



Using corpora for reference level descriptions of the CEFR and the CEFR-J

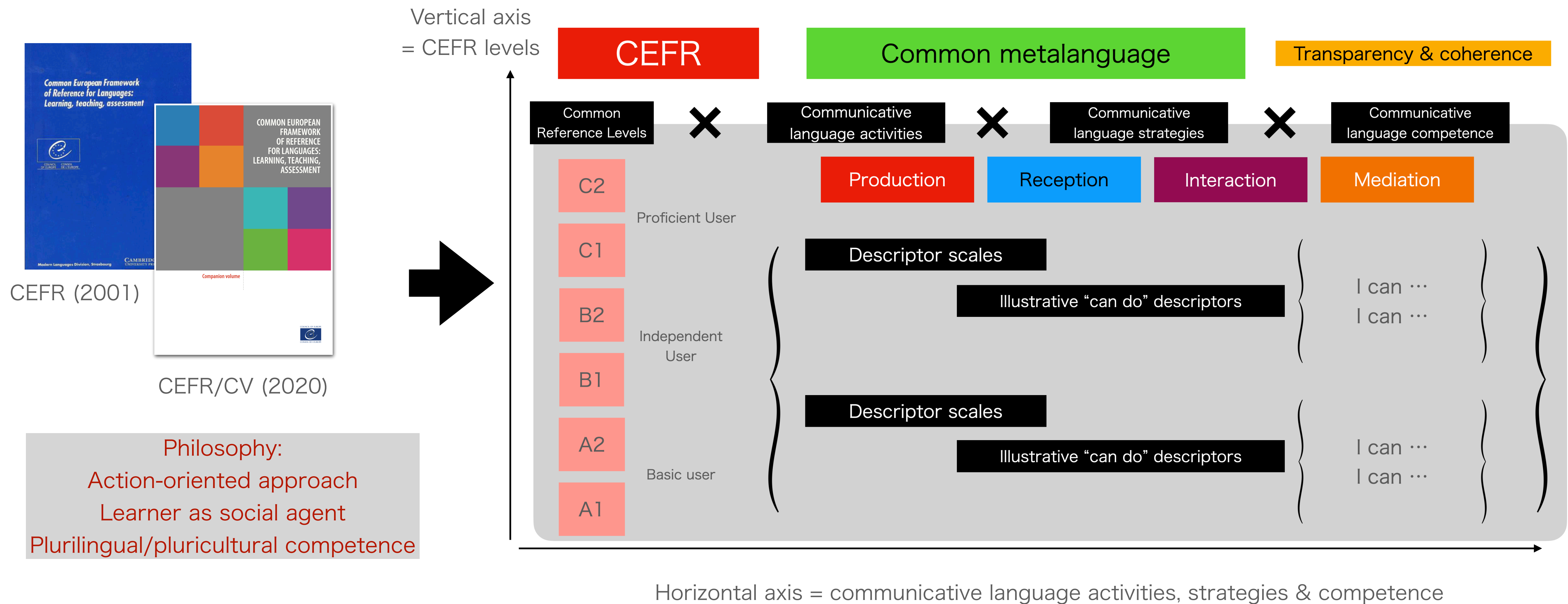
Online seminar series: International Perspectives on Corpus Technology for Language Learning

Yukio Tono, Tokyo University of Foreign Studies

30 September via Zoom



CEFR as a descriptive tool



What is “Reference Level Descriptions”?

- “The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) is potentially applicable to all the languages taught in Europe and does not, therefore, relate to any specific one. However, authors of textbooks, syllabus designers and language teachers have found its specifications to be insufficiently precise. Reference Level Descriptions (RLDs) language by language have therefore been drawn up to provide reference descriptions based on the CEFR for individual languages.”
- “These RLDs are made up of “words” of a language rather than general descriptors. Reference levels identify the forms of a given language (words, grammar and so on), mastery of which corresponds to the competences defined by the CEFR. They transpose the CEFR descriptors into specific languages, level by level, from A1 to C2.”

