Toward constructing a corpus with CEFR-based sentence level annotations

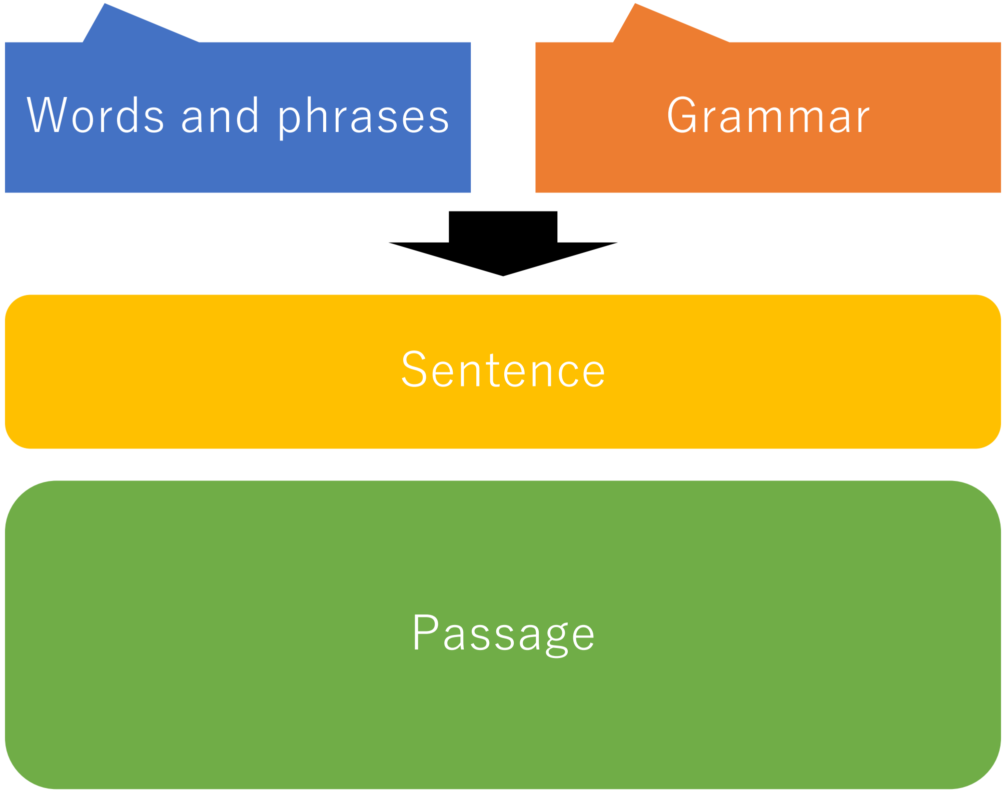
Satoru Uchida, Yuki Arase, Tomoyuki Kajiwara

December 7th, 2021

Building CEFR-graded resources for second and foreign language learning (GR4L2)

Introduction: CEFR

CEFR can-dos:
I can speculate about causes, consequences and hypothetical situations. （B2: Writing）

