Estimating Lexical Complexity from Document-Level Distributions

LREC-COLING 2024

Sondre Wold[1], Petter Mæhlum[1], Oddbjørn Hove[2]

[1]University of Oslo, [2]Helse Fonna

Introduction

- Existing methods for complexity estimation are typically developed for entire documents.

Introduction

- 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.

Introduction

- 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.

Introduction

- 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.

Introduction

- 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.

Methods

Based on two assumptions:

1. Words of high lexical complexity appear more frequently in documents with high levels of complexity.

Methods

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.

Methods

. . . consequently, if we can *i)* show that a document-level complexity measure can separate documents according to their assumed complexity

Methods

. . . consequently, if we can *i)* show that a document-level complexity measure can separate documents according to their assumed complexity, then *ii)* the average complexity score of the documents in which a lemma occurs will say something about the complexity of that particular lemma.

Data Collection 1

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

Data Collection 1

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

The first two collections were assumed to be of simple and medium complexity:

- Children's books:
  - 3 695 books, both literary and non-fiction, written between 1950 and 2023. Collected using the dhlab API from the Norwegian National Library.
- News articles:
  - 111 579 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.

Data Collection 2

The next two were assumed to be of medium and high complexity:

- Encyclopedia entries:
  - 17 033 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:
  - 2 726 openly available legislative decision proposals from the Norwegian parliament.

LIX-score

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.

- LIX is developed for Swedish, but has a history of use for Norwegian as well.

Overall statistics

Category	**LIX**
Very easy	20
Easy	30
Medium difficulty	40
Difficult	50
Very difficult	60

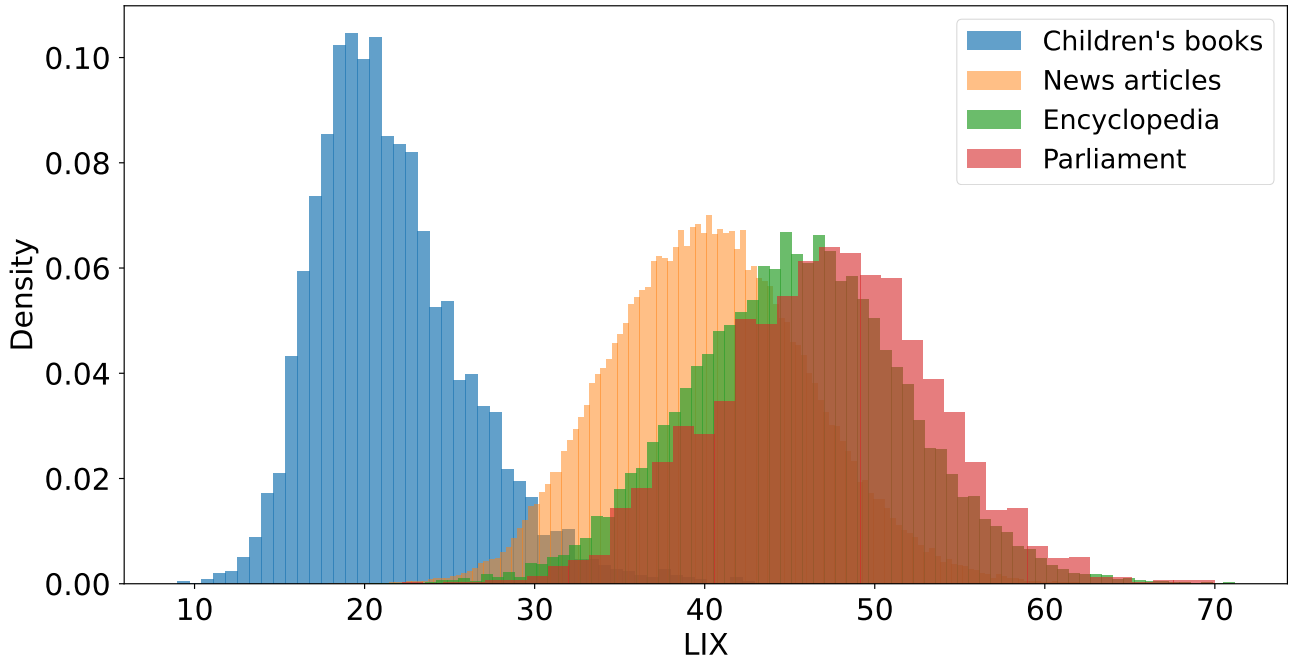
Table: Interpretation of LIX scores on a five-step scale of complexity.