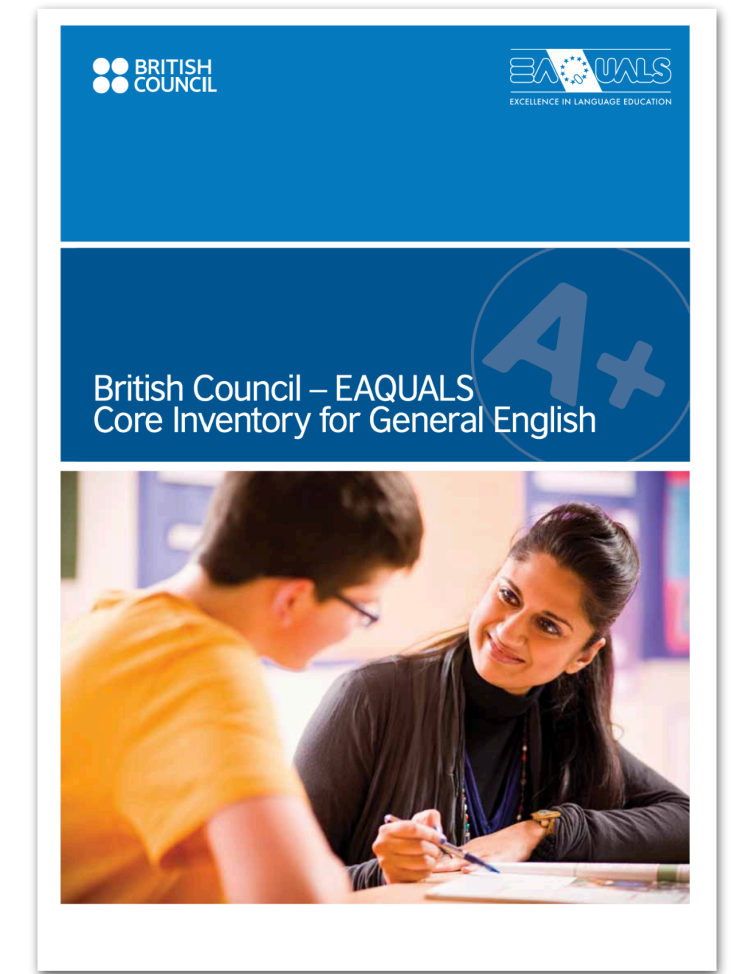
Methods of RLDs

- There are various ways of identifying the profiles, i.e. the forms of a language in terms of vocabulary, grammar, and text for each CEFR level.
- **Intuitive method:** (e.g.) Language specialists produce the language profiles largely using their insight as expert users and teachers of the language.
- **Qualitative method:** (e.g.) Profiles can be produced based on the surveys of available CEFR-based course books and expert judgements about what to include for each CEFR level.
- **Quantitative method:** (e.g.) Counting the occurrences of words and grammar items in CEFR level-classified texts and identifying the points where they are first introduced or consistently used.

RLD projects for English

- **British Council/EAQUALS Core Inventory for General English**
 - Course book analysis and expert judgements — “textbook input”
- **English Profile** (<http://www.englishprofile.org/>)
 - Cambridge Learner Corpus — profiling based on “learner output”
- **Global Scale of English** by Pearson
 - Rasch analysis of expert language teachers’ judgements of learning points — calibrated “teachers’ perceptions”

Each project has its own method of RLDs, which might produce different results.
Also, we cannot replicate the study because they do not make the data publicly available.



CEFR-J: Corpus approaches to RLDs

Replicate the construction of the CEFR & explore corpus-based RLDs

- Phase I (2008 - 2011): **Developing & scaling the CEFR-J descriptors**
 - Creating can-do descriptors for:
 - 10 sub-levels: [Pre-A1/A1.1/I.2/I.3/A2.1/2.2/B1.1/I.2/B2.1/2.2]
 - 5 modes of communication: [listening/ reading/ spoken interaction/ spoken production/ writing]
 - Scaling descriptors using Item Response Theory (IRT) 2-parameter model
 - 5,468 students (lower secondary: 1,685, upper secondary: 2,538, university: 1,245)
- Publication of the **CEFR-J Descriptors** (March, 2011)



<https://www.cefr-j.org>

CEFR-J: Corpus approaches to RLDs

Replicate the construction of the CEFR & explore corpus-based RLDs

- Phase II (2012 - 2015): RLD part 1
 - **The CEFR-J Wordlist**
 - **The CEFR-J Grammar Profile**
 - **The CEFR-J Text Profile**
- English companion website (mainly for data download):
<http://www.tufs.ac.jp/ts/personal/corpuskun/wiki/index.php?CEFR-J%20RLD>

The CEFR-J Wordlist & Collocation Dataset

- **The CEFR-J Wordlist Version 1.6 (2016-2022):**
 - Based on the ELT textbook analysis in Asian regions/countries (China, Korea, Taiwan, where English was taught at primary school.)
 - Compared the results against EVP by English Profile and merged
- **The CEFR-J Collocation dataset (released in September 2022):**
 - Extracted collocation frames from BNC syntactically parsed by Stanza (<https://stanfordnlp.github.io/stanza/>):
 - *amod* (adj + noun) / *nounmod* (noun + noun) / *obj* (verb + noun) / *advmod* adj (adv + adj) / *advmod* verb (adv + verb)
 - Each collocation pair has the information about dispersion (DP) and association measures (MI/ MI₂/ MI₃/ t_score/ z_score/ logDice/ log_likelihood/ chi_squared)
 - ADJ+NOUN: 135,939 pairs / VERB+NOUN: 114,582 pairs / NOUN+NOUN: 72,340 pairs
 - ADVERB+VERB: 43,992 pairs /ADVERB+ADJ: 16,180 pairs