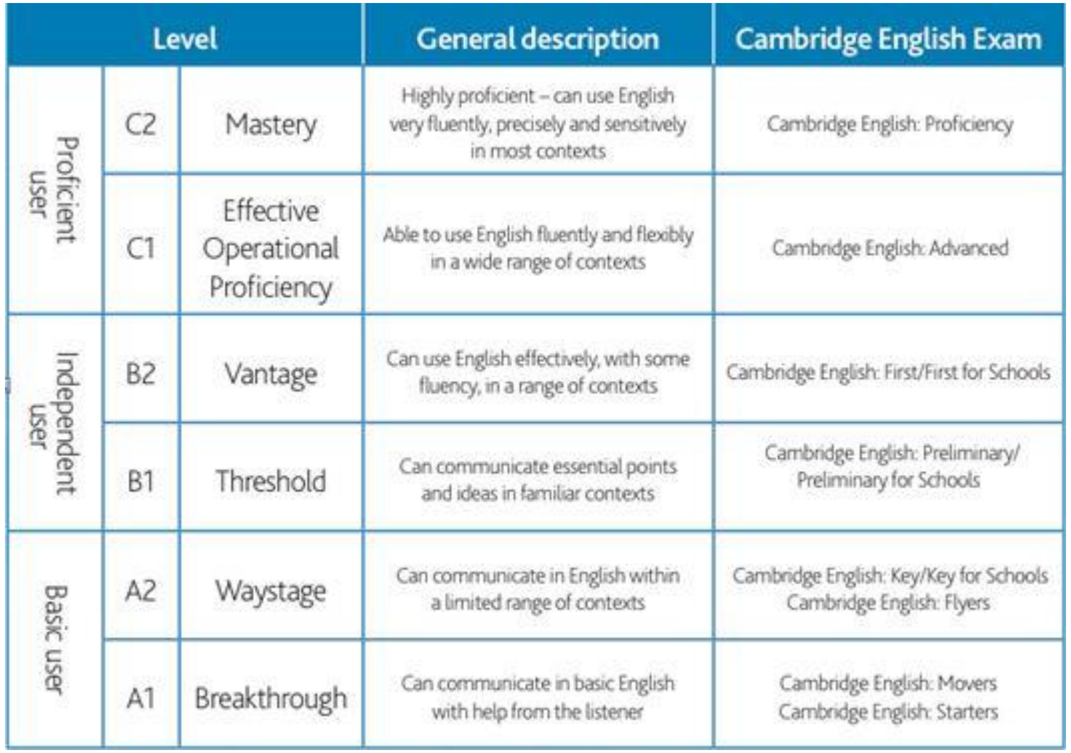


Words and phrases

Grammar

Sentence

Passage

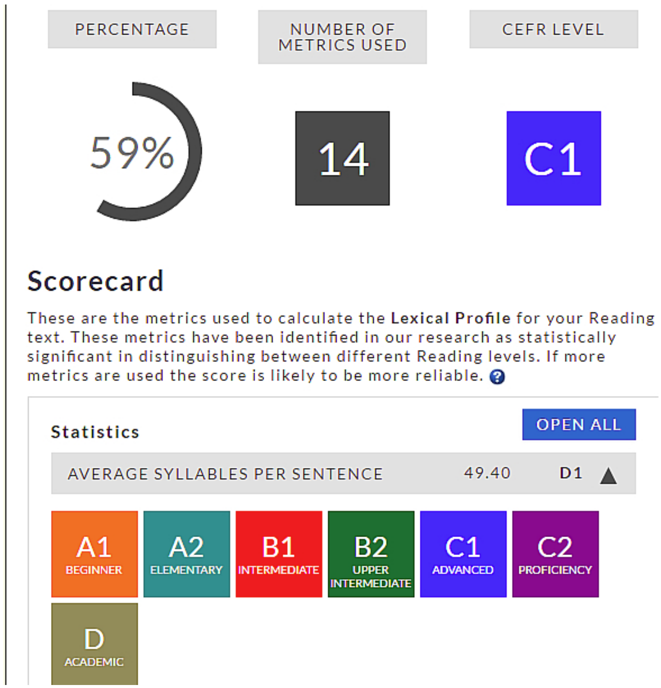


Level			General description	Cambridge English Exam
Proficient user	C2	Mastery	Highly proficient – can use English very fluently, precisely and sensitively in most contexts	Cambridge English: Proficiency
	C1	Effective Operational Proficiency	Able to use English fluently and flexibly in a wide range of contexts	Cambridge English: Advanced
Independent user	B2	Vantage	Can use English effectively, with some fluency, in a range of contexts	Cambridge English: First/First for Schools
	B1	Threshold	Can communicate essential points and ideas in familiar contexts	Cambridge English: Preliminary/ Preliminary for Schools
Basic user	A2	Waystage	Can communicate in English within a limited range of contexts	Cambridge English: Key/Key for Schools Cambridge English: Flyers
	A1	Breakthrough	Can communicate in basic English with help from the listener	Cambridge English: Movers Cambridge English: Starters

https://www.englishprofile.org/the-cefr/cefr-for-teachers-learners

Introduction: CEFR-level estimation

Text Inspector (Bax, 2012)

