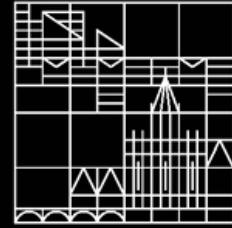




Universitat  
Pompeu Fabra  
Barcelona

Universität  
Konstanz



# An Open Multilingual System for Scoring Readability of Wikipedia

**Mykola Trokhymovych**

[mykola.trokhymovych@upf.edu](mailto:mykola.trokhymovych@upf.edu)

**Indira Sen**

[indira.sen@uni-konstanz.de](mailto:indira.sen@uni-konstanz.de)

**Martin Gerlach**

[mgerlach@wikimedia.org](mailto:mgerlach@wikimedia.org)



# Readability & Wikipedia

## What is Readability?

The readability concept aims to capture how easy it is to read a given text. It is usually defined as the sum of all factors that affect a reader's understanding, reading speed, and level of interest.

(Dale and Chall, 1949)

## Readability of Wikipedia

“[in English Wikipedia the] overall readability is poor.”

(Lucassen et al., 2012)

“Most of the health information [in English Wikipedia] remains written at a level above the reading ability of average adults.”

(Brezar and Heilman, 2019)

# Measuring readability is challenging

## Common approaches

are readability formulas

e.g., Flesch Reading Ease

(FRE) Score Formula

$$206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$$

## Wikipedia exists in more than 300 languages:

- Adapting readability formulas to each language not scalable/feasible
- Lack of multilingual models/tools
- Scarcity of multilingual ground-truth data
  - Standard datasets are not open (e.g. Newsela)
  - Mostly datasets are for English (no data for most languages)

# In this research we aim to ...

build a multilingual model that can automatically access the readability of the text of Wikipedia articles across languages.



Тунель Кохання —  
Die Bahnstrecke Klewan—  
愛情隧道 (烏克蘭語: Тунель  
Т. Кох The **Tunnel of Love** (**Ukrainian**:  
К. 諾州 Тунель Кохання, *Tunel*  
公司 *Kokhannya*) is a section of  
且不 industrial railway located near  
К. **Klevan, Ukraine**, that links it with

ukwiki	8.83	👩
dewiki	8.37	👨
zhwiki	12.1	👩🎓
enwiki	4.35	👶

Readability score

# A new multilingual dataset

## WikiReaD (Wikipedia Readability Dataset)

What? Pairs of encyclopedic articles in two readability levels (simple, hard) for 14 languages.

How? Matching Wikipedia articles with a simplified version.

- Simple English Wikipedia (>100K matched articles)
- Children encyclopedias (10–10K)
  - Vikidia (11 languages such as Greek, Spanish, French, Portuguese, Armenian, etc.)
  - Klexikon (German)
  - Txikipedia (Basque)
  - Wikikids (Dutch)



Simple English  
WIKIPEDIA



Vikidia Txikipedia

# Example



## Easy (Simple Wikipedia)

Title : Flag of Thailand

### Sentences :

The Flag of Thailand has five stripes red, white, and blue.

Red–white–blue stand for nation–religion–king, an unofficial motto of Thailand.

The flag was made official on 28 September 1917 by a royal decree.

The Thai name for the flag is ธงไตรรงค์ ("Thong Trairong"), which means "tricolour flag"

