

An automatic annotation toolchain to recognize, quantify and visualize occurrences of CEFR-graded vocabulary and grammar patterns in English learner texts

Viet Phe Nguyen

Ye Yao

Ronja Laarmann-Quante

Torsten Zesch

Andrea Horbach

Stefan Keller

Use-Case


×

Upload xmi file

Drag and drop file here

Limit 200MB per file

Browse files

 ex1txt.xmi

11.6KB

×

Select Type:

A2 ×

B1 ×

A1 ×

⊕ ▾

C2 ×

Choose used theme:

☒ light

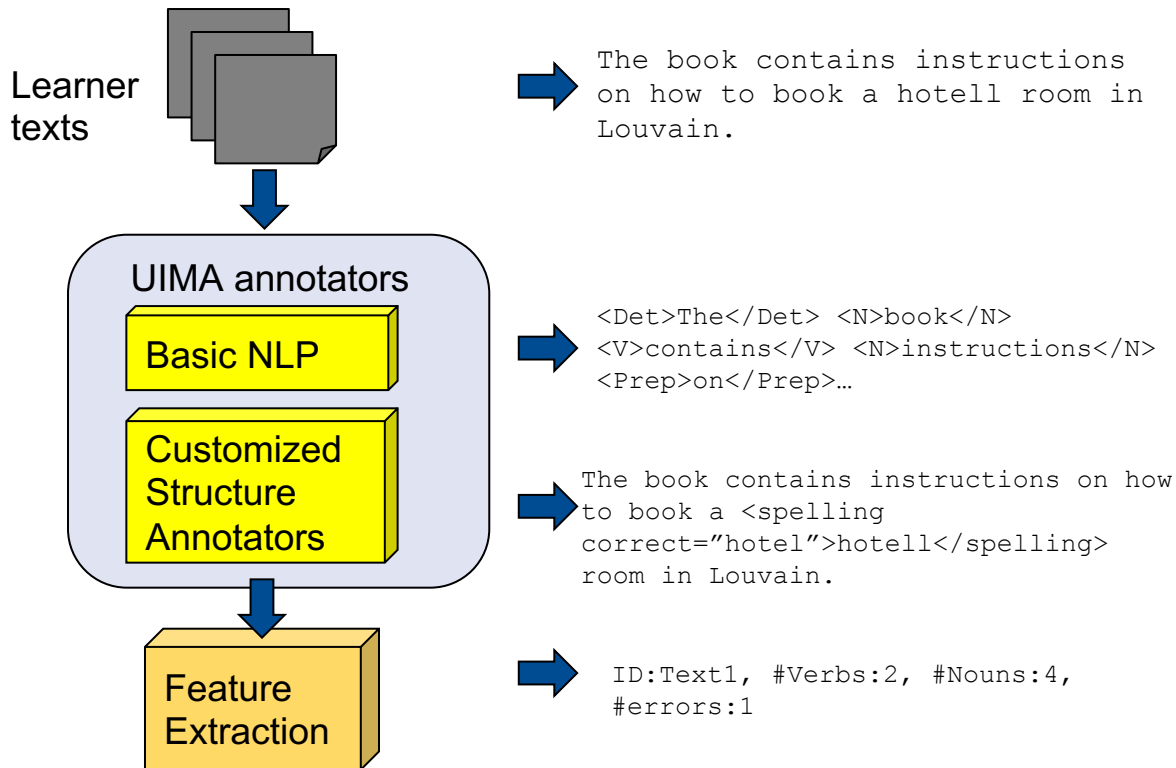
☐ dark

A1 B1 C2 A2

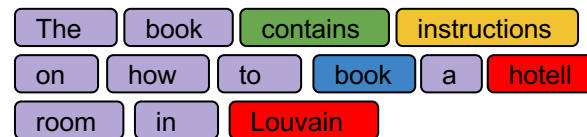
The book contains instructions on how to book a hotell room in Louvain
for a phantasmagorical monster.

Project TrACE - Training Assessment Competencies in English.

