



Kathy Rastle



Marc Brysbaert



Marco Marelli

Maria Korochkina

What children learn from books, and how to talk to teachers about it



**Economic
and Social
Research Council**

maria.korochkina@rhul.ac.uk
<https://mariakna.github.io/>



research ED

The goal of researchED is to bridge the gap between research and practice in education. Researchers, teachers, and policy makers come together for a day of information-sharing and myth-busting.

We aim to bring together as many parties affected by educational research – e.g. teachers, academics, researchers, policy makers, teacher-trainers – in order to establish healthy relationships where field-specific expertise is pooled usefully.

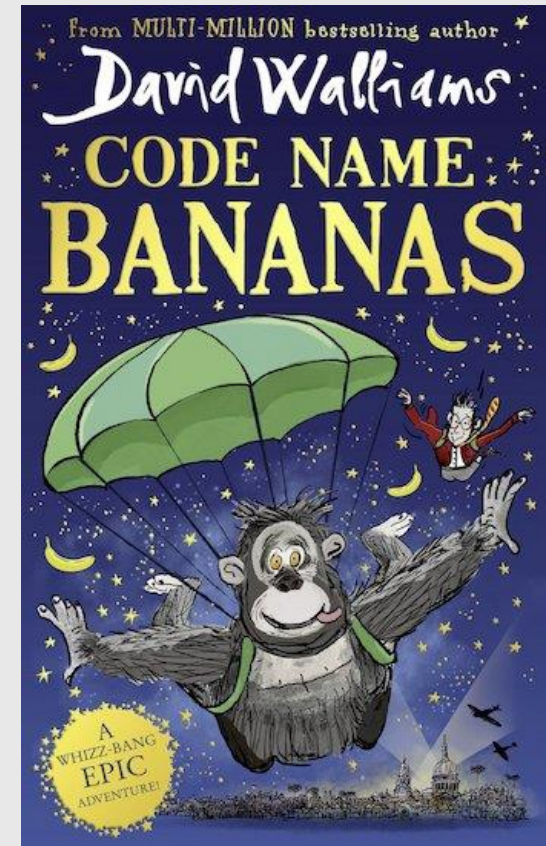
<https://researched.org.uk/>

The complexity of reading



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

“Then a mischievous thought flashed across her eyes, and she pursed her lips together and pushed her tongue forward”.

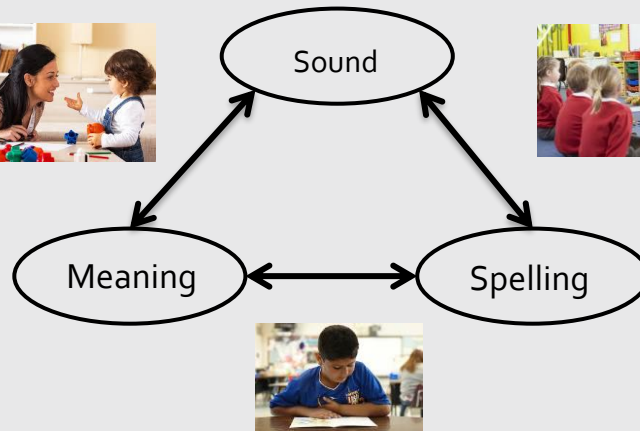


The journey to skilled reading



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

Oral language
foundations



Phonics
instruction

Text experience

What challenges does text experience pose and what opportunities does it bring?

Keep an eye out for the lightbulbs!