Overall statistics

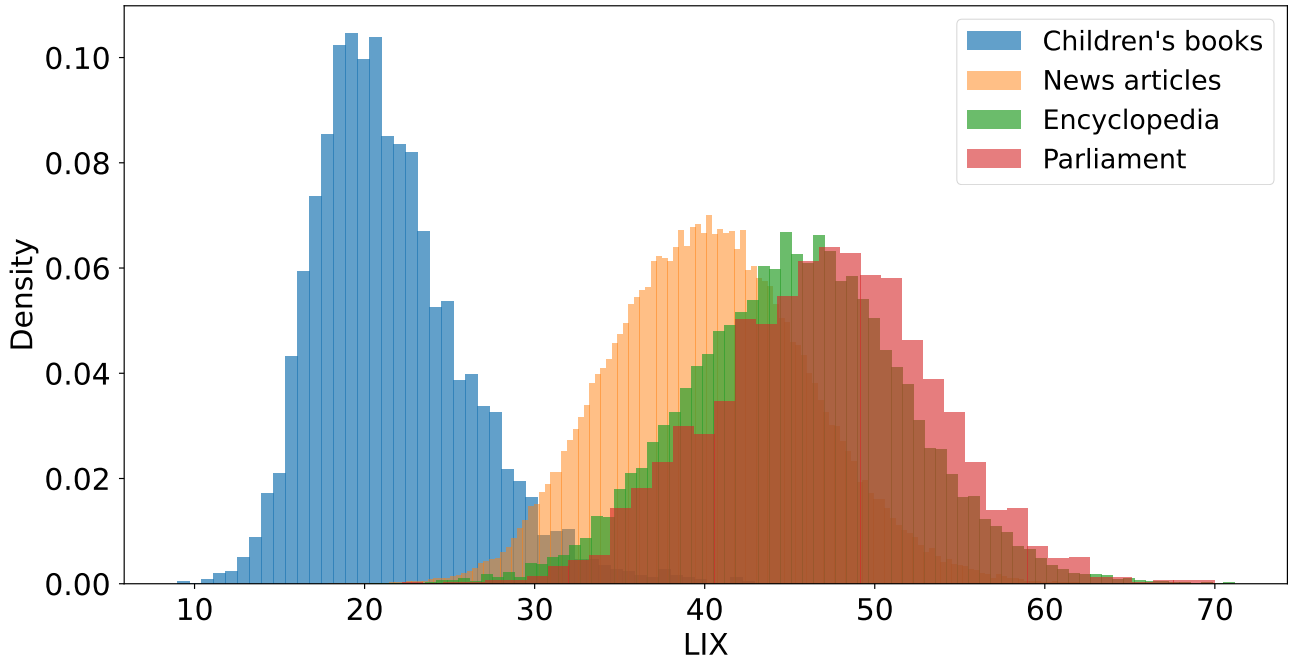
Dataset	**#**	**LIX**
Children	3695	$\mu = 21.57, \sigma = 4.56$
News	111579	$\mu = 40.32, \sigma = 5.82$
Encyclopedia	17033	$\mu = 45.40, \sigma = 6.40$
Parliament	2726	$\mu = 47.04, \sigma = 6.36$
Total	135033	$\mu = 40.58, \sigma = 6.94$

Table: Statistics for the different corpora with count (#), mean (μ) and standard deviation (σ) after pre-processing.

Distribution of LIX scores



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.

