slovenščina 2.0, letnik 10, številka 2

Slo2.0

CROWDSOURCING FOR LANGUAGE LEARNING AND LINGUISTIC RESOURCE CREATION

Množičenje v podporo jezikovnemu učenju in razvoju jezikovnih virov

2022_2

Univerza *v Ljubljani*
Filozofska fakulteta

Slovenščina 2.0
Letnik/Volume 10, Številka/Issue 2, 2022

ISSN: 2335-2736

Glavna urednika/Editors-in-Chief
Špela Arhar Holdt, Vojko Gorjanc

Gostujoča urednika tematske številke/Special Issue Guest Editors
Lionel Nicolas, Verena Lyding

Tehnična urednica/Managing Editor
Eva Pori

Prelom/Layout
Aleš Cimprič

Založila/Published by
Založba Univerze v Ljubljani

Za založbo/For the Publisher
Gregor Majdič, rektor Univerze v Ljubljani

Izdala/Issued by
Znanstvena založba Filozofske fakultete Univerze v Ljubljani;
Center za jezikovne vire in tehnologije Univerze v Ljubljani

Za izdajatelja/For the Issuer
Mojca Schlamberger Brezar, dekanja Filozofske fakultete

Publikacija je brezplačna./Publication is free of charge.

Publikacija je dostopna na/Avaliable at: https://journals.uni-lj.si/slovenscina2

Revija izhaja s podporo Javne agencije za raziskovalno dejavnost Republike Slovenije./
This journal is published with the support of the Slovenian Research Agency (ARRS).

Content

EDITORIAL

ARTICLES

Crowdsourcing for language learning and linguistic resource creation

Lionel NICOLAS
Eurac Research, Institute for Applied Linguistics, Bolzano

Verena LYDING
Eurac Research, Institute for Applied Linguistics, Bolzano

The current special issue of the journal *Slovenščina 2.0* focuses on the newly explored combination of crowdsourcing, language learning and linguistic resource creation. It contains five articles and one project report, providing insightful discussions on several aspects of this combination, as well as results which help us understand its versatile potential and the challenges to address in order to better exploit it.

This issue is directly related to the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) and constitutes the first milestone of its follow-up initiative the D4Collect Dariah Working Group. EnetCollect was a large network project funded as a COST Action which started in February 2017 with the objective of creating a research and innovation community to explore the subject. After the end of enetCollect's funding period, a Dariah Working Group called D4Collect was created to keep the community together and continue to foster and coordinate further research on language

Nicolas, L., Lyding, V.: Crowdsourcing for language learning and linguistic resource creation. Slovenščina 2.0, 10(2): 1–4.

1.20 Uvodnik / Editorial

DOI: https://doi.org/10.4312/slo2.0.2022.2.1-4

learning combined with crowdsourcing techniques. This special issue features an international set of 14 authors from 13 different countries (see the author list below), and most of the discussions and results presented are the outcome of efforts and collaborations either initiated or intensified in the context of enetCollect. This issue of the journal was also graciously supported in the review process by another international set of 10 enetCollect members from 10 different countries (see the reviewer list below). As guest editors of this special issue, we would like to thank the authors and reviewers for their contribution and, at the same time, express our gratitude to the main editors of *Slovenščina 2.0* for offering to center the 2022 special edition on the combination of crowdsourcing, language learning and linguistic resource creation.

The first article by Volodina, Alfter and Lindström Tiedemann on the Swedish language explores the concepts of core vocabulary learning, non-expert crowdsourcing, CEFR assignments and comparative judgments. More precisely, it investigates the theoretical and practical issues connected to identifying core vocabulary at different levels of linguistic proficiency using statistical approaches combined with crowdsourcing. At the same time, the authors investigate whether crowdsourcing second language learners' rankings can be used in a comparative judgment setting for assigning CEFR levels to unseen vocabulary.

In the second article by Zingaro Kuhn, Arhar Holdt, Kosem Tiberius, Koppel and Zviel-Girshin, the authors describe the development and use of a game-with-a-purpose (GWAP) to crowdsource a pedagogical corpus of example sentences showcasing different types of problems (sensitive content, offensive language, structural problems) for Dutch, Estonian, Slovene and Brazilian Portuguese. They provide a design based on initial experiments focusing on the crowdsourcing suitability of a GWAP in which players identify and classify problematic sentences, and point out problematic excerpts. They also present the methodology for data preparation in terms of source corpora selection, pedagogically oriented GDEX (Good Dictionary EXamples) configurations, and the creation of lemma lists, with a special focus on common and language-dependent decisions.

The third article by Graën explores the concept of generating language learning exercises from parallel corpora and crowdsourcing the actions of users (both learners and teachers) to improve the quality of the corpora. The article provides a blueprint for such a generation mechanism and details three main challenges to tackle when implementing it. It also discusses the fact that, through triangulation, user actions can be transferred to language pairs other than the original ones if multiparallel corpora are used as a source to generate exercises.

The fourth article by Hatipoğlu, Delibegović Džanić, Gajek and Miloshevska addresses the awareness and popularity of crowdsourcing solutions among language learners before and after the COVID-19 pandemic. More precisely, they show that the changes brought about by COVID-19 to educational systems worldwide noticeably impacted language learners' habits and attitudes towards the use of crowdsourcing materials in Turkey, Bosnia and Herzegovina, the Republic of North Macedonia and Poland. They also discuss how, among other factors, students' reduced interactions with teachers and peers added to their workload, and that the lack of support on the part of institutions led them to take more responsibility for their own learning.

The fifth article, by Gajek, shows how the eTwinning European framework, originally designed to facilitate collaboration among schools in Europe and beyond, can be considered as an extensive educational crowdsourcing activity, and demonstrates that teachers can effectively use crowdsourcing in educational practice. Accordingly, the author undertakes two kinds of analyses: a global analysis of the features of the framework in light of crowdsourcing principles, and a local analysis of a selection of outstanding projects submitted for evaluation for national awards in Poland.

In the final article, Nicolas and Lyding review the enetCollect project itself and examine to what extent it met its network-oriented and research-oriented goals that aimed at nurturing a new research and innovation community in order to foster the long-term exploration of the combination of crowdsourcing and language learning. They also introduce the D4Collect Dariah Group that they created to follow-up on enetCollect.

Author list

- David ALFTER, Université Catholique de Louvain, Belgium
- Špela ARHAR HOLDT, University of Ljubljana, Slovenia
- Nihada DELIBEGOVIĆ DŽANIĆ, University of Tuzla, Bosnia and Herzegovina
- Elżbieta GAJEK, University of Warsaw, Poland
- Johannes GRAËN, University of Zurich, Switzerland
- Çiler HATIPOĞLU, Middle East Technical University, Turkey
- Kristina KOPPEL, Institute of the Estonian Language, Estonia
- Iztok KOSEM, University of Ljubljana, Slovenia
- Therese LINDSTRÖM TIEDEMANN, University of Helsinki, Finland
- Lina MILOSHEVSKA, University of Information Science and Technology, North Macedonia
- Carole TIBERIUS, Institute for Dutch Lexicology, Netherlands
- Elena VOLODINA, University of Gothenburg, Sweden
- Tanara ZINGANO KUHN, University of Coimbra, Portugal
- Rina ZVIEL-GIRSHIN, Ruppin Academic Center, Israël

Reviewer List

- Petra BAGO, University of Zagreb, Croatia
- Andrew CAINES, University of Cambridge, United Kingdom
- Gülşen ERYIĞIT, Istanbul Technical University, Turkey
- Thomas FRANÇOIS, Catholic University of Louvain, Belgium
- Andrea HORBACH, University of Hagen, Germany
- Sviatlana KARPAVA, University of Cyprus, Cyprus
- Ivana KONJIK, Institute for the Serbian Language of SASA, Serbia
- Kristina KOPPEL, Institute of the Estonian Language, Estonia
- Julia OSTANINA-OLSZEWSKA, Pedagogical University of Kraków, Poland
- Edina ŠPAGO-ĆUMURIJA, Dzemal Bijedic University of Mostar, Bosnia and Herzegovina

Crowdsourcing ratings for single lexical items: a core vocabulary perspective

Elena VOLODINA
University of Gothenburg, Sweden

David ALFTER
University of Gothenburg, Sweden; Université Catholique de Louvain, Belgium

Therese LINDSTRÖM TIEDEMANN
University of Helsinki, Finland

In this study, we investigate theoretical and practical issues connected to differentiating between core and peripheral vocabulary at different levels of linguistic proficiency using statistical approaches combined with crowdsourcing. We also investigate whether crowdsourcing second language learners' rankings can be used for assigning levels to unseen vocabulary. The study is performed on Swedish single-word items.

The four hypotheses we examine are: (1) there is core vocabulary for each proficiency level, but this is only true until CEFR level B2 (upper-intermediate); (2) core vocabulary shows more systematicity in its behavior and usage, whereas peripheral items have more idiosyncratic behavior; (3) given that we have truly core items (aka anchor items) for each level, we can place any new unseen item in relation to the identified core items by using a series of comparative judgment tasks, this way assigning a "target" level for a previously unseen item; and (4) non-experts will perform on par with experts

Volodina, E., Alfter, D., Lindström Tiedemann, T.: Crowdsourcing ratings for single lexical items: a core vocabulary perspective. Slovenščina 2.0, 10(2): 5–61.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.5-61

in a comparative judgment setting. The hypotheses have been largely confirmed: In relation to (1) and (2), our results show that there seems to be some systematicity in core vocabulary for early to mid-levels (A1-B1) while we find less systematicity for higher levels (B2-C1). In relation to (3), we suggest crowdsourcing word rankings using comparative judgment with known anchor words as a method to assign a "target" level to unseen words. With regard to (4), we confirm the previous findings that non-experts, in our case language learners, can be effectively used for the linguistic annotation tasks in a comparative judgment setting.

Keywords: core vocabulary and language learning, non-expert crowdsourcing, single lexical items, CEFR levels, comparative judgment

1 Introduction

We set out to explore two broader questions in this study, both in the context of second language acquisition: The *first question* concerns theoretical and practical issues connected to differentiating between core and peripheral vocabulary at different levels of linguistic proficiency – that is, which vocabulary is critical for learners to know at a particular level (i.e. learners *need to know* it) versus which vocabulary is *good to know*. In other words, is there common core vocabulary for learners at different levels, and does it behave differently from peripheral vocabulary? In connection to this, we apply statistics and crowdsourcing to examine whether there are any particular word behavior patterns that can help us differentiate between core and peripheral vocabulary, which we study through hypotheses 1–3, as introduced in Section 3.2.

The *second question* concerns theoretical and practical aspects of using second language learners as crowdsourcers for the task of linguistic annotation. In particular, can we use second language learners to rank vocabulary according to difficulty? We experiment with crowdsourcing as a method to identify the receptive proficiency level of previously unseen vocabulary items (henceforth called *unknown items*) in relation to confirmed core (and peripheral) items, and compare teachers' and learners' votes (hypothesis 4 in Section 3.2).

In essence, we ask the following overarching question: Can we use crowdsourcing to identify core and peripheral vocabulary for a certain level? The study is partly motivated by the practical need to classify unseen vocabulary by target proficiency levels as necessary input for the automatic generation of learning materials, and/or for automatic assessment of learner production. We start from a simple assumption that if we ask crowdsourcers to explicitly compare two items at a time, of which one is *core* (with a confirmed level) and the other is a new item (i.e. with an unknown level), then the latter will end up having a rank close to the core items of the level of proficiency which it belongs to. Thus, if we have good *anchor words* (i.e. established core words per level), *unknown* words should appear in relative proximity to the anchor words of the corresponding level after a round of comparisons and votes (Example 1). The current study is designed to investigate how true this assumption is.

Example 1: *Illustration of relative ranking of an unknown item*

For example, given the following nouns with known “target” levels (core/anchor items)

`A1 – party; A2 – view; B1 – variety; B2 – purchase; C1 – reliability`

we need to place the noun `pillar` relative to the vocabulary above.

To do that, we compare `pillar` to each of the words/or groups of words (i.e. Is `party` more difficult than `pillar` or vice versa? Is `view` more difficult than `pillar` or vice versa?) and collect votes from crowdsourcers. Based on the votes, we assign “difficulty scores” to each item in each comparison task. After collecting three to five votes for each possible mini-task, we can see where the collected scores point us. For example, in this hypothetical case it might have pointed to the proximity of scores between `reliability` and `pillar`, and hence the appropriateness of `pillar` at C1 level.

Two broader theoretical questions arise in connection with such an endeavor. One is the well-known issue of *what core vocabulary actually is* (e.g. Stein, 2017; Carter, 1982). The other is a relatively new topic connected to the *reliability of crowdsourcing non-expert judgments* as a method of producing linguistic annotations (e.g. Paquot et al., 2022; Alfter et al., 2021).

In short, we assume that **core** vocabulary at a certain proficiency level is *vocabulary known by all learners of* ***that*** *target language at* ***that*** *particular level*. In our current experiment we focus on items known receptively, i.e. items which can be understood but the learners do not need to be able to use them productively yet. We focus on lexical word classes (nouns, verbs, adjectives and adverbs) and further assume that all items that do not belong to the "core" vocabulary at a particular level but occur in texts aimed at learners of these levels are **peripheral** vocabulary (i.e. good-to-know).

The design of the study is inherited from Alfter et al. (2021), where best-worst scaling was used to crowdsource the relative difficulty of multiword expressions and compare annotations from second language professionals (*experts*), on the one hand, and from second language learners (*non-experts*), on the other. While the main focus of the study by Alfter et al. (2021) was to see how the design of a crowdsourcing task may influence the reliability of the linguistic annotation by experts and non-experts, the main task of the current study is to see whether anchor words (single lexical items) per level will be ranked consistently close together, and thus may serve as anchors to derive the levels of unseen words. If confirmed, such a property can be exploited by other languages for the (inexpensive) creation of similar resources. The secondary task of the current study is to confirm findings by Alfter et al. (2021) about experts and non-experts being able to produce comparable annotations in comparative judgment settings, this time tested on *single lexical items* instead of multiword expressions.

The study is performed on Swedish, but the methodology presented here is applicable to any language. In Section 2 we start with a short note on the notion of core vocabulary, how it has been applied to language learning, as well as a short introduction to crowdsourcing non-expert judgments. Sections 3, 4 and 5 introduce the experimental setup, item selection and practical issues. The results and analyses are presented in Section 6, followed by the discussion and conclusions in Sections 7 and 8.

2 Related work

In this section, we present some of the earlier work related to the focus the current study. There are three main axes that we explore: core vocabulary from a theoretical perspective (2.1), core vocabularies for language learning (2.2), and crowdsourcing linguistic annotations using non-experts (2.3).

2.1 Core vocabulary – a theoretical perspective

Core vocabulary may be assumed to comprise lexical items that are known to all users of a language and thus form a shared vocabulary that all users would be able to use and understand, which echoes the *Basic Language Cognition* theory of Hulstijn (2019). Therefore, it is useful, both theoretically and practically, to understand what makes a lexical item a core item and which properties are characteristic of these.

Several paradigms have been proposed for testing language vocabulary for lexical coreness, e.g. Lehmann (1991) or Bell (2013). Carter (1982) lists the following properties of core vocabulary (presented here in a significantly shortened form):

- Collocational span, i.e. core items will collocate with a wide number of other items, e.g. *fat* book, *fat* cat, etc.
- Semantic neutrality, i.e. core vocabulary will exhibit less stylistically colored and/or less specific meaning than other items with shared semantics, e.g. *thin* versus *skinny*, *undersized*, *scraggy*.
- Definitional power, i.e. core vocabulary tends to be used to explain other vocabulary, e.g. *smile* being used to explain *grin*, *smirk*, *beam*. Here core vocabulary will enter syntactic constructions to explain non-core items, e.g. *non-core noun (individual)=adjective+core noun (a single person);* e.g. *non-core verb (stroll)=core verb+adverbial (walk in a relaxed way),* etc.
- High placement in semantic networks, i.e. core vocabulary items tend to be hypernyms to a number of hyponyms, e.g. *flower* to *tulip*, *rose*, etc.
- Antonymy, i.e. core vocabulary often has an antonymous counterpart, which is less common in non-core vocabulary.

- A cognitive basis reflecting the (semantic and sociolinguistic) norms of the usage, i.e. more normative (unmarked) use is characteristic of core vocabulary, e.g. male species versus female: *lion* vs *lioness*.

The above criteria suggest that core vocabulary is *useful* in many fields, and the concept has been researched and applied in fields like lexicography (West, 1953; Brezina and Gablasova, 2015), language learning (Carter, 1987), comparative historical linguistics (Swadesh, 1971), diachronic lexicostatistics (Márquez, 2007), speech pathology (Crosbie et al., 2006) and other areas. Stein (2017, p.760) argues that *usefulness* is "a *function* of core vocabulary" and not vice versa (i.e. not all useful vocabulary can qualify to be part of core vocabulary). Similarly, the *high frequency* of the core vocabulary is a reflection of the *usefulness* of core vocabulary. Stein (2017) warns against using frequency and usefulness as defining characteristics of core vocabulary. These properties may be used as a proxy for identifying core items, but one needs to keep in mind that not all frequent or useful items belong to the core vocabulary; and not all core items are equally frequent or equally useful (cf. *zip-code*, *bread* or *toothbrush*).

Besides, lexis shows resistance to systematization (Carter, 1982), which implies fuzziness in the definition of the core vocabulary. Some items could exhibit two of the six properties above, and yet be considered core, while others may exhibit all six, altogether leading to different degrees of coreness. Dixon (1971) also claimed that adjectives and adverbs are harder to categorize in this respect.

2.2 Core vocabularies for language learning

Numerous attempts to identify the core, or common, vocabulary for language learners have been made, some prominent examples being the General Service List (West, 1953; Brezina and Gablasova, 2015), the English Vocabulary Profile (Capel, 2015), the Routledge series[1] of most frequent core vocabulary for learners (e.g. Familiar, 2021; Lonsdale and Le Bras, 2009), and a series of Kelly lists for several languages

[1] https://www.routledge.com/Routledge-Frequency-Dictionaries/book-series/RFD?a=1

(Kilgarriff et al., 2014). Strategies for the selection of lexical items for inclusion have been different in different resources – *from* strict frequency indications based on various types of corpora *to* combinations of intuitions, judgments of importance, frequency indications and overlaps between concepts in the different languages (see Kilgarriff et al. (2014) for the latter). All of the lists claim that the identified vocabulary is useful for learners. The connection between the *objective token frequency* and the notion of *usefulness*, however, is not always clear (cf. Stein, 2017). Nonetheless, even though such lists will never be beyond criticism at a theoretical level, they make it possible, with a certain degree of objectivity, to address some central assumptions about vocabulary and its hypothetical importance to language as a system and to language learners in particular.

Several lists have been compiled for Swedish with language learners in mind, such as SVALex (François et al., 2016), SweLLex (Volodina et al., 2016), NyLLex (Holmer and Rennes, 2022), the Kelly-list (Volodina and Johansson Kokkinakis, 2012a, 2012b) and SweVoc (Mühlenbock and Johansson Kokkinakis, 2012), each of which has been compiled on different corpora. The unifying lexical unit for these lists is the *lemgram*,[2] i.e. a combination of lemma, its part-of-speech and its inflectional paradigm, e.g. lemgrams *can, verb* (can-could) and *can, verb* (can-canned) will have two separate entries in a list.

Holmer and Rennes (2022) compared two lists, SVALex and SweVoc (both generated from reading materials), with NyLLex (also based on reading comprehension texts) for overlaps, and identified that they have approximately 52–68% overlap. There was a 40% overlap between SweLLex (based on learner essays) and NyLLex. This suggests that the overlapping 40–50–60% of vocabulary definitely belongs to the core vocabulary that is useful for learners. This is not to say that other, non-overlapping, items do not belong to core vocabulary. In the non-overlapping cases, there are other characteristics that would qualify vocabulary to be included in the core, as outlined in Section 2.1. Holmer and Rennes (2022) further correlated the indications of proficiency levels in SVALex, where CEFR level indications are inherited

[2] For better readability we use the shortened term lemma in the rest of the article to refer to lemgrams.

from the texts used for teaching at these levels, and the readability levels (1–6) used in the NyLLex resource, identifying that approximately 20% of the vocabulary items per level overlap at exactly the same levels (i.e. CEFR A1 in SVALex with Level 1 in NyLLex).

Lexical resources like the ones described here are very valuable for language teaching and for the development of teaching materials. However, there will always be some items that have not been included in the lists, or have not been marked for appropriateness at certain levels of proficiency (or readability). Teachers, test developers, and assessors alike will thus need a method that would allow them to place new lexical items in relation to the items on the list. We are experimenting with ways to address this issue in this study, where we use experts and non-experts to classify unseen (unknown) vocabulary in relation to items of known levels in a crowdsourcing experiment.

2.3 Crowdsourcing linguistic annotation from experts versus non-experts

Kullenberg and Kasperowski (2016) have shown that use of non-experts for scientific projects has been increasing drastically since 2010, primarily in the fields related to natural sciences and medicine. Their analysis demonstrates that non-experts are successfully used for data collection and classification, and are able to perform expert tasks on par with experts. However, the use of non-experts for *linguistic analysis/annotation* is much less researched and continues to pose methodological questions. Below follows a small overview of studies involving crowdsourcing non-expert judgments for linguistic annotation at different levels of linguistic analysis.

Kosem et al. (2018) used a crowd for the task of *sense disambiguation of collocations* in a dictionary project. Their results show that the crowd agreed in 83% of cases, and that the benefits of using a crowd for linguistic annotation are much higher than the costs of employing experts. Lau et al. (2014) employed crowdsourcing for *grammaticality judgments on a sentence level* (binary judgments and gradual ones). The users were filtered through the use of five control items which the authors knew the answers to. Annotations from users who failed to pass the test were not considered in the analysis. The rest of the (filtered) crowd demonstrated consistency in annotations. Unfortunately, the study did not explicitly compare the output from experts and non-experts. De Clercq et al. (2014) designed an experiment involving experts and non-experts for the task of *ranking documents by readability* using crowdsourcing for the non-experts. Experts annotated the documents for readability directly, while non-experts were given a relative ranking task, i.e., determine which one of two texts is more readable. They found that the non-experts and experts agreed to a large extent (with a Pearson correlation coefficient of 0.90).

Similarly, Alfter et al. (2021) and Lindström Tiedemann et al. (2022) compared the judgments of experts and non-experts on the task of *ranking* Swedish *multiword expressions by difficulty*, explicitly studying the reliability of second language learners (non-experts) as annotators. The experts performed a direct annotation with CEFR levels, in addition to the crowdsourcing experiment, while the non-experts (learners) only participated in the crowdsourcing experiment, in which they were asked to indicate the “easiest” and “hardest” of four multiword expressions, a technique called *best-worst scaling* (Louvière et al., 2015). The study found that experts and non-experts agreed to a large extent (with Pearson correlation coefficients between 0.81 and 0.93). Alfter et al. (2022) adopted the same methodology as in Alfter et al. (2021) – albeit for French, and on *word senses*. They arrived at the same conclusions: non-native speakers (non-experts) and native speakers (experts) largely agree about the difficulty of word senses. This again lines up with previous research investigating the reliability of non-experts in tasks normally requiring expert knowledge. Paquot et al. (2022) set essay assessment into a comparative judgment paradigm, employing both trained assessors (experts) and non-trained academics (non-experts). The results clearly show that the two groups exhibit high similarity in their assessments, thus demonstrating that an untrained crowd can be used reliably for essay assessment tasks.

The short overview presented above demonstrates the use of non-experts for annotation tasks on different linguistic levels: multiword expressions and collocations (Alfter et al., 2021; Kosem et al., 2018), sentences (Lau et al., 2014; Alfter et al., 2022) and texts

(De Clerq et al., 2014; Paquot et al., 2022). A number of these studies have contrasted the use of non-experts and experts, demonstrating that the task design has a crucial impact on the reliability of the annotation results, and in particular that the setting of comparative judgments (e.g. easier–more difficult) yields a high correlation between experts and non-expert annotations. Only one study to date has explicitly tested the use of second language learners for annotation tasks (Alfter et al., 2021), although there were also non-native annotators in Alfter et al. (2022). In the current study we replicate the experimental setting from Alfter et al. (2021), using second language learners alongside second language professionals, but for annotation on the relative difficulty of single lexical items, looking for additional proof to support the findings in Alfter et al. (2021) and Lindström Tiedemann et al. (2022).

3 Methodology and experimental setup

To perform the experiment we need to separate vocabulary items we can observe at each particular CEFR level[3] in our data into core and peripheral (i.e. non-core) vocabulary for that level.

3.1 Core vocabulary for each level of proficiency

The attempts at identifying core vocabulary useful for language learning leads us to asking probably the most intriguing question in connection to this study: **Is there a core vocabulary for each level of proficiency?**

Stubbs (2001, p. 41) defines core words as "…known to all native speakers of the language […] that portion of the vocabulary which speakers could simply not do without". We adapt Stubbs' definition of core vocabulary to our context as follows: **core** vocabulary at a certain proficiency level is *vocabulary known by all learners of* ***that*** *target language at* ***that*** *particular level*. In our current experiment we focus on items known receptively, i.e. items which can be understood, but the learners do not need to be able to use productively yet.

[3] Levels here are represented by the scale used in the Common European Framework of Reference (CEFR, COE 2001), representing 6 levels: A1 (beginner), A2, B1, B2, C1, and C2 (near native).

Given this definition of core vocabulary, we assume that most items from closed (functional) word classes should by default belong to the core vocabulary. Therefore, we focus on lexical word classes which do not demonstrate a similar stability historically, due to the tendency to develop new senses, new collocations, and include new members through borrowings or word formation mechanisms.

Further, we assume that all items that do not belong to the "core" vocabulary at a particular level but occur in texts aimed at learners of these levels, are, by the definition above, **peripheral** vocabulary (i.e. good-to-know) or maybe even incidental (i.e. appearing at a level prematurely due to some "non-pedagogical" reasons or needs, e.g. archaic forms in poetry).

Previous research on core vocabulary indicates that items at proficiency levels above B1 do not belong to the common language core vocabulary. Hulstijn's (2019) theory of *Basic Language Cognition* suggests that all speakers of a certain language, even first language speakers, could manage with the vocabulary and grammatical structures that roughly correspond to what second language learners can be expected to have acquired by the time they complete B1 level. This fact suggests that it might be more challenging to identify items that *all* learners at B2–C2 levels would need to know. The idiosyncrasy of the vocabulary which learners at these levels acquire depends on the interests of the learners, professional specialization, and many other aspects, since at these levels lexical variety, lexical sophistication and specialized vocabulary become the most dominant vocabulary features. In fact, the attitude to core vocabulary becomes explicitly negative in the research on advanced language learning and academic writing, where general purpose language is no longer a focus (e.g. Granger and Larsson, 2021). At this level general vocabulary (often also high-frequency vocabulary) is expected to be replaced by more formal specialized alternatives.

One way or another, we see a strong incentive:

(1) To identify core vocabulary that we may expect all learners at that level to acquire – as an input to the theoretical discussions on the nature of core vocabulary, as well as an input to practical applications within ICALL and assessment. In search of strategies to achieve this,

we work with a set of tools – CEFR-tools – and apply personal judgments to separate vocabulary into core and periphery (see Section 4).

(2) To identify such items among core items for each level that could function as reliable *anchors* or *anchor words* when developing a method for placement of unknown items on a scalar vocabulary list. We prefer in this connection to use the term *anchor* instead of *core* for these items (see Section 4.2). *Anchor* is less charged and avoids the unnecessary confusion between practical exercises like the one we perform in this study and the ongoing theoretical discussions about the nature of core vocabulary in general.

All items that have been observed in our corpora at particular levels but which cannot be classified as core are classified by default as non-core, i.e. peripheral – which optionally may be further classified into more subclasses, e.g. as incidental vocabulary.

Words that have not been observed in our corpora have *unknown* status, and eventually need to be classified into either one of the CEFR levels, or outside the CEFR scope.

3.2 Hypotheses and study overview

Based on the short overview of the related work presented above, we have formulated four *hypotheses* for the current experiment:

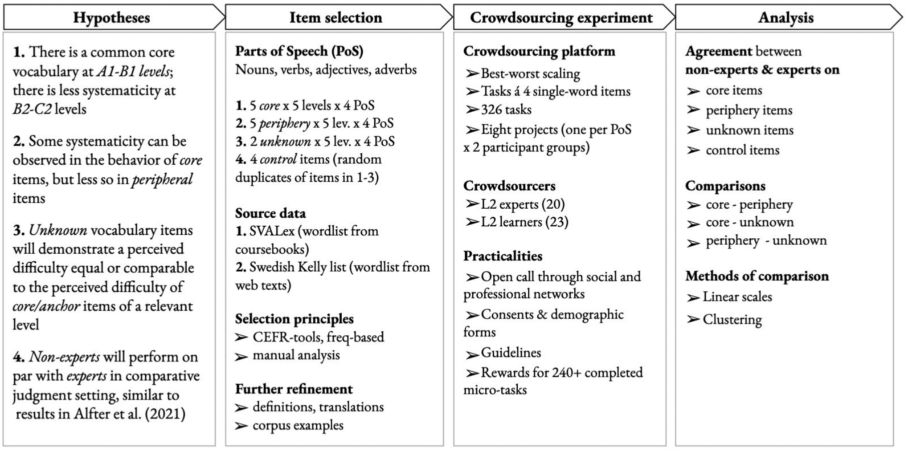
1. There is a common *core* vocabulary at *A1–B1 levels*; there is less systematicity at *B2–C2* levels, a hypothesis that is based on assumptions in the Basic Language Cognition theory of Hulstijn (2019).
2. Some systematicity can be observed in the behavior of the *core* items, but less so in the *peripheral* items.
3. Through crowdsourced comparative judgments, *unknown*[4] vocabulary items will demonstrate a perceived difficulty (expressed in numerical scores) equal or comparable to the perceived difficulty of *anchor* items of a particular level (see Example 1).
4. *Non-experts* will perform on par with *experts* in a comparative judgment setting, similar to the results in Alfter et al. (2021).

[4] *Unknown* in this context means that the item is not represented in the CEFR-graded lexical resource we have at hand.

The overarching procedure for the experiment is straightforward, even though its implementation – technically and theoretically speaking – is challenging, as will become clear from the text that follows:

- From textbooks, select five (5) *core/anchor* items and five (5) *peripheral* ones per level of proficiency (A1–C1).
- From general language corpora, select two (2) *unknown* items per five (5) frequency bands – i.e. items not represented in the textbooks, but that represent different frequency bands (e.g. 1–1,000 most frequent items; 1,001–2,000; etc.) in other resources.
- Mix all item types in tasks for comparison of *perceived difficulty* of items against each other using best-worst scaling (Louviere et al., 2015).
- Collect votes separately for *experts* (second language professionals of the language in question, Swedish in our case) and *non-experts* (second language learners of the language in question, Swedish in our case).
- Analyze the resulting order of items, focusing on the behavior of the *core/anchor* items, *peripheral* items and *unknown* items using linear scales, and clustering as means of visualization.
- Analyze the resulting order comparing *experts* and *non-experts* as annotators, calculating correlations between the two groups.

An overview of the study setup is shown in Figure 1. The sections below expand on each of the steps of the study.



Hypotheses

1. There is a common core vocabulary at *A1-B1 levels*; there is less systematicity at *B2-C2* levels

2. Some systematicity can be observed in the behavior of *core* items, but less so in *peripheral* items

3. *Unknown* vocabulary items will demonstrate a perceived difficulty equal or comparable to the perceived difficulty of *core/anchor* items of a relevant level

4. *Non-experts* will perform on par with *experts* in comparative judgment setting, similar to results in Alfter et al. (2021)

Item selection

Parts of Speech (PoS)
Nouns, verbs, adjectives, adverbs

1. 5 *core* x 5 levels x 4 PoS
2. 5 *periphery* x 5 lev. x 4 PoS
3. 2 *unknown* x 5 lev. x 4 PoS
4. 4 *control* items (random duplicates of items in 1-3)

Source data
1. SVALex (wordlist from coursebooks)
2. Swedish Kelly list (wordlist from web texts)

Selection principles
➢ CEFR-tools, freq-based
➢ manual analysis

Further refinement
➢ definitions, translations
➢ corpus examples

Crowdsourcing experiment

Crowdsourcing platform
➢ Best-worst scaling
➢ Tasks á 4 single-word items
➢ 326 tasks
➢ Eight projects (one per PoS x 2 participant groups)

Crowdsourcers
➢ L2 experts (20)
➢ L2 learners (23)

Practicalities
➢ Open call through social and professional networks
➢ Consents & demographic forms
➢ Guidelines
➢ Rewards for 240+ completed micro-tasks

Analysis

Agreement between **non-experts & experts on**
➢ core items
➢ periphery items
➢ unknown items
➢ control items

Comparisons
➢ core - periphery
➢ core - unknown
➢ periphery - unknown

Methods of comparison
➢ Linear scales
➢ Clustering

Figure 1: Overview of the study.

4 Item selection

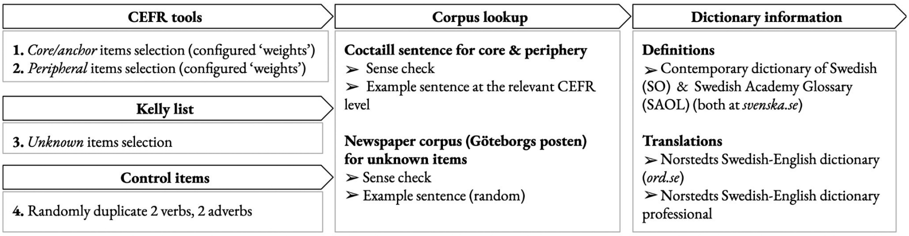

CEFR tools
1. *Core/anchor* items selection (configured 'weights')
2. *Peripheral* items selection (configured 'weights')
Kelly list
3. *Unknown* items selection
Control items
4. Randomly duplicate 2 verbs, 2 adverbs
Corpus lookup
Coctaill sentence for core & periphery
➢ Sense check
➢ Example sentence at the relevant CEFR level
Newspaper corpus (Göteborgs posten) for unknown items
➢ Sense check
➢ Example sentence (random)
Dictionary information
Definitions
➢ Contemporary dictionary of Swedish (SO) & Swedish Academy Glossary (SAOL) (both at *svenska.se*)
Translations
➢ Norstedts Swedish-English dictionary (*ord.se*)
➢ Norstedts Swedish-English dictionary professional

Figure 2: Overview of item selection procedure.

Figure 2 graphically represents the process of item selection for the experiment. For each lexical part-of-speech (PoS) – noun, verb, adjective, adverb – we selected 12 items per CEFR level, split into three different groups:

a) Core/anchor items of a certain CEFR level in the coursebook data in Coctaill (five items) (4.2);
b) Peripheral items of a certain CEFR level in the coursebook data (five items) (4.3);
c) Unknown items which should not appear in the coursebook data at all, with some exceptions (two items) (4.4).

Among the selected items, we also randomly selected two control items in two of the parts-of-speech (i.e. four items in total) to control for the systematicity and reliability of the annotations. These items are duplicates of the already included verbs (*underskatta* 'underestimate'; *förebygga* 'prevent') and adverbs (*således* 'consequently'; *sammanfattningsvis* 'summing up'). The hypothesis with control items is that if the annotations are chaotic, then these items will end up far from each other on the resulting linear scale. If, on the contrary, they end up close to each other, we can assume that the ranking is replicable even with new participants and therefore reliable. We suspected that peripheral items may be more unsystematically annotated and therefore included three control items for periphery and only one for core (*således* 'consequently').

Core/anchor and peripheral lexical items for the experiment were selected from **Coctaill** (Volodina et al., 2014), a corpus of coursebooks for Swedish as a second language published in Sweden between 1997 and 2014, and intended for adult learners. The coursebooks in Coctaill have been linked to CEFR levels with the help of teachers and these levels have been projected to the texts in each book and consequently to lexical items in those texts. This way we can see on which levels lexical items occur in the coursebooks. As a means of filtering the items in Coctaill and helping us select the best candidates for core and periphery we used *CEFR tools*[5] (see 4.1).

The unknown lexical items were picked from the **Swedish Kelly-list** (Volodina and Johansson Kokkinakis, 2012a, 2012b). The Kelly list includes CEFR level indications which are based on frequency bands of approximately 1,500 items (cf. Kilgarriff et al., 2014) and we picked two items per level and PoS from A1–C1.

CEFR proficiency levels focus on communicative abilities and should primarily be thought of as a continuum (COE, 2018, p. 34, cf. Ortega, 2012). Communicative skills can often be achieved through different grammatical and lexical means, and hence it can be difficult to link specific lexical items (single or multiword) to a particular proficiency level. Still, lexical control (COE, 2001, p. 112) and vocabulary size are clearly part of the linguistic competences which a learner has to acquire (COE, 2001, p. 108). The original CEFR publication claimed that detailed lists of vocabulary should be possible to specify for each language (e.g. *Threshold level* 1990) (COE, 2001, p. 30) and encouraged attempts to link communicative tasks to specific vocabulary (COE, 2001, p. 33). The authors note: “Users of the Framework may wish to consider and where appropriate state:

- which lexical elements (fixed expressions and single word forms) the learner will need/be equipped/be required to recognise and/or to use;
- how they are selected and ordered.” (COE, 2001, p. 112).

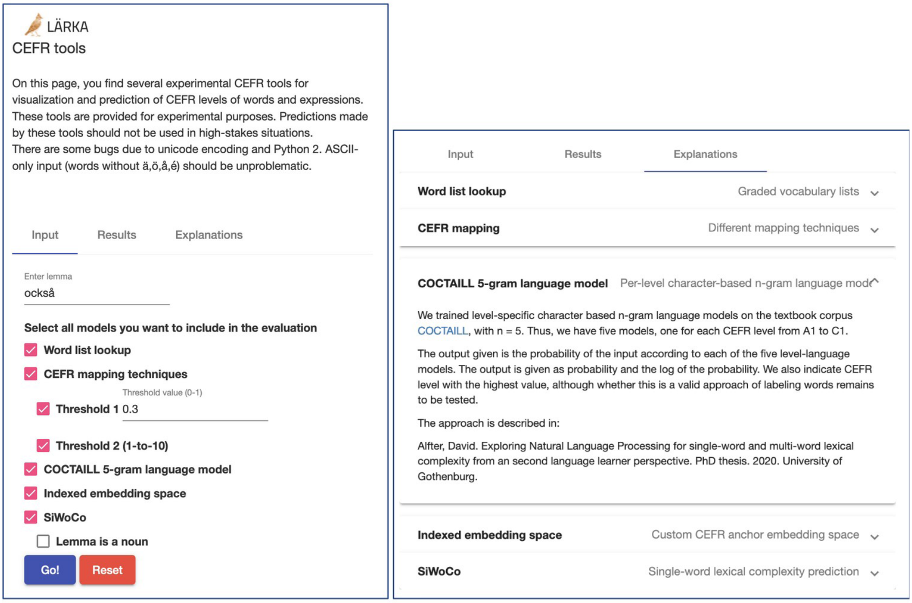
In Alfter et al. (2021) and Lindström Tiedemann et al. (2022) items were linked to their first level of occurrence in the coursebooks, regardless of how many books they appeared in or whether they recurred at later levels. This method of level assignment may be too simplistic,

[5] https://spraakbanken.gu.se/larkalabb/cefrtools (publication under preparation).

which is one of our reasons for investigating both core and periphery vocabulary in this study.

4.1 CEFR Tools

To select items we used output from *CEFR tools* (Alfter, 2021), as illustrated in Figure 3, which shows how lemmas (i.e. base forms) are used at different levels in the coursebook corpus Coctaill (Volodina et al., 2014) and in learner essays (the SweLL pilot corpus, Volodina et al., 2016). It also predicts the level of unseen items with the Coctaill LM-score and the SiWoCo-score (see below for more information).



LÄRKA
CEFR tools

On this page, you find several experimental CEFR tools for visualization and prediction of CEFR levels of words and expressions. These tools are provided for experimental purposes. Predictions made by these tools should not be used in high-stakes situations.
There are some bugs due to unicode encoding and Python 2. ASCII-only input (words without ä,ö,å,é) should be unproblematic.

Input | Results | Explanations

Enter lemma
också

Select all models you want to include in the evaluation

☑ **Word list lookup**
☑ **CEFR mapping techniques**
Threshold value (0-1)
☑ **Threshold 1** 0.3
☑ **Threshold 2 (1-to-10)**
☑ **COCTAILL 5-gram language model**
☑ **Indexed embedding space**
☑ **SiWoCo**
☐ **Lemma is a noun**

Go! | Reset

Input | Results | Explanations

Word list lookup — Graded vocabulary lists

CEFR mapping — Different mapping techniques

COCTAILL 5-gram language model — Per-level character-based n-gram language mod

We trained level-specific character based n-gram language models on the textbook corpus COCTAILL, with n = 5. Thus, we have five models, one for each CEFR level from A1 to C1.

The output given is the probability of the input according to each of the five level-language models. The output is given as probability and the log of the probability. We also indicate CEFR level with the highest value, although whether this is a valid approach of labeling words remains to be tested.

The approach is described in:

Alfter, David. Exploring Natural Language Processing for single-word and multi-word lexical complexity from an second language learner perspective. PhD thesis. 2020. University of Gothenburg.

Indexed embedding space — Custom CEFR anchor embedding space

SiWoCo — Single-word lexical complexity prediction

Figure 3: CEFR tools in Lärka (Alfter et al., 2019).

CEFR tools use various algorithms and techniques to indicate (or predict, depending on the algorithm) CEFR levels for Swedish words, both for known and unknown vocabulary (Alfter, 2021).

Word list lookup returns the level of first occurrence in SVALex (François et al., 2016) for receptive vocabulary and SweLLex (Volodina et al., 2016) for productive vocabulary.

CEFR mapping techniques uses two threshold techniques to derive a level from the underlying distribution across levels, one based on a variable threshold (Threshold 1 in Figure 3, fixed at 0.3[6]; Alfter et al., 2016) and one based on a fixed threshold (Threshold 2 (1-to-10) in Figure 3; Hawkins and Filipovic, 2012). The first threshold technique assigns as the level that at which a word occurs 30% more often that at the previous level. The 1-to-10 threshold technique assigns as level that at which a word occurs at least ten times as often at the previous level. CEFR-mappings do not produce predictions for unseen words, being based on observed frequencies, but may deviate from the first level of occurrence.

COCTAILL 5-gram language model (Coctaill LM for short) uses character-based n-gram language models trained on subparts of the COCTAILL corpus, with one language model per CEFR level. For prediction, each of the language models calculates the probability of the word belonging to the language model and the highest scoring model is used as a prediction. This model can also predict levels for unseen words, i.e. words not included in the coursebook data (*Coctaill*).

Indexed embedding space uses two models that are trained by injecting the CEFR levels as *words* into the embedding space (cf. Alfter et al., 2016, 2021; Wang et al., 2018): first of all, a linear model in which the training data is used as-is, and secondly, a shuffled model, in which the training data was shuffled prior to training. The results show that the shuffled model seems to generalize better.

Finally, *SiWoCo* (Single Word Complexity) automatically extracts numerous word-level features to predict both a receptive and a productive level at which the word should be possible to understand and produce, respectively, and can predict levels for unseen words as well (Alfter and Volodina, 2018).

In addition to the CEFR tool scores, we calculated the following metrics based on the automatic predictions: homogeneity, majority level and percentage agreement.

We define *homogeneity* as a weighted score that takes into account the divergence in levels from the majority level. *The majority level* is defined as the level that most methods agreed upon. In cases of a tie,

[6] The threshold value is fixed at 0.3 (a value which can be adapted) but the underlying frequency distributions are transformed to fit into the interval [0,1], thus in effect this is a variable threshold, even if the value is fixed.

the majority level is taken as the first level (alphabetically). Example: If out of eight predictions, seven methods resulted in a prediction of A1 and one method in A2, the homogeneity would be greater than if seven methods produced A1 and one method produced B1, since the difference from A1 to A2 is smaller than the difference from A1 to B1. The maximum value is 1, while the minimum value is bound by the number of predictions and can be negative (with a minimum of around -1.8 for eight predictors, depending on the predictions). Homogeneity does not contain information about which level the agreement was on, and distances between levels are treated as equal, as the purpose of this measure is to measure the proximity of predictions (i.e. the homogeneity score is the same if seven predictions say A1 and one says B1, or if seven say B2 and one produces C2).

More formally, homogeneity *H* for item *x* with CEFR predictions $p \in P$ is calculated as shown in Equation 1:

$$H(x) = \frac{c(m(P))}{c(P)} - \Sigma_{p \in P, p \neq m(P)} \left(\frac{|p - m(P)|}{c(P)} * c(p)\right) \qquad (1)$$

where *c(P)* is the count of predictions for item *x*, *m(P)* is the majority level in *P* (the first item alphabetically in case of ties), *c(m(P))* is the count of the majority level, and *c(p)* is the number of times *p* was predicted by different methods.

We also use *percentage agreement* (Scott, 1955) as a score to indicate the agreement of the different predictions. An agreement of 1 means that all predictions agree. Note that this score does not contain information about which level the agreement was on, and is calculated as the count of the *majority level* divided by the number of predictions. For example, if out of eight predictions seven methods produce A1 and one method B1 (or for that matter A2, B2 or C1), the agreement would be ⅞.

Finally, we supplemented the data with frequency information from SVALex, namely the relative frequency overall, per CEFR level and the number of documents which contained the item.

4.2 Selecting core items

To pick core items we started by selecting a *receptive majority level* in CEFR-tools (e.g. A1), since we were primarily interested in the receptive proficiency of the learners. We then selected the highest possible level of *agreement* among all items at that level. If this did not give enough items to work with, we added the next agreement level and carried on like this until we had enough items. Below we give some examples to demonstrate this approach, see also Table 1.

(1) *Också* 'also' was picked as an A1 core adverb, and hence the receptive majority level was A1 in CEFR tools. The *agreement score* was filtered to be as high as possible and for this item the receptive agreement score is 0.875 which is basically as high as it is possible to get without the agreement being complete (1). In addition, we note that the receptive *homogeneity score* for this item is 0.75 and hence also very high.

(2) *Ledsen* 'sad' was picked as a core A2 adjective. The receptive majority level was estimated to be A2 in CEFR tools. The agreement score was filtered as high as possible and for this item it is also 0.875, while the homogeneity score is 0.625. Both scores are thus very high, but homogeneity is slightly lower than for the adverb *också*. The item was picked anyway since there was a lack of (appropriate) items with higher homogeneity – higher homogeneity was observed for rather international words that we considered inappropriate for our experiment, such as *modern* 'modern', *obligatorisk* 'obligatory', and *ironisk* 'ironic'.

We tried to keep the level in *Coctaill LM* (Coctaill-based Language Model) and *SiWoCo* (Single Word Complexity prediction model) the same as the level under selection whenever possible. We excluded any items where the scores were more than one level out from the receptive majority level we had selected.

(1) The core item *också* (adverb A1 core) has Coctaill LM A1 and SiWoCo A2 and was hence within the range of what we allowed in these measurements.