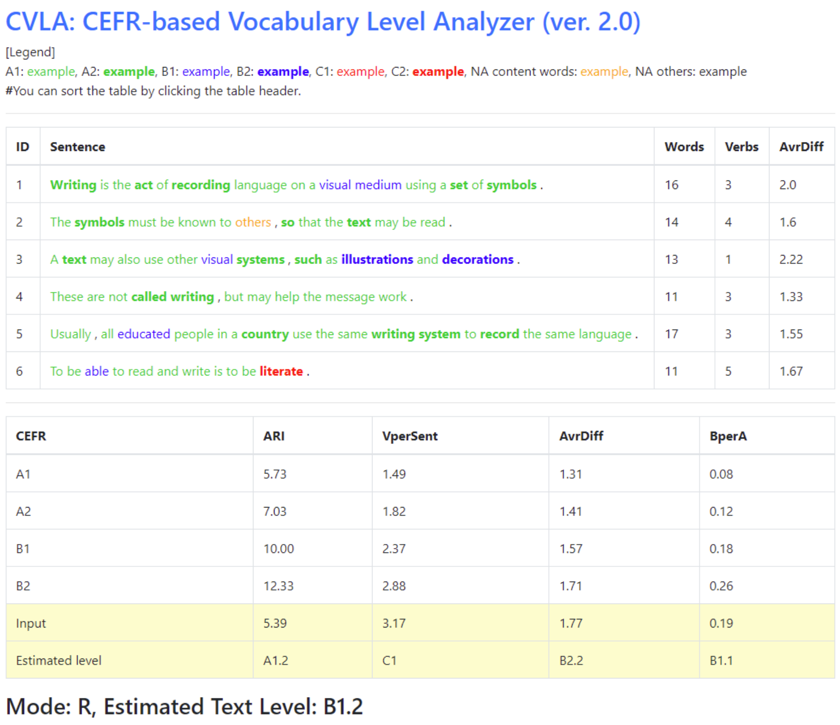


Scorecard

These are the metrics used to calculate the **Lexical Profile** for your Reading text. These metrics have been identified in our research as statistically significant in distinguishing between different Reading levels. If more metrics are used the score is likely to be more reliable.

AVERAGE SYLLABLES PER SENTENCE 49.40 D1

CVLA (Uchida ＆ Negishi, 2018)



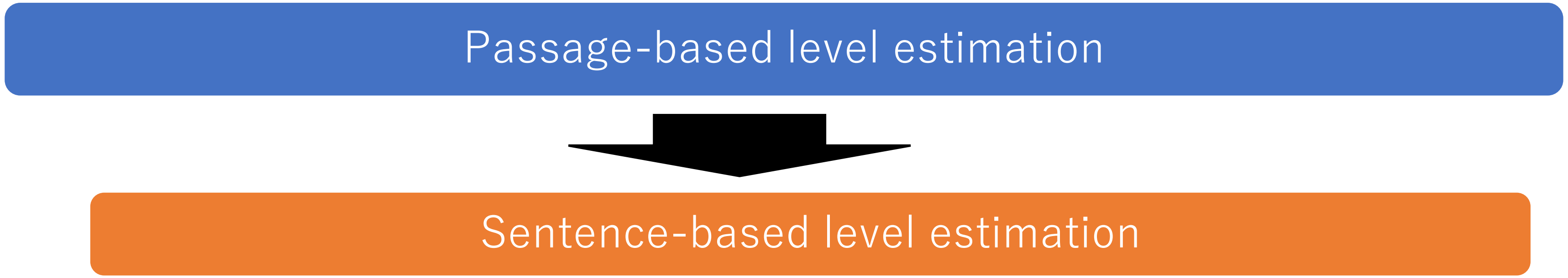
CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0)

[Legend]
A1: example, A2: **example**, B1: example, B2: **example**, C1: example, C2: **example**, NA content words: example, NA others: example
#You can sort the table by clicking the table header.

ID	Sentence	Words	Verbs	AvrDiff
1	**Writing** is the **act** of **recording** language on a visual medium using a **set** of **symbols** .	16	3	2.0
2	The **symbols** must be known to others , **so** that the **text** may be read .	14	4	1.6
3	A **text** may also use other visual **systems** , **such** as **illustrations** and **decorations** .	13	1	2.22
4	These are not **called writing** , but may help the message work .	11	3	1.33
5	Usually , all educated people in a **country** use the same **writing system** to **record** the same language .	17	3	1.55
6	To be able to read and write is to be **literate** .	11	5	1.67

CEFR	ARI	VperSent	AvrDiff	BperA
A1	5.73	1.49	1.31	0.08
A2	7.03	1.82	1.41	0.12
B1	10.00	2.37	1.57	0.18
B2	12.33	2.88	1.71	0.26
Input	5.39	3.17	1.77	0.19
Estimated level	A1.2	C1	B2.2	B1.1

