

# Estimating Lexical Complexity from Document-Level Distributions

Sondre Wold<sup>1</sup>, Petter Mæhlum<sup>1</sup>, Oddbjørn Hove<sup>2</sup>

<sup>1</sup> University of Oslo, Department of Informatics, Language Technology Group (LTG)

<sup>2</sup> Helse Fonna



## Motivation

- ▶ Existing methods for complexity estimation are typically developed for entire documents.
- ▶ This limitation in scope makes them inapplicable for shorter pieces of text, such as health assessment tools.
- ▶ Answering questions like “How much discomfort do the obsessions cause?” demands that the respondent meet various linguistic requirements, including vocabulary knowledge and syntactic skills.
- ▶ We develop a two-step approach for estimating *lexical* complexity that does not rely on any pre-annotated data, targeting the Norwegian language.
- ▶ Our method can be used to suggest lexical substitutions of lower complexity.

## Our Method

Based on two assumptions:

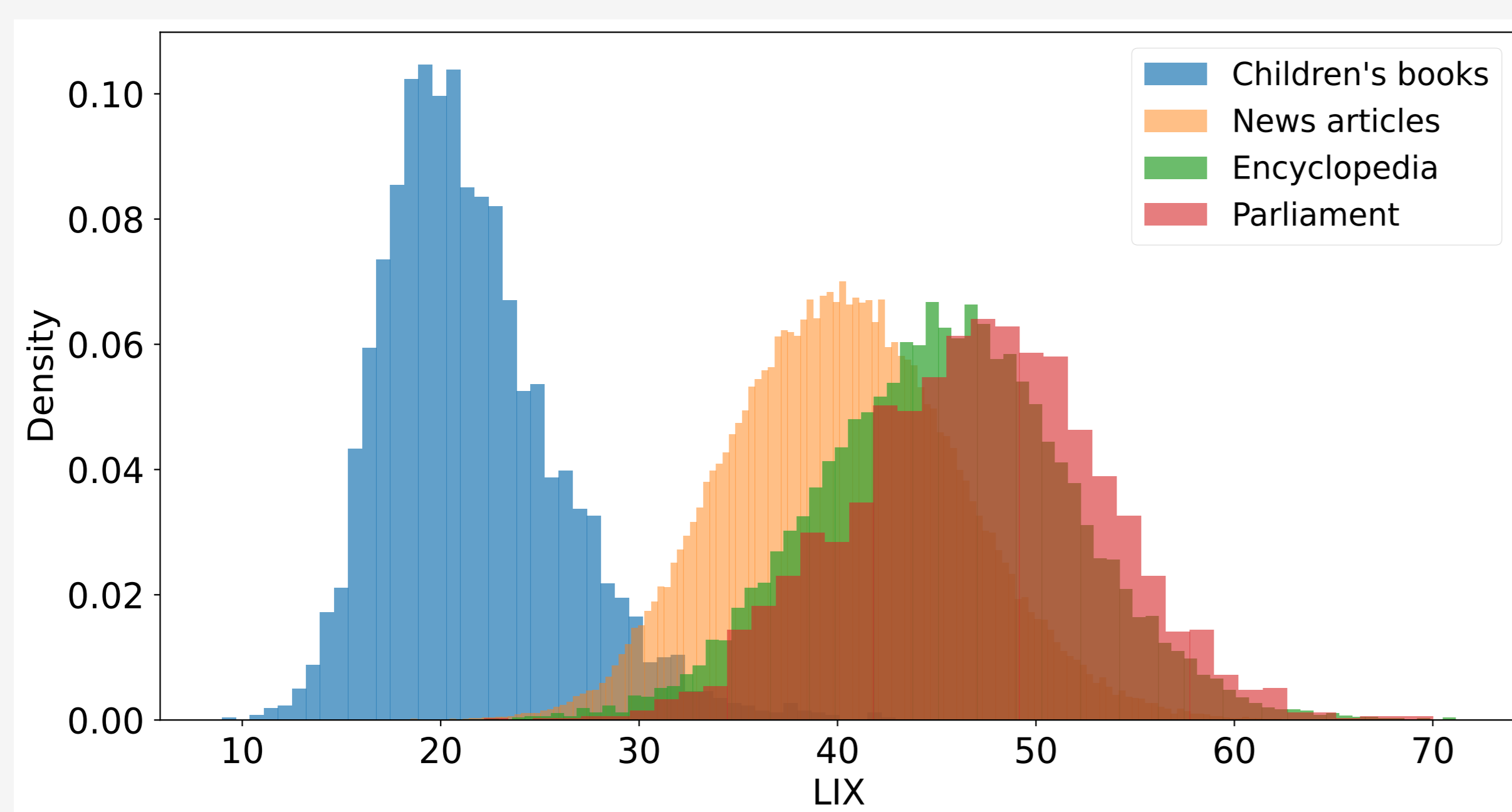
1. Words of high lexical complexity appear more frequently in documents with high levels of complexity.
2. If a document-level complexity measure can separate documents based on complexity, then this metric contains information on the complexity of individual words.