Profiling based on both INPUT and OUTPUT corpora

Profiling method	Corpus	Description	Size
INPUT	• CEFR Course Book Corpus	• 96 CEFR-based course books published in the UK, with CEFR level classifications	1,801,549
	• Corpus of English textbooks published in Japan	• Government-authorised secondary school textbooks	1,158,525
OUTPUT	• JEFL-CEFR Corpus	• written learner corpus; 10,038 student essays; CEFR-level classified version • Also proofread version available	669,281
	• NICTJLE-CEFR Corpus	• spoken learner corpus; 1,281 interview transcripts; CEFR-level classified version	763,289 (Interviewee's parts only)

CEFR-J Grammar Profile

- A list of English grammar items taught at primary & secondary schools
 - 263 items + sentence patterns (interrogative, negative)
 - REGEX pattern queries for all the grammatical items (by Yasutake Ishii)
- Frequency data was obtained from the following corpora:
 - INPUT: CEFR course book / English textbook in Japan
 - OUTPUT: JEFLL (written) / NICTJLE (spoken)
- Determining the CEFR level by the RANGE and FREQ of the items
- Determining the criterial features using machine learning
 - See Tono (2015) for further detail

CEFR Level Checker

<https://lr-www.pi.titech.ac.jp/gradesystem/>

English Level Checker ~ 英文レベル判定 ~

Textbook

Essay

In this talk, I will report on the CEFR-J project and how corpora are used for selecting criterial language features for characterizing CEFR levels. The CEFR-J project aims to localise the CEFR in the context of English language teaching in Japan. We replicated the construction and scaling of illustrative descriptors, in the same way as the original CEFR, and published a set of 100 validated Can Do descriptors as the CEFR-J Can Do list (Tono, 2012). These CEFR-J descriptors were used as one of the references in revising the CEFR itself in the Companion Volume (2020). The project went on to conduct so-called “Reference Level Descriptions (RLDs),” whose purpose was to identify lexical, grammatical and textual features representing each of the CEFR-J levels. We aimed to develop a valid method of profiling CEFR levels using both coursebook corpora as input and learner corpora as output. Currently, we are working with a group of teachers at primary and secondary schools to use the CEFR-J resources to teach English and examine how learning takes place in the learning environment supported by the CEFR-J descriptors and their accompanying resources. At the same time, all the related textual data such as the textbooks used in the school, the classroom discourse, students’ group and pair work, students’ final spoken or written proficiency tests are being made into corpus data in order to closely examine the changes taking place as the support and intervention using the CEFR-J resources will fundamentally change teachers’ perspectives of teaching as well as actual students’ outcomes.

255/10 ✓

submit

clear

本Webアプリの利用者は、本Webアプリにデータを送信した時点で、[免責事項](#)を 了解したものとします。

→

English Level Checker ~ 英文レベル判定 ~

Textbook

Essay

In this talk, I will report on the CEFR-J project and how corpora are used for selecting criterial language features for characterizing CEFR levels. The CEFR-J project aims to localise the CEFR in the context of English language teaching in Japan. We replicated the construction and scaling of illustrative descriptors, in the same way as the original CEFR,

255/10 ✓

submit

clear

文数	単語数	使用文法項目数
8	255	31

使用文法項目	頻度
14 定冠詞	22
105 動詞-ing形	12
150 等位接続詞	11
88 to不定詞(to DO)	6
169 名詞を後置修飾する現在分詞	4
13 不定冠詞	3
58 時制・相(現在)(be動詞)	3
65 時制・相(過去)(一般動詞)	3
168 名詞を前置修飾する現在分詞	3
59 時制・相(現在)(一般動詞・3人称単数以外)	2

さらに表示

CEFRランク

C1

間違ったCEFRランクを訂正する

使用語彙レベル割合

A1

A2

B1

機能語

CEFR-J Text Profile

Text profile measures (Mizushima, Arase, & Uchida, 2016)

Common measures	Lexical profile measures	Complexity measures	Grammar measures
word length (1 to 3 letters)	Average difficulty	sum_D_score	avg_G-item
word length (4 to 6 letters)	A1_per	avg_D_score	G-item_per
Word length (7 letters +)	A2_per	sum_L_score	
Average word length	B1_per	avg_L_score	
Types	B2_per	avg_MaxDepth	
TTR	C1_per		
Mean Length of Sentence	C2_per		

D_score: depth x difficulty level

L_score: depth x word length

http://www.tufs.ac.jp/ts/personal/corpuskun/cefr-j/textprofile_score.pdf

CEFR Level Checker

<https://cvla.langedu.jp/cvla.py>

- Application of text profile measures, developed by Satoru Uchida, Kyushu University.
- Online tool for estimating a CEFR level of the input text
- Using a linear model with ARI (readability), VperSent (the number of verbs per sentence), AvrDiff (average CEFR level of vocabulary) and BperA (the ratio of B-level words against A-level words) as predictors.
- Quite robust on texts longer than 500 words

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.

