Expert judgments versus crowdsourcing in ordering multi-word expressions

David Alfter
Språkbanken
University of Gothenburg
Sweden
david.alfter@gu.se

Therese Lindström Tiedemann
Department of Finnish, Finno-ugrian and Scandinavian Languages
University of Helsinki
Finland
therese.lindstromtiedemann@helsinki.fi

Elena Volodina
Språkbanken
University of Gothenburg
Sweden
elena.volodina@gu.se

1 Introduction

Many of the challenges in automatically driven solutions for language learning boil down to the lack of data and resources based on which we can develop language learning materials or train models. Resources like the English Vocabulary Profile (Capel, 2010, 2012; Cambridge University Press, 2015) are a luxury that cost a lot of time and resources to create, and for most languages such resources do not exist. Crowdsourcing has been suggested as a potential method to overcome these challenges. Recently, a European network enetCollect[1] (Lyding et al., 2018) has been initiated to stimulate synergies between language learning research and practice on the one hand, and crowdsourcing on the other. New initiatives have arisen as a result, e.g. using implicitly crowdsourced learner knowledge for language resource creation (Nicolas et al., 2020), crowdsourcing corpus cleaning (Kuhn et al., 2019), development of the Learning and Reading Assistant LARA (Habibi, 2019). However, there are many questions that need to be investigated and answered with regards to methodological issues arising from using crowdsourcing as a method in/for language learning.

In this paper, we raise some methodological questions about crowdsourcing in the context of second language (L2) learning material creation. To go back to the example of the English Vocabulary Profile – could we generate something similar for other languages without involving lexicographers and experts? For example, given a set of some unordered vocabulary items (e.g. phrases), how can we order them by difficulty/complexity and split them into groups appropriate for teaching at different levels of linguistic proficiency? Could a crowd help us in this scenario? Who can be "the crowd" in that case? Are the results reliable?

We focus on whether a crowd of non-expert crowdsourcers can be used to generate language learning materials and how the annotations by experts, L2 professionals such as L2 Swedish teachers, assessors and researchers, i.e. people with formal training in teaching and assessing L2 Swedish irrespective of whether they have Swedish as L1 or L2,[2] compare to the annotations by non-experts, by which we mean L2 Swedish speakers (L2 speakers for short). On a more general note, we investigate whether crowdsourcing as a method can be *reliably* applied to language learning resource building using a mixed crowd.

As an object of study, we investigate a way to order vocabulary by its difficulty for learners, both for academic purposes and for practical use in teaching scenarios. Previously, first language (L1) materials (Kilgarriff et al., 2014; West, 1953; Brezina and Gablasova, 2015) or second language textbooks/essays (François et al., 2014, 2016; Tack et al., 2018) were used to identify vocabulary relevant for learners at different levels of proficiency. We are exploring both a way to validate the previously obtained results, and an alternative way of establishing vocabulary appropriate for different levels of proficiency.

To do that, we use a selection of multi-word expressions (MWE) and ask experts and non-experts to arrange the MWEs by perceived difficulty. The crowdsourcing part of the experiment is designed to test which *intuitions* people have about the relative difficulty of *understanding* MWEs. In this design, we do not expect our participants to have any explicit knowledge about language learning theories, instead relying on their intuitive comparative judgments. Intuitive comparative judgments (including ranking items against each other) has been earlier proven to be easier and more reliable

[1] https://enetcollect.eurac.edu/

[2] By L2 Swedish we mean Swedish as a second (third, fourth, ...) language and as a foreign language

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
L2 speakers-L2 professionals	0.95	0.93	0.92
L2 speakers-CEFR experts	0.93	0.81	0.84
L2 professionals-CEFR experts	0.94	0.85	0.86

Table 1: Agreement between voter groups in the crowdsourcing experiment

than assigning items to a category (e.g. a level of proficiency) (Lesterhuis et al., 2017). We hypothesize that given an unordered list of expressions, using crowdsourcing, we can derive a list ordered by difficulty that can be used in language teaching. We surmise that difficulty and proficiency are correlated, thus one might expect more difficult expressions to be learned at later stages of language development.