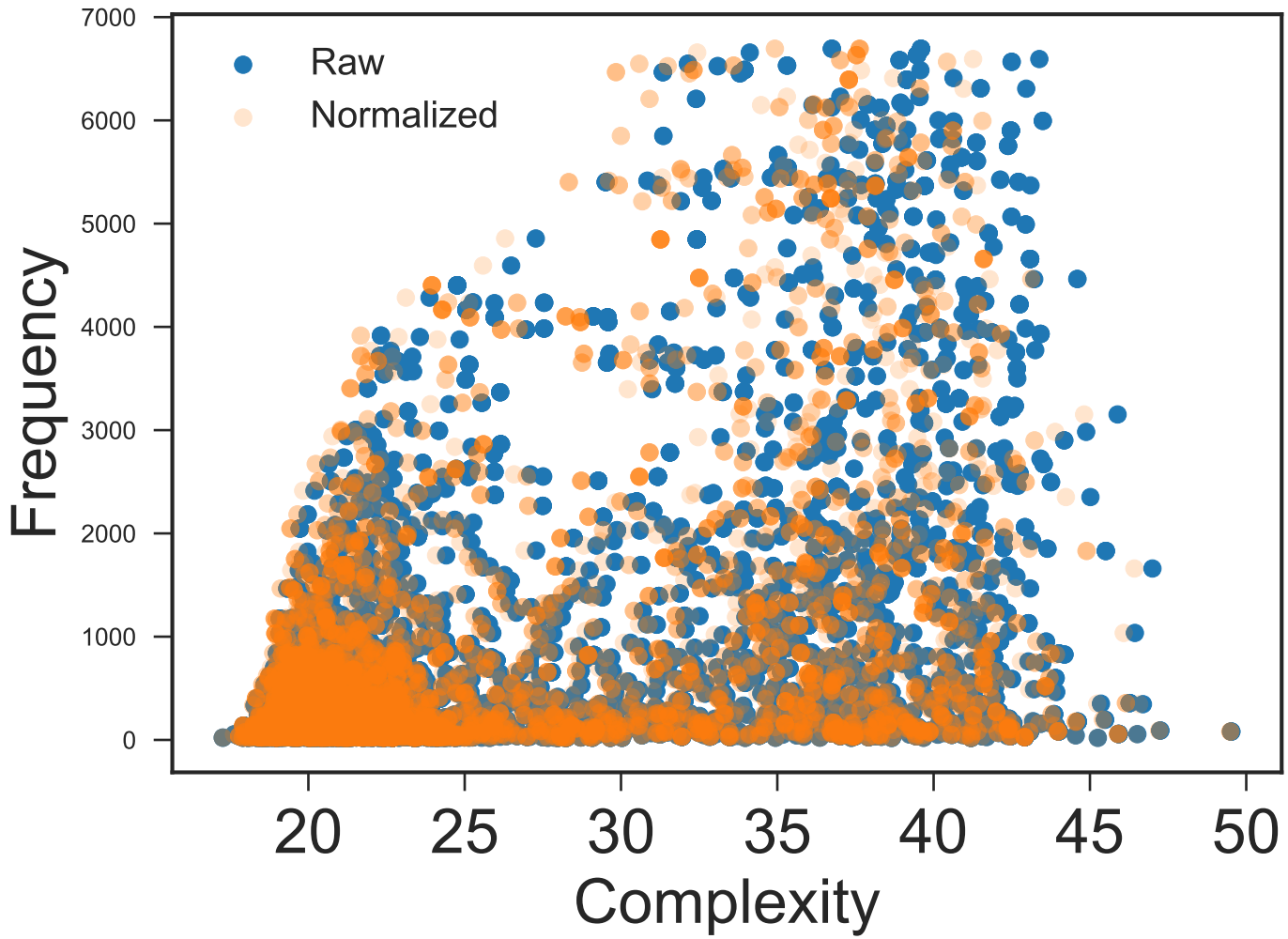
Creating a lexical complexity metric 1

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