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON



CYP-LEX

The Children and Young People's Books Lexicon



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

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



1,200 popular books, 400 books per age band

7-9 years

10-12 years

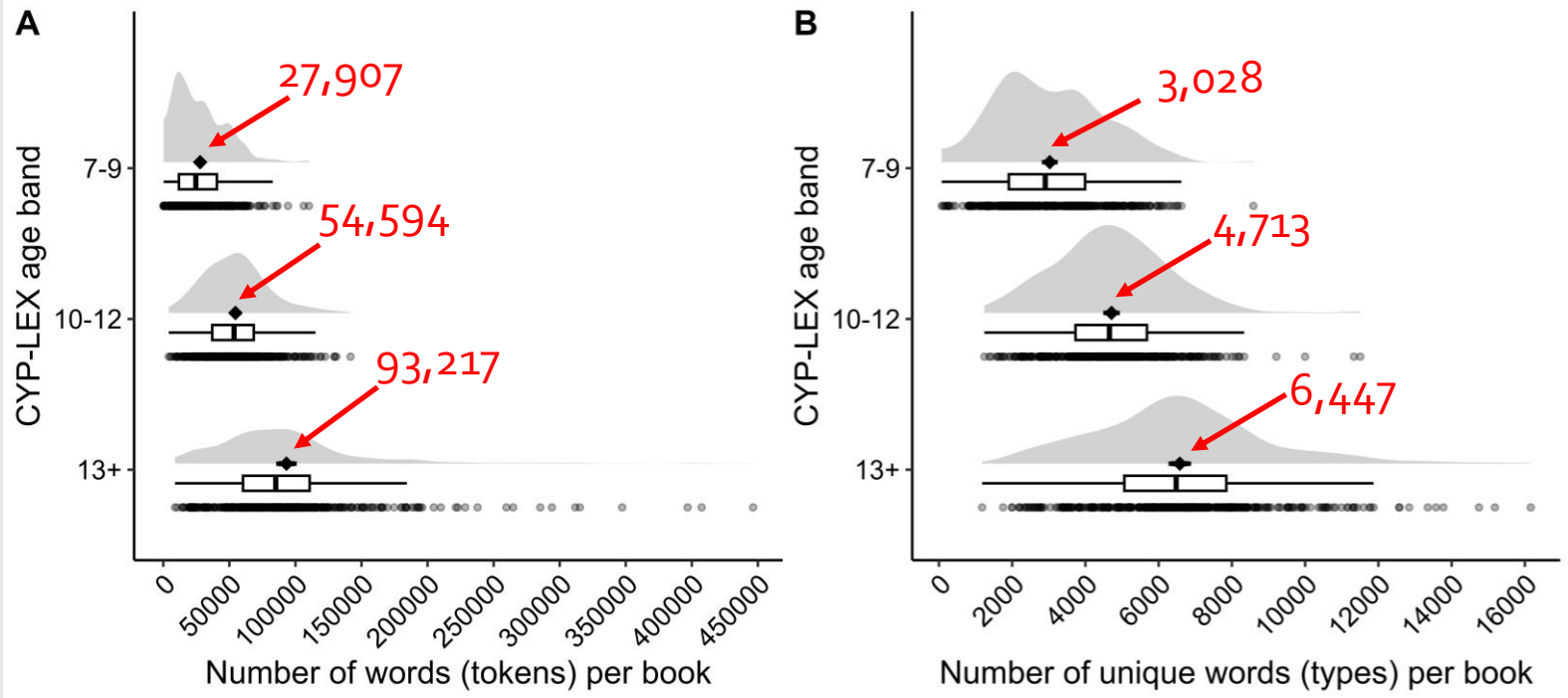
13+ years



Many distinct words in each age band



- Over **70 million words** and **over 100K** distinct words in 1,200 books
- **50K+** distinct words in the 7-9 age band alone



- Vast numbers of distinct words in books
- **Memorising words by rote is not an option**

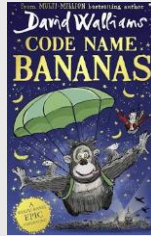
Yet, these words aren't very useful for understanding



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

“Then a mischievous
thought flashed
across her eyes, and
she pursed her lips
together and
pushed her tongue
forward”.

“Then a
her , and
she her
and
her
”.



- Children will quickly learn to recognise these words by sight
- Recognising every second word effortlessly **will not be enough to understand** the text

Many words may be unfamiliar



Percentage of CYP-LEX words **not** encountered on TV

	CBeebies (<i>up to 6 yrs</i>) + CBBC (<i>6-12 yrs</i>) 63,081 words	9 BBC channels 159,235 words
7-9 age band	28%	
10-12 age band	40%	
13+ age band		21%

- Children encounter many words in books that are **not in their spoken vocabulary**
- This occurs from the **earliest** years of independent reading



Children **will be needing support** to understand the words they encounter in books

Most words are not used repeatedly



	% words used more than 100 times	% words used less than 50 times
7-9 age band	12%	81%
10-12 age band	14%	79%
13+ age band	16%	77%

- Increasing % of frequently used words as books become more advanced
- **Not enough exposure** to learn to recognise most words by sight



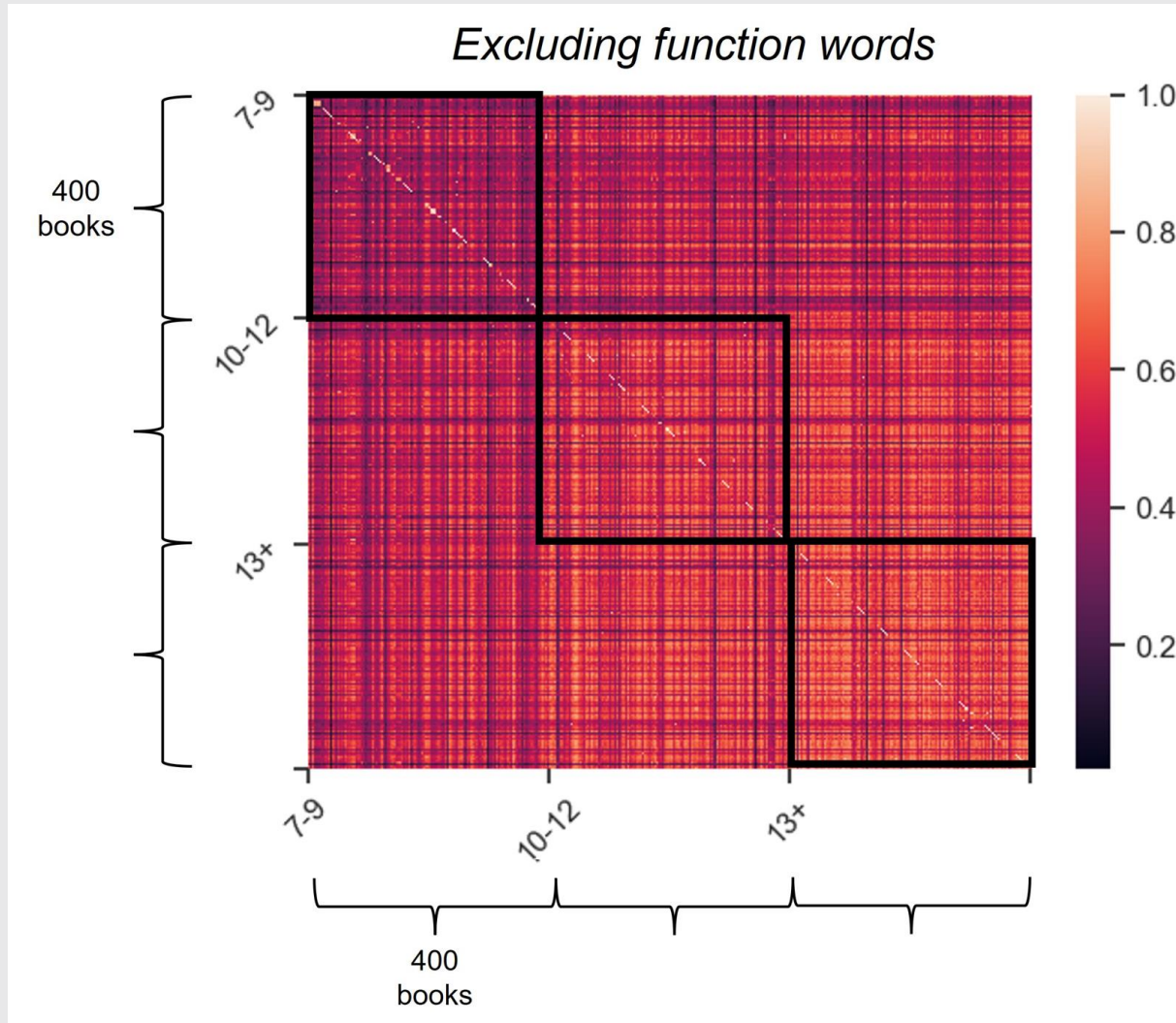
It is crucial that children acquire
strong decoding skills early on