Mode: R, Estimated Text Level: B1.2



Passage-based level estimation

Sentence-based level estimation

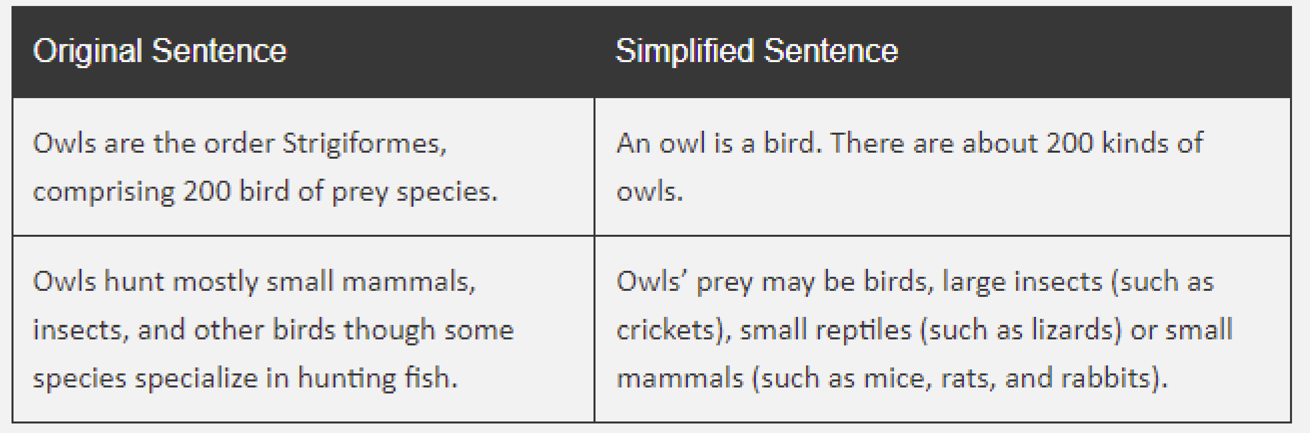
Why sentence level? (1)
Classroom reality

- Compared with paper, screens may also drain more of our mental resources while we are reading and make it a little harder to remember what we read when we are done. Whether they realize it or not, people often approach computers and tablets with a state of mind less conducive to learning than the one they bring to paper. And e-readers fail to re-create certain tactile experiences of reading on paper, the absence of which some find unsettling. (Authentic Reader, L1)

Many students understand the main idea of this passage, but some find this sentence particularly difficult.

Why sentence level? (2):
Useful for Simplification tasks

- Sentence simplification (cf. Alva-Manchego, Scarton, & Specia 2020)



Original Sentence	Simplified Sentence
Owls are the order Strigiformes, comprising 200 bird of prey species.	An owl is a bird. There are about 200 kinds of owls.
Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.	Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

http://nlpprogress.com/english/simplification.html

Lack of data that can be used for this kind of tasks especially for educational purposes

The purpose of this presentation

- To present necessity and challenges of sentence-based level annotation
- To show the overview of our dataset (under construction)
- To show the results of our preliminary experiment

Challenges for sentence level annotation(1)

- **Sentence level ≠ Vocabulary level**

【Grammar】

- The man closed the door. vs The door was closed by the man.

→Consist of mostly the same words but the sentence with **passive voice** is more advanced.

【Idiom】

- Don’t shoot the messenger.

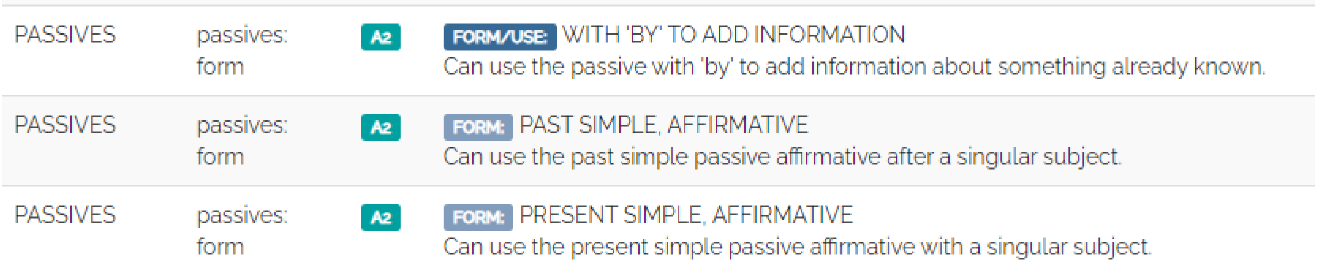
→ Words are easy but the meaning is idiomatic.