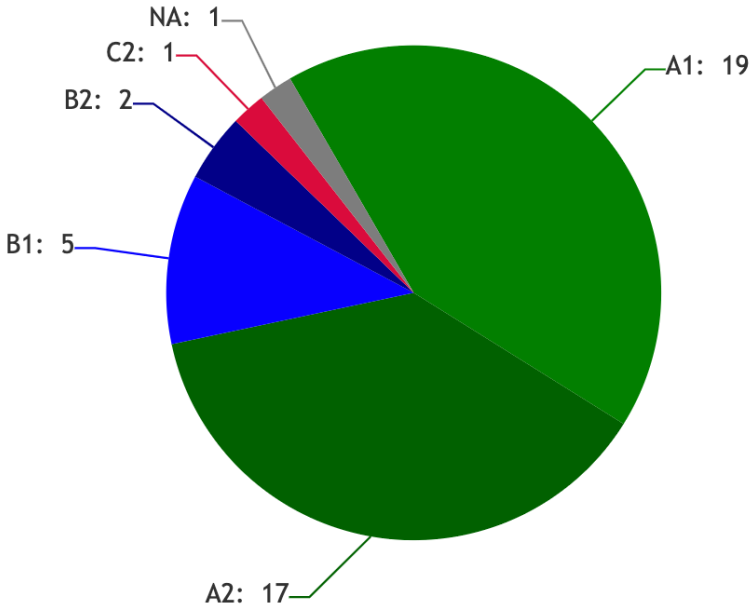
ID	Sentence	Words	Verbs	AvrDiff
1	Writing is the act of recording language on a visual medium using a set of symbols .	16	3	2.0
2	The symbols must be known to others , so that the text may be read .	14	4	1.6
3	A text may also use other visual systems , such as illustrations and decorations .	13	1	2.22
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

CanvasJS Trial

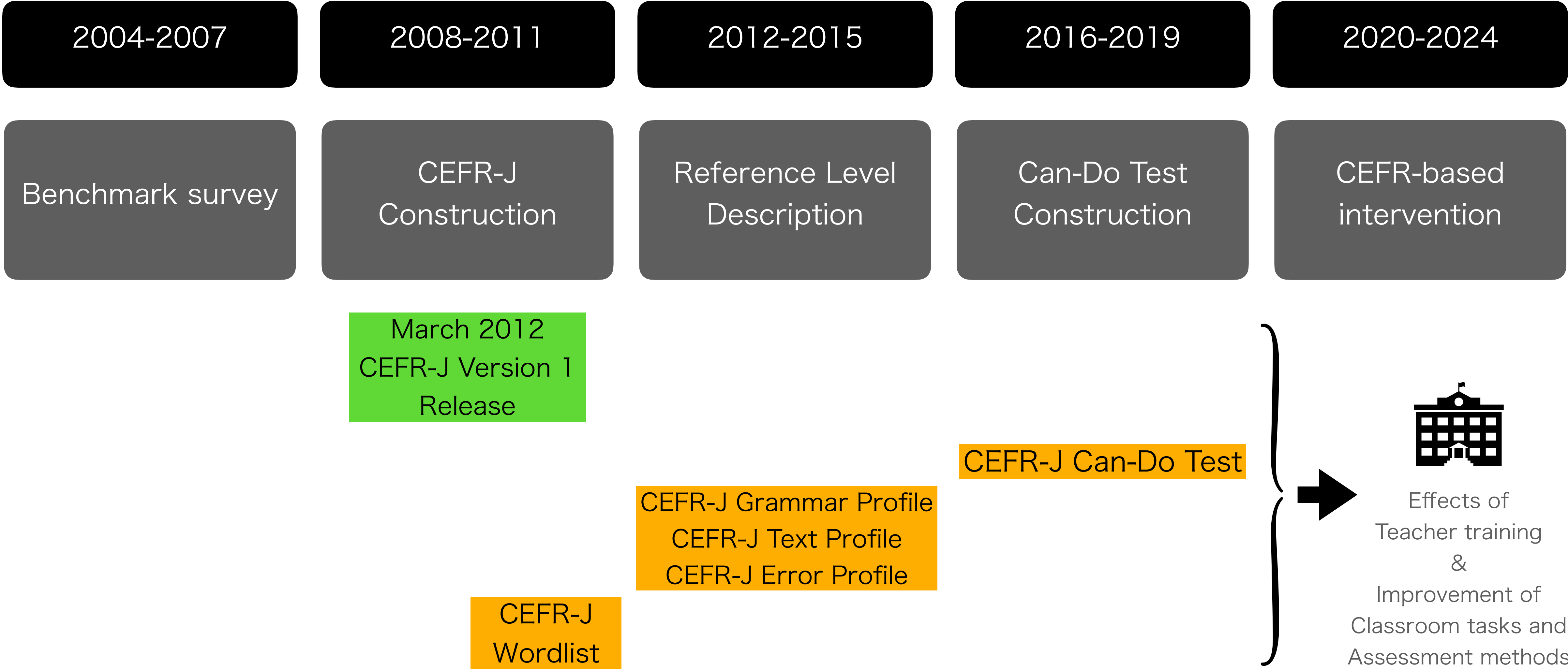
The ratio of CEFR levels (Content words)



