

Crowdsourcing for language learning: looking for potential

Elena Volodina, University of Gothenburg, Sweden

Louvain-la-Neuve, 3 April 2019

Crowdsourcing

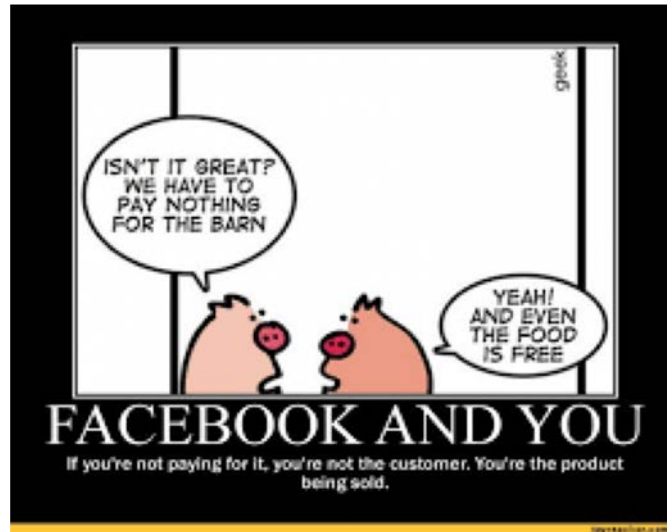
The Rise of Crowdsourcing

Remember *outsourcing*? Sending jobs to India and China is so 2003. The new pool of *cheap labor*: everyday people using their spare cycles to create content, solve problems and even do corporate R & D.

Jeff Howe, 2006, Wired magazine

About the (crowd)sourcing

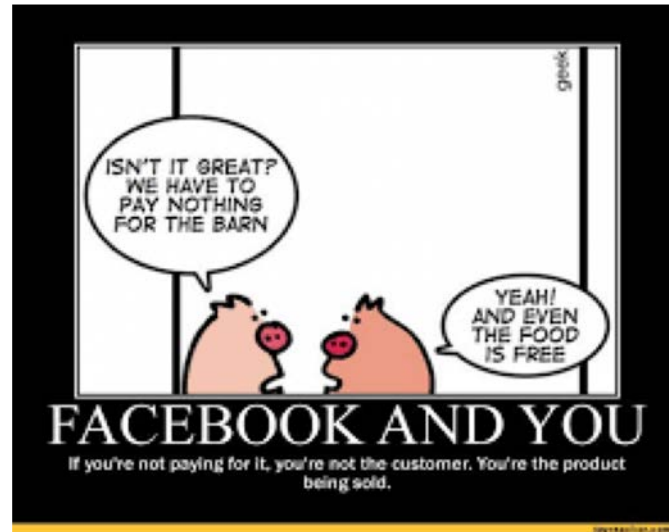
Isn't it great?
We have to
pay nothing
for the barn



YEAH!
And even
the food
is free

If you are not paying for it, you are not the customer.
You are the product being sold.

