
Erasmus Mundus Master's Degree in Language and Communication Technologies
*MSc in Natural Language Processing, University of Lorraine*
*MSc in Language Science and Technology, Saarland University*

# NLP Methods to Automate Language Learning through Extensive Reading

Sémagramme Seminar

15 December 2020

**Author:**
Siyana Pavlova

**Supervisors:**
Prof. Dr. Günter Neumann
Saadullah Amin (unofficial)
Prof. Dr. Josef van Genabith

# Overview

- Goal and Motivation
- Similar Systems
- Building the solution
  - Functional Requirements
  - System Architecture
  - Data
  - Topic classification experiment
- Demo
- System details
- User study
- Future work

# Goal

Develop a language learning application, which uses Extensive Reading to build a vocabulary list and Spaced Repetition to aid users learn vocabulary items.

Automation of the content creation and processing should be used wherever possible.

# Motivation

- Why Extensive Reading (ER)
- Why Spaced Repetition
- Why automation

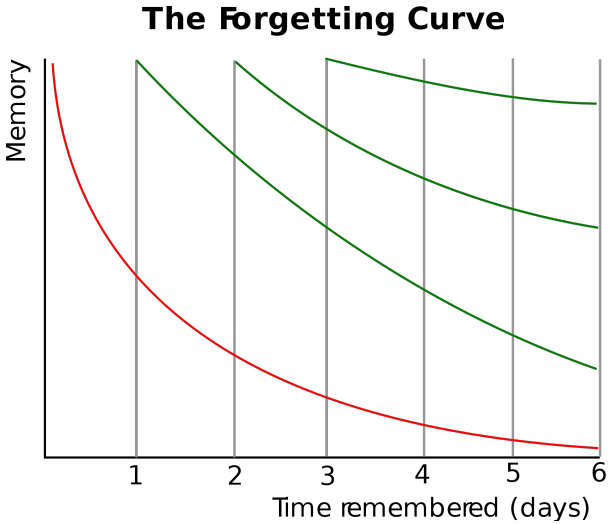
# Extensive Reading (ER)

- What is it?
  - The independent reading of large quantity of text for information or pleasure
  - Can be opposed to Intensive Reading
    - Students spending a lot of time reading short, difficult texts under the supervision of a teacher, the emphasis being on the students developing reading and language skills
  - Some argue that learning a foreign language should not be the primary goal [Kaufmann, 2003]
- Why ER?
  - ER beneficial to foreign language acquisition [Mason and Krashen, 1997]
  - More beneficial to reading speed and comprehension than Intensive Reading [Bell, 2001]

[Image source: https://favpng.com/png_view/book-clip-art-book-vector-graphics-illustration-image-png/HFpqpm47]

# Spaced Repetition

- What is it?
  - Newly acquired information gets forgotten at progressively larger intervals over time
  - Goal of Spaced Repetition: revise items just before they are forgotten so that they remain in memory
- Why Spaced Repetition
  - Shown to be beneficial to foreign language acquisition [Atkinson, 1972]




[Image source: https://upload.wikimedia.org/wikipedia/commons/4/4e/ForgettingCurve.svg]

# Why automation

- Vast amount of languages
- Vast amount of language pairs
- Vast amount of texts in some/many languages

# Similar systems

| Criteria | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| **Extensive Reading** | | | | | | | | | |
| **Vocabulary Builder** | | | | | | | | | |
| **Spaced Repetition** | | | | | | | | | |
| **Reasonable free version** | | | | | | | | | |
| **Automated Content Creation** | | | | | | | | | |

# Similar systems

| Criteria | Extensive Reading Systems | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | LingQ | Bliu Bliu* | LWT** | | | | | |
| Extensive Reading | ✓ | ✓ | ✓ | | | | | |
| Vocabulary Builder | ✓ | ✓ | ✓ | | | | | |
| Spaced Repetition | ✓ | ╳ | ╳ | | | | | |
| Reasonable free version | ╳ | - | ✓ | | | | | |
| Automated Content Creation | ╳ | Unk | ╳ | | | | | |

*Discontinued since 2016
**Learning with texts

# Similar systems

| Criteria | Extensive Reading Systems | | | Spaced Repetition Systems | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LingQ | Bliu Bliu* | LWT** | Memrise | Duolingo | Anki | Lingvist | Glossika | |
| Extensive Reading | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | |
| Vocabulary Builder | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | |
| Spaced Repetition | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Reasonable free version | ✗ | - | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | |
| Automated Content Creation | ✗ | Unk | ✗ | ✗ | ✓ | ✗ | Unk | ✓ | |

*Discontinued since 2016
**Learning with texts

# Similar systems

| Criteria | Extensive Reading Systems | | | Spaced Repetition Systems | | | | | Proposed Solution |
|---|---|---|---|---|---|---|---|---|---|
| | LingQ | Bliu Bliu* | LWT** | Memrise | Duolingo | Anki | Lingvist | Glossika | |
| Extensive Reading | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Vocabulary Builder | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| Spaced Repetition | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Reasonable free version | ✗ | - | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| Automated Content Creation | ✗ | Unk | ✗ | ✗ | ✓ | ✗ | Unk | ✓ | ✓ |