CEFR Level	Count
A1	19
A2	17
B1	5
B2	2
C2	1
NA	1

CanvasJS.com

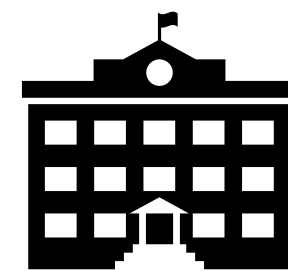
CEFR-J: What's next?



CEFR-J Kyoto Project

2021

2022

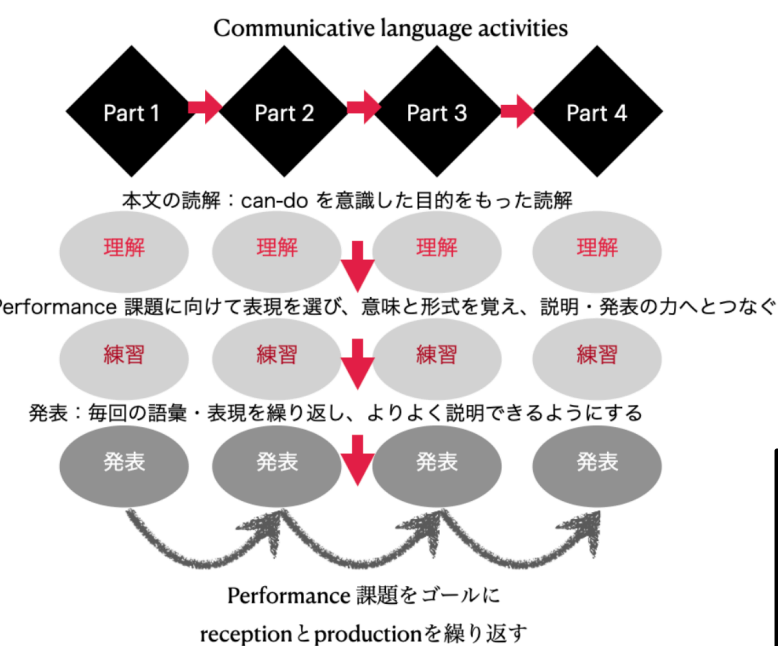


Higashi-Maizuru High

Teacher Workshop

- Understanding the CEFR can do descriptors
- How to improve your lesson using can do's
- How to assess learners using performance tests