Challenges for sentence level annotation(2)

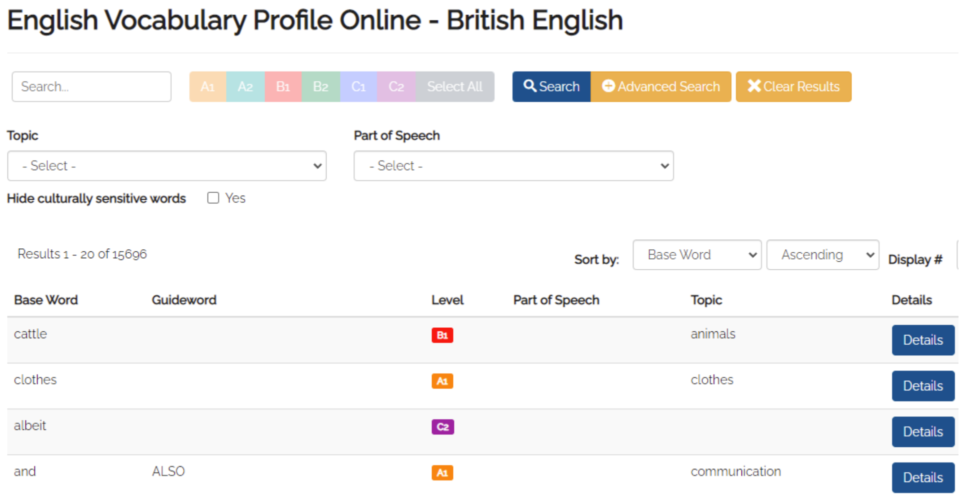
【Vocabulary and grammar and CEFR levels】

English Profile (Harrison, J & Barker (Eds.), 2015)

CEFR-J project (Negishi & Tono, 2014)



PASSIVES	passives: form	A2	FORM/USE: WITH 'BY' TO ADD INFORMATION Can use the passive with 'by' to add information about something already known.
PASSIVES	passives: form	A2	FORM: PAST SIMPLE, AFFIRMATIVE Can use the past simple passive affirmative after a singular subject.
PASSIVES	passives: form	A2	FORM: PRESENT SIMPLE, AFFIRMATIVE Can use the present simple passive affirmative with a singular subject.



English Vocabulary Profile Online - British English

Topic: - Select -

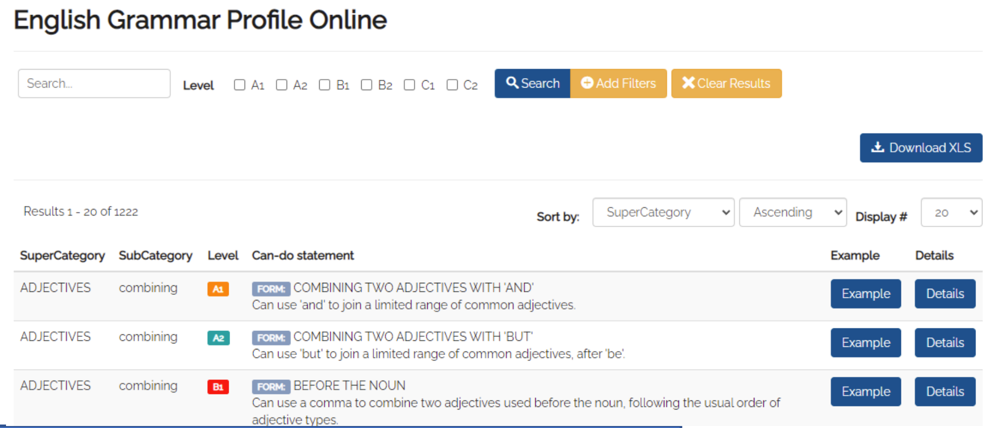
Part of Speech: - Select -

Hide culturally sensitive words □ Yes

Results 1 - 20 of 15696

Sort by: Base Word Ascending Display #

Base Word	Guideword	Level	Part of Speech	Topic	Details
cattle		B1		animals	Details
clothes		A1		clothes	Details
albeit		C2			Details
and	ALSO	A1		communication	Details