(2) The adjectival core item *ledsen* (A2) has a Coctaill LM measure of A2 and SiWoCo A2.

If there were still items with a level prediction from one of the measurements in CEFR tools that was more than one level above the receptive majority level we excluded those as well.

Table 1: *Scores from CEFR tools for the selection methods including four examples, and their corpus frequencies*

CEFR tools and Corpus frequency look up	core range (rule of thumb)	också.adv 'also' A1.core	ledsen.adv 'sad' A2.core	granska.vb 'inspect' B2.per	olycka.nn 'accident' A1.per
Word list lookup					
first occurrence (receptive) SVALex	actual lev.	A1	A2	B2	A1
first occurrence (receptive) SenSVALex	actual lev.	A1	A2	B2	A1
CEFR mapping					
threshold 0.3 (receptive)	actual lev.	A1	A2	C1	B1
threshold 1-to-10 (receptive)	actual lev.	A1	A2	B2	A1
Coctaill language model					
5-gram (prediction)	±1 level	A1	A2	A2	A2
Indexed embedding space					
linear model	±1 level	A1	A2	A2	A2
shuffled model	disregarded	A1	B2	C1	B2
SiWoCo prediction					
receptive (prediction)	actual lev. (±1)	A2	A2	A1	A1
Majority level					
receptive	actual lev.	A1	A2	B2	A1
productive	> actual lev.	A1	A2	C1	B2
Homogeneity					
receptive	0.6 - 1.0	0.75	0.625	-0.75	-0.375
productive	disregarded	0.2	1	-0.2	0.2
Agreement					
receptive	0.7 - 1.0	0.875	0.875	0.375	0.5
SVALex frequency look up					
receptive freq (total-relative)	top freq	32,751,295	714,988	78,403	646,122
receptive freq (total-docs)	top freq	422	18	5	20
receptive freq (level-relative)	top freq	44,397,093	1,556,716	61,296	31,368
receptive freq (level-docs)	top freq	27	5	1	1
Coctaill coursebook inclusion					
# books at the current level (max.≈4/lev.)	> 1 book	3	3	1	1
# books at the next level up (max.≈4/lev.)	≥ 1 book	4	3	2	1

Furthermore, we excluded *productive majority levels* that were lower than the selected receptive majority level, i.e. if we had selected B1 as the receptive majority level the productive majority level based on the SweLL pilot corpus should not be A1 or A2, but could be B1–C2.

(1) For *också* (adverb A1 core) the productive majority level was A1.
(2) *Ledsen* (adjective A2 core) had A2 as the productive majority level.

For adjectives we actually sidestepped our guideline regarding the productive majority level for some core items, and selected some items which have a lower productive majority level than the receptive majority level, e.g. *duktig* 'clever, capable' (A2 core adjective) which has A1 as the productive majority level, most probably due to the word being very frequent in spoken language as a form of praise.

Apart from the more international words mentioned above, *duktig* was the only A2 adjective with relatively high overall scores that suited our needs. It appears in six coursebook documents at the A2 level and on all higher levels, too.

We selected our *core lexical items* based on the items remaining after filtering according to the above principles. If there were still a lot of items to choose from we inspected the frequency information per book. Core items should occur in more than one book, or at least in more than one text. For nouns we tried to make sure that the items were from different topics, and that the items appeared to be reasonably well related with the core topics of that CEFR level according to the CEFR documentation (COE, 2001; 2018). The same was not possible with all PoS, since there were not always so many items per level that we could choose from or no clear topical domain.

(1) *Också* (adverb A1 core) appeared in 27 documents at A1-level and appeared frequently in documents at all other levels in the coursebooks.
(2) The adjectival core A2 item *ledsen* appeared in 5 documents on the A2-level and was present in documents on all levels above A2, and not at all on A1. Three of four books on A2-level contained the word.

After having selected potential core items we checked all items in Coctaill to see how they were used in actual texts in Coctaill.

4.3 Selecting peripheral items

Peripheral items were similarly picked by first choosing the *receptive majority level*, e.g. A1. We then selected that the lemma should ideally only appear in one (or a maximum of two) documents at that level, and we also checked the number of documents on the next level up that used the item. We deselected the current level in Coctaill LM and selected as low *agreement* as possible. Finally, the *productive majority level* should not be the same as the selected receptive majority level, but rather it should be higher in accordance with the assumption that productive proficiency often comes after receptive proficiency. This approach proved very difficult for C1 adverbs, since there were too few items, and most of them have been used as core.

(1) The verb *granska* 'to inspect/check' was picked as a peripheral B2 item. It fulfilled the requirement of appearing only in one document at that level, and also appearing in rather few documents on the next level (four documents). Coctaill LM was A2, and thus different to the level aimed at, and the receptive agreement score was 0.375, and hence very low. Receptive homogeneity was also very low, -0.75. The majority productive level was C1, and hence higher than the level aimed at.

(2) The noun *olycka* 'accident' was picked as a peripheral A1 item. It occurred in only one document at that level, and also appeared in only one document at A2, after which it became slightly more common appearing in five documents at B1-level, seven at B2, and six at C1. Coctaill LM was A2, and hence slightly higher than the level aimed at. The receptive agreement score was 0.5 and receptive homogeneity was -0.375. The majority productive level was B2, and hence clearly higher than A1.

4.4 Selecting unknown items

Finally, we selected two lemmas that were not in the coursebooks, for each level and part-of-speech, from the Kelly list from each frequency band (1–5) associated with the CEFR levels A1–C1. We had trouble choosing adverbs, especially at the A1-level, since most of the adverbs in Kelly were also included in Coctaill. Sometimes we thus made an exception and selected items that were also in Coctaill, but from higher CEFR levels and with very low frequency.

4.5 Translations, definitions and examples

All lemmas were selected with a given part-of-speech and in a certain sense, effectively disambiguating polysemous words. For each item, we selected example sentences from Coctaill (*core* and *periphery*) and from the *Göteborgsposten* corpus[7] (for *unknown* items).

We also provided definitions of the Swedish words, based on two dictionaries: *Svensk ordbok* (SO, Contemporary dictionary of the Swedish Academy) or *Svenska akademiens ordlista* (SAOL, The Swedish Academy Glossary), both available at svenska.se, and translations to English, primarily based on Norstedts Swedish-English online dictionary (ord.se), but supplemented with some additional translations from the dictionary *Norstedts svensk-engelska ordbok professionell* (Norstedts Swedish-English dictionary, professional edition).

Example sentences, definitions and translations were included in the crowdsourcing experiment as an extra feature (if you clicked on one of the items, these would appear).[8]

5 Crowdsourcing experiment

The current study is a replication of Alfter et al. (2021) with regard to the use of best-worst scaling for crowdsourcing linguistic annotation (Louviere et al., 2015). The same number of items per project has been selected (60), and the same redundancy-reducing combinatorial algorithm has been used, resulting in the same number of micro-tasks per project (326). Likewise, we deploy the projects on the pyBossa[9] platform based on an open-source customizable framework for crowdsourcing tasks developed by SciFabric. We apply the same strategy to convert votes from the crowdsourcers into linear scales for further exploration.

Three things differ: Unlike the previous study, we (1) focus on the ranking of single items (as opposed to multiword expressions in

[7] *Göteborgsposten* – a newspaper published in Gothenburg.

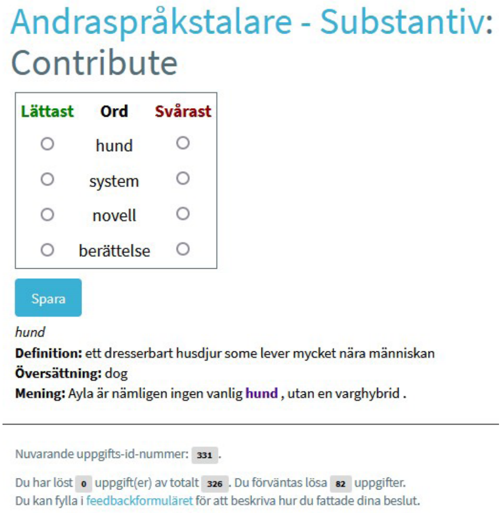
[8] All of the corpora are available through Korp (Borin et al., 2012).

[9] https://pybossa.com

Alfter et al., 2021), (2) we investigate the behavior of core, peripheral and unknown vocabulary (as opposed to the focus on the effects of design of an annotation task), and (3) we explore clustering as a method for visualizing and disentangling the results of the linear scale approach.

5.1 Practicalities and implementation

We implemented the *best-worst scaling projects* on pyBossa, a crowdsourcing platform. For each participant group (expert and non-expert) we set up separate projects for nouns, verbs, adjectives and adverbs, in total eight pyBossa projects (4 x 2 groups). Each project contained 326 micro-tasks (see Figure 4 for an example of a micro-task). Each micro-task contained four single lexical items and a possibility to mark one of them as the easiest (to the left) and one of them as the most difficult (to the right). Clicking on an item would open a field below the task showing the lexical item's definition, translation and an example of its use in a sentence from Coctaill (core and periphery) or from the *Göteborgsposten* newspaper corpus (unknown). The participants could see how many tasks they had completed, as well as open a feedback form and leave a message for us.



Andraspråkstalare - Substantiv: Contribute

Lättast	Ord	Svårast
○	hund	○
○	system	○
○	novell	○
○	berättelse	○

hund

Definition: ett dresserbart husdjur some lever mycket nära människan

Översättning: dog

Mening: Ayla är nämligen ingen vanlig **hund** , utan en varghybrid .

Nuvarande uppgifts-id-nummer: 331 .

Du har löst 0 uppgift(er) av totalt 326 . Du förväntas lösa 82 uppgifter.
Du kan fylla i feedbackformuläret för att beskriva hur du fattade dina beslut.

Figure 4: Example of a micro-task for nouns in pyBossa.

Following the traditions of crowdsourcing, we issued an *open call for participation* (Fort, 2016). All participants were recruited using our professional and social networks. A small *reward*[10] was promised to any participant who completed at least 240 micro-tasks, which was estimated to take a maximum of two hours, with an estimated 30 seconds per micro-task, and in fact took 2 hours on average.[11] Our intention was to collect at least three votes per micro-task from L2 learners (from now on 'non-experts') and three votes per micro-task from L2 professionals (i.e. L2 experts), in accordance with the findings in Alfter et al. (2021).

Before starting the projects, participants were asked to give their *consent*[12] to collect some *demographic information* about them. The latter included information about their gender, year of birth, country of residence, highest education level, native language(s), self-assessed level in Swedish, and an email for linking their pyBossa accounts with the demographic profiles as well as for further contact. For those who marked 'L2 professional' (from now on 'experts'), an additional question was asked about teaching experience counted in the number of years and level/type of teaching (elementary school, high school, Swedish for adults, etc). The demographic information was necessary to separate the participants into experts and non-experts and thus pursue our research interests (hypothesis 4).

On completion of the form, participants received an email with links to the relevant pyBossa projects and *guidelines*[13] in Swedish. Swedish was used in the guidelines as a way to filter participants with insufficient knowledge of the language (we aimed at B1 or more advanced speakers of Swedish). The guidelines included information on the purpose of the experiment, instructions on how to c*reate a pyBossa account*, details about the four part-of-speech-based *pyBossa projects* and explanations of *how to complete micro-tasks* (see Table 2 for the exact formulation of the task).

[10] All participants who completed 240 micro-tasks received a digital voucher for the Amazon online store.

[11] Based on the average time per task as detailed in Section 6.4. However, it should be noted that not every participant completed 240 tasks and that the time per task contains outlier values which skew the actual values.

[12] Consents and socio-demographic information were collected via an online form,

[13] Guidelines: https://docs.google.com/document/d/1gROsxmo4UPoe-bOPKYKJ6Z58tYKMrSwnJgt-Vn1GHL0/edit?usp=sharing

Table 2: *Excerpt of Guidelines with the definition of the task*

Guidelines (in Swedish)	Translation into English
3.1 Beskrivning av uppgifterna [...] Du får se fyra (4) ord i taget, och din uppgift är att markera vilket ord som är **svårast att förstå** av de fyra, och vilket som är **lättast att förstå** av dessa fyra (relativ svårighetsgrad). Med "förstå" menar vi att *kunna förstå ord i en text som man läser på egen hand*. [...]	**3.1 Description of the task** [...] You will see four (4) words at a time, and your task is to mark which word is **the most difficult to understand** out of the four, and which one is **the easiest to understand** (relative difficulty). By "understand" we mean *to be able to understand the word in a text that you read on your own*. [...]
Efter vi har samlat in röster (rankningar) från flera deltagare, kan vi analysera ifall **intuitionerna** om ordens svårighetsgrad stämmer mellan andraspråkstalare och lärare / forskare. Du behöver alltså inte fundera mycket på varför du ser ett ord som lättare eller svårare än ett annat utan använd din intuition framför allt. Men om det är något speciellt som du tycker spelar in i din bedömning så får du gärna kommentera det i feedbackformuläret. Du kan lämna återkoppling via formuläret flera gånger. Det är anonymt.	After collecting votes (rankings) from several participants, we can analyze whether **intuitions** about the difficulty of the words coincide between second language speakers and teachers/researchers. You need not think a lot why you see a word as easier or more difficult compared to another, instead please primarily use your intuition. If you feel there is something that influences your judgments, feel free to comment in the feedback form. You can leave comments several times. It is anonymous.

The projects were open for a month, during which we successfully reached the desired *number of votes* for each of the eight projects.

5.2 Demographic information

A total of 43 participants were recruited through the open call, of those 23 were non-experts ('L2 learners') and 20 experts. Tables 3a and 3b (and the graphs in Appendix 1 for better visualization) present the detailed demographic statistics. One learner left all fields blank, and therefore the total counts in the 'L2 learner' column add up to 22 for all rows.

We can see that women were far better represented in both groups than men, as were university-level or higher educated participants. Among learners, Finnish, Dutch and English were the most represented first languages, but even other languages, such as French, German, Polish, Russian and Ukrainian occurred. For L2 experts, we have a majority with Swedish as their first language, followed by Finnish, and a few other languages, including Bosnian, Dutch, English, German. The participants reported several countries of residence, including Belgium, Finland, France, Germany, Sweden and the UK.

The presence of Finland as a country of residence and Finnish as a mother tongue is not surprising: Swedish is an official language in Finland, is an obligatory school subject for all pupils (either as an L1 or L2), and is required in some occupations. This explains the number of both L2 learners and L2 experts with Finnish as their first language, and the number of residents in Finland. We were positively surprised to see representatives from other countries than Sweden and Finland, and L2 experts who have mother tongues other than Swedish or Finnish.

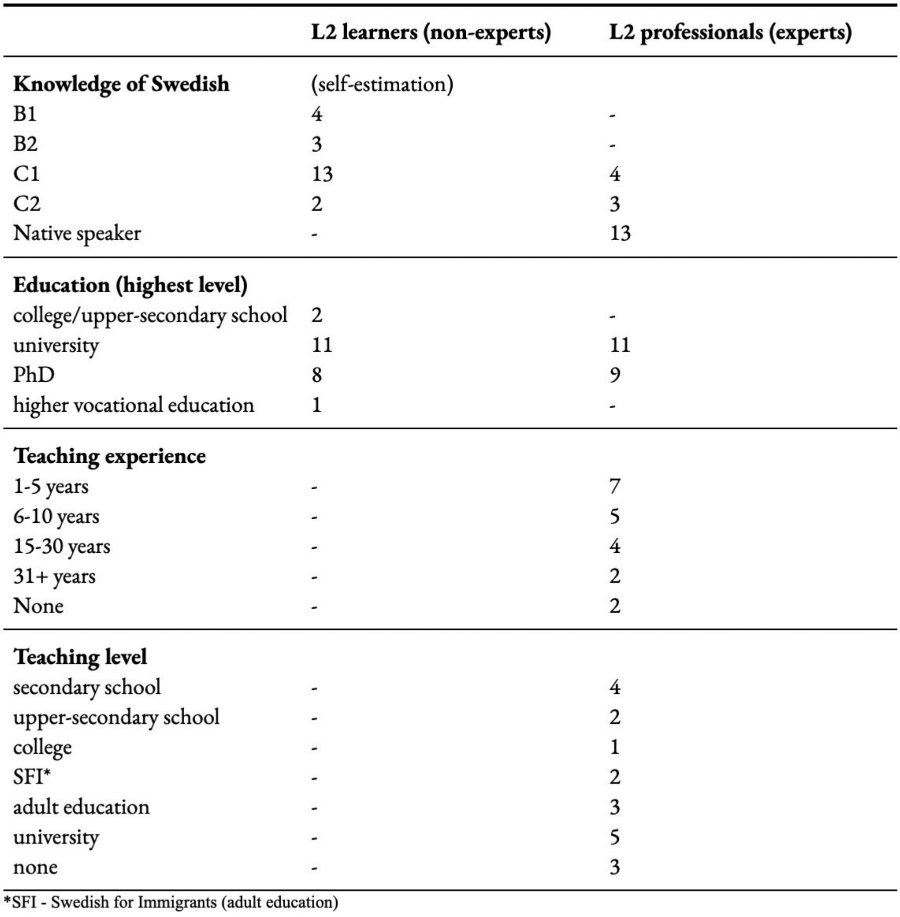
As we intended, L2 learners below B1 did not participate, but those at advanced levels were well represented (15 participants out of 23). The levels are assessed by the learners themselves based on their experience and our short explanations. We thus need to keep in mind that the votes provided by the non-experts in this study come from advanced language users, and the results might, potentially, differ if we had a majority of non-experts at B1 level. Language experts are predominantly native speakers of Swedish, but also include seven participants indicating that their level is C1 or C2.

The majority of the L2 experts indicated that they are in one way or another involved with teaching Swedish proficiency courses as well, and thus can be assumed to understand what makes vocabulary relevant or difficult for learners. More than half of them teach Swedish proficiency courses for adults, and one third of them to children at secondary school level.

Table 3a: *Demographic information about the participants*

	L2 learners (non-experts)	L2 professionals (experts)
Total	23	20
Gender		
male	9	3
female	12	17
other	1	-
Age		
... -20	1	-
21-30	8	5
31-40	5	6
41-50	6	2
51-60	1	3
60- ...	1	4
Mother tongues (L1)		
Bosnian/Swedish	-	1
Dutch	4	1
Dutch/Flemish	1	-
English	4	1
Finnish	7	4
French	1	-
German	1	1
Polish	1	-
Russian	2	-
Swedish	-	12
Ukrainian/Russian	1	-
Country of residence		
Belgium	4	2
Finland	8	6
France	-	1
Germany	1	1
Sweden	8	10
UK	1	-

Table 3b: *Demographic information about the participants*



	L2 learners (non-experts)	L2 professionals (experts)
Knowledge of Swedish	(self-estimation)	
B1	4	-
B2	3	-
C1	13	4
C2	2	3
Native speaker	-	13
Education (highest level)		
college/upper-secondary school	2	-
university	11	11
PhD	8	9
higher vocational education	1	-
Teaching experience		
1-5 years	-	7
6-10 years	-	5
15-30 years	-	4
31+ years	-	2
None	-	2
Teaching level		
secondary school	-	4
upper-secondary school	-	2
college	-	1
SFI*	-	2
adult education	-	3
university	-	5
none	-	3

***SFI - Swedish for Immigrants (adult education)**

6 Results and analysis

6.1 Control items

Four items among the verbs and adverbs were duplicated as control items (one core item and three periphery items), to see whether the crowdsourcers annotated items consistently. If they did, then these items should appear next to each other in the linear rankings.

For adverbs, the two occurrences of the core item *således* ('consequently') appear at ranks 55 and 56 for non-experts and 51 and 52 for experts, which shows that they were ranked consistently. The other adverbial control word *sammanfattningsvis* (‘summing up’) appears at ranks 44 and 46 for non-experts and ranks 34 and 35 for experts, again showing consistent rankings.

The verb *underskatta* (‘underestimate’) appears at ranks 32 and 33 for non-experts and ranks 39 and 40 for experts, once again showing good ranking consistency in both groups. However, the second control verb, *förebygga* (‘prevent’), appears at ranks 35 and 42 for non-experts and at ranks 30 and 35 for experts. This verb shows more variation than the other items, especially for non-experts, with a difference of seven ranks (five for experts) on a scale of 60, amounting to ≈10% (≈8%). This could be due to the fact that it appeared at two levels in the coursebooks, and not only at the level it was picked for but also the level *before* it (B2). Another reason could be the co-occurrence with items against which the two occurrences of *förebygga* (förebygga_1 and förebygga_2) have been compared in the experiment. Both alternative explanations should be tested further, as should items’ distributions in the coursebooks (see Appendix 2) to see whether we can find a way to prevent ‘förebygga’-cases in the future applications of this method.

To conclude, the use of control items has shown that crowdsourcers generally vote very systematically. Even when there seems to be a difference in perception of items’ difficulty, the difference in the resulting scalar ranking is relatively modest. We consider, therefore, the resulting ranking (and the method to generate this type of ranking) reliable for our purposes.

6.2 Linear scale

Here we present the results obtained by aggregating the votes for each item into linear scales. To calculate the linear ranking, each time an item was marked as the easiest within a mini-task it received a score of 1, and if marked as the most difficult one it received a score of 3 (see Figure 4 for an example of a mini-task). The unmarked two items received a score of 2. After all votes were collected, the average scores per item were calculated and used for linear ranking (column

'Linear score' in Table 4). As a result, we can explore the positioning of unknown items relative to core items. Table 4 illustrates an excerpt of the resulting linear scale for the unknown word *likaledes* ('likewise, also') and its four closest core neighbors (periphery and unknown neighbors omitted). This example clearly demonstrates that the most probable level that we can expect *likaledes* to appear at and be understood at is C1, both according to teacher votes and to learner votes.

Table 4: *An excerpt of a linear ranking of the unknown item* likaledes *('likewise, also') with a window of four closest core items around it*

Lemma	Linear score	CEFR	Coreness	Rank
Learners				
sammanfattningsvis (' to sum up')	2.35	C1	core	44
jämförelsevis ('comparatively')	2.38	C1	core	45
likaledes ('likewise, also')	2.45	_	**unknown**	48
därigenom ('in that way')	2.50	C1	core	50
bevisligen ('demonstrably')	2.54	C1	core	51
Teachers				
ytterst ('farthest out')	2.26	B2	core	47
därigenom ('in that way')	2.28	C1	core	48
följaktligen ('consequently, accordingly')	2.42	C1	core	50
likaledes ('likewise, also')	2.73	_	**unknown**	55
bevisligen ('demonstrably')	2.74	C1	core	56

We then calculate the correlation between the expert and non-expert rankings by using the Pearson correlation coefficient. Overall, the rankings are quite correlated, ranging from 0.77 (Pearson correlation coefficient) to 0.94 overall. Table 5 gives an overview of the correlation coefficients by part-of-speech, as well as a more detailed overview by core, peripheral and unknown words. It shows that experts and non-experts agreed most on verbs (0.94) and least on nouns (0.77). Appendix 4 gives further details of the correlations by level.

Table 5: *Pearson correlation coefficients between experts and non-experts by part-of-speech and core, peripheral and unknown*

	Overall	Core	Peripheral	Unknown
Nouns	0.77	0.84	0.60	0.52
Verbs	0.94	0.95	0.90	0.78
Adjectives	0.92	0.92	0.88	0.84
Adverbs	0.91	0.90	0.90	0.92

Note. The grey background indicates where there is a reasonably high correlation ≥0.84.

We see a pattern in the correlations of the core, peripheral and unknown vocabulary: participants agreed most on core vocabulary for three of the parts-of-speech (verbs, adjectives, adverbs), a bit less on peripheral vocabulary, and least on unknown vocabulary. Adverbs are the only category where this trend seems reversed, with most agreement on the unknown vocabulary. However, this category also shows the least variation overall, with coefficients around 0.90. This could be related to the fact that we also had some trouble picking items for this part-of-speech since there were simply fewer adverbs to choose from in the data. It could also be an idiosyncratic result and we would need more data to confirm the reason for this difference between adverbs and the other groups, or if more data would result in the same trend as for the other parts-of-speech.

6.3 Clustering

While the linear scale gives a good overview of the relative difficulty of items, it remains a one-dimensional representation. As we want to explore how to assign levels to words of unknown level based on anchor words of known level, such a representation might not suffice. As illustrated in Figure 5, increasing the dimensionality can uncover relations that are not visible at lower dimensions. In the figure the green stars represent words of unknown level, while the blue dots and brown squares represent anchor words of known level. From looking at the one-dimensional (i.e., linear scale) visualization, it is rather difficult to attach a level to those words of unknown level based on the proximity of anchor words. Even the two-dimensional representation does not show a clear trend. However, the three-dimensional representation shows clearly that one of the words of unknown level is very close to the blue dot anchor words, while the other one is close to the brown square anchor words. While we acknowledge that vectors based on pairwise comparisons from the linear scales will show high degrees of intercorrelation, we explore this technique in an attempt to untangle the low-dimensional representation and to visualize the results.

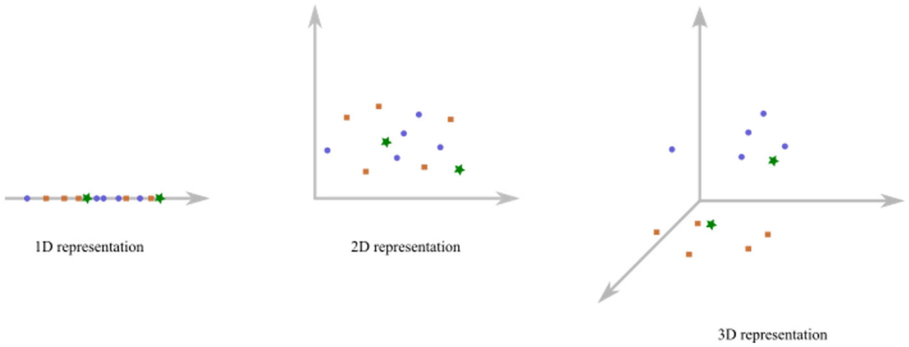


Figure 5: Clustering results in 1-, 2- and 3-dimensional representations.

Following this intuition, we use the distances from the linear scale to represent our words in high dimensional space. To do so, we calculate the distance between each pair of words, so that each word has 59 distances, one for each of the other words, plus a distance of zero to itself. These distances are then interpreted as coordinates in a 60-dimensional space.

By using this high dimensional representation, we want to see whether we can assign levels to words of unknown levels. As a first step, we perform a clustering analysis on the core and peripheral data in order to see whether clustering might be a viable choice for assessing the difficulty of unknown words. As we assume the levels of core and peripheral vocabulary to be known and valid, we can use these labels to see to what extent a clustering algorithm generates the expected results, and for the clustering in this step we use KMeans (McQueen, 1967).

Tables 6a–6d show the overall confusion matrices by group (experts and non-experts) and part-of-speech (nouns, verbs, adjectives and adverbs), excluding words of unknown level, since their true level is not known. Numbers in bold indicate cases where the clustering algorithm assigned most of the elements in a class to the correct class.

Table 6a: *Adverbs, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**6**	3	0	0	0		**8**	3	0	0	0
A2	2	**3**	1	1	0		0	1	0	2	2
B1	2	2	**8**	3	0		2	4	**5**	2	0
B2	0	1	1	3	5		0	1	3	**4**	2
C1	0	1	0	3	**5**		0	1	2	2	**6**

Table 6b: *Adjectives, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	1	0	0	0		**6**	1	0	0	0
A2	5	5	0	4	0		4	4	0	3	1
B1	0	2	**8**	3	1		0	3	**5**	2	0
B2	0	2	2	3	5		0	0	1	1	4
C1	0	0	0	0	**4**		0	2	4	4	**5**

Table 6c: *Verbs, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	2	0	0	0		**4**	2	0	0	0
A2	3	**4**	1	1	0		3	**4**	1	0	0
B1	0	0	**4**	3	2		1	3	**4**	3	0
B2	1	3	3	3	2		0	1	1	**4**	3
C1	1	1	2	3	**6**		2	0	4	3	**7**

Table 6d: *Nouns, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	2	0	0	0		**5**	2	0	0	0
A2	4	4	2	2	2		4	4	2	2	2
B1	1	2	**4**	2	1		1	2	**4**	2	1
B2	0	1	3	3	2		0	1	3	3	2
C1	0	1	1	3	**5**		0	1	1	3	**5**

As can be gathered from the confusion matrices, the clustering tends to perform well on the extremes of the scale (levels A1 and C1) but also around B1, with most occurrences of these items correctly clustered.

In the second step, we want to assign levels to words with an unknown level. In order to do so, we use another clustering algorithm, the k-nearest-neighbors (*k-NN*; Fix and Hodges, 1989) algorithm to see which anchor words are closest to the unknown words. We then predict the level of the unknown word as the majority level of its five closest neighbors. Tables 7a–7d present the results of this analysis.

Table 7a: *Clustering results for unknown adverbs*[14]

Adverbs	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
enbart ('solely, only')	A1	B1	B1
således ('consequently')	A1	C1	C1
förvisso ('certainly')	A2	C1	C1
såklart ('absolutely')	A2	B1	B2
mestadels ('mostly')	B1	C1	B1
sedermera ('afterwards')	B1	C1	C1
tillika ('moreover')	B2	C1	C1
fortsättningsvis ('henceforth')	B2	B2	C1
likaledes ('likewise')	C1	C1	C1
massvis ('lots of')	C1	B2	C1

Table 7b: *Clustering results for unknown adjectives*

Adjectives	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
vag ('vague')	A1	B2	B2
uppenbar ('obvious')	A1	B2	B1
facklig ('trade union')	A2	B2	B2
rättslig ('legal')	A2	B2	B2
skeptisk ('skeptical')	B1	B1	B1
nyliberal ('neo-liberal')	B1	B1	B1
medborgerlig ('civil')	B2	B2	B1
byråkratisk ('bureaucratic')	B2	B1	B1
välmående ('healthy')	C1	B1	B1
ovannämnd ('above-mentioned')	C1	B2	B2

[14] The grey background in Tables 7a–7d marks agreement between the two groups or between at least one group and the CEFR-level predicted in the Swedish Kelly-list.

Table 7c: *Clustering results for unknown verbs*

Verbs	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
kapa ('hijack')	A1	B1	B1
ämna ('intend to')	A1	B1	B1
förespråka ('advocate')	A2	B1	B1
tillhandahålla ('supply')	A2	B1	B1
erinra ('remind')	B1	B1	B1
påvisa ('prove')	B1	B1	B1
avlägsna ('remove')	B2	B1	B1
genomsyra ('permeate')	B2	B1	B1
understödja ('support')	C1	B1	B1
beskåda ('regard')	C1	B1	B1

Table 7d: *Clustering results for unknown nouns*

Nouns	Cf. Kelly level	Predicted level	
		L2 speakers	Experts
medlemsstat ('member state')	A1	B1	B1
pelare ('pillar')	A1	B1	B1
upphovsrätt ('copyright')	A2	B2	B1
penningpolitik ('monetary policy')	A2	B2	B1
fildelare ('file sharer')	B1	B2	B1
antagande ('assumption')	B1	B2	B1
mervärde ('surplus')	B2	B2	B1
sökmotor ('search engine')	B2	B1	A2
vapenvila ('cease-fire')	C1	B1	B1
dotterbolag ('subsidiary company')	C1	B1	B1

The predicted levels largely overlap between the two voter groups (Tables 7a–7d), with differences within one level, except for the adverb *mestadels* ('mostly'), which is predicted to be both B1 and C1. All unknown words are further predicted as at least B1 (except for *sökmotor* 'search engine' which is predicted as A2 by the experts), which confirms our findings for linear scales, where unknown words end up in the middle and the end of the scale.

For adverbs, the non-expert clustering perfectly aligns with their ranking: A2 words (according to the clustering) have ranks lower than B1 words (according to the clustering), which in turn have ranks lower than B2 words (according to the clustering). For the expert clustering, this is almost true: *såklart* (‘absolutely’) is at rank 16 in the ranking while *enbart* (‘solely, only’) is found at rank 27. However, *enbart* is predicted as B1 and *såklart* B2.

For adjectives, the non-expert clustering also perfectly aligns with their ranking, as all words predicted as B1 come before words predicted as B2 in the ranking. This also holds true for the expert clustering.

For verbs, clustering is perfectly identical for both groups, and all words are predicted to be of level B1 in both groups.

For nouns, the non-expert clustering again perfectly aligns with their ranking, as B1 words (according to the clustering) are all found before B2 words (according to the clustering) in the ranking. This is not the case for the expert clustering, as all words are predicted as B1 except for *sökmotor* (‘search engine’) which is predicted as A2, yet is found at rank 44 according to the linear scale, while *pelare* (‘pillar’) – predicted as B1 – is ranked at 27.

Finally, if we compare the predicted levels to the levels assigned to these words in Kelly, we can see a rather large discrepancy in most of the cases, especially concerning the lowest levels, which could indicate that frequency alone may not be sufficient as a predictor of CEFR levels.

6.4 Time investment

Figure 6 shows the average time needed per task for non-experts and experts in seconds.[15] The box shows the first and third quartiles (lower and upper lines of the box), the orange line dividing the box indicates the median, the whiskers show the minimum and maximum values (outliers not counted), and the dots show outlier values. We hypothesize that

[15] Outliers of more than 100 seconds are excluded in order to improve the readability of the figure. Extreme outliers go up to 3,500 seconds.

outliers indicate moments when participants were interrupted during the experiment (e.g. went to get coffee or answered the phone) without closing the program. As we see, despite some obvious outliers, the average projected time of 30 seconds per task is well met.

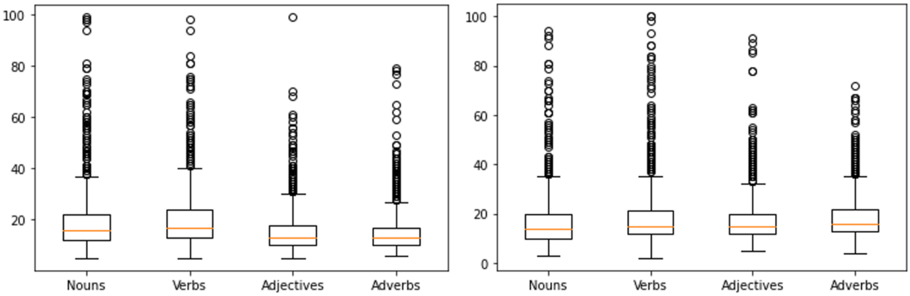


Figure 6: Boxplots of time taken per task in seconds, for learners (left) and experts (right).

6.5 Qualitative analysis

The easiest 10 items (Table 8) in all four parts-of-speech were mainly core items, according to both learners and experts. For nouns and verbs, 90–100% of the easiest ten items were core items, but they ranged from A1–B1 in the case of the learners' ranking of nouns whereas the experts picked A1–A2 core items as the easiest ten nouns, and learners and experts picked A1–A2 core items as 90% of the ten easiest verbs (see Appendix 3). For adjectives and adverbs, 80% of the easiest ten items were picked from the core items by learners and 60–70% by the experts. All of the core items picked among the easiest ten adjectives and adverbs were from A1–A2 in the coursebooks.

Even among the 20 easiest items, it was mainly core items that were picked by both learners and experts in all four parts-of-speech. However, these items range from A1–C1 and there are more peripheral items compared to the easiest 10. Among the easiest 20 adverbs experts even include one of the unknown items (*såklart* 'absolutely'), an item picked from A2 in the Kelly list and hence from the second frequency band.

Looking instead at how the core items were ranked, we see that they appear primarily among the 20 easiest items in all four parts-of-speech

Table 8: *Core items among the items ranked as the 10 easiest items by learners and experts*

	Learners	Experts
Nouns	A1: *pappa, kaffe, klocka, dag, frukost* (daddy, coffee, clock, day, breakfast)	A1: *pappa, kaffe, klocka, frukost, dag*
	A2: *kilo, doktor, mage, kött, flygplan* (kilo, doctor, belly, meat, airplane)	A2: *doktor, flygplan, mage, kött, kilo*
Verbs	A1: *äta, gå, titta, stå, heta* (eat, go, look, stand, be named)	A1: *gå, titta, äta, heta, stå*
	A2: *cykla, kontakta, trycka, meddela* (bike, contact, push, inform)	A2: *cykla, kontakta, meddela, trycka*
Adjectives	A1: *liten, glad, stor, bra, halv* (small, glad, big, good, half)	A1: *liten, bra, stor, glad, halv*
	A2: *grön, lycklig, vuxen* (green, happy, grown-up)	*A2: grön*
Adverbs	A1: *också, hemma, där, ibland* (also, at home, there, sometimes)	A1: *sedan, också, ibland, hemma, där* (since, …)
	A2: *ej, dit, inne, var* (not, there, inside, where)	*A2: dit, ej*

Note. The grey background indicates that learners and experts agreed on these items as being among the 10 easiest core items.

and for both learners and experts (Table 9). Conversely, we find very few core items among the most difficult items (20–30%) in all four parts-of-speech and both for learners and experts. Furthermore, the core items which are among the most difficult are from B1–C1 and never A1–A2, whereas the easiest core items range from A1–B2 for learners, A1–C1 for experts.

Table 9: *Dispersion of the core items in the ranking experiment according to the number of core items among items number 1–20, 21–40 and 41–60, including the levels predicted for those levels based on CEFR tools*

	1–20	21–40	41–60
Noun core items (learners)	14 (A1–B2)	7 (B1–C1)	4 (B2–C1)
Noun core items (experts)	11 (A1–B1)	9 (B1–C1)	5 (B1–C1)
Verb core items (learners)	12 (A1–B1)	6 (B1–C1)	6 (B1–C1)
Verb core items (experts)	12 (A1–B1)	7 (B1–C1)	5 (B1–C1)
Adjective core items (learners)	13 (A1–B2)	8 (B1–C1)	4 (B2–C1)
Adjective core items (experts)	13 (A1–C1)	8 (A2–C1)	4 (B2–C1)
Adverb core items (learners)	13 (A1–B2)	5 (B1–B2)	6 (B2–C1)
Adverb core items (experts)	12 (A1–B2)	6 (A2–C1)	6 (B1–C1)

The items which were ranked as the ten most difficult contain very few core items, and the core items which appear here are usually C1 (the learners chose eight of these, the experts five, see Table 10). These words do not contain any clearly international items, and both learners and experts agree to a large extent, although the former include three more items than the latter. There is also one item that is only included by the experts (*anvisning* 'directions, instructions'), whereas learners include four items which the experts did not include: *bedra, förebygga, värdesätta* and *följaktligen.*

Table 10: *Core items among the items ranked as the 10 most difficult by learners and experts*

	Learners	Experts
Nouns	*utformning* ('design')	*utformning* *anvisning* ('directions')
Verbs	*bedra* ('deceive') *förebygga* ('prevent') *värdesätta* ('value')	-
Adjectives	*övergripande* ('comprehensive') *nedlåtande* ('condescending')	*övergripande* *nedlåtande*
Adverbs	*bevisligen* ('demonstrably') *följaktligen* ('consequently')	*bevisligen*

Note. The grey background indicates that the same words were seen as among the 10 most difficult by learners and experts.

7 Discussion

The way we designed the experiment shows that frequency-based statistical measures and predictions offered by CEFR-tools can, indeed, help stratify vocabulary into (so-called) core and periphery items. The items we have picked as core based on the ranges in Table 1 behaved differently from those that we chose as periphery. It is important to bear in mind that, apart from pure frequency ranges, we also applied the principles of topicality and/or usefulness where many candidate items were available or, conversely, where too few were available. In general, we have seen that the core/anchor items per level have been confirmed as such on a linear scale by voters with different linguistic backgrounds. These are the items that could be effectively used for further experiments on a method of assigning unseen items to a proficiency level.

We can thus claim that we have developed a strategy to identify words capable of being reliable anchors, namely, using *CEFR tools* by applying various statistical measures. The ranges that we have experimented with, have helped us capture the coreness of certain vocabulary items for each level, as confirmed by both expert and non-expert ratings. However, we have also realized that cognates and internationally recognizable items give a false sense of simplicity, can easily mislead and should not be used for experiments of this type (cf. Lindström Tiedemann et al., 2022). The item selection and crowdsourcing experiment have enriched us with a list of items that we recommend using in the future for assigning unknown vocabulary to the target levels. The core items used in our experiment are listed in Appendix 5, both in Swedish and translated into English. It would be most interesting to see whether the same items represent coreness in other languages for the corresponding levels.

Empirical analysis of unknown items in relation to our anchor words has shed new light on **how frequencies, usefulness/topicality, coreness and language learning may be related** (see 6.3). Frequency has been claimed to be a consequence of being core, and not vice versa (Stein, 2017), although frequency is often taken as a proxy of coreness. The problem with this type of simplification is that not all core items are frequent (e.g. the frequency of *Tuesday* compared to *Friday*; *brown* and *white*) thus frequency may lead to contentious results.

To demonstrate the last claim, the unknown items that we selected from the Swedish Kelly list have been related to CEFR levels based on frequency bands. Our results, however, show that these levels very rarely coincide with the levels predicted through positions on a linear scale, even though the learners and experts in this experiment were in high agreement about their relative difficulty. We take this to mean that although frequency is important in learning new vocabulary, CEFR levels (and ‘coreness’ of the items) cannot simply be related to frequency bands. This finding is highly relevant to second language acquisition, since vocabulary assessment tends to rely on testing vocabulary in relation to frequency bands. Most importantly, frequencies in general corpora may be irrelevant for L2 contexts, whereas L2-relevant corpora are likely to contain more reliable frequency indications.

Unknown items coming from the Kelly list were found only once among the easiest 20 items and only among the expert judgments. Unknown items were, instead, very highly represented among those ranked as the most difficult 20, or even the most difficult 10 (see Appendix 3), which also serves as an indication that at least levels below B1 were poorly predicted by the Kelly list, based on frequencies from general corpora.

The results of the study suggest that the same setup, but limited to two–three anchor items per level (e.g. 10 anchor items in total) and one–two unknown words (e.g. 12 items for a "project" in total), could help resolve the question of an unknown word and its placement on a CEFR scale for learning and assessment purposes. Since we already know the relations (easier–more difficult) between the anchor words, the number of micro-tasks would be dramatically reduced by only testing these relations for the unknown words. A suggestion for item placement could thus be achieved in a very limited time. Moreover, our findings indicate that we can let any user with sufficient knowledge of Swedish vote, without controlling for their background (i.e. native speakers, trained experts or non-native speakers). Testing this approach as a quick method for resolving level assignment for previously unseen vocabulary items is planned in future work. A number of questions need to be addressed in this context, for example:

- How many micro-tasks are optimal? How much time will it take to place one new item?
- How many votes do we need? (cf. Alfter et al., 2021)
- How stable is the ranking? Does decreasing the number of votes affect placement reliability?

Carter (1982, p. 46) summarizes his theoretical analysis of the nature of core vocabulary by saying that "...no single criterion can be taken to produce definitely a core vocabulary item. Rather some combination can help define the strength of the 'coreness' but it will also, to some extent, be affected by the purposes for which a definition of a core lexis is sought.” Based on Carter (1982) we may stipulate that the testing paradigm for core vocabulary (Section 2.1) will not be applicable in its entirety to all types of core vocabularies. For example, a cognitive basis will be less critical for pedagogical uses of core vocabulary; for general lexicography, definitional power and semantic network placement would figure most prominently; while for diachronic lexicostatistics, semantic neutrality and frequency will be the most important properties.

We would like to round off this discussion by quoting Borin (2012, p. 63): “It is perhaps not surprising that there should be so little overlap among different kinds of ‘core vocabularies’, since they aim at capturing different aspects of ‘coreness’”. We thus do not propose that the suggested anchor words that we claim to be core for the second language of learners will be universal in all settings.

8 Conclusions

Returning to our **hypotheses**, we can now confirm the following:

1. There is a common core vocabulary at *A1–B1 levels*; there is less systematicity at *B2–C1* levels.

Analyzing the correlation between learners and experts we could see that the correlation was generally higher in the core items for A1–B1 than for B2–C1. In fact, we even noticed that the correlation was sometimes higher for the whole A1–B1 group than within a particular level (cf. Appendix 4). We believe this to be an indication that there is a core which relates quite well to the A1–B1 levels in coursebooks, but that the precise order in which these items occur in the coursebooks and how they would be ranked by learners or experts might not coincide as well for each level. This is also related to the fact that we assume that proficiency is a continuum rather than something that can be clearly divided into discrete levels.

2. Some systematicity can be observed in the behavior of the *core* items, but less so in *peripheral* items.

Core items very clearly appear mainly among the items which are ranked among the easiest. This is likely to be because both peripheral items and unknown items in this experiment do not belong to the core vocabulary, and that different learners vary in their knowledge of these words and hence variations in the ranking of these words.

3. Through crowdsourced comparative judgments, *unknown* vocabulary items will demonstrate a perceived difficulty (expressed in numerical scores) equal or comparable to the perceived difficulty of *anchor* items of a particular level.

The crowdsourcing experiment has shown that sensible levels can be assigned to words of unknown level based on comparative judgments of unknown words against anchor words, as illustrated in Table 4. This seems to confirm that we can use the perceived difficulty of unknown vocabulary items for the assignment of levels based on the closeness of the difficulty of nearby words. However, while this methodology allows for the identification of levels for words included in the experiment, it is not possible to easily calculate actual levels for new words, as the linear ranking and clustering are performed on the distance of each word to every other word (and thus voting to calculate distances between all anchor words and the unknown item is a prerequisite of this approach). The experiments presented in this work also show that **learners** are perfectly aligned with regard to their assessments of the difficulty of words with an unknown level and the subsequent linear scale projection and clustering.

4. *Non-experts* will perform on par with *experts* in a comparative judgment setting.

This question yields a convincing "yes" in response. The study has confirmed previous findings that L2 learners can be used in the same way as experts, given a carefully designed comparative judgment setting.

We found that our method of selecting core items worked well in establishing anchors. To ensure their reliability it is particularly important to make sure that their meaning cannot be derived from international words which appear in many European languages or cognates in English, such as the Swedish *doktor* ('doctor').

While we do not yet have an inexpensive cheap method for ranking items in relation to explicit CEFR levels, there is a good chance that with the knowledge we have gained one will be soon available. We have seen that clustering can indeed be used to derive sensible levels for words of unknown level; in the future it would be interesting to calculate, for example, the distances between the unknown words in a cluster and the cluster center to see whether this gives any hints as to the coreness of these items.

It is especially encouraging that we can use learners for this type of linguistic annotation, and in fact our results indicate that learners *might* be more attuned to the relative difficulty of words than experts are, since their rankings more often coincide with the coursebook levels. This may be due to the fact that we operationalized *difficulty* in terms of the CEFR, and the CEFR levels specifically target learners. It would thus be logical for learners to be more sensitive to these levels than native speakers and language professionals, a finding also hinted at in Alfter et al. (2022).

Acknowledgments

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1, and by *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

Alfter, D. (2021). Exploring natural language processing for single-word and multiword lexical complexity from a second language learner perspective. PhD thesis. University of Gothenburg.

Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E., & Pilán, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 1–7).