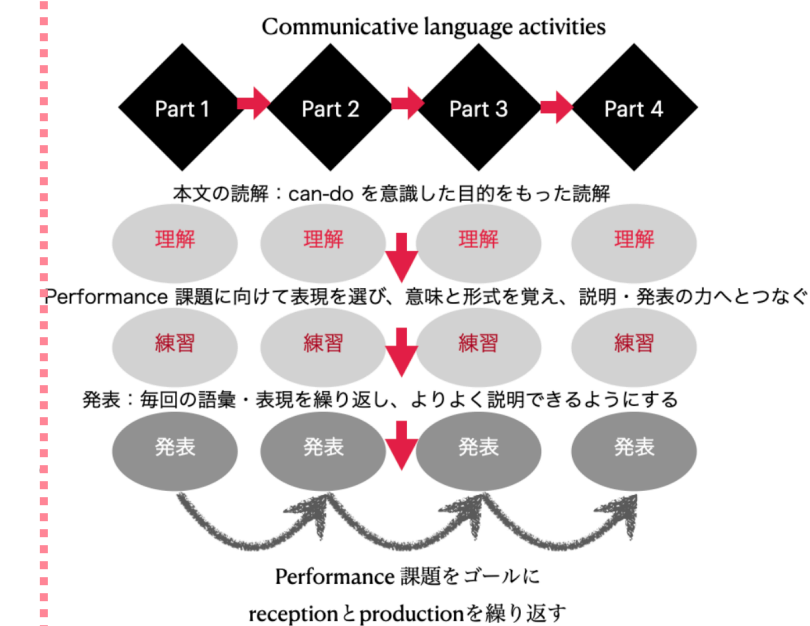
Lesson 5



Classroom Corpus

Performance Test

Lesson 8



Classroom Corpus

Performance Test

Classroom Discourse Corpus

Performance Test Corpus

Term 1

Term 2

Term 3

Feedback

Data Collection

Data Collection