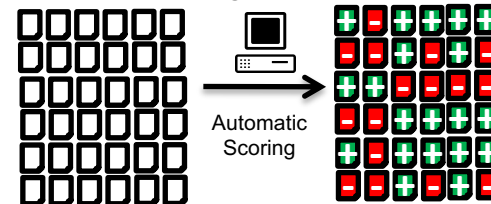
Broader Context: LiFT - Linguistic Features in Text



Assisted Scoring / Manual Inspection



Automatic Scoring / Machine Learning



Motivation

Complexity of vocabulary and grammar as an indicator for language complexity/learner proficiency (Vögelin et al. 2019):

It is very cold. vs. *The temperature is below freezing.*

simple



complex

Don't forget me! vs. *Don't you dare forget me!*

Part 1:

Annotating Vocabulary Information

Annotating Vocabulary Information

Task: Annotate lexical items with a frequency/complexity measure such as CEFR level

The book contains instructions on how to book a hotel

A1 **A1** **B1** **C2** **A1** **A1** **A1** **A2** **A1** **?**

room in Louvain for a phantasmagorical monster.

A1 **A1** **?** **A1** **A1** **?** **B1**

Annotating Vocabulary Information

Task: Annotate lexical items with a frequency/complexity measure such as CEFR level

The **book** contains instructions on how to **book** a hotell
A1 A1 B1 C2 A1 A1 A1 A2 A1 ?
room in Louvain for a phantasmagorical monster.
A1 A1 ? A1 A1 ? B1

Challenges:

- **Words with several word senses**
 - book - noun (A1) vs book - verb (A2): disambiguate via linguistic preprocessing (POS tagging)
 - book for reading (A1) vs book for writing (B1): use lower level by default: WSD difficult because of ill-defined word-sense definitions

Annotating Vocabulary Information

Task: Annotate lexical items with a frequency/complexity measure such as CEFR level

The book contains instructions on how to book a **hotell**

A1 A1 B1 C2 A1 A1 A1 A2 A1 ?

room in Louvain for a phantasmagorical monster.

A1 A1 ? A1 A1 ? B1

Challenges:

- **misspellings - *hotell*:**
 - per default no level assigned
 - use spellchecking during linguistic preprocessing

Annotating Vocabulary Information

Task: Annotate lexical items with a frequency/complexity measure such as CEFR level

The book contains instructions on how to book a hotel

A1 **A1** **B1** **C2** **A1** **A1** **A1** **A2** **A1** **?**

room in **Louvain** for a phantasmagorical monster.

A1 **A1** **?** **A1** **A1** **?** **B1**

Challenges:

- **Named Entities:** do not occur in word lists → no level

Annotating Vocabulary Information

Task: Annotate lexical items with a frequency/complexity measure such as CEFR level

The book contains instructions on how to book a hotel

A1 **A1** **B1** **C2** **A1** **A1** **A1** **A2** **A1** **?**

room in Louvain for a **phantasmagorical** monster.

A1 **A1** **?** **A1** **A1** **?** **B1**

Challenges:

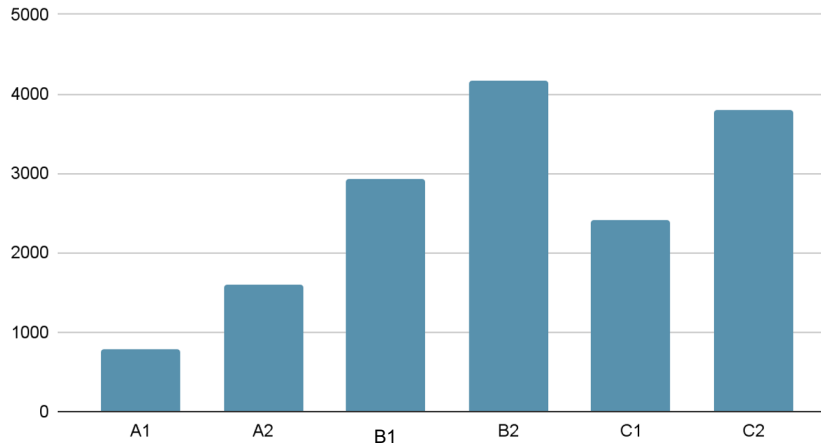
- other words not covered in CEFR:
 - typically rare words beyond learner English
 - if no level is assigned: hard to distinguish between misspellings, NEs, rare words

Overview

About 16000 lexical items (word - POS pairs) extracted from English Vocabulary Profile