Alfter, D., Cardon., R., & François, T. (2022). A Dictionary-based Study of Word Sense Difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, (pp. 17–24).

Alfter, D., & Volodina, E. (2018). Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 79–88).

Alfter, D., Lindström Tiedemann, T., & Volodina, E. (2021). Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. *Northern European Journal of Language Technology* (Vol. 1). doi: 10.3384/nejlt.2000-1533.2021.3128

Alfter, D., Borin, L., Pilán, I., Lindström Tiedemann, T., & Volodina, E. (2019). Lärka: from language learning platform to infrastructure for research on language learning. In *Selected papers from the CLARIN Annual Conference 2018, 8–10 October 2018, Pisa* (pp. 1–14). Linköping University Electronic Press.

Bell, H. (2013). Core Vocabulary. In C. Chapelle (Ed.) *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell.

Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In *Shall We Play the Festschrift Game?* (pp. 53–65). Springer, Berlin, Heidelberg.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In: *Proceedings of LREC 2012* (pp. 474–478). Istanbul: ELRA.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1–22.

Capel, A. (2015). The English Vocabulary Profile. *English profile in practice, 5*, 9–27.

Carter, R. (1982). A note on core vocabulary. In Stubbs M. and Carter R. (Eds.), *Nottingham Linguistic Circular, 11*(2), 39–51.

Carter, R. (1987). Is there a core vocabulary? Some implications for language teaching. *Applied linguistics, 8*(2), 178–193.

Council of Europe [COE]. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe [COE]. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (19. 10. 2021)

Crosbie, S., Pine, C., Holm, A., & Dodd, B. (2006). Treating Jarrod: A core vocabulary approach. *Advances in Speech Language Pathology, 8*(3), 316–321.

De Clercq, Orphée, Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering, 20*(3), 293–325.

Dixon, Robert MW. (1971). A method of semantic description. *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, 436–471.

Familiar, L. (2021). *A frequency dictionary of contemporary Arabic fiction: core vocabulary for learners and material developers*. Routledge.

Fix, E., & Lawson Hodges, J. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/ Revue Internationale de Statistique, 57*(3), 238–247.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

François, T., Volodina, E., Pilán, I., & Tack, A. (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)* (pp. 213–219).

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of THING. *Journal of English for Academic Purposes, 52*, 100999.

Hawkins, J. A., & Filipović, L. (2012). Criterial Features in L2 English. In *English Profile Studies 1*. Cambridge: Cambridge University Press.

Holmer, D., & Rennes, E. (2022). NyLLex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level. In *Proceedings of the 13th Language Resources and Evaluation Conference* (*LREC*)*, 20 – 25 June 2022, Marseille* (pp. 1326–1331). Retrieved from http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.141.pdf

Hulstijn, J. H. (2019). An individual differences framework for comparing nonnative with native speakers: Perspectives from BLC theory. *Language Learning, 69*, 157–183.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Bondi Johannessen, J., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., & Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation, 48*(1), 121–163.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018). Collocations dictionary of modern Slovene. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 989–997). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Retrieved from https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/118/211/2939

Kullenberg, C., & Kasperowski, D. (2016). What is citizen science? – a scientometric meta-analysis. PloS one, *11*(1), e0147152.

Lau, J. H., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the annual meeting of the cognitive science society, 36*(36).

Lehmann, H. (1991). Towards a core vocabulary for a natural language system. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Retrieved from https://aclanthology.org/E91-1053.pdf

Lindström Tiedemann, T., Alfter, D., & Volodina, E. (2022). CEFR-nivåer och svenska flerordsuttryck [= CEFR levels and Swedish Multiword Expressions]. In S. Björklund, B. Haagensen, M. Nordman & A. Westerlund (Eds.), *Svenskan i Finland 19,* Vaasa:: Svensk-Österbottniska Samfundet r.f. (pp. 218–233). Retrieved from https://www.doria.fi/handle/10024/185549

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.

Louviere, J. J., N. Flynn, T., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*.

Márquez, M. F. (2007). Renewal of core English vocabulary: A study based on the BNC. *English Studies, 88*(6), 699–723.

Mühlenbock, K., H., & Johansson Kokkinakis, S. (2012). SweVoc-a Swedish vocabulary resource for CALL. In *Proceedings of the SLTC 2012 workshop on NLP for CALL, 25th October 2012, Lund* (pp. 28–34). Linköping University Electronic Press.

Ortega, L. (2012). Interlanguage complexity. In Kortmann, B. & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). De Gruyter.

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating. *Language Learning*.

Scott, William A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. doi: 10.1086/266577

Stein, G. (2017). Some thoughts on the issue of core vocabularies: A response to Vaclav Brezina and Dana Gablasova: Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 38*(5), 759–763.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell publishers.

Swadesh, M. (2017). *The origin and diversification of language*. Routledge.

Volodina, E., & Johansson Kokkinakis, S. (2012a). Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1040–1046).

Volodina, E., & Johansson Kokkinakis, S. (2012b). Swedish Kelly: Technical Report. GU-ISS-2012-01. The Swedish Language Bank, Gothenburg University.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016). SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 76–84).

Volodina, E., Pilán, I., Rødven Eide, S., & Heidarsson, H. (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* (pp. 128–144).

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., Carin, L. (2018). Joint Embedding of Words and Labels for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 2321–2331).

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.

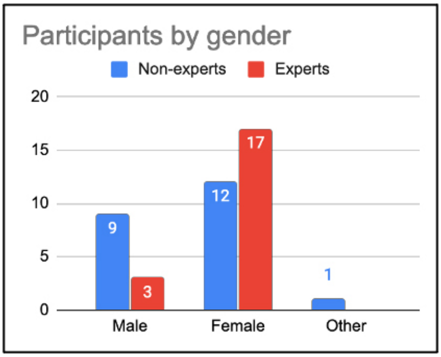
Ocene posameznih leksikalnih elementov, pridobljene z množičenjem: perspektiva osrednjega besedišča

V raziskavi preučujemo teoretična in praktična vprašanja, povezana z razlikovanjem med osrednjim in obrobnim besediščem na različnih ravneh jezikovnega znanja z uporabo statističnih pristopov v kombinaciji z množičenjem. Obenem ugotavljamo, ali je mogoče razvrstitve oseb, ki se učijo drugega jezika, uporabiti za določanje ravni nepoznanega besedišča. Raziskava je izvedena na enobesednih enotah v švedščini.

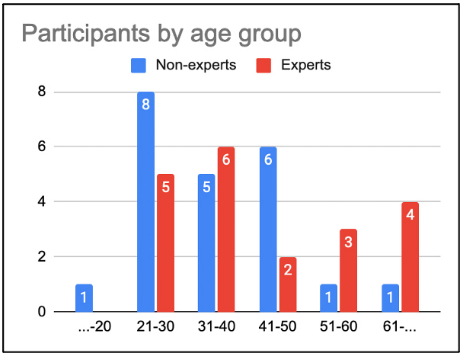
Preučujemo štiri hipoteze: (1) za vsako raven znanja obstaja osrednje besedišče, vendar to velja le do ravni B2 po CEFR (višja srednja raven); (2) osrednje besedišče kaže večjo sistematičnost v rabi, medtem ko se robni elementi obnašajo bolj idiosinkratično; (3) glede na to, da imamo za vsako raven na voljo ključne elemente (t. i. sidrne elemente), lahko vsako novo nepoznano besedo postavimo ob bok omenjenim ključnim elementom z vrsto primerjalnih ocenjevalnih nalog in tako določimo "ciljno" raven za prej nepoznano besedo; in (4) osebe s pomanjkljivim znanjem se bodo v primerjalnem ocenjevanju odrezale enakovredno osebam z dobrim znanjem. Hipoteze smo v veliki meri potrdili: V povezavi z (1) in (2) naši rezultati kažejo, da obstaja določena sistematičnost pri jedrnem besedišču za začetne in srednje ravni (A1-B1), medtem ko smo pri višjih ravneh (B2-C1) opazili manj sistematičnosti. Pri točki (3) predlagamo, da se kot metoda za dodelitev "ciljne" ravni nepoznanim besedam uporabi množičenje ocen besed z uporabo primerjalne presoje in s pomočjo poznanih sidrnih besed. Glede (4) potrjujemo predhodne ugotovitve, da je mogoče za naloge jezikovnega označevanja v okviru primerjalne presoje učinkovito uporabiti nestrokovnjake, v našem primeru učence jezika.

Ključne besede: osrednje besedišče in učenje jezika, množičenje pri nestrokovnjakih, posamični leksikalni elementi, ravni CEFR, primerjalna presoja

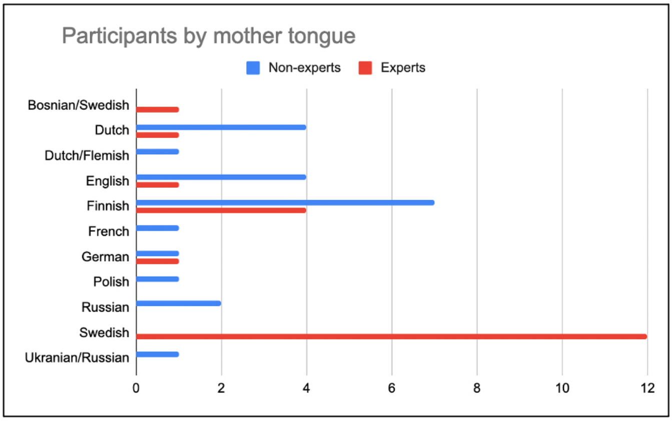
Appendix 1: Demographic information about the participants shown in graphs (1a–h)



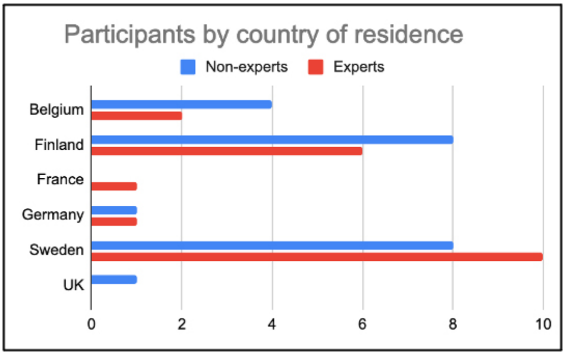
Appendix 1a.



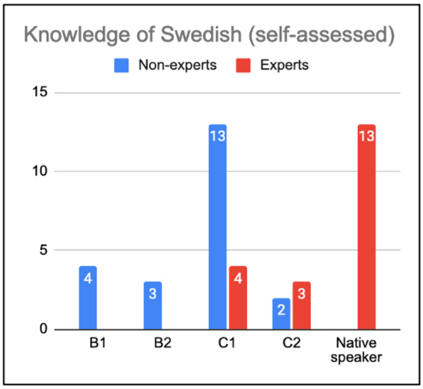
Appendix 1b.



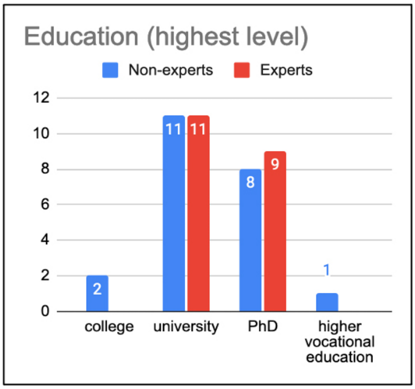
Appendix 1c.



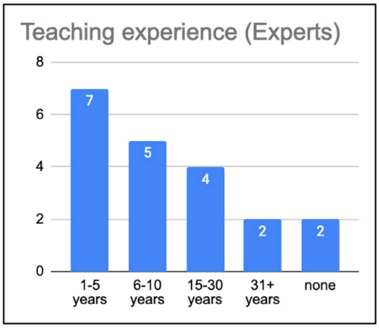
Appendix 1d.



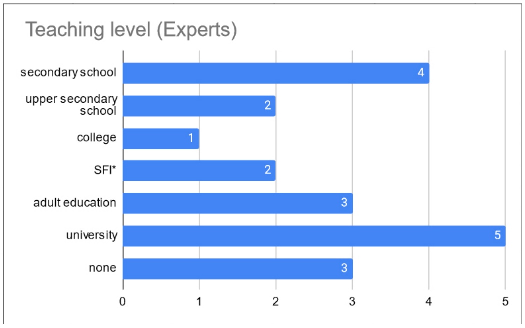
Appendix 1e.



Appendix 1f.



Appendix 1g.



Appendix 1h.

Appendix 2: Control items

	Level	Core / periph.	Agreement	Homogeneity	Product. majority level	Coctaill LM	Documents	Books
således *‘consequently’*	B2	core	0.5	-0.5	C1	B2	B2: 2 C1: 2	B2: 1 C1: 2
sammanfattningsvis *‘summing up’*	C1	periphery	0.5	-0.25	B1	B2	C1: 2	C1: 1
underskatta *‘underestimate’*	C1	periphery	0.75	0.375	C1	C1	C1: 1	C1: 1
förebygga *‘prevent’*	C1	periphery	0.5	-0.75	C1	C1	B2: 1 C1: 1	B2: 1 C1: 1

Note. Information about control items.

Appendix 3: Statistics for the included items

	Easiest 10 (1-10)	Easiest 20 (1-20)	Hardest 10 (51-60)	Hardest 20 (41-60)
Nouns (learners)	**90% core (A1–B1)** 10% periphery (A1)	**70% core (A1–B2)** 30% periphery (A1–C1)	10% core (C1) **50% periphery (A2–C1)** 40% unknown	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–B2)
Nouns (experts)	**100% core (A1–A2)**	**55% core (A1–B1)** 45% periphery (A1–C1)	20% core (C1) **50% periphery (A2–C1)** 30% unknown (A2–B2)	25% core (B1–C1) **40% periphery (A2–C1)** 35% unknown (A2–C1)
Verbs (learners)	**90% core (A1–A2)** 10% periphery (A1)	**60% core (A1–B1)** 40% periphery (A1–B2)	10% core (B2) 30% periphery (B2–C1) **60% unknown (A1–C1)**	30% core (B1–C1) 30% periphery (B1–C1) **40% unknown (A1–C1)**
Verbs (experts)	**90% core (A1–A2)** 10% periphery (A1)	**60% core (A1–B1)** 40% periphery (A1–B2)	0% core **50% periphery (A2–C1)** **50% unknown (A2–C1)**	25% core (B1–C1) 30% periphery (A2–C1) **45% unknown (A1–C1)**
Adjectives (learners)	**80% core (A1–A2)** 20% periphery (A1–A2)	**65% core (A1–B2)** 35% periphery (A1–B2)	20% core (C1) 30% periphery (B2–C1) **50% unknown (A1–C1)**	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–C1)
Adjectives (experts)	**60% core (A1–A2)** 40% periphery (A1–A2)	**65% core (A1–C1)** 35% periphery (A1–B1)	20% core (C1) **40% periphery (B1–C1)** **40% unknown (A1–C1)**	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–C1)
Adverbs (learners)	**80% core (A1–A2)** 20% periphery (A1)	**65% core (A1–B2)** 35% periphery (A1–B2)	20% core (C1) 30% periphery (A2–C1) **50% unknown (A1–B2)**	30% core (B2–C1) **35% periphery (A2–C1)** **35% unknown (A2–C1 + AB)**
Adverbs (experts)	**70% core (A1–A2)** 30% periphery (A1)	**60% core (A1–B2)** 35% periphery (A1–A2) 5% unknown (A2)	10% core (C1) 30% periphery (A2–C1) **60% unknown (A2–C1 + AB)**	30% core (B1–C1) 30% periphery (A2–C1) **40% unknown (A2–C1 + AB)**

Note. Core, periphery and unknown items by percentage in the 10 and 20 easiest items and the 10 and 20 most difficult items.

Appendix 4: Correlations

		Core	Peripheral
Nouns	A1	0.88	0.67
	A2	-0.36	0.84
	B1	0.72	0.55
	B2	**0.14**	**0.20**
	C1	0.75	**0.13**
	A1–B1	0.84	0.82
	B2–C1	0.58	**0.16**
Verbs	A1	0.66	0.93
	A2	0.91	0.95
	B1	0.99	0.77
	B2	0.85	0.93
	C1	0.39	0.94
	A1–B1	0.99	0.83
	B2–C1	0.73	0.92
Adjectives	A1	0.55	0.57
	A2	0.52	0.86
	B1	0.86	0.66
	B2	0.94	0.91
	C1	0.89	0.71
	A1–B1	0.93	0.88
	B2–C1	0.95	0.75
Adverbs	A1	-0.57	0.82
	A2	0.67	0.98
	B1	0.68	0.43
	B2	0.63	0.90
	C1	0.74	0.76
	A1–B1	0.81	0.89
	B2–C1	0.83	0.82

Note. Pearson correlation coefficients by part-of-speech and level, core, peripheral and unknown (correlation of learner and expert rankings).

In comparing the correlation scores between levels and between core and periphery we have marked the reasonably high (≥(-)0.72) correlations with a grey background, but a maximum of one per row (i.e. either core or periphery) unless the two values are both >0.9, to make it easier to see which correlations were highest. Correlations marked in bold are particularly low and to be seen as negligible, at ≤(-)0.30.

Appendix 5: Swedish core items with English translations

Level	Swedish core word	English translation
ADVERB		
A1	också	also
A1	där*	there
A1	ibland	sometimes
A1	sedan	then
A1	hemma	at home
A2	var*	where
A2	ej	not
A2	verkligen	really
A2	inne	in; indoors
A2	dit	there
B1	gradvis	gradually
B1	troligen	very likely, probably
B1	alltför	far too, much too
B1	näst	last (but one), second (best)
B1	vanligtvis	usually
B2	ytterst	farthest out
B2	ingenstans	nowhere
B2	möjligen	possibly
B2	således	consequently
B2	säkerligen	certainly
C1	därigenom	in that way
C1	jämförelsevis	comparatively
C1	sammanfattningsvis	to sum up
C1	följaktligen	consequently; accordingly
C1	bevisligen	demonstrably
ADJECTIVE		
A1	liten	small; little
A1	halv*	half
A1	stor	large
A1	glad	happy
A1	bra	well, alright
A2	duktig	capable
A2	ledsen	sad
A2	lycklig	happy

A2	vuxen	adult, grown-up
A2	grön*	green
B1	hemsk	ghastly, terrible
B1	offentlig	public
B1	besviken	disappointed
B1	van	used, accustomed
B1	stilla	calm
B2	kollektiv*	collective, *here*: public (as in public transport)
B2	orättvis	unjust; unfair
B2	tacksam	grateful
B2	relevant*	relevant
B2	främst	foremost
C1	nonchalant*	nonchalant, careless, negligent
C1	förstådd	understood
C1	nedlåtande	condescending, patronizing
C1	acceptabel*	acceptable
C1	övergripande	comprehensive
NOUN		
A1	kaffe*	coffee
A1	dag*	day
A1	pappa*	father, dad
A1	klocka*	watch; clock; at x o'clock
A1	frukost	Breakfast
A2	flygplan	airplane
A2	doktor*	doctor
A2	mage	stomach
A2	kilo*	kilo; kilogram
A2	kött	meat; (flesh)
B1	samarbete	co-operation
B1	ledare*	leader; head; chief
B1	distans*	distance
B1	djurliv	animal life; wildlife
B1	matvana (nb. singular)	eating habits
B2	resurs*	resource; means
B2	avsked	dismissal; goodbye
B2	existens*	existence; livelihood
B2	tillit	confidence; reliance
B2	folkgrupp	ethnic group

C1	anvisning	directions; instructions
C1	flexibilitet*	flexibility
C1	klyfta	gap
C1	utformning	design; shaping
C1	enhet	unit; unity
VERB		
A1	gå	walk
A1	heta	be called, be named
A1	stå	stand
A1	titta	look, glance
A1	äta*	eat
A2	trycka	press; squeeze, oppress sb
A2	meddela	inform sb; let sb know
A2	lägga	put, place; lay
A2	kontakta*	contact, get in touch with
A2	cykla*	cycle; (informal) bike
B1	föreslå	propose, suggest
B1	fokusera*	focus
B1	utvidga	widen; extend; expand; enlarge
B1	stärka	strengthen, confirm, starch
B1	anordna	organize, arrange
B2	brottas	wrestle, fight
B2	tillgodose	meet, satisfy; supply
B2	bedriva	carry on, pursue
B2	klistra	paste, stick
B2	bifoga	enclose, attach
C1	underskatta	underrate, underestimate
C1	värdesätta	value, estimate
C1	bedra	deceive, cheat, be unfaithful to
C1	belysa	light up, illuminate
C1	förebygga	prevent, forestall

Note. Swedish core items for each level and part-of-speech with their translations into English. These are all of our core items, selected as specified in Section 4.2.

Items marked with an asterisk (*) either have cognates in English or the same international loanword is present in English. This is likely to affect how easy English speakers find these items, and hence maybe they should not be seen as anchor items. Cells with a dark grey background are words which learners and experts agreed were among the 10 easiest items, and light grey background marks items which were among the 10 most difficult items by both learners and experts (see Section 6.5).

Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game

Tanara ZINGANO KUHN
Research Centre for General and Applied Linguistics, University of Coimbra

Špela ARHAR HOLDT
Faculty of Arts, University of Ljubljana; Faculty of Computer and Information Science, University of Ljubljana

Iztok KOSEM
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Carole TIBERIUS
Dutch Language Institute

Kristina KOPPEL
Institute of the Estonian Language

Rina ZVIEL-GIRSHIN
Ruppin Academic Center

One way to stimulate the use of corpora in language education is by making pedagogically appropriate corpora, labeled with different types of problems (sensitive content, offensive language, structural problems). However, manually labeling corpora is extremely time-consuming and a better approach

Zingano Kuhn. T., Arhar Holdt, Š., Kosem, I., Tiberius, C., Koppel, K., Zviel-Girshin, R.: Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game. Slovenščina 2.0, 10(2): 62–100.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.62-100

should be found. We thus propose a combination of two approaches to the creation of problem-labeled pedagogical corpora of Dutch, Estonian, Slovene and Brazilian Portuguese: the use of games with a purpose and of crowdsourcing for the task. We conducted initial experiments to establish the suitability of the crowdsourcing task, and used the lessons learned to design the Crowdsourcing for Language Learning (CrowLL) game in which players identify problematic sentences, classify them, and indicate problematic excerpts. The focus of this paper is on data preparation, given the crucial role that such a stage plays in any crowdsourcing project dealing with the creation of language learning resources. We present the methodology for data preparation, offering a detailed presentation of source corpora selection, pedagogically oriented GDEX configurations, and the creation of lemma lists, with a special focus on common and language-dependent decisions. Finally, we offer a discussion of the challenges that emerged and the solutions that have been implemented so far.

Keywords: crowdsourcing, game with a purpose, example sentences, pedagogical corpus

1 Introduction

Evidence of authentic language use is fundamental for language learning. One way to access this evidence is through the use of examples from corpora, i.e., large collections of texts produced in natural contexts, saved in electronic form. However, these corpora may include sensitive content or offensive language, in addition to exhibiting structural problems. While such use is unquestionably authentic, some teachers or material developers might consider it to be inappropriate for their needs, thus finding it necessary to manually filter the corpus before applying authentic examples to pedagogical contexts, which is a laborious task.

To facilitate and stimulate the use of corpora in education we propose creating problem-labeled pedagogical corpora. This way, the process of example selection could be significantly streamlined. At the same time, instead of deleting potentially problematic content from the corpus we will label it, thus leaving the choice of the use of certain examples dependent on the needs and contexts of use of teachers and didactic material developers. The types of problems to be labeled are: vulgar, offensive, sensitive content, grammar/spelling problems, incomprehensible/lack of context.

Creating such corpora is challenging due to at least three reasons. Firstly, the process of labeling sentences in corpora is extremely time-consuming, if done manually. Secondly, automatic labeling can also be demanding given the polysemic nature of words. Thirdly, sensitivity and offensiveness are rather subjective concepts. Our proposal is thus to use the help from the crowd to achieve this task. For that, we are currently developing CrowLL – Crowdsourcing for Language Learning,[1] a multi-mode, multi-language (Dutch, Estonian, Slovene, and Portuguese) digital game. In this game, the players will be offered two examples (automatically extracted from existing corpora) and prompted to choose one (or both, or even none) that they consider to be appropriate for language teaching purposes. They will be asked to categorize the problem(s) of the example that has not been chosen and point out the constituent parts of the sentence that they consider to be problematic. With the output obtained from the players, we will compile problem-labeled pedagogical corpora for the languages mentioned above. These corpora can be used for the development of auxiliary language learning resources, such as Sketch Engine for Language Learning – SKELL (Baisa and Suchomel, 2014),[2] dictionaries and teaching materials; and, within Natural Language Processing, for the creation of datasets aimed at training machine learning algorithms for the compilation of larger pedagogical corpora.

Data preparation plays a crucial role in any crowdsourcing project that deals with the creation of language learning resources. Indeed, the quality and structure of the input data, together with the type of

[1] The research group carrying out the Crowdsourcing Corpus Filtering for Pedagogical Purposes project, within which the Crowdsourcing for Language Learning (CrowLL) game is being developed, originated under the umbrella of the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) COST Action (CA 16105). It is currently composed of seven members from six countries (Brazil, Estonia, Israel, Netherlands, Slovenia, and Portugal) and encompasses four languages (Dutch, Estonian, Slovene, and Portuguese). See https://ucpages.uc.pt/celga-iltec/crowll/ for further information on the project.

[2] SKELL is a free language learning tool that provides automatic summaries of corpus data, namely, examples, collocations and thesaurus. Available at https://skell.sketchengine.eu (30. 8. 2022).

task, have a direct impact on the quality of the output. Consequently, our research question in this paper is: What is the methodology of data preparation that is required to attend to the needs of a crowdsourcing game dealing with identification of offensive language, sensitive content and structural problems in authentic language material? We present the steps taken, the decisions made, the challenges faced and the solutions found to create the methodology for preparing a dataset of 10,000 sentences per language to develop and internally test the CrowLL game. For that, we use three key elements: source corpora, from where the sentences to be labeled by the players will be extracted; Good Dictionary Examples – GDEX (Kilgarriff et al., 2008) configurations, which automatically identify more pedagogically-suited examples in the source corpora and assign scores to the sentences; and lemma lists, which define the sentences to be extracted from the corpora. After the game is developed and tested with real users, the methodology of data preparation itself can also be evaluated.

The paper builds on our previous work within the enetCollect COST action.[3] We have previously established the motivation for a gamified approach to the labeling of examples in pedagogical corpora. We have developed the idea, formulated research questions, conducted initial tests with the crowd to establish the suitability of the crowdsourcing task, and used the lessons learned to design both the game flow and a work plan for the implementation. We have presented different stages of this work at conferences, as available in Kuhn et al. (2021) and Zviel-Girshin et al. (2021). In this paper, we focus on the newest development, namely on the first stage of the game preparation that primarily addresses issues related to the (corpus) data needed for the game. While the paper builds upon our previous work, it also presents a new, summative view and describes various applicative methodological decisions that were tested on different languages to ensure further usability of our proposed model, both by other languages and for purposes other than the CrowLL game development.

This paper is structured as follows. Section 2 reviews different approaches to the identification of good examples for the creation of pedagogical corpora. Section 3 introduces crowdsourcing and gamification,

[3] https://www.cost.eu/actions/CA16105/ (28. 10. 2022)

specifically within the context of language learning. Section 4 presents the CrowLL game, firstly reporting on our previous crowdsourcing experiment, whose results have led to the adoption of the Games with a Purpose (von Ahn, 2006) approach. Section 5 describes the methodology for data preparation in detail, and Section 6 analyzes and discusses the results.

2 Pedagogical corpora and language examples

Text corpora are collections of authentic (written or spoken) texts in electronic form, sampled to represent a specific type of language use (e.g. Gries, 2009; Sinclair, 2005). Corpus texts are typically equipped with metadata and linguistic information on different levels, increasing their value for different purposes in applied linguistics, natural language processing, and other fields that benefit from analyzing language data. In this paper, we focus on the field of language education, where the importance and value of corpora have been firmly established (Boulton, 2017; Callies, 2019; Römer, 2009; Vyatkina and Boulton, 2017). Corpora can be used by researchers and teachers for the creation of teaching and testing materials, language resources (such as learners' dictionaries), or directly by students, as classroom work with authentic language facilitates bottom-up language learning (Osborne, 2002).

It has been established (e.g. Callies, 2019) that direct use of corpora for teaching purposes is still not very widespread for a series of reasons, among which is skepticism about the quality and appropriateness of the data, especially because corpora are usually compiled for carrying out research, not for language teaching. Attempts to address this problem and promote the use of corpora for teaching have led to the emergence of specialized *pedagogical corpora*, i.e., corpora prepared specifically for language learning purposes (Chambers, 2016, p. 364). One of the main characteristics of a pedagogical corpus is the need for "pedagogic mediation" (Braun, 2005), which takes into consideration a set of factors related to the learners and the learning context. For purposes of good example selection, for instance, we argue that one type of monitoring could focus on identification of possible structural (grammar and spelling) problems as well as sensitive/offensive content, which might be problematic when presented to learners without the mediation of the teacher.

The creation of pedagogical corpora is a costly and time-consuming endeavor; however, the process can be supported by the automatization of certain procedures. One possible approach is to clean elements considered to be problematic for pedagogical purposes from existing corpora, such as offensive words and structural errors (misspellings, grammar errors).

In reference to the former, one area that has invested extensively in the identification of offensive language is natural language processing (NLP), mainly with research on the automatic detection of hate speech, with the aim to contribute to monitoring abusive behavior on the internet (e.g., social media, comments on media channels). Some examples of efforts on this topic are specific evaluation tasks at SemEval (International Workshop on Semantic Evaluation),[4] such as OffensEval[5] (Zampieri et al., 2019; Zampieri et al., 2020), and the Workshop on Online Abuse and Harms (WOAH),[6] currently in its 6th edition (2022). An impressive amount of research on the subject has been carried out in NLP, as can be seen, for example, in Poletto et al. (2020). This survey presents an up-to-date, systematic review of the available resources on hate speech, with detailed analysis, some of the current weaknesses, and goals for improvement. According to the authors, it is a complement to previous surveys, in particular, Lucas (2014), Wiegand and Schmidt (2017), and Fontana and Nunes (2018) (Poletto et al., 2020, p. 479). Datasets, such as the ones available on the dedicated webpage Hate Speech Dataset Catalogue (Vidgen and Derczynski, 2020),[7] and lexica, such as HurtLex (Bassignana et al., 2018), are some of the resources developed in NLP that could be used as a source of keywords for corpus cleaning. This approach consists of using blacklists containing swear words, vulgarisms, and words related to sensitive content in order to remove from the corpus sentences where these words occur (see below for a combined use of blacklists and GDEX). That means

[4] https://semeval.github.io/ (28. 10. 2022)

[5] https://sites.google.com/site/offensevalsharedtask/home (28. 10. 2022)

[6] https://www.workshoponlineabuse.com (28. 10. 2022)

[7] https://hatespeechdata.com/ (28. 10. 2022)

the “clean” corpus would not contain any sentences with those words. Another contribution from NLP to corpus cleaning would be through the application of offensive identification models at the sentence level, thus eliminating from the source corpus sentences automatically identified as offensive. However, one of the challenges in computational approaches to this subject is that other aspects, above and beyond the linguistic surface, have a crucial influence in the determination of what offensiveness is. Schmidt and Wiegand (2017) present a few works that seek to incorporate context to hate speech detection, but acknowledge that in certain difficult cases the method fails, so more investigation is needed. Relatedly, Poletto et al. (2020) point out a shortcoming of not considering the pragmatic aspects of swearing when evaluating hate speech – the production of false positives.

Whatever perspective is adopted with regard to identifying offensiveness, either at a word or sentence level, we have argued (Kuhn et al., 2021) that the total elimination of sentences from the corpus should be avoided because: 1. very few words are problematic in all of their senses and contexts, and 2. teachers and didactic material developers should be free to use whatever examples they find useful for their various needs. We thus propose to label potentially problematic data in pedagogical corpora instead of removing it.

For structural errors, automatic error detection (following different methods), has been widely adopted. For instance, Reynaert (2006) adopts a corpus-induced corpus clean-up approach to detect typos in texts. Rather than dictionaries, the lexicon used in the clean-up process consists of typos found in large corpora. However, Xu and Chamberlain (2020) have shown that some problems identified as structural errors by automatic error detection methods might not be actual mistakes, but rather spelling and grammatical variations based on the context of use. They argue that humans are still required to perform the clean-up task, and thus developed a game (Cipher) in which players are asked to identify different types of errors in texts and annotate them.

A more lexically-oriented approach to the compilation of pedagogical corpora refers to the adoption of sophisticated methods that automatically analyze texts according to several criteria to identify good examples. These good examples can then be gathered in a pedagogical corpus. The current state-of-the-art in corpus linguistics is Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008), available as a feature in the Sketch Engine (Kilgarriff et al., 2004, 2014) corpus query system. The general idea of GDEX is to provide a list of suitable, good-quality candidate corpus sentences that lexicographers can directly add into the dictionary as illustrative examples. At the heart of GDEX is a rule-based formula that assigns a numerical score to each corpus sentence based on how well it meets the pre-defined criteria. The criteria can determine, for instance, the length of the sentence, the number of words in the sentence, the frequency of word forms or lemmas in the corpus, the presence or absence of certain elements in the sentence, and so on. The scoring formula (with additional parameters) constitutes a so-called GDEX configuration. There are two groups of classifiers used in the configuration: hard and soft. Hard classifiers include a very high penalty giving sentences a very low score, resulting in pushing them to the bottom of the candidate list. Soft classifiers either penalize sentences or award bonus points, helping to rank good dictionary example candidates. As a result, GDEX lists all example candidates in descending order and can also be used to filter out all sentences below a certain threshold (Kosem et al., 2019).

A GDEX-based methodology has already been used to create pedagogical SKELL (Sketch Engine for Language Learning) corpora for Russian, Estonian (Koppel, Kallas, et al., 2019), English, German, Italian and Czech. This entails filtering a source corpus with a GDEX configuration, leaving only the sentences that meet all the criteria of good dictionary examples and removing the rest. But creating corpora by eliminating data brings out the shortcomings we mention earlier in this paper. The English noun *ass*, for example, can refer either to a body part, a donkey or a stupid/annoying person.[8] Since in some instances it may be considered problematic, it might be added to the blacklist. In that case, all sentences containing the word *ass* are removed from the corpus regardless of the word's meaning. This is not ideal for either lexicographers, who want to illustrate all the meanings of a word in a dictionary, or teachers, who should be given the choice to decide what they want to use for teaching, considering the

[8] https://www.macmillandictionary.com/dictionary/british/ass (30. 8. 2022).

students' characteristics, such as level, age, and background and relevance to the course topic.

Building on GDEX, Stanković et al. (2019) adopted machine learning to identify good candidate examples for Serbian. First, they analyzed lexical and syntactic features in a corpus compiled of illustrative examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary. They then identified 14 features relevant for the task (character-based, token-based and syntactic features) and prepared a gold dataset of good examples. Sentences from the prepared dataset, represented as feature-vectors, were used for a supervised machine learning model, which was then used in a GDEX classifier for contemporary Serbian sentences. A decision-tree classifier trained on the data predicted whether a certain corpus sentence is a good candidate for an illustrative example for the given dictionary headword or not, with an accuracy of 83% for both positive and negative samples (Šandrih, 2020).

Another tool to automatically identify good examples based on a series of criteria and using both rule-based and machine-learning approaches is HitEx. The combined approach was designed to assess the readability and suitability of (initially coursebook) material for teaching Swedish as L2 (Pilán et al., 2013, 2014; Pilán et al., 2016). For this task, 61 features of different types were used: length-based (e.g. number of tokens and characters), lexical (e.g. CEFR[9]-annotated wordlists), morphological (e.g. part-of-speech), syntactic (dependency relation tags), and semantic features (e.g. number of senses of a specific word). Candidate sentences were first ranked according to these features, and the 100 highest-ranked sentences were given to the machine-learning model for classification. The sentences were classified according to their proposed suitability for students at a certain CEFR level, and returned in the order of their heuristic ranking. Using the complete feature set at the document level, the tool obtained 81% accuracy, however, the classification accuracy for sentences was only 63.4%, presumably because the amount of context was too limited for the features to capture differences between the sentences.

[9] Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press (2001).

Taken together, it can be concluded that the creation of pedagogical corpora can be challenging in at least two ways: 1. manually monitoring large amounts of texts is extremely time-consuming, and consequently, expensive; and 2. automatization of processes to support compilation has limitations due to the very nature of language. As mentioned above, one of the main shortcomings of rule-based approaches to automatic corpus cleaning, such as the method used for the development of SKELL corpora, lies in the fact that many of the words in the blacklists used as a reference to exclude sentences from a corpus are polysemic. Moreover, the automatic identification of structural problems does not take into consideration language variation. Finally, the NLP field has acknowledged that further investigation and development are needed in order to include contextual aspects to automatic offensiveness identification, with current methods still falling short.

As a result, human verification of sentences is required. More importantly, from our perspective, pedagogical corpora should be *labeled* for potentially problematic content rather than *cleaned* from it. In order to streamline the verification of the sentences for the creation of problem-labeled pedagogical corpora, we have decided to ask the crowd for help. It was in this context that the Crowdsourcing Corpus Filtering for Pedagogical Purposes project was created.

3 Crowdsourcing and gamification

Crowdsourcing is a technique for gathering data or performing large-scale tasks which is often based on the framework of collective intelligence (Lévy, 1997). Concepts related to crowdsourcing include co-creation, open innovation, and user innovation (Chesbrough, 2006; Prahalad and Ramaswamy, 2000; Von Hippel and Katz, 2003). The benefits of crowdsourcing have been thoroughly established (Aitamurto et al., 2011; Buecheler et al., 2010; Lew, 2014; Morschheuser et al., 2017; von Ahn and Dabbish, 2008), and success stories can be found in various fields, from astronomy (e.g. Zooniverse; Simpson et al., 2014) to business. Language-related use of crowdsourcing is found in NLP (e.g. for tasks such as named entity recognition and entity linking), but also in fields such as lexicography (e.g. Arhar Holdt et al., 2018; Kosem et al., 2018) and more recently in language learning.

The role of crowdsourcing and its potential in language education has been investigated by enetCollect (the European Network for Combining Language Learning and Crowdsourcing Techniques), a large European network project funded as a COST action. The action addressed the pan-European challenge of fostering the language skills of all citizens regardless of their social, educational, and linguistic backgrounds. Its focus was on exploring the possibilities of how to use crowdsourcing to enhance the production of learning materials to cope with both the increase in demand for learning a second language (for migration, business, and tourism purposes), and the demand for more accessible materials in the many languages that are of interest to learners.

As the enetCollect research has confirmed, combining crowdsourcing and language learning is not a new undertaking, and it is possible to merge them to mass-produce language resources for any language in which a crowd of language learners can be involved (Arhar Holdt et al., 2021; Bédi et al., 2019; Lyding et al., 2018; Nicolas et al., 2020). Several language learning portals based on crowdsourcing have gathered huge multilingual audiences. Although this paper is not the platform for a detailed presentation of any of these portals, we offer some data to provide an insight into the scale of the crowd they were able to reach between 2017–2018 (Gorovaia, 2018). Rosetta Stone, the oldest of the portals and founded in 1992, attracted 75,720,000 users. Babbel, which opened in 2007, gathered 20,000,000 users. Mango Languages, launched in 2007, attracted 300,000 users. LiveMocha, which began in 2007, had 12,000,000 users in 2016. Busuu, which started in 2008, reached an audience of 70,000,000, while Duolingo, launched in 2011, had 300,000,000. Duolingo is notable for having built one of the world's most popular language-learning apps while hiring only a handful of language experts. Each day, it provides millions of sentence examples and exercises to users, almost all of them created by its 300 million or so volunteers. All of these portals are educational business entities, which confirms that educational businesses are able to attract users. The content they provide may facilitate and improve teaching, and crowdsourcing may be used to help to create resources for additional educational areas or new languages.

An important aspect of crowdsourcing is crowdsourcer motivation, i.e. finding the best method for a specific crowdsourcing task that will attract enough people and ensure their participation until the end of the task. Lew (2014) states there are three types of motivation: psychological, social, and economic. Psychological motivation is driven by the expectation that participants will find the task psychologically satisfying or personally fulfilling. Social motivation relies on the desire of individuals to interact with others who share similar interests, contribute to the community, or improve a certain skill. Economic motivation involves financial benefits for the participants who can, for example, receive micropayments for successfully completed tasks (see Rumshisky, 2011).

A method that relies heavily on the psychological motivation of the participants, and aims to make completing the task pleasurable, is a game with a purpose (GWAP). GWAPs are “games that are fun to play and at the same time collect useful data for tasks that computers cannot yet perform” (Hacker & von Ahn, 2009, p. 1208). They have been increasingly used to crowdsource data to create lexical infrastructures of different types, and examples of GWAP include Dodiom (Eryiğit et al., 2022), Jeux de Mots (Lafourcade, 2007), Phrase Detective (Chamberlain et al., 2008), ZombiLingo (Guillaume et al., 2016), Jinx (Seemakurty et al., 2010), Game of Words (Arhar Holdt et al., 2020), and Cipher (Xu and Chamberlain, 2020).

In sum, when applied in the right circumstances, to the right crowd, and using a method and motivation best suited for a specific task, crowdsourcing can deliver very useful outcomes. It is, however, important to note that successful completion of a crowdsourcing task also requires a careful analysis of the related goals, the problem-solving environment, the expertise required, complementary activities and capabilities, and the competitive environment (Aitamurto et al., 2011; Morschheuser et al., 2017; Pe-Than et al., 2015).[10]

[10] There is evidence that crowdsourcing tasks are sometimes not well-defined, or are given to the “wrong” unskilled/untrained crowd that cannot complete the task.

4 The crowdsourcing for language learning game – CrowLL

4.1 Background

In 2019 we carried out an experiment on the use of crowdsourcing for corpus filtering in which we asked the crowd to identify offensive sentences for pedagogical purposes (Kuhn et al., 2021). The sentences to be judged were automatically extracted from corpora of Brazilian Portuguese, Dutch, Serbian, and Slovene, and the participants were from Brazil, Netherlands, Serbia, and Slovenia, respectively. This study has revealed that the crowd considered to be offensive sentences which, although not directly formulated as such, expressed misogyny, religiously-offensive content, violence towards children, or contained topics related to war and politics. The study has also shown that sentences with explicitly rude content were not necessarily considered to be inappropriate.

These revealing results support our understanding that offensiveness and sensitivity are subjective and that their expression through language involves mechanisms that go beyond the explicit use of swear words. The findings of the experiment have also indicated that crowdsourcing seems to be an adequate technique to deal with such a contentious topic. Nevertheless, the traditional approach used in the experiment, namely, via the Pybossa crowdsourcing platform,[11] was considered to be rather unappealing by the participants, and thus we decided to experiment with the Games with a Purpose approach. This has also been adopted to address a similar topic by High School Super Hero (Bonetti and Tonelli, 2020, 2021), a game currently under development that focuses on the linguistic annotation of abusive language to collect data for hate speech detection. However, while GWAPs have been used for various purposes in different fields (cf. section 3), the use of games to monitor offensiveness and sensitive content in authentic examples is still in its infancy.

One additional point should be made. Given that some participants in our experiment considered sentences with structural problems inappropriate for language learning, we decided to include this type of problem in the game, in addition to offensiveness and sensitive content.

[11] https://pybossa.com (30. 8. 2022)

4.2 CrowLL

The Crowdsourcing for Language Learning (CrowLL) game is under development for Brazilian Portuguese,[12] Dutch, Estonian, and Slovene. The idea for CrowLL was originally inspired by the Matchin game (Hacker and von Ahn, 2009). In this, two players compete with each other to guess which of the two pictures that are shown to them their opponent will choose. If their predictions match, they score points. According to Hacker and von Ahn (2009), this game mechanism can be used to elicit user preferences. Harris (2014) has also shown that asking about the partner's opinion leads to better results with regard to both parties giving the same answers than when the players make decisions based on their own opinions. Given that our interest in the game is to find out what examples players consider to be offensive, have sensitive content or have structural problems, this in fact includes asking players to make judgements that can vary from one person to another. Thus, the selection of a game mechanism that elicits the users' opinions and preferences seems to be a viable solution.

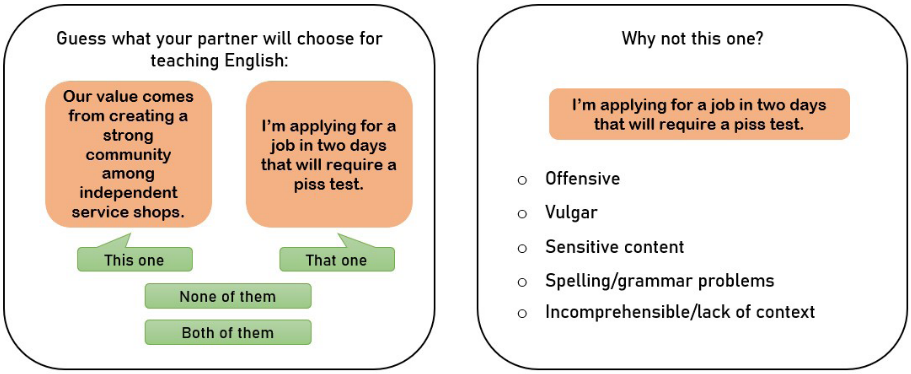
Nevertheless, we have also opted to offer a single-player mode. Although with this mode, the game might not benefit from the advantages put forth by the dual-player mode, the organizational factors have led us to opt to start with the development of the solo mode. Namely, the computational implementation of the solo mode requires less time and is, consequently, less expensive.

In terms of the type of crowdsourced work, Morschheuser et al. (2017) propose a categorization of crowdsourcing types based on the framework presented by Geiger and Schader (2014). Based on this, we consider CrowLL as a crowdrating game, given that "crowdrating systems commonly seek to harness the so-called *wisdom of crowds* (Surowiecki, 2005) to perform collective assessments or predictions. In this case, the emergent value arises from a huge number of homogeneous 'votes'" (Morschheuser et al., 2017, p. 27).

With CrowLL, the definition of whether a sentence is problematic or not, to which category of problem it belongs, and what constituent part of the sentence is problematic will emerge from the majority consensus.

[12] European Portuguese will be included later.

CrowLL will be a collaborative game with three levels. In level 1 (I'm curious!), players identify appropriate sentences for language teaching (Figure 1). In level 2 (I'm eager to help!), they categorize the sentences that have not been chosen (i.e., considered to be inappropriate), ranging from grammar/spelling problems to issues of offensiveness and sensitivity (Figure 2). In level 3 (I'm feeling enthusiastic!), players mark in the sentence what they consider to be problematic. Players can choose to play the full game cycle (all levels), a combination of two levels, or only one level.


Guess what your partner will choose for teaching English:

Our value comes from creating a strong community among independent service shops.

I'm applying for a job in two days that will require a piss test.

Why not this one?

I'm applying for a job in two days that will require a piss test.