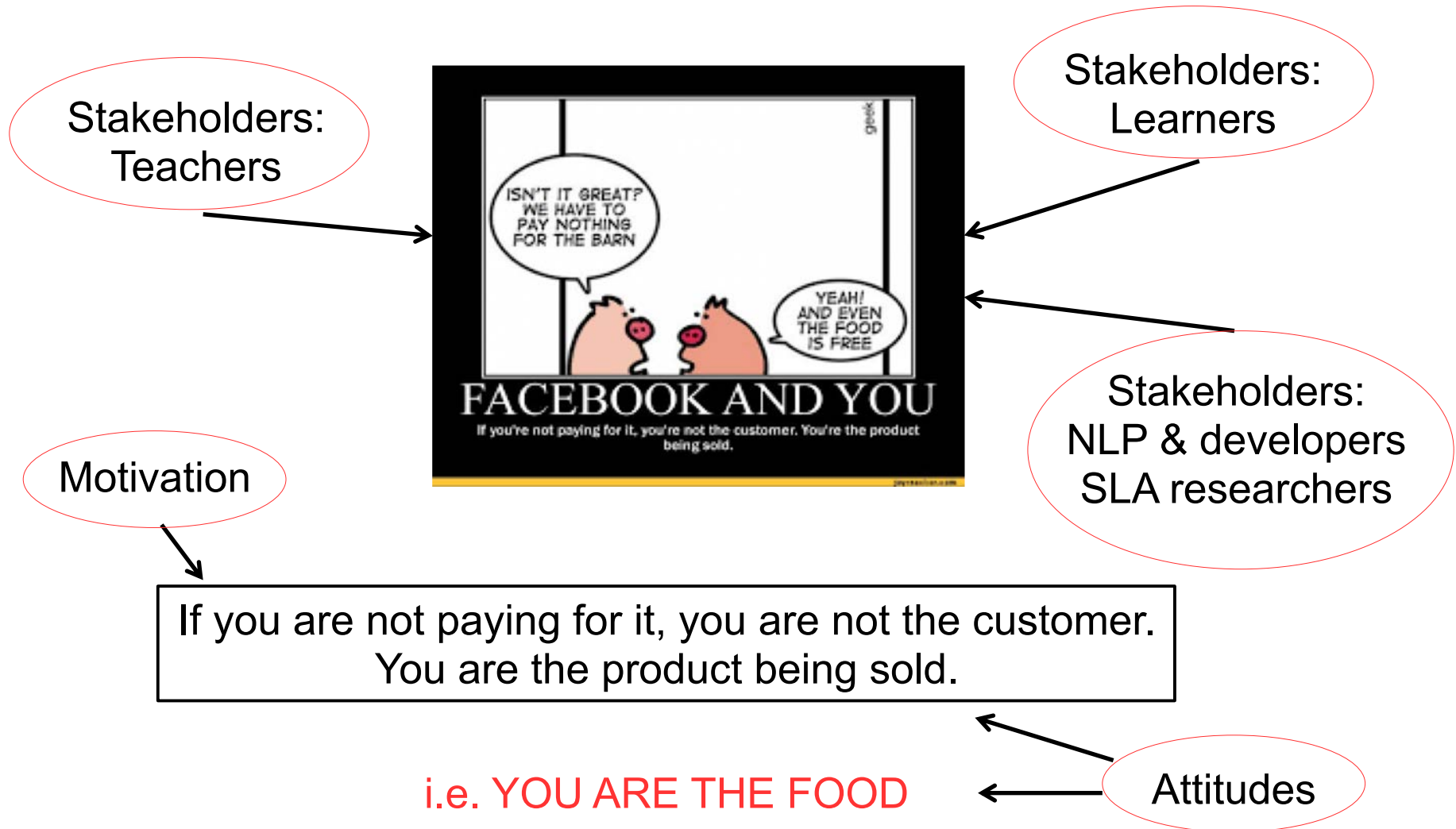
About the (crowd)sourcing



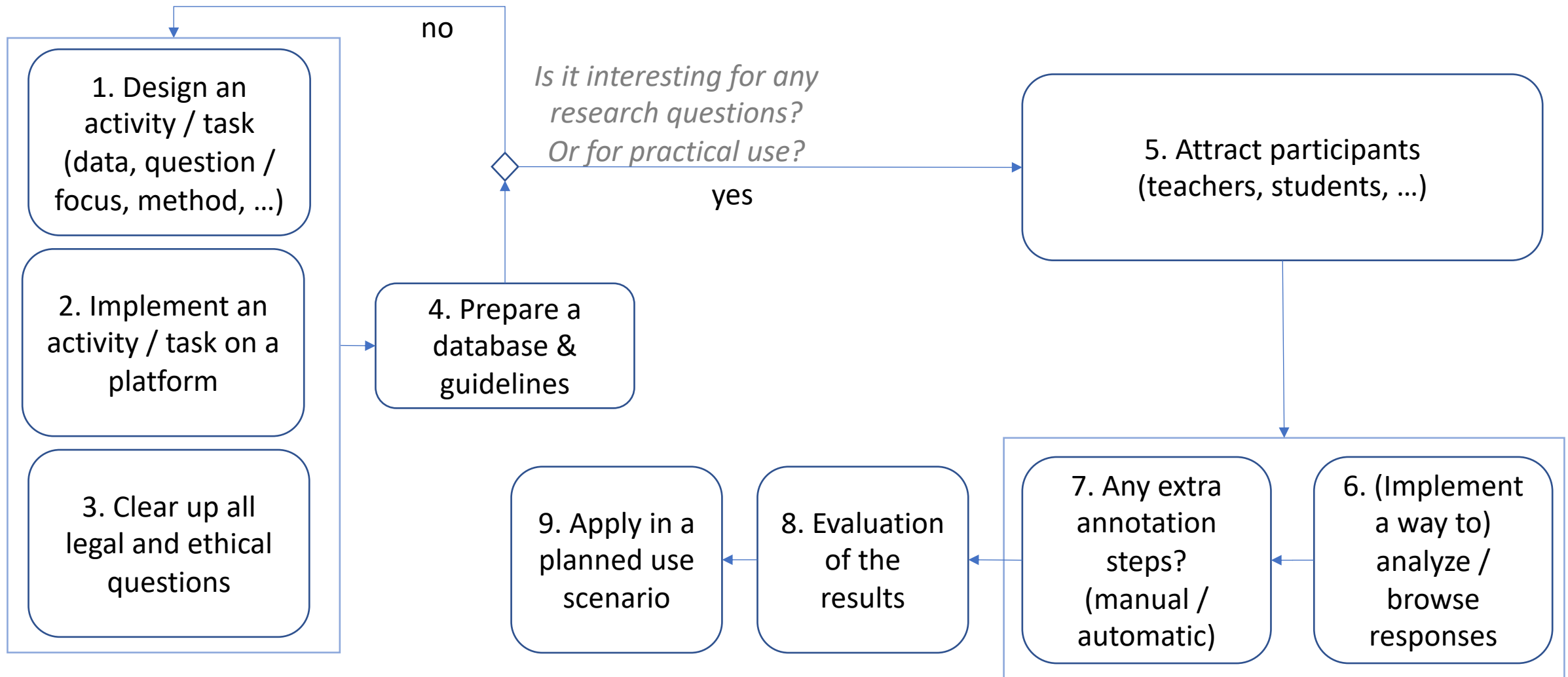
If you are not paying for it, you are not the customer.
You are the product being sold.

i.e. YOU ARE THE FOOD

About the (crowd)sourcing



Crowdsourcing L2 resources/materials: steps



Crowdsourcing for language learning

- Crowdsourcing for language learning
 - Assessing the quality of text questions (*Chinkina & Meurers, 2017*)
 - FeedBook (*Ramon Ziaj, Bjoern Rudzewitz, Kordula De Kuthy, Florian Nuxoll & Detmar Meurers, 2018*)
 - DuoLingo (*Settles, Brust, Gustafson, Hagiwara & Madnani, 2018*)
 - ...
 - but --> not much to go, in fact, to validate crowdsourcing for LL
- We need to have a proof-of-concept that
 - crowdsourcing is valid for annotating/creating language learning data/resources

Idea

- Rank a set of expressions relevant for language learners by difficulty.
- Can this be done through crowdsourcing?
- How?
 - We need a simple task.
 - Manageable workload.
 - (Relatively) reliable results.

Who and when?

- STSM (Elena Volodina, Ljubljana, June 2018)
– planning the experiment
- STSM (Jaka Čibej, Gothenburg, September 2018)
– setting up the experiment
- Preparations and WG1 Workshop (Gothenburg, October–December 2018) –
conducting the experiment and presenting the results



David Alfter



Jaka Čibej



Iztok Kosem



Elena Volodina

What?

- A **crowdsourcing experiment** to rank English multi-word expressions (MWE) according to their difficulty (L2 levels of proficiency)
 - *to burn the midnight oil*
 - *to be absorbed in something*
 - *to add insult to injury*
 - *to be able to do something*

Why MWE ?

- MWEs/formulaic language characterizes learners of more advanced levels
(*e.g. Paquot & Granger 2012, Suñer 2018, Thewissen 2013, Forsberg and Bartning, 2010; Erman et al., 2016*)
- Define a scope for generation of exercises/tests and testing reading materials for appropriateness
- Where to get this data? And how?
 - Automatic annotation
 - Manual annotation
 - Crowdsourcing

MWE experiment: overview

- English Vocabulary Profiles
- Pybossa
- Best-worst scaling
- 5 votes for each task
- Clustering & ranking

The screenshot shows the British English Vocabulary Profile search interface. It features a sidebar on the left with options for British and American English, level selection (A1 to C2), and an advanced search section with filters for category, part of speech, grammar, usage, topic, prefix, and suffix. The main area displays search results for 'A1-C2' on page 1 of 40, listing various phrases with their corresponding CEFR levels indicated by colored boxes.

British English | **American English**

Choose level:

- ☐ A1
- ☐ A1-A2
- ☐ A1-B1
- ☐ A1-B2
- ☐ A1-C1
- ☒ A1-C2
- ☐ A2 only
- ☐ B1 only
- ☐ B2 only
- ☐ C1 only
- ☐ C2 only

[Browse A-Z](#)

OR

Enter a word or phrase

ADVANCED SEARCH ⤴

Category: ⤴

Part of speech: ⤴

Grammar: ⤴

Usage: ⤴

Topic: ⤴

Prefix: ⤴

Suffix: ⤴

[Clear filters](#)

☐ Hide culturally sensitive words

[Search](#)

Search results for A1-C2 (max. number of

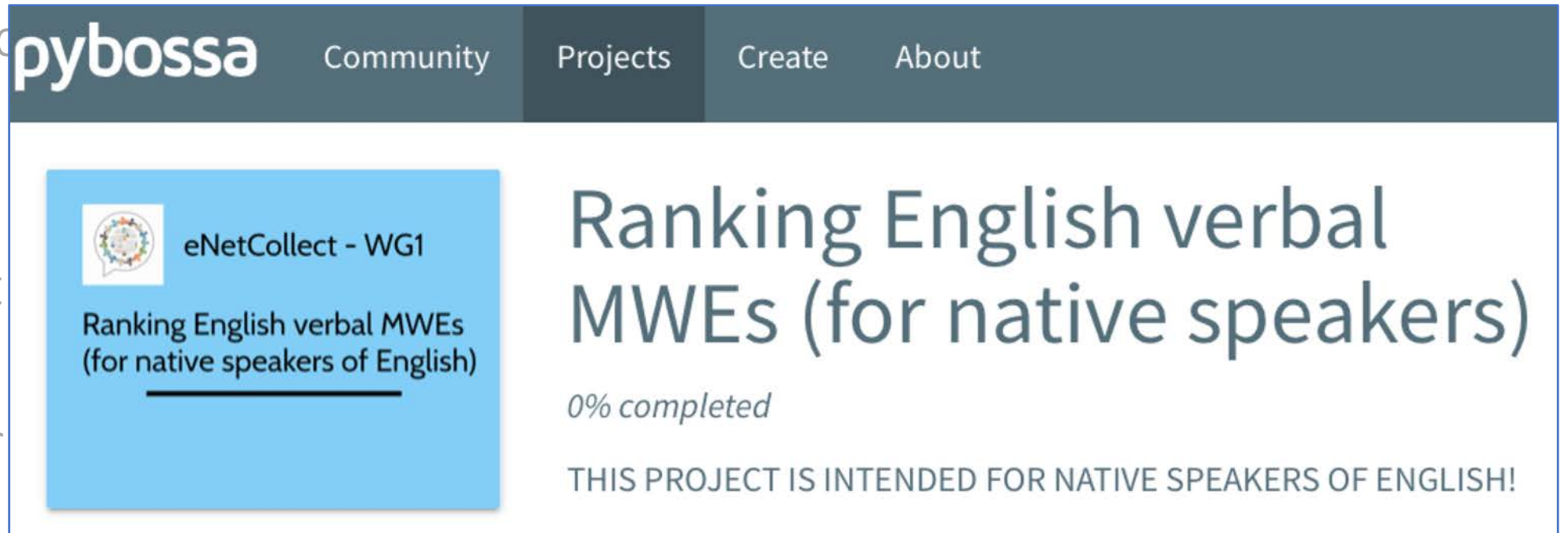
1 2 3 ... 40 > Page 1 of 40

Core results:

- be able to do sth **A2**
- How/What about ...? **A2**
- How/What about ...? **B1**
- be (just) about to do sth **B1**
- above all **B1**
- Absolutely not. **C2**
- be absorbed in sth **B2**
- accept responsibility/blame **B2**
- by accident **B1**
- of your own accord **C2**
- in accordance with sth **C1**
- take account of sth **B2**
- take into account sth **B2**
- on account of sth **B2**
- by all accounts **C1**
- on no account; not on any account **C2**
- accustomed to sth/doing sth **C1**
- legal action **C1**
- out of action **C1**
- course of action **C1**
- in actual fact **B2**
- add insult to injury **C2**
- in addition (to) **B1**

MWEs

- English Vocabulary
- Pybossa
- Best-worst
- 5 votes for
- Clustering & ranking



The screenshot shows the Pybossa website interface. At the top is a dark blue navigation bar with the 'pybossa' logo and links for 'Community', 'Projects', 'Create', and 'About'. The 'Projects' link is highlighted. Below the navigation bar, the main content area features a project card for 'eNetCollect - WG1'. The card has a light blue background and contains a circular logo with a globe. The project title is 'Ranking English verbal MWEs (for native speakers of English)', followed by a horizontal line. To the right of the card, the project title is repeated in a larger font, followed by '0% completed' and the text 'THIS PROJECT IS INTENDED FOR NATIVE SPEAKERS OF ENGLISH!'.

pybossa Community Projects Create About

eNetCollect - WG1

Ranking English verbal MWEs
(for native speakers of English)

Ranking English verbal
MWEs (for native speakers)

0% completed

THIS PROJECT IS INTENDED FOR NATIVE SPEAKERS OF ENGLISH!

MWEs

- English Vocabulary Prof
- Pybossa
- Best-worst scaling
- 5 votes for each task
- Clustering & ranking

Easiest	Expression	Hardest
<input type="radio"/>	it goes without saying	<input type="radio"/>
<input type="radio"/>	kill time	<input type="radio"/>
<input type="radio"/>	lose your temper	<input type="radio"/>
<input type="radio"/>	beat about/around the bush	<input type="radio"/>

[Save](#)

Current task ID number: **688569** .

You have solved **0** task(s) out of a total of **326** . You are expected to solve **82** .

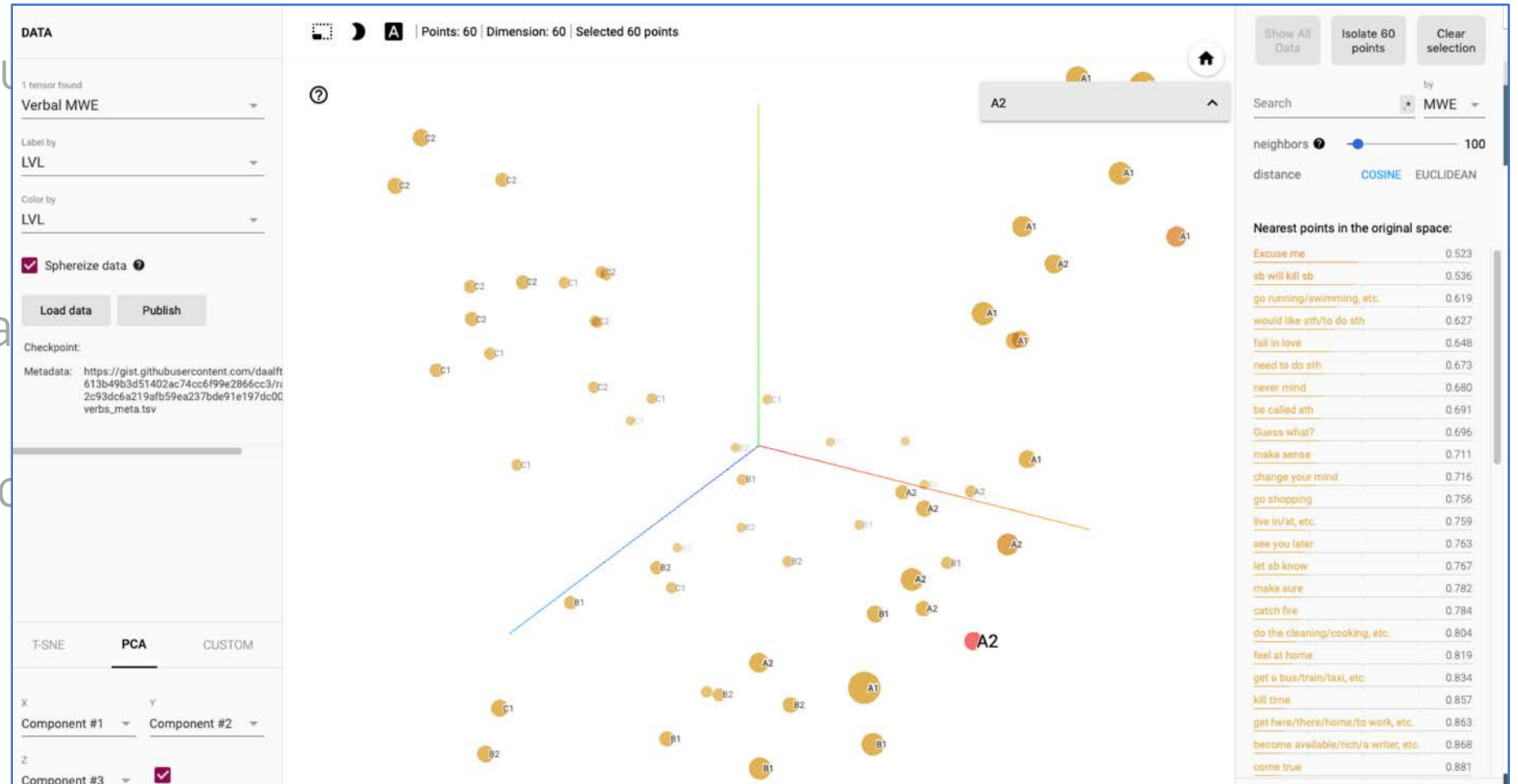
You can fill in [the feedback questionnaire](#) to describe how you made your decisions.

MWEs

- English Vocabulary Profiles
- Pybossa
- Best-worst scaling
- 5 votes for each task. Voter profile: non-native English speaker
- Clustering
& ranking

MWEs

- English Vocabulary
- Pybossa
- Best-worst scaling
- 5 votes for each
- Clustering & ranking



MWE clustering

MWE	Excuse me	Guess what?	be called sth	beat about/around the bush	become available/rich/a writer, etc.	break the ice	break the law	bring a lump
Excuse me	0.0	1.0	1.2857142857142858	1.7142857142857142	0.8571428571428571	1.0	1.0	1.4285714285714286
Guess what?	1.0	0.0	1.0	1.2857142857142858	0.5714285714285714	1.2857142857142858	0.8571428571428571	1.4285714285714286
be called sth	1.2857142857142858	1.0	0.0	1.1428571428571428	0.8571428571428571	0.5833333333333334	0.5714285714285714	1.1428571428571428
beat about/around the bush	1.7142857142857142	1.2857142857142858	1.1428571428571428	0.0	1.4285714285714286	1.3333333333333333	0.42857142857142855	1.1666666666666667
become available/rich/a writer, etc.	0.8571428571428571	0.5714285714285714	0.8571428571428571	1.4285714285714286	0.0	1.0	0.8571428571428571	1.0
break the ice	1.0	1.2857142857142858	0.5833333333333334	1.3333333333333333	1.0	0.0	0.8571428571428571	0.85
break the law	1.0	0.8571428571428571	0.5714285714285714	0.42857142857142855	0.8571428571428571	0.8571428571428571	0.0	1.0
bring a lump to your throat	1.4285714285714286	1.4285714285714286	1.1428571428571428	1.1666666666666667	1.0	0.85	1.0	0.0
burn the midnight oil	2.0	2.0	1.2	1.0	1.5	1.0526315789473684	1.4	0.7142857142857142
can't/couldn't help doing sth	1.7142857142857142	1.5714285714285714	1.0	1.0	1.1666666666666667	0.42857142857142855	0.5714285714285714	1.2857142857142858
catch fire	1.0	1.0	0.7142857142857143	1.5714285714285714	1.1428571428571428	0.7142857142857143	0.8	1.0
change your mind	1.1428571428571428	0.14285714285714285	1.0	1.0	0.6428571428571429	1.2857142857142858	0.4	1.1666666666666667
come true	1.0	1.0	1.0	1.4285714285714286	0.5	1.0714285714285714	0.7894736842105263	0.8333333333333333
crack a joke	1.8	1.1428571428571428	1.1428571428571428	0.5714285714285714	1.4285714285714286	1.0	0.42857142857142855	0.8571428571428571
cross your mind	1.0	1.0	1.5	1.0	0.7142857142857143	0.5714285714285714	0.7142857142857143	0.5714285714285714
do the cleaning/cooking, etc.	0.8571428571428571	1.0	1.0	0.5714285714285714	1.0	0.7142857142857143	1.3333333333333333	1.0
draw a conclusion	1.037037037037037	0.5714285714285714	1.0	0.7142857142857143	1.0	1.1428571428571428	0.42857142857142855	1.0
drive sb mad/crazy, etc.	1.0	0.5714285714285714	0.14285714285714285	0.7142857142857143	1.0	0.2857142857142857	1.0	1.0
face the music	1.8571428571428572	1.5714285714285714	1.0	0.6428571428571429	1.1428571428571428	1.0	0.7142857142857143	1.0
fall flat	1.2857142857142858	0.8571428571428571	0.42857142857142855	1.0	1.7142857142857142	0.7142857142857143	0.4	1.0
fall in love	0.8	1.0	0.8571428571428571	1.1428571428571428	0.8571428571428571	1.1428571428571428	1.0	1.4285714285714286
feel at home	0.2	1.1428571428571428	0.14285714285714285	0.7142857142857143	0.6666666666666666	0.8571428571428571	0.7142857142857143	1.3571428571428572
follow suit	1.6428571428571428	1.5714285714285714	1.5714285714285714	1.0	1.2857142857142858	0.5714285714285714	0.8571428571428571	1.0
get a bus/train/taxi, etc.	0.7142857142857143	0.5714285714285714	1.4285714285714286	0.5714285714285714	0.4	1.0	0.5714285714285714	2.0

