cjvt Center za jezikovne vire in tehnologije

Ranking Multi-Word Expressions by Difficulty

Set-up and Results of the Crowdsourcing Experiment

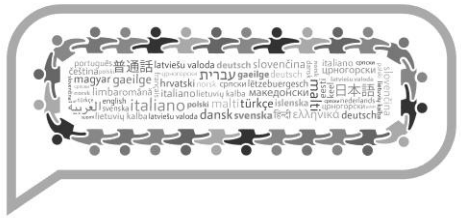
Jaka Čibej

Centre for Language Resources and Technologies, University of Ljubljana

Göteborg, 5 December 2018

Univerza v Ljubljani

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY



Idea

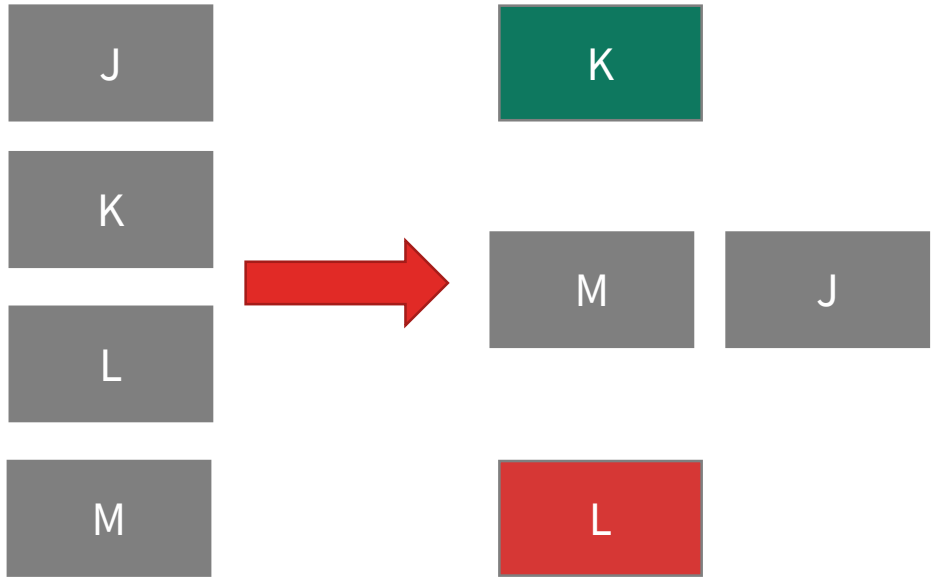
- **Rank a set of expressions by difficulty.**
- **Can this be done through crowdsourcing?**
- How?
- We need a simple task.
- Manageable workload.
- (Relatively) reliable results.

Ranking – How?

- **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?**
- Ranking a (small) subset of tasks, e.g. 4?
  - Again, which combinations?