Figure 1: Levels 1 and 2 of CrowLL.

Initially, the dual-player mode should involve two human players. However, 'the cold-start problem', i.e., the lack of an opponent to start a game (Dulačka et al., 2012 as cited in Pe-than et al., 2015) has made us think of alternatives. Indeed, it can be a challenge to find a playing partner at any given time, especially in the case of small language communities such as some of those for which this game is being developed. Therefore, we propose two solutions, a synchronous and an asynchronous mode. In the synchronous mode, players will play against bots with pre-recorded answers. Players are rewarded when their predictions match the pre-recorded answers. With the asynchronous approach, we will offer delay mechanics (Pe-Than et al., 2015). Here, players will choose packages containing sentences previously judged by others, and players will be rewarded once their answers are confirmed by others at a later time. Depending on whether the game is played in single- or dual-player mode, some of the questions will have to be changed and the scoring will also be different.

We have several ideas with regard to incentive through scoring mechanisms, ranging from offering an individual score that stems from consecutive work, to keeping a record of a cooperative score that shows the agreement of the player in teams/partnerships (so-called normalization motivation, according to Preist et al., 2014), including displaying scoreboards of the player's country's ranking position in comparison to the other countries (Olympic Games style). In this way, the game can be competitive on an individual level, while at the same time cooperative on the team level.

5 Methodology of data preparation

In order to start testing the game so that adjustments and development can be made before the official public release, we have decided to create an initial dataset of 10,000 sentences per language. The data extraction procedure involves – from each of the source corpora – the use of GDEX and a lemma list to extract the sentences. However, before proceeding with the extraction, a series of actions are required:

1. Definition of the source corpora from which sentences will be extracted;
2. Provision of pedagogically oriented GDEX configurations;
3. Creation of lemma lists to extract sentences from the corpora.

Next, we will explain each action in more detail.

5.1 Source corpora

One of the crucial guidelines for choosing our source corpora was that they were at least in some part openly available. This way, the resulting labeled datasets can be shared with and used by others. This decision aims at contributing to overcome one of the main problems in the area of language resource development, namely the lack of open-source data for many languages, as noted, for example, by Vajjala (2022) with regard to research on automatic readability assessment.

For Dutch and Brazilian Portuguese, we use the respective corpora of the Timestamped JSI web corpus, which is a family of web corpora created from IJS newsfeed by the Jozef Stefan Institute, in Slovenia, for 18 languages (Trampuš and Novak, 2012). Corpora in this family comprise news articles continuously crawled from RSS feeds. Both corpora are available in Sketch Engine. The Dutch corpus covers texts originating from the Netherlands and Belgium from 2014 to 2021. The whole corpus, totaling approximately 1.3 billion words, will be used. The Portuguese corpus covers texts from 2014 to 2021, published online in different countries, totaling over 4.5 billion words. As we are first developing CrowLL for Brazilian Portuguese, we only used texts marked with Brazil as a source country, thus making a subcorpus of 3,202,820,993 words.

For Estonian we use the Estonian National Corpus 2021 (Koppel and Kallas, 2022), which is the latest and largest corpus of written texts of modern Estonian. The texts span the period from 1990 to 2021. The most extensive part of the Estonian National Corpus 2021 is the Estonian Web Corpora, i.e. texts crawled from the web. It contains eleven sub-corpora (i.e. Web 2013, Web 2017, Web 2019, Web 2021, Feeds 2014-2021, Wikipedia 2021, Wikipedia Talk 2017, the Open Access Journals (DOAJ), Literature, Balanced Corpus, and the Reference Corpus) totaling 2.3 billion words.

For Slovene we use Gigafida 2.0 (Krek et al., 2020), the most recent version of the reference written corpus of Slovene. It contains 38,310 texts and 1,134,693,333 words. The texts span the period from 1991 to 2018, and cover newspapers, internet resources (the texts collected using the IJS Newsfeed service; Trampuš and Novak, 2012), magazines, fiction, non-fiction (such as textbooks), and various other texts. Newspaper texts represent nearly half of the corpus (47.8% of tokens), followed by internet texts (28%) and magazines (16,5%).

5.2 Pedagogically oriented GDEX configurations

In section 2, we introduced GDEX (Good Dictionary Examples) (Kilgarriff et al., 2008). While the Sketch Engine team has made general GDEX configurations for a number of languages available on their platform, GDEX configurations can be specially devised to better fit specific purposes, depending on the objectives of the project at hand. As the objective of the CrowLL game is to have the crowd help to create problem-labeled corpora for language learning, the sentences to be presented to the crowd for labeling have to be previously prepared to fit the pedagogical purpose. In order to do this automatically, we have opted to use pedagogically oriented GDEX configurations.[13] Slovene and Estonian have adopted configurations that have been previously devised for pedagogical purposes, while Dutch and Portuguese have built on existing pedagogically oriented configurations.

The Slovene GDEX configuration was originally devised for lexicographic projects at the Centre for Language Resources and Technologies, and more specifically this includes the Slovene Lexical Database and Collocations Dictionary of Modern Slovene (Gantar et al., 2016; Kosem et al., 2011; Kosem et al., 2012; Kosem et al., 2013). The initial lexicographically oriented GDEX configuration was also used for pedagogical purposes, i.e. in the preparation of examples for exercises in the Pedagogical Corpus Grammar (Arhar Holdt et al., 2011; Arhar Holdt et al., 2017).

The Estonian configuration was originally devised for extracting examples for the Estonian Collocations Dictionary (Kallas et al., 2015) aimed at learners of Estonian as a foreign language on the B2-C1 level. The configuration was later used to create a corpus – the etSkELL corpus – that only includes sentences that meet all the pre-defined criteria (i.e. have a GDEX score above 0.5). The etSkELL corpus is now also used as a source corpus in the Estonian SKELL, as well as in the language portal Sõnaveeb for presenting the users a set of authentic corpus examples (Koppel, Kallas et al., 2019; Koppel, Tavast et al., 2019; Koppel, 2020).

For Dutch, special GDEX configurations were developed in the context of the project Woordcombinaties[14] (Word combinations) which is

[13] While we are aware that some fields of the NLP area are devoted to related issues that could potentially contribute to the automatic identification of pedagogical sentences or even to enhancing GDEX configurations, such as automatic normalization, automatic error detection, and readability assessment, a decision was made to adopt or adapt existing versions of GDEX configurations as a first step towards identifying candidate sentences for pedagogical purposes. Moreover, and relatedly, it is outside the scope of this paper to explore other approaches to further enhance GDEX configurations.

[14] https://woordcombinaties.ivdnt.org/

targeted at advanced language learners (Colman and Tiberius, 2018). For this project, a minimal configuration was defined only using the classifiers not surrounded by round brackets in Table 1, as well as a more restrictive configuration also incorporating the classifiers in between brackets. Lexicographers in the project Woordcombinaties have access to both configurations, and both are being used. For the initial dataset for CrowLL a combination of the two configurations will be used, to bring the Dutch configuration more in line with the configurations for the other languages.

The GDEX configuration that was devised for academic Portuguese in the context of a design of a dictionary for university students (Kuhn, 2017) is the basis for the development of the configuration for data extraction. Given the pedagogical aspect of the academic configuration, adjustments were mostly made according to the characteristics of the type of language, i.e., from academic to general language. Additional development might take place in the future.

Out of the four languages, Estonian has carried out a study especially developed to evaluate its GDEX configuration, while the other languages have relied on the successful and extensive use of the configurations by lexicographers and other users. The output of the Estonian GDEX configuration has been assessed by lexicographers and L2 learners of Estonian. The two types of annotators performed a task to determine whether authentic and unedited corpus sentences would be suitable as example sentences for learners' dictionaries on the B2-C1 level. The results of the assessment showed that both types of annotators considered as many as 85% of the corpus sentences chosen by the Estonian GDEX configuration as good examples, confirming the premise that the methodology GDEX uses to select the examples is reliable (Koppel, 2019). The pre-existing Slovene GDEX configuration adopted in our methodology has been widely tested by lexicographers and successfully implemented in the development of other resources, such as a pedagogical grammar, as noted above. For Dutch, the configuration used is a combination of two configurations that have been tested extensively by a team of lexicographers within the Woordcombinaties project. The Portuguese GDEX configuration for the game is actually the only one that has not been previously tested, as it consists of an adaptation of an existing configuration. However, the configuration that was used as the basis has been carefully devised and used by other users (for example, when integrated in the Sketch Engine tool).

As mentioned in section 2, GDEX configurations consist of two types of classifiers: hard and soft. Sentences are evaluated against those classifiers and scores are calculated accordingly, based on the weighted sum. Hard classifiers serve to severely penalize sentences, separating the good from the (really) bad ones. Soft classifiers, on the other hand, penalize or give bonuses to the sentences, thus contributing to ranking qualitatively more similar sentences. For the present project, some classifiers are used in all languages, while others are language-dependent. Table 1 provides an overview of the classifiers used in the configurations of the four languages of the game.

Hard classifiers (in bold in Table 1) mean that the evaluation of these features in the sentences weighs heavily on their score. A sentence must start with a capital letter and finish with a period, an exclamation mark or a question mark to be considered a **whole sentence**. For pedagogical purposes, it is crucial that only whole sentences are extracted from the source corpora. The **blacklist – illegal characters** classifier is used to detect the sentences containing strings with unwanted characters such as parts of the program code (<tag>) or URLs (//), because such sentences are not wanted in pedagogically oriented content. Spam texts are usually machine-generated, and thus are not appropriate for language learning. With the **blacklist – spam** classifier, sentences containing words in this blacklist get a very low score. In addition to spam texts, other characteristics of texts found on the web can be counterproductive for pedagogical purposes, such as the presence of typos and misspellings. In order to filter those sentences out, a **minimum frequency for tokens** is established. Another aspect to be considered in a pedagogical example is its length. Very long sentences can compromise intelligibility, i.e., “examples that are intelligible (to the users) are those that are not too long and do not contain complex syntax or rare or specialized vocabulary” (Kosem et al., 2019, p.120), while very short sentences might lack context and lose informative value (ibid.). Thus, sentences that do not fit between the **minimum and maximum sentence length** values get a high penalty.

Table 1: *Overview of the classifiers used in pedagogically oriented configurations for Slovene, Dutch, Estonian and Brazilian Portuguese (adapted from Kosem et al., 2019)*

Classifier	Slovene	Dutch	Estonian	Brazilian Portuguese
whole sentence	X	X	X	X
blacklist - illegal characters	X	X	X	X
blacklist - spam	X		X	X
minimum frequency for tokens	X (3)	X (20)	X (5)	X (5)
minimum and maximum sentence length	X (7 and 60)	X (<30)	X (4 and 20)	X (7-30)
graylist – bad words	X	(X)	X	X
optimal sentence length	X (15-40 tokens)	X (9-12 tokens)	X (6-12 tokens)	X (10-18 tokens)
penalty for long words	X (longer than 12 characters)			X (longer than 12 characters)
penalty for rare characters	X	X	X	X
penalty for capital letters	X	X (part of rare characters)	X	
penalty for tokens with mixed symbols	X	X	X	X
penalty for proper nouns	X	(X)	X	X
penalty for pronouns	X		X	
penalty for sentence initial words	X (list of words provided)	(X)	X	
penalty for sentence initial phrase	X	(X)	X	
penalty for sentence initial tags		(X)	X	
penalty for rare words	X (fewer than 1,000 hits in the corpus)	(X)	X (fewer than 1,000 hits in the corpus)	X (fewer than 500 hits in the corpus)
penalty for commas	X (3 or more)		X (2 or more)	X (2 or more)
penalty for abbreviations		(X)		
penalty for sentences without a finite verb			X	
penalty for more than two occurrences of *que* (that, which)				X

As can be seen in Table 1, the use of soft classifiers (in non-bold in Table 1) varies among the languages, with optimal sentence length and graylist – bad words being used in all of them. Sentences within the **optimal sentence length** get a higher score than the other sentences outside this interval, and are thus ranked higher up among all the sentences. Length values vary from language to language, and have been defined based on what each language considered to be the optimal sentence length interval for pedagogical purposes.[15]

Words in the **graylist – bad words** are compared against the sentences in a corpus and, if any word is found, the sentence is penalized. Evaluation of the settings has shown that this penalization is enough to push such potentially problematic sentences lower down the ranking, but still not too low in case the penalization is unjustified (polysemous words, etc.). This means that sentences with higher scores (in the upper part of the list) will probably not contain explicitly offensive words, that sentences with very low scores (at the bottom of the list) will probably contain offensive words, and that the ones in the middle might or might not contain them. While we want the players to assess the sentences from the upper and lower parts and possibly confirm that they are non-problematic and problematic, respectively, one of the most interesting contributions from the players will be the evaluation of sentences pertaining to exactly this grey, middle area, where one can expect to find explicitly offensive lemmas, offensive lemmas that are polysemous and not being used in an offensive manner, offensive sentences with no overtly offensive lemmas, and sentences with sensitive content. This type of evaluation is still not well performed by computers, so we need humans to do it.

The Slovenian graylist contains 1,909 words (nouns, adjectives, verbs and adverbs) that were identified in several lexicographic and linguistic projects as vulgar or (potentially) offensive. For Portuguese, there are two graylists of explicitly offensive and vulgar items (nouns, adjectives and verbs), one consisting of lemmas and another one of word forms and strings (e.g., *fodid.+*), totaling 91 items. These lists result from manual

[15] It was observed that different languages differ in the average sentence length due to various reasons such as word formation (e.g. compounds in Estonian are mainly written as one word, as opposed to two or more words in Slovene), existence of articles etc.

evaluation and editing of the list of taboo lemmas and word forms created by the Sketch Engine team for the default Portuguese GDEX that they have devised. Words related to cultural aspects, such as those related to religion or nationalities, that were not offensive or vulgar but had probably been included because of their potential to spark hate speech, were discarded. In addition, new offensive or vulgar items were added, but further editing can be carried out if necessary. The Estonian graylist contains 1,472 words (nouns, adjectives, verbs), consisting of words tagged as vulgar, offensive, colloquial, and slang in the EKI Combined Dictionary (Langemets et al., 2022), swear words in foreign languages (e.g. *fuck*), their adapted variants (e.g. *fakk*, *pohui* 'похуй'), and words written differently from the written language norm. The Dutch configuration uses a graylist of 93 words which is based on words labeled as vulgar or offensive in the Algemeen Nederlands Woordenboek.[16] If needed, the Dutch graylist will be further refined in the future.[17]

Other classifiers relevant in the context of language learning are **penalties for long words**, **rare characters**, **tokens with mixed symbols** and **capital letters**. This is based on the assumption that longer words, too many rare characters and capital letters as well as the occurrence of non-words have an impact on reading complexity. For pedagogical purposes, a penalty can also be given to **proper nouns** in order to give priority to sentences without (or with few) of these, as in many cases the named entities in those sentences might not be known to the learners. The same applies to **abbreviations** which learners may not necessarily be familiar with. Penalizing **pronouns** can also help, as sentences with many pronouns are often too anaphoric and lack context for proper understanding.

[16] https://anw.ivdnt.org/search (30. 8. 2022). Note the ANW is a dictionary under construction, and thus new words (including words labelled as vulgar or offensive) are continuously being added. The current GDEX configuration for Dutch uses the words labelled as vulgar or offensive in the ANW at the time the GDEX configuration was defined for the project Woordcombinaties.

[17] As can be noticed, there is a considerable difference between the number of lemmas in the graylists for different languages. More thorough studies on problematic vocabulary were conducted for Slovenian and Estonian, and more extensive word lists were obtained as a result. It should be noted that these graylists contain lemmas that are problematic only in part, e.g. in one of their senses. Consequently, the penalization of sentence(s) containing the word(s) is milder. Using different approaches to graylists will open possibilities to compare them at the end of the study.

Another type of classifier uses lists containing **words** and **phrases** that should not occur **in** a **sentence-initial position**. These words and phrases are heavily penalized because in previous manual evaluations of extracted sentences for Slovene, Estonian and Dutch, several sentence-initial words and phrases were identified that are a good signal that the sentence is contextually dependent on the previous sentence(s), and is thus less suitable to be used as a standalone component for pedagogical purposes. Similarly, certain **sentence-initial tags** can be penalized, e.g. conjunctions, because sentences starting with conjunctions are often anaphoric.

Furthermore, sentences containing less frequent words tend to be considered inadequate to serve as examples of language use in pedagogical contexts, as such words are likely not known to the learners and might act as a distraction. The **penalty for rare words** classifier penalizes sentences with words whose frequency is below a certain threshold, so these sentences get lower scores. The use of too many commas in a sentence might be indicative of complexity, so the **penalty for commas** is a classifier that penalizes sentences if they have more than a defined number. A **penalty for sentences without a finite verb** can help to filter out less typical sentences. The grammar of the Estonian language (Erelt and Metslang, 2017), for instance, states that a typical sentence contains a finite verb and phrases (collocations) that go with the verb.

Portuguese adopts a separate **penalty for more than two occurrences of *que*** (*that*, *which*). This classifier has been created to avoid sentences with too many subordinate or relative clauses, because high syntactic complexity makes understanding more difficult, which is something to be avoided in pedagogical examples.

5.3 Lemma lists

To ensure at least partial comparability of the multilingual results, we decided to extract the data using lemmata, comparable across the participating languages. For this purpose, we first prepared a list of 100 words in English using the criteria described below. In the second step, we translated the list to Slovene, Brazilian Portuguese, Dutch, and

Estonian, reporting on problems with translation equivalents, as well as their frequency in the corresponding source corpora. We discuss some of these issues in Section 6.

We wanted to include lemmata that were of different relevance for labeling in the context of the CrowLL task: (a) words that were clearly (on the surface and in the vast majority of the meanings) offensive or vulgar, for example: *nigger, whore, bitch, retarded, to fuck, to piss*; (b) words that were offensive or vulgar in some of the meanings, as well as words with potentially sensitive content, for example: *cow, drunk, suicide, fanatic, depressed, to molest*; (c) words that would typically not be considered offensive, vulgar or sensitive from the perspective of our labeling task, for example *year, world, service, new, to say, to see.* Vocabulary from the first group would typically make it to blacklists, and thus a blacklist-based methodology would automatically filter out corpus occurrences with these words before they would be included in any teaching material. Here, we are including it to test the hypothesis that these corpus occurrences would also be marked as inappropriate by the crowd. On the other hand, non-problematic words are included to test the complementary premise. The most interesting for our task, however, are words in group (b). The lemmata list thus includes 20 words from groups (a) and (c) and 60 words from group (b).

The seed lemmata were selected using the translation into English of a list of words that were identified during the creation of a GWAP called Game of Words (Arhar Holdt et al., 2021). This game prompts the players to provide synonyms and collocations for different Slovene words, with the implicit purpose to clean the noise from two automatically created databases comprising openly available lexical information for Slovene. As the game is aimed at young(er) users, not only vulgar and offensive words were removed from the list of potential prompts, but also words with sensitive content that could cause the player unnecessary discomfort. The criteria for removal were based on existing resources, such as dictionaries, and privately compiled lists by researchers or journalists (ibid., p. 43). Semantically, the removed words covered a) human features, such as race, nationality, gender, age, sexual orientation, religious and political beliefs, migration status, social status, education, handicap, bodily and mental features etc., as well as b) sensitive topics, such as violence, illness, death, addiction, sex, excretions, etc. Offensive, vulgar, and potentially sensitive words for CrowLL were selected based on these categories, while non-problematic words were chosen from the most frequent words in English Web 2020, available on Sketch Engine.

The majority (50) of the included lemmata are nouns, 25 are verbs and 25 are adjectives. An example of seed lemmata with labels and translations is provided in Table 2.

Table 2: *Common lemma list and its translations to Slovene, Estonian, Brazilian Portuguese and Dutch*

Category	Type	English	POS	Slovene	Estonian	Brazilian Portuguese	Dutch
Race	B	black-skinned	A	temnopolt	mustanahaline	negro	zwart
Race	B	native	N	domorodec	pärismaalane	índio	autochtoon
Race	B	racist	A	rasističen	rassistlik	racista	racistisch
Race	A	nigger	N	črnuh	neeger	crioulo	neger
sexual orientation	B	homosexual	A	homoseksualen	homoseksuaalne	homossexual	homoseksueel
sexual orientation	B	straight	A	heteroseksualen	heteroseksuaalne	heterossexual	heteroseksueel
sexual orientation	B	lesbian	A	lezbičen	lesbiline	lésbica	lesbisch
sexual orientation	A	faggot	N	peder	pede	bicha	flikker
violence	B	to murder	V	umoriti	mõrvama	assassinar	vermoorden
violence	B	brutal	A	brutalen	brutaalne	brutal	brutaal
violence	B	to bully	V	ustrahovati	kiusama	intimidar	intimideren
violence	B	to torture	V	mučiti	piinama	torturar	martelen
violence	B	to rape	V	posiliti	vägistama	estuprar	verkrachten
violence	B	to beat	V	pretepati	peksma	bater	slaan
violence	B	to molest	V	zlorabljati	ahistama	molestar	lastigvallen
violence	B	to shoot	V	ustreliti	tulistama	atirar	schieten
non-problematic	C	time	N	čas	aeg	tempo	tijd
non-problematic	C	way	N	način	viis	maneira	manier
non-problematic	C	to include	V	vključiti	sisaldama	incluir	omvatten
non-problematic	C	good	A	dober	hea	bom	goed

As mentioned at the outset of this section, the data extraction procedure to obtain 10,000 sentences is meant for game development and initial tests. More specifically, the procedure will be performed as follows. We will use GDEX configurations to extract the top 200 sentences per lemma of the lemma list so that we have a buffer in case of duplicates. We will then verify those 20,000 sentences and reduce them to 10,000 sentences per language.

Once we have this data, we will proceed with manual annotation of the sentences with the labels from the game (non-problematic/problematic; category of the problem), which will allow us to evaluate the labeling system and the quality of the input data, and propose adjustments to the resources and the game if necessary. These annotated sentences will comprise manually annotated pedagogical corpora, and will be available as part of the CLARIN Language Resources Family. They will also be fed into the game to be used for scoring mechanism development, such as the scores given by comparison with other players and asynchronous play, for implementation of the dual-player mode, as pre-recorded answers for a bot, and as input data for the game.

When the game is launched, additional data will be required as input. The extraction of this data will follow a slightly different approach, given that we want the crowd to label as many sentences from the source corpora as possible. With the source corpora, pedagogically oriented GDEX configurations, and tested labeling system and gameplay, data input for the game will be extracted as follows. First, we will GDEX the corpus, i.e., run the GDEX configuration to assign GDEX scores to all sentences in the corpus. We will then extract sentences in batches, with varying GDEX scores, i.e., a certain number of sentences with the highest scores, medium scores and low scores. These sentences will be input into the game for players to play. Once the game is tested with actual players, an evaluation of the methodology of data preparation can be carried out.

6 Analysis and discussion

One of the main aspects that might have an impact on the results of the initial test with annotation of 10,000 sentences is that the resources that were used for data preparation present different levels of development. While Estonian and Slovene use source corpora that have been carefully compiled in the context of other projects, with rich metadata and advanced annotation, Dutch and Portuguese use automatically compiled web corpora with no human curation and POS-tagged by the Sketch Engine team. It should be acknowledged that these differences in the development of the resources might influence the quality of the input data (extracted sentences), with consequent reflection on the quality of the output data (annotated sentences).

Preparing the common lemma list posed many challenges, becoming an iterative process in which English words were proposed, translated to the target languages and then – based on the suitability of the translation equivalents – accepted or replaced. A discussion was needed if for one or more target languages a translation equivalent was not suitable from the perspective of form, meaning, connotation or frequency.

To ease the data extraction, we aimed for a list of single-word lemmata for all target languages. We thus avoided English prompts that would require multiword translations. For example, for the English verb *to fuck off* not all languages had single-word translations (Slovene: *odjebati,* Estonian: *perse käima*, Portuguese: *ir se foder,* Dutch: *opsodemieteren*), therefore we replaced it with the verb *to fuck* (Slovene: *jebati,* Estonian: *keppima*, Portuguese: *foder*, Dutch: *neuken*). More permissive were our decisions when it came to the part-of-speech of the translation equivalents. For most of the cases, providing translation equivalents of the same POS was unproblematic. In rare instances where the POS of otherwise the most suitable translation candidate did not match, we kept it on the list. For example, some English adjectives in Estonian are actually case forms of a noun, e.g. *depressioonis* ‘in depression’ (not ‘depressed’). When examining the occurrences of the lemmata in the source corpora, we also noticed that some POS differences stemmed from the features of the taggers used to annotate the data (e.g., the Portuguese equivalent *retardado* for the English adjective *retarded* occurs erroneously tagged as verbs (participle) in the Portuguese corpus). While such problems would have to be considered when extracting the data, they did not influence the selection of the candidates for the common lemma list.

Important for the list was the connotation of the translation equivalents. When the target language did not have a translation equivalent with comparable sensitivity, the English word was replaced. For example, the English noun *bimbo* for an 'attractive but unintelligent or frivolous young woman' did not have a suitable single-word translation in Portuguese, so we replaced it with a (more offensive) *slut* (Slovene: *cipa*, Estonian: *libu*, Portuguese: *vagabunda*, Dutch: *slet*). Other semantic differences, such as nuances in the meaning(s) of the translated words were accepted, as we did not want to create a list that would be overly curated, artificial, and methodologically difficult to expand with further lemmata and to other languages. In situations where more semantically suitable translation equivalents were possible, we opted for the one that was less polysemic (for example, for the English noun *corpse,* we chose the Portuguese *cadáver* and not *corpo* which has a wider use).

Finally, the translation equivalents were checked for their frequency in the corresponding source corpora. According to our methodology, we needed at least 100 heterogenous corpus examples per lemma, but to have enough data to select from we aimed to extract 200. Especially in "cleaner" corpora, such as the Slovene source corpus Gigafida, the offensive and vulgar words were rare, but nearly all proposed lemmas had over 200 occurrences. We decided to keep the noun *asshole* with a Slovene translation *pezde* (198 occurrences in the Slovene source corpus) and replace the adjective *transsexual* (less than 10 occurrences in the Dutch source corpus) with a more frequently occurring *transgender*.

Once the game is fully operational, a series of issues need to be considered. For example, it is important to ensure the rapid implementation of the game's results into practice. This requires both a set of clear parameters on what a minimum number – as well as a maximum number – of user responses per example is, what level of agreement is required, etc., as well as automatic tools or algorithms for regular data analysis and summarization. All this helps to increase the quantity of crowdsourced data, as more examples can be added to the game (and at the same time the sufficiently examined ones removed) on a regular basis. Technical aspects should also be paid enough attention, meaning the server should have enough capacity and storage space to cater for heavy usage, which can partly be addressed by conducting rigorous stress tests before the launch of the game. Last but not least, a detailed promotion plan needs to be prepared in advance, including the steps on how to not only attract users, but also keep them long term.

7 Conclusions

In this paper, we proposed a methodology of data preparation for the development of the Crowdsourcing for Language Learning (CrowLL) game, from which data will be collected through crowdsourcing to create problem-labeled pedagogical corpora for Dutch, Estonian, Slovene, and Brazilian Portuguese. For this process a series of decisions had to be made, from the choice of source corpora, to GDEX configuration development and lemma list creation. By describing the methodology and reflecting on the challenges posed and solutions found, it is our intention to provide researchers sharing common interests with a model that can be applied to other languages, and potentially to other purposes.

The next steps of our project involve the extraction of sentences for the game, full implementation of the game, collection of answers (from actual players), statistical analysis of labeled data, and design and administration of a user survey to evaluate the game design and user experience. With the players' answers, we will compile problem-annotated corpora and develop other auxiliary language learning resources, such as SKELL for all the languages. After that, we plan to start the third stage of the project, in which we will use the problem-labeled corpora to create the basis for the future development of machine-learning training models to automatize identification and labeling of problematic content, thus contributing to the further and faster creation of pedagogical corpora.

Acknowledgments

The authors acknowledge the financial support received from the Portuguese national funding agency, FCT – Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020) and the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene, and project funding No. J7-3159, Empirical foundations for digitally-supported development of writing skills). The research received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731015. This study has also been supported by the CLARIN Resource Families Project Funding.

References

Aitamurto, T., Leiponen, A., & Tee, R. (2011). *The promise of idea crowdsourcing–benefits, contexts, limitations* [White paper]. Nokia Ideas project.

Arhar Holdt, Š., Kosem, I., & Gantar, P. (2017). Corpus-based resources for L1 teaching: The case of Slovene. In *Handbook on digital learning for K-12 schools* (pp. 91–113). Springer, Cham. doi: 10.1007/978-3-319-33808-8_7

Arhar Holdt, Š., Kosem, I., Krapš Vodopivec, I., Ledinek, N., Može, S., Stritar Kučuk, M., Svenšek, T., & Zwitter Vitez, A. (2011). *Pedagoška slovnica pri projektu Sporazumevanje v slovenskem jeziku: K16 – Standard za korpusno analizo slovničnih pojavov*. Ljubljana: Ministrstvo za šolstvo in šport: Amebis. Retrieved from http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik16/Kazalnik_16_Pedagoska_slovnica_SSJ.pdf

Arhar Holdt, Š., Logar, N., Pori, E., & Kosem, I. (2021). "Game of Words": Play the game, clean the database. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of the EURALEX XIX congress: Lexicography for inclusion, 7–11 September, Aleksandroupolis, Greece* (Vol I., pp. 41–49). Retrieved from https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p041-049.pdf

Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. *Proceedings of the eighth workshop on recent advances in Slavonic natural language processing, RASLAN 2014* (pp. 63–70). Retrieved from https://nlp.fi.muni.cz/raslan/2014/12.pdf

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. *CEUR Workshop proceedings,* 1–6. Retrieved from http://ceur-ws.org/Vol-2253/paper49.pdf

Bédi, B., Chua, C., Habibi, H., Martinez-Lopez, R., & Rayner, M. (2019). Using LARA for language learning: a pilot study for Icelandic. In F. Meunier, J. van de Vyver, L. Bradley & S. Thouësny (Eds.), *CALL and complexity: short papers from EUROCALL 2019* (pp. 33–38). Research-publishing.net. doi: 10.14705/rpnet.2019.38.982

Bonetti, F., & Tonelli. S. (2020). A 3D role-playing game for abusive language annotation. *Workshop on games and natural language processing* (pp. 39–43). Retrieved from https://aclanthology.org/2020.gamnlp-1.6

Bonetti, F., & Tonelli. S. (2021). Challenges in designing games with a purpose for abusive language annotation. *Proceedings of the first workshop on bridging human–computer interaction and natural language processing* (pp. 60–65). https://aclanthology.org/2021.hcinlp-1.10

Boulton, A. (2017). Corpora in language teaching and learning: Research timeline. *Language Teaching*, *50*(4), 483–506. doi: 10.1017/S0261444817000167

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47–64. doi: 10.1017/S0958344005000510

Buecheler, T., Sieg, J. H., Füchslin, R. M., & Pfeifer, R. (2010). Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. In H. Fellermann, M. Dörr, M. Hanczyc, L. L. Laursen, S. Maurer, D. Merkle, P-A. Monnard, K. Stoy, S. Rasmussen (Eds.), *Artificial live XII: proceedings of the twelfth international conference on the synthesis and simulation of living systems* (pp. 679–686). MIT Press. doi: 10.21256/zhaw-4094

Callies, M. (2019). Integrating corpus literacy into language teacher education. In S. Götz, J. Mukherjee (Eds.), *Learner corpora and language teaching* (pp. 245–263). John Benjamins Publishing Company. doi: 10.1075/scl.92.12cal

Chamberlain, J., Poesio, M., & Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. *Proceedings of the international conference on semantic systems (I-Semantics'08)* (pp. 42–49). Retrieved from https://www.jonchamberlain.com/media/doc/Chamberlain-2008Phrase.pdf

Chambers, A. (2016). Written language corpora and pedagogic applications. In F. Farr, L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 362–375). Routledge. doi: 10.4324/9781315657899.ch26

Chesbrough, H. W. (2006). *Open innovation: The new imperative for creating and profiting from technology.* Harvard Business School Press.

Colman, L., & Tiberius C. (2018). A good match: A Dutch collocation, idiom and pattern dictionary combined. *Proceedings of the XVIII EURALEX international congress: Lexicography in global contexts* (pp. 233–246). Retrieved from https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2952-1-10-20180820.pdf

Erelt, M., & Metslang, H. (2017). *Eesti keele süntaks.* Eesti keele varamu III. Tartu Ülikooli Kirjastus. Retrieved from https://dspace.ut.ee/handle/10062/70510

Eryiğit, G., Şentaş, A., & Monti, J. (2022). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering* (pp. 1–33). doi: 10.1017/S1351324921000401

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, *29*(2), 200–225. doi: 10.1093/ijl/ecw014

Gorovaia, N. (2018). *Behavior of users on the crowdsourcing platforms*. [Poster session]. EnetCollect WG3/WG5 meeting, October 24–25, Leiden, Netherlands.

Gries, S. (2009). What is corpus linguistics? *Language and Linguistics Compass*, *3*, 1–17. doi: 10.1111/j.1749-818X.2009.00149.x

Guillaume, B., Fort, K., & Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3041–3052). Retrieved from https://aclanthology.org/C16-1286

Hacker, S., & von Ahn, L. (2009). Matchin: eliciting user preferences with an online game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1207–1216). doi: 10.1145/1518701.1518882

Harris, C.G. (2014). The beauty contest revisited: measuring consensus rankings of relevance using a game. *Proceedings of the first international workshop on gamification for information retrieval – GamifIR@ECIR '14* (pp. 17–21). doi: 10.1145/2594776.2594780

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. *Proceedings of the eLex 2015 conference* (pp. 1–20). Retrieved from https://elex.link/elex2015/proceedings/eLex_2015_01_Kallas+etal.pdf

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36. doi: 10.1007/s40607-014-0009-9

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX international congress* (Vol. 1, pp. 425–432). https://tinyurl.com/yckr9w8s

Kilgarriff, A., Rychlý, P., Smrz, P., & D. Tugwell (2004). The Sketch Engine. *Proceedings of the eleventh EURALEX international congress, EURALEX 2004* (pp. 105–116). Retrieved from https://tinyurl.com/mvrp4ymy

Koppel, K. (2019). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks. *Lähivõrdlusi. Lähivertailuja, 29*, 84–112. doi: 10.5128/LV29.03

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele*. Doktoritöö, Tartu Ülikool. Retrieved from https://dspace.ut.ee/handle/10062/67138

Koppel, K., & Kallas, J. (2022). *Eesti keele ühendkorpus 2021*. doi: 10.15155/3-00-0000-0000-0000-08D17L

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V., & Michelfeit, J. (2019). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. *Proceedings of the eLex 2019 conference* (pp. 763–782). Retrieved from https://zenodo.org/record/3612933#.Yywd1XZBy70

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. *Proceedings of the eLex 2019 conference* (pp. 434–452). Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Kosem, I. (2012,). Using GDEX in (semi)-automatic creation of database entries [Conference presentation]. *SKEW-3, 3rd international Sketch Engine workshop*, *21–22 March, 2012.*

Kosem, I., Gantar, P., & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. *Proceedings of the eLex 2013 conference* (pp. 32–48). Retrieved from http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf

Kosem, I., Husák, M., & McCarthy, D. (2011). GDEX for Slovene. *Proceedings of eLex 2011* (pp. 151–159). Retrieved from http://www.dianamccarthy.co.uk/files/Kosemetal-paper.pdf

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography, 32*(2), 119–137. doi: 10.1093/ijl/ecy014

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: The reference corpus of written standard Slovene. *Proceedings of the twelfth language resources and evaluation conference* (pp. 3340–3345). Retrieved from https://aclanthology.org/2020.lrec-1.409

Kuhn, T. Z. (2017). *A design proposal of an online corpus-driven dictionary of Portuguese for university students* [Doctoral dissertation, Universidade de Lisboa]. Retrieved from http://hdl.handle.net/10451/32013

Kuhn, T. Z., Šandrih Todorović, B., Holdt, Š. A., Zviel-Girshin, R., Koppel, K., Luís, A.R., & Kosem, I. (2021). Crowdsourcing pedagogical corpora for lexicographical purposes. *Proceedings of the XIX EURALEX congress: Lexicography for inclusion* (Vol. II., pp. 771–779). Retrieved from https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol2-p771-779.pdf

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *Proceedings of SNLP'07: 7th international symposium on natural language processing*. Retrieved from https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883

Langemets, M., Hein, I., Jürviste, M., Kallas, J., Kiisla, O., Koppel, K., Leemets, T., ..., & Tubin, V. (2022). *EKI ühendsõnastik 2022*. doi: 10.15155/3-00-0000-0000-0000-08C0AL

Lévy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace*. Plenum Trade. New York.

Lew, R. (2014). User-generated content (UGC) in online English dictionaries. *OPAL, 4*, 8–26. Retrieved from https://pub.ids-mannheim.de//laufend/opal/opal14-4.html

Lyding, V., Nicolas, L., Bédi, B., & Fort, K. (2018). Introducing the European network for combining language learning and crowdsourcing techniques (enetcollect). In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds.), *Future-proof CALL: language learning as exploration and encounters– short papers from EUROCALL* (pp. 176–181). Research-publishing.net. doi: 10.14705/rpnet.2018.26.833

Morschheuser, B., Hamari, J., Koivisto, J., & Maedche, A. (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies, 106*, 26–43. doi: 10.1016/j.ijhcs.2017.04.005

Nicolas, L., Lyding, V., Borg, C., Forăscu, C., Fort, K., Zdravkova, K., Kosem, I., ..., & HaCohen-Kerner, Y. (2020). Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. *Proceedings of the 12th language resources and evaluation conference* (pp. 268–278). Retrieved from https://aclanthology.org/2020.lrec-1.34

Osborne, J. (2004). Top-down and bottom-up approaches to corpora in language teaching. language and computers. In U. Connor, T. A. Upton (Eds.), *Applied Corpus Linguistics*. A Multidimensional Perspective (pp. 251–265). Brill. doi: 10.1163/9789004333772_015

Pe-Than, E. P. P., Goh, D. H. L., & Lee, C. S. (2015). A typology of human computation games: an analysis and a review of current games. *Behaviour & Information Technology*, *34*(8), 809–824. doi: 10.1080/0144929X.2013.862304

Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. ArXiv. doi: 10.48550/arXiv.1603.08868

Pilán, I., Volodina, E., & Johansson, R. (2013). Automatic selection of suitable sentences for language learning exercises. *20 Years of EUROCALL: Learning from the past, looking to the future: 2013 EUROCALL Conference Proceedings* (pp. 218–225). Retrieved from https://aclanthology.org/W14-1821.pdf

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 174–184). Retrieved from https://aclanthology.org/W14-1821

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources & Evaluation*, *55*(2), 477–523. doi: 10.1007/s10579-020-09502-8

Prahalad, C. K., & Ramaswamy, V. (2000). Co-opting customer competence. *Harvard Business Review*. Retrieved from https://hbr.org/2000/01/co-opting-customer-competence

Preist, C., Massung, E., & Coyle, D. (2014). Competing or aiming to be average? Normification as a means of engaging digital volunteers. *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing (CSCW '14)* (pp. 1222–1233). doi: 10.1145/2531602.2531615

Reynaert, M. (2006). Corpus-induced corpus clean-up. *Proceedings of the fifth international conference on language resources and evaluation* (pp. 87–92). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/229_pdf.pdf

Römer, U. (2009). Using general and specialised corpora in language teaching: Past, present and future. In M. C. Campoy, B. Belles-Fortuno & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp.18–35). Continuum Publishing Corporation.

Šandrih Todorović, B. (2020). *Impact of text classification on natural language processing applications*. [Универзитет у Београду].

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). doi: 10.18653/v1/W17-1101

Seemakurty, N., Chu, J., von Ahn, L., & Tomasic, A. (2010). Word sense disambiguation via human computation. *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 60–63). doi: 10.1145/1837885.1837905

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. *Proceedings of the 23rd international conference on world wide web*, 1049–1054. doi: 10.1145/2567948.2579215

Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxbow Books. Retrieved from https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D., & Marković, A. (2019). SASA dictionary as the gold standard for good dictionary examples for Serbian. *Proceedings of the eLex 2019 conference* (pp. 248–269). Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_14.pdf

Trampuš, M., & Novak, B. (2012). The internals of an aggregated web news feed. *Proceedings of 15th multiconference on information society 2012 (IS-2012)*. Retrieved from http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf

Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5366–5377). Retrieved from https://aclanthology.org/2022.lrec-1.574

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12): e0243300. doi: 10.1371/journal.pone.0243300

von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94. Retrieved from https://www.cs.cmu.edu/~biglou/ieee-gwap.pdf

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. doi: 10.1145/1378704.1378719

Von Hippel, E., & Katz, R. (2002). Shifting innovation to users via toolkits. *Management science*, *48*(7), 821–833.

Vyatkina, N., & Boulton, A. (2017). Corpora in language teaching and learning. *Language Learning and Technology*, *21*(3), 1–8.

Xu, L., & Chamberlain, J. (2020). Cipher: a prototype game-with-a-purpose for detecting errors in text. *Workshop games and natural language processing* (pp. 17–25). Retrieved from https://aclanthology.org/2020.gamnlp-1.3

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the 13th international workshop on semantic evaluation (SemEval-2019)* (pp. 75–86). doi: 10.18653/v1/S19-2010

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, C. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *Proceedings of the 14th international workshop on semantic evaluation.* Retrieved from https://arxiv.org/abs/2006.07235

Zviel-Girshin, R., Kuhn, T. Z., Luís, A. R., Koppel, K., Šandrih Todorović, B., Holdt, Š. A., Tiberius, C., & Kosem, I. (2021). Developing pedagogically appropriate language corpora through crowdsourcing and gamification. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation: short papers from EUROCALL 2021* (pp. 312–317). doi: 10.14705/rpnet.2021.54.1352

Priprava podatkov pri množičenju v pedagoške namene: primer igre CrowLL

Eden od načinov za spodbujanje uporabe korpusov pri jezikovnem izobraževanju je izdelava pedagoško primernih korpusov, označenih z različnimi vrstami problematik (občutljiva vsebina, žaljiv jezik, strukturne težave). Ker je ročno označevanje korpusov zelo časovno potratno, je potrebno poiskati boljši pristop. Predlagamo kombinacijo dveh pristopov k oblikovanju problemsko označenih pedagoških korpusov nizozemščine, estonščine, slovenščine in brazilske portugalščine: uporabo iger z namenom množičenja. Z udeleženci smo izvedli začetne poskuse, da bi ugotovili, če je naloga množičenja ustrezna, pridobljene izkušnje pa smo uporabili za oblikovanje igre *Crowdsourcing for Language Learning (CrowLL)*, v kateri igralci prepoznavajo problematične povedi in segmente ter jih razvrščajo. V prispevku se osredotočamo na pripravo podatkov, saj ima ta korak ključni pomen pri vsakem projektu množičenja, ki obravnava ustvarjanje jezikovnih učnih virov. Predlagamo metodologijo za pripravo podatkov, podrobno predstavljamo izbiro izvornih korpusov, pedagoško usmerjene konfiguracije GDEX in oblikovanje seznamov lem, s posebnim poudarkom na pogostih in od jezika odvisnih odločitvah. Za konec ponujamo razpravo o izzivih, ki smo jih zasledili, in o rešitvah, ki smo jih do sedaj že uvedli.

Ključne besede: množičenje, igra z namenom, vzorčni stavki, pedagoški korpus

Learning languages from parallel corpora: a blueprint for turning corpus examples into language learning exercises

Johannes GRAËN

Institute of Computational Linguistics, University of Zurich &
Department of Swedish, University of Gothenburg

This work describes a blueprint for an application that generates language learning exercises from parallel corpora. Word alignment and parallel structures allow for the automatic assessment of sentence pairs in the source and target languages, while users of the application continuously improve the quality of the data with their interactions, thus crowdsourcing parallel language learning material. Through triangulation, their assessment can be transferred to language pairs other than the original ones if multiparallel corpora are used as a source.

Several challenges need to be addressed for such an application to work, and we will discuss three of them here. First, the question of how adequate learning material can be identified in corpora has received some attention in the last decade, and we will detail what the structure of parallel corpora implies for that selection. Secondly, we will consider which type of exercises can be generated automatically from parallel corpora such that they foster learning and keep learners motivated. And thirdly, we will highlight the potential of employing users, that is both teachers and learners, as crowdsourcers to help improve the material.

Keywords: ICALL, language learning exercises, parallel corpora, data-driven learning

Graën, J.: Learning languages from parallel corpora: a blueprint for turning corpus examples into language learning exercises. Slovenščina 2.0, 10(2): 101–131.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.101-131

1 Overview

The generation of language learning exercises based on parallel corpus material requires the combination of several techniques and strategies. First of all, in order to automatically assess corpus material regarding its suitability for language learning exercises, we need to annotate it using standard techniques of Natural Language Processing (NLP), such as tokenization, lemmatization, part-of-speech tagging, and named entity recognition. In addition, we want to annotate the vocabulary used in those examples with the lowest proficiency level required to comprehend single lexical items of the target language that the learners want to acquire. The use of NLP techniques for computer-assisted language learning (CALL) is commonly referred to as ICALL (intelligent CALL) due to the numerous components of artificial intelligence (AI) that are applied in NLP methods (Lu, 2018).

Concerning parallel corpora (Section 2), we can take advantage of the expected parallelism between individual corpus units in the target language and the native language (L1) of the learner, or another foreign language (L2) in which the learner is sufficiently proficient. The latter case might be advantageous if there is a close typological relation between the target language and the L2. Take, for instance, a Finnish learner of Portuguese, who is already an advanced learner of Italian. In that case, examples from a parallel corpus of Portuguese/Italian will likely have more similarities regarding vocabulary and structure than a parallel corpus of Portuguese/Finnish.

The adequacy of the corpus material in particular sentences for different learner proficiency levels has received considerable attention in recent years (Pilán, 2018; Tack, 2021). A multitude of factors determine whether learners of a particular proficiency level are likely to comprehend a sentence or not. In the case of parallel sentence pairs, we will not only estimate the required proficiency level for each of the sentences individually, but also take into account the way it has been translated, independent of the translation direction. Employing interlingual word-level correspondences and intralingual syntactic relations between single words, we will derive grammatical correspondences, which, in turn, can be classified in terms of proficiency levels (Section 3).

Data-driven learning (Section 4) is a well-explored technique supporting language learner autonomy. The main idea is to let learners explore authentic language material on their own, which will make them observe patterns, turn those into hypotheses and then corroborate these with the help of search tools. Those patterns can relate to any linguistic level, such as lexicon, morphology, or syntax. While the idea of learning languages utilizing language material (as opposed to learning by prescribed rules) has been around for several decades, and its efficacy has been experimentally substantiated, the use of parallel corpora for that purpose has received significantly less attention (Lawson, 2001; Bluemel, 2014; Montero Perez et al. 2014, to name a few).

