, transcaucasian, detouring, retrreaded, reappraised, mudholes, ashmolean, salvagers, tweeddale, coexisting, toweling, merde, mistrustfulness, hanratty, dereliction.

# Many new words in each age band

About 25,000-31,000 words

10-12 vs. 7-9:

73% encountered  $\leq$  3 times

13+ vs. 10-12:

74% encountered  $\leq$  3 times

A dense cluster of colorful words from the CYP-LEX lexicon, primarily in green, blue, and purple, illustrating the vocabulary used by different age groups. The words include: osteichthyes, christmastides, traceries, frazzles, morello, chairmanship, strength, unconsuming, undervests, deplete, scath, georgians, reclusiveness, darfur, polygons, reexamine, pinpricking, unirrigated, unmysterious, calvaire, earworms, islandless, schmuck, stalklike, ibsen, liaison, mirth, preternaturally, langleys, crotched, unreled, resection, ismael, ramified, unoriginal, summersets, handley, ebbtide, warne, knobstick, enrolls, catheters, chaises, warne, unoriginal, tabla, detent, wroth, knibled, rete, fairmont, ingres, beguilingly, kleiner, Dustin, biomass, smokable, mortem, kirkyard, lightship, regicide, benefaction, dirts, torbay, krug, avocations, englishness, libeled, emboldening, quia, englischness, garble, zebrawood, mesopotamians, donas runch, inrushings, laughers, signa, choriambs, uses, turnstone, rethreaded, reappraised, sapwood, flowerets, derouing, detouring,石榴石, transcaucasian, goneness, toweling, mistrustfulness, mudholes, ashmolean, salvagers, coexisting, merde, disgorging, hanratty, dereliction, tweeddale.

# Many new words in each age band

About 25,000-31,000 words

10-12 vs. 7-9:

73% encountered  $\leq$  3 times

osteichthyes christmastides  
unconsumming morello traceries frazzles  
scath georgians reclusiveness chairmanship strength  
polygons reexamine pinpricking unreveled darfur  
earworms islandless cheapened unirrigated calvaire  
ibsen laisee schmuck preternaturally dunnocks stalklike  
ebbtide ramified merth langleys  
chaises warne unoriginal crotched resection ismael  
beguilingly kleiner knobstick unrelaxed handley  
lightship libeled tabla detent wroth catheters  
avocations regicide biomass smokable handlings  
quia englishness libealed rete mortem kirkyard  
detouring inrushings garble benefaction kirkby  
gomeness toweling signa choriambs mortem  
coexisting merde mistrustfulness mesopotamians uses  
toweling turnstone reappraised donas runch  
merde dereliction rethreaded sapwood transcaucasian  
disgorger hanratty mudholes inkpots  
dereliction turnstone rethreaded hanratty tweeddale

13+ vs. 10-12:

74% encountered  $\leq$  3 times

outshooting hemophilic cubiculo privatisation pontificating  
remolded lezzies infanticides wisewoman valborg  
squabs Dempster cankerworm bratlings  
rhapsodical teethlike outhomers eliding crappier  
glasses empathizing wornout appre ciatory gathings  
supinely sultrier depositaries myocarditis gutturals  
olid porringer habitua degradations selenic un honed vising  
mcphail sphygmic innit panniered fees  
caimans spizy Ephesus skimping swahili tarlton  
musts parzival widener odfa pollack liaison  
shirred jostle overfond libellers reactant semblable  
earlyish unprizable lomouc magnetical hassett  
denay tipplers studier simborsk dominica floodwall  
cantons expropriating ramification vassalage  
sickened Doddypoll cbd fibreboard lexicographers  
stomatitis mendota galí avoiders homozygous  
prototyping housemother graticules misstatement  
sexagenarians divisa manteca  
overarches landhold lawtons retranfer

# The importance of morphology

# The importance of morphology

- Morphological knowledge is an important heuristic for vocabulary growth and is thus crucial for the development of reading expertise