As a document level measure of complexity, we use the LIX Score (Björnsson, 1968)

$$LIX = \frac{A}{B} + \frac{C * 100}{A},$$

where  $A$  is the number of tokens,  $B$  the number of sentences and  $C$  is the number of words with  $> 6$  letters.

## Distribution of LIX Scores



We verify that the samples are unlikely to have been drawn from the same distribution using a 2-way Kolmogorov–Smirnov test between the corpora, essentially showing that the LIX metric can separate documents into categories that match or intuition about the complexity of these documents.

## Data Collection

We collect and process documents from four different sources that are *assumed* to be of different complexity:

- ▶ **Children's books:**
  - ▶ 3695 books, both literary and non-fiction, written between 1950 and 2023.
- ▶ **News articles:**
  - ▶ 111579 articles from the 2019 version of the Norwegian Newspaper Corpus. We include articles from ten different publications ranging from typical tabloids to more traditional prints, and specialized publications focusing on a single topic, like economics.
- ▶ **Encyclopedia entries:**
  - ▶ 17033 texts from the Great Norwegian Encyclopedia (SNL), entries written by domain experts for the general public on a wide range of topics.
- ▶ **Texts from the Norwegian parliament:**
  - ▶ 2726 openly available legislative decision proposals from the Norwegian parliament.

While the children's book collection is not openly available, the necessary data to calculate the complexity of these books can be accessed using the DH-LAB API service, available as a Python library, from the Norwegian National Library.