Learners benefit from corpus tools that are easy to use and visually help them to explore the respective content. Corpus search activities are either learner-driven, in the case of autonomous learners or open exercises, or instructor-driven, when learners are given concrete tasks to perform. While a learner already needs to have acquired a certain level of autonomy for the former case, the latter requires some form of feedback from the teacher in case the learners have not understood the motivation behind those tasks. That is why we are going one step further and use sentence pairs retrieved from corpora for generating language learning exercises (Section 5). Having annotated and aligned parallel sentences facilitates a whole new range of exercise types.

The term crowdsourcing is often associated with the idea of a large number of people doing voluntary work. Voluntariness, however, needs to be seen with respect to the motivation of the volunteers. Whether they are contributing out of interest, are getting paid for their work, or need to participate for other reasons (e.g. to pass a course) makes a difference concerning the results we expect to get. In addition to motivation, we can distinguish, whether crowdsourcers are consciously contributing or not, and thus providing explicit or implicit feedback (Wang et al., 2019). As opposed to amateur scientists participating in research projects, which is typically referred to as “citizen science”, crowdsourcers can be lay people with no expert knowledge (Section 6).

Having briefly discussed all the relevant topics, we proceed to describe the envisaged architecture for the application in Section 7 addressing the previously described challenges. The corpus retrieval functionality has been implemented and fed with parallel sentences from the OpenSubtitles corpus (Lison and Tiedemann, 2016) in 21 language pairs, namely every combination of the Catalan, English, French, German, Italian, Spanish and Swedish part of that corpus. We named it PaCLE (Parallel Corpora for Language Learning Exercises) and used it in several experiments, one of which we describe in Graën et al. (in press).

2 Parallel corpora

In a previous work (Zanetti, Volodina, and Graën 2021), we describe two challenges of automated exercise generation, namely reducing the ambiguity of generated exercises with the help of NLP methods, and the selection of appropriate sentences from corpora. In both cases, parallel corpora will be of great avail.

Parallel corpora consist of at least two datasets that refer to the same sequence of language material. The typical cases are bilingual or multilingual corpora, where those datasets correspond to translations of some material. The original material can be one of the datasets but does not necessarily need to be part of the corpus. As for the material, most parallel corpora consist of plain text, but parallel corpora of audio recordings also exist, which are often accompanied by transcripts, such as the Parallel Audiobook Corpus[1] (Ribeiro 2018). What is more, corpora consisting of several layers in the same language, such as the just-mentioned Parallel Audiobook Corpus which comprises recordings of different speakers reading the same books, also meet the condition of parallelism. Finally, learner corpora that comprise not only the learners' writings but also a normalized or corrected version of their text productions are also covered by the term parallel corpus.

Unlike parallel corpora, so-called comparable corpora do not necessarily possess parallel structures, but merely share the same topics per corresponding unit (e.g., articles). Wikipedia[2] can be seen as a comparable corpus, since a correspondence relation between languages can be established for individual articles (McEnery and Xiao, 2007; Otero and López, 2010; Barrón-Cedeno et al., 2015).

[1] https://datashare.is.ed.ac.uk/handle/10283/3217

[2] https://www.wikipedia.org/

2.1 Sources

Many parallel corpora have been made freely available over the last few decades. The largest source of parallel corpus material is arguably the OPUS collection[3] (Tiedemann, 2009, 2012). We have recompiled a small number of existing parallel corpora of different text types and languages (including low-resource languages such as Romansh and Swiss German) into a common format that allows for hierarchical correspondence annotation (Graën 2018) on any of the three levels that each of the individual corpora has, namely documents, sentences and words (i.e. tokens) (Graën et al., 2019).

At first, parallel corpora were compiled from publicly available translations. In several countries with more than one official language, documents from the respective authorities need to be translated from their original language to all other official ones. Typical examples of such corpora are the Canadian Hansards (Gale and Church, 1991, 1993), parliamentary debates in English and French, or the Belgisch Staatsblad (Vanallemeersch 2010), publications from the Belgian government in Dutch and French. In countries like Switzerland with three official languages (on the federal level) and multinational organizations such as the United Nations or the European Union, multilingual translations are produced that can and have been turned into corpora (Koehn, 2005; Rafalovitch et al., 2009; Eisele and Chen, 2010; Volk et al., 2010, 2016; Scherrer et al., 2014; Ziemski et al., 2016).

2.2 Alignment

The individual correspondence of textual units (e.g. sentences or words) is called an alignment, as is the process of deriving these correspondence relations. While the correspondence on the document level is typically derived by metadata (e.g. book chapters, webpages, external identifiers such as numbers assigned to documents), the identification of corresponding sentences and words requires dedicated tools. The performance of sentence alignment depends to a large part on how many one-to-one correspondences there are – that is, one sentence in one language translated to exactly one sentence in the other

[3] https://opus.nlpl.eu/

language. If there are numerous one-to-many relations or sentences without correspondence in the other language, so-called null alignments, the alignment performance can be significantly lower. A number of commonly used tools and methods exist to improve alignment performance (e.g. Varga et al., 2005; Braune and Fraser, 2010; Sennrich and Volk, 2010), and new methods keep being developed (Thompson and Koehn, 2019; Jiang et al., 2020).

For word alignment, the respective language pairs play an important role. As a rule of thumb, languages with similar structures and word formation yield better results. If bilingual alignments of more than two languages are combined, two scenarios are possible. Either all alignments agree, which suggests good quality of the individual bilingual alignments, or there are discrepancies between the pairwise alignments, which indicates that one or more of the alignments are erroneous, as not all identified correspondences can be correct in this case (cf. Graën et al., 2019). An approach of rotating triangulation can be used in this case to combine several bilingual alignments into a single harmonized multilingual one, and thus improve alignment quality.

In the same vein, the combination of different alignment techniques helps improve alignment quality. Ensemble methods such as the one presented by Steingrímsson, Loftsson, and Way (2021) have an advantage over the individual alignment methods, as seen in performance metrics such as the score or the alignment error rate (see Tiedemann, 2011, Section 2.6). Modern sentence aligners achieve better results by employing pre-trained multilingual neural language models (see Jalili Sabet et al., 2020; Dou and Neubig, 2021).

Alignment information in a corpus can be aggregated to derive a distribution from a single lexical unit in the source language to different units in the target language. The translation variants determined and quantified in this way help us select the right context, including word sense (see Section 3). We used these distributions to calculate a semantic relation between word pairs by means of translation variants (Graën and Schneider, 2020). Figure 1 shows a visualization from the tool that we created for learners to explore the semantics of translation variants from corpora.

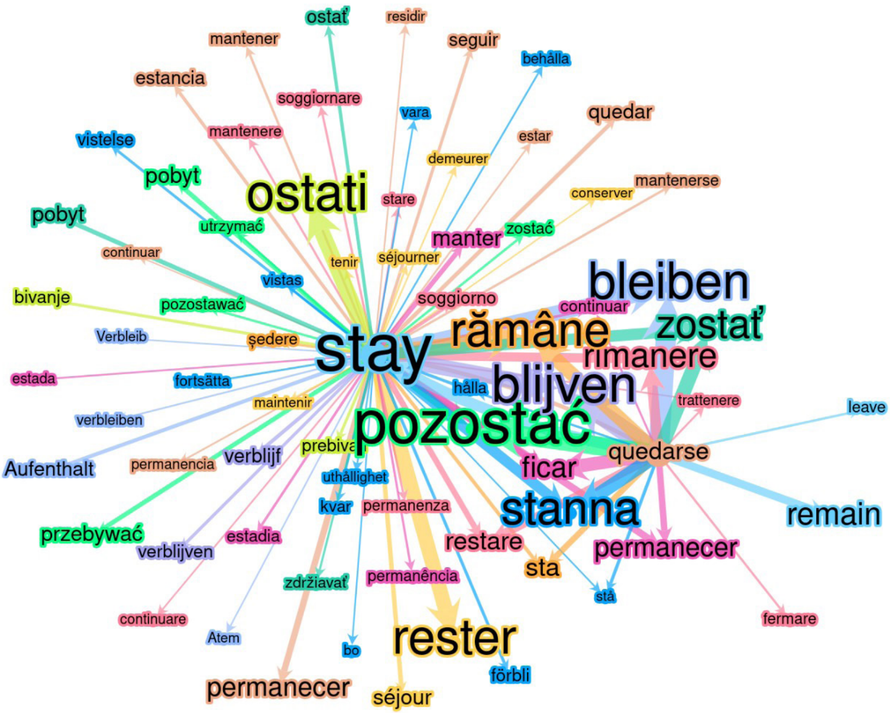


Figure 1: Shared and unique translation variants for English ‘stay’ and Spanish ‘quedarse’ in various languages. Word frequencies are expressed by the size of nodes and alignment probabilities by the thickness of edges. Individual languages are color-coded.

3 Learner proficiency

Like any other skill, learning a language starts with the first contact with the target, and eventually ends with its mastery. In between, there is a continuum that can be subdivided into a scale of proficiency levels defined by capabilities that a learner is required to achieve. Several standards of scaling exist and can be approximately mapped to each other, as they all define waypoints on the journey of acquiring a foreign language.

The proficiency of an individual learner can be measured in several dimensions, the two most prominent ones being reception vs. production and oral vs. written. The Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) subdivides “language activities” into reception and production as primary activities and interaction and mediation as secondary ones (Council of Europe 2001, Section 2.1.3).

Figure 2 replicates Figure 1 from the Common European Framework of Reference for Languages, which divides the proficiency scale into three coarse-grained levels (basic, independent and proficient user), each of which is further subdivided into two levels. We will henceforth refer to the six levels from A1 to C2 as CEFR levels. The CEFR scale has become a ubiquitous measure of language learning proficiency, and courses now indicate which level can be obtained after successfully finishing them, while job offers use them to specify proficiency requirements.

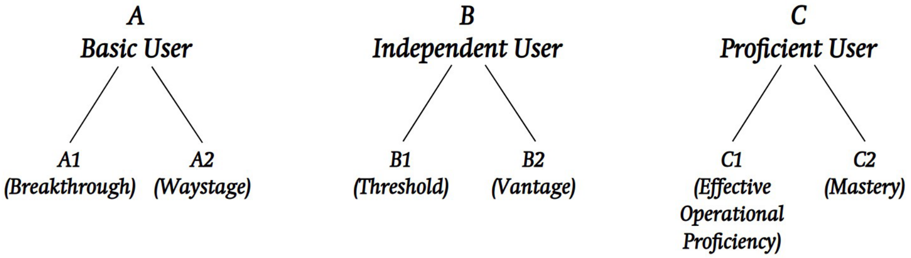


Figure 2: The “Common Reference Levels” as defined by (Council of Europe, 2001).

In the field of CALL, a multitude of research has been using the CEFR levels for various purposes, e.g. for the classification of texts (see Pilán et al., 2017) or the prediction of learner proficiency (Gaillat et al., 2022). The CEFRLex project[4] (François et al., 2016) provides mappings from lexical entries to distributions of CEFR levels for several languages. Those distributions stem in most cases from an analysis of textbooks. Each textbook is dedicated to a particular proficiency level, and the appearance of lexical entries (words and expressions) in the respective textbooks is represented as a frequency distribution. This distribution undergoes a normalization step to account for peaks of low-frequency entries, which is typically due to particular topics involving those entries (Dürlich and François, 2018).

We compared the English EFLLex from the CEFRLex resources (Dürlich and François, 2018) with two other lexical resources for

[4] https://cental.uclouvain.be/cefrlex/

English, namely the Pearson Global Scale of English (Pearson, 2017) and the Cambridge English Vocabulary Profile (Cambridge University Press, 2015), and found that they all agree to a large extent regarding the assigned CEFR level per lexical entry (Graën et al., 2020). The main difference between EFLLex and the other two resources is that the latter distinguish word senses, from which we had to abstract away for the sake of comparability by choosing the lowest level per entry, which typically corresponds to the most frequently used sense.

The word "stay" with the sense "to live in a place for a short time as a visitor or guest", for example, is classified by the Global Scale of English as beginner level (A1) on the CEFR scale. The same word is also used with the sense "to continue to be in a particular state, and not change", which is classified as an intermediate level (B1). Multiword expressions such as the phrasal verbs "stay on" or "stay out of" rank even higher (B2).

Apart from lexical resources, the frequency of a lexical unit in a general corpus and its length in terms of characters are also good indicators for the corresponding proficiency level. The relation between these two properties is illustrated by Zipf's law of abbreviation: shorter words are more frequently used and frequently used words tend to be shorter in general.

In addition to comparing EFLLex with other English resources, we also proved the hypothesis that "similar words in two languages, i.e. good direct translations, should have similar CEFR levels" (Graën et al., 2020, Section 3.5) by combining three monolingual CEFRLex resources, namely EFLLex for English, FLELex for French (François et al., 2014) and SVALex for Swedish (François et al., 2016), into one multilingual resource with the help of alignment probabilities obtained from a large parallel corpus (Graën, 2018), which we then used together with the raw CEFR level provided by EFLLex to predict the CEFR level of lexical entries from the above-mentioned lexical resources, the Pearson Global Scale of English and the Cambridge English Vocabulary Profile.

With the knowledge of how to identify words in different languages whose CEFR levels are strongly correlated, we can use one of the CEFRLex resources to project CEFR levels from one language onto another for which no equivalent resource exists. For multilingual corpora, as a matter of course we can project jointly from several languages for which CEFR-graded lexical resources are available.

4 Data-driven learning

A typical way for a learner to start learning an unfamiliar language is through language classes with the help of textbooks. Once an exercise in the textbook has been solved, however, it cannot be reused in a meaningful way, as doing exactly the same exercise more than once is a tedious task. To keep learners motivated, teachers need not only to have access to a large repertoire of different learning activities, including exercises, but also need a constant supply of novel language material.

A quarter of a century ago, Wilson (1997) identified "two major problems" in creating a language course. Both have to do with the availability of sufficient language learning material. The first one is about meeting "the needs of students with different abilities", while the second one addresses the need to provide "enough exercises to ensure that a student is confronted by a different set of examples whenever he or she uses the language learning program". In Wilson's view, "corpora present a unique and unexploited resource" in this context.

Boulton and Cobb (2017) performed a meta-analysis of publications studying the effects of data-driven learning, and concluded that this technique is both efficient and effective. In a previous study on the same topic (Cobb and Boulton, 2015), the authors state that for data-driven learning to succeed, "massive but controlled exposure to authentic input is of major importance, as learners gradually respond to and reproduce the underlying lexical, grammatical, pragmatic, and other patterns implicit in the languages they encounter".

5 Language learning exercises

Language learning exercises aim at improving the language skills of learners, which, at first glance, seems to be an obvious truism, though not all exercises are equally effective in all contexts. Under some conditions, the learning effect can be small to nonexistent, if, for example, the learner is overchallenged by an exercise and cannot solve it. Laufer and Ravenhorst-Kalovski (2010) evaluate the vocabulary size required for an “adequate reading comprehension” of regular texts in a foreign language, but also underline that the text type plays a role in this, and that texts with “a large proportion of technical and jargon vocabulary” might be more challenging to comprehend. On the other extreme, underchallenging the learner can also lead to them quickly losing motivation (Mousavian Rad et al., 2022).

For learning to be effective, exercises should thus be neither too simple nor too difficult for the learner in question. Language learners differ in various dimensions, e.g. in age (from elementary school pupils to language students at university level, or adult learners), motivation (intrinsic or extrinsic), current proficiency level in the target language (beginner to advanced), previous language learning experiences (e.g. of similar L2s), their metalinguistic knowledge, etc. Furthermore, the settings in which exercises are done also vary: in-class exercises vs. exercises done at home, individual or group exercises, low-stake (ungraded) vs. high-stake (graded) activities, and so on.

In the best case, teachers take into account all these properties when devising exercises as part of the curriculum, which, optimally, consists of complementary exercises and planned repetitions (cf. Nation and Webb, 2011; Nakata and Webb, 2016).

5.1 Limitations for automatically generated exercises

When it comes to generating language learning exercises automatically, that is by an algorithm instead of a human, only a small number of all possible exercise types are eligible, and even fewer can be reliably assessed programmatically. First of all, we want to limit ourselves to the interaction of a single learner with the (interactive) exercise. Observing a group of learners when they are interacting, e.g. in a role-play exercise, and providing feedback to the individual participants is something that language teachers are used to; this is, however, far beyond what can be automated today, despite the continuous advance of language technology. If human-human interaction is our target, communication is best channeled through the computer and the exercise is defined in a way such that communication is mostly controlled by the software.

This kind of language learning has been the subject of several publications in the field of computer-mediated communication (CMC). According to Heift and Vyatkina (2017), “CMC has shown to have many features similar to face-to-face language classroom interactions such as clarification requests and feedback”.

Another limitation to note is that we will exclusively work with written text. Oral exercises require additional technologies, speech recognition for productive exercises and speech generation for receptive ones, which add to the likelihood of the software making a mistake when generating the exercise or assessing the user input. There are, however, existing tools for supporting the oral part of language learning, e.g. in the area of computer-assisted pronunciation training (CAPT) (Fouz-González, 2015; Schwab and Goldman, 2018).

Our third and last limitation concerns the user input. Natural language processing techniques are – in their current state – not capable of semantically interpreting free-form answers reliably, especially if the input provided, which is the users’ textual output, deviates significantly from the training material, which for a large share of the available languages still are newspaper texts and other official documents. Texts produced by language learners comprising potentially innovative lexical and grammatical components typically yield a significantly higher error rate when being processed by such models. Assuming that we could process texts produced by learners without making annotation errors, we would still struggle to provide learners with the helpful feedback that a human teacher could. Existing tools that accept free-form textual input provide selective feedback on spelling and grammatical constructions. A machine-generated exercise where the learner continues a story for which only the beginning is given – with automated feedback provided by an algorithm on writing style, text structure, and word choice– is unlikely to be available soon.

5.2 Exercises from parallel corpora

As we have annotated corpus material, we can support the comprehension of text by simple means, such as color-coding different parts of speech, showing additional information when the user hovers over a particular token, interactively displaying syntactic relations (e.g. marking subject and object relations of verbs or pointing out the respective base verbs for separated particles in languages such as German or Swedish). In parallel corpora, we can also highlight translation equivalents with the help of alignments (as we do in multilingwis, see Clematide et al., 2016; Graën et al., 2017) or combine alignments and syntax to retrieve meaningful chunks of words (as in Zanetti et al., 2021).

In an earlier work (Alfter and Graën, 2019) we present the prototype of a game to train particle verbs in English and Swedish. A virtual currency is used for motivational purposes. The user earns credits for correctly guessed particles and loses them if they are wrong, while different types of hints can be "bought" by using credits. Parallel data used by the application is extracted from the CoStEP corpus (Graën et al., 2014), which is based on Europarl (Koehn, 2005), and annotated in an unsupervised way. Particle verbs are classified with respect to their proficiency level based on EFLLex (Dürlich and François, 2018) and SVALex (François et al., 2016).

Our work described in Zanetti, Volodina, and Graën (2021) introduces a novel type of sentence reordering exercise. We address the issue of potentially erroneous alignment of function words and the (sometimes) unclear correspondence of functional parts by merging single tokens to chunks based on their syntactic relations. We extracted sentences from the OpenSubtitles corpus (Lison and Tiedemann, 2016), processed them with standard natural language processing pipelines, and used language-specific readability measures to estimate the complexity of sentences.[5]

6 Crowdsourcing

A crowdsourcing application known by many people is "recaptcha" (Von Ahn et al., 2008), a word recognition task that users have to solve before they are allowed to proceed to the actual web content they requested. These puzzles have a dual purpose: by solving them, the users primarily prove that they are human, but at the same time they provide

[5] A prototype of the envisaged exercise type can be tested here: https://codepen.io/gi0/pen/vYLJYjp.

human judgments on words that are unknown to the recaptcha system, thus contributing to a dataset that can be used to train OCR algorithms.

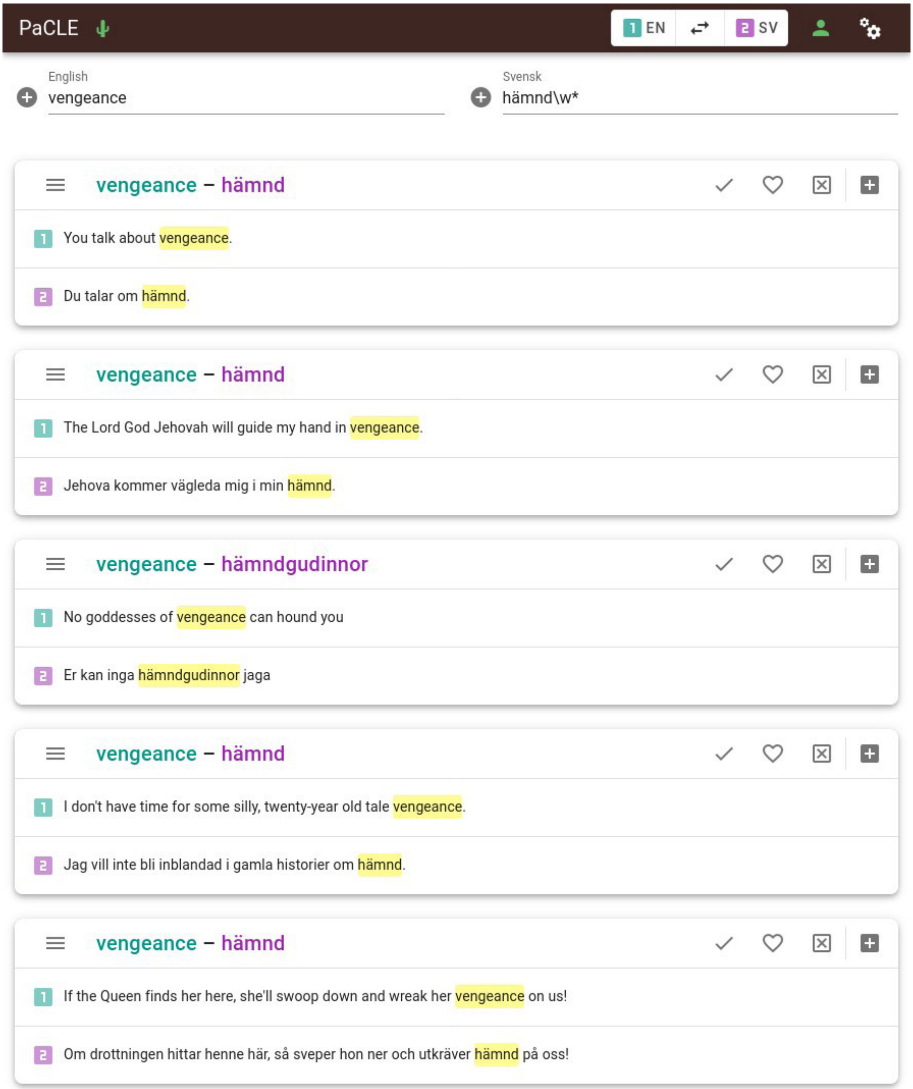
Apart from this prototypical example, where crowdsourcing is used "along the way", there are tools for creating crowdsourcing experiments and having people solve a large number of tasks.[6] Users of those applications typically spend a considerable amount of time performing a large number of tasks. Here, the recruitment of crowdsourcers plays a key role. One can disseminate information and ask people to volunteer, or require university students to contribute a particular number of tasks, as is frequently done for publications about crowdsourcing experiments.

The crowdsourcing taxonomy by Geiger et al. (2011) can be employed to classify existing crowdsourcing approaches into four different categories, based on: 1) who are the contributors, or rather which type of contributors are wanted for the application in question, and if they have to show their capacity for the given task first; 2) to which degree a user can access the contributions of other users; 3) how the contributions of different users are aggregated or selected; and 4) whether or under which circumstances contributions are remunerated. For cases where no remuneration is available, the authors list as potential motivational factors "passion, fun, community identification, or personal achievement".

Another dimension is defined by the degree to which the participants are conscious as to whether they are contributing their efforts towards a particular goal. Most cases can be unequivocally assigned to one extreme or the other. Any paid crowdsourcing work is by definition explicit, unless the participants are paid for a different task than the one whose data is actually being crowdsourced. At the other extreme, analyzing log files to see how users interact with some software is a good example of implicit crowdsourcing (Wang et al., 2019). In between we have situations with no explicit tasks and where users might or might not know that they are contributing data through their interactions with software.

[6] E.g. the open PyBossa (https://pybossa.com/) or Amazon Mechanical Turk (https://www.mturk.com/) for paid microservices.

7 The application



vengeance – hämnd

1 You talk about vengeance.

2 Du talar om hämnd.

vengeance – hämnd

1 The Lord God Jehovah will guide my hand in vengeance.

2 Jehova kommer vägleda mig i min hämnd.

vengeance – hämndgudinnor

1 No goddesses of vengeance can hound you

2 Er kan inga hämndgudinnor jaga

vengeance – hämnd

1 I don't have time for some silly, twenty-year old tale vengeance.

2 Jag vill inte bli inblandad i gamla historier om hämnd.

vengeance – hämnd

1 If the Queen finds her here, she'll swoop down and wreak her vengeance on us!

2 Om drottningen hittar henne här, så sveper hon ner och utkräver hämnd på oss!

Figure 3: The PaCLE application showing five examples for a parallel corpus search in the English-Swedish part of OpenSubtitles. Matching parts are highlighted. The use of advanced regular expressions is supported.

The blueprint for the application that we describe in this work can be split into two phases: First, an offline phase, in which sentence pairs are extracted from parallel corpora, processed with (language-specific)

NLP techniques, assessed regarding their usefulness in language learning and, finally, added to a database. Second, an online phase, in which a web application interacts with two types of users, namely teachers and learners.[7] The application allows users to perform searches in the corpus examples using metadata (e.g. the source of the respective example) and derived measures (e.g. the estimated target proficiency levels) as filters. The retrieved sentence pairs can then be manually reviewed and turned into learning exercises. In Graën et al. (in press), we used an early prototype of the application in a language-learning class and analyzed the students' use of the tool and other technologies. Figure 3 shows the user interface.[8]

One criterion for filtering out sentences in the offline phase is that they are not immediately comprehensible to the reader without the contexts in which they appear in the corpus. Pilán et al. (2017) provide an extensive overview of measures that can be employed for selecting corpus examples suitable for use in educational contexts. Some of the measures they list do not require sentences to be excluded *a priori*, but rather determine for which type and proficiency of learners they can be used (e.g. measures concerning grammatical or lexical complexity). In addition to monolingual criteria that are applied to one part of a parallel corpus,[9] we define measures on sentence pairs that determine whether those pairs are added to the database and measures that are used in the online phase for making a selection that fits the requirements of a particular configuration (languages, search terms, learner proficiency level, exercise type, etc.).

A measure that can be used in both phases is the degree of equivalence between the two sentences in terms of syntactic structures and lexical items that are used as translations of each other. By

[7] We do not envisage providing two different applications or user modes for teachers and learners, as we conceive autonomous language learners as their own teachers and, beyond that, have no means to distinguish them technically.

[8] We started developing the web application with desktop clients in mind. We discourage using the application on mobile phones as, from our perspective, the attention span on those devices is often lower, less information can be displayed (although today's mobile phones typically have a high resolution), and user input is not as precise and fluent as with regular keyboards and pointing devices.

[9] We do not distinguish between source and target languages at that stage. Later on, when selecting corpus examples in the online phase, we usually prefer the target language to be the one that is more comprehensible.

calculating structural equivalence in terms of the relative frequency that the structure in question is used in a parallel corpus in relation to the overall number of structures identified in both sentences, we obtain a ratio (values between 0 and 1) for which we define a threshold for inclusion in the database. For lexical items, a similar formula is used. Higher values of both measures mean that we expect the sentence pair in question to show more frequently used structural and lexical correspondence and, consequently, represent a more direct translation (as opposed to a freer one with less frequent correspondences and, hence, lower values).

7.1 Corpora

While a variety of parallel corpora can be obtained easily, e.g. downloaded directly from the OPUS collection (Tiedemann, 2009, 2012), not all of them are equally suited for language learning purposes. For a corpus to fit the needs of learners, in the optimal case, it should comprise language material that a) is adequate for the proficiency level of said learners, b) comprises the material to be learned (lexical elements, grammatical constructions, and so on), c) be sufficiently large so that the application can choose from a large number of examples, and d) be of interest to the learner. The latter point is unequivocally learner-dependent, but we expect that there are domains that are generally better received than others (e.g. law texts vs. fiction).

One source of parallel texts that we found particularly useful for the purpose of language learning is the OpenSubtitles corpus (Lison and Tiedemann, 2016) which we used in Zanetti, Volodina, and Graën (2021), but also for the PaCLE application. It consists of translated subtitles for a large number of movies. Translations are contributed by users who can also review the work of other users. A large number of subtitles is available for most of the available 62 languages, but for some languages – such as Bengali, Georgian, or Tagalog – the coverage is quite low, and insufficient for our purposes.

Besides the large size and coverage of many language pairs with this corpus, subtitles have the advantage that “[they] cover various genres and time periods and combine features from spoken language corpora and narrative texts including many dialogs, idiomatic expressions, dialectal expressions and slang" (Tiedemann, 2012).

Similar to OpenSubtitles, we find a richer vocabulary and less formal language in corpora of transcribed speech, such as the parliamentary proceedings of the European Union (Koehn 2005), the Canadian Hansards (described in Gale and Church, 1991, 1993) or the TED Talks corpus (Reimers and Gurevych, 2020).

Corpora compiled from legislative texts, patents, technical manuals, medication leaflets, and other more restricted text types might be helpful for particular learning tasks and more advanced learners, but they are hardly suited for most learners with lower proficiency levels. We can also expect to find considerably fewer appearances of offensive language, often abbreviated as PARSNIP, than in monolingual corpora (Dekker et al., 2019) for the same reason.

7.2 Data preparation

Modern NLP applications use language models that can perform several annotation tasks simultaneously. Performance measures show that those joint models outperform traditional pipeline approaches (Qi et al., 2020). The standard tasks for such models to perform are tokenization, lemmatization, part-of-speech tagging, and syntactic dependency parsing. Other tasks include morphological analysis, named-entity recognition, and word-sense disambiguation, all of which provide valuable information for the creation of language learning exercises.

Some corpora are provided pre-aligned (typically on the sentence level), but there are corpora indicating alignment only on a higher level, such as documents or chapters. In such cases we need to perform document alignment first, followed by sentence alignment to obtain parallel sentences. The correspondence of documents is to a large extent corpus-specific, and thus no out-of-the-box solutions can be employed (Graën, 2018, Section 4.1). In the case of multiparallel corpora, we might want to apply approaches that produce consistent multilingual alignments (Graën, 2018, Section 4.3).

We also need the retrieved and annotated sentence pairs to be word-aligned. By combining the results of different aligners and different types of aligners (probabilistic measures vs. word embeddings), we obtain the most reliable alignment links. We then group the correspondence links between single tokens using syntactic relations as described in Zanetti et al. (2021). After this, function words such as prepositions or particles that often have no correspondence in another language are part of larger units for which we can assert correspondence with higher precision. The groups we build with the help of dependency and alignment relations often correspond to phrases, but this is not necessarily always the case.

Alignment probabilities calculated on the whole corpus or obtained from another source help us to identify idiomaticity (Schneider and Graën, 2018). In support verb constructions, for example, the correspondence of the aligned nouns, which are frequently direct objects of the verb in question, is a very strong one; that is, we expect it to be the prototypical translation equivalent, while the correspondence of the governing verbs is often an infrequent one (but it can also be the case that the same support verb is used). The English support verb

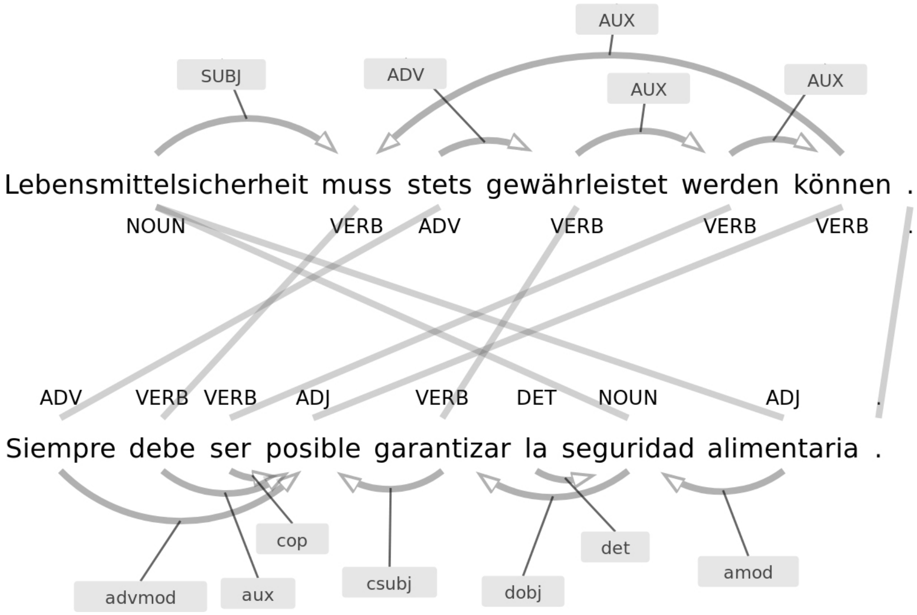

AUX
SUBJ
ADV
AUX
AUX
Lebensmittelsicherheit muss stets gewährleistet werden können .
NOUN
VERB
ADV
VERB
VERB
VERB
.
ADV
VERB
VERB
ADJ
VERB
DET
NOUN
ADJ
.
Siempre debe ser posible garantizar la seguridad alimentaria .
cop
advmod
aux
csubj
dobj
det
amod

Figure 4: Sentence pair in German and English with different syntactic structures, which is highlighted by the heavily crossing alignment links. Here, language-dependent label sets have been used instead of Universal Dependencies.

construction "(to) take a walk", for example, and the Spanish one "dar un paseo" ("give a walk") are common translations of each other. The nouns "walk" and "paseo" also show a high alignment probability in any parallel English-Spanish corpus. However, "take" is only a good translation of "dar" as part of a limited number of other expressions other than "(to) take a walk" / "dar un paseo" (e.g. "take a step" and "dar un paso").

7.3 Example selection

For the selection of adequate sentence pairs, we envisage using classifiers like the ones described in Pilán et al. (2017), Pilán (2018) and Tack (2021) for the individual sentences. In addition to the estimated proficiency levels, we will compare the aligned groups of tokens. Noun phrases that translate to noun phrases are arguably less challenging than completely diverging structures. By aggregating syntactic structures and calculating conditional probabilities from the observed frequencies in a large parallel corpus, we can say how likely it is for a particular syntactic structure in one language to be translated to another structure in the other language. The main idea here is that structural correspondences with higher probabilities will be more advantageous for language learning. Nonetheless, non-standard or less frequent correspondences will certainly be of interest for more advanced learners (Figure 4 shows an example).

7.4 Exercise generation

The combination of two sentences including word alignment paves the way for a whole new range of exercise types. At the same time, we can use the information of word and phrase correspondence to improve common monolingual exercises. For cloze tests, for instance, we can use the translation of the sentence in question to identify distractors that are unlikely to accidentally fit in the gap.

Contrastive exercises look for similarities and differences between the source and target language, and thus foster metalinguistic awareness. Properties that could be the focus of such exercises are morphological features (e.g. grammatical genders), the order of syntactic elements (e.g. the position of modifying adjectives relative to their governor), or the use of discourse markers.

In the parallel reordering exercise presented in Zanetti et al. (2021) and in the gap-filling exercise with parallel clues presented in Alfter and Graën (2019), the source language serves as an anchor for the learner. Truly multilingual exercises are those where there is no distinction between source and target languages. One example is a gap-filling or cloze exercise in the style of bundled gaps (Wojatzki et al., 2016) but with word pairs (or triples, ...) in two (or three, ...) different languages. A potential way to find good distractors is to generate different inflections of the original words that have been replaced by the gaps. Alternatively, homographs or false friends can be used with non-parallel sentences to focus on differences and similarities.

7.5 Crowdsourcing aspects

The way the application is intended to be used is threefold. First, we envisage an autonomous learner – i.e. a more advanced learner with a good command of technology – to use the application for looking up words, expressions, or grammatical constructions in context together with their translations. In this scenario, we use the annotation and alignment layers obtained during corpus preparation to let the user interactively explore the examples that they found. Learners can add particular examples to (named) collections, mark their favorites and report entire sentence pairs, individual annotations, or alignment links that they consider false or dubious.

In the second scenario, teachers look up examples relevant to their respective topics, with respect to both content and language. They group examples in collections from which they can feed the in-class exercises that they prepare. Sharing those collections between teachers and collaborating on the creation of language learning material is facilitated by the application (e.g. by just copying an URL and sending it to other teachers or students).

The third scenario goes one step further. Here, teachers use collections of corpus examples to generate exercises. Generated exercises can be reviewed and discarded as needed, but the parallelism in the exercise types should generally result in higher precision, so good accuracy can be expected. Teachers then share those exercises with their students who, in turn, can also provide feedback in terms of reporting any errors or discrepancies in the example items.

In all scenarios, users should be able to fix errors for themselves, such as by correcting spelling mistakes in the original corpus material, or propose changes that can be reviewed by other users. The simplest solution that does not require a dedicated user or group to review all proposals is to explicitly ask other users and let them up- or downvote the (proposed) changes. In cases with a clear tendency of mostly upvotes, the solution would be automatically accepted and replace the original example. The current prototype allows users to edit the actual examples, accept or reject them, and put them on a list of favorites, which is meant to keep those examples that learners consider valuable to them.

The type of crowdsourcing envisaged for the different scenarios is both explicit and implicit. Explicit crowdsourcing involves error correction and the categorization of annotations as dubious. When users are explicitly asked by the application for their opinions on changes proposed by other users, they are also explicitly contributing their knowledge. The collaborative elaboration of language learning material falls in the category of crowd annotation.

When users mark their favorite examples or remove elements from their collections, they contribute in an implicit way. We can only guess why examples have been removed; it might be due to errors in the examples themselves, their annotation, because they are not comprehensible for the individual learner, or they simply do not match the topic in question. In cases of doubt, we can always turn those choices into explicit questions with which we ask other users for clarification.

It is important to note that all crowdsourcing tasks are designed to stem from intrinsic motivation. The added value of using the application for self-learning – which is the corpus search function or the assistance provided with the creation of learning exercises – needs to convince learners and teachers to voluntarily contribute to the project.

8 Conclusions

We have discussed a blueprint for an application that generates language learning exercises from parallel corpora. To this end, we have outlined the required methods and techniques, and described how it is envisaged they will work together in the final application.

Moreover, we have argued how the ensemble of annotation and alignment of parallel corpora can be employed to reduce the uncertainty about potential errors in automatically generated exercises. What is more, the use of parallel material paves the way for a multitude of novel exercise types that encourage learners to contrast target and source languages, and thus strengthen their metalinguistic capabilities.

In short, with the help of implicit and explicit crowdsourcing, we expect language learning material to gradually improve over time.

Acknowledgments

This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1 184212 through the project "From parallel corpora to multilingual exercises: Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications", conducted at Pompeu Fabra University in Barcelona (with Grael, Grup de Recerca en Aprenentatge i Ensenyament de Llengües, and at the University of Gothenburg (with Språkbanken Text).

References

Alfter, D., & Graën, J. (2019). Interconnecting Lexical Resources and Word Alignment: How Do Learners Get on with Particle Verbs? In *Proceedings of the 22nd Nordic Conference of Computational Linguistics (NODALIDA)* (pp. 321–26). Turku, Finland: Linköping University Electronic Press. Retrieved from https://www.aclweb.org/anthology/W19-6135

Barrón-Cedeno, A., España Bonet, C., Boldoba Trapote, J., & Márquez Villodre, L. (2015). A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora* (pp. 3–13). Association for Computational Linguistics.

Bluemel, B. (2014). Learning in Parallel: Using Parallel Corpora to Enhance Written Language Acquisition at the Beginning Level. *Dimension, 31*, 48.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning,* 67(2), 348–393.

Braune, F., & Fraser, A. (2010). Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING): Posters* (pp. 81–89). Association for Computational Linguistics (ACL).

Cambridge University Press. 2015. English Vocabulary Profile. Retrieved from https://www.englishprofile.org/wordlists

Clematide, S., Graën, J., & Volk, M. (2016). Multilingwis – a Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora. In G. Corpas Pastor (Ed.), *Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia Computacional y Basada En Corpus: Perspectivas Monolingües y Multilingües* (pp. 447–455). Geneva: Tradulex. doi: 10.5167/uzh-120153

Cobb, T., & Boulton, A. (2015). Classroom Applications of Corpus Analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 478–497). Cambridge University Press. doi: 10.1017/CBO9781139764377.027

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

Dekker, P., Zingano Kuhn, T., Šandrih, B., Zviel-Girshin, R., Arhar Holdt, Š., & Schoonheim, T. (2019). Corpus Filtering via Crowdsourcing for Developing a Learner's Dictionary. In I. Kosem & S. Krek (Eds.), *Proceedings of the eLexicography in the 21st Century (eLex 2019): Smart Lexicography, 1–3 October* 2019, *Sintra, Portugal* (pp. 84–85). Brno: Lexical Computing CZ, s.r.o.

Dou, Z.-Y., & Neubig, G. (2021). Word Alignment by Fine-Tuning Embeddings on Parallel Corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL), 19–23 April 2021.*

Dürlich, L., & François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In N. Calzolari et al. (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation, 7–12 May 2018, Miyazaki, Japan.* European Language Resources Association (ELRA).

Eisele, A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In N. Calzolari et al. (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), 17–23 May*

2010, *Valletta, Malta* (pp. 2868–2872). European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/volumes/L10-1/

Fouz-González, J. (2015). Trends and Directions in Computer-Assisted Pronunciation Training. *Investigating English Pronunciation*, 314–342.

François, T., Fairon, C., & Watrin, P. (2016). CEFRLex: A Graded Lexical Resource for French Foreign Learners. Retrieved from http://cental.uclouvain.be/cefrlex/

François, T., Gala, N., Watrin, P., & Fairon, C. (2014). FLELex: A Graded Lexical Resource for French Foreign Learners. In N. Calzolari et al. (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, *26–31 May*, *Reykjavik, Iceland* (pp. 3766–3773). European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L14-1

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). SVALex: A CEFR-Graded Lexical Resource for Swedish Foreign and Second Language Learners. In N. Calzolari et al. (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), May 2016, Portorož, Slovenia* (pp. 213–219). Retrieved from https://aclanthology.org/L16-1032.pdf

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2022). Predicting CEFR Levels in Learners of English: The Use of Microsystem Criterial Features in a Machine Learning Approach. *ReCALL, 34*(2), 130–146.

Gale, W. A., & Church, K. W. (1991). A Program for Aligning Sentences in Bilingual Corpora. In D. E. Appelt et al. (Eds.), *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), 18–21 June 1991*, *Berkeley, California, USA* (pp. 177–184). Stroudsburg, PA, USA. Association for Computational Linguistics (ACL). doi: 10.3115/981344.981367

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics, 19*(1), 75–102.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., & Schader, M. (2011). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In *AMCIS 2011 Proceedings - All Submissions: Virtual Communities and Collaborations* (p. 430).

Graën, J. (2018). Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning. PhD thesis. University of Zurich.

Graën, J., Alfter, D., & Schneider, G. (2020). Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. In *Proceedings of the 12th*

Language Resources and Evaluation Conference (LREC), 2020, Marseille, France (pp. 346–355). Marseille, France: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.43

Graën, J., Bach, C., & Cassany, D. (in press). Using a Bilingual Concordancer to Promote Metalinguistic Reflection in the Learning of an Additional Language: The Case of B1 Learners of Catalan. In *n/a*. Peter Lang.

Graën, J., Batinic, D., & Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In J. Ruppenhofer & G. Faaß (Eds.), *Proceedings of the 12th edition of the Conference on Natural Language Processing (KONVENS)* (Vol 1, pp. 222–227). Stiftung Universität Hildesheim. GSCL, ÖGAI, DGfS, Clarin-D, University of Hildesheim. doi: 10.5167/uzh-99005

Graën, J., Kew, T., Shaitarova, A., & Volk, M. (2019). Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection. In P. Bański et al. (Eds.), *Challenges in the Management of Large Corpora (CMLC)*. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9020

Graën, J., Sandoz, D., & Volk, M. (2017). Multilingwis. Explore Your Parallel Corpus. In J. Tiedemann & N. Tahmasebi (Eds.), *Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA), May 2017, Gothenburg, Sweden* (pp. 247–250). Association for Computational Linguistics (ACL). doi: 10.5167/uzh-137129

Graën, J., & Schneider, G. (2020). Exploiting Multiparallel Corpora as a Measure for Semantic Relatedness to Support Language Learners. In D. Levey (Ed.), *Strategies and Analyses of Language and Communication in Multilingual and International Contexts* (pp. 153–167). Cambridge Scholars Publishing.

Heift, T., & Vyatkina, N. (2017). Technologies for Teaching and Learning L2 Grammar. *The Handbook of Technology and Second Language Teaching and Learning*, 26–44.

Jalili Sabet, M., Dufter, P., Yvon, F., & Schütze, H. (2020). SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In B. Webber, T. Cohn, Y. He & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, November 2020, online* (pp. 1627–1643). Association for Computational Linguistics (ACL). Retrieved from https://www.aclweb.org/anthology/2020.findings-emnlp.147

Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020). Neural CRF Model for Sentence Alignment in Text Simplification. In D. Jurafsky, J. Chai, N.

Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *July 2020*, *online* (pp. 7943–7960). Association for Computational Linguistics (ACL). doi: 10.18653/v1/2020.acl-main.709

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit*, *5*, 79–86. Asia-Pacific Association for Machine Translation.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension.

Lawson, A. (2001). Collecting, Aligning and Analysing Parallel Corpora. *Small Corpus Studies and ELT: Theory and Practice. Amsterdam, John Benjamins*, 279–309.

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In N. Calzolari et al. (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), May 2016, Portorož, Slovenia*. European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1147/

Lu, X. (2018). Natural Language Processing and Intelligent Computer-Assisted Language Learning (ICALL). In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). John Wiley & Sons, Ltd. doi: 10.1002/9781118784235.eelt0422

McEnery, T., & Xiao, Z. (2007). Parallel and Comparable Corpora: The State of Play. *Corpus-Based Perspectives in Linguistics* 6.

Montero Perez, M., Paulussen, H., Macken, L., & Desmet, P. (2014). From Input to Output: The Potential of Parallel Corpora for CALL. *Language Resources and Evaluation, 48*(1), 165–189.

Mousavian Rad, S. E., Roohani, A., & Mirzaei, A. (2022). Developing and Validating Precursors of Students' Boredom in EFL Classes: An Exploratory Sequential Mixed-Methods Study. *Journal of Multilingual and Multicultural Development*, 1–18. doi: 10.1080/01434632.2022.2082448

Nakata, T., & Webb, S. (2016). Vocabulary Learning Exercises: Evaluating a Selection of Exercises Commonly Featured in Language Learning Materials. In *SLA Research and Materials Development for Language Learning*, 139–154. Routledge.

Nation, I. S. P., & Webb, S. 2011. *Researching and Analyzing Vocabulary*. Heinle, Cengage Learning Boston, MA.