## Hard (English Wikipedia)

Title : Flag of Thailand

### Sentences :

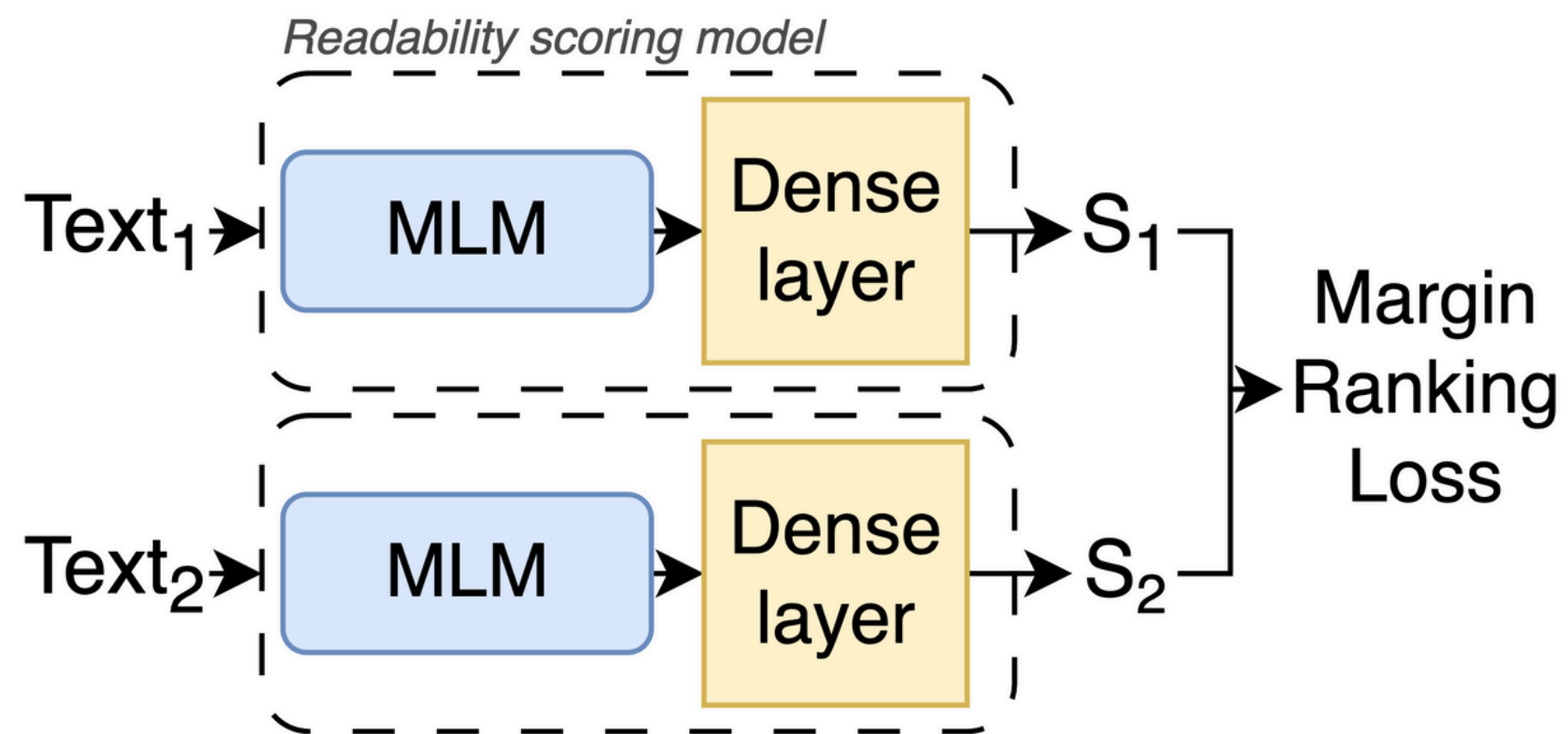
The flag of Thailand (Thai: ธงไตรรงค์; thong trai rong, meaning 'tricolour flag') shows five horizontal stripes in the colours red, white, blue, white and red, with the central blue stripe being twice as wide as each of the other four.

The design was adopted on 28 September 1917, according to the royal decree issued by Rama VI.

Since 2016, that day is a national day of importance in Thailand celebrating the flag.

....

# A new readability scoring model



- xlm-roberta-longformer as the base MLM
- ~100 languages support
- Siemease architecture for training:
  - Training is based on pairwise ranking, but still able to score individual texts
- Loss: Margin Ranking Loss
- Fine-tuning only using data in English:
  - No language-specific fine-tuning for other languages! (zero-shot)

# Model performance

**Evaluation metric:** Ranking accuracy (RA)

**Explanation:** RA is a percentage of pairs that are ranked correctly

**Setup 1:** Evaluation on English.

- Simple Wikipedia (training corpus): RA=0.976
- Vikidia (out-of-corpus): RA=0.991