2 Experiment

Data For the experiment, we selected 180 multiword expressions (MWEs) split into three different groups: (1) interjections, fixed expressions and idioms; (2) verbal MWEs; and (3) adverbial, adjectival and non-lexical MWEs.[3] Each of the MWE groups consists of 60 expressions, with 12 items for 5 levels of proficiency according to the Common European Framework of Reference (CEFR, Council of Europe (2001), A1–C1 levels), i.e. $12 \cdot 5 = 60$ expressions. The items and their associated CEFR levels come from the COCTAILL corpus (Volodina et al., 2014) where course book texts are marked for the levels of the CEFR courses at which these texts are used. Identification of MWEs and sense disambiguation was performed automatically using the Sparv pipeline (Borin et al., 2016) . The first level of the text where an MWE appears is associated with the MWE (a hypothesized "target level"). Each MWE expression represents one sense according to the Saldo lexicon (Borin et al., 2013). Each MWE is presented in isolation and in its dictionary form. We manually added explanations for each of the 180 MWEs used in the experiment.

Methodology Instead of volunteers annotating each MWE with a target CEFR level (a task that requires in-depth knowledge of the CEFR), and following previous results showing that relative comparative judgments are easier than assigning items to a category (Lesterhuis et al., 2017), we opted to use best-worst scaling (Louviere et al., 2015) for the crowdsourcing task.

In best-worst scaling, one is presented with a group of items to rank and is asked to rank one of the items as the "best" / "easiest" and one as the "worst" / "hardest". Presenting 4 items to the annotator will then result in 5 out of 6 possible relations.

For evaluation, we project the crowdsourced data onto a linear scale by calculating the average score given to expressions. For each group of four items, the "hardest" item will be given a score of 3, the "easiest" item a score of 1 and the two unrated items a score of 2. By averaging these scores for each item and each group it occurs in, we arrive at a linearly ordered scale of items.

Study design We asked L2 speakers and L2 professionals (separately) to perform three crowdsourcing experiments (one per MWE group) set up as best-worst scaling tasks. It was an open call which attracted 27 L2 speakers and 23 L2 professionals to participate in the study. In parallel, we asked three CEFR experts (incl. in the L2 professionals above) to carry out both a crowdsourcing experiment and a direct annotation task, explicitly labeling the 180 items for the target (lowest) CEFR level at which the item can be expected to be understood by an L2 learner.

3 Results

Below, we report (1) agreement between crowdsourcers by professional background; (2) inter-annotator agreement for the direct labeling experiment; (3) intra-annotator agreement for CEFR experts who participated in both the direct labeling and in the crowdsourcing experiment; and (4) the relation between the number of votes and the results.

Agreement between crowdsourcers by professional background Table 1 shows the Spearman rank correlation coefficient for the three sets of MWEs and the three groups of participants on the

[3]For the sake of conciseness and spatial limitations, we refer to group 1 as "interjections", to group 2 as "verbs" and to group 3 as "adverbs".

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
Tolerance 0	15	21	13
Tolerance 1	61	58	65

Table 2: Percentage agreement (%) between annotations of CEFR experts in explicit mode of annotation

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
Expert 1	0.91	0.93	0.89
Expert 2	0.85	0.61	0.73
Expert 3	0.80	0.52	0.55

Table 3: Spearman rank correlation coefficients for intra-annotator agreement for CEFR experts comparing implicit and explicit modes of annotation

linear scale projection: L2 speakers, general L2 professionals (excluding the 3 CEFR experts) and CEFR experts . Spearman rank correlation coefficient has a range from -1 to +1 where -1 indicates a perfect negative correlation; 0 indicates no correlation; and +1 indicates perfect positive correlation.