English Grammar Profile Online

Level □ A1 □ A2 □ B1 □ B2 □ C1 □ C2

Results 1 - 20 of 1222

Sort by: SuperCategory Ascending Display # 20

SuperCategory	SubCategory	Level	Can-do statement	Example	Details
ADJECTIVES	combining	A1	FORM: COMBINING TWO ADJECTIVES WITH 'AND' Can use 'and' to join a limited range of common adjectives.	Example	Details
ADJECTIVES	combining	A2	FORM: COMBINING TWO ADJECTIVES WITH 'BUT' Can use 'but' to join a limited range of common adjectives, after 'be'.	Example	Details
ADJECTIVES	combining	B1	FORM: BEFORE THE NOUN Can use a comma to combine two adjectives used before the noun, following the usual order of adjective types.	Example	Details

It is not simply a combination of the two to determine the sentence level, as topic and other factors are also involved.

Requirements for sentence level annotation

- Sentences should be **stand-alone**

Sentences with referential expressions are not suitable

These are called constructed languages also known as Oral Sects.

Sentences that require external knowledge are not suitable

Our approach: Data collection (1)

CEFR-based Sentence Level Annotation Dataset

- **Sources**

✓Newsela-auto (Jiang et al., 2020)
✓Wiki-auto (ibid.)
✓SCoRE (Sentence Corpus of Remedial English) (Chujo, Oghigian, & Akasegawa, 2015)
→**20,000** sentences in total (currently **5,000**)

- **Length**

✓5~30 words

Sentences with referential expressions are not suitable

- **Sampling**

✓The first sentences of each paragraph (the first paragraph was not used from Wiki)

- **Filtering**

✓Deleted sentences with punctuation marks such as “, [, (

Our approach: Data collection (2)

CEFR–based Sentence Level Annotation Dataset

- **Filtering using named-entity tags**

Using Stanza (https://stanfordnlp.github.io/stanza/);

Included expressions with that are marked as DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL

Included proper names that are in our whitelist (e.g. Japan, Japanese, English, American, Africa, Tokyo, John, Paul)

Sentences with other entity labels are excluded (EVENT, PERSON, ORG, WORK_OF_ART etc.)

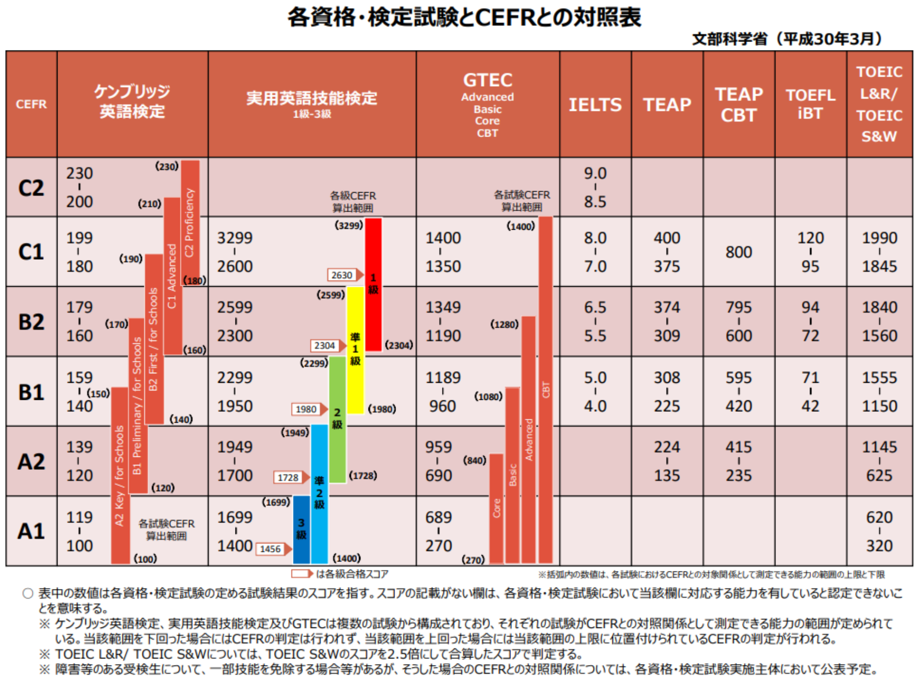
Sentences that require external knowledge are not suitable

Our approach: Annotation procedure

- 6 levels based on CEFR levels with sample sentences

ID	Sentence	CEFR
1	I want to see the cherry blossoms .	A1
2	You can sit with us .	A1
3	All of the children ate ice cream under the hot sun .	A1
4	If I were a king , I 'd make peace .	A2
5	I know you would like me to visit , but we ca n't afford the airfare this year .	A2
6	The move is part of a large change in education .	A2