# The importance of morphology

- Morphological knowledge is an important heuristic for vocabulary growth and is thus crucial for the development of reading expertise
- Children's morphological knowledge is shaped through their experience with written text

# The importance of morphology

- Morphological knowledge is an important heuristic for vocabulary growth and is thus crucial for the development of reading expertise
- Children's morphological knowledge is shaped through their experience with written text



Understanding the nature of text experience is critical for understanding what children can learn about individual morphemes and how

# Etymological approach: MorphoLex

# Etymological approach: MorphoLex

MorphoLex size: 68,624 words  
ELP complete list: 79,672 words

57,133 words (54%)  
in MorhoLex  
48,560 words are not

CYP-LEX size: 70,287,217 tokens & 105,693 types

	7–9	10–12	13+
$N$ words	11,162,653	21,837,794	37,286,770
Average $N (\sigma)$ 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 $N (\sigma)$ unique words per book	3,028 (1,452)	4,713 (1,550)	6,447 (2,366)

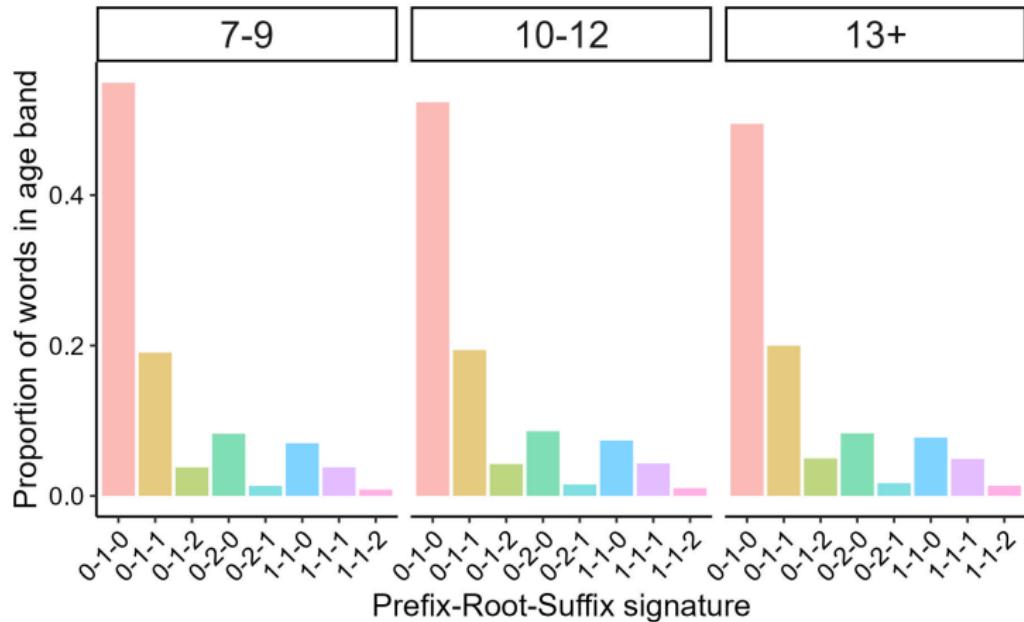
39,149 words (74%)  
in MorhoLex  
13,702 words are not

47,363 words (67%)  
in MorhoLex  
23,582 words are not

54,557 words (60%)  
in MorhoLex  
36,423 words are not

At least half of words are multimorphemic

# At least half of words are multimorphemic



# Properties of multimorphemic words

# Properties of multimorphemic words

Multimorphemic words...

# Properties of multimorphemic words

Multimorphemic words...

- are less common than monomorphemic words, but their frequency increases with book target age

# Properties of multimorphemic words

Multimorphemic words...

- are less common than monomorphemic words, but their frequency increases with book target age
- appear in fewer books than monomorphemic words, but their CD increases with book target age

# Properties of multimorphemic words

Multimorphemic words...