*Discontinued since 2016
**Learning with texts

# Functional Requirements

**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.

# Functional Requirements

**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.

# Functional Requirements

**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.

# Functional Requirements

**FR1.** User should be able to read texts in their target language. These texts should:

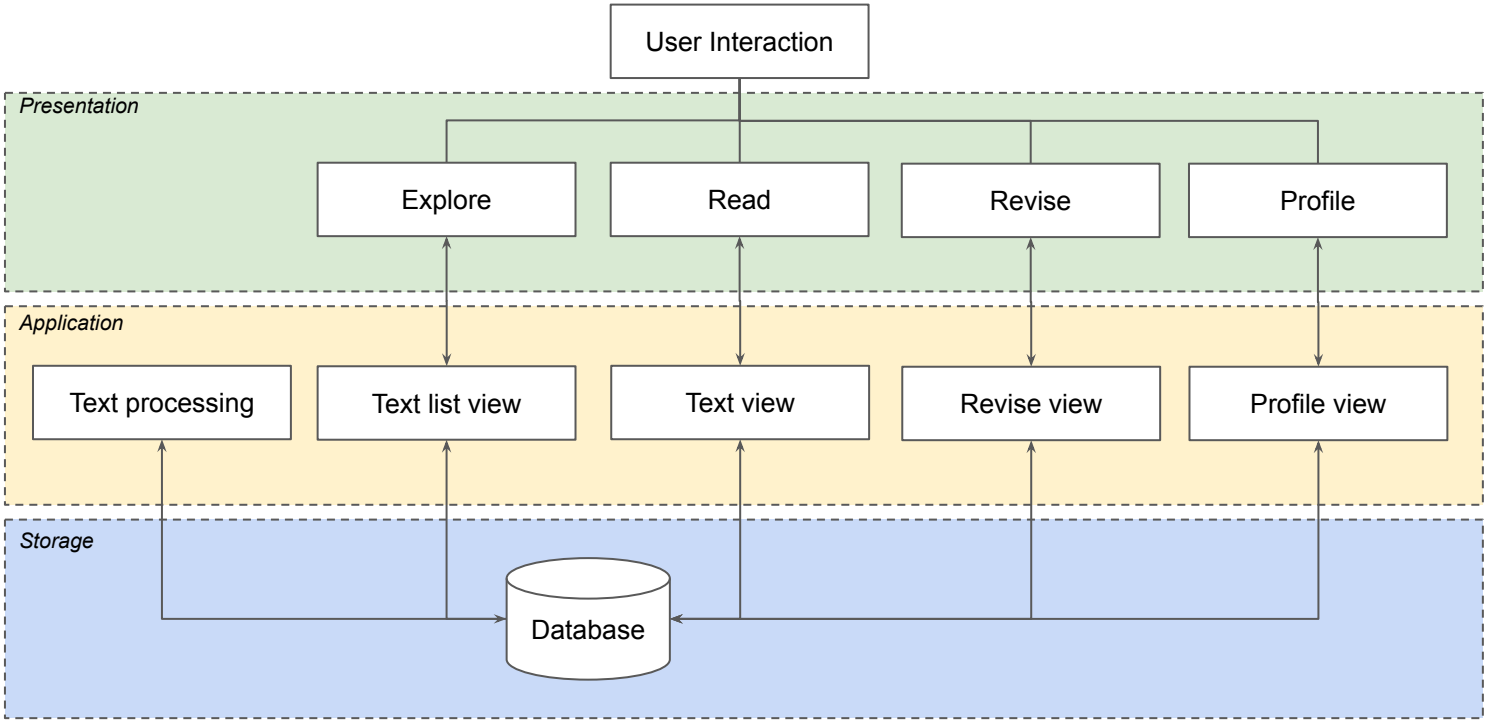
a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.

# System Architecture



# System Architecture

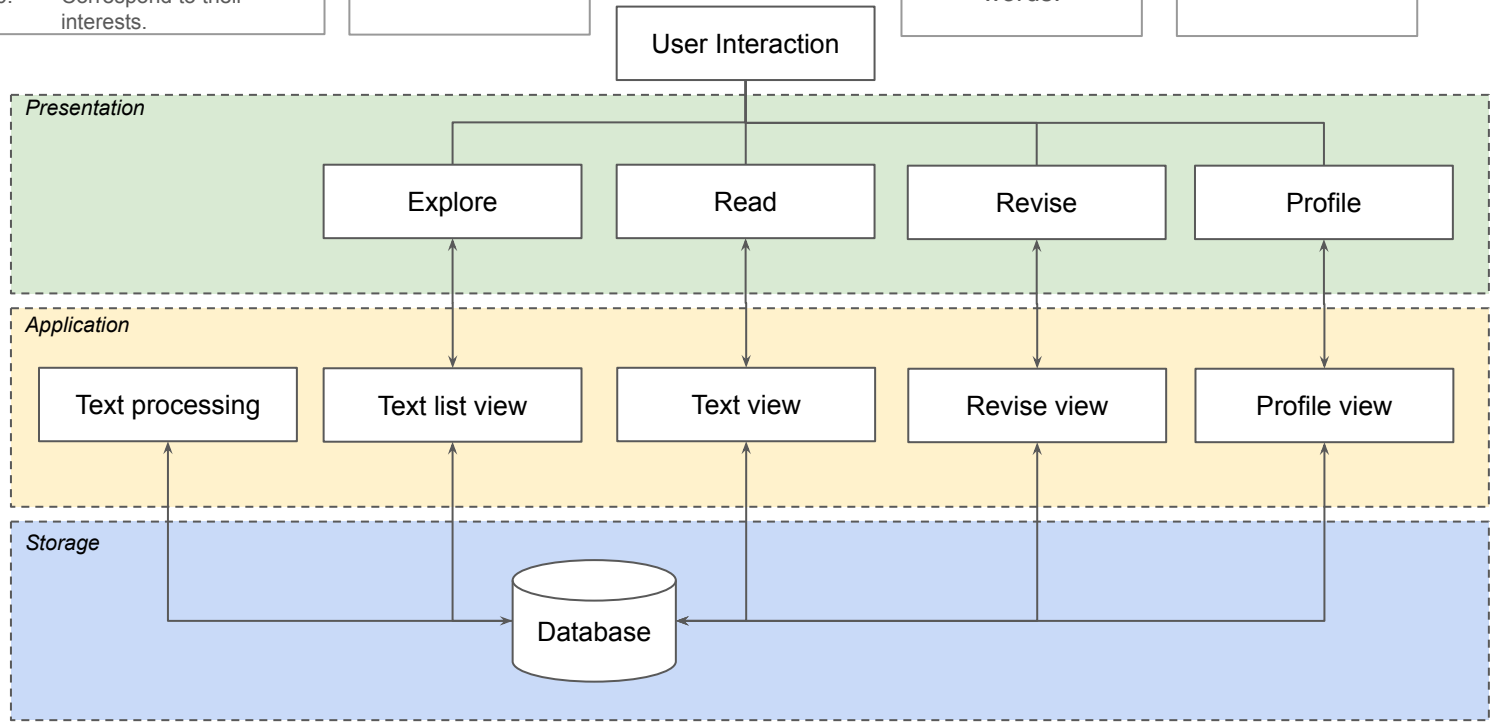
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