Distribution across CEFR levels

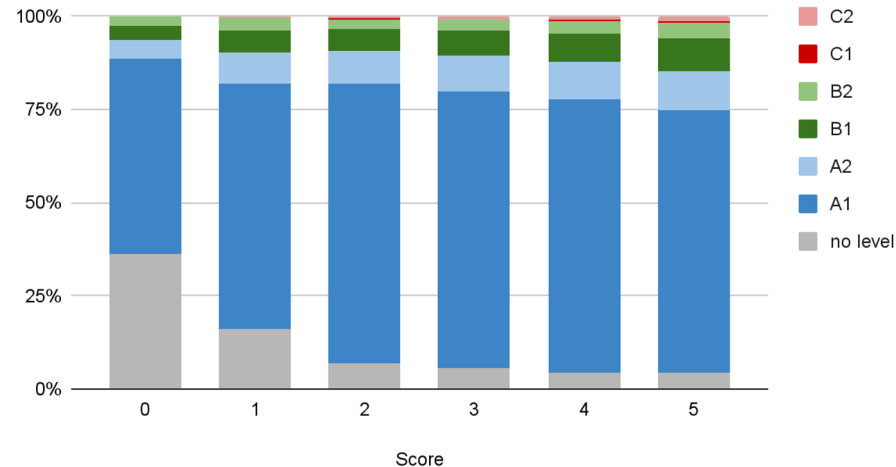


Evaluation on Essay Data

Distribution of CEFR rated vocabulary per essay score

- independent writing task for L2 English learners - MEWS dataset
- MEWS – Measuring English Writing at Secondary Level - A Binational Comparative Study.

Results on MEWS data



Part 2:

Grammar Patterns

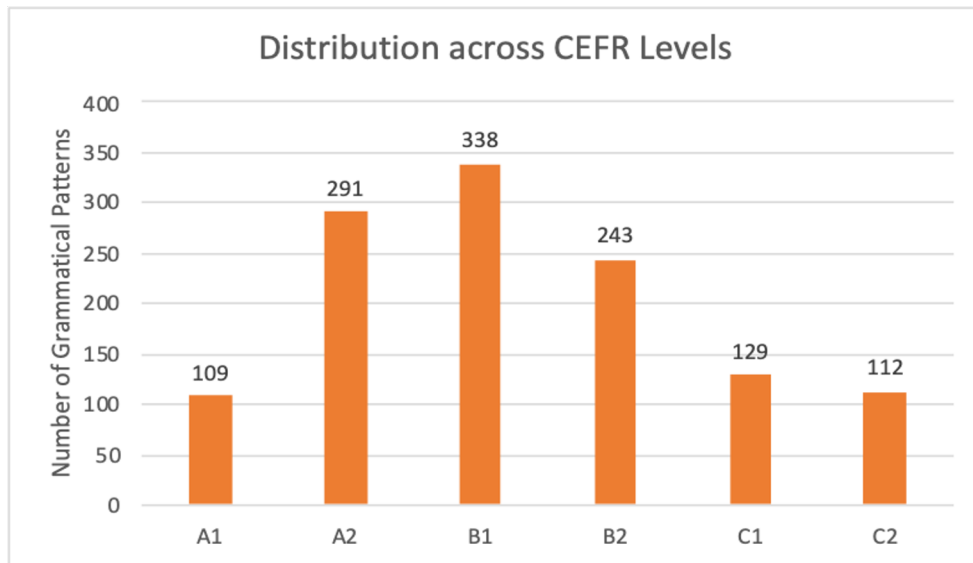
Examples for **Adjectives**, Subcategory **Superlatives**

Level	Can-Do Statement	Example
A1	MY BEST FRIEND Can use the irregular superlative adjective 'best' in the phrase 'my best friend'.	She's my best friend.
A2	WITH 'THE MOST' Can form superlative adjective phrases using 'the most', with longer adjectives of two or more syllables.	It is the most famous place in Edinburgh
B2	WITH 'BY FAR' Can use the premodifier 'by far' to make a superlative adjective stronger.	When I was a child, Christmas morning was by far the most exciting and happiest moment.
C1	WITH POSTMODIFIER AND NOUN Can use a postmodifier to make the superlative stronger, in the structure superlative + noun + postmodifier ('possible', 'ever', 'by far').	we want to present ourselves in the best way possible.

English Grammar Profile



- 1,222 Grammatical Structures
- 19 Categories, e.g. Adjectives, Nouns, Questions, Present, Reported Speech



Identifying Grammatical Structures in Text

Level	Can-Do Statement	Example
A1	MY BEST FRIEND Can use the irregular superlative adjective 'best' in the phrase 'my best friend'.	She's my best friend.

“(M|m)y best friend”

→

She's my best friend.

I meet my best friend every day.

