MWEs and Crowdsourcing
Outline and Results

Jaka Čibej, David Alfter, Iztok Kosem, Elena Volodina
eNetCollect Annual Meeting
14 March 2019, Lisbon
What and why?

• 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*
• Manual annotation with CEFR-levels (A1 … C2) is time-consuming and difficult
• Is there a simpler way to generate a ranked list?
  • Through crowdsourcing, perhaps?
  • Let’s try and see.
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
English Vocabulary Profile

- 10 expressions per CEFR level
- 60 verbal MWEs
  - to burn the midnight oil
  - it goes without saying
- 60 adverbial MWEs
  - Happy New Year!
  - by all accounts

MWEs and Crowdsourcing: Outline and Results

14 March 2019
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:

  J  K  M  L

- 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!
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).
Predictions

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

• **THEN:**
  • Tasks per crowdsourcer: 82 (per project)
  • Time per crowdsourcer: 0.68 hours, which equals 40.75 minutes
PyBossa Interface

<table>
<thead>
<tr>
<th>Easiest</th>
<th>Expression</th>
<th>Hardest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a lot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>once upon a time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>as it happens</td>
<td></td>
</tr>
<tr>
<td></td>
<td>deadly dull/serious, etc.</td>
<td></td>
</tr>
</tbody>
</table>

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
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

<table>
<thead>
<tr>
<th>MWE</th>
<th>CEFR</th>
<th>average_rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>burn the midnight oil</td>
<td>C2</td>
<td>2,781818182</td>
</tr>
<tr>
<td>go against the grain</td>
<td>C2</td>
<td>2,771428571</td>
</tr>
<tr>
<td>grasp the nettle</td>
<td>C2</td>
<td>2,745454545</td>
</tr>
<tr>
<td>follow suit</td>
<td>C2</td>
<td>2,681818182</td>
</tr>
<tr>
<td>throw in the towel</td>
<td>C2</td>
<td>2,657142857</td>
</tr>
<tr>
<td>beat about/around the bush</td>
<td>C1</td>
<td>2,636363636</td>
</tr>
<tr>
<td>keep sb on their toes</td>
<td>C2</td>
<td>2,63</td>
</tr>
<tr>
<td>nothing ventured, nothing gained</td>
<td>C2</td>
<td>2,608695652</td>
</tr>
<tr>
<td>go from strength to strength</td>
<td>C1</td>
<td>2,59047619</td>
</tr>
<tr>
<td>face the music</td>
<td>C1</td>
<td>2,545454545</td>
</tr>
<tr>
<td>bring a lump to your throat</td>
<td>C2</td>
<td>2,536363636</td>
</tr>
<tr>
<td>have a rest/shower/walk, etc.</td>
<td>A2</td>
<td>1,438095238</td>
</tr>
<tr>
<td>get here/there/home/to work, etc.</td>
<td>A1</td>
<td>1,409090909</td>
</tr>
<tr>
<td>do the cleaning/cooking, etc.</td>
<td>A1</td>
<td>1,4</td>
</tr>
<tr>
<td>get a bus/train/taxi, etc.</td>
<td>A1</td>
<td>1,381818182</td>
</tr>
<tr>
<td>go running/swimming, etc.</td>
<td>A2</td>
<td>1,32</td>
</tr>
<tr>
<td>see you later</td>
<td>A1</td>
<td>1,260869565</td>
</tr>
<tr>
<td>live in/at, etc.</td>
<td>A1</td>
<td>1,217391304</td>
</tr>
<tr>
<td>Excuse me</td>
<td>A1</td>
<td>1,209090909</td>
</tr>
<tr>
<td>go shopping</td>
<td>A1</td>
<td>1,18</td>
</tr>
</tbody>
</table>
Merging the Results

- **Method 2:** Clustering and multi-dimensional visualization using vector embeddings

- 60x60 matrix of average distances between expressions
Clustering

- **Method 1: Linear scale using average ranks**
  - k-means with n = 6
  - 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

<table>
<thead>
<tr>
<th>Assigned ↓</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True -→</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>C1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

Verbal MWEs
Visualization

- Tensorflow embedding projector
- [https://tinyurl.com/enet CollectVerbalMWE](https://tinyurl.com/enet CollectVerbalMWE)
- [https://tinyurl.com/enet CollectAdverbialMWE](https://tinyurl.com/enet CollectAdverbialMWE)
Conclusion

- Crowdsourcing for generating language learning resources?
  - possible
  - minimal crowdsourcer training
  - results comparable to expert annotations

- Future work
  - difference between native and non-native crowdsourcers?
  - what about language learners?
  - similar experiments for other languages
Thank you.

Jaka Čibej  
jaka.cibej@cjvt.si