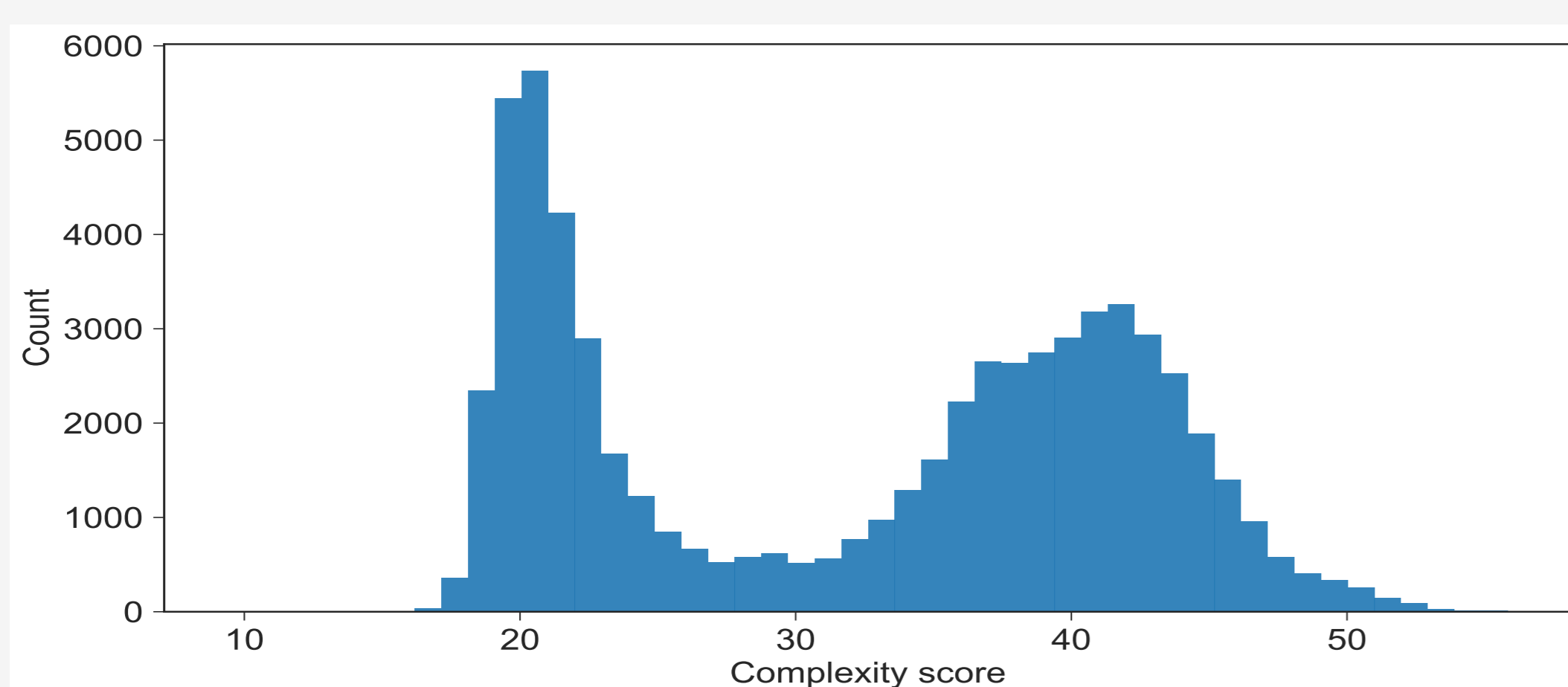
## Creating a Lexical Complexity Metric

Based on Assumption 1 and the effectiveness of the LIX as a document-level complexity metric, we define our lexical complexity score (CS) as:

$$CS(\text{lemma}) = x * (1 - \frac{n}{m}),$$

where  $x$  is the median LIX score of the  $n$  documents in which this lemma occurs, and  $m$  is the total number of documents. This is essentially discounting the median with the proportion of the documents in which this lemma occurs.

- ▶ We want to push high-frequency words to the lower rangers of the distribution.
- ▶ We only focus on content words with the following parts of speech: nouns, verbs, adjectives, and adverbs.



## Evaluations

We can pair the CS with a word-embedding model to generate substitutions of lower complexity. For two samples from Yale-Brown Obsessive Compulsive Scale inventory:

- ▶ ...**medfører** *betydelig svekkelse / sosiale eller / arbeidsmessig utfoldelse* ‘...causes substantial impairment in social or occupational performance’
- ▶ *Hvor mye* **ubehag** *medfører tvangstankene?* ‘How much discomfort do the obsessions cause?’

Lemma		CS	#	Lemma		CS	#
<i>bety</i>	‘means’	33.09	18330	<i>tretthet</i>	‘tiredness’	21.55	228
<i>resultere</i>	‘result’	40.77	1675	<i>skyldfølelse</i>	‘guilt’	29.64	198
<b>medføre</b>	‘cause’	41.67	3984	<i>smerte</i>	‘pain’	31.71	2685
<i>tilsi</i>	‘entail’	41.88	2190	<i>irritasjon</i>	‘irritation’	32.98	435
<i>vanskeliggjøre</i>	‘convolute’	46.07	199	<i>stress</i>	‘stress’	34.86	784
<i>nødvendiggjøre</i>	‘necessitate’	47.17	22	<b>ubehag</b>	‘discomfort’	38.37	392

## Coleman-Liau Index (CLI)

We compare LIX and CLI and find that all pair-wise distances are still significant, but with smaller margins. The CLI equation is as follows:

$$0.0588 * L - 0.296 * S - 15.8,$$

where  $L$  is the average number of letters per 100 words and  $S$  is the average number of sentences per 100 words.

## Complexity Score and Word-Level Features

- ▶ We do not observe any correlation between our scores and word-level features such as length and the number of syllables.
- ▶ Words of different lengths are evenly spread across the complexity spectrum and words with more syllables do not receive higher scores through our method, with the exception of short words being somewhat more frequent in the lower ranges.

## Data, Code and Interactive Model

- ▶ [https://github.com/SondreWold/lexical\\_complexity\\_estimation](https://github.com/SondreWold/lexical_complexity_estimation)

## Conclusion and Limitations

- ▶ Some words are almost exclusively used in one category, which might obfuscate the score. E.g ‘budgeriar’ has a lower CS than ‘bird’, because it is a common illustration in children's books.
- ▶ We find that we can construct a lexical complexity score without the use of any annotation efforts.
- ▶ We show how it can be used on the fly to generate substitution suggestions for simplifying mental health assessment tools, and improving their cognitive accessibility.
- ▶ We find that lexical complexity does not correlate with word-level features such as length and the number of syllables.