As can be gathered from Table 1, the highest correlations can be found between non-experts (i.e. L2 speakers) and the general group of "L2 professionals" across all of the three MWE groups, while the correlations between non-experts and "CEFR experts" (i.e. the subgroup of L2 professionals) are the lowest among all the three MWE groups. We can thus say that non-experts and experts (i.e. the general group of L2 professsionals) in our experiment agree very well on the relative difficulty of MWEs, followed by the general group of L2 professionals and CEFR experts, while L2 speakers and CEFR experts tend to agree to a lesser extent. Despite these marginal fluctuations, we can see strong correlations between all of the tested target groups across all the three sets of tested MWEs. This indicates that intuitions about the difficulty of MWEs are more or less shared across all tested groups, despite the differences in background and professional competence. It seems that we can confirm that non-experts (i.e. L2 speakers lacking expertise in a subject (e.g. language assessment)) can be seen as on par with experts for tasks requiring high competence, something that has also been shown in approaches in citizen science (Kullenberg and Kasperowski, 2016).

Inter-annotator agreement for the direct labeling experiment If we look closer at the simple and extended percentage agreement between the CEFR experts in the 'direct' labeling experiment, we can see that agreement is generally quite low for simple agreement (Tolerance 0 in Table 2). With a tolerance of zero, one counts exact agreement between the annotators (e.g. the same item has been assigned to the same CEFR level). However, if one relaxes the tolerance level to 1 (extended percentage agreement), meaning that positive agreement also includes cases where annotators differed by only one level (e.g. one annotator said the item was A2 while another annotator said the item was B1), we can see that agreement drastically improves, as illustrated in Table 2.

In general, this gives us a picture that expert judgments are not ideal and that reaching an exact agreement between them is possibly an unattainable target, which also confirms the results from essay evaluation according to the CEFR-scale (see e.g. Díez-Bedmar (2012)). Given that direct labeling is a subjective and cognitively challenging task, more opinions than one are required (cf Snow et al., 2008; Carlsen, 2012). Furthermore, the MWEs in our experiments are de-contextualized which might further complicate decisions.

This speaks in favor of assuming tolerance level 1 since the assigned levels describe a continuum of proficiency rather than strict categories (Council of Europe, 2018, p. 34). A hypothesis in connection to this is that disagreement outside tolerance 1 may indicate items that are on the periphery of the lower CEFR level, while items within tolerance 1 constitute the core vocabulary on the lower level. This is something to be explored in future research.

Intra-annotator agreement for CEFR experts from both experiment setups Results of agreement between the explicit ranking of each individual CEFR expert and their own individual implicit judgment from the crowdsourcing experiment show mixed results (Table 3). Expert 1 is very consistent in both annotation methods, and all annotators seem to agree with themselves most for MWE group 1. This could indicate that expert 1 is the one with the most experience of working with CEFR-levels. The inconsistency of the results for the same CEFR expert indicates that the expert reasons differently when using different methods, and that the way of reasoning influences the results. It has been

Group	Sample size	Spearman
Non-experts	1	0.96
	2	0.98
Experts	1	0.97
	2	0.99
Mixed	1	0.99
	2	0.99

Table 4: Spearman rank correlation for different groups and sample sizes

previously shown that explicit scoring is more subjective and cognitively demanding than assessing by comparing two samples to each other (Lesterhuis et al., 2017), which also seems to be confirmed in this experiment. This indicates that we should not compare the two types of annotation and that expert judgment can only give reliable annotation if a reasonably large number of experts is used to counter-balance a potential subjective bias. How large a number constitutes a "reasonable amount" is still an open question.

Number of votes We have tested sub-sampling from each crowdsourcer group separately as well as across groups. While two votes produce results quite close to those obtained with three votes, we surmise that a higher number of votes produces more stable results. Table 4 illustrates the average Spearman rank correlation coefficient for non-experts, experts and mixed group when sampling one respectively two votes. The results are compared to the full vote results. For reasons of space, the numbers indicated are averaged over the three different MWE groups; results per MWE group reflect the same tendency. As can be gathered from table 4, the mixed group sampling leads to a higher correlation than per-group sampling.

4 Discussion and concluding remarks

Among the burning questions in emerging crowdsourcing projects within the domain of language learning two methodological questions remain the most important at the moment:
1. Who can be the crowd – with regards to the *background* of crowdsourcers? and
2. How can *reliable* annotations be achieved with regards to design, number of answers and number of contributors?