Books vary greatly in the words they use



- In each age band, 30% of distinct words appear in **1 book out of 400**
- Most of these words are used **once** in that book

Books vary greatly in the words they use



Books vary greatly in the words they use



- In each age band, 30% of distinct words appear in **1 book out of 400**
- Most of these words are used **once** in that book
- **Low similarity** in vocabulary **across the individual books**
- In the 7-9 age band, books are less similar to one another than in the other age bands



Each book contains many words that are not encountered in any other book

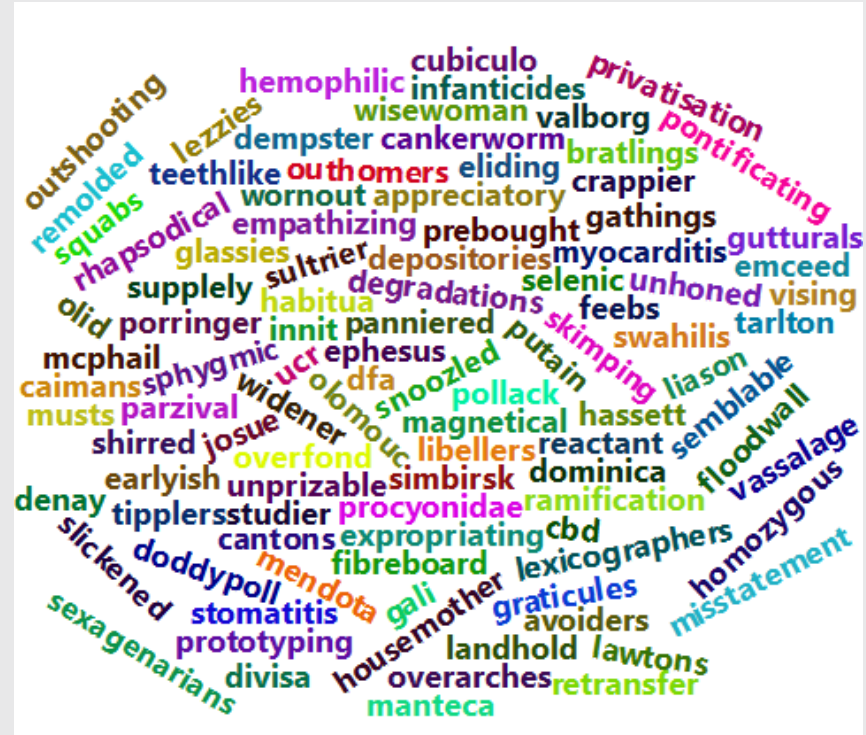


- **Each book is a challenge**
- To encounter as many different words as possible as often as possible, children **must read widely**

Over 70% of “new” words are used infrequently...



... and many contain several morphemes!



Children will need **strong morphological knowledge** to understand new words in books

The power of morphology



- Most English words are built by **recombining stems and affixes**

cleaner, cleanly, unclean
teacher, banker, builder

- Morpheme knowledge is also crucial for computing the meanings of **unfamiliar** words

bright + -ify → brightify

- Limited time for explicit teaching of morphology, so morpheme knowledge often acquired through **text experience**

Few complex words are used repeatedly



- Roughly **half of all distinct words** in each age band are complex
- But **few complex words are used repeatedly** or in many books

	7-9	10-12	13+
Occur 5 times or less	50%	42%	35%
Occur 100 times or more	8%	11%	15%

- Children are **likely to see** a complex word, but **unlikely to ever see it again!**



- **Difficult** to learn to recognise complex words by sight
- It is critical to be able to **break words apart**