Otero, P. G., & González López, I. (2010). Wikipedia as Multilingual Source of Comparable Corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC* (pp. 21–25). Citeseer.

Pearson. (2017). GSE Teacher Toolkit. Retrieved from https://www.english.com/gse/teacher-toolkit/user/lo

Pilán, I. (2018). *Automatic Proficiency Level Prediction for Intelligent Computer-Assisted Language Learning*. PhD thesis. University of Gothenburg.

Pilán, I., Volodina, E., & Borin, L. (2017). Candidate Sentence Selection for Language Learning Exercises: From a Comprehensive Framework to an Empirical Evaluation. *Revue Traitement Automatique Des Langues. Special Issue on NLP for Learning and Teaching. 57*(3), 67–91.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, *July 2020*, *online* (pp. 101–108). Association for Computational Linguistics (ACL). doi: 10.18653/v1/2020.acl-demos.14

Rafalovitch, A., & Dale, R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proceedings of the Machine Translation Summit*, *12*, 292–299.

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics (ACL). doi: 10.18653/v1/2020.emnlp-main.365

Ribeiro, M. S. (2018). Parallel Audiobook Corpus (version 1.0), University of Edinburgh. School of Informatics. doi: 10.7488/ds/2468

Scherrer, Y., Nerima, L., Russo, L., Ivanova, M., & Wehrli, E. (2014). SwissAdmin: A Multilingual Tagged Parallel Corpus of Press Releases. In N. Calzolari et al. (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, *26–31 May*, *Reykjavik, Iceland.* European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L14-1

Schneider, G., & Graën, J. (2018). NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners' Collocational Skills. In I. Pilán, E. Volodina, D. Alfter & L. Borin (Eds.), *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018*

(NLP4CALL), *November 2018, Stockholm, Sweden* (pp. 69–78). LiU Electronic Press. doi: 10.5167/uzh-157985

Schwab, S., & Goldman, J.-P. (2018). MIAPARLE: Online Training for Discrimination and Production of Stress Contrasts. In K. Klessa et al. (Eds.), *Proc. 9th Int. Conf. Speech Prosody, 13–16 June 2018*, *Poznań, Poland* (pp. 572–576). doi: 10.21437/SpeechProsody.2018-116

Sennrich, R., & Volk, M. (2010). MT-Based Sentence Alignment for OCR-Generated Parallel Texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*, *31 October – 5 November 2010,* Denver, Colorado, USA. Association for Machine Translation in the Americas (AMTA). Retrieved from https://aclanthology.org/2010.amta-papers.14.pdf

Steingrímsson, S., Loftsson, H., & Way, A. (2021). CombAlign: A Tool for Obtaining High-Quality Word Alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, *31 May – 2 June 2021*, *Reykjavik, Iceland, Sweden, online* (pp. 64–73). Linköping University Electronic Press, Sweden. Retrieved from https://aclanthology.org/2021.nodalida-main.7

Tack, A. (2021). *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. PhD thesis.

Thompson, B., & Koehn, P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), November 2019, Hong Kong, China* (pp. 1342–1348). Association for Computational Linguistics (ACL). Retrieved from https://aclanthology.org/D19-3.pdf

Tiedemann, J. (2009). News from OPUS – a Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, *5*, 237–248.

Tiedemann, J. (2011). *Synthesis Lectures on Human Language Technologies 2*. Morgan & Claypool. doi: 10.2200/S00367ED1V01Y201106HLT014

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari et al. (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, *May 2012*, *Istanbul, Turkey* (pp. 2215–2218). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Vanallemeersch, T. (2010). Belgisch Staatsblad Corpus: Retrieving French-Dutch Sentences from Official Documents. In N. Calzolari et al. (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), May 2010, Valletta, Malta* (pp. 3413–3416). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/758_Paper.pdf

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., & Nagy, V. (2005). Parallel Corpora for Medium Density Languages. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, N. Nikolov (Eds.), *Proceedings of Recent Advances in Natural Language Processing (RANLP), 21–23 September 2005, Borovets, Bulgaria* (pp. 590–596). Retrieved from http://lml.bas.bg/ranlp2005/

Volk, M., Amrhein, C., Aepli, N., Müller, M., & Ströbel, P. (2016). Building a Parallel Corpus on the World's Oldest Banking Magazine. In *KONVENS*. s.n. doi: 10.5167/uzh-125746.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., & Ruef, B. (2010). Challenges in Building a Multilingual Alpine Heritage Corpus. In N. Calzolari et al. (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), 17–23 May 2010, Valletta, Malta.* European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/110_Paper.pdf

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). Recaptcha: Human-Based Character Recognition via Web Security Measures. *Science, 321*(5895), 1465–68.

Wang, C., Daneva, M., Van Sinderen, M., & Liang, P. (2019). A Systematic Mapping Study on Crowdsourced Requirements Engineering Using User Feedback. *Journal of Software: Evolution and Process, 31*(10), e2199.

Wilson, E. (1997). The Automatic Generation of CALL Exercises from General Corpora. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora (Applied linguistics and language study)* (pp. 116–30).

Wojatzki, M., Melamud, O., & Zesch, T. (2016). Bundled Gap Filling: A New Paradigm for Unambiguous Cloze Exercises. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, June 2016, San Diego, CA* (pp. 172–81). Association for Computational Linguistics (ACL). doi: 10.18653/v1/W16-0519

Zanetti, A., Volodina, E., & Graën, J. (2021). Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data. *International Journal of TESOL Studies, 3*(2), 55–71.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus V1.0. In N. Calzolari et al. (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), May 2016, Portorož, Slovenia.* European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1561.pdf

Učenje jezikov iz vzporednih korpusov: zasnova za spreminjanje korpusnih primerov v vaje za učenje jezikov

Članek opisuje arhitekturo aplikacije, ki iz vzporednih korpusov generira vaje za učenje jezika. Poravnava besed in vzporedne strukture omogočajo samodejno ocenjevanje stavčnih parov v izvornem in ciljnem jeziku, medtem ko uporabniki aplikacije s svojimi interakcijami nenehno izboljšujejo kakovost podatkovne zbirke in tako množičijo vzporedno jezikovno učno gradivo. S pomočjo triangulacije se lahko njihovo ocenjevanje prenese tudi na druge jezikovne pare, če kot vir uporabimo več vzporednih korpusov.

Da bi lahko takšna aplikacija delovala, je treba nasloviti več izzivov. V nadaljevanju bomo obravnavali tri. Prvič, v zadnjem desetletju se je nekaj pozornosti posvetilo vprašanju, kako v korpusih prepoznati ustrezno učno gradivo. Podrobno bomo opisali, kako na to vpliva struktura vzporednih korpusov. Drugič, katere vrste vaj je mogoče samodejno ustvariti iz vzporednih korpusov, tako da spodbujajo učenje in ohranjajo motivacijo učencev. In tretjič, kakšne so možnosti vključevanja uporabnikov, tj. učiteljev in učencev, kot množice, ki bi pomagala izboljšati gradivo.

Aplikacijo, ki jo opisujemo v članku, smo delno implementirali in preizkusili v različnih eksperimentalnih okoljih. Več funkcij, ki bodo vključene v končno programsko opremo, smo razvili in ovrednotili ločeno. Za implementacijo vseh delov, ki so podrobno opisani v tem dokumentu, pa je potrebno še veliko dela in razpoložljivost dejanskih učiteljev in učencev za namene preskušanja. Da bi lahko potrdili želene pozitivne učinke prispevkov uporabnikov, bo treba končne aplikacije uporabljati dalj časa, kar predstavlja še dodaten izziv.

Ključne besede: ICALL, vaje za učenje jezikov, vzporedni korpusi, učenje na podlagi podatkov, množičenje

Crowdsourcing and language learning habits and practices in Turkey, Bosnia and Herzegovina, the Republic of North Macedonia and Poland in the pre-pandemic and pandemic periods

Çiler HATIPOĞLU
Faculty of Education, Middle East Technical University

Nihada DELIBEGOVIĆ DŽANIĆ
University of Tuzla

Elżbieta GAJEK
University of Warsaw

Lina MILOSHEVSKA
University of Information Science and Technology

The popularity of online crowdsourcing platforms was slowly increasing among language learners before the pandemic, but COVID-19 changed the educational systems worldwide. This study aims to uncover whether or not, and if 'YES', how the attitudes and habits of language learners concerning the

Hatipoğlu, Ç., Delibegović Džanić, N., Gajek, E., Miloshevska, L.: Crowdsourcing and language learning habits and practices in Turkey, Bosnia and Herzegovina, the Republic of North Macedonia and Poland in the pre-pandemic and pandemic periods. Slovenščina 2.0, 10(2): 132–181.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.132-183

use of crowdsourcing materials in Turkey, Bosnia and Herzegovina, the Republic of North Macedonia and Poland changed during the pandemic.

To compare the pre-and during the covid crowdsourcing tool usage, the cross-culturally appropriate questionnaire utilised in the pre-COVID-19 period was used again. The collected data were analysed qualitatively and quantitatively to identify the differences between the periods.

The study's findings showed that the shift from face-to-face to online learning significantly affected the development of crowdsourcing platforms worldwide and their employment in the studied countries. The results also demonstrated that a combination of factors, such as reduced interactions with teachers and peers, an increase in workload, and a lack of support on the part of institutions, led to students taking responsibility for their learning. The number and characteristics of the popular platforms changed from country to country since expectations from students varied.

Keywords: crowdsourcing, language learning, COVID-19, pre-pandemic period, post-pandemic period

1 Introduction

Crowdsourcing, in Estellés-Arolas et al.'s (2015, p. 33) definition, is a problem-solving and task realisation model where thanks to harnessing collective intelligence, creative solutions to complex problems are found. Due to the success and usefulness of the initiative and its products, the number of fields embracing it (e.g., tourism, architecture, artificial intelligence) and researchers focusing on the concept (Lyding et al., 2018; Rodosthenous et al., 2019) have been steadily increasing (Xu et al., 2022). The popularity of online crowdsourcing platforms in language teaching and learning was slowly rising before the COVID-19 pandemic (Arhar Holdt et al., 2020; Gajek, 2020; Hatipoğlu et al., 2020; Miloshevska et al., 2021). Studies done in Turkey (TUR), Bosnia and Herzegovina (B&H), the Republic of North Macedonia (RNM), and Poland (POL) showed that language teachers used them both as in- and out-of-class materials, and students employed them as tools helping them to sharpen their skills and knowledge in the target languages and become more autonomous learners (Hatipoğlu et al., 2020, 2021; Miloshevska et al., 2021).

The COVID-19 pandemic has changed, however, the educational systems worldwide and the established teaching and learning practices.

Face-to-face classes were abruptly suspended in almost all countries, and this led to the disruption of "the original teaching plans of schools in these countries and regions" (Chen et al., 2020, p. 1). During the first COVID-19 period of online education (2019-2020 spring semester, especially after March 2020), teachers and students had to abandon their familiar settings and quickly adapt to the new environments, which was a stressful process for all involved parties (Akat and Karataş, 2020; Krajka, 2021). Teachers, who up to that point were experts in their fields but did not frequently use digital technology, had to learn about new tools and systems and modify their teaching methods, techniques, materials and assessment practices (Hatipoğlu et al., 2021). Unfortunately, many of the changes were done randomly or opportunistically, without being aware of and, therefore, not following any of the established Computer Assisted Language Learning (CALL) models (Bax, 2003; Hampel and Stickler, 2005). Students also had to adjust to the new, mainly solitary online environment where they were deprived of social contact with their peers and teachers, and could not expect constant support from their institutions (Miloshevska et al., 2020; Trung et al., 2020).

Studies related to the status of tertiary education during the second and third semesters of online learning and teaching in some countries (e.g., Australia, the USA, and Canada in Hickling et al., 2021; Latvia in Baranova et al., 2020) showed that both teachers and students successfully settled into new routines and started following practices that were more suitable for the prolonged period online education. But what about the foreign language students in TUR, B&H, RNM and POL? What were their learning and teaching experiences during the COVID-19 pandemic with regard to using online/digital crowdsourcing materials/platforms?

This study was conducted to find out whether there have been any changes related to the use of crowdsourcing materials for language learning and teaching purposes (e.g., sources such as *Wikipedia, Duolingo, Kahoot,* Online dictionaries, social media sites) in TUR, B&H, RNM and POL during the different phases of the COVID-19 pandemic (for more detailed examples and explanations, see the Literature Review section). These four countries were selected as focal points since some previous studies (Delibegović Džanić et al., in press; Hatipoğlu, 2021;

Miloshevska et al., 2020, 2021) showed that the national ministries of education and the university administrations planned and organized instruction in K-12 and tertiary levels differently. The expectations and requirements from students were also somewhat different in these countries. To reach our goal, the results of the Miloshevska et al.'s (2021) study, which are based on data from the pre- and initial COVID-19 periods, are compared with new data sets collected in the later phases of the pandemic (i.e., from December 2021 to March 2022, that is 2021-2022 spring semester).

2 Literature review

The term "crowdsourcing" was first coined by the American journalist Jeff Howe (2006) in an article for *Wired* magazine. The term was developed further, defined, and exemplified in his book *Crowdsourcing: How the Power of the Crowd is Driving the Future of Business* (Howe, 2008). In his work, Howe (2006, 2008) describes how the internet and the development of Web2.0 tools broke down the traditional way of doing work as well as the employer-employee relationships. He argued that thanks to the collaborative nature of the newly developed digital tools, companies, institutions, and even individuals, just by posting an open call, could now benefit from the wisdom of the usually heterogeneous crowd on the internet (e.g., volunteers, experts, even amateur enthusiasts) to find solutions for challenging problems, create new products, sort pictures and a multitude of other tasks. The reward the contributors receive in this setting depends on the company posting the call, the project's nature, and the crowd's interests. It could be either intangible (e.g., recognition or prestige within a group with specific interests, being entertained because they like playing a particular game) or tangible (e.g., money). The first of these practices is known as "microworking crowdsourcing" (e.g., people add entries to *Wikipedia*, but they are not paid), while the latter is called "benevolent crowdsourcing" (e.g., Amazon's Mechanical Turk pays individuals for their work[1]). This model of doing things and/or completing projects was first used in the business environment, but has evolved and spread and is now being used for different purposes in fields

[1] www.mturk.com

as diverse as geography (See et al., 2014), medicine (King et al., 2013), and multimedia (Soleymani and Larson, 2013). Nowadays, anyone can post videos on *YouTube* or *TikTok*, those who feel competent are free to write book reviews on Amazon, and ambitious, fearless artists submit their T-shirt designs to *Threadless* and wait for the crowd's verdict.

One other field where crowdsourcing has started gaining momentum and is being used more frequently in recent years is education. Since education theories, methods and techniques, as well as the learner profiles (e.g., daily routines, interests, social, cultural and language backgrounds) are getting increasingly diverse, conventional education, where traditional classrooms and textbooks limit students' experiences, is now being challenged and replaced by various other practices. Rapid developments in technology and greater respect for diversity in learning needs mean that the "wisdom of the crowd" is being considered by a growing number of competitive educational organizations (Çebi, 2018; Solemon et al., 2013; Wang, 2016). Crowdsourcing is used in various ways to support innovative education, and research shows that with such practices it is possible to create and offer authentic in- and out-of-class activities (Chen and Luo, 2014; Hui et al., 2014), innovative learning and teaching resources (Farasat et al., 2017), and context and student group-specific support (Goel, 2017; Shaikh et al., 2017; Weld et al. 2012).

Despite the advantages associated with the use of crowdsourcing in education, research also shows that one field where its use was not fully incorporated before the COVID-19 pandemic was language teaching and learning. Two main reasons have been identified as to why the inclusion of crowdsourcing activities in language education was still in its initial stage at this time: 1) the lack of knowledge on the part of the teachers, which led to 2) disinterest and gaps in students' knowledge related to them. In a study conducted by Arhar Holdt et al. (2019), where the researchers collected data from 1,129 language teachers from more than 30 countries, it was found that quite a significant number of the participants were not familiar with the concept of crowdsourcing, and therefore they were using a very small number of crowdsourcing activities in their classrooms. Maybe this is why, several years earlier, Odo (2016) published an article targeting language teachers and comparing the advantages and disadvantages of using crowdsourcing materials. He also presented some lists aiming to show language teachers how such materials could be used in the classroom, the stages of the lessons where they could be incorporated, what teachers are expected to do to encourage the use of such materials, and the available and useful crowdsourcing resources. Odo (2016) completed his article by arguing that the "potential of these resources is immense. Ignoring the possibilities for our classroom is a missed opportunity for our students to join a trend that could revitalize our language teaching and their learning" (p. 23).

2.1 Crowdsourcing research before COVID-19

The number of studies examining language learners' views of crowdsourcing was even more limited before the COVID-19 pandemic. The authors are aware of just four papers that specifically focus on language learners' views of crowdsourcing platforms (Gajek, 2020; Hatipoğlu et al., 2020; Miloshevska et al., 2021; Mospan, 2018), and this indicates a significant gap in the field, since it is essential that students (i.e., end users) accept the validity of a new product and begin to use it. Rafiee and Abbasian-Naghneh (2019, p. 1) maintain that there are "complex relationships between the perceived usefulness, perceived ease of use, e-learning motivation, online communication self-efficacy and language learners' acceptance and readiness of e-learning". Stated differently, knowing which crowdsourcing materials are employed by language learners, as well as when and how, is vital information not only for language teachers but also for platform creators, since it will aid them in developing and recommending resources to help students with their learning and progress.

Online teaching and/or blended learning were part of the educational system long before the COVID-19 pandemic. What is more, technology has often been used to support the continuity of teaching and learning in areas suffering from natural disasters (e.g., earthquakes, floods) (Baytiheh, 2018) or a lack of resources (e.g., large classes) (Krajka, 2021). These were, however, implementations in a limited number of places or in periods that were carefully planned and followed well-designed stages and procedures. The rest of the education, the vast bulk of it, both around the world and in TUR, B&H, RNM and POL was done face-to-face (Miloshevska et al., 2020).

2.2 Crowdsourcing research during COVID-19

The spread of COVID-19 and its identification as a pandemic led to sudden lockdowns in many countries. This required changes in all established ways of teaching and learning, including language education. All stakeholders in the educational institutions suddenly found themselves “in a new reality, with technology-mediated instruction of different kinds substituting for traditional face-to-face teaching” (Krajka, 2021, p. 112). Neither teachers nor students had access to the resources, methods and techniques they were used to, so they had to use a system that many of them were testing for the first time, i.e., online learning and teaching. As a result of this sudden but compulsory change, there was a boom in the development, usage and research related to the use of crowdsourcing materials during the COVID-19 period. When Kansal et al. (2021) used Google Trends analysis to uncover the platforms of online teaching and learning that made remote learning around the world possible, they found that there had been significant growth in the number of such platforms in just a year. They also reported that the “existing assets of educational establishments have effectively converted conventional education into new-age online education with the help of virtual classes and other key online tools in this continually fluctuating scholastic setting” (Kansal et al., 2021, p. 418). That is, faced with the harsh reality of lockdowns, platform developers, teachers, students, and researchers were all trying to find ways to help formal education continue.

The research done during the COVID-19 period can be placed mainly in three sometimes overlapping categories. The studies in the first group focus on uncovering, classifying and/or listing the platforms that could be used in such circumstances (Chen et al., 2020; Kansal et al., 2021; Reimers et al., 2020). Reimers et al. (2020, p. 2), for instance, prepared an annotated selection of “online educational resources to support the continuity of teaching and learning during the 2019-20 COVID-19 Pandemic with education leaders around the world". The list of resources was compiled based on the responses of 333 informants from 99 countries. They asked stakeholders participating in the survey to identify online educational resources that they had found helpful in supporting education continuity up to that point, and classified them into *Curriculum Resources, Professional Development Resources* or *Tools*. They also used Pellegrino and Hilton's (2012) taxonomy to provide information related to the foreign languages, skills, and subjects that can be taught using the materials, as well as the grades of students that could benefit from them and whether or not they developed the interpersonal and intrapersonal skills of the users.

To be able to construct a valid evaluation index that could be used in the case of other emergencies similar to the COVID-19 pandemic, Chen et al. (2020) asked users in China to review their experiences with online education platforms before and after the outbreak of the disease using criteria such as access speed, reliability, timely transmission of video data, course management, communication and interaction, and learning and technical support. The study focused on the performance of the seven most popular platforms in the country: *Chaoxing Learning*, *DingTalk*, *MOOC*, *Tencent Meeting*, *TIM*, *WeChatWork*, and *Zoom Cloud*. The analysis of the data showed that before the pandemic what users expected from a good platform were characteristics such as good access speed, reliability, and smooth transmission of video data. However, after the outbreak of COVID-19, when all classes moved online, they were more concerned with course management, communication and interaction, and the quality of the learning and technical support services of the platforms. In Chen et al.'s (2020) article, overall "Chaoxing Learning had the poorest user experience and DingTalk performed best" (p. 28).

The second group of studies tried to uncover the general benefits of using certain platforms as main or supplementary materials for language learners (Ali, 2022; Krishnan et al., 2020; Nadhifah and Puspitasari, 2021). When Krishnan et al. (2020) looked at how free online resources were used by language learners during the pandemic, they found two crucial facts. One, the user-friendly technologies that were freely available on the internet gained popularity during the COVID-19 crisis (i.e., they were used much more frequently and by a bigger number of students). Two, the educational lives of many students who reported experiencing economic problems during the pandemic were saved by freely available online resources, such as online dictionaries, *YouTube* videos, foreign language material development platforms, and grammar checkers. Nadhifah and Puspitasari (2021), who reviewed the effects of remote learning on students' study habits (before specifically focusing on *Duolingo*), maintain that the use of online platforms during the pandemic forced students to become more "responsible learners", and that the "pandemic condition urged them to conduct a self-regulated language learning by utilizing and optimizing the relevant media to learn" (p. 303).

The third group of studies examined whether, and if so how, certain online resources helped language learners develop specific language skills and sub-skills, such as listening, speaking, and pronunciation, and types of knowledge, including vocabulary and grammar (e.g., Krajka, 2021; Li and Xu, 2015; Nadhifah and Puspitasari, 2021; Trinh et al., 2021; Tsai, 2019; Waicekawsky et al., 2020). Nadhifah and Puspitasari (2021) studied the effects of *Duolingo* on the development of the structural knowledge of students with low and intermediate proficiency levels. They found that while intermediate-level students did not think they benefited much from the exercises on the platform, low-level learners stated that *Duolingo* helped them develop their grammar knowledge in English with tasks that were fun and appropriate for their level.

Trinh et al. (2021, p. 28), who worked with Vietnamese language learners, and Waicekawsky et al. (2020), whose participants were Argentinian EFL students, looked at the effects of another group of online resources that were frequently employed by foreign language learners during the pandemic – online dictionaries. Trinh et al.'s (2021) participants, who were native speakers of a tonal language (Alvez, 2006) and for whom speaking patterns in English are usually tricky, reported benefits such as improved intonation, pronunciation and grasp of vocabulary items' meaning. Consequently, the majority of the 300 junior students who participated in Trinh et al.'s (2021) study demonstrated a strong preference for online dictionaries over paper ones.

The concise literature review in this section demonstrates the striking differences in the use of crowdsourcing materials in educational settings before and during the COVID-19 pandemic. The current study aims to contribute to this area of research and examines the potential changes in TUR, B&H, RNM and POL.

The specific research questions that this study aims to answer are:

(1) What digital crowdsourcing resources did students in TUR, B&H, RNM and POL know about and use to learn foreign languages in the pre- and initial COVID-19 period (Period1, P1) versus the late COVID-19 period (Period2, P2)?

(2) Were there any changes in the frequencies, attitudes, contexts of use, and habits related to crowdsourcing materials of language learners in TUR, B&H, RNM and POL in P2 when compared to P1?

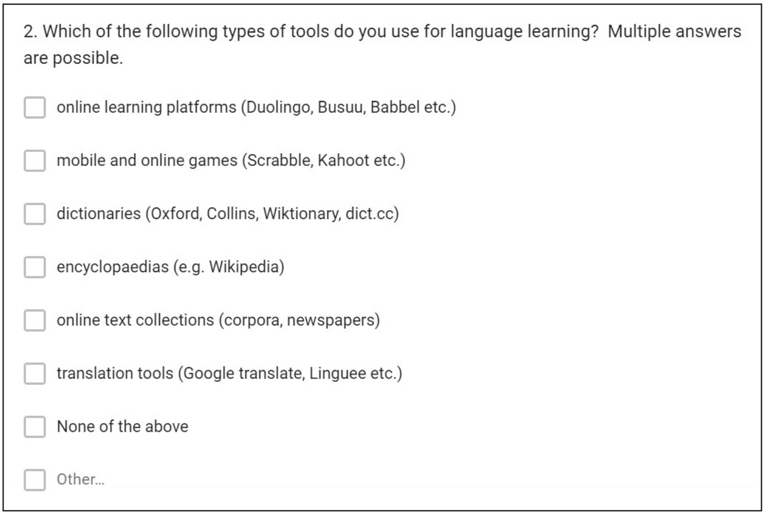
3 Methodology

3.1 Data Collection

In this study, the main aim was to uncover whether there have been any changes related to the use, attitudes, habits and contexts of use of crowdsourcing materials from the pre- to during the COVID19 periods by language learners in TUR, B&H, RNM and POL. To achieve this goal, the results of the authors' earlier study (Miloshevska et al., 2021) for which the data were collected in the pre- and during the emergency online teaching period in the Spring 2020 semester are compared with the new data collected in Spring 2022.

To ensure a reliable and valid comparison across countries and periods, the questionnaire designed for the initial study was utilized again, since it proved to be a cross-culturally appropriate data collection tool eliciting high-quality data enabling researchers to answer their research questions.

The written data collection tool employed in both studies had two sections, A and B. The 11 questions in Section A aimed to gather detailed information about the participants’ use of crowdsourcing tools and platforms. Nine of the 11 items in this section were checkbox questions where the participants could select multiple answers from a list of options (see Figure 1 for an example question). There was also one Likert scale item and one open-ended item. The Likert scale item asked participants to rate the crowdsourcing platforms they used from “Very enjoyable” (5) to “Not enjoyable at all” (1) and “I have not used it” (0). On the other hand, in the open-ended item the participants were asked to give information about their previous contributions to various crowdsourcing platforms.



2. Which of the following types of tools do you use for language learning? Multiple answers are possible.

□ online learning platforms (Duolingo, Busuu, Babbel etc.)

□ mobile and online games (Scrabble, Kahoot etc.)

□ dictionaries (Oxford, Collins, Wiktionary, dict.cc)

□ encyclopaedias (e.g. Wikipedia)

□ online text collections (corpora, newspapers)

□ translation tools (Google translate, Linguee etc.)

□ None of the above

□ Other...

Figure 1: Example checkbox question used in the study.

Section B of the questionnaire included six questions, and it aimed to collect data related to the participants' backgrounds. Four of the six questions were checkbox items, and two were open-ended.

3.2 Data Analysis

The collected data sets were analyzed both qualitatively and quantitatively to identify even the most minor changes between the compared periods in the studied four countries. The quantitative analyses were done using SPSS, where various descriptive (e.g., frequencies, percentages) tests were performed. The qualitative data were evaluated following the procedures proposed by Miles and Huberman (1994, pp. 58-69).

The tested hypothesis was that there had been a significant change in both teaching and learning habits in the pre-and pandemic periods, and that crowdsourcing platforms gained popularity during the COVID-19 period. This hypothesis was based on the findings of a study (Miloshevska et al., 2020) showing that teachers in B&H, NM, POL and TUR, similarly to their colleagues around the world, were forced to use almost all the digital tools they had at their disposal at this difficult time, especially during the emergency online teaching period in the Spring 2020 semester. At the same time, language learners were forced to independently use different crowdsourcing tools and platforms to catch up with the requirements of their institutions.

3.3 Participants

A total of 396 university students participated in the study. The participants in Study1 (Period1, Spring 2020) were 211 students from TUR (N=43, 20.4%), B&H (N=69, 32.7%), RNM (N=42, 19.4%) and POL (N=58, 27.5%) (see Table 1a). Their age range was 18-39, although 98.1% of them were 18-25 years old (Age Group 1: 18-21 years old, N=109, 51.7%; Age Group 2: 22-25 years old, N=98, 46.4%). Only 1.9% of the informants were in Age Group 3 (Range: 26-39; N=4).

Table 1a: *Participants in Period 1 (P1) (Spring 2020)*

	TUR	B&H	RNM	POL	ALL
Males (M)	12 (27.9%)	17 (24.6%)	27 (65.9%)	12 (20.7%)	69 (33%)
Females (F)	31 (72.,1%)	52 (75.4%)	14 (34.1%)	45 (79.3%)	142 (67%)
Prefer not to say	0	0	0	0	0
TOTAL	**43 (20.4%)**	**69 (32.7%)**	**41 (19.4%)**	**58 (27.5%)**	**211 (100%)**

As can be seen in Table 1a, 67% (N=142) of the participants were female, while 33% (N=69) were male. The informants from TUR, B&H and POL were training to become foreign language teachers, while the participants from RNM were Information and Communication, Engineering, and Computer Science Engineering students learning English for specific purposes. The smaller number of male participants in the study reflected the gender distribution of students at the Faculties of Education in TUR, B&H and POL (Can Daşkın & Hatipoğlu, 2019).

Table 1b: *Participants in Period 2 (P2) (Spring 2022)*

	TUR	B&H	RNM	POL	ALL
Males (M)	28 (46.7%)	11 (23.9%)	33 (67.3%)	5 (16.7%)	77 (42%)
Females (F)	32 (53.3%)	31 (67.4%)	16 (32.7%)	23 (76.7%)	102 (55%)
Prefer not to say	0	4 (8.7%)	0	2 (6.6%)	6 (3%)
TOTAL	**60** (32.4%)	**46** (24.9%)	**49** (26.5%)	**30** (16.2%)	**185 (100%)**

To have comparable informant groups to the first study (i.e., P1), the data in Study2 (Spring 2022) were collected from the same institutions and faculties. The total number of participants in Study2 was 185 – TUR (N=60, 32.4%), B&H (N=46, 24.9%), RNM (N=49, 26.5%) and POL (N=30, 16.2%) (see Table 1b) – and their age range was 17-40. Similarly to Study1, most of the students (94%) were 18-25 years old (Age Group 1: 18-21, N=114, 62%; Age Group 2: 22-25, N=60, 32%), and only 2% were in the 35-40 age group. Among the 185 participants, 55% (N=102) were female, and 42% (N=77) were male; 3% (N=6) of the informants ticked "Prefer not to say" as an answer to this question.

To check whether the students' language proficiency affected the type of crowdsourcing tools they utilized for language learning, the participants in both phases of the study were asked to self-evaluate using CEFR levels and criteria (Council of Europe, 2001).

As shown in Table 2a, in Study1 about two-thirds (65.4%) of the participants placed themselves in the *Proficient Users* (C1=79, 37.4% or C2=59, 28%) category, while 18.4% identified themselves as *Independent Users* (B1=6, 2.5% or B2=33, 15.6%). Only a small number of the participants from RNM stated they were *Basic Users* (A1=4, 1.9%; A2=3, 1.3%).

Table 2a: *Self-reported level of proficiency of the participants in Period 1 (Spring 2020)*

	TUR		B&H		RNM		POL		ALL	
	n	%	n	%	n	%	n	%	n	%
A1					4	**9.8**			4	**1.9**
A2					3	**7.3**			3	**1.4**
B1			2	**2.9**	4	**9.8**			6	**2.8**
B2	3	**7.0**	15	**21.7**	10	**24.4**	5	**8.6**	33	**15.6**
C1	8	**18.6**	27	**39.1**	13	**31.7**	31	**53.4**	79	**37.4**
C2	23	**53.5**	19	**27.5**	5	**12.2**	12	**20.7**	59	**28.0**
No answer	9	**20.9**	6	**8.7**	2	**4.9**	10	**17.2**	27	**12.8**
All	43	**100.0**	69	**100.0**	41	**100.0**	58	**100.0**	211	**100.0**

When the students participating in our Period 2 study were asked to evaluate their language proficiency, 99% of them placed themselves in either the *Proficient Users* (C1=91, 49.2% or C2=56, 30.3%) or *Independent Users* (B1=5, 2.7% or B2=31, 16.8%) categories (see Table

2b). Only one student from RNM chose the A1 level, and there was one student who did not respond to this question. So, similarly to the participants in P1, the students we are dealing with in Study 2 are also mainly advanced learners of English, and thus the preferences and experiences discussed in this paper are more relevant to learners with more advanced skills in the target languages.

Table 2b: *Self-reported level of proficiency of the participants in Period 2 (Spring 2022)*

	TUR		B&H		RNM		POL		ALL	
	n	%	n	%	n	%	n	%	n	%
A1					1	**2.0**			1	**0.5**
A2									0	**0.0**
B1			1	**2.2**	4	**8.2**			5	**2.7**
B2	4	**6.7**	11	**23.9**	15	**30.6**	1	**3.3**	31	**16.8**
C1	32	**53.3**	18	**39.1**	20	**40.8**	21	**70.0**	91	**49.2**
C2	24	**40.0**	15	**32.6**	9	**18.4**	8	**26.7**	56	**30.3**
No answer	0	**0.0**	1	**2.2**	0	**0.0**	0	**0.0**	1	**0.5**
All	60	**100.0**	46	**100.0**	49	**100.0**	30	**100.0**	185	**100.0**

4 Results and discussion

Before the COVID-19 pandemic, crowdsourcing materials were a relatively new phenomenon. Their use was beginning to gain pace, but they were still not often used in the educational context (Chen et al., 2020; Jiang et al., 2018) or in the four countries examined in this study (i.e., TUR, B&H, RNM and POL) (Miloshevska et al., 2021). The rapid switch from face-to-face to online learning, however, surprised and forced students, teachers, and institutions to alter their teaching and learning practices (Delibegović Džanić et al., in press; Hatipoğlu et al., 2022). What about the crowdsourcing resources that students use to learn languages? Did they change from the pre- to the late COVID-19 periods? Were there any changes in the frequencies, attitudes, contexts of use, and habits related to the crowdsourcing materials used by language learners in TUR, B&H, RNM and POL in P2 when compared to P1?

This study aims to answer these questions by comparing the crowdsourcing materials students from TUR, B&H, RNM, and POL knew about and used to learn foreign languages in P1 and P2. It was hoped that comparing students' answers in P1 and P2 would provide clues about the immediate and prolonged effects of online learning on the students' habits, and would enable different stakeholders in education to create more suitable and productive learning environments for the current and following generations of students based on empirical information coming from four distinct countries.

Analysis of the students' answers in P1 and P2 revealed some general tendencies observed across the four countries and certain country-specific changes (i.e., different countries were affected differently by the pandemic). One common feature was the **increase in the number of platforms listed by the students** in P2. In our first study, the total number of platforms reported by the participants was 26 (see Appendix A; for more details, see Miloshevska et al., 2021). Among those, POL students stated that they had used 14, TUR and B&H students 13 and the participants from RNM had used 8 (i.e., apart from the RNM students, the participants coming from the other three countries had experience with roughly the same number of online platforms). The number of platforms listed in P2 was 92 (i.e., 3.5 times more than in P1), and the percentage of students who said they had never used any crowdsourcing materials went from 6.6% in P1 down to 2.2% (see Appendix B). This finding can, on the one hand, be explained by Chen et al.'s (2020) claim that after the outbreak of COVID-19 the number of mobile online platforms increased because of the market demand for online education and the rise in the number of online platform users. Our participants could list more digital resources because more platforms catering to their needs had been created, and they could choose and use the ones they needed. Another plausible explanation for the observed sharp rise in the number of the listed online resources could be the new "strong technology literacy" (Ali, 2022, p. 202) of the students that was fostered by the prolonged online teaching and learning environment. In the second period examined in this study, P2, the students were still at home, away from their university campuses, with limited or no access to their teachers, peers, and university libraries. This meant that the resources and skills they used to depend on were partially or entirely inaccessible to them. But they had already had some experience with online learning, and they knew they had to develop new skills and find new resources to help them reach their goals in the new environment. And that is what they did. They improved their technology literacy and started searching for and using tools that best suited their needs.

Different to P1, there were clear differences between the number of crowdsourcing platforms used by the participants in the four countries. The data show that in P2 the TUR students reported using 60, RNM 26, B&H 23, and POL 23 of these platforms (i.e., in P2, TUR students used 4.6 times more crowdsourcing platforms, RNM participants 3.3, B&H 1.8 and POL 1.6). Based on these findings, it can be argued that TUR students were affected the most by the changed teaching mode, RNM students were affected moderately, and B&H and POL were affected the least. One explanation for the sharp rise in the number of online resources employed by TUR students in P2 might come from a study conducted by Delibegović Džanić et al. (in press) in TUR, B&H and RNM. In that study, students were asked to talk about the positive changes brought by online education and TUR students, like the one quoted in Example 1, frequently stated that one

Example 1: *TUR Student 71*

> *...positive effect and advantage might be my experiences about using web 2.0 tools, computer and doing effective search on net to get my answer and do my assignments more fruitful.*

That is, TUR students viewed their experiences with different online resources as something positive, as something that gave them a chance to improve their computer and digital literacy skills.

Among the 26 crowdsourcing sites listed in P1, six were used by the students in all four countries and with relatively high frequency (i.e., *Wikipedia* (N=158, 74.9%), *Kahoot* (N=133, 63%), *Duolingo* (N=130, 61.6%), *Khan Academy* (N=49, 23.2%), *Memrise* (N=43, 20.4%, *Busuu* (N=21, 10%). The remaining 20 platforms were usually rarely employed, and if they were, that usage was country-specific (i.e., they were employed in only one of the studied countries, e.g., *Rosetta Stone* in TUR; *Flocabulary* in B&H; *Quizlet* and *Anki* in POL) (for more details, see Miloshevska et al., 2021). In P2, five sources were used in all the studied countries: *Duolingo* (N=75, 40.5%), *Google Translate* (N-40, 21.6%), *Kahoot* (N=32, 17.3%), *Wikipedia* (N=32, 17.3) and *YouTube* (N=31, 16.8%). Similarly to P1, the remaining 87 platforms were utilized less frequently and not across all of the examined countries (see Appendix B).

As shown in Appendix B, the order of popularity and the characteristics of the most frequently used individual platforms changed from P1 to P2. Among the top six resources listed in P1, three were still at the top in the later period – *Duolingo* (N=75, 40.5%), *Kahoot* (N=32, 17.3%) and *Wikipedia* (N=32, 17.3%). However, the number of students who reported using them was much smaller. *Wikipedia*, the overwhelming favorite crowdsourcing resource before and during the first COVID-19 period, as well as *Kahoot* (N=32, 17.3%), *Khan Academy* (N=5, 2.7%), *Memrise* (N=3, 1.6%), and *Busuu* (N=2, 1.1%), were not the go-to sites in P2 anymore. In contrast, platforms such as *Google Translate* (N=40, 21.6%) and *YouTube* (N=31, 16.8%), which just one student in P1 mentioned, were now the second and fifth most popular crowdsourcing sites, respectively, for the students in TUR, B&H, RNM and POL.

Analyses of the contents and aims of the resources listed by the students in P2 showed that they could be grouped under seven categories (see Table 3). The biggest of those categories, as in P1, is still the **language learning and teaching platforms** (e.g., *Duolingo, Rosetta Stone*) (N=121, 25.9%), but together with those students reported using resources such as **online dictionaries** (e.g., *Cambridge Online Dictionary, Tureng*) (N=82, 17.5%), **professional development and collaboration**

Table 3: *Crowdsourcing resource sub-categories in P2*

	CATEGORIES	N	%
1.	Language learning and teaching platforms	121	25.9
2.	Online dictionaries	82	17.5
3.	Professional development and collaboration platforms	70	15
4.	Game-based platforms	65	13.9
5.	(Digital) TV channels and news media websites	62	13.2
6.	Translation and grammar monitoring platforms	58	12.4
7.	Social media messaging apps	6	1.3
8.	None of the above	4	0.9
	ALL	**468**	**100.0**

resources (e.g., *Anki, Wikipedia, Udemy*) (N=70, 15%), **game-based platforms** (e.g., *Kahoot, Scrabble*) (N=65, 13.9%), **(digital) TV channels and news media websites** (e.g., *Netflix, TwitchTV, YouTube, BBC websites*) (N=62, 13.2%), **translation and grammar monitoring platforms** (N=58, 12.4%), and **social media messaging apps** (N=6, 1.3%), which were mentioned by only a few students or not mentioned at all in P1.

A closer look at the listed platforms showed that in contrast to platforms such as *Wikipedia, Duolingo, Memrise, Khan Academy* and *Busuu* that offer general information or guidance related to learning foreign languages, in P2 the students started searching for and using more resources that catered to their country-specific and/or individual needs, and could fill in the gaps created by the lack of regular, in-person interaction with the most reliable sources of information, i.e., their lecturers, classmates and on-campus libraries.

4.1 Language Learning and Teaching Platforms

Students who participated in the P2 study listed 14 language learning and teaching platforms (LLTP) in total, and they formed 25.9% of all mentioned resources (121/468) (see Appendix B). Among those, *Duolingo* was the most popular tool (overall mentioned by 40.5% of the students in P2) and the only one named by the participants in all four countries (TUR: N=26, 43.3%; B&H: N=16, 34.8%; RNM: N=19, 38.8%; POL: N=14, 46.7%). However, when compared with P1, it was seen that even its popularity dropped 1.5 times in P2, as in P1, it was mentioned by 61.6% of the students.

One possible explanation for the fall in popularity of *Duolingo* in P2 in TUR, B&H, RNM and POL could come from Nadhifah and Puspitasari (2021), who worked with beginner- and intermediate-level undergraduate students in Indonesia. The students used *Duolingo* to learn English during the COVID-19 period as a self-learning tool. The results of the study showed that while beginner-level users felt satisfied with *Duolingo* since it was fun, easy to use and helped them develop their knowledge related to basic structures in English, the intermediate-level students reported that the platform did not really help in improving their target language skills. They maintained that it was a bit boring and too easy, but the more critical problem for them was the lack of discussion boards on the application. That is, the already isolated learners did not have a chance to share their experiences with each other while using *Duolingo*, and felt they "needed a place to share and to interact with the other users about their experience during using this application" (Nadhifah and Puspitasari, 2021, p. 308). As such, two things in Nadhifah and Puspitasari's work (2021) are particularly relevant to the current study. First, the students who participated in our P2 study were predominantly advanced learners of English (79.5% of the informants classified themselves as proficient users). *Duolingo*, a novel and exciting platform to use in P1, was now not satisfying their needs as advanced learners in P2. Second, the opportunity for social interaction, which has been shown to motivate students in self-regulated learning (Zimmerman and Schunk, 2001), was missing in *Duolingo*. This aspect of the *Duolingo* that was mentioned as a notable disadvantage by Indonesian students might have been a critical drawback for TUR, RNM, B&H and POL students when choosing a self-regulated learning platform in P2, too.

Among the remaining 13 platforms, *Quizlet* (listed by 14.1% of the students in P2), which only POL students mentioned in P1, was listed by the TUR, RNM and POL participants in P2. It was the second most popular platform overall in P2, and the most popular platform in POL (70%) once more. *Quizlet* is described as a "multi-facet CALL software" (Toy, 2019, p. 26) that can also be used as an online learning platform by both teachers and language learners. One reason why it was used by POL, TUR and RNM students in P2 could be the fact that it combines the benefits of classroom interactivity with personal self-study (Kose et al., 2016), and when using *Quizlet*, students can learn at their own pace and meet their individual needs better, in a fun manner. As seen in Example 2, it looks as if these features of *Quizlet* appealed to the student in three of the studied countries, and they started using it more in P2.

Example 2: *RNM student 16 (from Delibegović Džanić et al., in press)*

> *I have more free time since I can organize my time more freely.*

Khan Academy, Memrise and *Rosetta Stone* were listed by TUR and RNM students, while the remaining nine platforms were mentioned by either one or two students in a single country (e.g., *Lingodeer* in TUR, *Lingvist* in RNM).

4.2 Online Dictionaries

"The importance of dictionaries in language learning is indisputable" (Jin and Deifel, 2013, p. 515), as they help language learners understand new words' meanings, (contextual) usage, and grammatical features. With the creation of online dictionaries, students can now not only read and/or try to guess the pronunciation, intonation, and stress patterns of the words they encounter, but can also listen to and practice saying them. Jin and Deifel, in their 2013 study, claimed that "the emergence of online dictionaries has noticeably influenced the way students learn a foreign language" (p. 515).

Despite these benefits and claims, online dictionaries and thesauruses were not listed among the crowdsourcing materials students had used to learn foreign languages in our first study. This picture changed dramatically in P2, where they were the second most frequently mentioned group of resources (see Table 4). Students listed 18 online dictionaries and thesauruses in total, and more than half of the students in POL (N=20, 66.6%), TUR (N=34, 56.7%) and B&H (N=54.3%) said they were using these to learn foreign languages. The exceptional group was the RNM students, among whom only three (6.1%) listed any online dictionaries and thesauruses.