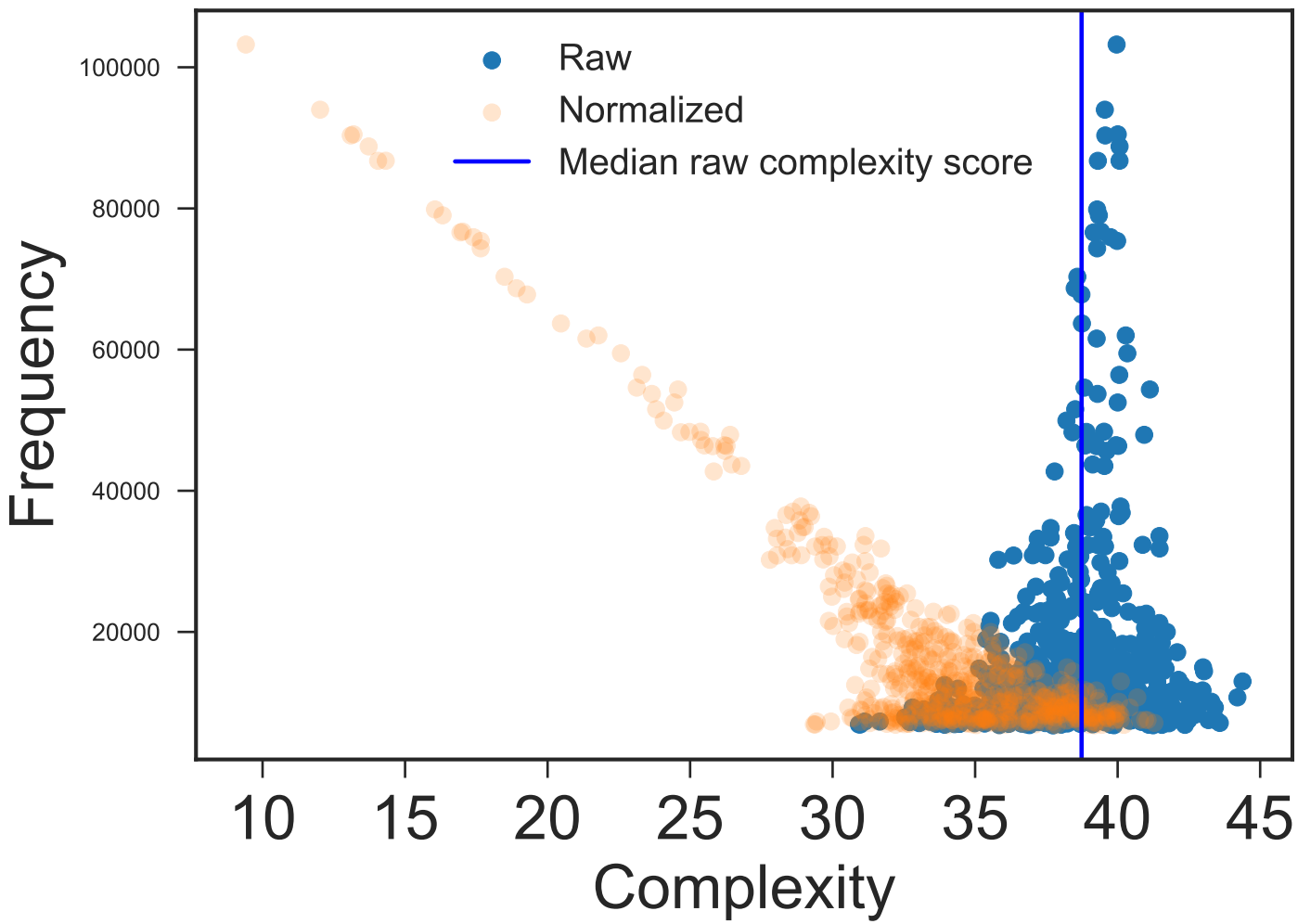
$$CS(lemma) = x * (1 - (\frac{n}{m})),$$

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.