Pre-requisites for affix learning



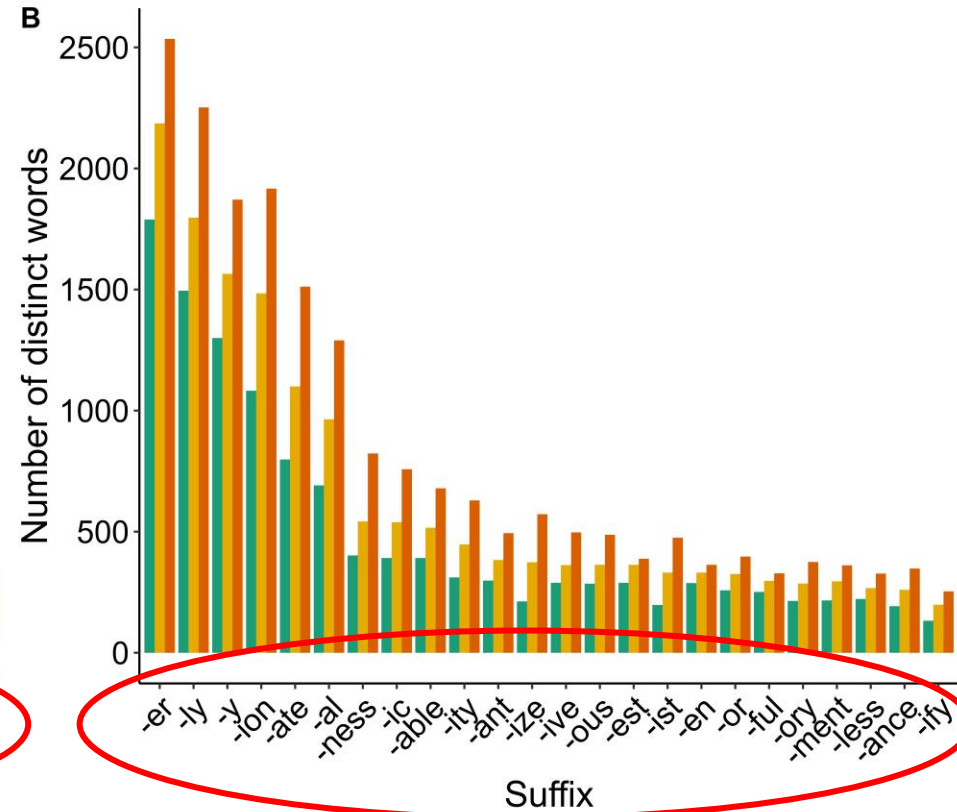
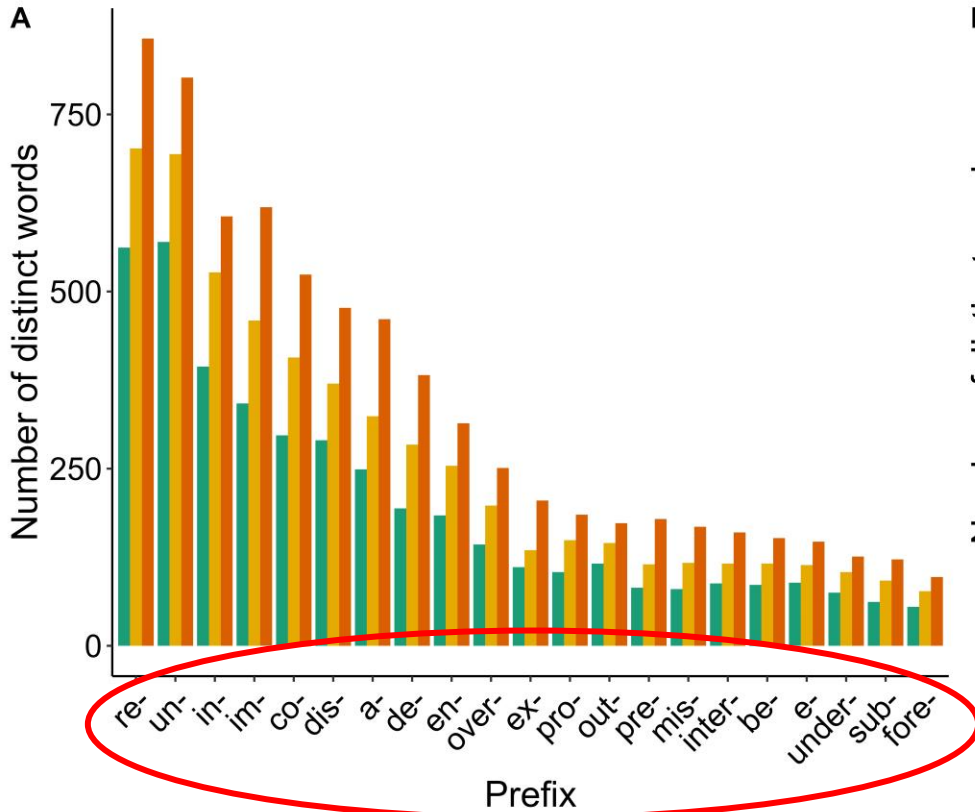
<u>u</u> nknown	<u>s</u> ubconscious
<u>u</u> nfair	<u>s</u> ubheading
<u>u</u> nafrail	<u>s</u> uboptimal
<u>u</u> nlikely	<u>s</u> ubjugate
<u>u</u> nconvinced	<u>s</u> ubmit
<u>u</u> nsecure	<u>s</u> ubject
<u>u</u> nwell	<u>s</u> ubside (sub + -sidere)

- Must have **consistent** meaning transformation
- Must occur with a high number of **distinct stems**
- Must be **detectable**