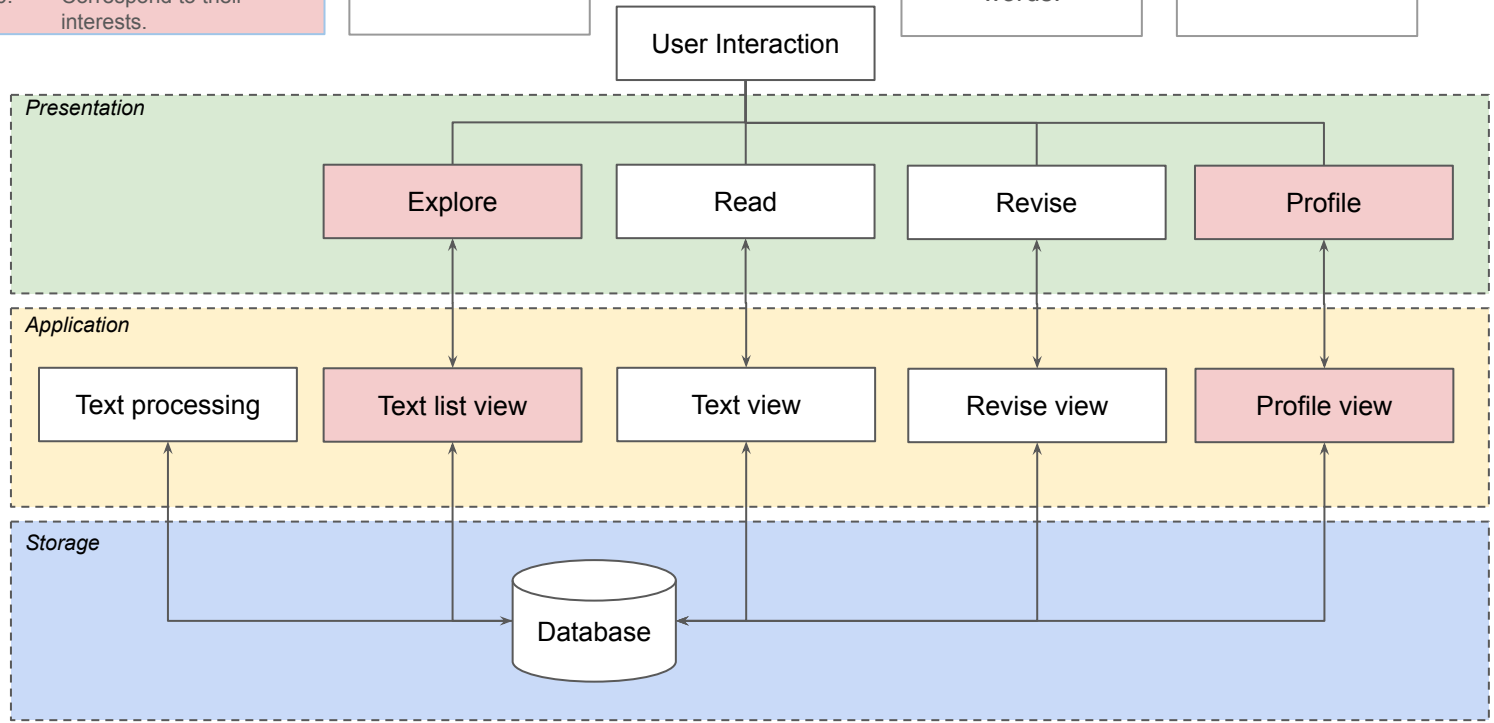
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

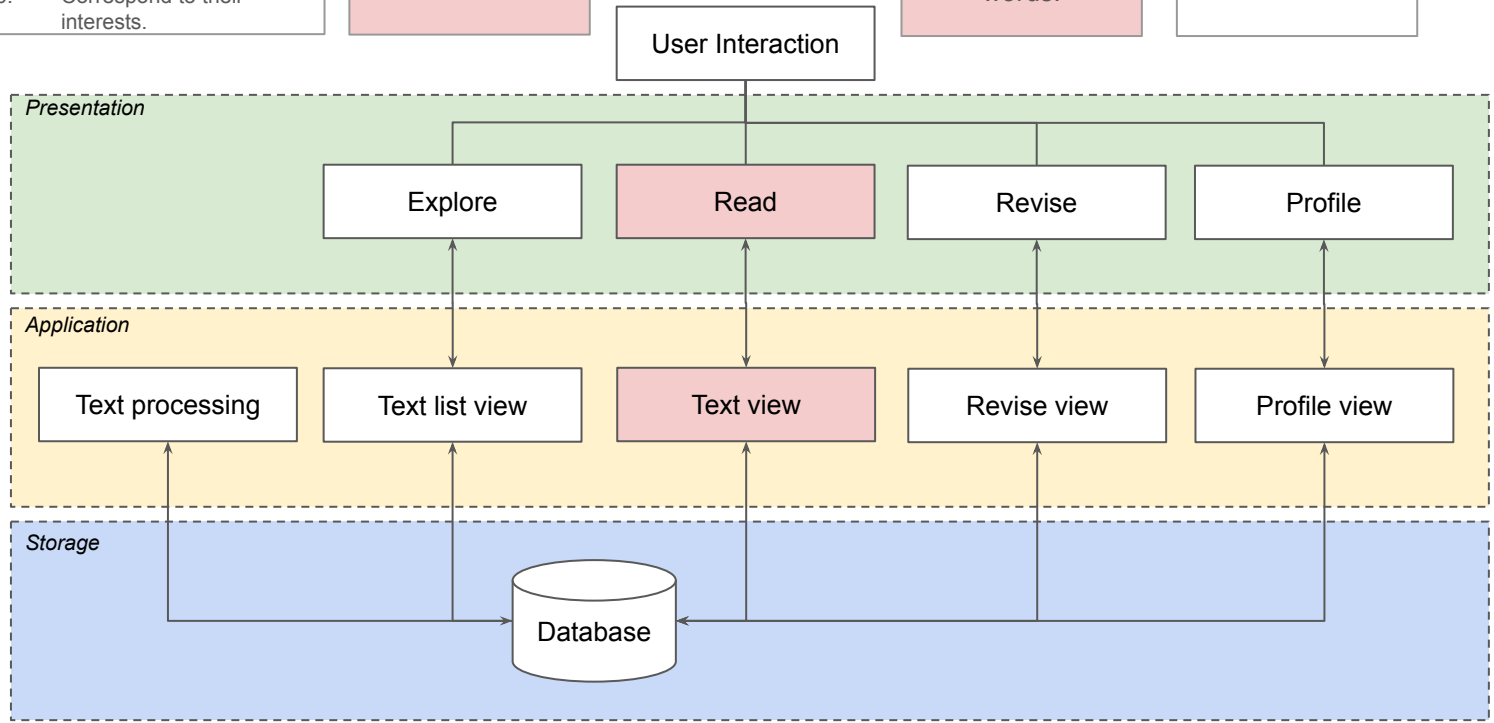
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