MWE ranking

MWE	CEFR	average_rank
burn the midnight oil	C2	2,777027027
grasp the nettle	C2	2,769736842
go against the grain	C2	2,75862069
throw in the towel	C2	2,68707483
beat about/around the bush	C1	2,647058824
follow suit	C2	2,644295302
keep sb on their toes	C2	2,620437956
nothing ventured, nothing gained	C2	2,6125
go from strength to strength	C1	2,584507042
bring a lump to your throat	C2	2,553333333
face the music	C1	2,549668874
fall flat	C1	2,503311258
get a grip (on yourself)	C1	2,493421053
hit the roof	C2	2,455172414
let off steam	C2	2,434482759
keep a low profile	C1	2,364238411
crack a joke	C1	2,353333333
get sth straight	C1	2,268115942
take it for granted	B2	2,253424658
lose your temper	B2	2,208333333
it goes without saying	B2	2,2
keep sb posted	C1	2,190789474
to cut a long story short	C1	2,15625
make up your mind or make your mind up	B1	2,151898734

Data

- English Vocabulary Profile (EVP, *Capell 2010, 2012*)
 - <http://vocabulary.englishprofile.org/staticfiles/about.html>
 - user: englishprofile, password: vocabulary
- MWEs are defined in terms of "phrases", "phrasal verbs" & "idioms" in EVP
- Verbal MWEs: 10 per CEFR level = 60 items
 - *to burn the midnight oil*
 - *it goes without saying*
- Adverbial MWES: 10 per CEFR level = 60 items
 - *Happy New Year!*
 - *by all accounts*

The screenshot displays the English Vocabulary Profile (EVP) search interface. It features a sidebar on the left for selecting the language (British or American English) and the CEFR level (A1 to C2). The main area shows search results for A1-C2, with a list of 40 items. Each item is a phrase followed by a colored box indicating its CEFR level. For example, 'be able to do sth' is marked A2, 'How/What about ...?' is marked A2, 'be (just) about to do sth' is marked B1, 'above all' is marked B1, 'Absolutely not.' is marked C2, 'be absorbed in sth' is marked B2, 'accept responsibility/blame' is marked B2, 'by accident' is marked B1, 'of your own accord' is marked C2, 'in accordance with sth' is marked C1, 'take account of sth' is marked B2, 'take into account sth' is marked B2, 'on account of sth' is marked B2, 'by all accounts' is marked C1, 'on no account; not on any account' is marked C2, 'accustomed to sth/doing sth' is marked C1, 'legal action' is marked C1, 'out of action' is marked C1, 'course of action' is marked C1, 'in actual fact' is marked B2, 'add insult to injury' is marked C2, and 'in addition (to)' is marked B1. The interface also includes an 'ADVANCED SEARCH' section with filters for Category, Part of speech, Grammar, Usage, Topic, Prefix, and Suffix, and a 'Search' button.

British English | American English

Choose level:

- ☐ A1
- ☐ A1-A2
- ☐ A1-B1
- ☐ A1-B2
- ☐ A1-C1
- ☒ A1-C2
- ☐ A2 only
- ☐ B1 only
- ☐ B2 only
- ☐ C1 only
- ☐ C2 only

Browse A-Z

OR

Enter a word or phrase

ADVANCED SEARCH

Category: phrases

Part of speech: --Any--

Grammar: --Any--

Usage: --Any--

Topic: --Any--

Prefix: --Any--

Suffix: --Any--

Clear filters

Hide culturally sensitive words

Search

Search results for A1-C2 (max. number of 40)

1 2 3 ... 40 > Page 1 of 40

Core results:

- be able to do sth **A2**
- How/What about ...? **A2**
- How/What about ...? **B1**
- be (just) about to do sth **B1**
- above all **B1**
- Absolutely not. **C2**
- be absorbed in sth **B2**
- accept responsibility/blame **B2**
- by accident **B1**
- of your own accord **C2**
- in accordance with sth **C1**
- take account of sth **B2**
- take into account sth **B2**
- on account of sth **B2**
- by all accounts **C1**
- on no account; not on any account **C2**
- accustomed to sth/doing sth **C1**
- legal action **C1**
- out of action **C1**
- course of action **C1**
- in actual fact **B2**
- add insult to injury **C2**
- in addition (to) **B1**

CEFR – Common European Framework of Reference



EVP data

- Labeled by lexicographers / teachers → EXPERTS
- Based on Cambridge Learner Corpus (CLC) - a corpus of L2 English
- Using our system of ranking – how well can “a crowd” perform → NON-EXPERTS

The screenshot displays the Cambridge Learner Corpus (CLC) search interface. On the left, there are tabs for 'British English' and 'American English'. Below these, a 'Choose level:' section lists various proficiency levels (A1, A1-A2, A1-B1, A1-B2, A1-C1, A1-C2, A2 only, B1 only, B2 only, C1 only, C2 only) with radio buttons. A 'Browse A-Z' link and an 'Enter a word or phrase' search bar are also present. An 'ADVANCED SEARCH' section includes dropdown menus for 'Category' (set to 'phrases'), 'Part of speech', 'Grammar', 'Usage', 'Topic', 'Prefix', and 'Suffix', each with a 'Clear filters' button. A checkbox for 'Hide culturally sensitive words' and a 'Search' button are at the bottom left. The right side shows 'Search results for A1-C2 (max. number of)' with a pagination bar indicating 'Page 1 of 40'. A list of 'Core results' follows, each preceded by a bullet point and a colored label (A2, B1, B2, C1, C2) indicating the proficiency level. The results include phrases like 'be able to do sth', 'How/What about ...?', 'be (just) about to do sth', 'above all', 'Absolutely not.', 'be absorbed in sth', 'accept responsibility/blame', 'by accident', 'of your own accord', 'in accordance with sth', 'take account of sth', 'take into account sth', 'on account of sth', 'by all accounts', 'on no account; not on any account', 'accustomed to sth/doing sth', 'legal action', 'out of action', 'course of action', 'in actual fact', 'add insult to injury', and 'in addition (to)'.