Few affixes are used with many different stems



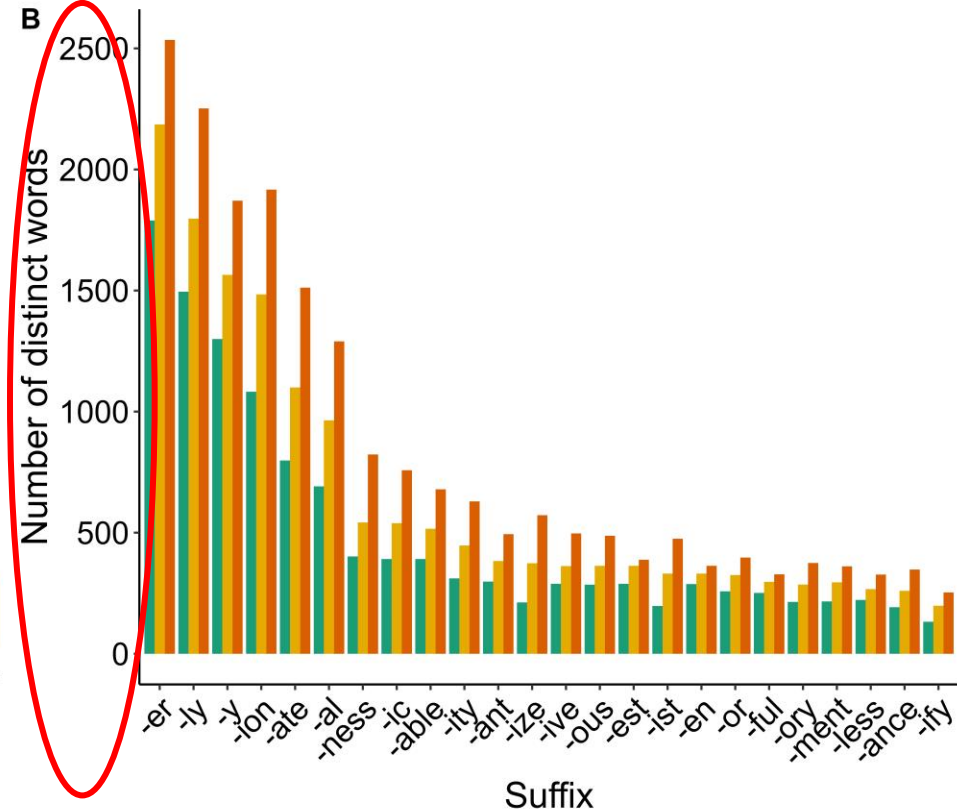
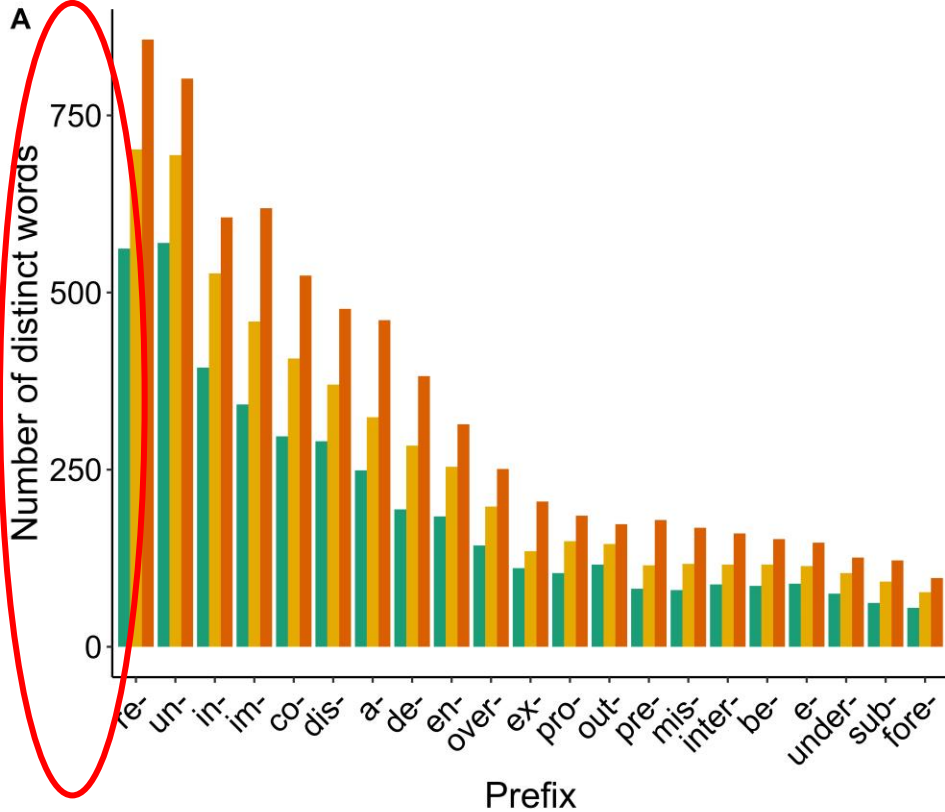
Age band 7-9 10-12 13+



Few affixes are used with many different stems



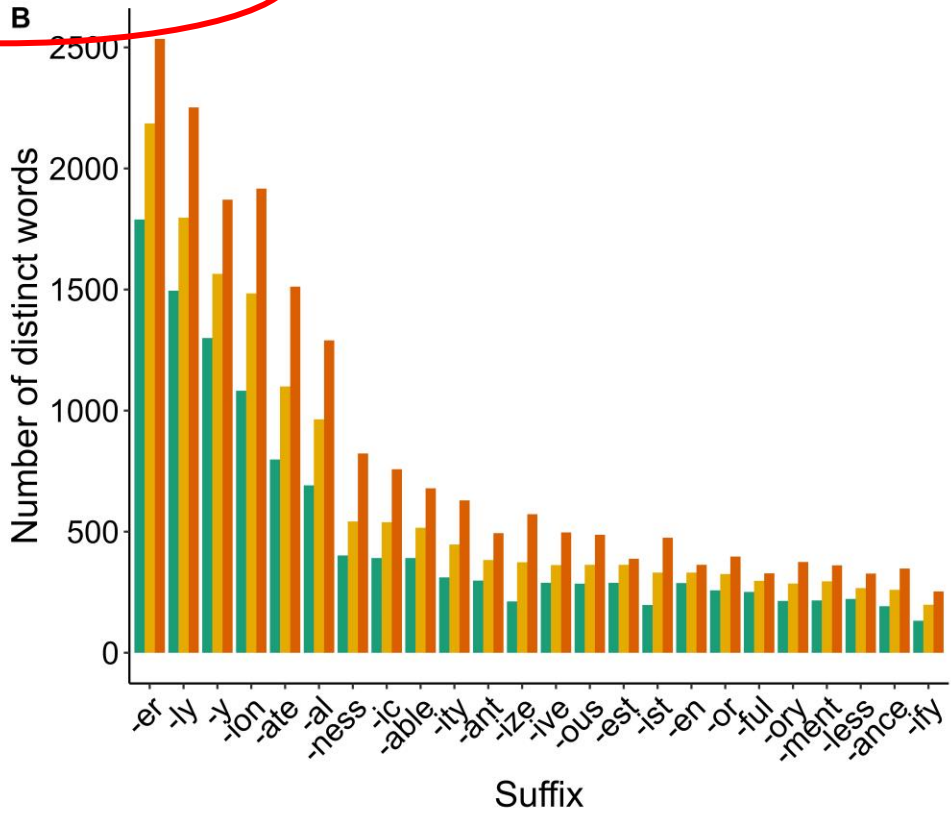
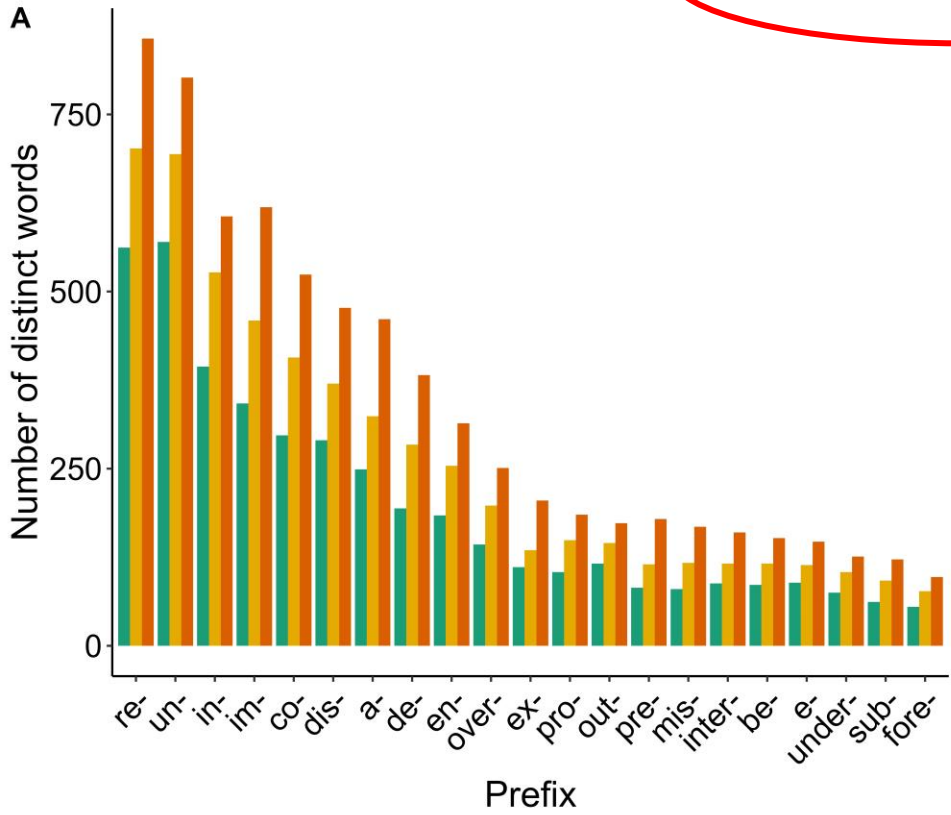
Age band 7-9 10-12 13+