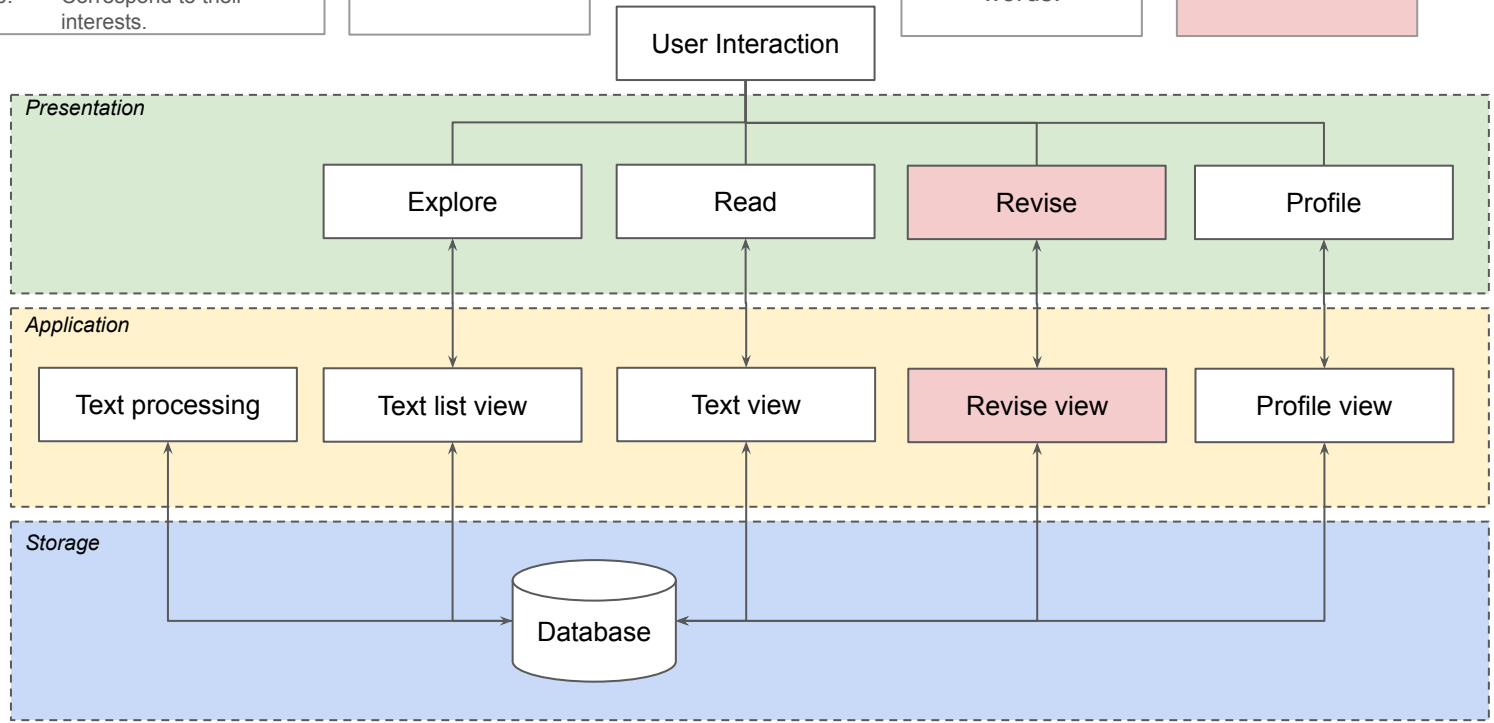
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

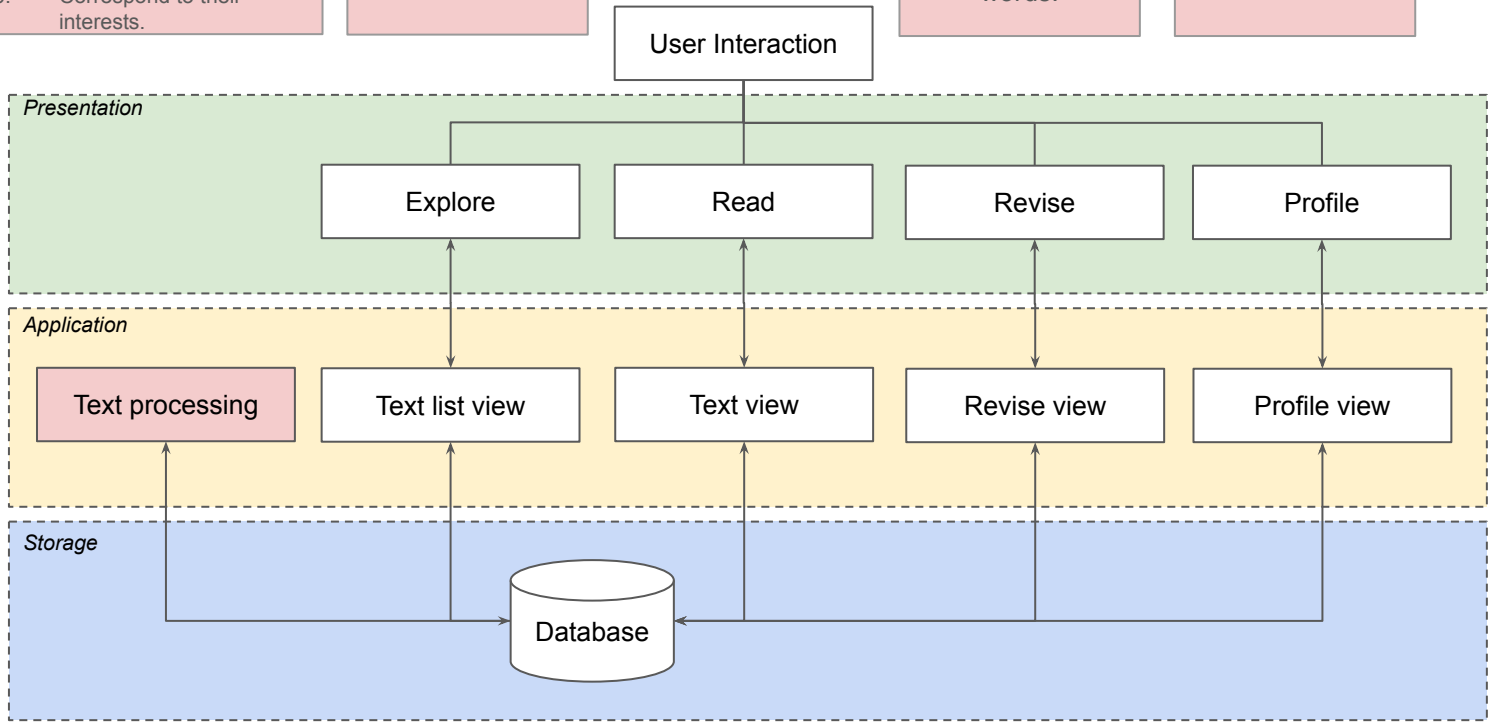
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