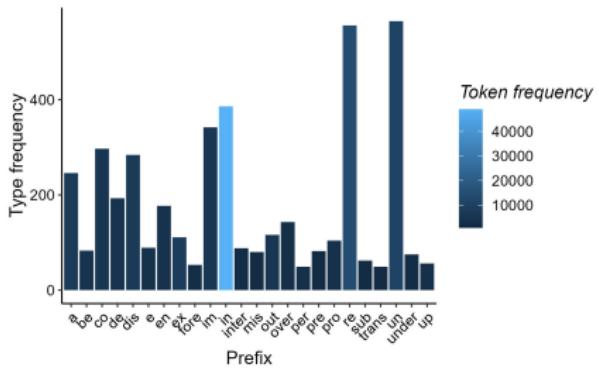
- are less common than monomorphemic words, but their frequency increases with book target age
- appear in fewer books than monomorphemic words, but their CD increases with book target age
- constitute the majority of words missing in SUBTLEX-UK or younger age bands, with the number and frequency of these words increasing with book target age

# Prefix frequency and contextual diversity

7–9 age band

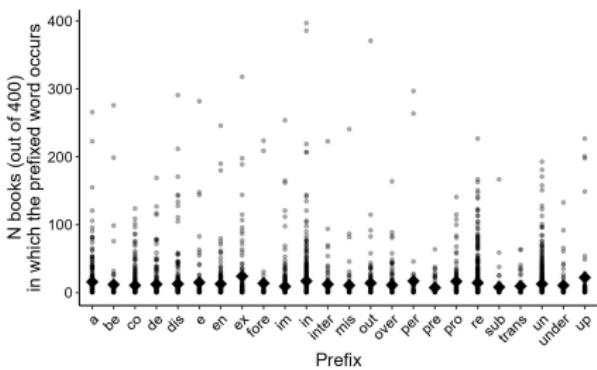
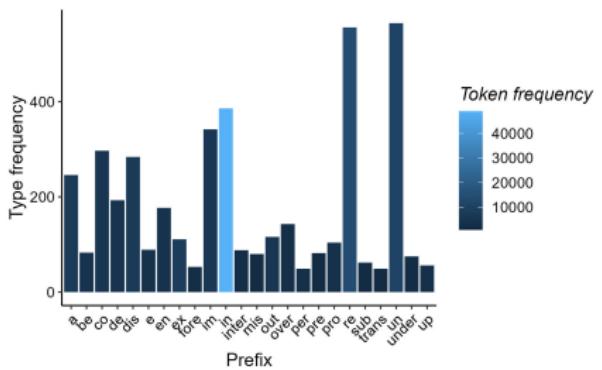
# Prefix frequency and contextual diversity