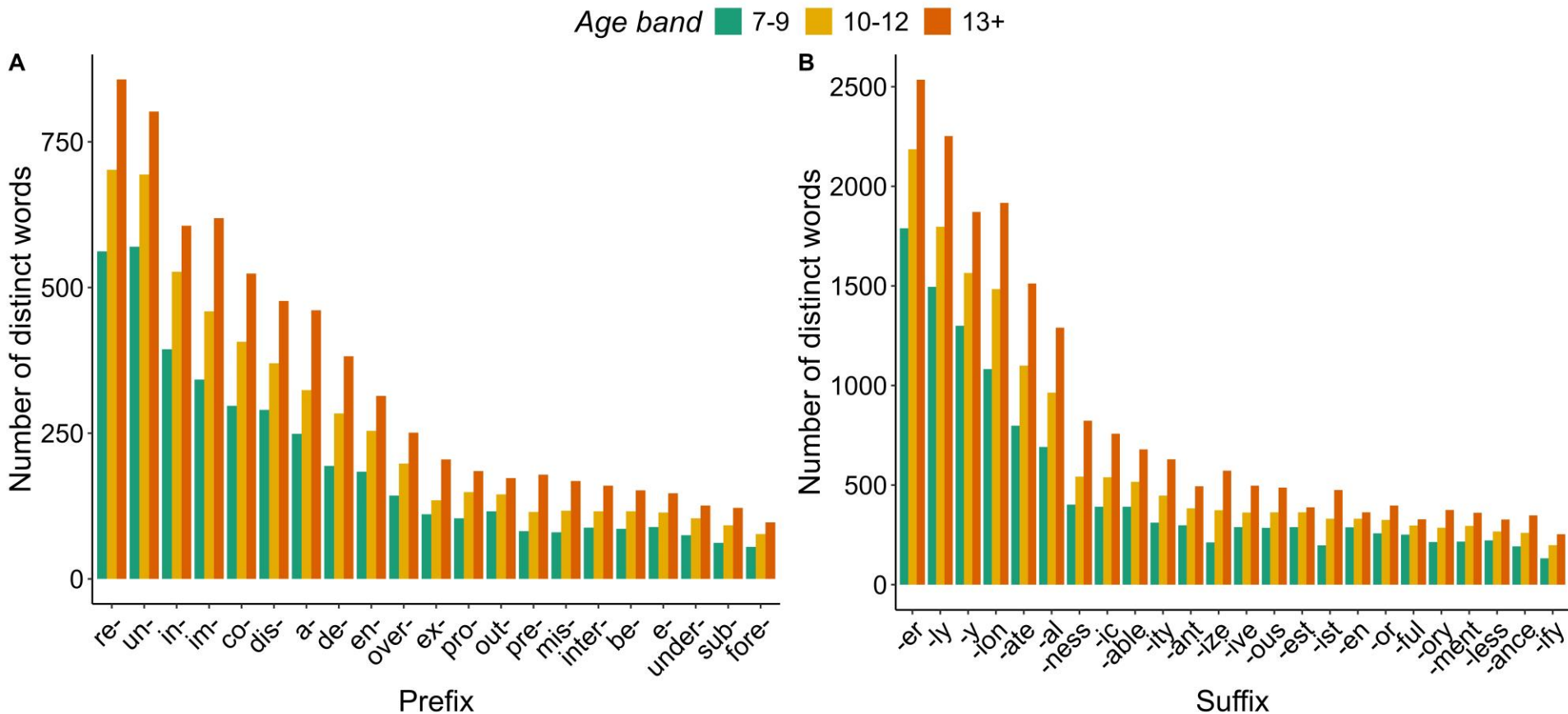
Few affixes are used with many different stems