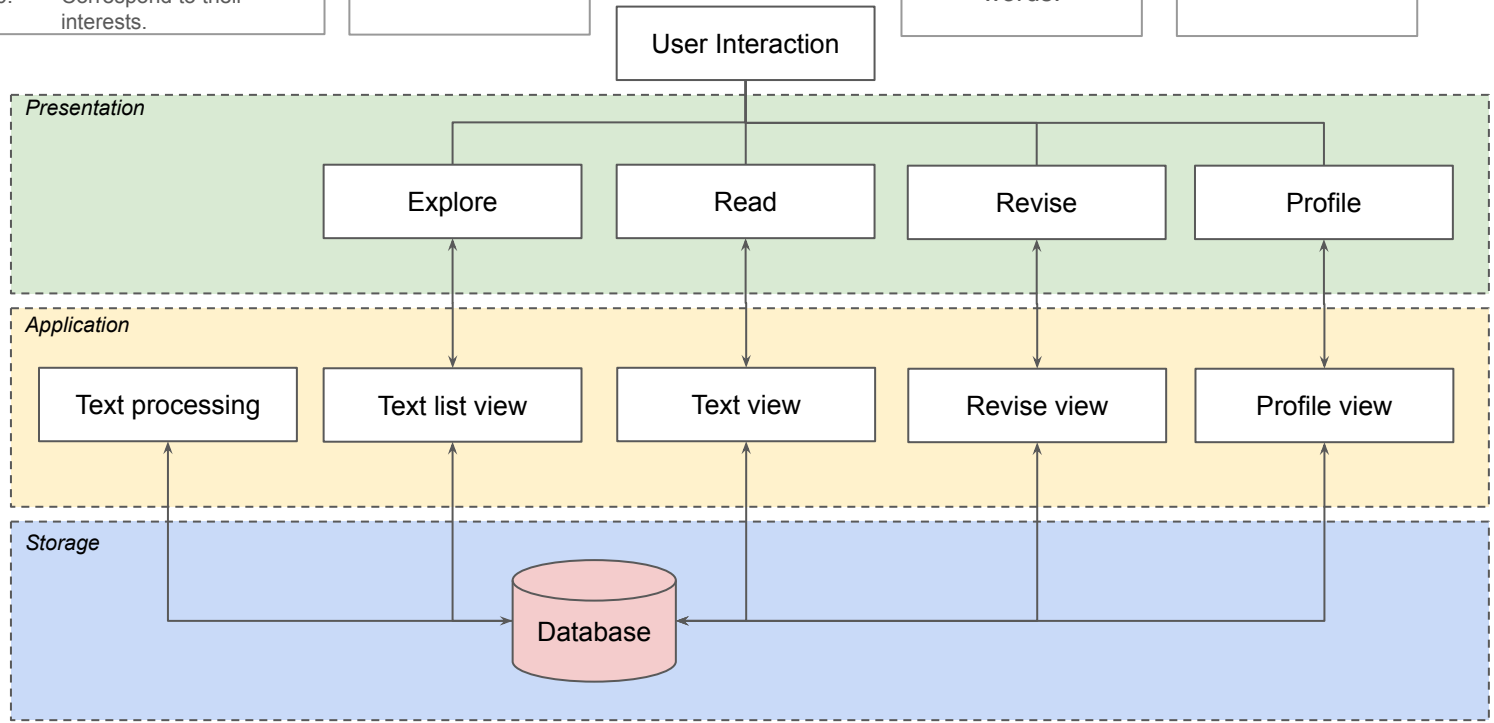
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# System Architecture

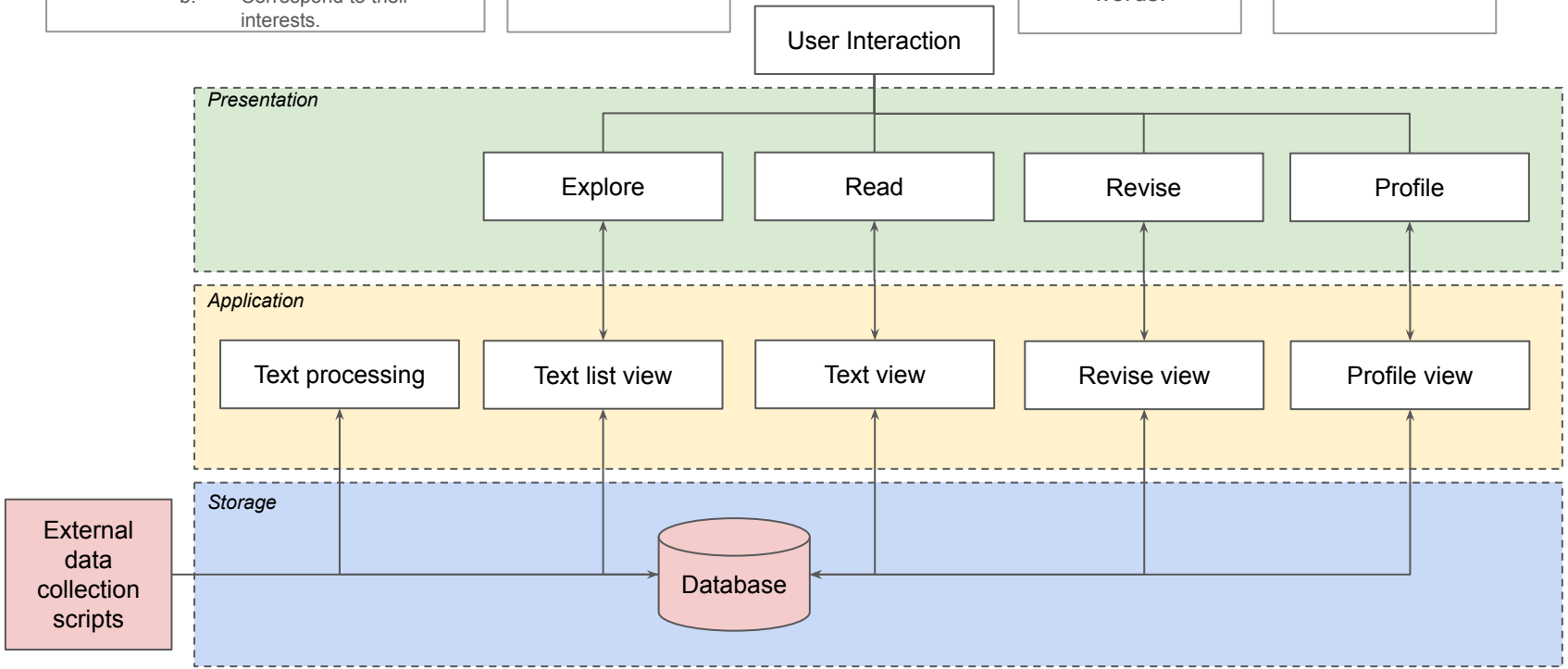
**FR1.** User should be able to read texts in their target language. These texts should:

a. Be appropriate for their language level.
b. Correspond to their interests.

**FR2.** User should be able to mark unknown words.

**FR3.** User should be able to get translations for unknown words.

**FR4.** User should be able to revise new vocabulary.



# Data

- Sources
  - Wikipedia
  - Hurraki (for Easy German)
  - Global Voices* (only for experiments)
- Each article from Wikipedia and Hurraki annotated with
  - Difficulty level (0 - easy, 1 - standard)
  - Topic
- Data available on GitHub

| Topic | Standard English | Easy English | Standard German | Easy German |
|---|---|---|---|---|
| Arts and Culture | 100 | 100 | 89 | 50 |
| Sports | 100 | 100 | 84 | 50 |
| Science and Technology | 100 | 100 | 91 | 50 |
| Travel | 100 | 100 | 87 | 50 |
| Food | 100 | 100 | 81 | 50 |
| Business, Economics, Politics | 100 | 100 | 87 | 50 |

[data available at https://github.com/siyanapavlova/thesis-dataset]

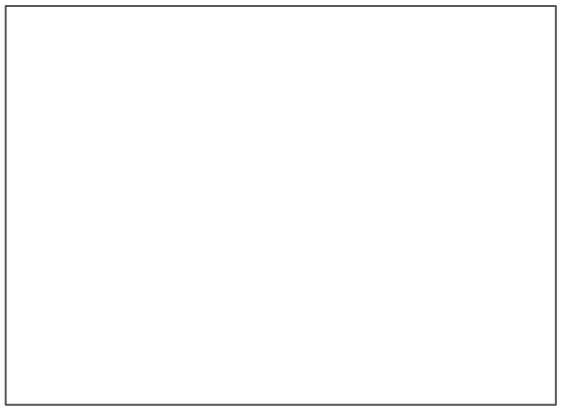
# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

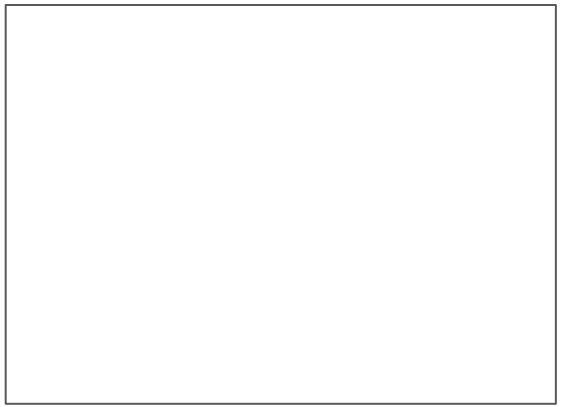
Combined



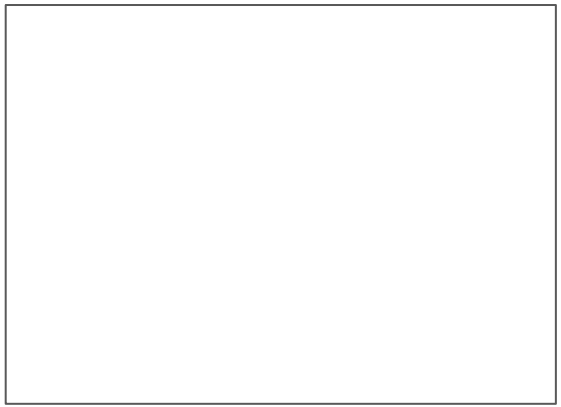
# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

Combined



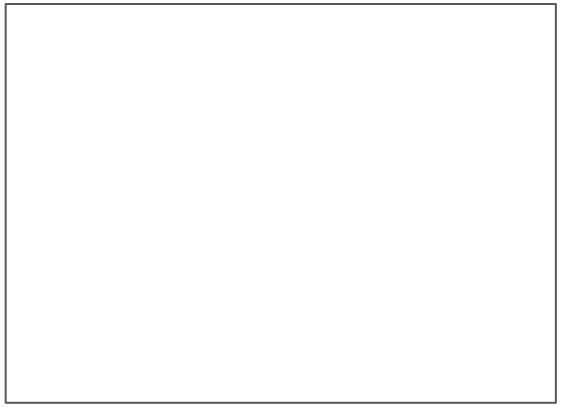
Global Voices



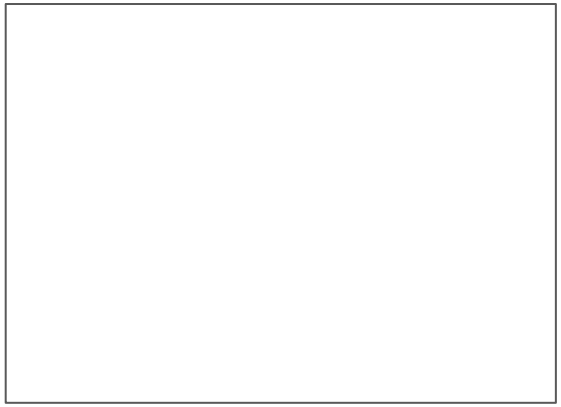
# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