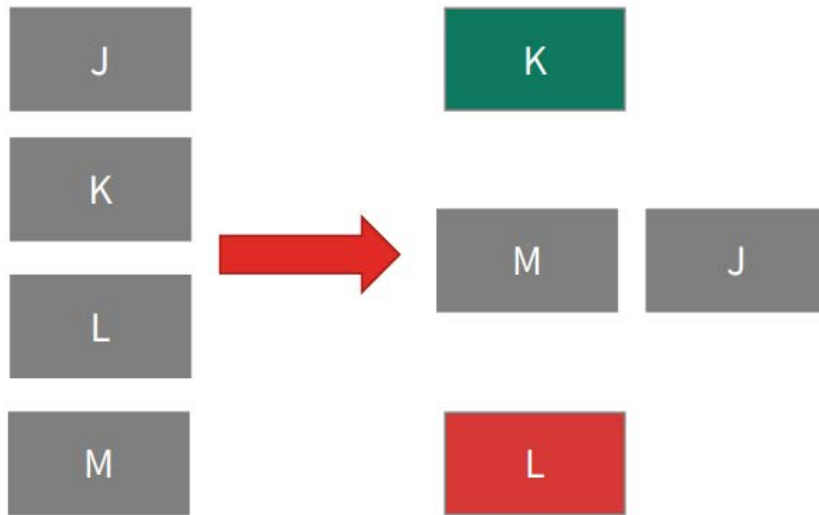
Over - to set up and results

How to rank?

- Ranking the entire list?
 - Task can't be divided between multiple participants.
- Ranking a subset of tasks?
 - Combinations might affect results.
 - Still not very user-friendly.
 - Difficult to merge?
 - Which combinations?

Best-Worst Scaling

- Ranking method
- Choosing the best and worst unit in a combination of (ideally) 3–4 candidates
- Example:



- 6 possible binary relations between the 4 elements
 - $J \sim K, J \sim L, J \sim M, K \sim L, K \sim M, L \sim M$
- **BWS with 4 elements**
 - $K = 3, M = 2, J = 2, L = 1$
 - $J < K, J > L, J \sim M, K > L, K > M, L < M$
 - 5 out of 6 relations (83 %)
 - (at least) 2 clicks
- **Ranking all 4 elements:**
 - 6 out of 6 relations (100 %)
 - (at least) 4 clicks
 - twice the workload!

Selecting the Optimal Number of Combinations

- We go through all combinations and choose only the ones where no relation is repeated (in order to avoid tasks where we get too many repeated relations, which are practically useless).
 - We continue by selecting tasks with only 1 repeated relation, then 2, then 3, then 4, then 5 (until we cover all possible binary relations).
 - Why?
 - To minimize the number of (completely) redundant tasks.
- 60 expressions
 - 1,770 binary relations
 - 1,362 (77%) relations covered with **non-repetitive combinations**.
 - **33 combinations** where 1 relation is already known.
 - **50 combinations** where 2 relations are already known.
 - **12 combinations** where 3 relations are already known.
 - **3 combinations** where 4 relations are already known.
 - **1 combination** where 5 relations are already known.

Tasks

- 60 expressions per project
 - 487,635 combinations (for combinations of 4 units)
 - 1,770 binary relations
- 326 tasks per project (to include all binary relations between the expressions)
- 77% are non-repetitive.
- 23% are partially repetitive (as little as possible).

Final Set of Tasks

- 326 tasks
- 77% are non-repetitive.
- 23 % are partially repetitive (as little as possible).

PREDICTIONS:

IF:

- Number of crowdsourceers: **20**
- Average response time: **30 seconds**
- Responses per task: **5**

THEN:

- Time per crowdsourceer: **0.68 hours**, which equals **40.75 minutes**

PyBossa Interface

Easiest	Expression	Hardest
<input type="radio"/>	a lot	<input type="radio"/>
<input type="radio"/>	once upon a time	<input type="radio"/>
<input type="radio"/>	as it happens	<input type="radio"/>
<input type="radio"/>	deadly dull/serious, etc.	<input type="radio"/>

Save

as it happens

Meaning: something that you say in order to introduce a surprising fact

Example: As it happens, her birthday is the day after mine.

Current task ID number: 689222 .

You have solved 1 task(s) out of a total of 326 . You are expected to solve 82 .

You can fill in [the feedback questionnaire](#) to describe how you made your decisions.

<https://pybossa.com>

PyBossa Interface

- phone compatibility (not too wide or too long, etc.)
- user-friendly (or is it?)
- foreseen error scenarios - warnings helped limit any technical mistakes during annotation
 - e.g. only one ticked expression,
 - same expression in both columns

mnozicenje.cjvt.si says

Please tick an expression in each column before saving.

OK

Easiest	Expression	Hardest
<input type="radio"/>	a lot	<input type="radio"/>
<input type="radio"/>	once upon a time	<input type="radio"/>
<input type="radio"/>	as it happens	<input type="radio"/>
<input type="radio"/>	deadly dull/serious, etc.	<input type="radio"/>

Save

as it happens

Meaning: something that you say in order to introduce a surprising fact

Example: As it happens, her birthday is the day after mine.

Current task ID number: 689222 .

You have solved 1 task(s) out of a total of 326 . You are expected to solve 82 .

You can fill in [the feedback questionnaire](#) to describe how you made your decisions.

Guidelines

- Decide which expression is the most difficult/easiest for a language learner **to produce**.
- In case of a tie, choose one.
- Do not overthink the decision.
- Try not to spend more than 30 seconds per task.
- (No mention of the English Vocabulary Profile OR CEFR-levels!)
 - crowdsourcers only relied on their intuition
- 26 participants, mostly linguists and NLP experts
 - 24 non-native speakers of English, 2 native speakers

Results - Metadata

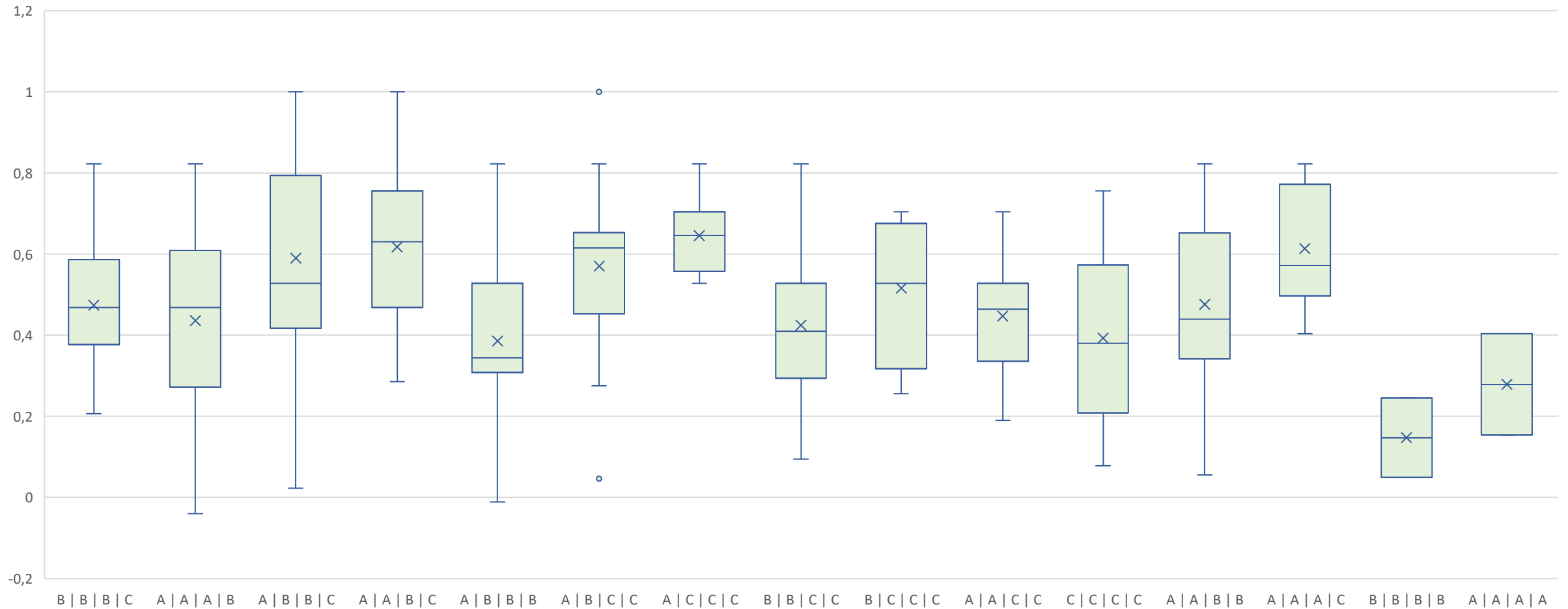
- 2 projects with 326 tasks
- Up to 7 responses per task (at least 5).
- A total of 26 annotators.