**Setup 2:** Evaluation on Non-English data (zero-shot cross-lingual transfer)

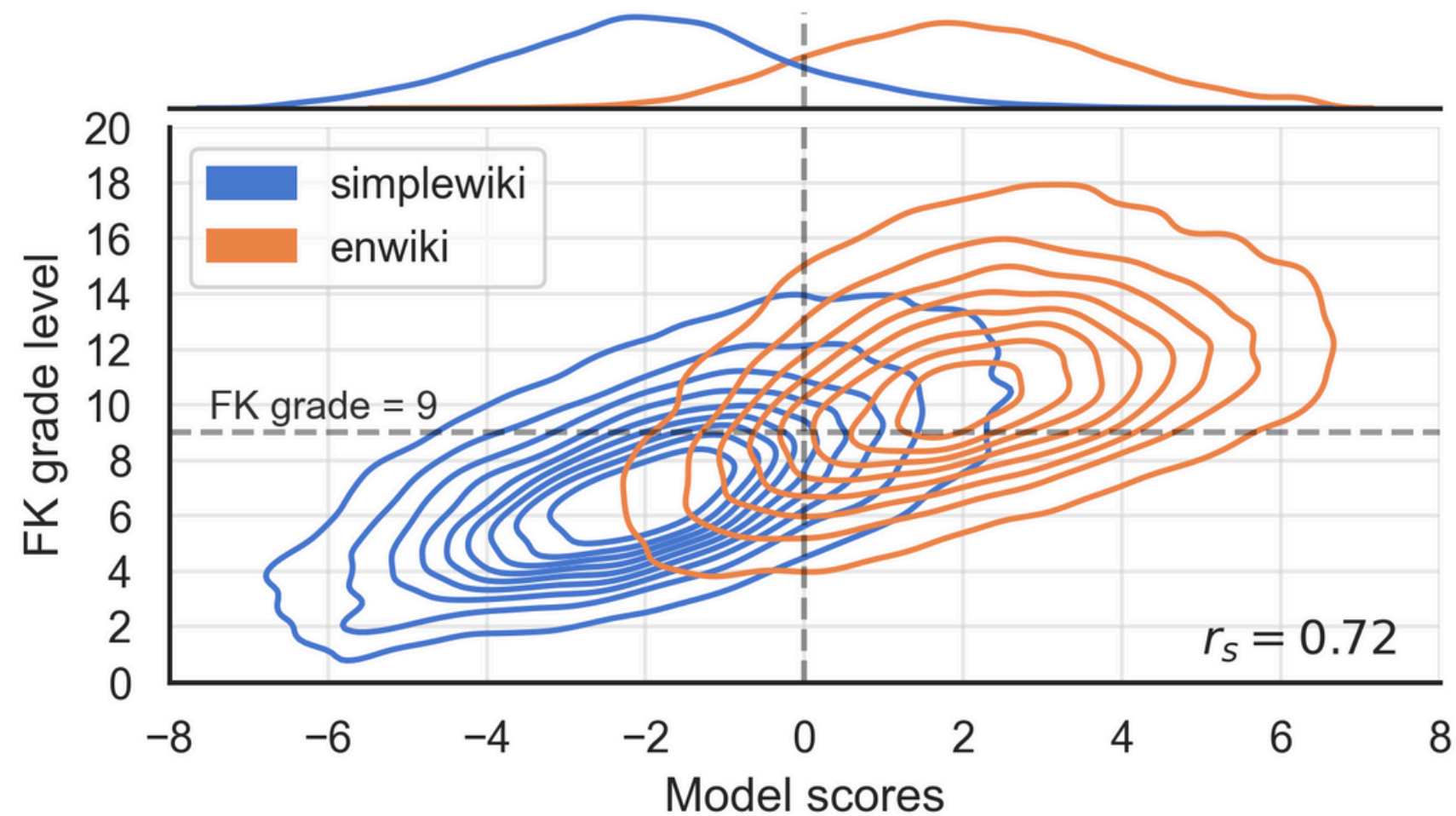
- RA > 0.8 for all languages
- RA > 0.9 for 10/15 datasets

**Setup 3:** Evaluation on reference benchmarks (e.g. Lee&Vajjala, ACL 2022)

- Outperform reference benchmarks
- VikidiaFr: 0.978 (ours) vs 0.811 (NPRM)
  - OneStopEnglish: 0.974 (ours) vs 0.878 (NPRM)