7–9 age band



# Prefix frequency and contextual diversity

7–9 age band

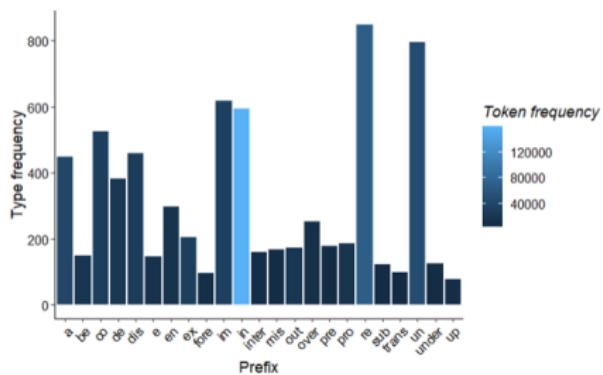


# Prefix frequency and CD increase with book target age

13+ age band

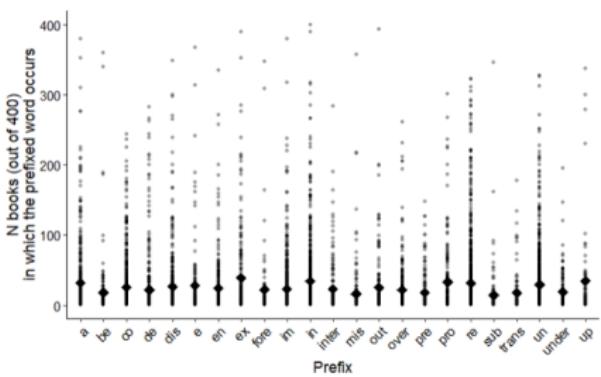
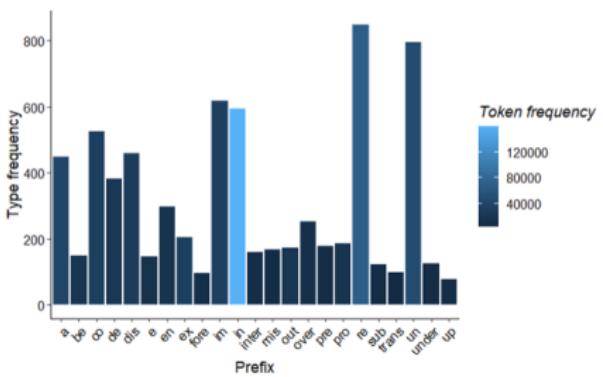
# Prefix frequency and CD increase with book target age

13+ age band



## Prefix frequency and CD increase with book target age

13+ age band

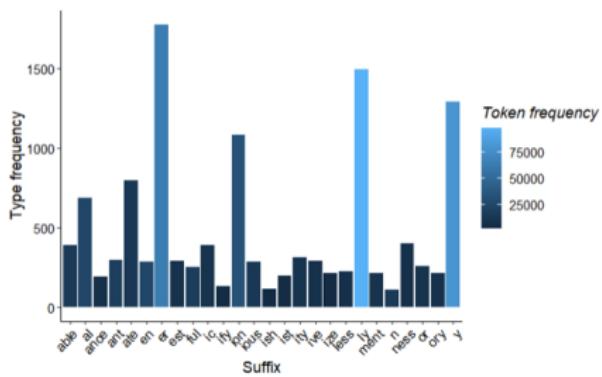


# Suffix frequency and contextual diversity

7–9 age band

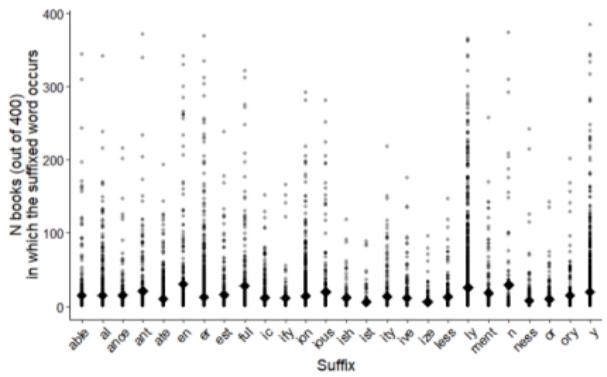
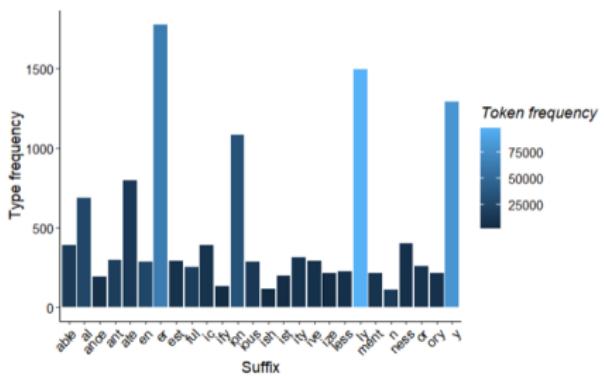
## Suffix frequency and contextual diversity