Metadata	Adverbs	Verbs
Mean response time	47.4 seconds	50.38 seconds
Median response time	22.9 seconds	26.67 seconds
Total time spent on tasks	27.88 hours	31.25 hours
Mean response time (no outliers over 30 seconds)	18.54 seconds	20.12 seconds
Median response time (no outliers over 30 seconds)	18.3 seconds	20.02 seconds
Total time spent on tasks (no outliers over 30 seconds)	7.26 hours	7.24 hours
Time per crowdsourcer (no outliers over 30 seconds)	0.28 hours	0.29 hours

Results – Agreement (Verbs)

- Inter-annotator agreement (Krippendorff's Alpha)

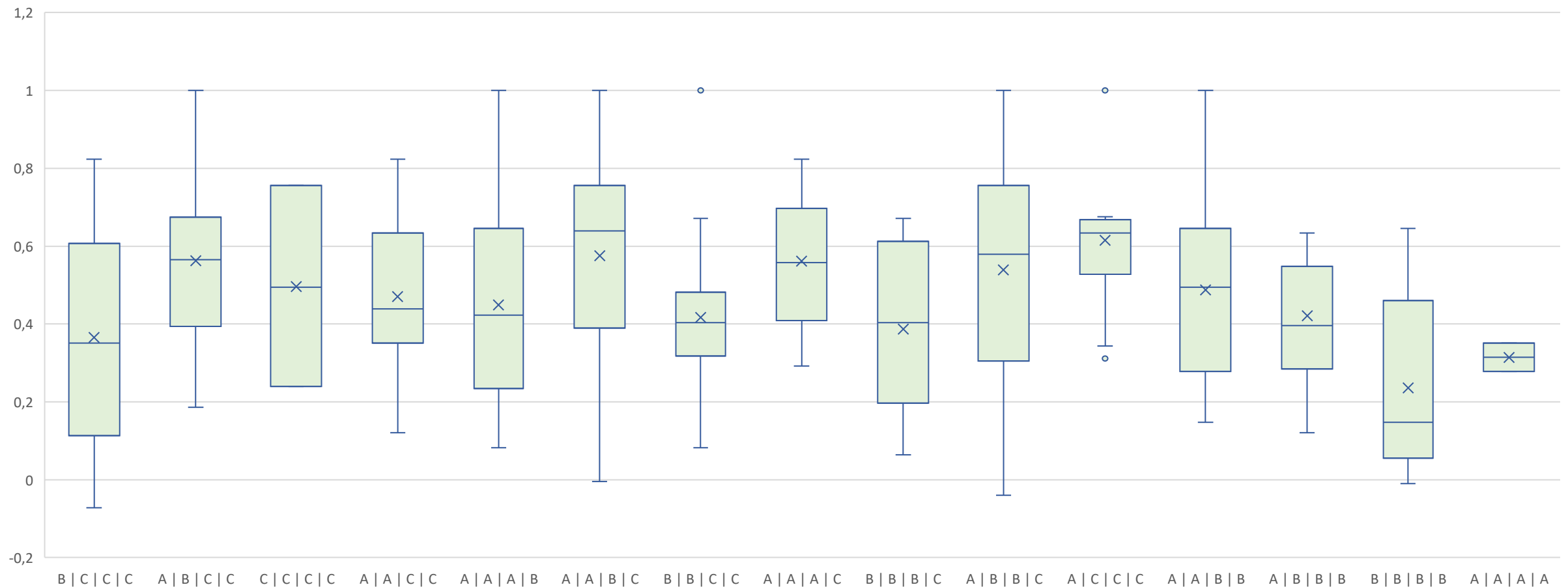
Krippendorff's Alpha by CEFR-combinations (Verbs)



Results – Agreement (Adverbs)

- Inter-annotator agreement (Krippendorff's Alpha)

Krippendorff's Alpha by CEFR-combinations (Adverbs)



Merging the Results

- Method 1: Linear scale using average ranks
 - a more brute-force approach
 - take all annotations for a specific expression (regardless of the expressions it appears with)
 - average the sum to get the expression's average rank
 - the premise: harder/easier expressions should more frequently be annotated as more difficult (rank 3) or easier (rank 1)

Linear Scale

MWE	CEFR	average_rank
burn the midnight oil	C2	2,781818182
go against the grain	C2	2,771428571
grasp the nettle	C2	2,745454545
follow suit	C2	2,681818182
throw in the towel	C2	2,657142857
beat about/around the bush	C1	2,636363636
keep sb on their toes	C2	2,63
nothing ventured, nothing gained	C2	2,608695652
go from strength to strength	C1	2,59047619
face the music	C1	2,545454545
bring a lump to your throat	C2	2,536363636

have a rest/shower/walk, etc.	A2	1,438095238
get here/there/home/to work, etc.	A1	1,409090909
do the cleaning/cooking, etc.	A1	1,4
get a bus/train/taxi, etc.	A1	1,381818182
go running/swimming, etc.	A2	1,32
see you later	A1	1,260869565
live in/at, etc.	A1	1,217391304
Excuse me	A1	1,209090909
go shopping	A1	1,18

MWE	CEFR	average_rank
burn the midnight oil	C2	2,781818182
go against the grain	C2	2,771428571
grasp the nettle	C2	2,745454545
follow suit	C2	2,681818182
throw in the towel	C2	2,657142857
beat about/around the bush	C1	2,636363636
keep sb on their toes	C2	2,63
nothing ventured, nothing gained	C2	2,608695652
go from strength to strength	C1	2,59047619
face the music	C1	2,545454545
bring a lump to your throat	C2	2,536363636
fall flat	C1	2,527272727
get a grip (on yourself)	C1	2,481818182
let off steam	C2	2,466666667
hit the roof	C2	2,428571429
crack a joke	C1	2,381818182
keep a low profile	C1	2,372727273
take it for granted	B2	2,247619048
lose your temper	B2	2,241666667
get sth straight	C1	2,228571429
it goes without saying	B2	2,19047619
keep sb posted	C1	2,172727273
to cut a long story short	C1	2,139130435
make up your mind	B1	2,130434783
cross your mind	B2	2,127272727
draw a conclusion	B2	2,095552174
break the ice	B2	2,091666667
can't/couldn't help do sth	B1	2,07826087
get rid of sth	B1	2,057142857
break the law	B2	2,019047619
catch fire	B1	2
drive sb mad/crazy, etc.	B2	1,980952381
kill time	B2	1,963636364
keep in touch	B1	1,904761905
help yourself (to sth)	B1	1,904761905
fall in love	B1	1,754545455
change your mind	B1	1,752380952
be called sth	A1	1,75
make sense	B2	1,747826087
let sb know	B2	1,747826087
come true	B1	1,745454545
give sb a call/ring	B2	1,73
never mind	B2	1,7
feel at home	B1	1,660869565
make sure	B2	1,652173913
become available/ric	B2	1,647619048
give a party	B2	1,6
would like sth/to do sth	A1	1,6
sb will kill sb	B2	1,580952381
need to do sth	A1	1,539130435
Guess what?	B2	1,514285714
have a rest/shower/walk	B2	1,438095238
get here/there/home	A1	1,409090909
do the cleaning/cook	A1	1,4
get a bus/train/taxi, etc.	A1	1,381818182
go running/swimming	B2	1,32
see you later	A1	1,260869565
live in/at, etc.	A1	1,217391304
Excuse me	A1	1,209090909
go shopping	A1	1,18

Ranking

- Method 1: Linear scale using average ranks
 - Adverbial MWEs: 41.7% accuracy
 - Verbal MWEs: 50.0% accuracy
 - most misclassifications between neighboring levels!
 - e.g. A1 ~ A2, C1 ~ C2, but no A1 ~ C2

Verbal MWEs

Assigned ↓ True ->	A1	A2	B1	B2	C1	C2
A1	6	1	2	1	0	0
A2	2	5	3	0	0	0
B1	0	1	3	5	1	0
B2	0	1	1	5	3	0
C1	0	0	0	3	2	5
C2	0	0	0	1	0	9