Normalization of low frequency terms (bottom 5%

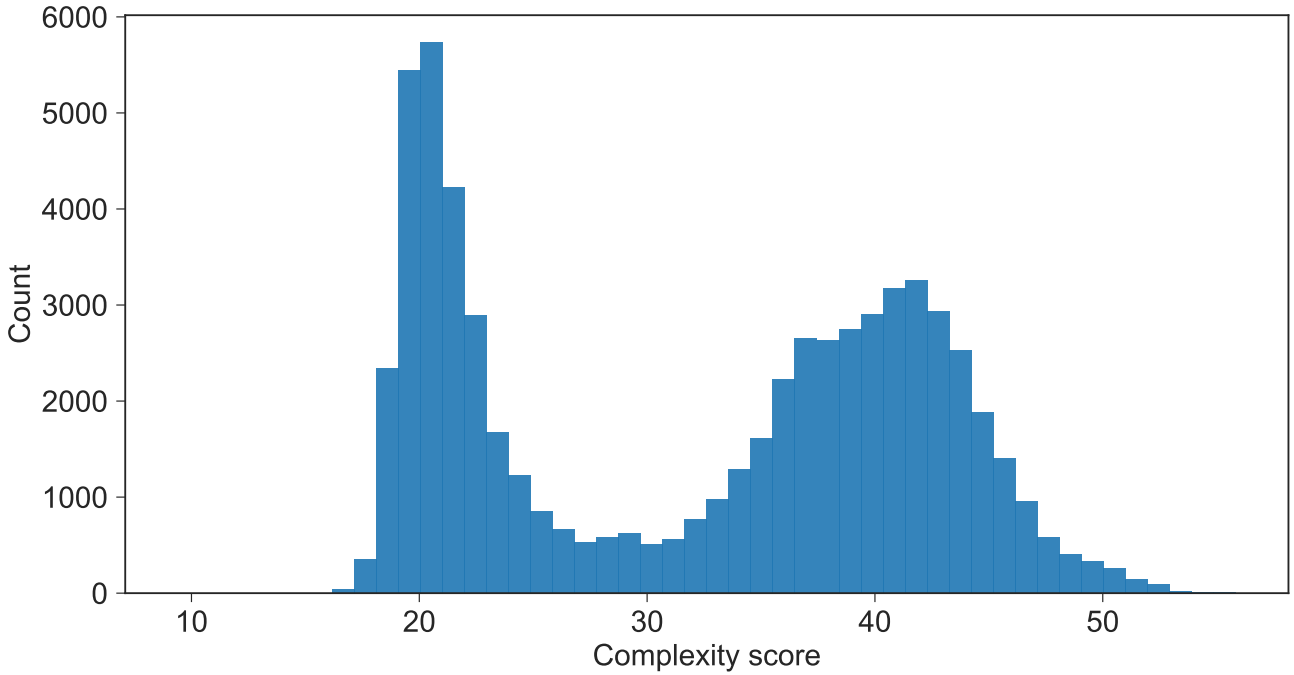


Normalization of high frequency terms (top 5%)



Creating a lexical complexity metric 2

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



Qualitative evaluation

We can pair the CS with a word-embedding model to generate possible lexical substitutions that has a lower complexity. We look at samples from the Yale-Brown Obsessive Compulsive Scale inventory and (YBOS) the Eating Disorder Examination Questionnaire (EDE-Q 6.0).

Evaluation Example 1, EDE-Q

- *Har du prøvd å følge **bestemte** regler for hva eller hvordan du spiser (f.eks. en kalorigrense)...?* 'Have you tried to follow **specific** rules for what or how you eat (e.g. a calorie limit)...?'