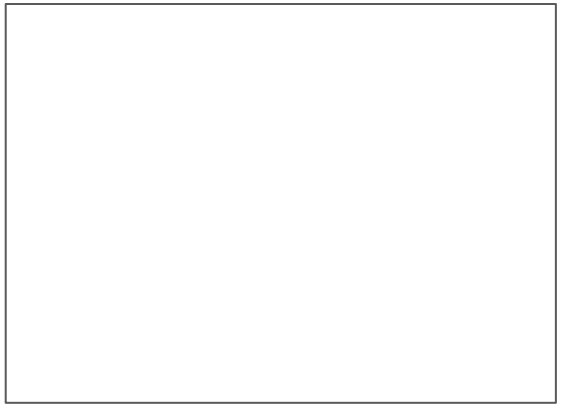
Combined



Global Voices



Wikipedia



# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

Combined

Accuracy: 0.542

Global Voices

Wikipedia

# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

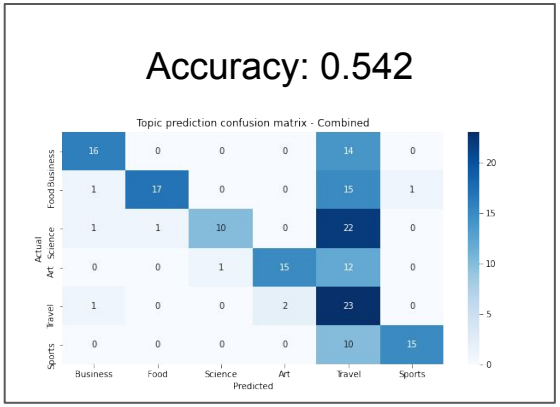
| Combined | Global Voices | Wikipedia |
| --- | --- | --- |
| Accuracy: 0.542 | Accuracy: 0.218 | Accuracy: 0.844 |

# Topic classification experiment

- 600 Standard English texts from Wikipedia, 6 topics
- 587 English articles from Global Voices, 6 topics
- TF-IDF vectorizer + SVM classifier
- 3 models; 85:15 train-test split for each

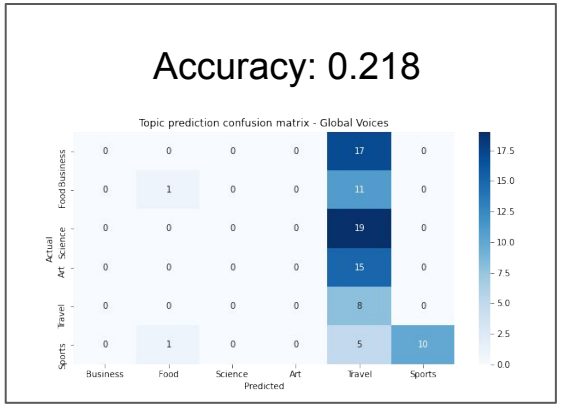
## Combined

Accuracy: 0.542



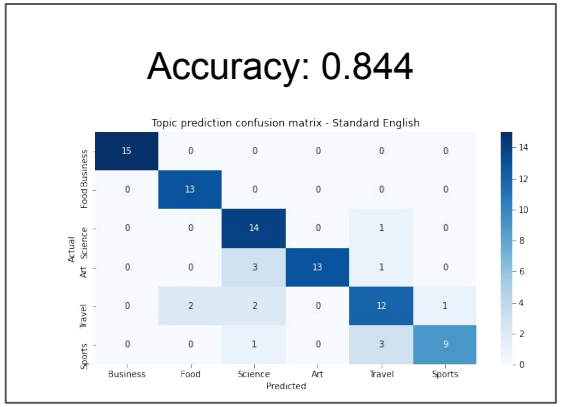

## Global Voices

Accuracy: 0.218




## Wikipedia

Accuracy: 0.844



# Demo

https://learn-with-lily.herokuapp.com/

# Demo System Details

- 819 texts in German (519 standard and 300 easy)
- Leitner system for spaced repetition
- Multiple choice quiz used for revision
  - Distractor answers generated from words already in the user’s vocabulary
- Relevance prioritised over difficulty
- Most frequent translation used to make translations
  - Panlex Lite corpus used as a dictionary

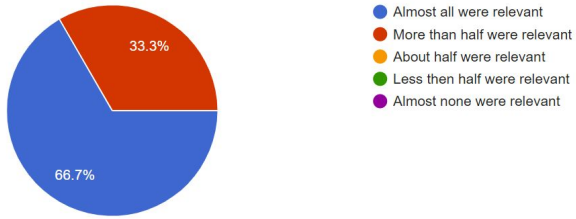
# User study

- Users asked to:
  - register and use the platform for three days
  - read texts and use the revision mode a few times each day
  - Complete a survey at the end
- Survey details:
  - 10 multiple choice and six free answer questions in four categories:
    - Use of other language learning platforms
    - Appropriate content
    - Ease of use
    - Overall
- Six participants
  - Master’s or PhD students or young professionals
  - Proficient or native English speakers
  - Speak at least one more language fluently

# User study results - Appropriate content

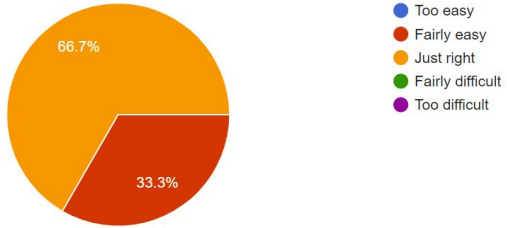
Overall, did the texts in "Explore" match the topics you selected as being interested in in your profile?

6 responses



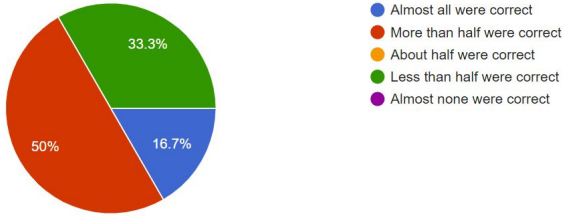

Overall, was the text difficulty appropriate for the language level you selected in your profile?