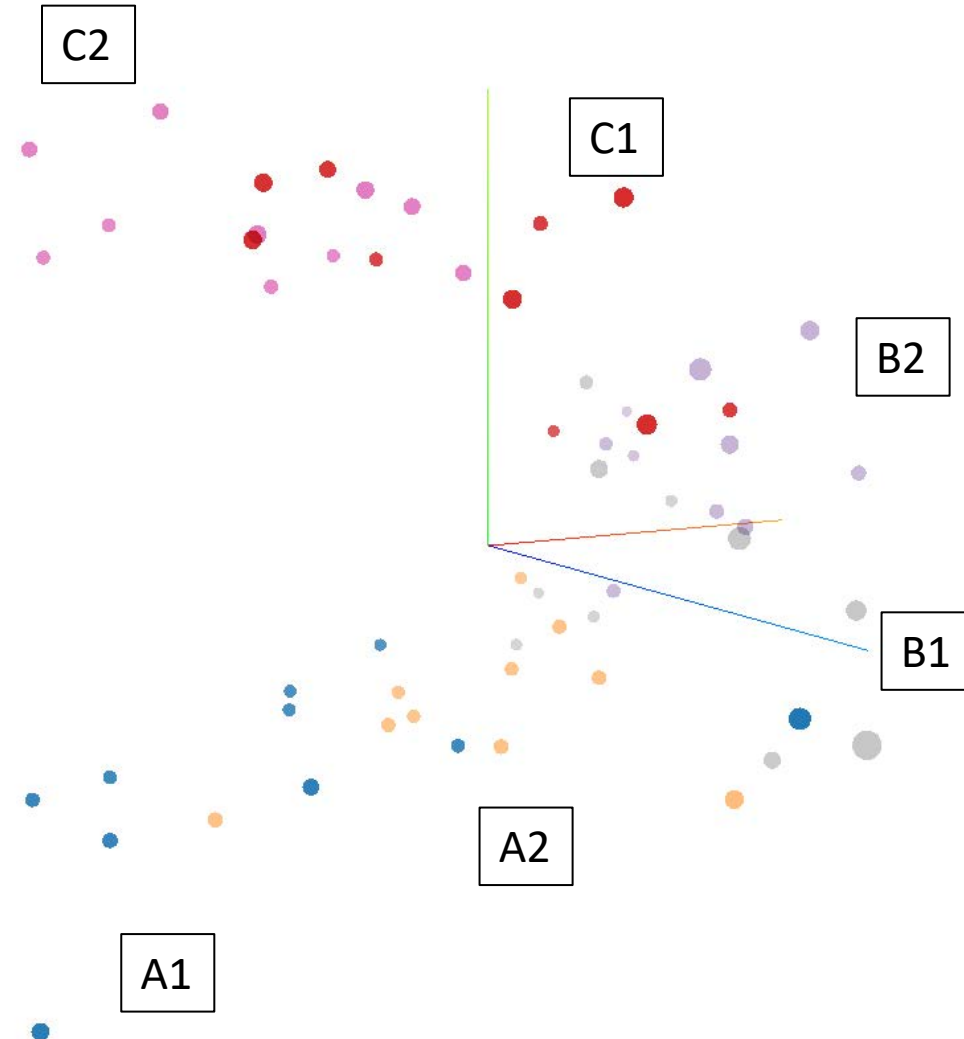
Merging the Results

- Method 2: Clustering and multi-dimensional visualization using vector embeddings
- 60x60 matrix of average distances between expressions

MWE	Excuse me	Guess what?	be called sth	beat about/around the bush	become available/rich/a writer, etc.	break the ice	break the law	bring a lump to your th
Excuse me	0.0	1.0	1.2857142857142858	1.7142857142857142	0.8571428571428571	1.0	1.0	1.4285714285714286
Guess what?	1.0	0.0	1.0	1.2857142857142858	0.5714285714285714	1.2857142857142858	0.8571428571428571	1.4285714285714286
be called sth	1.2857142857142858	1.0	0.0	1.1428571428571428	0.8571428571428571	0.5833333333333334	0.5714285714285714	1.1428571428571428
beat about/around the bush	1.7142857142857142	1.2857142857142858	1.1428571428571428	0.0	1.4285714285714286	1.3333333333333333	0.42857142857142855	1.1666666666666667
become available/rich/a writer, etc.	0.8571428571428571	0.5714285714285714	0.8571428571428571	1.4285714285714286	0.0	1.0	0.8571428571428571	1.0
break the ice	1.0	1.2857142857142858	0.5833333333333334	1.3333333333333333	1.0	0.0	0.8571428571428571	0.85
break the law	1.0	0.8571428571428571	0.5714285714285714	0.42857142857142855	0.8571428571428571	0.8571428571428571	0.0	1.0
bring a lump to your throat	1.4285714285714286	1.4285714285714286	1.1428571428571428	1.1666666666666667	1.0	0.85	1.0	0.0
burn the midnight oil	2.0	2.0	1.2	1.0	1.5	1.0526315789473684	1.4	0.7142857142857143
can't/couldn't help doing sth	1.7142857142857142	1.5714285714285714	1.0	1.0	1.1666666666666667	0.42857142857142855	0.5714285714285714	1.2857142857142858
catch fire	1.0	1.0	0.7142857142857143	1.5714285714285714	1.1428571428571428	0.7142857142857143	0.8	1.0
change your mind	1.1428571428571428	0.14285714285714285	1.0	1.0	0.6428571428571429	1.2857142857142858	0.4	1.1666666666666667
come true	1.0	1.0	1.0	1.4285714285714286	0.5	1.0714285714285714	0.7894736842105263	0.8333333333333334
crack a joke	1.8	1.1428571428571428	1.1428571428571428	0.5714285714285714	1.4285714285714286	1.0	0.42857142857142855	0.8571428571428571
cross your mind	1.0	1.0	1.5	1.0	0.7142857142857143	0.5714285714285714	0.7142857142857143	0.5714285714285714
do the cleaning/cooking, etc.	0.8571428571428571	1.0	1.0	0.5714285714285714	1.0	0.7142857142857143	1.3333333333333333	1.0
draw a conclusion	1.037037037037037	0.5714285714285714	1.0	0.7142857142857143	1.0	1.1428571428571428	0.42857142857142855	1.0
drive sb mad/crazy, etc.	1.0	0.5714285714285714	0.14285714285714285	0.7142857142857143	1.0	0.2857142857142857	1.0	1.0
face the music	1.8571428571428572	1.5714285714285714	1.0	0.6428571428571429	1.1428571428571428	1.0	0.7142857142857143	1.0
fall flat	1.2857142857142858	0.8571428571428571	0.42857142857142855	1.0	1.7142857142857142	0.7142857142857143	0.4	1.0
fall in love	0.8	1.0	0.8571428571428571	1.1428571428571428	0.8571428571428571	1.1428571428571428	1.0	1.4285714285714286
feel at home	0.2	1.1428571428571428	0.14285714285714285	0.7142857142857143	0.6666666666666666	0.8571428571428571	0.7142857142857143	1.3571428571428572
follow suit	1.6428571428571428	1.5714285714285714	1.5714285714285714	1.0	1.2857142857142858	0.5714285714285714	0.8571428571428571	1.0
get a bus/train/taxi, etc.	0.7142857142857143	0.5714285714285714	1.4285714285714286	0.5714285714285714	0.4	1.0	0.5714285714285714	2.0

Visualization

- Tensorflow embedding projector
- <https://tinyurl.com/enetCollectVerbalMWE>
- <https://tinyurl.com/enetCollectAdverbialMWE>



Analysis – core vs outliers

1	MWE ▼	CEFR ▼	average_rank ▼
2	be dead (set) against sth/doing sth	C1	2,849673203
3	thick and fast	C2	2,823129252
4	be poles apart	C2	2,742857143
5	be/run counter to sth	C2	2,676056338
6	behind the times	C1	2,619047619
7	as far as sb is concerned	B2	2,554744526
8	I/you/he, etc. had better do sth	A2	2,538461538
9	go downhill	C2	2,514084507
10	deadly dull/serious, etc.	C1	2,510948905

Graphic representation?