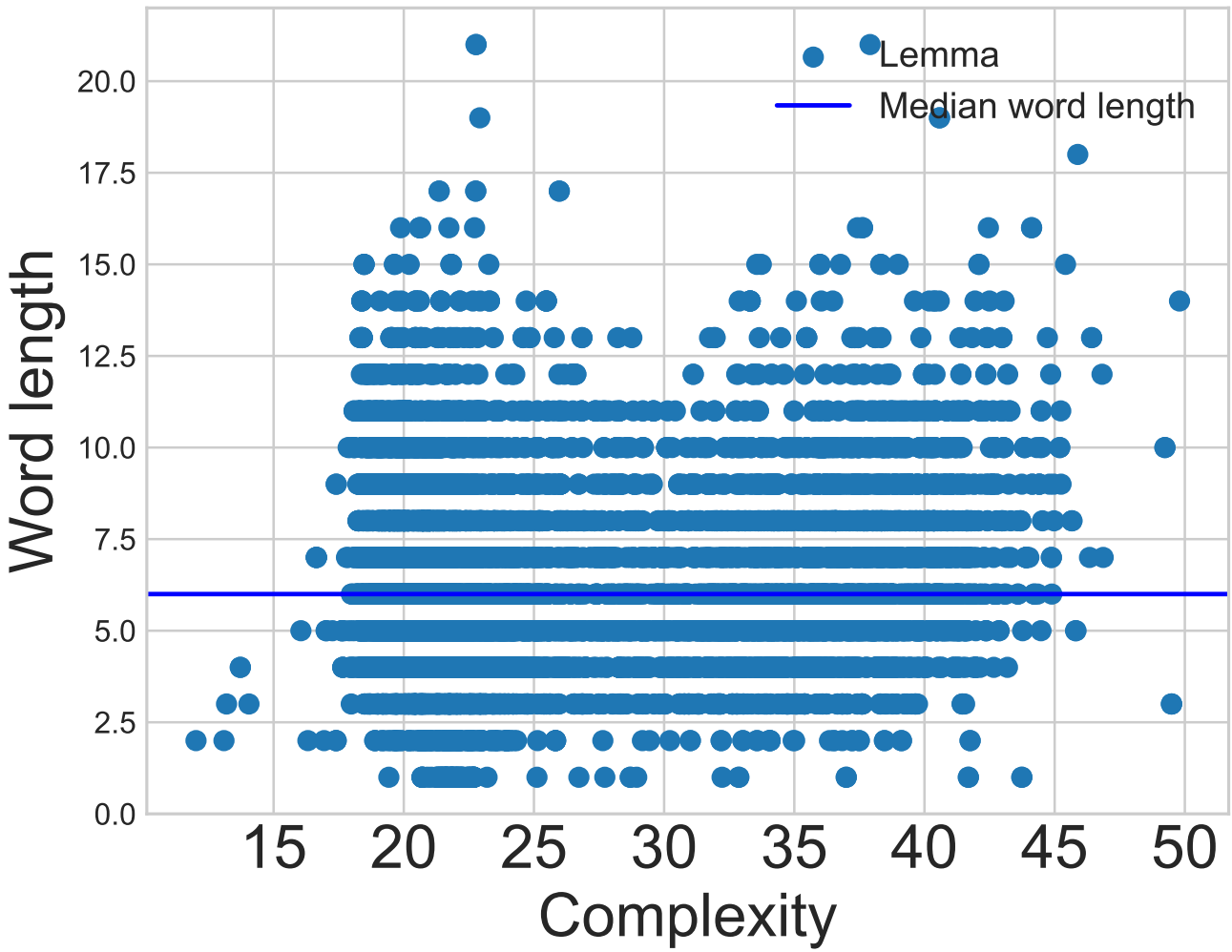
Lemma		**CS**	#
viss	'certain'	34.72	7009
bestemt	'specific'	37.40	3851
enkelt	'some'	38.51	10316
akseptabel	'acceptable'	41.16	732
nøytrale	'neutral'	41.61	827
spesifikk	'specific'	41.98	1142