Age band 7-9 10-12 13+



Few affixes are used with many different stems



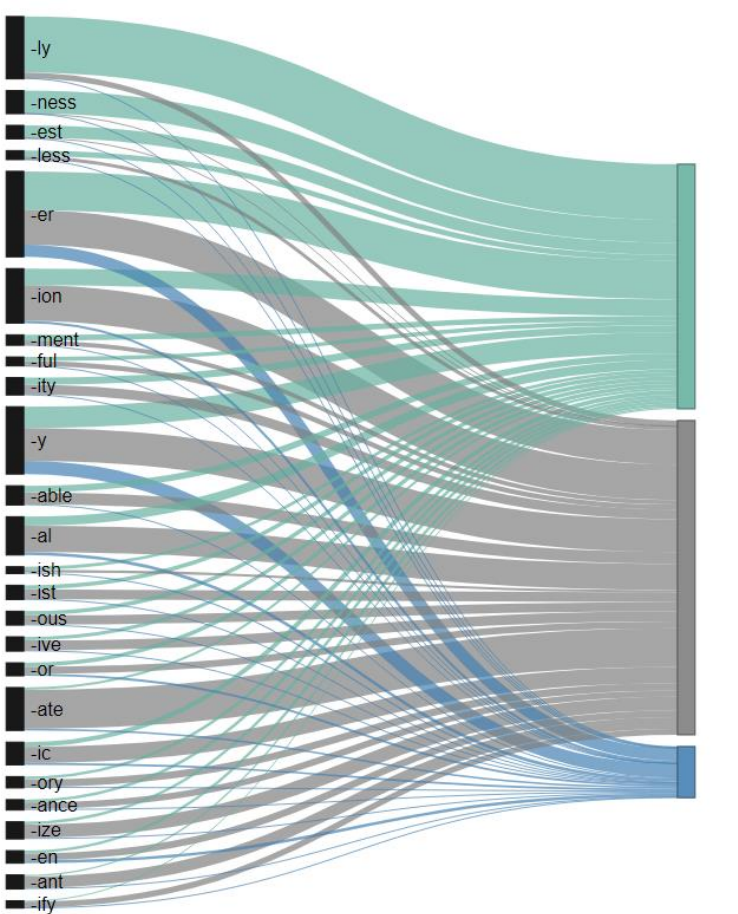
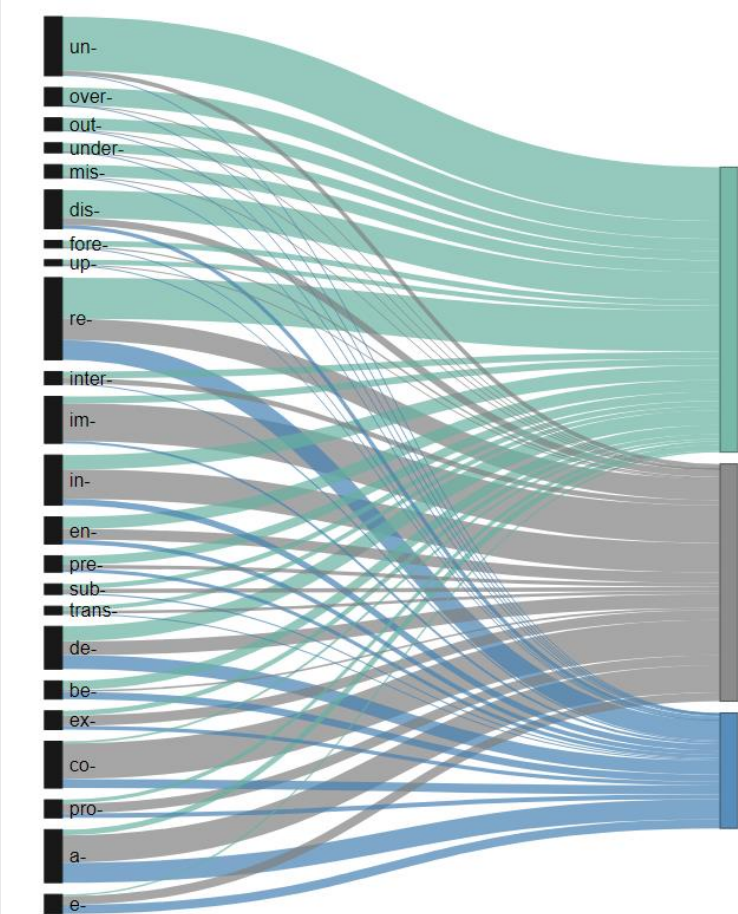
- **Limited exposure** before 13+ texts
- Only a few affixes are frequent: *un-*, *re-*, *in-*, *-er*, *-ly*, *-y*, *-ate*

Morpheme detectability analysis



Prefixes

Suffixes



- Prefixed words detectable with RegEx
- Prefixed words not detectable with RegEx
- Words incorrectly parsed as prefixed

- Suffixed words detectable with RegEx
- Suffixed words not detectable with RegEx
- Words incorrectly parsed as suffixed

Few affixes are easy to detect



Easy to detect

un- (unknown, unwise, undo)

-ly (warmly, openly, friendly)

Mostly undetectable

in- (inject, include, involve)

-ate (facilitate, allocate, irrigate)

Often undetectable or difficult to parse

-y (gravity, trinity, comply, rely, subsidy)

Pseudo-affixation

-er (corner, brother, number)



- Many complex words **will not add to a reader's experience** of the affixes
- The **opportunity** for affix learning via text is **limited**

A case for morphology instruction?..



- Complex words comprise a large proportion of words in children's books
- Beyond a handful of affixes, morpheme knowledge will be **difficult to acquire** from text



- Is there value in **more systematic** morphology instruction?
- There is **potential for substantial impact** on vocabulary acquisition and reading comprehension
- Yet, there may be significant **challenges in implementation**

A closer look at the 13+ books...



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

32 prose books from the AQA and EdExcel specifications for English Literature GCSE



How do these books compare to the popular books?