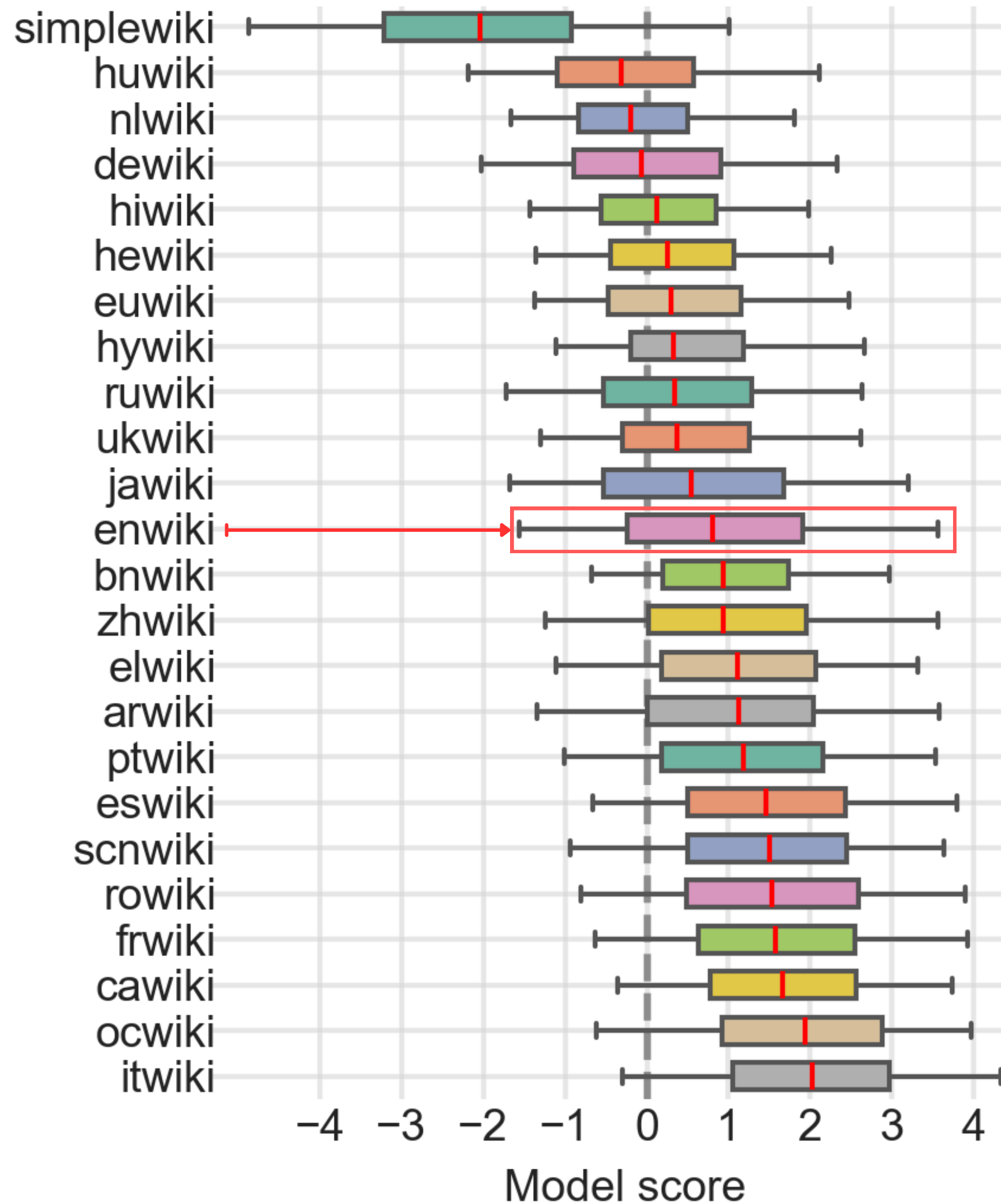
# Model scores correlate with existing readability formulas



A model score of approx. 0 separates texts from simplewiki (easy) and enwiki (hard).

This corresponds to Flesch-Kincaid grade level (FKGL)  $\sim 9$ .