Best-Worst Scaling

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



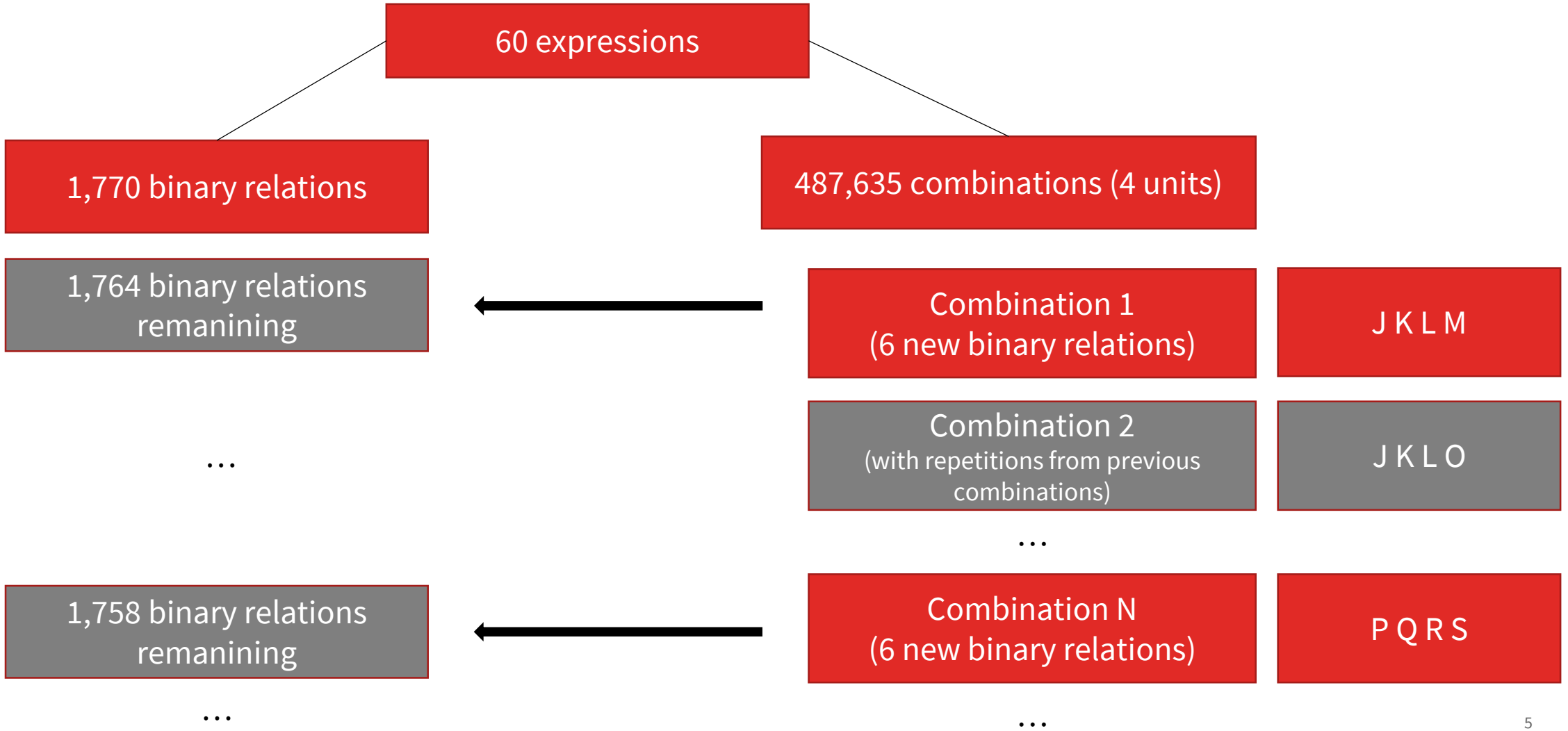
- 6 possible binary relations between the 4 elements
  - J ~ K, J ~ L, J ~ M, K ~ L, K ~ M, L ~ M
- **BWS with 4 elements**
  - K = 3, M = 2, J = 2, L = 1
  - J < K, J > L, J ~ M, K > L, K > M, L < M
  - 5 out of 6 relations (83 %)
  - (at least) 2 clicks
- **Ranking all 4 elements:**
  - 6 out of 6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

Best-Worst Scaling

- 60 expressions
- How many tasks to rank them all?
- Which combinations?
- We based our selection process on the premise that we need the information on all binary relations to make sure the final ranking is as accurate as possible.

- **60 expressions**
  - 1,770 binary relations
  - **487,635 possible combinations of 4 elements**
    - too much work
    - too much **repetition**
    - **no point in collecting information on binary relations multiple times from the same crowdsourcers**

Selecting the Optimal Number of Combinations



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.

Final Set of Tasks

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

PREDICTIONS:

IF:

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

THEN:

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

PyBossa Interface

Easiest	Expression	Hardest
○	a lot	○
○	once upon a time	○
○	as it happens	○
○	deadly dull/serious, etc.	○

Save

as it happens

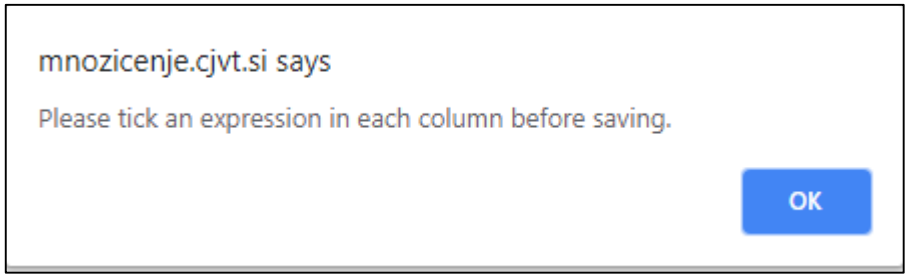
Meaning: something that you say in order to introduce a surprising fact

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.