- Japanese grade scale and the CEFR correspondence with English tests are provided for reference.
- Conducted some trial annotations to select two annotators with sufficient experiences in language education



各資格・検定試験とCEFRとの対照表

文部科学省（平成30年3月）

CEFR	ケンブリッジ英語検定	実用英語技能検定 1級-3級	GTEC Advanced Basic Core CBT	IELTS	TEAP	TEAP CBT	TOEFL iBT	TOEIC L&R/ TOEIC S&W
C2	230-200			9.0-8.5				
C1	199-180	3299-2600	1400-1350	8.0-7.0	400-375	800	120-95	1990-1845
B2	179-160	2599-2300	1349-1190	6.5-5.5	374-309	795-600	94-72	1840-1560
B1	159-140	2299-1950	1189-960	5.0-4.0	308-225	595-420	71-42	1555-1150
A2	139-120	1949-1700	959-690		224-135	415-235		1145-625
A1	119-100	1699-1400	689-270					620-320

は各級合格スコア

※括弧内の数値は、各試験におけるCEFRとの対象関係として測定できる能力の範囲の上限と下限

○ 表中の数値は各資格・検定試験の定める試験結果のスコアを指す。スコアの記載がない欄は、各資格・検定試験において当該欄に対応する能力を有していると認定できないことを意味する。

※ ケンブリッジ英語検定、実用英語技能検定及びGTECは複数の試験から構成されており、それぞれの試験がCEFRとの対照関係として測定できる能力の範囲が定められている。当該範囲を下回った場合にはCEFRの判定は行われず、当該範囲を上回った場合には当該範囲の上限に位置付けられているCEFRの判定が行われる。

※ TOEIC L&R/ TOEIC S&Wについては、TOEIC S&Wのスコアを2.5倍にして合算したスコアで判定する。

※ 障害等のある受検生について、一部技能を免除する場合等があるが、そうした場合のCEFRとの対照関係については、各資格・検定試験実施主体において公表予定。

Overview of our dataset (1)

Sentence	A	B
At some point the temple was forgotten and overgrown by jungle.	B1	A2
She and Dold say training for shows keeps captive whales and dolphins mentally and physically healthy.	B1	A1
Janet keeps her sewing room cluttered.	A2	A2
After an international uproar, and facing a suit by preservationists, a developer who planned a condo on the site sold the property to the state for $27 million.	C1	C1
Western isn't the first university to use college students to help younger kids.	A2	A2
The concept forces all sides in a disagreement to communicate and understand one other instead of resorting to violence.	B1	B1
In February 1982, two television antennas were added to the tower.	A2	A2
This study only looked at 42 people, a relatively small sample.	A2	A2
There were mice scratching in the walls.	A1	A1
Her daughter wants to become a personal trainer.	A1	A1
The province is divided into 6 districts and 12 municipalities.	A2	A2

Overview of our dataset (2)

		B						
		1(A1)	2(A2)	3(B1)	4(B2)	5(C1)	6(C2)	Total
A	1(A1)	**93**	169	29	2			293
	2(A2)	23	**419**	312	31	2		787
	3(B1)	8	403	**1236**	320	9		1976
	4(B2)	1	50	652	**610**	74		1387
	5(C1)		2	65	264	**115**	5	451
	6(C2)			2	40	59	**5**	106
	Total	125	1043	2296	1267	259	10	5000

Diff	Count	Cumulative ratio
0	2478	0.50
1	2281	0.95
2	232	0.99
3	9	1

r=0.68

#In the following analyses, sentences with more than 1 level difference between annotators are excluded. Those with one point difference will be treated as upper class (e.g. If the annotations are A1 and A2, then this sentence is treated as A2).