Highly dense vocabulary in the GCSE books



- GCSE books are **half as long as** popular books, but contain a **similar number of distinct words**
- GCSE books are much **less homogenous** in the words they use



- **More** vocabulary through **less** text in the GCSE books
- May be **harder to understand the text as a whole**

Many unfamiliar words in the GCSE books



- Only **33%** of the distinct words **occur regularly** in popular books
- The remaining **67%** are **used sparsely**
- **3,000** distinct words **never used on 9 BBC TV channels over 3 years**
 - *poulterer, bonneted, dowerless, bedight, sepulchre, catechize*
 - *brusquely, docilely, imploringly, beatifically, superciliously*



- GCSE texts **will stretch** even those who read widely
- Weaker readers **may not be able to engage** with the GCSE texts at all



Many unfamiliar words in GCSE books are new roots



- Children **often** encounter unfamiliar words in books
- In popular books, many of these “new” words are **morphologically complex**
→ Meanings **can be derived** from the words’ constituents: [mourn] + [-ful] + [-ly]
- In GCSE books, most “new” words are **new roots**
→ Meanings **cannot be derived** from smaller units
 - *aspidistra, crimplene, beseech, coccidia, gambol*



- Pupils **must rely on context or instruction** to understand these words
- Deriving meaning from context requires **advanced language and reading skills**

Conclusions



- 💡 Books offer a wonderful **opportunity** to build vocabulary
- 💡 Yet, book vocabulary is **challenging** from the get-go
- 💡 Children need **strong foundational reading skills** to access popular books
- 💡 Children need to **read widely** to build reading proficiency
- 💡 There is a **partnership** between reading skills and reading motivation
- 💡 Children with good foundational reading skills will be able to read, understand what they are reading, and derive pleasure from books, **leading to a virtuous cycle**

Dissemination & impact



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

Quarterly Journal of Experimental Psychology
OnlineFirst

Open Access

Sage Journals

© Experimental Psychology Society 2024, Article Reuse Guidelines
<https://doi.org/10.1177/17470218241229694>

Original Article



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

Maria Korochkina ¹, Marco Marelli², Marc Brysbaert ³, and Kathleen Rastle ¹

rastlelab.com/blog

Maria Korochkina & Kathy Rastle · Mar 12 · 5 min read

What Words do Children Encounter When They Read for Pleasure?

The ability to read opens up worlds. Reading enables children to progress into post-primary education and provides the basis for lifelong learning and prosperity into adulthood. Importantly, the [journey](#) to becoming a skilled reader requires not only high-quality classroom instruction but also many years of practice through independent book reading.

Gold, silver and tin: standards of evidence p. 4 Cognitive load theory under attack? p. 20 The science on reading fluency p. 28

Nomanis

Reading | Teaching | Learning | Connecting Issue 17, June 2024

TESCO 4G 7:05 PM 62%

You reposted

Alex Quigley
@AlexJQuigley

There are a small number of high frequency words that feature in over half of the words in children's books (54%) - however the smaller subset of rare words is where the complexities and comprehension lies (notice the difference!). By @Kathy_Rastle & Maria Korochkina at #rED24

3:01 PM · Sep 7, 2024 · 2,691 Views

1 11 16 1

x.com

Dissemination & impact



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON



Powerful data ▾

Effective literacy ▾

Research & evidence

dev.literacy.fft.org.uk/app/workbench/wordusage

Welcome nicholas.pattman@fft.org.uk Logout

CYP-LEX Search

[Choose Columns](#) [Hide Help](#)

The CYP-LEX data set
The Children and Young People's Books-Lexicon (CYP-LEX) is a lexical data set derived from books popular with children and young people in the United Kingdom. It includes 1,200 books evenly distributed across three age bands (7-9, 10-12, 13+) and comprises over 70 million tokens (words) and over 105,000 types (unique words).

The Zipf frequency score
Scores are provided as Zipf-transformed frequencies, calculated as follows:

$$\text{Zipf} = \log_{10} \left(\frac{\text{raw frequency count} + 1}{N \text{ tokens in millions} + N \text{ types in millions}} \right) + 3.0$$

For full details of the data please see the [Open Science Framework](#) for CYP-LEX (opens in a new window).

Data fields

- Word**
The word (token) used to uniquely identify the data
- Score**
The overall Zipf score of the word
- Lemma**
The lemma (root / head) word
- Cyplex 7-9 yr old (rank)**
The CYP-LEX Zipf score for the 7-9 year old band (and the rank of this value within the field)
- Cyplex 10-12 yr old (rank)**
The CYP-LEX Zipf score for the 10-12 year old band (and the rank of this value within the field)
- Cyplex 13+ yr (rank)**
The CYP-LEX Zipf score for the 13 years plus band (and the rank of this value within the field)
- CBeebies Score**
The Zipf score CBeebies subtitles of the CYP-LEX words

10/04/2024

Dissemination & impact



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

<https://doi.org/10.31219/osf.io/vg8c3>

The vocabulary barrier in the General Certificate of Secondary Education (GCSE) in English Literature

Maria Korochkina and Kathleen Rastle

Department of Psychology, Royal Holloway, University of London, United Kingdom

Vocabulary Sep 21, 2024 · by Alex Quigley

The rare vocabulary problem in English Literature

Emerging research from Maria Korochkina and Kathleen Rastle, entitled 'The Vocabulary Barrier in the General Certificate of Education (GCSE) in English Literature', has shown the vocabulary in the GCSE English

Literature texts is uniquely rare. They reveal it would be *access even for avid teen readers*, whilst weaker readers have a *chance to engage with the language of the texts*.

tes
magazine

Teaching & Learning ▾ Scotland Leadership ▾ Newsletters Jobs and more ▾

What makes GCSE English lit so hard for students?

New research has analysed GCSE literature texts against popular fiction – and found three key reasons why so many young people struggle with the qualification

4th August 2024, 8:00am

Maria Korochkina and Kathleen Rastle



Morphology in children's books: What's there and

what's useful for learning?

Maria Korochkina^{1*} and Kathleen Rastle¹

Thank you!



ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

maria.korochkina@rhul.ac.uk
<https://mariakna.github.io/>