6 responses



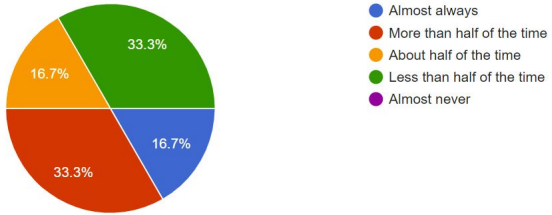

Overall, did you find that the translations offered were correct?

6 responses




How often would you say the translations offered helped you to understand the meaning of the text?

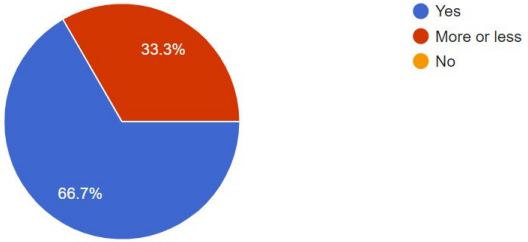
6 responses



# User study results - Ease of use

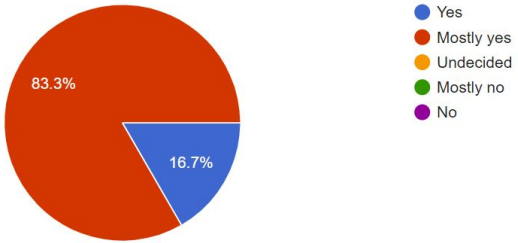
Would you say that the platform was untuitive to use?

6 responses




Did you enjoy using the platform?

6 responses



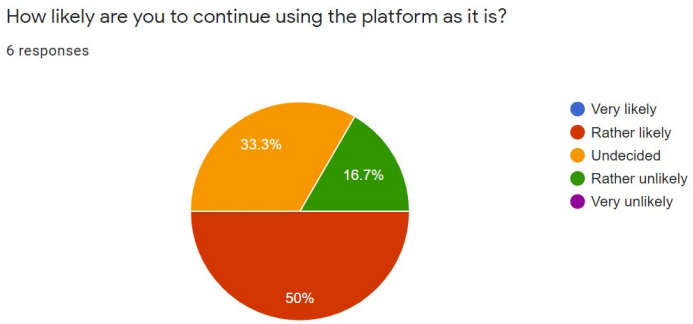

Comments and suggestions:

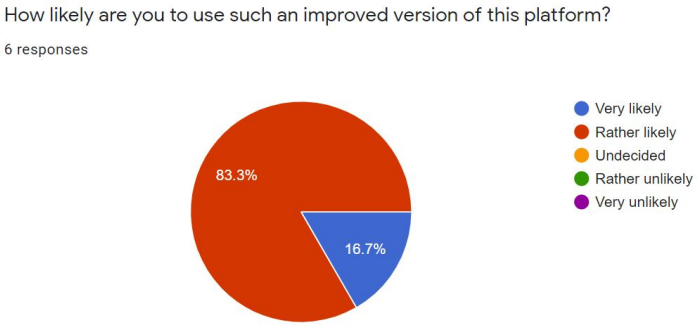
- “Done” button confusing
- Faster loading
- Possibility to enter notes
- Images displayed along texts

# User study results - Continued use

Users want to see:

- Walkthrough/guidelines
- User tailored texts
- Higher granularity of levels



# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

Learn from users

...

# Future Work

More and better content
- More topics
- More levels
- More languages
- Better translations

How to proceed with complete beginners?
- Vocabulary courses (Memrise-style)?
- “Nature method” texts?
- Both?

More revision modes

Allow users to edit vocabulary and suggest translations

Audio support (for words and for texts)

Pre-test

Gamification

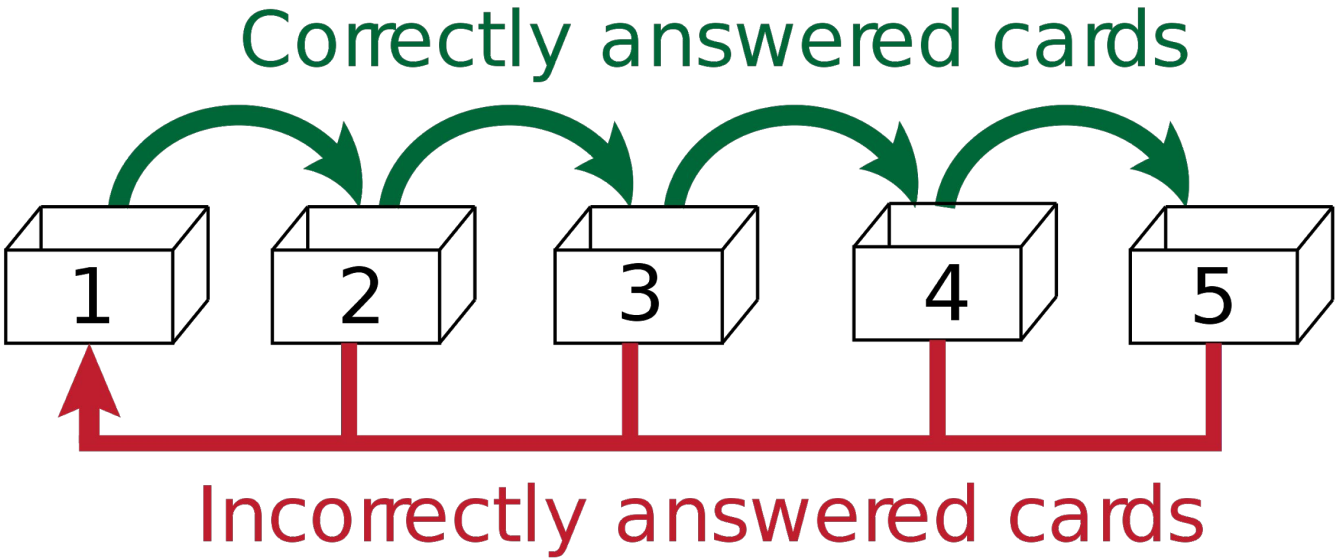
Learn from users

...

# Thank you for your attention!

Questions?

# The Leitner System (1972)




[Image source: https://en.wikipedia.org/wiki/Leitner_system#/media/File:Leitner_system_alternative.svg]