Evaluation Example 2, YBOS

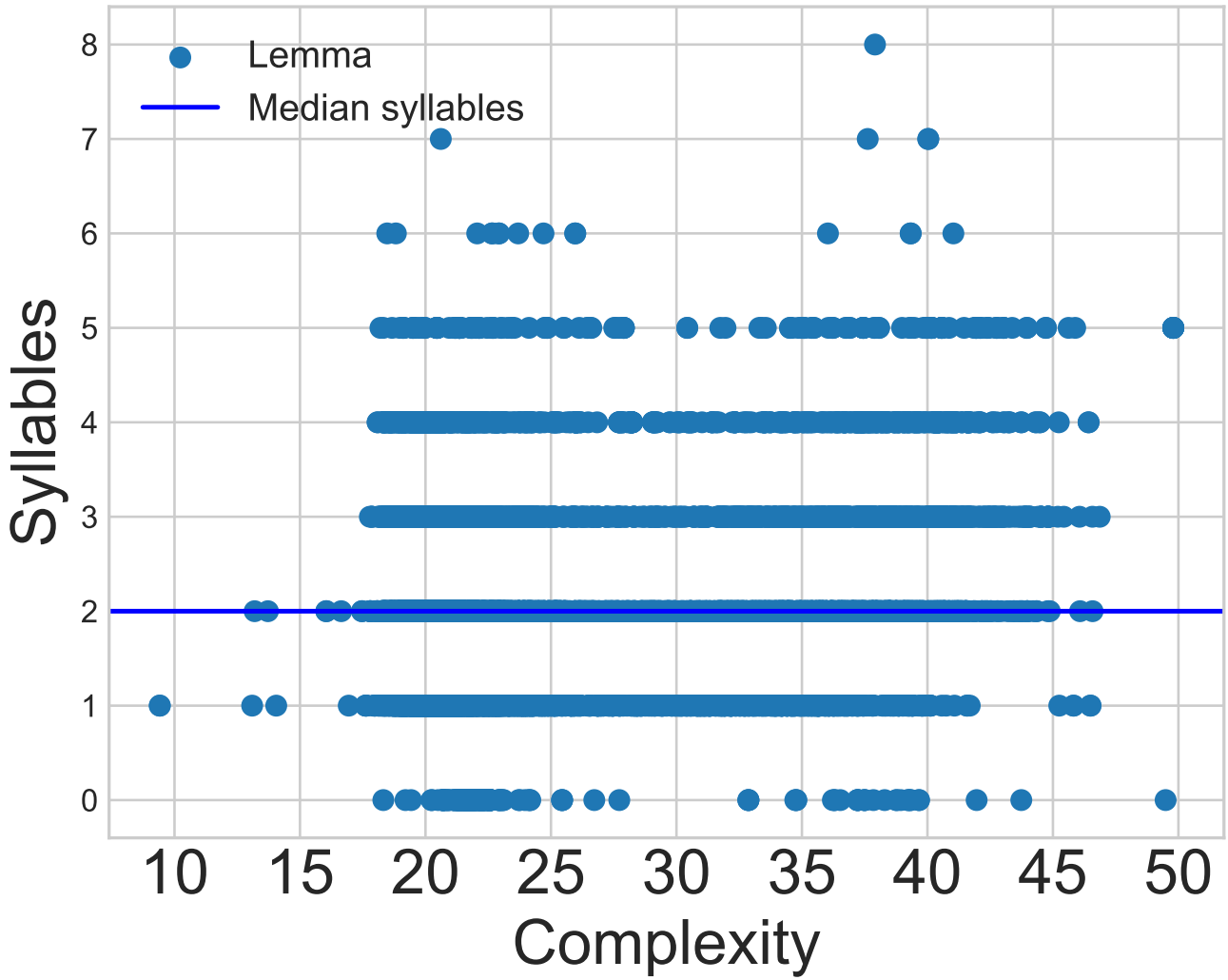
- *Hvor mye **ubehag** medfører tvangstankene?* 'How much **discomfort** do the obsessions cause?'

Lemma		**CS**	#
tretthet	'tiredness'	21.55	228
skyldfølelse	'guilt'	29.64	198
smerte	'pain'	31.71	2685
irritasjon	'irritation'	32.98	435
stress	'stress'	34.86	784
ubehag	'discomfort'	38.37	392

Complexity and word length



Complexity and morphology



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, except for short words being somewhat more frequent in the lower ranges.

Conclusion and limitations

- Some words are almost exclusively used in one category, which might obfuscate the score. E.g. ‘budgerigar’ 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.

Thank you for your attention!