A closer look at the types and characteristics of the listed dictionaries shows that students not only consulted the "known"/"global" sources (e.g., *Cambridge*, *Oxford*, *Longman*), but they also started depending more on locally created online dictionaries (e.g., *Tureng* for TUR students, *DIKI* for the POL group) where entries related to language-/culture-specific terms, idioms and phrases, usually missing from the "general" sources, are included. Two such examples are the *Tureng Online Dictionary* (https://tureng.com/tr/turkce-ingilizce) initiated by a Turkish translation company, and *DIKI: Słownik Angielsko-Polski, Słownik Angielski Onlin*e (www.diki.pl), whose webserver is in Warsaw, Poland. *Tureng* was the second most frequently used dictionary by TUR students after the *Cambridge Online Dictionary*, and as shown in Figure 2 it includes translations for language-specific idiomatic expressions such as *"Ellerine sağlık".* This phrase, whose literal translation is *"Health to your hands"*, is a speech act that native speakers of Turkish use to compliment and express gratitude to their interlocutors simultaneously. Entries related to such phrases are included in *Tureng*, and if the speakers of the language think that their translations, definitions, and explanations should be broadened and/or refined, they can do that via a specific tab/function on the platform (see Figure 2). This, in turn, means that language learners have dictionaries on which they can rely

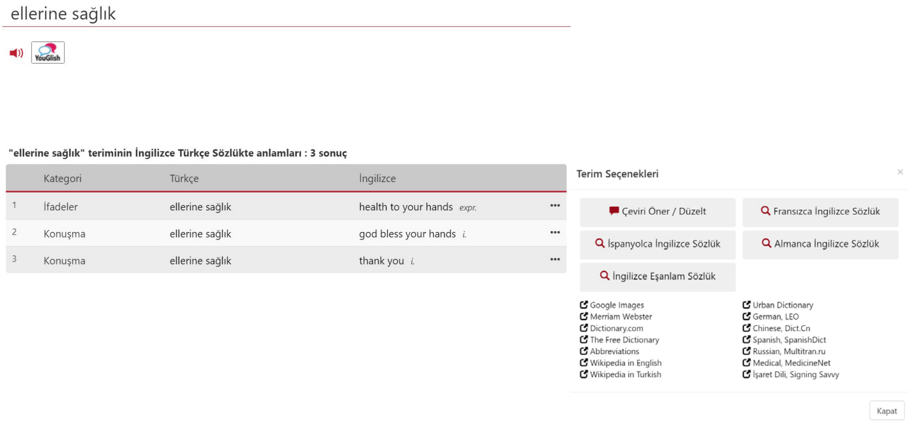
Table 4: *Online Dictionaries used for language learning in P2*

			TUR		B&H		MAC		POL		ALL	
		Tools	n	%	n	%	n	%	n	%	n	%
1.	3.	BAB.LA					1	**2.0**			1	**0.5**
2.	10.	Cambridge (Online) Dictionary	13	**21.7**	2	**4.3**					15	**8.1**
3.	13.	Diki							6	**20.0**	6	**3.2**
4.	20.	English idioms and phrases			1	**2.2**					1	**0.5**
5.	29.	Glosbe			5	**10.9**			1	**3.3**	6	**3.2**
6.	39.	Linguee							1	**3.3**	1	**0.5**
7.	41.	Longman (Online) Dictionary	1	**1.7**	1	**2.2**					2	**1.1**
8.	51.	One Look Thesaurus (online)	2	**3.3**							2	**1.1**
9.	52.	Online dictionaries	4	**6.7**	12	**26.1**			9	**30.0**	25	**13.6**
10.	54.	Oxford Online Dictionary	3	**5.0**	4	**8.7**					7	**3.8**
11.	55.	Ozdic	2	**3.3**							2	**1.1**
12.	58.	Pons							2	**6.7**	2	**1.1**
13.	64.	Relatedwords.org	1	**1.7**							1	**0.5**
14.	69.	SpanishDict					1	**2.0**			1	**0.5**
15.	75.	TheFreeDictionary	1	**1.7**							1	**0.5**
16.	76.	Tureng (online dictionary)	7	**11.7**							7	**3.8**
17.	82.	Urban Dictionary					1	**2.0**			1	**0.5**
18.	88.	Word Reference							1	**3.3**	1	**0.5**
		ALL	34	56.7	25	54.3	3	6.1	20	66.7	82	44.3

for translating phrases that they frequently use in their first language, want to use in their target language texts, but usually are not found in other dictionaries. Such dictionaries save time and maybe allow them to complete their work faster.

Another essential characteristic of some of the dictionaries listed by the students was that they were based on and/or benefited from the research done in corpus linguistics (e.g., *Ozdic*, *Relatedwords.org*, *Bab.la*, *English Idioms and phrases*). These new generation dictionaries are based on available corpora (e.g., the British National Corpus), and are regularly updated using internet searches to ensure "the most up-to-date usage for fast changing areas of language"[2]. Another advantage of these dictionaries is that they present easily searchable information related to **collocations** which are words or phrases that are

often used with another word or phrase, in a way that sounds correct to people who have spoken the language all their lives but might not be expected from the meaning, e.g., "*a hard frost*" but not "*a strong frost*" in English. (Cambridge Online Dictionary)[3]



ellerine sağlık

"ellerine sağlık" teriminin İngilizce Türkçe Sözlükte anlamları : 3 sonuç

	Kategori	Türkçe	İngilizce	
1	İfadeler	ellerine sağlık	health to your hands *expr.*	•••
2	Konuşma	ellerine sağlık	god bless your hands *i.*	•••
3	Konuşma	ellerine sağlık	thank you *i.*	•••

Terim Seçenekleri

Figure 2: TURENG Dictionary.[4]

Such information is essential for language learners, since research shows that even advanced learners of English have problems mastering collocations (Laufer and Waldman, 2011).

[2] https://ozdic.com/

[3] https://dictionary.cambridge.org/dictionary/english/collocation?q=collocations

[4] https://tureng.com/en/turkish-english

In addition to the above, this new generation of dictionaries (e.g., *Ozdic*) present the material in context with grammar and register information (e.g., daily/informal vs academic vs formal writing) as well as natural word combinations and alternatives. All of these help learners write using more "native-like language", and they are able to access and check that information quickly and for free.

All these features of the online dictionaries combined with the effects of "forced partial or complete isolation" during the second COVID-19 online learning period can explain the sharp increase in the use of these sources. Deprived of access to their teachers, peers and libraries, students had to find new, fast and reliable means to help them with the tasks at hand. Rundell (2014, p. 1) argues that "with easy access to numerous free reference sites, users searching for lexical information have a huge variety of options", and they choose online dictionaries because they include all the information contained in paper dictionaries but also materials that go "far beyond the traditional focus of 'the dictionary'" (Rundell, 2014, p. 6). That is, they include,

> language games, pedagogically-oriented videos, downloadable teaching materials, a weekly column on new words, and an active blog with regular contributions on a variety of language issues from both Macmillan's own editors and over a hundred guest bloggers. (Rundell, 2014, p. 6).

Stated differently, online dictionaries include many essential features that paper dictionaries, peers and lecturers provided in some way or another during the face-to-face teaching periods.

Besides the listed advantages of online dictionaries for language learners, another reason for using such a considerable number and wide variety of these resources might be the heavy course load, and the high number of homework projects assigned to students during the semesters taught online. In a study conducted in TUR, B&H and RNM (Delibegović Džanić et al., in press), students complained about the much heavier workload with the online teaching model, and how they struggled to complete their assignments even though they were studying much harder (see Example 3).

Example 3: *TUR Student 20*

> *I was studying regularly, but now, it is hard for me to focus on my homework not only because it is online, but also I have more course load than before. It is hard for me to catch up with all of the courses.*

During the online teaching periods, students were deprived of the systems they knew well and worked well for them (i.e., face-to-face classes where they worked closely with their lecturers and peers). They were on their own, and had a greater workload to deal with. Rundell (2014) and Trinh et al. (2021, p. 29) compared paper and digital dictionaries and argued that one specific advantage of the latter is that their "users can access and search large amounts of information quickly". Similarly, Li and Xu (2015) maintain that online resources would gradually replace bulky and outdated paper dictionaries, because with digital dictionaries the information retrieval rate is fast and using them is less time-consuming.

In short, during the first COVID-19 period the students did not use online dictionaries, and perhaps did not even know about many of them. However, it looks as if the combination of factors such as lack of access to known and reliable sources and an increase in workload during the second COVID-19 period forced students to look for new resources that would help them complete their assignments in a quick, reliable and high-quality manner, and thus they turned to online dictionaries.

4.3 Professional Development and Collaboration Platforms (PDCP)

Platforms that aim to or can support (pre-service) language teachers with their development as educators were included in the 'professional development and collaboration platform' (PDCP) category. These platforms had either one, a combination of or all of the content and characteristics listed below:

(i) Include resources (e.g., lessons, videos, interactive learning modules, texts) that directly support users in acquiring knowledge and skills.

(ii) Allow users to build online courses on various topics.
(iii) Contain course development tools that platform users can use to upload materials that foreign language learners might find helpful (e.g., texts, audios, videos, PowerPoint presentations, PDFs, ZIP files, source code for developers).
(iv) Allow users to engage and interact with each other via online discussion boards.

In P2, students listed 28 different PDCP platforms, which formed 15% of all crowdsourcing resources (see Table 3 and Appendix B). Apart from *Wikipedia* and *Fiszkoteka*, none of the remaining 26 PDCP were listed in our P1 study. Among the 28 platforms, TUR students stated that they used 15 (e.g., *Udemy* (N=2), *Wordwall* (N=2) and *Fandom* (N=2)), POL informants six (e.g., Anki (N=2), *Fiszkoteka*), the B&H (e.g., *Eng Vid, FunEasyLearn*) and RNM (e.g., *Coursera, Flocabulary)* groups five each.

There were a large number of platforms listed in P2 (N=28), but apart from *Wikipedia* (N=32, 17.3%), which the participants in all four countries named, each of the other resources were only mentioned by informants from one country. In our opinion, this emphasizes once more not only the richness of such resources (Chen et al., 2020; Kansal et al., 2021), but also the search of students for materials that best suit their needs. It thus looks as if the second COVID-19 period was a period of self-discovery, context discovery and switching from a teacher dependent to a more autonomous learner profile (also see Hatipoğlu et al., 2022).

In TUR, B&H, and POL the students were pre-service language teachers, and they all had to create high-quality work rapidly and by themselves in the second and third COVID-19 semesters. Each country followed, however, different rules and regulations regarding teaching policies at the tertiary level (Miloshevska et al., 2020) and practicum courses (Ersin et al., 2020; Krajka, 2021) during the lockdown periods. The participants in our study were also in a unique position, since while trying to expand their knowledge and English skills they also had to think about the best resources to help them develop the most suitable materials for the students in their practicum classes. Additionally, there were native language and cultural differences between the participant groups and the students with whom they were expected to interact in their school practice classes. Therefore, each group of students were on their own journey of discovery. They had to assess and understand what was required from them in their unique contexts, and then search for and identify the resources that catered best to their needs. Because of the lack of recommended platforms by the various ministries of education (Krajka, 2021), it is likely that the novice users of crowdsourcing platforms did not get it right the first time and had to search for and find something that would better fit their needs. Hence the large number of platforms listed in this category.

A closer look at the platforms employed by the students showed that they were varied in quality (i.e., from the most general to the more specific ones), content, information presentation, teaching and assessment styles and practices. The first category (i.e., ***General Resources***) of PDCP included electronic libraries and encyclopedias (*Wikipedia*), where students could find entries on numerous topics, academic and non-academic journals, and educational and general-interest books. The materials from these platforms were then either used in students' projects or as texts that could be taken and adapted to the needs of students in their practicum classes as they were of different ages and had different proficiency levels.

The second group of ***PDCP*** were the ***'Job specialized platforms'*** that allow users to create courses and materials tailored to their students' needs. These are platforms like *Udemy*, where educators have a comprehensive collection of tools (e.g., videos, source code for developers, PowerPoint presentations, PDFs, audio, ZIP files and any other content that learners might find helpful) that they can use to build online courses on specific topics. These platforms also allow course/material creators to engage and interact with their students and colleagues via online discussion boards.

The richest sub-group of PDCP was Group 3, which included tools with which materials for more specialized purposes could be created. These aim at developing particular types of knowledge (grammar, vocabulary) or skills for foreign language learners (speaking, writing, test-taking skills), and include platforms such as *Worldwall,*[5]

[5] https://wordwall.net/

Spike Notes,[6] *Easy Languages*,[7] *engVid*,[8] *FunEasyLearn*,[9] *Coursera*,[10] *Flocabulary*,[11] *Vocaroo*,[12] *Fiszkoteka*[13] (for the full list see Appendix B). The list of platforms in this group was the longest, proving once again that students were trying to find the ones that fit their needs the best.

What is more, depending on the severity of the pandemic in the examined countries and the level of access of the students to the internet and the required equipment (e.g., PC computers, laptops, smartphones), the various ministries of education planned and organized education in K-12 and tertiary education differently. In TUR, for instance, after it became clear that a considerable number of K-12 students had either no or limited access to the internet, in addition to strengthening the infrastructure of the already existing Educational Informatics Network (EBA) and introducing EBA-TV, the Ministry of National Education collaborated with the Turkish Radio and Television Corporation (TRT) and started airing the K-12 lessons on TRT channels (Özer, 2020). This change in policy forced pre-service teachers to prepare different materials for different modes of practicum applications, which in turn required the usage of a bigger number of crowdsourcing materials.

4.4 Game-based Platforms

Kahoot was the only game-based platform listed by the students in P1 (Miloshevska et al., 2021). It was a popular tool in all of the studied countries. Overall, it was used by 63% of the participants, but it was a particularly popular platform in POL and TUR, where respectively, 93.1% and 83.7% of the students stated that they had used it to learn languages (see Appendix A).

In P2, game-based platforms were the fourth most frequently used resource, while in P1 they formed only 13.9% of all mentioned resources (see Table 3). For P2 the participants listed nine platforms,

[6] https://www.sparknotes.com/
[7] https://www.easy-languages.org/
[8] https://www.engvid.com/
[9] https://www.funeasylearn.com/
[10] https://www.coursera.org/
[11] https://www.flocabulary.com/
[12] https://vocaroo.com/
[13] https://fiszkoteka.pl/

among which *Kahoot* was the most frequently mentioned one once again. However, when compared with P1, it can be seen that both overall and for individual countries, *Kahoot*'s popularity dropped significantly. In P2, it was used overall by 3.7 times fewer students and by only 25% of TUR, 20% of the POL, 17.4% of the B&H and only 6.1% of RNM students (see Table 5).

Table 5: *Game-based platforms used in P2*

			TUR		B&H		RNM		POL		ALL	
		Tools	n	%	n	%	n	%	n	%	n	%
1.	28.	Gamepedia	1	1.7	1	2.2					2	1.1
2.	34.	Kahoot	15	25	8	17.4	3	6.1	6	20	32	17.3
3.	44.	Minecraft			1	2.2	1	2			2	1.1
4.	53.	Online games	1	1.7	5	10.9	11	22.4	2	6.7	19	9
5.	60.	Quizizz	1	1.7							1	0.5
6.	67.	Scrabble			3	6.5					3	1.6
7.	71.	Steam language games	1	1.7							1	0.5
8.	83.	Video games	2	3.3			2	4.1			4	2.2
9.	89.	Word search (Puzzles)			1	2.2					1	0.5
		TOTAL	21	35	19	41.3	17	34.7	8	26.7	65	35.1

Another sub-category listed by participants in all four countries was the generic 'online games' group (N=19.9%). The category with the third highest percentage was a generic one, too – 'video games' (N=4, 2.2%). The presence of these two categories in the collected corpus was interpreted as the respondents saying "there are many online and video games that we use but not one, in particular, that is worth mentioning here". The remaining six games were mentioned by three or fewer students.

It has been known for a while now that online and video games can be effective language-learning tools (Hung, 2019; McNeil, 2020; Thorne and Reinhardt, 2008), since they offer benefits such as engaging dialogues, the opportunity to listen to various accents in English/ the target language, exposure to a variety of grammar structures, new vocabulary, stress relief, and also the possibility of making new friends all at once. Therefore, in a study done before the COVID-19 pandemic, Hung (2019) argued that the "use of learning games in educational contexts has expanded significantly, leading to the emergence of game-based learning as a recognized field of study" (p. 89). However, it looks as if, at least for this specific group of students, the situation changed with the outbreak of the pandemic. The changed conditions led to the replacement of online or video games with online dictionaries, grammar-checking programs, (digital) TV channels and social media platforms. In our opinion, this shift happened because the latter group of platforms responded to the overwhelmed students' needs faster, and provided a richer set of materials.

4.5. (Digital) TV Channels and News Media Websites

In P1, *YouTube* and 'movies and books' were entries listed by one B&H and one TUR student, respectively, while 'news media websites' were not mentioned at all. Stated differently, '(digital) TV channels and news media websites' (TVC&NMW) was an almost non-existent category in the pre-COVID-19 study. The picture changed in P2 when a total of twelve TVC&NMW were listed 62 times (13.2% of all mentioned resources) (see Table 6). Overall, 33.5% of the participants stated that they used one or a combination of these to learn languages.

When the lists of (digital) TV channels and news websites were compared, we saw that TV channels were used much more frequently. They formed 89% of the TVC&NMW resources (N=55), and websites comprised the remaining 11% (N=7). The most popular digital channel, as well as a resource in this group, was *YouTube.* It was listed in all four countries and by 16.8% (N=31) of all students. There might be two reasons why *YouTub*e, which was mentioned only once in P1, became such a popular resource during the pandemic. First, with its ever-growing content, *YouTube* has turned into an enormously rich library where users can easily find incredible amounts of information presented through multimodal means (Bloom and Johnston, 2010). In a period when not all teachers and institutions could supply all of the needed materials to their students, *YouTube* became a great alternative or supplement to books and lectures. Second, as early as 2013, Clarkson, in her book entitled *Usage of Social Network Sites amongst University Students*, argued that millions of people use platforms such

Table 6: *TV Channels and New Media Websites used in P2 TV Channels and New Media Websites used in P2*

			TUR		B&H		RNM		POL		ALL	
		Tools	n	%	N	%	n	%	n	%	n	%
1.	5.	BBC Learning English	1	**1.7**			1	**2.0**	1	**3.3**	3	**1.6**
2.	16.	DW Deutsch lernen	2	**3.3**							2	**1.1**
3.	46.	Movies					7	**14.3**			7	**3.8**
4.	48.	Netflix	7	**11.7**			2	**4.1**			9	**4.9**
5.	49.	News Websites	**1**	**1.7**							**1**	**0.5**
6.	57.	Podcasts	2	**3.3**							2	**1.1**
7.	73.	Ted Talks	1	**1.7**	1	**2.2**					2	**1.1**
8.	78.	TV5monde	2	**3.3**							2	**1.1**
9.	79.	Twitch.tv					1	**2.0**			1	**0.5**
10.	84.	Voice of America (VOA)	1	**1.7**							1	**0.5**
11.	91.	Younglish	1	**1.7**							1	**0.5**
12.	92.	YouTube	14	**23.3**	9	**19.6**	6	**12.2**	2	**6.7**	31	**16.8**
		ALL	**32**	**53.3**	**10**	**21.7**	**17**	**34.7**	**3**	**10.0**	**62**	**33.5**

as *Facebook* and *YouTube* to connect with each other based on shared interests, political views, or activities. That is, in a period of social isolation, *YouTube* became a safe space for learning communities “where everyone has a voice [and] anyone can contribute” (Educase Learning Initiative, 2006, p. 2).

Other popular resources in the TVC&NMW group were *Netflix* (N=9, 4.9%) and the generic category ‘movies’ (N=7, 3.8%). When discussing *Netflix*, some students listed generic categories such as ‘movies and/or series on *Netflix*’ while others specifically mentioned series like *A Life on Our Planet*, *Explained*, and *History 101* as beneficial resources for foreign language development. When discussing the benefits of watching movies and soap operas, Bhusal et al. (2020) argued that besides being “very good time passing activities”, they could also be good motivators and, when related to our areas of study, further help with learning some additional content. They can also help foreign language learners with their vocabulary, pronunciation and listening comprehension.

What is interesting about the specifically mentioned programs on *Netflix* is that they are typically short (e.g., *Explained* is less than 20 min long) but tackle some key topics (e.g., *A Life on Our Planet* follows David

Attenborough, who maps the sharp decrease in our planet's biodiversity; *Explained* focuses on issues such as money, the mind, and voting) that are also usually taught in foreign language classes. As such, one reason for the popularity of those programs could be their versatility. That is, by watching these programs, language learners can improve their target language knowledge, while pre-service language teachers might also use them as teaching materials in their practicum classes.

Four news media websites (including *BBC Learning English* and *Voice of America*) were mentioned by the students in P2, and they formed 3.8% (N=7) of all resources. The number of these sites is relatively small, but keeping in mind that they were not mentioned at all in P1, it is encouraging to know that the students were searching for and exploring new resources that have been proven to help others with their foreign language proficiency development. In a study conducted with Indonesian students learning English at the tertiary level, Barella and Linarih (2020) asked participants to listen to the news on various news websites as an extensive listening activity twice a week, and to keep listening logs where they note the names of the websites, the types and lengths of the news shows, and ask and answer questions related to the content of the material they listened to. Similarly to the participants in our study, Indonesian students found *VOA Learning English* and *BBC News*, as well as the *CNN* and *National Geographic* websites, useful sources in helping them develop their foreign language skills. Ninety percent of the students in Barella and Linarih's (2020 study stated that learning English while listening to the news increased their motivation and made learning fun. They also argued that the extensive listening exercises helped improve their listening and speaking skills and expanded their vocabulary.

A close look at the data also showed that there were important differences between the four studied countries in the use of TVC&NMW. Within the TUR group, 53.3% of all students stated that they had used TVC&NMW to learn languages, and listed 10 of the 12 resources in this group. In the remaining three countries, both the number of tools listed and the percentages of the students who employed them were much lower. The RNM group listed three sources, while the B&H and POL groups listed only two. When the percentages of the students helped by these tools are compared, it can be seen that 34.7% of the RNM, 21.7% of B&H and only 10% of POL students listed those resources. Once again, it can be seen that online teaching had a different effects in the examined countries.

4.6 Translation and Grammar Monitoring Programs

Writing in a foreign language is a complex and challenging task (Uluçay and Hatipoğlu, 2017). To write acceptable texts in their non-native language, students must know the target language's spelling, punctuation, and grammar (Hatipoğlu and Algi, 2018). They must also master the register and genre-specific vocabulary and use them appropriately (e.g., collocations, idioms, proverbs). Finally, after creating the first draft, they must revise, reorganize, and edit their texts, keeping in mind language and culture-specific rhetoric rules (Bakry and Alsamadani, 2015; Sokolik, 2003).

Because of all these difficulties associated with writing in a foreign language, and due to the increased rate of communication in English in the last few decades, several companies and research groups have created and developed various pieces of software that can help learners of foreign languages with their grammar, translation and paraphrasing in their target language. Some popular programs are *Google Translate, Grammarly*, and *ReversoContext.*

Table 7: *Translation and Grammar Monitoring Tools used in P2*

			TUR		B&H		RNM		POL		ALL	
		Tools	n	%	n	%	n	%	n	%	n	%
1.	11.	Conjugato							1	3.3	1	0.5
2.	30.	Google Translate	8	13.3	16	34.8	16	32.7			40	21.6
3.	31.	Grammarly	5	8.3			3	6.1	1	3.3	9	4.9
4.	59.	Quillbot (paraphrasing tool)					1	2.0			1	0.5
5.	65.	ReversoContext							6	20.0	6	3.2
6.	77.	Turnitin (feedback)	1	1.7							1	0.5
		ALL	14	23.3	16	34.8	20	40.8	8	26.7	58	31.4

However, in P1, *Grammarly* and *Google Translate* were listed by just one student and from only one country (TUR). None of the other translation and computer-mediated corrective feedback digital tools were mentioned. That is, before and during the first emergency COVID-19 period, such tools were not among the ones students in the studied countries were (often) using to learn or improve their foreign languages. This was not the case in P2 (see Table 7), when, overall, 31.4% of students started utilizing translation and corrective feedback tools to improve their target languages. In the second study, those tools were listed by 40.8% of RNM, 34.8% of B&H, 26.7% of POL and 23.3% of TUR students. The participants in the second study listed six tools, and *Google Translate* was the most popular one in TUR, B&H and RNM. The preferred translation tool for the POL students was *Reverso-Context*, which was listed by 20% of them.

One explanation for why students started using translation programs more during the second COVID-19 period, as mentioned above, could be the need to create good-quality papers in a short time, and research in the field shows that such programs can help in this regard. Tsai (2019) worked with native speakers of Chinese learning English and asked them to (1) write an essay in Chinese, (2) draft the same essay in English, (3) translate the Chinese essay into English using *Google Translate*. When the self-written and Google translated texts of the students were compared, it was found that the ones created using *Google Translate* "presented a number of components of significantly higher writing quality than those of students' SW (self-written) texts, by having more words, fewer mistakes in spelling and grammar, and fewer errors per words" (Tsai, 2019, p. 510). There were also more advanced-level words in the texts created with *Google Translate*.

The most popular grammar monitoring program in the examined countries was *Grammarly,* which was listed by TUR, RNM and POL students (N=9, 4.9%). *Grammarly* is a cloud-based typing assistant that identifies duplicate content and reviews grammar, vocabulary, mechanics (spelling, punctuation errors), as well as language style and delivery mistakes (Bailey and Lee, 2020; Barrot, 2020). One of the more critical advantages of this program mentioned in the literature is that it reduces the errors related to "vocabulary usages (diction), language use (grammar), and mechanics of writing (spelling and punctuation)" (Ghufron and Rosyida, 2018, p. 395; also see Bailey & Lee, 2020; Barrot, 2020). However, it is usually found to be less effective in improving "the content and organization of students' EFL writing" (Ghufron and Rosyida, 2018, p. 395). It thus looks as if the participants in our study employed *Grammarly* to quickly clean up their texts with regard to problems related to vocabulary, grammar usage, spelling and punctuation in order to concentrate more on the content and organization of their work, as well as on issues stemming from the possible influence of their cultural and first language norms of writing.

What is more, in a study done with Indonesian English education study program students, it was shown that when used to teach reading and writing in English as a foreign language, *Grammarly*, in combination with digital tools such as *Telegram*, *WhatsApp*, *Google Meet*, *YouTube*, and a *Plagiarism Checker*, made a positive contribution not only to the development of the students' proficiency in their target language, but also to the improvement of their knowledge related to these new digital tools and their self-esteem and belief in themselves (Setyowati et al., 2021).

The lockdown periods during COVID-19 took students away from the known and conventional face-to-face classrooms and pushed them into the technology-based online instruction environment. With the uncertainty of when the pandemic would end, they had two options: to give up and freeze their semesters, or to adapt as quickly as possible and continue fighting and learning. From the information gathered in the current study, it looks as if many of the students chose the latter approach.

4.7 Social Media Messaging Apps

The 'social media messaging apps' (SMMA) category was a non-existent category in our first study. None of the 211 students in P1 mentioned any SMMA as tools they had used to learn foreign languages. In P2, the participants listed SMMA six times, accounting for 3.2% of all resources in the second study (see Table 8). Still, the number of learners who found them helpful in supporting their target language development was relatively small compared to the other categories in the current study.

Another important fact related to the use of this category is that, once again, the bulk of the students who treated them as foreign language learning tools were from TUR (four out of six students, 67%). Only one student from B&H and one from POL stated that they used *Reddit.com* and *Discord*, respectively, while none of the students from the RNM listed any SMMA. One possible reason for the observed difference could be the more positive approach to using SMMA in educational settings and teacher education in Turkey in the last two decades. Such applications are seen as an intelligent employment of existing technologies in the classroom (Mendez et al., 2009, p.1), and it is believed that teacher education can benefit from such applications in two ways. First, SMMA can be used to enhance (pre- and in-service) teachers' learning and preparation for the job, and second, in language classrooms, where social media tools could make the learning environment more engaging and benefit the teaching of the target language (Albion, 2008). Balçıkanlı (2010), who examined the effects of social networking on pre-service English teachers' metacognitive awareness and teaching practices in Turkey, found that they both were positively affected.

Table 8: *Messaging Apps used in P2*

			TUR		B&H		RNM		POL		ALL	
		Messaging Apps	n	%	n	%	n	%	n	%	n	%
1.	8.	Bottled	1	1.7							1	0.5
2.	14.	Discord							1	3.3	1	0.5
3.	56.	Plotagon	1	1.7							1	0.5
4.	63.	Reddit.com	1	1.7	1	2.2					2	1.1
5.	68.	Slowly (Twitter app)	1	1.7							1	0.5
		ALL	4	6.8	1	2.2			1	3.3	6	3.2

The introduction of this new group of tools among the resources that students use to learn foreign languages could also be seen as a support for Chic and Benson's (2020) claim that online media awareness was particularly high during the COVID-19 pandemic, since most of the world had to adopt digital, online means of working. Language learners, like others, had to start reading, writing and learning online (i.e., they had to start doing things differently), because this was the only way of accessing the information they needed during this time. What is more, they saw the SMMA, as Barlett-Brag (2006, p. 3) predicted and described, as a "range of applications that augments group interactions and shared spaces for collaboration, social connections, and aggregates information exchanges in a web-based environment". Once again, students showed flexibility and initiative. When the world required them to change, young people analyzed the situation correctly and adapted accordingly.

5 Conclusions

Education systems around the world have changed because of the COVID-19 pandemic, but what about the learning of foreign languages in this context?

To contribute to answering a part of this important question, the current study focused on four distinct countries – TUR, B&H, RNM and POL – and tried to find answers to the following two research questions:

(1) What digital crowdsourcing resources did students in TUR, B&H, RNM and POL know about and use to learn foreign languages in the pre- and first- COVID-19 (P1) versus late COVID-19 periods (P2)?

(2) Were there any changes in the frequencies, attitudes, contexts of use, and habits related to crowdsourcing materials of language learners in TUR, B&H, RNM and POL in P2 when compared to P1?

The results of the study show that in general terms the answers to these questions are, respectively, "many and varied" and "yes". The attitudes, knowledge and use of crowdsourcing materials by language learners changed from the beginning towards the later phases of the pandemic. Overall, the study's findings show that the shift from face-to-face to online learning because of COVID-19 significantly affected the development and use of crowdsourcing materials in the studied countries.

In parallel to some earlier studies in the field (e.g., Krishnan et al., 2020), our findings show that the second COVID-19 period (P2) was marked by the use of a much richer range of digital resources when compared with P1. This, in our opinion, points to three facts. One, as claimed in some previous studies (Chen et al., 2020; Kansal et al., 2021), the number of such resources increased, and the students could find, test, and select the ones that fit their needs better. Second, with the experience they gained in P1, students became more skillful in searching for, finding and using such resources. Learners now had a stronger technological literacy and could use not just one or two but many digital platforms. Three, the students became more autonomous learners, able to better understand their specific contexts and what was expected from them in these (i.e., create high-quality work within a short period of time on your own). There was a realization that they had to switch to a self-regulated learning program because they were the only people who knew what they needed. Therefore, they had to plan and monitor their own actions. They had to search for, find and use the platforms that they thought best fit their needs; and, in the end, they had to reflect on the outcomes of their actions. Since different countries had different expectations from their students, each group needed to follow different paths. Therefore, only a limited number of the listed resources overlapped among the examined countries. The results once again showed students' ability to read their contexts well, and their success in adapting accordingly.

It was also seen that there was not only an increase in the number of resources used, but there was also a change in the students' expectations from the platforms and, as a result, in their features. Resources that were popular in P1 (e.g., *Wikipedia, Kahoot*) became less popular in P2, and the ones that were not mentioned at all or very rarely mentioned in P1 came to the fore (e.g., online dictionaries, *YouTube*, social media platforms). There was a general shift from the more general resources to the more needs and country specific ones (e.g., the use of *DIKI* by POL and *Tureng* by TUR students). The ones that presented more tailor-made opportunities for the personal development of the students were the ones that were selected.

It looks as if the students' attitudes towards and expectations from crowdsourcing platforms also changed in P2. They were not only good sources of information for the students but also safe spaces where users were able to connect with people with similar interests and views, and jointly work on different projects (i.e., they were spaces where "communities of practices" were formed, see Wenger, 1998). The study also seems to support the claim (Schunk and Zimmerman, 2001) that social interaction is a prerequisite for an increase in student motivation and progress in self-regulated learning. Platforms that did not allow for such collaboration lost popularity (e.g., *Duolingo* in P2). The results also showed that a combination of factors such as isolation, lack of access to familiar resources (e.g., teachers, peers, university libraries), increase in workload, and lack of support on the part of institutions might have led to this shift.

The study also found that when it comes to the use of crowdsourcing materials, the changes observed in TUR were much more noticeable than in B&H, RNM and POL. When compared with P1, in P2, TUR students listed the widest variety and biggest number of digital tools among the four groups of students. Factors that might have contributed to the observed differences could be varying policies related to the education practices in K-12 and tertiary education in the examined countries, and decisions related to practicum classes, workload requirements, and cultural differences.

It is hoped that the findings of the study will serve as potential guidelines for language teachers who plan to incorporate crowdsourcing activities into their in- and out-of-class activities in the future. However, they might also provide essential feedback to both groups of platform creators: the ones who aim to design resources that are valid cross-culturally and those who are seeking to create platforms that would cater to language learners with more specialized needs and interests. It is believed that incorporating crowdsourcing resources in language curricula can provide students with more in- and out-of-class collaboration opportunities and more active language learning, which, in turn, will lead to the development of more independent, active and confident language learners.

Despite the careful collection and analysis of the data sets discussed in this study, it should be mentioned that this research was based on questionnaire data gathered from relatively small and unequal (e.g., across countries and genders) samples of students in TUR, B&H, RNM and POL. Therefore, studies where bigger samples and additional data collection tools (e.g., interviews, observations, exam results and student projects) are employed should be conducted to enhance our understanding of the real changes in the use of crowdsourcing materials for foreign language learning, not only in our four countries but also in others that had similar experiences during the COVID-19 pandemic. Studies where the viewpoints of university lecturers, teachers, university/school administrators or other stakeholders are examined should also be done so that we get a more detailed and realistic picture related to the shifts and/or changes that occurred with regard to the use of crowdsourcing tools in before and after COVID-19 in foreign language teaching and learning.

References

Akat, M., & Karataş, K. (2020). Psychological effects of COVID-19 pandemic on society and its reflections on education. *Electronic Turkish Studies*, *15*(4), 1–13. doi: 10.7827/TurkishStudies.44336

Albion, P. (2008). Web 2.0 in teacher education: two imperatives for action. *Computers in the Schools*, *25*(3/4), 181–198. doi: 10.1080/07380560802368173

Ali, Z. (2022). 21st-century learning: Understanding the language learning strategies with technology literacy among L2 learners. *Journal of Nusantara Studies (JONUS)*, *7*(2), 202–220. doi: 10.24200/jonus.vol7iss2pp202-220

Arhar Holdt, Š., Zviel-Girshin, R., Gajek, E., Durán-Muñoz, I., Bago, P.; Fort, K.; Hatipoğlu, Ç., ..., & Zanasi, L. (2020). Language Teachers and Crowdsourcing: Insights from a Cross-European Survey. *Rasprave, 46*(1), 1–28. Retrieved from https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=353213

Bailey, D., & Lee, A. R. (2020). An Exploratory Study of Grammarly in the Language Learning Context: An Analysis of Test-Based, Textbook-Based and Facebook Corpora. *TESOL International Journal*, *15*(2), 4–27. Retrieved from https://www.tesol-international-journal.com/

Bakry, M. S., & Alsamadani, H. A. (2015). Improving the persuasive essay writing of students of Arabic as a foreign language (AFL): Effects of self-regulated strategy development. *Procedia-Social and Behavioral Sciences*, *182*, 89–97. doi: 10.1016/j.sbspro.2015.04.742

Balçıkanlı, C. (2010). The Effects of Social Networking on Pre-service English Teachers' Metacognitive Awareness and Teaching Practice. Unpublished Ph.D. Dissertation. Gazi University.

Baranova, T., Kobicheva, A., & Tokareva, E. (2021). Total transition to online learning: students' and teachers' motivation and attitudes. In D. Bylieva, A. Nordmann, O. Shipunova & V. Volkova (Eds.), *Knowledge in the Information Society* (pp. 301–310). Springer, Cham. https://link.springer.com/content/pdf/10.1007/978-3-030-65857-1.pdf

Barella, Y., & Linarsih, A. (2020). Extensive listening practice in EFL classroom with variety of news websites. *Pedagogy: Journal of English Language Teaching*, *8*(1), 43–50. doi: 10.32332/pedagogy.v8i1.1961

Barrot, J. S. (2020). Integrating technology into ESL/EFL writing through Grammarly. *RELC Journal*, 0033688220966632. doi: 10.1177/0033688220966632

Bartlett-Bragg, A. (2006). Reflections on pedagogy: Reframing practice to foster informal learning with social software. Retrieved from http://matchsz.inf.elte.hu/tt/docs/Anne20Bartlett-Bragg.pdf

Bax, S. (2003). CALL – past, present and future. *System*, *31*, 13–28. doi: 10.1016/S0346-251X(02)00071-4

Baytiheh, H. (2018). Online learning during post-earthquake school closures. *Disaster Prevention and Management*, *27*(1), 215–227. doi: 10.1108/DPM-07-2017-0173

Bhusal, S., Niroula, A., & Kafle, R. (2020). Quarantine: A Period of Self-discovery and Motivation as Medical Student. *JNMA: Journal of the Nepal Medical Association*, *58*(227), 536–539. doi: 10.31729/jnma.5005

Bloom, K., & Johnston, K. M. (2010). Digging into YouTube videos: Using media literacy and participatory culture to promote cross-cultural understanding. *Journal of Media Literacy Education*, *2*(2), 113–123. doi: 10.23860/jmle-2-2-3

Can Daşkın, N., & Hatipoğlu, Ç. (2019). A proverb learned is a proverb earned: Proverb instruction in EFL classrooms. *Eurasian Journal of Applied Linguistics*, *5*(1), 57–88. doi: 10.32601/ejal.543781

Chen, Z., & Luo, B. (2014). Quasi-Crowdsourcing Testing for Educational Projects. In *Companion Proceedings of the 36th International Conference on Software Engineering* (pp. 272–275). ACM Press. Retrieved from https://dl.acm.org/doi/pdf/10.1145/2591062.2591153

Chen, T., Peng, L., Jing, B., Wu, C., Yang, J., & Cong, G. (2020). The impact of the COVID-19 pandemic on user experience with online education platforms in China. *Sustainability*, *12*(18), 7329, 1–31. doi: 10.3390/su12187329

Chik, A., & Benson, P. (2020). Commentary: Digital language and learning in the time of cronavirus. *Linguistics and Education*, *62*, 100873. doi: 10.1016/j.linged.2019.100750

Clarkson, K. (2013). *Usage of Social Network Sites amongst University Students*. Grin Verlag.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University.

Çebi, A. (2018). Teachers' Perceptions Toward Technology Integration into the Language Teaching Practices. *Journal of Narrative and Language Studies*, *6*(11), 150–177.

Delibegović Džanić, N., Hatipoğlu, Ç., Milosevska, L., & Gajek, E. (in press). Has online learning changed the way we teach and study?: Student evaluation of teachers' pedagogical skills during the first COVID-19 period and potential change in their learning habits. *Folia Linguistica et Litteraria*.

DIKI. Słownik Angielsko-Polski, Słownik Angielski Online. Retrieved from www.diki.pl

Educase Learning Initiative. (2006). 7 Things You Should Know About YouTube. Retrieved from https://library.educause.edu/-/media/files/library/2006/9/eli7018-pdf.pdf

Ersin, P., Atay, D., & Mede, E. (2020). Boosting preservice teachers' competence and online teaching readiness through e-practicum during the COVID-19 outbreak. *International Journal of TESOL Studies*, *2*(2), 112-124. https://doi.org/10.46451/ijts.2020.09.09

Estellés-Arolas, E., Navarro-Giner, R., & González-Ladrón-de-Guevara, F. (2015). Crowdsourcing Fundamentals: Definition and Typology. In F. J. Garrigos-Simon, I. Gil-Pechuán & S. Estelles-Miguel (Eds.), *Advances in Crowdsourcing* (pp. 33–48). Springer.

Farasat, A., Nikolaev, A., Miller, S., & Gopalsamy, R. (2017). Crowdlearning: Towards Collaborative Problem-Posing at Scale. In *Proceedings of the Fourth ACM Conference on Learning@ Scale* (pp. 221–224). ACM Press. doi: 10.1145/3051457.3053990

Gajek, E. (2020). Crowdsourcing in language learning as a continuation of CALL in varied technological, social, and ethical contexts. In K. M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds.), *CALL for widening participation: Short papers from EUROCALL 2020* (pp. 75–80). Research-publishing.net. Retrieved from https://research-publishing.net/publication/978-2-490057-81-8.pdf

Ghufron, M. A., & Rosyida, F. (2018). The role of Grammarly in assessing English as a Foreign Language (EFL) writing. *Lingua Cultura*, *12*(4), 395–403. doi: 10.21512/lc.v12i4.4582

Goel, D. (2017). Because learning should be Chimple: How storytellers and artists can help kids read and write. Retrieved from http://www.edexlive.com/people/2017/oct/11/because-learningshould-be-chimple-how-storytellers-and-artists-can-help-kids-read-and-write-1326)

Hampel, R., & Stickler, U. (2005). New skills for new classrooms: Training tutors to teach languages online. *Computer Assisted Language Learning*, *18*(4), 311–326. doi: 10.1080/09588220500335455

Hatipoğlu, Ç., & Algi, S. (2018). Catch a tiger by the toe: Modal hedges in EFL argumentative paragraphs. *Educational Sciences: Theory and Practice*, *18*(4), 957–982. doi: 10.12738/estp.2018.4.0373

Hatipoğlu, Ç., Gajek, E., Milosevska, L., & Delibegović Džanić, N. (2020). Crowdsourcing for Widening Participation and Learning Opportunities: A view from pre-service language teachers' window. In K. M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds.), *CALL for widening participation: Short papers from EUROCALL 2020* (pp. 81–87). Retrieved from https://research-publishing.net/publication/978-2-490057-81-8.pdf

Hatipoğlu, Ç., Gajek, E., Milosevska, L., & Delibegović Džanić, N. (2021). Student evaluation of teachers' pedagogical skills during the first COVID-19 period. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley & S. Thouësny (Eds.), *CALL and professionalisation: Short papers from EUROCALL 2021* (pp. 119–125). Retrieved from https://research-publishing.net/publication/978-2-490057-97-9.pdf

Hatipoğlu, Ç., Gajek, E., Delibegović Džanić, N., & Milosevska, L. (2022). Comparative analysis of students' learning in the first and second semester of COVID-19 related online education in Türkiye, Poland, Republic of North Macedonia and Bosnia and Herzegovina. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny & M. J. Whelpton (Eds.), *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022* (pp. 154–161). Retrieved from https://doi.org/10.14705/rpnet.2022.61.1451

Hickling, S., Bhatti, A., Arena, G., Kite, J., Denny, J., Spencer, N. L., & Bowles, D. C. (2021). Adapting to teaching during a pandemic: Pedagogical adjustments for the next semester of teaching during COVID-19 and future online learning. *Pedagogy in Health Promotion*, *7*(2), 95–102. doi: 10.1177/2373379920987264

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–4.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Hui, J. S., Gerber, E. M., & Dow, S. P. (2014). Crowd-based design activities: helping students connect with users online. In *Proceedings of the 2014 conference on Designing Interactive Systems* (pp. 875–884). doi: 10.1145/2598510.2598538

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, *126*, 89–104. doi: 10.1016/j.compedu.2018.07.001

Jiang, Y., Schlagwein, D., & Benatallah, B. (2018). A review of crowdsourcing for education: State of the art of literature and practice. *PACIS 2018 Proceedings* (p. 180). Retrieved from https://aisel.aisnet.org/pacis2018/180

Jin, L., & Deifell, E. (2013). Foreign language learners' use and perception of online dictionaries: A survey study. *Journal of Online Learning and Teaching*, *9*(4), 515–533.

Kansal, A. K., Gautam, J., Chintalapudi, N., Jain, S., & Battineni, G. (2021). Google Trend Analysis and Paradigm Shift of Online Education Platforms during the COVID-19 Pandemic. *Infectious Disease Reports*, *13*, 418–428. doi: 10.3390/idr13020040

King, A. J., Gehl, R. W., Grossman, D., & Jensen, J. D. (2013). Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches. *Cancer Epidemiology*, *37*(6), 979–984. doi: 10.1016/j.canep.2013.09.004.

Köse, T., Çimen, E., & Mede, E. (2016). Perceptions of EFL learners about using an online tool for vocabulary learning in EFL classrooms: A Pilot Project in Turkey. *Procedia-Social and Behavioural Sciences, 232*(4), 362–372. doi: 10.1016/j.sbspro.2016.10.051

Krajka, J. (2021). Teaching Grammar and Vocabulary in COVID-19 Times: Approaches Used in Online Teaching in Polish Schools during a Pandemic. *JALT CALL Journal*, *17*(2), 112–134. doi: 10.29140/jaltcall.v17n2.379

Krishnan, I. A., Ching, H. S., Ramalingam, S., Maruthai, E., Kandasamy, P., De Mello, G., Munian, S., & Ling, W. W. (2020). Challenges of learning English in 21st century: Online vs. Traditional during covid-19. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *5*(9), 1–15. doi: 10.47405/mjssh.v5i9.494

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. doi: 10.1111/j.1467-9922.2010.00621.x

Li, L., & Xu, H. (2015). Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners. *Lexikos, 25*, 191–209. doi: 10.5788/25-1-1295

Lyding, V., Nicolas, L., Bédi, B., & Fort, K. (2018). Introducing the European network for combining language learning and crowdsourcing techniques (enetcollect). In P. Taalas, J. Jalkanen, L. Bradley & S. Thouesny (Eds.), *Future-proof CALL: language learning as exploration and encounters (Short papers from EUROCALL 2018)* (pp. 176–181). Research-Publishing.net.

McNeil, L. (2020). Implementing digital game-enhanced pedagogy: Supportive and impeding language awareness and discourse participation phenomena. *ReCALL*, *32*(1), 106–124. doi: 10.1017/S095834401900017X

Mendez, J. P., Curry, J., Mwavita, M., Kennedy, K., Weinland, K., & Bainbridge, K. (2009). To friend or not to friend: Academic interaction on Facebook. *International Journal of Instructional Technology & Distance Learning, 6*(9), 33–47.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE Publications.

Miloshevska, L., Gajek, E., Delibegović Džanić, N., & Hatipoğlu, Ç. (2020). Emergency online learning during the first Covid-19 period: Students' perspectives from Bosnia and Herzegovina, North Macedonia, Poland and Turkey. *Explorations in English Language and Linguistics (ExELL)*, *8*(2), 101–143. doi: 10.2478/exell-2021-0002

Miloshevska, L., Delibegović Džanić N., Hatipoğlu Ç., & Gajek E. (2021). Crowdsourcing for language learning in Turkey, Bosnia and Herzegovina, Republic of North Macedonia and Poland. *Journal of Narrative and Language Studies*, *9*(16), 106–121. Retrieved from https://nalans.com/index.php/nalans/article/view/391

Mospan, N. (2018). Mobile teaching and learning English–A multinational perspective. *Teaching English with Technology*, *18*(3), 105–125. http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-263522eb-dbc5-4ee6-b561-519e55b8ca57

Nadhifah, U. N., & Puspitasari, D. (2021). Learning English Through Duolingo: Narrating Students' Experience During Covid-19 Pandemic Time. *Ethical Lingua: Journal of Language Teaching and Literature*, *8*(1), 302–310. doi: 10.30605/25409190.280

Odo, D. M. (2016). Crowdsourced Language Learning: Lessons for TESOL from Online Language-Learning Enthusiasts. *English Teaching Forum*, *54*(4). 14–23. Retrieved from https://files.eric.ed.gov/fulltext/EJ1123197.pdf.

Özer, M. (2020). Educational policy actions by the Ministry of National Education in the times of COVID-19 pandemic in Turkey. *Kastamonu Eğitim Dergisi [Kastamonu Education Journal]*, *28*(3), 1124–1129. doi: 10.24106/kefdergi.722280

Pellegrino, J. W., & Hilton, M. L. (Eds). (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the Twenty-First Century*. National Academies Press.