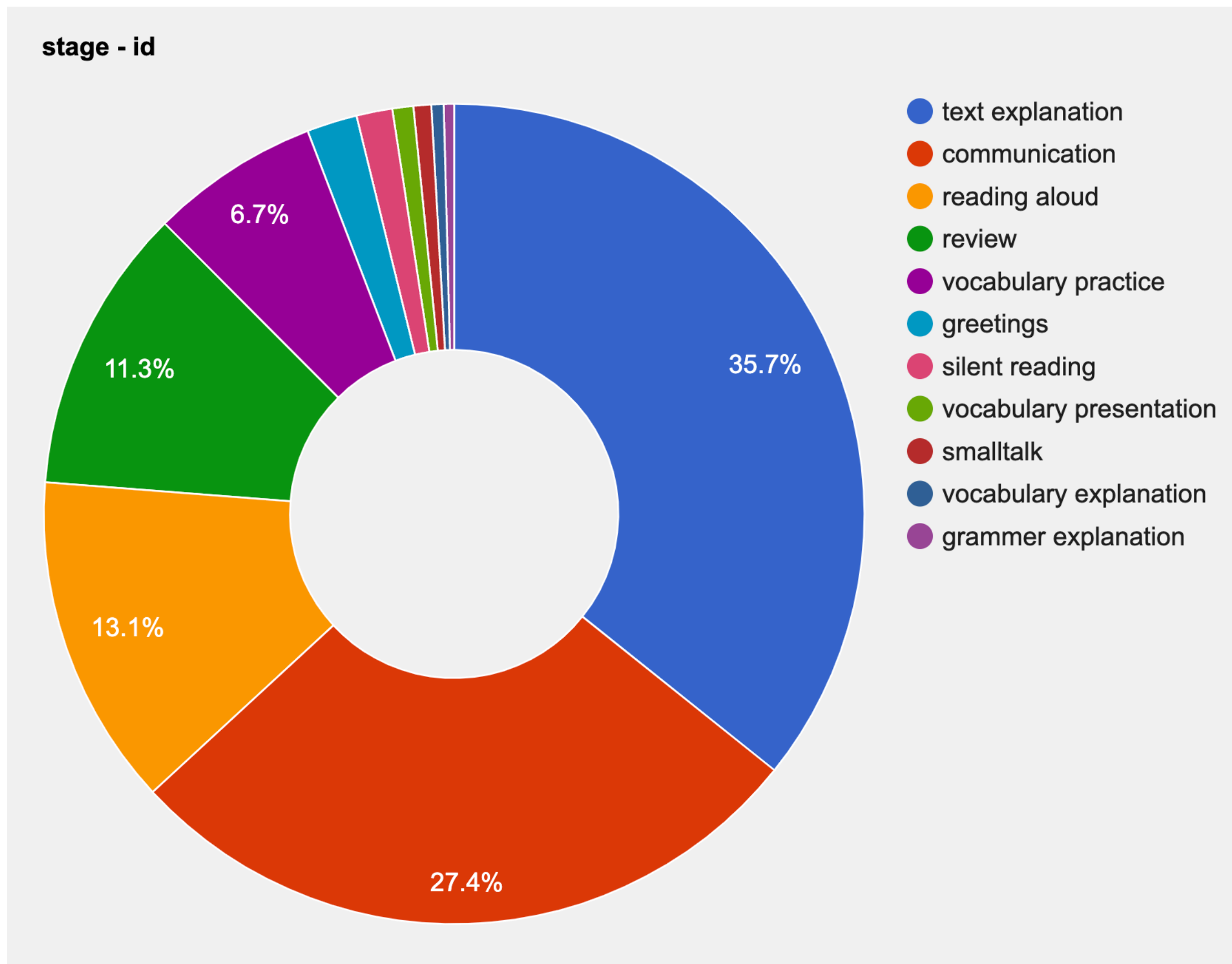
Workshop

Can-Do Test Corpus

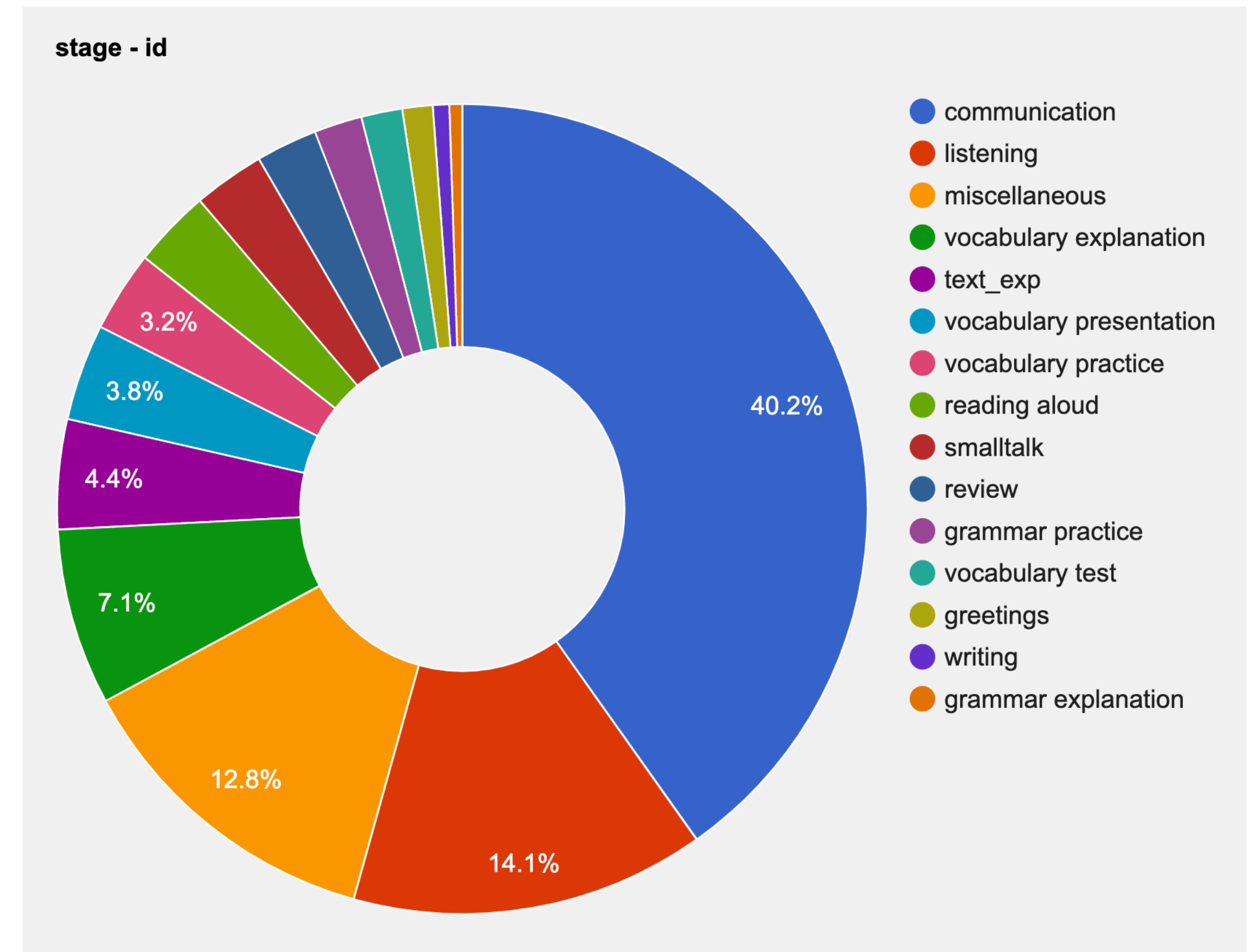
Lesson 5 vs. Lesson 8

The percentage of “text explanation in Japanese” decreased and the time spent on “genuine communication” increased.