# Wikipedia is difficult to read across most languages



In English Wikipedia, more than half the articles exceed the average readability level of American adults (~grade level 8–9) (Brezar&Heilman 2019)

Overall readability in most Wikipedias are similar to English Wikipedia

# Contributions

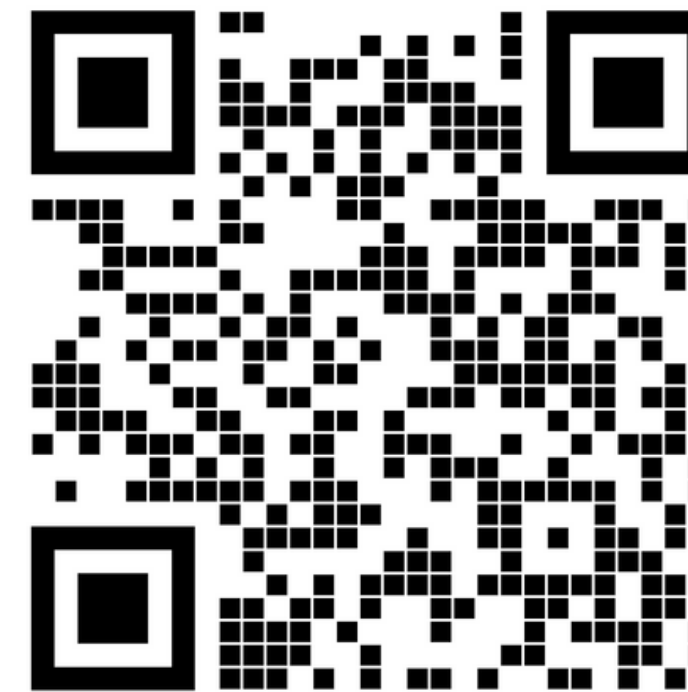
- Release a new **open multilingual dataset** covering 14 languages
- Develop a **single multilingual model** for text readability scoring, working in **zero-shot cross-lingual transfer**
- Present the first systematic overview of the **state of readability of Wikipedia** articles **beyond English**
- Provide a **public API endpoint** of the model for use by readers, editors, and researchers



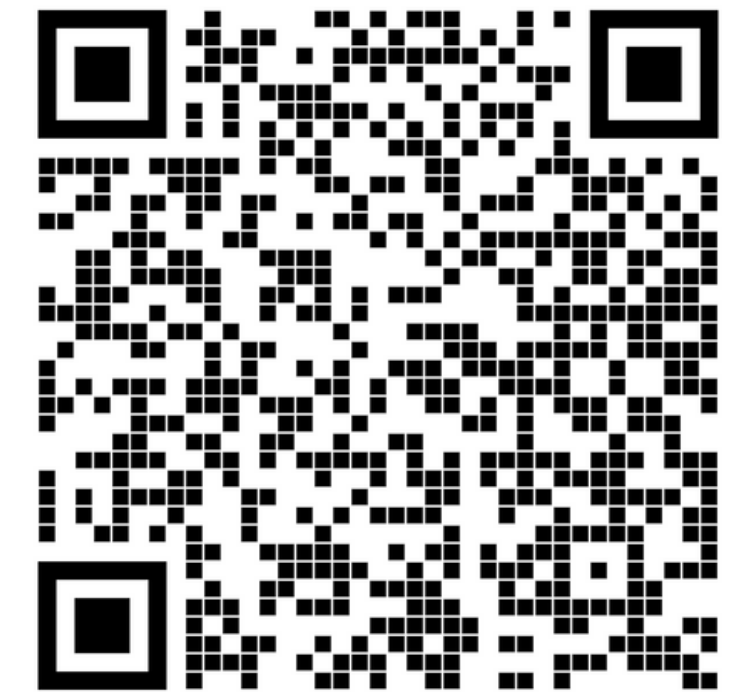
The image shows a screenshot of a Wikipedia article snippet for 'Тунель Кохання' (The Tunnel of Love). The snippet includes text in multiple languages: Ukrainian, German, Chinese, and English. A readability score overlay is present on the right side of the snippet, showing scores for different language editions and a readability score for the English edition.

Language	Readability Score
ukwiki	8.83
dewiki	8.37
zhwiki	12.1
enwiki	4.35

Readability score



WikiReaD dataset



API endpoint



**Thank you!**