The biggest gap that we have tried to fill with this study concerns the first (1) question, i.e. whether crowdsourcing as a method in language learning (within a limited domain of L2 resource annotation) could be used without explicit control for the background of the crowd.

Our results convincingly show that non-experts can perform on par with experts. We have seen that crowds with different backgrounds agree very well with each other, in comparison to previous research where CEFR raters of essays have often reached fairly low agreement (Díez-Bedmar, 2012). Note here that these conclusions are true of annotation carried out in a *comparative judgment* or *best-worst scaling* setting whereas previous work on essay rating has been done based on scales (e.g. the CEFR-scale) similar to our direct-labeling experiment . To confirm our findings, similar experiments need to be repeated for other languages, other types of problems (e.g. annotation of texts for difficulty/readability), and other sub-problems of a given problem (e.g. annotation of single vocabulary items for difficulty).

In relation to question (2) the *reliability* of annotations, we have seen how the design of an annotation task influences the results. Clearly, a more traditional method of annotation – using expert judgments – produces less reliable results than crowdsourced comparative judgments/best-worst scaling rankings. We have seen that experts do not agree with themselves when using comparative judgments versus categorical judgments, whereas the comparative judgment setting leads to homogeneous results between all groups of crowdsourcers regardless of their background.

Furthermore, we explored how the number of votes influences the results and we found that with only two votes, the difference in results on a scale 1-60 is insignificant in comparison to three votes. We found that sampling from a mixed-background group tends to produce more stable results.

Future studies could investigate whether the same methodology produces the same results when applied to e.g. single word expressions or essays. Another direction might be how to partition an unordered, unlabeled set of expressions into different proficiency levels based e.g. on clustering results. This might be achieved by adding certain *anchor* expressions to the experiment, i.e. expressions of which one knows with a sufficient degree of certainty their true label (i.e. target level). Further, one might want to investigate how core and peripheral vocabulary can be identified based on different kinds of annotations.

References

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Vaclav Brezina and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1):1–22.

Cambridge University Press. 2015. English Vocabulary Profile. https://www.englishprofile.org/wordlists. Accessed: 2019-11-11.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.

Annette Capel. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

Cecilie Carlsen. 2012. Proficiency level—a fuzzy variable in computer learner corpora. *Applied Linguistics*, 33(2):161–183.

Council of Europe. 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Press Syndicate of the University of Cambridge.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. www.coe.int/lang-cefr. Accessed 18.06.2020.

María Belén Díez-Bedmar. 2012. The Use of the Common European Framework of Reference for Languages to Evaluate Compositions in the English Exam Section of the University Admission Examination. *Revista de Educación*, 357:55–79.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *LREC*, pages 3766–3773.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219.

Hanieh Habibi. 2019. LARA Portal: a Tool for Teachers to Develop Interactive Text Content, an Environment for Students to improve Reading Skill. In *Proceedings of the 12th annual International Conference of Education, Research and Innovation*, pages 8221–8229.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Tanara Zingano Kuhn, Peter Dekker, Branislava Šandrih, Rina Zviel-Girshin, Špela Arhar, and Tanneke Schoonheim Holdt. 2019. Crowdsourcing corpus cleaning for language learning resource development. In *EUROCALL Conference 2019*, page 163.

Christopher Kullenberg and Dick Kasperowski. 2016. What is citizen science?–a scientometric meta-analysis. *PloS one*, 11(1):e0147152.

Marije Lesterhuis, San Verhavert, Liesje Coertjens, Vincent Donche, and Sven De Maeyer. 2017. Comparative judgement as a promising alternative to score competences. In *Innovative practices for higher education assessment and measurement*, pages 119–138. IGI Global.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Verena Lyding, Lionel Nicolas, Branislav Bédi, and Karën Fort. 2018. Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing techniques (enetcollect). *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL*, 2018:176–181.

Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. 2020. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 268–278, Marseille, France. European Language Resources Association.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the Third workshop on NLP for computer-assisted language learning*, pages 128–144.

Michael Philip West. 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green.