Overview of our dataset (3)

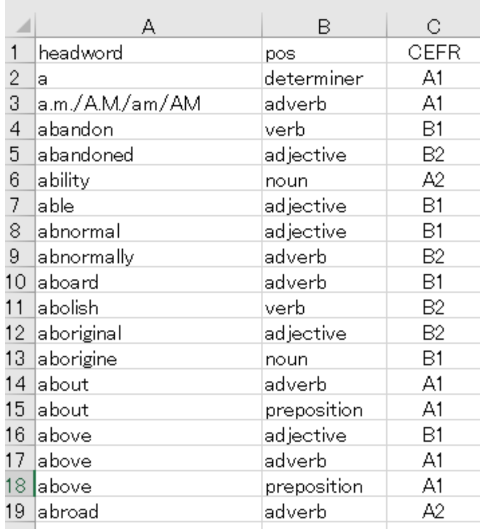
Level	# of examples	Length (# of words)	Dependecy distance (average)	Depth of constituency tree (average)
1 (A1)	93	8.0	2.4	6.5
2 (A2)	611	9.9	2.5	7.6
3 (B1)	1951	14.5	2.8	9.3
4 (B2)	1582	17.8	3.0	10.2
5 (C1)	453	18.8	3.0	10.6
6 (C2)	69	18.2	3.0	10.2

Overview of our dataset (4)

CEFR-J wordlist and EVP（for C1 and C2 levels）

CEFR-based Sentence Level Annotation Dataset

	A1	A2	B1	B2	C1	C2
A1	78.4%	15.9%	3.8%	1.9%	0.0%	0.0%
A2	65.9%	21.0%	10.2%	2.4%	0.5%	0.0%
B1	52.0%	23.7%	17.0%	6.3%	0.5%	0.5%
B2	40.0%	24.2%	23.0%	9.9%	1.4%	1.6%
C1	35.6%	23.4%	24.3%	12.5%	1.9%	2.2%
C2	31.4%	25.7%	29.3%	8.9%	2.7%	1.9%



	A	B	C
1	headword	pos	CEFR
2	a	determiner	A1
3	a.m./A.M./am/AM	adverb	A1
4	abandon	verb	B1
5	abandoned	adjective	B2
6	ability	noun	A2
7	able	adjective	B1
8	abnormal	adjective	B1
9	abnormally	adverb	B2
10	aboard	adverb	B1
11	abolish	verb	B2
12	aboriginal	adjective	B2
13	aborigine	noun	B1
14	about	adverb	A1
15	about	preposition	A1
16	above	adjective	B1
17	above	adverb	A1
18	above	preposition	A1
19	abroad	adverb	A2

Preliminary experiment

- Training data: 5,000 sentences

- Test data: 120 sentences with accurate annotation

- Model: BERT-base

- Task: 6 class classification

Results

Prediction

		A1	A2	B1	B2	C1	C2	Total	Recall
Annotation	A1	4	9	1				14	28.6%
	A2	1	8	16	1			26	30.8%
	B1		3	23	8	1		35	65.7%
	B2		2	10	14			26	53.8%
	C1			1	9	3		13	23.1%
	C2				1		5	6	83.3%
Total		5	22	51	33	4	5	120	
Precision		80.0%	36.4%	45.1%	42.4%	75.0%	0.0%		
Precision (±1)		100.0%	90.9%	96.1%	93.9%	75.0%	0.0%		

A1~B1 tend to be higher

B2-C2 tend to be lower

# (±1): cases when diff±1 are treated as correct

Accuracy: 47.5% (57/120)
Accuracy (±1): 94.2% (113/120)

The estimations are not significantly out of line.
However, it is a challenge to distinguish neighboring levels.

Summary

- CEFR–based Sentence Level Annotation Dataset

-> 20,000 sentences with CEFR levels (currently 5,000).
->Stand-alone sentences are selected.
->Annotated by two experienced language educators.

- **Preliminary experiments reveals:**

->BERT-base model is good at “rough” estimation.
->However, it is not good at distinguishing neighboring levels.

- **The updated version (20,000 sentences)**

-> Will be released in the near future.

References

- Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135-187.
- Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial EFL learners. *Multiple affordances of language corpora for data-driven learning*, 109-130.
- Bax, S. (2012). Text Inspector: Online text analysis tool. Available at: https://textinspector.com/.
- Harrison, J., & Barker, F. (Eds.). (2015). *English profile in practice* (Vol. 5). Cambridge University Press.
- Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020). Neural CRF Model for Sentence Alignment in Text Simplification. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Negishi, M., & Tono, Y. (2014). An update on the CEFR-J project and its impact on English language education in Japan. In *5th International Conference of the Association of Language Testers in Europe (ALTE), Paris, France, April* (pp. 10-11).
- Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Tono, Y. & Isahara, H. (eds.) *Proceedings of Asia Pacific Corpus Linguistics Conference, 4*, 463-467.

Thank you very much for your attention!