My best friend is called Peter.

Identifying Grammatical Structures in Text

Level	Can-Do Statement	Example
C1	WITH POSTMODIFIER AND NOUN Can use a postmodifier to make the superlative stronger, in the structure superlative + noun + postmodifier ('possible', 'ever', 'by far').	we want to present ourselves in the best way possible.

[superlative] [noun] [postmodifier] → We want to present ourselves in the best way possible.
This is the coolest thing ever.
This was the most exciting day by far.

Implementation

UIMA Ruta (Kluegl et al. 2016)

- rule language
- grammatical structures can be identified via patterns

```
(  
W{REGEXP("(?i)my")}  
"best"  
"friend"  
)  
{-> CREATE(GrammarProfile, "name"="MyBestFriend", "level"="A1");
```



Create UIMA annotation of any type

Implementation

- Can be combined with **NLP preprocessing** (POS, chunks, ...) and **wordlists**
- We can structure the code in a linguistically meaningful and re-usable way

```
(  
  Superlative  
  POS_NOUN  
  Postmodifier  
)  
{-> CREATE(GrammarProfile,  
  "name"="SuperlativeNounPostmodifier",  
  "level"="C1")};
```

```
DECLARE Superlative;  
(POS{FEATURE("PosValue", "JJS")}) {-> MARK(Superlative)};  
("most" POS_ADJ) {-> MARK(Superlative)};  
("least" POS_ADJ) {-> MARK(Superlative)};
```

```
DECLARE Postmodifier  
WORDLIST PostmodifierList='postmodifiers.txt';  
Document{-> MARKFAST(Postmodifier, PostmodifierList)};
```

possible
by far
ever
...

Upcoming Challenges

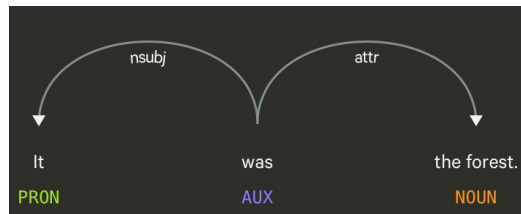
English Grammar Profile makes reference to **lexical range**, but is not aligned with the English Vocabulary Profile

Level	Can-Do Statement	English Vocabulary Profile
B1	COMPOUND ADJECTIVES Can use a limited range of compound adjectives ('good-looking', 'well-known')	good-looking: A2 well-known: A2
B2	COMPOUND ADJECTIVES Can use an increasing range of compound adjectives ('up-to-date', 'state-of-the-art')	up-to-date: B1 state-of-the-art: C1
C1	COMPOUND ADJECTIVES Can use a wide range of compound adjectives ('open-minded', 'above-mentioned', 'well-to-do', 'jaw-dropping')	open-minded: C1 above-mentioned: NA well-to-do: NA jaw-dropping: NA

Upcoming Challenges

Examples

- Linear patterns not always sufficient, e.g. to recognize questions



→include parsing

- “**Was** the forest nearby?” vs. “It **was** the forest.”
- Recognize particular **usages** of forms, e.g. present tense used as future
 - “The class **is** on Monday. It **starts** at 6:00 pm and **finishes** at 7:00 pm.”

Outlook

How can the annotations be used in practice?

- Assisted scoring experiments with teachers in training within TRACE project

How can others use it?

- Make available through LiFT toolkit

How can we extend it?

- Find/build resources for other languages

References

- Capel, Annette. "A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project." *English Profile Journal* 1 (2010).
- Capel, Annette. "Completing the English vocabulary profile: C1 and C2 vocabulary." *English Profile Journal* 3 (2012).
- Ferrucci, David, and Adam Lally. "UIMA: an architectural approach to unstructured information processing in the corporate research environment." *Natural Language Engineering* 10.3-4 (2004): 327-348.
- Hancke, Julia and Meurers, Detmar. "Exploring CEFR classification for German based on rich linguistic modeling". In Proceedings of the Learner Corpus Research (LCR) conference (2013).
- Kluegl, Peter, et al. "UIMA Ruta: Rapid development of rule-based information extraction applications." *Natural Language Engineering* 22.1 (2016): 1-40.
- Vögelin, C., Jansen, T., Keller, S., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39 (2019), 50-63.
<https://doi.org/10.1016/j.asw.2018.12.003>
- Zesch, Torsten, and Andrea Horbach. "ESCRITO-An NLP-Enhanced Educational Scoring Toolkit." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.