Analysis – core vs outliers

38	by the way	A2	1,814285714
39	after all	B1	1,791366906
40	as many as	C1	1,790697674
41	by accident/mistake, etc.	B1	1,787234043
42	as much/quickly/soon, etc. as possible	A2	1,781690141
43	first of all	A2	1,753333333
44	in front of sb/sth	A2	1,709219858

EVP miss?

How many are a crowd?

- Results with 2, 3, 4, 5 answers show that ratings are pretty similar...

<http://tiny.cc/29ui4y>

Feedback form

Please try to describe your reasoning in deciding on which MWE was the easiest or the most difficult in the MWE experiment.

13 responses

I tend to choose the most fixed expressions as "easiest" expressions, e.g. things like "a lot", "happy birthday"

The MWE that is like idioms and whose elements do not need to change the wordform are the easiest for learning, to my mind.

Since I actually haven't learned English, only acquired by watching television etc., I decided on my knowledge and intuition. If I knew the MWE, I chose them as the easiest and if I had never heard of it before and didn't understand the meaning behind it, I chose it to be the hardest.

Feedback form

Please try to describe your reasoning in deciding on which MWE was the easiest or the most difficult in the MWE experiment.

13 responses

I tend to choose the most fixed expressions as "easiest" expressions, e.g. things like "a lot", "happy birthday"

The MWE that is like idioms and whose elements do not need to change the wordform are the easiest for learning, to my mind.

Since I actually haven't learned English, only acquired by watching television etc., I decided on my knowledge and intuition. If I knew the MWE, I chose them as the easiest and if I had never heard of it before and didn't understand the meaning behind it, I chose it to be the hardest.

I picked a construction as difficult, if I did not know it myself, or judged it as difficult by its syntactic structure (eg. an object behind it, where you would not expect it), or if it requires an elaborate lexical knowledge. I picked as easy all constructions which had an easier syntactic structure or easier words

Expressions with figurative meaning were the most difficult ones, while no-figurative were easier to solve. Another criteria were existence of similar or identical concepts based on metaphor and monotony in my L1

Non-figurative simple expression were easier to solve. However, figurative expressions based on different conceptual metaphors and metonymies than those in my L1 were more challenging.

I think idiomatic expressions are more difficult to learn, and some structural fixed expressions are easier. Among idiomatic expressions, it is easier to learn those with the same metaphors (positive transfer native language knowledge), while idioms and collocations that use different cultural associations and metaphors than in a source language, are more difficult to learn. I based this attitude on my experience from my teaching methodology classes.

As a native English speaker, I have never before heard "grasp the nettle". Therefore, I rated it as the hardest every time I encountered it.

I did try not to overthink my decisions. Sometimes I thought about learning English myself, which MWE I know and when I learned them. Some MWE I didn't know I ranked 'the hardest'.

I also considered the length of the MWE as well as their semantics and whether the semantics of the whole can be deduced from the single parts of MWEs and if so, how easily. For example: The MWE "cut a long story short" seemed easier for me because the meaning can be grasped by a metaphorical expansion of "cut" while this is not possible with the MWE "to burn the midnight oil" (at least not for me).

It was not always easy to decide between the hardest and the easiest MWE, especially when the combinations of the MWEs displayed changed. One MWE I have ranked the hardest in one set of MWE wasn't the hardest MWE when displayed among other MWEs. Often, the decision was thus a relative one and was clearly influenced by the combinations of MWE one had to choose from.

Besides, I have learned new MWEs myself! Thank you! ;)

The easiest MWE is usually taught at beginning levels and has to do with themes that are dealt with in these levels. The most difficult MWE: a) is typically used in certain genres that are usually taught at higher levels; and/or b) is not frequent; and/or c) is fully opaque; and/or d) has more elaborated syntactic structure; and/or e) is used as a subordinator.

Any other comments

I found the experiment too long: I had expected 10 or 20 words. 82 is in my opinion too much for a task which is very monotonous, but at the same time requires some concentration (you cannot do it automatically)

The procedure of answering of questions is easy, but it gets automatic at a certain point. A possibility to do this task in several attempts (save and continue) is a good option to keep people involved. I stopped at the required number of 82 answers (actually did one additional :))

1) One should be reminded explicitly (during the task) that the task is about choosing the easiest or the hardest MWE to "produce". It is stated in the guidelines but I caught myself forgetting this while solving the task, sometimes I thought about 'understanding' the MWE and how difficult this might be for learners. Then again I remembered that the task was about the production of MWEs. For future tasks it might be helpful to have a short description (a reminder) above every task, like "the goal is to rate expressions based on how difficult they are for a language learner to produce" (sentence taken from the guidelines).

2) I wasn't stated explicitly which production mode to assume for the task: written or oral? During the task, I often decided based on the assumption the learner would produce a written text, but some MWE like "see you later" or "Guess what?" seemed more likely to occur in an oral production mode, e.g. in a talk with a friend. So I wasn't sure which mode to consider. As a consequence, I often switched the 'assumed modes' during the task in order to decide on the easiest and hardest MWEs.

Testing

<https://spraakbanken.gu.se/eng/crowdsourcingMWE>

A lot to discuss...

- Is ranking a valid way of viewing vocabulary?
- ...or – more generally - chunking it into level “portions”?
- Is EVP reliable in all respects, to start with?
- Does it matter how we present MWEs?
 - *get here/there/home/to work, etc.*
 - *can't/couldn't help doing sth*
 - *I/you/he, etc. had better do sth*

Same experiment on a different data – possible?

- How to cut into levels groups, then?
 - Would "seed" items with known levels help?
- Use crowd to confirm level labels?
 - i.e. automatic predictions vs crowd votes. But what then?

Insights

- It's (probably) a myth
- ...that crowdsourcing saves time and money
- It is a challenge
 - ...to find and motivate a crowd
 - ...to establish reliability of the results
- But
 - ...it may help cover broader range of participants (compared to 2-3 annotators)
 - ...helps avoid looking for trained teachers eager to annotate all material

Open questions

- How many votes are enough?
- How to identify „unreliable“ voters?
- How to attract and motivate a crowd?
- Difference between native and non-native crowdsourcers?
- What about language learners? At which level can they be used for this experiment?

Conclusion

- Crowdsourcing for generating language learning resources?
 - possible
 - minimal crowdsourcer training
 - results comparable to expert annotations
 - better combine several methods?
- Future work
 - similar experiments for other languages
 - large-scale experiments (> 60 items, and in the „wild“ with an „open call“)

MWEs and WG1 crowdsourcing workshop

- <https://spraakbanken.gu.se/eng/wg1-dec18-gbg>
 - Analysis of the experiment results
 - Assessment – can crowdsourcing be used for language learning resources and materials?
 - Who should be the crowd?
 - How many are a crowd?
 - Is it possible to set up a multi-lingual experiment?
 - Etc-etc

Reliability of tools & algorithms – and data!



Thank you!

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (pp. 265-283).
- Capel, A. (2010). A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.
- Capel, Annette. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3.
- Chinkina, M., & Meurers, D. (2017). Question Generation for Language Learning: From ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 334-344).
- Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.
- Ellis Nick C., Simpson-Vlach Rita , Römer Ute, Brook O'Donnell Matthew and Wulff Stefanie (2016). Learner corpora and formulaic language in second language acquisition research. In: Granger Sylviane , Gilquin Gaëtanelle, Meunier Fanny (eds) , *The Cambridge handbook of learner corpus research*, Cambridge University Press
- Flynn, T. N., & Marley, A. A. (2014). *Best-worst scaling: theory and methods* (Doctoral dissertation, Edward Elgar).
- Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels?: A pilot study of written L2 French. *EuroSLA monographs series 1. European Second Language Association*, p.133--158
- Howe, J. (2006). *The rise of crowdsourcing*. *Wired magazine*, 14(6), 1-4.
- Paquot, Magali & Sylviane Granger (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*. 32: 130–149. doi: 10.1017/S0267190512000098
- Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 56-65).
- Suñer, F. (2018). *The interplay of cross-linguistic differences and context in l2 idiom comprehension*. *Research in Language*.
- Thewissen, J. (2013) Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77-101.
- Ziai, R., Rudzewitz, B., De Kuthy, K., Nuxoll, F., & Meurers, D. (2018, November). Feedback Strategies for Form and Meaning in a Real-life Language Tutoring System. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018* (No. 152, pp. 91-98). Linköping University Electronic Press.