Lesson 5



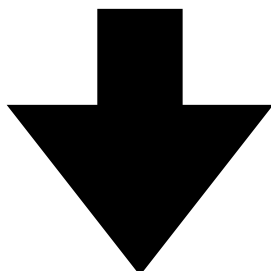
Lesson 8



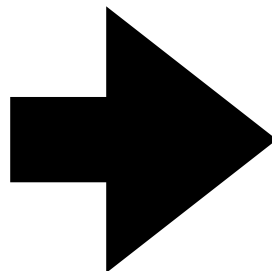
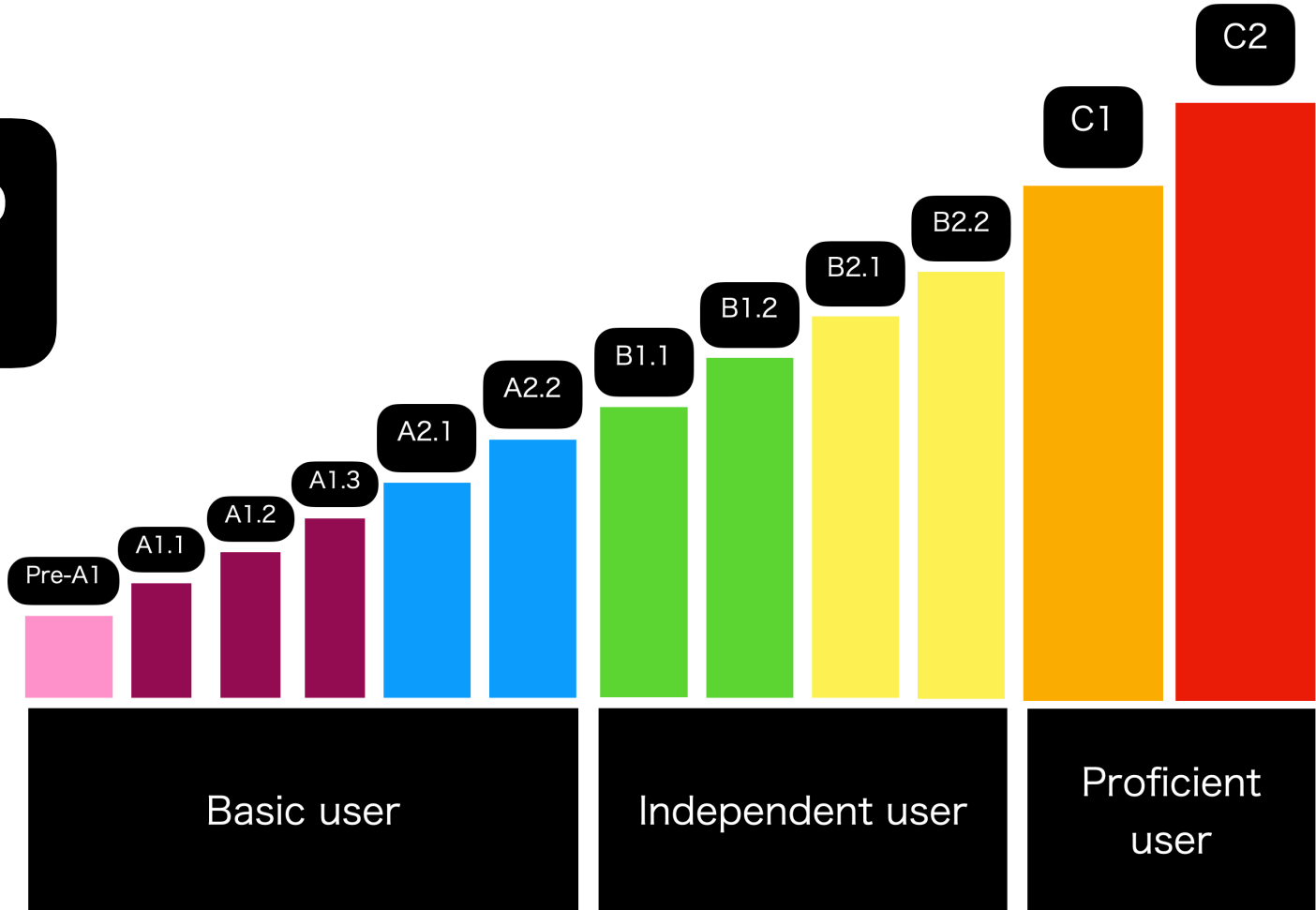
CEFR-J x 28 project at TUFs

Transforming CEFR-J Resources into 27 other languages

Wordlist
&
Phrase list
Pedagogical Corpora

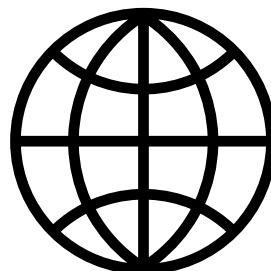


Can do Test



Exporting CEFR-J based resource pack for under-resourced languages

Other universities



International market

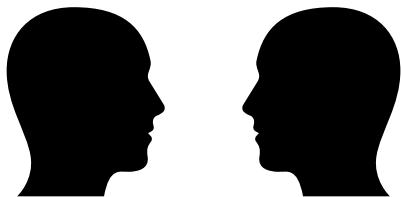
“Can Do”-based common syllabus

Monitoring progress using CEFR-J Can Do Test Batteries

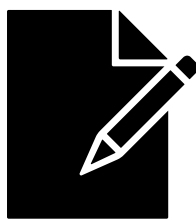


Online/On-demand Learning

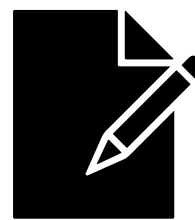
Vocabulary
+
Grammar
+
Text



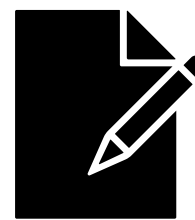
In-person Learning



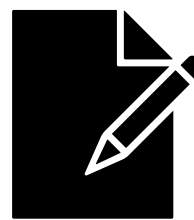
Year 1



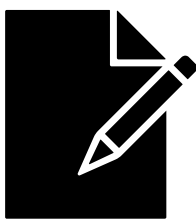
Year 2



Year 3



Year 4



CEFR-J based course syllabus for all the major languages taught at TUFs

Major Language
English
Japanese
+ language 4+
= Plurilingual competence

Thank you!

Email: y.tono@tufts.ac.jp