- **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
  - no ticked expression



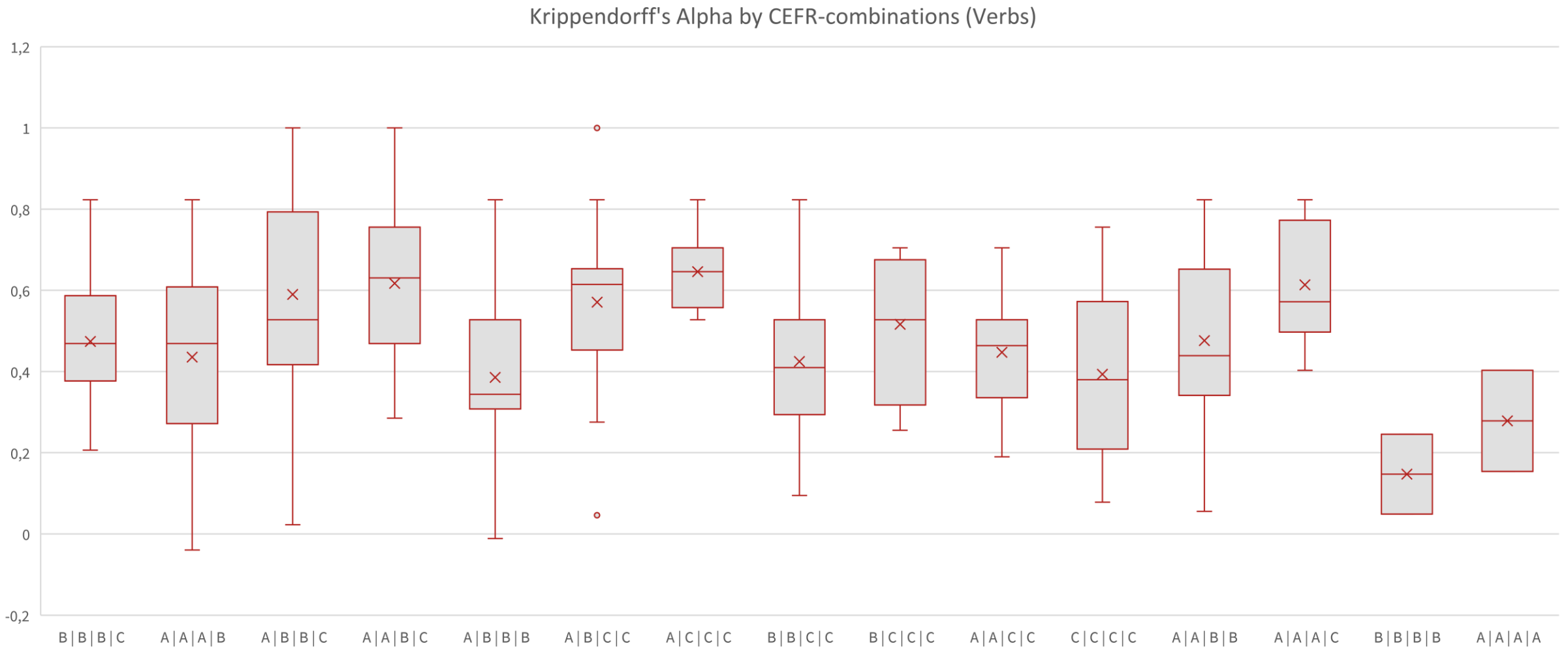
Results – Metadata

- 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**

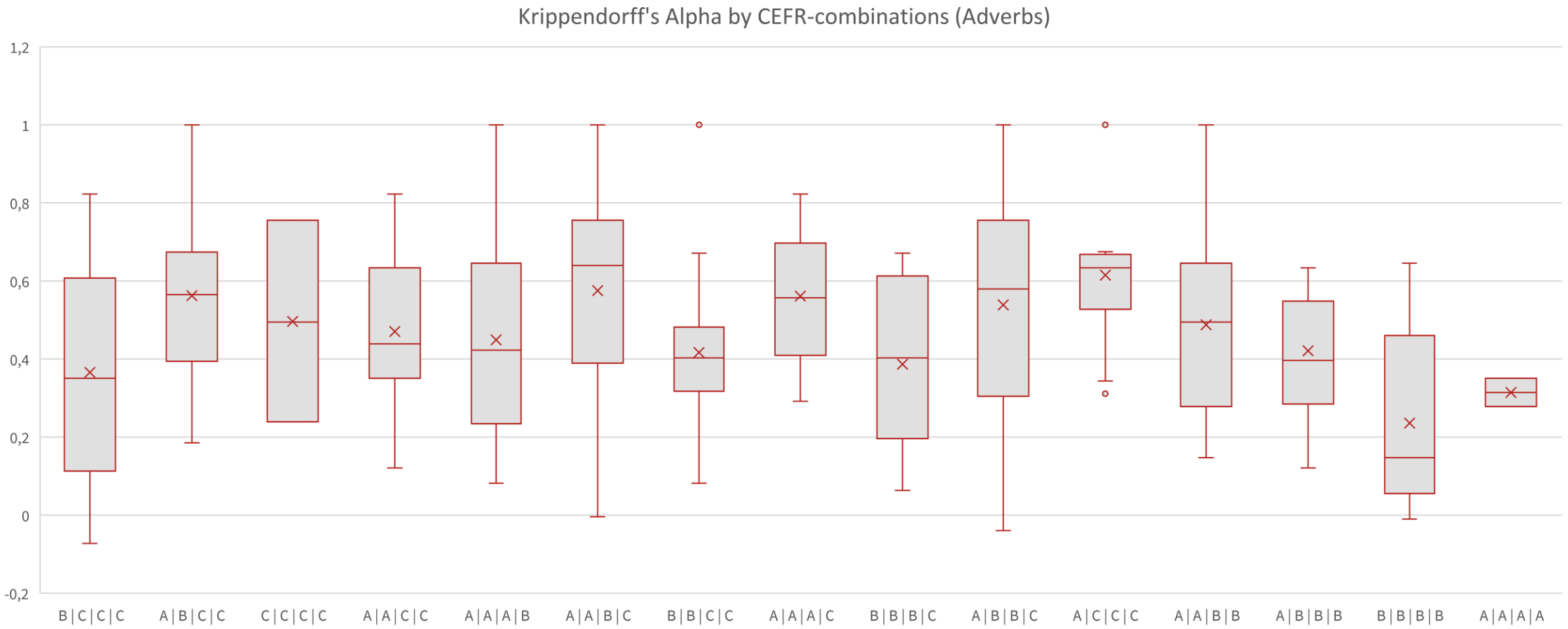
Results – Agreement (Verbs)

- Inter-annotator agreement (Krippendorff‘s Alpha)



Results – Agreement (Adverbs)

- Inter-annotator agreement (Krippendorff‘s Alpha)



Merging the Results

- Collection of ranked expressions obtained in two ways:
  - **Linear scale using average ranks** (presented by Jaka)
  - Clustering and multi-dimensional visualization using **vector embeddings** (presented by David)

- **The Linear Scale approach**
- **Download and view eNetCollect_Linear_Scales-verbs_and_adverbs.xlsx**
- a more brute-force approach
- take all annotations for a specific expression (regardless of the expressions it appears with) and average the sum to get the expression‘s average rank
- the premise: harder/easier expressions should consistently/more frequently be annotated as more difficult (rank 3) or easier (rank 1)

Thank you.

Jaka Čibej
jaka.cibej@cjvt.si

Center za
jezikovne vire
in tehnologije

Večna pot 113
1000 Ljubljana
Slovenija

www.cjvt.si
00386 14798299
info@cjvt.si