## 7–9 age band



## Suffix frequency and contextual diversity

## 7–9 age band

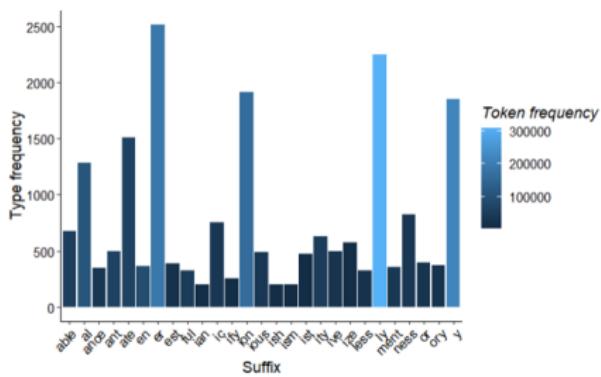


# Suffix frequency and CD increase with book target age

13+ age band

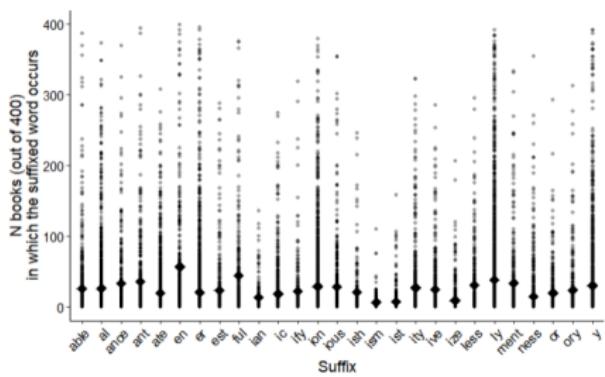
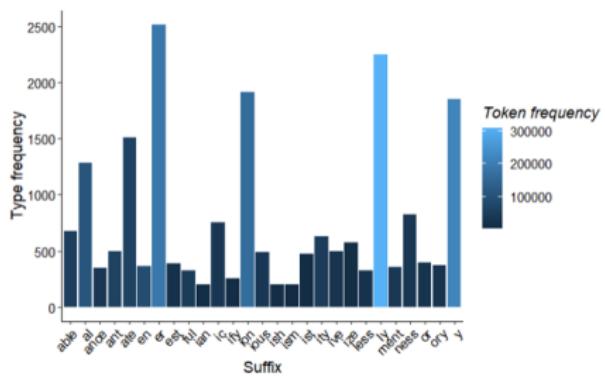
# Suffix frequency and CD increase with book target age

13+ age band



## Suffix frequency and CD increase with book target age

## 13+ age band



## Limitations of current approach & future directions

## Limitations of current approach & future directions

- Information available only for half of the CYP-LEX types

## Limitations of current approach & future directions

- Information available only for half of the CYP-LEX types
- Children have limited linguistic knowledge & operate on the orthographic principle ('sustain' vs. 'corner')

## Limitations of current approach & future directions

- Information available only for half of the CYP-LEX types
- Children have limited linguistic knowledge & operate on the orthographic principle ('sustain' vs. 'corner')
- Affixes differ in terms of semantic consistency ('-less' vs. '-ist')

## Limitations of current approach & future directions

- Information available only for half of the CYP-LEX types
- Children have limited linguistic knowledge & operate on the orthographic principle ('sustain' vs. 'corner')
- Affixes differ in terms of semantic consistency ('-less' vs. '-ist')



Use distributional semantic modelling to build affix meaning representations

## Limitations of current approach & future directions

- Information available only for half of the CYP-LEX types
- Children have limited linguistic knowledge & operate on the orthographic principle ('sustain' vs. 'corner')
- Affixes differ in terms of semantic consistency ('-less' vs. '-ist')



Use distributional semantic modelling to build affix meaning representations



Develop theoretically-driven metrics that capture the complexity of derivational regularities encountered in children's books

Korochkina, M., Marelli, M., Brysbaert, M., & Rastle, K. (2023). The Children and Young People's Books Lexicon (CYP-LEX): A large-scale lexical database of books read by children and young people in the United Kingdom. *Pre-print*.

<https://doi.org/10.31234/osf.io/nha8t>

*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>.