Rafiee, M., & Abbasian-Naghneh, S. (2019). E-learning: development of a model to assess the acceptance and readiness of technology among language learners. *Computer Assisted Language Learning*. doi: 10.1080/09588221.2019.1640255

Reimers, F., Schleicher, A., Saavedra, J., & Tuominen, S. (2020). Supporting the continuation of teaching and learning during the COVID-19 Pandemic. *OECD*, *1*(1), 1–38. Retrieved from https://globaled.gse.harvard.edu/files/geii/files/supporting_the_continuation_of_teaching.pdf

Rodosthenous, C., Lyding, V., König, A., Horbacauskiene, J., Katinskaia, A., Ul Hassan, U., Isaak, N., Sangati, F., & Nicolas, L. (2019). Designing a prototype architecture for crowdsourcing language resources. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/313258/paper4.pdf?sequence=1

Rundell, M. (2014). Macmillan English Dictionary: The End Of Print? *Slovenščina 2.0, 2*(2), 1–14. Retrieved from http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_02.pdf

See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., & Comber, A. (2014). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*. doi: 10.1016/j.isprsjprs.2014.06.016.

Setyowati, L., Sukmawan, S., & El-Sulukkiyah, A. A. (2021). Learning from home during pandemic: A blended learning for reading to write activity in EFL setting. *JEES (Journal of English Educators Society)*, *6*(1), 9–17. doi: 10.21070/jees.v6i1.662

Shaikh, U. U., Karim, S., & Asif, Z. (2017). Re-Thinking Vygotsky: Applying Social Constructivism to Asynchronous Online Courses utilising the Power of Crowdsourcing. In *Proceedings of the 21st Pacific Asia Conference on Information Systems* (p. 233). Langkawi. http://aisel.aisnet.org/pacis2017/233

Sokolik, M. (2003). Writing. In D. Nunan (Ed.), *Practical English language teaching (PELT)* (pp. 87–88). New York: McGraw Hill.

Solemon, B., Ariffin, I., Din, M. M., & Anwar, R. M. (2013). A review of the uses of crowdsourcing in higher education. *International Journal of Asian Social Science, 3*(9), 2066-2073. Retrieved from https://archive.aessweb.com/index.php/5007/article/view/2564

Soleymani, M., & Larson, M. (2013). Crowdsourcing for multimedia research (pp. 1111–1112). ACM Press, New York. doi: 10.1145/2502081.2502234.

Thorne, S. L., & Reinhardt, J. (2008). “Bridging activities”, new media literacies, and advanced foreign language proficiency. *CALICO Journal, 25*(3), 558–572. doi: 10.1558/cj.v25i3.558-572

Toy, F. (2019). *The Effects of Quizlet on Students’ and EFL Teachers’ Perceptions on Vocabulary Learning/Teaching Process*. MA Thesis, Süleyman Demirel University, Turkey.

Trinh, T. L. A., Tran, T. K. N., Vo, T. B. N., & Huynh, T. T. S. (2021). The difference effects of paper dictionaries vs. online dictionaries. *AsiaCALL Online Journal*, *12*(3), 28–38. Retrieved from https://asiacall.info/acoj/index.php/journal/article/view/34

Trung, T., Hoang, A. D., Nguyen, T. T., Dinh, V. H., Nguyen, Y. C., & Pham, H. H. (2020). Dataset of Vietnamese student’s learning habits during COVID-19. *Data in Brief*, *30*, 105682, 1–7. doi: 10.1016/j.dib.2020.105682

Tsai, Shu-Chiao. (2019). Using google translate in EFL drafts: a preliminary investigation, *Computer Assisted Language Learning,* 32(5/6), 510–526. doi: 10.1080/09588221.2018.1527361

Tureng Online Dictionary. Retrieved from https://tureng.com/tr/turkce-ingilizce

Uluçay, Ç., & Hatipoğlu, Ç. (2017). Cause markers in Turkish cause paragraphs. In Çiler Hatipoğlu, Erdem Akbaş & Yasemin Bayyurt (Eds.), *Metadiscourse across Genres: Uncovering Textual and Interactional Aspects of Texts* (pp. 223–249). Frankfurt: Peter Lang.

Waicekawsky, L., Laurenti, L., & Yuvero, F. (2020). Teaching ESP online during the COVID-19 pandemic: An account of Argentinian students on this teaching modality. In *SHS Web of Conferences,* (Vol. 88, p. 02002). EDP Sciences. doi: 10.1051/shsconf/20208802002

Wang, L. (2016). Employing Wikibook Project in a Linguistics Course to Promote Peer Teaching and Learning. *Education and Information Technologies, 21*(2), 453–470. doi: 10.1007/s10639-014-9332-x

Weld, D. S., Adar, E., Chilton, L. B., Hoffmann, R., Horvitz, E., Koch, M., ..., & Mausam, M. (2012, July). Personalised Online Education-A Crowdsourcing Challenge. In *HCOMP@ AAAI*. Retrieved from https://cond.org/hcomp12.pdf

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Xu, H., Wu, Y., & Hamari, J. (2022). What determines the successfulness of a crowdsourcing campaign: A study on the relationships between indicators of trustworthiness, popularity, and success. *Journal of Business Research*, *139*, 484–495. doi: 10.1016/j.jbusres.2021.09.032

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.

Množičenje ter navade in prakse učenja jezikov v Turčiji, Bosni in Hercegovini, Republiki Severni Makedoniji in na Poljskem v predpandemskem in pandemskem obdobju

Priljubljenost spletnih množičenjskih platform za poučevanje in učenje jezikov je pred pandemijo COVID-19 počasi naraščala. Študije, izvedene v Turčiji, Bosni in Hercegovini, Republiki Severni Makedoniji in na Poljskem, so pokazale, da so jih učitelji uporabljali tako kot orodje pri pouku in izven njega. Po drugi strani pa so jih učenci uporabljali kot pomoč pri izpopolnitvi svojih spretnosti in znanja ciljnih jezikov ter da bi postali bolj avtonomni. Izobraževalni sistemi po vsem svetu ter ustaljene prakse poučevanja in učenja pa so se vendar spremenili s pandemijo covid-19. Ta raziskava si prizadeva odkriti, ali so se med pandemijo COVID-19 spremenila stališča, konteksti uporabe, frekventnost in navade učencev jezikov v Turčiji, Bosni in Hercegovini, Republiki Severni Makedoniji in na Poljskem, in če »DA«, kako.

Da bi primerjali uporabo orodij za množičenje pred in med pandemijo covid-19 pri učencih jezika v omenjenih štirih državah, smo ponovno uporabili medkulturno ustrezen vprašalnik, ki smo ga pred tem že uporabili v obdobju pred pandemijo. Zbrane podatke smo kvalitativno in kvantitativno preučili, da bi odkrili tudi najmanjša odstopanja. Postavili smo hipotezo, da so platforme za množičenje postale bolj razširjene med pandemijo zaradi znatnih sprememb, povezanih s poučevanjem in učenjem jezikov. Hipoteza je temeljila na ugotovitvah raziskave, ki je pokazala, da so bili učitelji v Turčiji, Bosni in Hercegovini, Republiki Severni Makedoniji in na Poljskem, podobno kot njihovi kolegi po svetu, prisiljeni uporabljati skoraj vsa digitalna orodja, ki so jih imeli na voljo, zlasti v kriznem obdobju selitve poučevanja na splet pomladi leta 2020. Obenem so bili učenci jezikov prisiljeni samostojno uporabljati številna orodja in platforme množičenja, da bi sledili zahtevam izobraževalnih ustanov.

Rezultati so pokazali, da je prehod z učenja v živo na spletno učenje zaradi covida-19 pomembno vplival na razvoj platform za množičenje po vsem svetu in na uporabo virov za množičenje v državah, vključenih v raziskavo. Opaziti je bilo, da se ni povečalo le število uporabljenih virov, temveč so se spremenile tudi funkcije uporabljenih platform (tj. od bolj splošnih k bolj »prilagojenim potrebam in državam»). Rezultati so tudi pokazali, da je preplet dejavnikov, kot so sprememba načina poučevanja, manjša interakcija z učitelji in vrstniki, večja delovna obremenitev in pomanjkanje stalne podpore s strani izobraževalnih ustanov, privedel do tega, da so učenci sami prevzeli odgovornost za svoje učenje. Spoznali so, da so edini, ki vedo, kaj potrebujejo, in da so edini, ki si lahko pomagajo, zato so začeli iskati in uporabljati platforme, ki so najbolj ustrezale njihovim zahtevam. Ker so bila pričakovanja in potrebe učencev v preučevanih državah različna, so se število, pogostost in lastnosti priljubljenih platform od države do države spreminjali.

Upamo, da bodo izsledki raziskave služili kot morebitne smernice za učence in učitelje jezikov, ki nameravajo v svoje dejavnosti v razredu in zunaj njega vključiti dejavnosti množičenja. Rezultati bi obenem lahko predstavljali pomembne povratne informacije za ustvarjalce platform, ki si prizadevajo oblikovati vire, ki so medkulturno ustrezni, hkrati pa izpolnjujejo bolj posebne zahteve učencev jezikov v specifičnih okoliščinah. Menimo, da lahko vključitev množičenja v jezikovne učne načrte učencem omogoči več priložnosti za sodelovanje v razredu in zunaj njega ter učinkovitejše učenje jezika, kar bo posledično privedlo do razvoja bolj samostojnih, aktivnih in samozavestnih učencev jezika.

Ključne besede: množičenje, učenje jezikov, COVID-19, obdobje pred pandemijo, obdobje po pandemiji

Appendix A: Period 1: Crowdsourcing sites/tools used to learn foreign languages

	Crowdsourcing tools	TUR		B&H		RNM		POL		ALL	
		%	n	%	n	%	n	%	n	%	n
1	Wikipedia	37	86	40	58.0	29	70.7	52	89.7	158	74.9
2	Kahoot	36	83.7	31	44.9	12	29.3	54	93.1	133	63.0
3	Duolingo	23	53.5	40	58.0	20	48.8	47	81.0	130	61.6
4	Khan Academy	9	20.9	8	11.6	23	56.1	9	15.5	49	23.2
5	Memrise	10	23.3	7	10.1	3	7.3	23	39.7	43	20.4
6	Busuu	9	20.9	3	4.3	2	4.9	7	12.1	21	10.0
7	Quizlet							19	32.8	19	9.0
8	Storybird	4	9.3	8	11.6	2	4.9			14	6.6
9	Writeandimprove.com	1	2.3	5	7.2	2	4.9			8	3.8
10	Anki							6	10.3	6	2.8
11	Speakandimprove.com	1	2.3							1	0.5
12	Grammarly	1	2.3							1	0.5
13	Movies and books	1	2.3							1	0.5
14	Rosetta Stone	1	2.3							1	0.5
15	Voscreen	1	2.3							1	0.5
16	Insta.ling							1	1.7	1	0.5
17	Wordreference							1	1.7	1	0.5
18	Fiszkoteka							1	1.7	1	0.5
19	Lingo Hut							1	1.7	1	0.5
20	Kanji Study							1	1.7	1	0.5
21	Tandem language app							1	1.7	1	0.5
22	Flocabulary			1	1.4					1	0.5
23	Drops			1	1.4					1	0.5
24	English Club TV			1	1.4					1	0.5
25	Google translate			1	1.4					1	0.5
26	YouTube			1	1.4					1	0.5
27	None of them	0	0	10	14.5	4	9.8	0	0.0	14	6.6

Appendix B: Period 2: Crowdsourcing sites/tools used to learn foreign languages in alphabetical order

	Crowdsourcing tools	TUR		B&H		RNM		POL		ALL	
		n	%	n	%	n	%	n	%	n	%
1	activelylearn.com	1	**1.7**							1	**0.5**
2	Anki							2	**6.7**	2	**0.9**
3	BAB.LA					1	**2.0**			1	**0.5**
4	Babble	1	**1.7**							1	**0.5**
5	BBC Learning English	1	**1.7**			1	**2.0**	1	**3.3**	3	**1.4**
6	blogs					1	**2.0**			1	**0.5**
7	Books					4	**8.2**			4	**1.9**
8	Bottled	1	**1.7**							1	**0.5**
9	Busuu	2	**3.3**							2	**0.9**
10	Cambridge (Online) Dictionary	13	**21.7**	2	**4.3**					15	**7.1**
11	Conjugato							1	**3.3**	1	**0.5**
12	Coursera					1	**2.0**			1	**0.5**
13	diki							6	**20.0**	6	**2.8**
14	Discord							1	**3.3**	1	**0.5**
15	Duolingo	26	**43.3**	16	**34.8**	19	**38.8**	14	**46.7**	75	**35.5**
16	DW Deutsch lernen	2	**3.3**							2	**0.9**
17	Easy Languages	1	**1.7**							1	**0.5**
18	EDX	1	**1.7**							1	**0.5**
19	Eng Vid			1	**2.2**					1	**0.5**
20	English idioms and phrases			1	**2.2**					1	**0.5**
21	eTutor							1	**3.3**	1	**0.5**
22	Fandom	2	**3.3**							2	**0.9**
23	Fiszkoteka							1	**3.3**	1	**0.5**
24	Flocabulary					1	**2.0**			1	**0.5**
25	Forums	1	**1.7**							1	**0.5**
26	francaisfacile	1	**1.7**							1	**0.5**
27	FunEasyLearn			1	**2.2**					1	**0.5**
28	Gamepedia	1	**1.7**	1	**2.2**					2	**0.9**
29	Glosbe			5	**10.9**			1	**3.3**	6	**2.8**
30	Google translate	8	**13.3**	16	**34.8**	16	**32.7**			40	**19.0**
31	Grammarly	5	**8.3**			3	**6.1**	1	**3.3**	9	**4.3**
32	Hello talk	1	**1.7**							1	**0.5**

	Crowdsourcing tools	TUR		B&H		RNM		POL		ALL	
		n	%	n	%	n	%	n	%	n	%
33	isl collective.com	1	1.7							1	0.5
34	Kahoot	15	25.0	8	17.4	3	6.1	6	20.0	32	15.2
35	Khan Academy	4	6.7			1	2.0			5	2.4
36	Learn Spanish	1	1.7							1	0.5
37	Lingodeer	1	1.7							1	0.5
38	Lingua	1	1.7							1	0.5
39	Linguee							1	3.3	1	0.5
40	Lingvist					1	2.0			1	0.5
41	Longman Dictionary	1	1.7	1	2.2					2	0.9
42	Memrise	1	1.7			2	4.1			3	1.4
43	Mentimeter	1	1.7							1	0.5
44	Minecraft			1	2.2	1	2.0			2	0.9
45	mondly							1	3.3	1	0.5
46	Movies					7	14.3			7	3.3
47	Nearpod	1	1.7							1	0.5
48	Netflix (e.g., series like History 101, Explained, movies, etc.)	7	11.7			2	4.1			9	4.3
49	News websites	1	1.7							1	0.5
50	None of them	1	1.7	3	6.5					4	1.9
51	One Look Thesaurus (on-line)	2	3.3							2	0.9
52	Online dictionaries	4	6.7	12	26.1			9	30.0	25	11.8
53	Online games	1	1.7	5	10.9	11	22.4	2	6.7	19	9.0
54	Oxford Online Dictionary	3	5.0	4	8.7					7	3.3
55	Ozdic	2	3.3							2	0.9
56	Plotagon	1	1.7							1	0.5
57	Podcasts	2	3.3							2	0.9
58	Pons							2	6.7	2	0.9
59	Quillbot (paraphrasing tool)					1	2.0			1	0.5
60	Quizizz	1	1.7							1	0.5
61	Quizlet	3	5.0			2	4.1	21	70.0	26	12.3
62	Reading power	1	1.7							1	0.5
63	Reddit.com	1	1.7	1	2.2					2	0.9
64	Relatedwords.org	1	1.7							1	0.5
65	ReversoContext							6	20.0	6	2.8

Crowdsourcing tools		TUR		B&H		RNM		POL		ALL	
		n	%	n	%	n	%	n	%	n	%
66	Rosetta Stone	1	**1.7**			1	**2.0**			2	**0.9**
67	Scrabble			3	**6.5**					3	**1.4**
68	Slowly (Twitter app)	1	**1.7**							1	**0.5**
69	SpanishDict					1	**2.0**			1	**0.5**
70	Spike Notes	1	**1.7**							1	**0.5**
71	Steam language games	1	**1.7**							1	**0.5**
72	Teamspeak							1	**3.3**	1	**0.5**
73	Ted Talks	1	**1.7**	1	**2.2**					2	**0.9**
74	Test English	1	**1.7**							1	**0.5**
75	TheFreeDictionary	1	**1.7**							1	**0.5**
76	Tureng (Online Dictionary)	7	**11.7**							7	**3.3**
77	Turnitin (Feedback)	1	**1.7**							1	**0.5**
78	TV5monde	2	**3.3**							2	**0.9**
79	Twitch.tv					1	**2.0**			1	**0.5**
80	Udemy	2	**3.3**							2	**0.9**
81	Uncharted	1	**1.7**							1	**0.5**
82	Urban Dictionary					1	**2.0**			1	**0.5**
83	Video games	2	**3.3**			2	**4.1**			4	**1.9**
84	VOA (Voice of America)	1	**1.7**							1	**0.5**
85	Vocaroo							1	**3.3**	1	**0.5**
86	Websites			5	**10.9**					5	**2.4**
87	Wikipedia	5	**8.3**	10	**21.7**	15	**30.6**	2	**6.7**	32	**15.2**
88	Word reference							1	**3.3**	1	**0.5**
89	Word search			1	**2.2**					1	**0.5**
90	Wordwall	2	**3.3**							2	**0.9**
91	Younglish	1	**1.7**							1	**0.5**
92	YouTube	14	**23.3**	9	**19.6**	6	**12.2**	2	**6.7**	31	**14.7**
93	Zlibrary			1	**2.2**					1	**0.5**
	ALL	**171**		**108**		**105**		**84**		**468**	

Application of crowdsourcing in education on the example of eTwinning: the Polish experience

Elżbieta GAJEK
University of Warsaw

While eTwinning is focused on facilitating collaboration among schools in Europe and beyond, the extensive participation of over one million teachers from 44 countries makes the program an extensive educational crowdsourcing activity. In this paper the program which structures the related pedagogical approaches and practices will be analyzed and discussed in light of the crowdsourcing principles. Teachers and students participate in the program voluntarily. All collaborative activities, material production and publication of results which take place online and emphasize language learning fulfil the characteristics of the effective use of crowdsourcing in education. Two kinds of analyses are undertaken, a global analysis of the program features and local analysis of the selected projects. The global analysis relates the crowdsourcing practices to the eTwinning activities. The local analysis is based on the outstanding projects submitted for evaluation for national awards in Poland, further exemplified by activities and reference to the public sites of the projects. The aim of the text is to show that teachers may effectively use crowdsourcing in educational practice even when not primarily focused on its application.

Keywords: crowdsourcing, eTwinning, language learning

Gajek, E.: Application of crowdsourcing in education on the example of eTwinning: the Polish experience. Slovenščina 2.0, 10(2): 184–201.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.184-201

1 Introduction

It can be assumed that the eTwinning program framework fulfils the characteristics of crowdsourcing in education, which is illustrated below in a global analysis of its components.

The participation of schools in the eTwinning program can be perceived as a crowdsourcing activity at a local level, as the following actions are performed by a crowd working online: the materials used in the eTwinning project are produced in various languages (cf. 4.3.1. p.2) by the teachers and students (cf. 4.3.1. p.1) from the participating countries; the activities are initially negotiated by teachers with a greater or lesser contribution by the learners (cf. 4.3.1. p.3); communication and collaboration take place online (cf. 4.3.1. p.3) and mainly via the tools provided by the program, but the participants are free to choose other tools available at their disposal; finally the stakeholders publish the results of their collaborative work so that other teachers and learners can make use of them (cf. 4.3.1. p.4). In addition, teachers support one another by sharing responsibilities within the project (c.f. 4.3.1. p.5). This study reports on a detailed investigation into whether the program activities fulfil the requirements of using crowdsourcing in educational contexts globally and locally.

2 Crowdsourcing practices in education

Crowdsourcing is a technology-mediated form of collecting, creating and distributing data (Howe, 2006). Howe (2009, p. 280) identifies and defines four primary categories of crowdsourcing applications: (i) crowd wisdom or collective intelligence; (ii) crowd creation or user-generated content; (iii) crowd voting; and (iv) crowdfunding.

The first category, crowd wisdom, relies on the fact that a crowd's knowledge is greater than an individual's. As such, creating an appropriate environment where the crowd can present and share their knowledge is essential. However, this is no easy task, as privacy concerns (Yu, 2016), data breaches (Edwards et al., 2016), intellectual property infringements (Bettig, 2018), and in extreme cases even life-threatening behavior in the form of cyberbullying (Aboujaoude et al., 2015) can occur. That is why a level of ethical awareness among teachers who are the main contributors to this process in an educational context (Chou and Chen, 2016) is a prerequisite for the effective introduction of the novel methodology.

Crowdsourcing has been applied in education, where it is perceived as a "supplementary educational component to enhance the traditional in-class and online activities" (Zdravkowa, 2020). This model is based on engaging enthusiastic teachers on a voluntary basis. It ensures freedom of speech and actions limited only by ethical rules (Zdravkova, 2020). It can be defined as "a type of an (online) activity in which an educator or an educational organization proposes to a group of individuals via a flexible open call to directly help learning or teaching" (Jiang et al. 2018, p. 3). According to Jiang et al. (2018), crowdsourcing can be beneficial for education in four different ways: "creating educational contents (Resources), providing practical experience (Activities), exchanging complementary knowledge (Support), and augmenting abundant feedbacks (Evaluation)" (2018, p. 4). Zdravkova (2020, not paginated), on the other hand, emphasizes the role of a digital platform where "teachers are able to create and observe the pedagogical content, which is consistent to [*sic*] effective learning and teaching methodologies, and facilitates the objective and efficient assessment. Entitled by the increasing content, teachers will themselves become content consumers, as well."

Incorporating crowdsourced materials in ELT classes requires careful planning and detailed instructions for accessing online resources, as a certain level of computer and media literacy is required from both teachers and students. Jiang et al. (2018, p. 10) also suggest that "crowdsourcing educational contents collaboratively among online crowd requires time and effort in coordinating the writing and review process to ensure that the end-product is beneficial to learners and maintained moving forward" (Skaržauskaitė, 2012; Weld et al., 2012). Although this process requires a great deal of enthusiasm from both teachers and learners, it is believed that the benefits certainly outweigh some potential drawbacks. Skaržauskaitė (2012, p. 74) concludes that "crowdsourcing gives students real world experience in coming up with creative solutions to important problems. Students can apply classroom knowledge to real world problems and learn the ins and outs of their chosen fields from a practical perspective." Such learner contributions to the active construction of knowledge are a major feature of constructivism (McLeod, 2019; Zdravkova, 2020). However, to ensure learner safety and protection of their data, the participants need to be acquainted with privacy and data protection rules (Johnstone and Soares, 2014; Zdravkova, 2020).

Teachers' participation in crowd-oriented education thus requires responsibility and readiness for continuous professional development (Zdravkova, 2020), with support for the participants being provided by outstanding and experienced teachers (Sallis, 2014).

3 Description of the eTwinning program

The eTwinning[1] program is widespread in primary and secondary education (Gajek, 2007a, 2007b, 2021). Its aim is to empower learners and teachers, and it is focused on a holistic educational practice which allows various theoretical perspectives to be taken in the analysis of its structure, procedure, activities and outcomes. It applies blended learning pedagogy (Gajek, 2021) as well as constructionism[2] (Gajek, 2017) and constructivism[3] (Gajek, 2010). It also helps to develop key competences[4] (Gajek, 2009a 2009b) with teachers communicating, collaborating and producing materials together with their learners.

The aim of the program is to facilitate collaboration among schools in 44 countries of the EU and 17 beyond outside it, including Armenia, Azerbaijan, Norway, Iceland, Moldova, Jordan, Turkey, Tunisia, Serbia, Georgia, Lebanon, Bosnia and Herzegovina, Cyprus, and North Macedonia. Teachers voluntarily register on the eTwinning portal and search for partners. Then they work out a concept for a project, which they

[1] https://school-education.ec.europa.eu/en/etwinning

[2] Constructionism, in the context of learning, is the idea that people learn effectively through making things.

[3] Constructivism is a theory that says learners construct knowledge rather than just passively take in information. Social constructivism teaches that all knowledge develops as a result of social interaction and language use, and is therefore a shared, rather than an individual, experience. Cognitive constructivism states knowledge is something that is actively constructed by learners based on their existing cognitive structures.

[4] The key competences for lifelong learning include the following: literacy competence; multilingual competence; mathematical competence and competence in science, technology and engineering; digital competence; personal, social and learning to learn competence; citizenship competence; entrepreneurship[cultural awareness and expression (2019).

develop together. Next, teachers and learners prepare and share materials, communicate and collaborate on the project. Eventually, they evaluate the project and disseminate the results of their work. However, they are free to apply elements that are not harmonized with their partners, e.g. assessment criteria or levels of knowledge or skills attainment. All participants monitor the ethical aspects of intercultural communication. Examples of good practice are open to the public and shared among other teachers.

The main pedagogical approach applied in such projects is Content and Language Integrated Learning (CLIL), which means that language and content are learned simultaneously. This ensures linguistic development in one or more languages used by the partners while they work on any subject theme. There are projects especially focused on a language which is foreign to all participants, but these constitute a small share of the total. CLIL combined with ongoing collaboration focused on the end-product create a natural environment and context for language learning, as the foreign language is most often the only means of communication among partners. It is claimed that a language is learned more effectively when it functions as a tool for achieving other purposes (Kurcz, 2011). In eTwinning languages are learned in practice, through learning by doing in clear contexts, not only as school subjects.

In this program social and organizational incentives are emphasized, although there are also financial incentives for very active teachers in the form of reimbursement for participation in conferences, workshops, and dissemination events at regional, national and international levels. In two countries, Poland and Malta, the program is recommended in the curriculum.

The presentation of data on the participation of teachers in the program is based on the information displayed on www.etwinnig.net (accessed 25.04.2022 at 10:11 am) The eTwinning community involves 1,052,832 teachers employed in 233,087 schools and working on 138,621 projects, which creates a considerable community of contributors.

The statistics which demonstrate the level of participation are taken from the Polish eTwinning National Support Service in and https://etwinning.pl/o-programie/statystyki (January 2022).

The numbers represent the number of schools (S-), number of teachers (T-) and number of projects (P-). The leading country is Turkey (S-52,205, T-287,685, P-50,795), followed by Poland (S-19,358, T-79,556, P-36,118), Italy (S-11,178, T-94,652, P-33,440), Spain (S-16,721, T-77,790, P-30,272), France (S-22,143, T-68,501, P-26,719), Romania (S-9,985, T-34,167, P-24,569), Greece (S-10,658, T-33,086, P-18,696), Portugal (S-1,881, T-21,699, P-14,860), Germany (S-10,353, T-30,587, P-13,112), Slovakia (S-2,926, T-12,403, P-10,074), Czechia (S-4,257, T-11,875, P-9,563), Lithuania (S-1,992, T-11,497, P-9,222), Croatia (S-2,813, T-17,796, P-8,219), Bulgaria (S-2,996, T-10,574, P-7,923), Ukraine (S-1,560, T-2,966, P-5,265), Latvia (S-1,153, T-7,607, P-4,585), Finland (S-2,512, T-8,801, P-4,467), Albania (S-1,551, T-5,485, P-4,426), Slovenia (S-875, T-5,419, P-4,204), Sweden (S-3,540, T-12,660, P-4,108), Belgium (S-2,450, T-9,226, P-4,025), Georgia (S-1,162, T-2,095, P-3,937), Serbia (S-1,286, T-4,161, P-3,920), Holland (S-2,534, T-9,652, P-3,657), Estonia (S-1,004, T-5,454, P-3,576), Azerbaijan (S-1,001, T-3,503, P-3,552), North Macedonia (S-493, T-2,365, P-3,473), Hungary (S-2,358, T-5,867, P-3,043), Norway (S-1,787, T-6,352, P-2,952), Denmark (S-2,066, T-9,277, P-2,840), Austria (S-1,993, T-5,915, P-2,330), Cyprus (S-643, T-3,163, P-2,226), Armenia (S-623, T-2,125, P-1,999), Moldova (S-412, T-950, P-1,985), Malta (S-296, T-4,035, P-1,981), Tunisia (S-885, T-2,189, P-1,908), Ireland (S-1,820, T-3,659, P-1,824), Bosnia & Herzegovina (S-581, T-1,863, P-1,507), Jordan (S-408, T-1,133, P-1,398), Iceland (S-320, T-1,796, P-1,102), Luxemburg (S-151, T-691, P-372), Lebanon (S-80, T-185, P-97), and Lichtenstein (S-17, T-42, P-19).

It is worth mentioning that registration does not equate with active participation, or the creation and sharing of materials. On the one hand, the program is a great success. On the other, as it involves so much scaffolding, continuous professional development opportunities, technical support, support from peer teachers (who serve as active ambassadors of the idea), and support from the National Support Services and Central Support Service, producing new ideas and new practices based on crowdsourcing in education is very challenging.

4 Methodology

4.1 Aims

The aim of this analysis is to demonstrate that the eTwinning program can be perceived as a widespread application of crowdsourcing in education. First, a comparison of features of crowd-oriented education related to eTwinning practices (Table 1) is presented. Then crowd-oriented tasks and procedures are illustrated on the basis of examples of selected projects. Finally, the extensive use of digital tools for educational purposes is presented to reflect on another perspective of eTwinning – its application for crowdsourcing.

4.2 Global analysis of eTwinning practices as crowdsourcing activity

Table 1 shows the relation between the crowd-oriented educational features in relation to the eTwinning practices presented above.

Table 1: *Characteristics of general crowd-oriented educational practices applied to the eTwinning program*

Crowdsourcing practices	eTwinning features and activities
Existence of a crowd	Over 1,000,000 teachers, together with learners and school communities in participating countries create the crowd.
Crowd wisdom and collective intelligence	Anybody involved in the educational system in their country can contribute to the development of the project ideas and their implementation regardless of the languages used, the system organization, specific curricular requirements and educational traditions.
Freedom of speech and actions	The organizational framework is open to exploration. There are no limits related to time, content, number of participants, tools, products, etc. However, some recommendations are published only to facilitate participation by the teachers.
Organization of support	Initial support is provided via open models displayed by National and Central Support Service offices, project ambassadors via webinars, training sessions, contact meetings and conferences.
Readiness for continuous professional development	Participation in the project allows for professional development of the teachers through sharing experiences, resources, multilingual negotiations and discussions.
Inclusion of practical perspectives on the chosen fields	The projects undertake topics related to the close neighborhood of the learners, or they relate globally discussed themes to the learner’s neighborhood. They ensure linguistic variety in shared activities.

Crowdsourcing practices	eTwinning features and activities
Flexible digital area where the crowd can present and share their knowledge	TwinSpace is meant to be a place for contact, communication and storage of materials. It is the main area where participants meet, collaborate, get support, store project resources and evaluate the outcomes of their projects. But the use of technology is not limited to TwinSpace.
User-generated content (Resources)	Teachers, learners, and other invited stakeholders create the content, which may relate to the curriculum or go beyond it. The materials can then be published on websites that are open to the public for use by others.
Sharing practical experience (Activities)	Teachers, and to some extent learners, develop activities collaboratively and decide on the project outcomes, forms of evaluation and its dissemination.
Exchanging knowledge (Support)	Within the project participants share knowledge, experiences and resources at a local level, according to their specific needs and contexts.
Providing feedback (Evaluation)	At the end, each project is evaluated by stakeholders: teachers, learners, and others, such as parents via questionnaires and interviews.
Compliance with in-class teaching	The content and theme of the project need to comply with the curriculum, or go beyond it if there is a need for such actions.
Focus on ethical issues in education, GDPR and privacy policies	Clear guidelines on safety and etiquette are provided on the eTwinning portal.

It is worth noticing that such features of crowdsourcing as crowdfunding, crowdvoting, crowdshipping, and crowdsolving, etc. are not used in the eTwinning program.

To sum up, the features of eTwinning related to elements which characterize crowdsourcing clearly show how they are interrelated globally, and how crowdsourcing is applied in this program. They are also in line with theoretical approaches to the educational use of crowd-oriented activities.

4.3 Analysis of selected projects

4.3.1 Criteria of analysis of the projects

Analysis of local projects as crowdsourcing activities provides insights into how the general ideas are translated into in-class practices. As there are thousands of projects undertaken among European schools, it was necessary to specify criteria for the selection of projects for this study. These criteria were as follows:

1. the activities and materials had been designed and prepared by teachers and students together;
2. the language learning tasks involved various languages: native and foreign depending on the needs of the participants;
3. digital technology was used widely for communication, development collaboration and publishing of the materials;
4. the results of the project were published for further use by other teachers and learners;
5. teachers' workload was shared, thus enhancing professional development.

Although assessment is a very important part of education, it is not included in the analysis as it is not performed as a crowdsourcing activity in the eTwinning projects.

The examples of projects selected for analysis are from those submitted for the Polish national contest titled "Our Project 2022". They cover all age categories, from kindergarten to secondary education, to provide a variety of techniques used in eTwinning projects. All the materials and results of the analyzed projects are published on the internet and comply with the copyright and safety rules as all the participants (teachers and students through their legal representatives – that is, parents) agreed to publish the photographs and outcomes of their work. All of them are focused on the learning outcomes related to both the content and language or languages used in the projects. The projects for which resources were shown exclusively to adjudicators of the contest were not considered for analysis.

4.3.2 Identification of crowdsourcing practices in the projects

The four examples analyzed below fulfil the abovementioned criteria.

A Project for Children 3 to 6-Year-Old Learners

Natural Pharmacy is a project which focuses the young learners' attention on various treasures of the natural world, and the role of nature in our lives. Children learn about healing herbs by exploring their neighborhoods.

The partners in the project were from Slovakia, Turkey, Greek, Slovenia, Czechia, Spain, Estonia and Poland, with the results are presented at https://twinspace.etwinning.net/121166/home.

Ad 1. Projects with the youngest group of children require teachers to communicate and involve learners in those activities which they are capable of doing. However, children are very effective in the production of materials if they are properly supported.

Ad 2. English was the main language of communication and collaboration. The native languages of the participants were used but rarely documented in writing, as such young children do not possess the literacy skills needed to communicate in writing. More often the children were filmed while speaking their native or foreign languages.

Ad 3. Digital tools were used for all communication, activities, the production of materials and reporting, but TwinSpace remained the main area of collaboration.

Ad 4. The project was extensively presented on its public TwinSpace, mainly through photographs. The teachers got the parents' consent to publish the children's faces on the internet.

Ad 5. Teachers shared the responsibility to ensure the flow of work was done in parallel at all partner institutions.

A Project for 7 to 10-Year-Old Learners

Steamist is a very innovative and creative project about biomimicry. Teachers and students from Turkey, Poland, Portugal, Ukraine, Romania, Moldova, North Macedonia, Lithuania, Tunisia, Albania and Italy worked together on holistic topics which integrate themes from various school subjects, such as natural science, maths, languages, art, and digital technology. The students observed nature and found relations with various technological solutions that people use in their everyday lives, i.e. biomimicry. Here the public TwinSpace documents the work and products of the project: https://twinspace.etwinning.net/118413/home and https://steamist20.blogspot.com/.

Ad 1. Teachers and students communicated online synchronously and asynchronously to share their work on the materials and activities.

Ad 2. English was the main means of communication and collaboration, but the native languages of the participants were also used in parallel to explain the work in school communities, as well as for presentation and dissemination of the results.

Ad 3. Digital tools were used for all communication, activities, the production of materials and reporting, but TwinSpace remained the main area of collaboration. The results are presented on a blog.

Ad 4. Various materials such as video clips https://youtu.be/nVFPRyXt_Uk?t=18, a summary of the project https://steamist20.blogspot.com/ and many photographs can be used by other educators.

Ad 5. Teachers clearly stated that they learned a lot during the projects themselves, as biomimicry is not a popular topic in the curricula.

A Project for 11 to 15-Year-Old Learners

My Water Footprint is a project about the role of water in both nature and the human world. Its public website shows the partners, results, and activities, as well as illustrates the work involved in the project: https://twinspace.etwinning.net/120761/home. The participants come from 13 countries: Turkey, Croatia, Poland, Italy, Spain, Portugal, Ukraine, Belgium, France, Lithuania, Romania, Greece, and Bosnia and Herzegovina.

Ad 1. Teachers and students communicated online to share their work on the materials and activities.

Ad 2. Project members used English as the main language of communication and collaboration, but their native languages were also used to share their work in school communities, for presentation and dissemination of the results.

Ad 3. Digital tools were used for all communication, activities, the production of materials and reporting, but TwinSpace remained the main area of collaboration.

Ad 4. Various materials such as video clips https://youtu.be/Ar6Td-HqKf8?t=19, dynamic presentations[5] and reports document the

[5] https://docs.google.com/presentation/d/e/2PACX-1vSsSnKLdRJswX-VZk5FhyvV6fSP-b3x7x8O7qORu2Inzk-j8PfIK5kQInUUHSA6naSp8vasRjVJvaeAC/pub?start=false&loop=-false&delayms=3000#slide=id.gd864009de4_0_5

results of the project work. They are ready for further use by other stakeholders.

Ad 5. Teachers shared the responsibility to ensure the flow of the work.

A Project for 16 to 19-Year-Old Learners

The theme and title of the project, Wonderful Journey, refers to the importance of railway transport. The participants collaboratively prepared various materials, such as maps, posters, journeys itineraries, virtual trips, games and quizzes. Some of these products can be seen online at https://twinspace.etwinning.net/163327.

Ad 1. Teachers and students from Portugal, Italy, Czechia, Serbia, Finland, Greece, Turkey and Poland communicated online to share their work on the materials and activities.

Ad 2. They used English as a means of communication and collaboration, but the native languages were also used for local purposes, such as presentation and dissemination of the results.

Ad 3. Digital tools were used for all communication, activities, the production of materials and reporting, but TwinSpace remained the main area of collaboration.

Ad 4. The booklet[6] was one of the final results of the project, and shows the various perspectives that were applied, with focuses on ecology, economy, safety, comfort, and sustainability.

Ad 5. Teachers shared the responsibility to ensure the flow of the work, which is well demonstrated on the project website.

4.3.3 Discussion of the analysis of projects

The projects described briefly above only illustrate some of the activities undertaken in all good projects, and thus show features characteristic of many eTwinning projects. What is more, one of the criteria of selection was the presentation of the resulting content on TwinSpace, which is thus made available to the wider public. It has not been possible to provide detailed data about closed projects that are only available to the adjudicators of the Polish contest, as this would be contrary

[6] https://read.bookcreator.com/b3Rh2kc4eQUQFHXQNszpWWTwpsq2/hqSYLpGJQ8iZpS-3b6EhXjQ

to ethics and safety regulations. That is why the general discussion below not only presents an interpretation of the processes and activities observed in the projects outlined above, but also gives more holistic insights into the program as a crowdsourcing phenomenon.

Even such a limited presentation and analysis of project activities shows that they have a lot in common. In all projects both teachers and learners work together to prepare materials and tasks. However, their level of contribution depends on the age of learners, their computer literacy and ability to take part in making videos, and to participate in discussions independently.

In most of the projects subject to analysis and beyond, the main language of communication is English, and to a lesser extent German and French. In the case of Polish, it is the language of communication between schools in Poland and Polish schools abroad, such as in Latvia, Lithuania, and Ukraine, where a Polish community has been present for many years, as well as between schools hosting children from the recent waves of emigration, such in Great Britain. In all projects, language learning can be seen as one dimension of holistic learning which takes place during the related activities.

In all projects TwinSpace is the main digital platform for collaboration and communication in foreign languages. This is a place where partners are searched for and found, materials stored, ongoing reports written, oral and audio-visual communication is performed, project activities documented, and evaluation and dissemination applied. However, TwinSpace is not the only digital tool used in the projects, as it is shown below (cf. Table 2).

Open sharing of the materials and access to activities depends on the teachers involved. They need to compromise between ensuring the safety of their students, keeping parents' consent and the desire to present new educational practices and models in action to inspire other teachers. The open results, however, are usually prepared in such a way that they do not break the ethical regulations stipulated by the program.

A final but important point to note here is that the teachers ensure their own professional development through such collaboration. Teachers constantly improve their linguistic skills in natural interactions with other members of the projects, which is of special importance for teachers working with learners at lower proficiency levels. They develop intercultural competences while learning about other school systems, curricula and educational traditions, as well as how to adjust the content of their projects to various time and curricular constraints. They share responsibilities in a collaborative way, which is rarely observed in class as the European tradition of separating school subjects form one another rarely allows for intensive holistic collaboration across the curriculum. Teachers also get acquainted with various technological tools and find reasons for using them in their practice, finding support and advice if needed from others.

All in all, the eTwinning project has the characteristics of crowdsourcing applied to education in local, in-class, contexts with regard to the criteria that were applied: creating content, gaining practical experience, and sharing knowledge.

4.4 Digital technology in projects

The use of various digital tools available for participants also represent the preferences and purposeful actions taken by the crowd of teachers within the eTwinning program. Table 2 shows the technology, free of charge or commercial, applied in projects and used to achieve specific educational goals at various stages.

Table 2: *Digital tools for various purposes (Upgraded source: Gajek, 2021)*

Project activities	Examples of digital tools
Project posters and logos	Canva, Logomaker, Poster Maker, RedenForest, Postermywall, Tagull
Map of partner schools	Google Maps, MapLoco, Tripline, Zeemap, Pictramap
Timeline	Flipidity, Powtoon
Surveys, votes, evaluations	Google Forms, Pollmaker, Tricider, Wakelet, SurveryMonkey AnswerGarden
Checklists, lists of groups	Google sheets, TwinSpace Table
Partners' presentations	Wakelet, Kizoa, Biteable, Animoto, Animaker, MsMovieMaker, RedenForest, Phrase.it, Voki, VoiceThread, BlabbeRize, AvatarMaker, Avachara PhotoTalks, FactoryForAvatars, TellaGami, YouTube, Quik, Canva, Dotstorming, Chatterkid, Pixton

Project activities	Examples of digital tools
Schools' presentations	Emaze, Prezi, Kizoa, Visme Biteable, GoogleSlides, Piktochart, Animaker, MsMovieMaker, RedenForest, GoProQuik, YouTube, Knovio, Voki, Create Avatar, Scoompa Video
Making and editing videos	Scoompa, InShot, Flexiclip, Voki, Create Avatar, Animoto, Kioza, Animaker, Biteable, Flixpress, VivaVideo, TellaGami, Filmora, VivaVideo
Discussions and communication	TwinSpace Journal, Facebook Messenger, Twinspace Forums, Messenger Chat Groups, WhatsApp, Edmodo, Skype, Adobe Connect, Hang Outs, Zoom, Flipgrid, Doodle, Padlet
Taking notes and making mindmaps	Evernote, Mindmeister, Coggle
Cartoons and avatars	Avachara, Voki, Create Avatar, Avatar Maker, FactoryForAvatars, Toontastic, Phrase.it, PhotoTalks
Infographics	Pictochart, Genial.ly, Canva
Project blog	Blogspot, Blogger
Activity photos, photo editing	Padlet, Adobe, Spark Joomag, Canva, Story Jumper, Madmagz, Paint, Befunky, inCollage, Pixlr, Ipiccy, InShot, Tuerchen, Collage-maker,[7] Pixiz, Pizap, PhotoCollage
3D design	Thinkercad
Collaborative area	Jamboard, Padlet, Sway, Pearltrees, VoiceThread
Exchanging instructions	QR code tools
Comparing opinions	Mentimeter
Collages	BeFunky, pixi, Piccollage[8]
Classroom applications	ClassDojo, Google Classroom, ThingLink
Recording and editing audio	Audacity, Vocaroo
Storyboard creators	StoryboardThat
Games	LearningApps, Bookwidgets,[9] JigsawPlanet, Gimkit, Quizziz, Quizlet, Actionbound, Crosswordlabs.Com, Kahoot, Poll everywhere, Cram
Project websites	Weebly
Statistics	IBM Statistics 20, Surveymonkey, jamboard
Final products and publishing	Blogspot, Quik, Powtoon, Moviemaker, Bookcreator, Smore, Google Slides, Calameo, Genial.ly, Wordart, Storyjumper, YouTube, issuu, Flippity, Wordart, Prezzi, Yumpu

The list of digital tools used for educational purposes in eTwinning projects is extensive and constantly changing, as such tools frequently

[7] https://collage-maker.com/
[8] https://piccollage.com/
[9] https://www.bookwidgets.com/

appear and disappear. It is thus likely that crowd wisdom and collective intelligence are the best ways for teachers to get oriented as to the variety of products and platforms that currently available.

5 Conclusion

To conclude, the size of the eTwinning community and the three dimensions of the program investigated in this study, i.e., the global and local activities undertaken, as well as the various digital tools used for educational purposes, demonstrate that eTwinning is a valid example of crowdsourcing applied in education. Among others results, the linguistic outcomes and progress observed in both learners and teachers establish the added value of using crowdsourcing in this context. However, both the founders and participants of the program are barely aware of this fact, as in general teachers do not know much about how crowdsourcing relates to their work (Arhar Holdt et al., 2020). This shows that the cultural trend that encourages shared collaboration online, which is described under the umbrella term of crowdsourcing in education, may be embodied in various organizations and attract various stakeholders who appear able to translate the general into local. However, sometimes their vague ideas, intuitions and hopes with regard to crowdsourcing can be turned into concrete actions, and the success of these may gradually change educational practices on a large scale.

References

Aboujaoude, E., Savage, M. W., Starcevic, V., & Salame, W. O. (2015). Cyberbullying: Review of an old problem gone viral. *Journal of Adolescent Health, 57*(1), 10–18. doi: 10.1016/j.jadohealth.2015.04.011

Arhar Holdt, Š., Zviel-Girshin, R., Gajek, E., Durán-Muñoz, I., Bago, P., Fort, K., Hatipoglu, C., Kasperavičienė, R., ..., & Zanasi, L. (2020). Language Teachers and Crowdsourcing: Insights from a Cross-European Survey. *Rasprave Instituta za hrvatski jezik i jezikoslovlje, 46*(1), 1–28.

Bettig, R. V. (2018). *Copyrighting culture: The political economy of intellectual property*. Routledge.

Chou, H. L., & Chen, C. H. (2016). Beyond identifying privacy issues in e-learning settings –Implications for instructional designers. *Computers & Education, 103*(1), 124–133.

Edwards, B., Hofmeyr, S., & Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity, 2*(1), 3–14. doi: 10.1093/cybsec/tyw003

Gajek, E. (2007a). eTwinning przykładem e-learningu w oświacie. *e-mentor, 2*(19), 33–39.

Gajek. E. (2007b). eTwinning: the Polish experience. In: B. Beaven (Ed.), *IATEFL 2006, Harrogate Conference Selection*s. Canterbury: IATEFL Darvin College.

Gajek. E. (2009a). Ewaluacja kursu online: „Jak uczestniczyć w programie eTwinning?". In E. Gajek & P. Poszytek (Eds.), *eTwinning drogą do edukacji przyszłości.* Warszawa: Fundacja Rozwoju Systemu Edukacji.

Gajek, E. (2009b). Kompetencje kluczowe w projektach międzynarodowych programu eTwinning. In E. Gajek & P. Poszytek (Eds.), *eTwinning drogą do edukacji przyszłości.* Warszawa: Fundacja Rozwoju Systemu Edukacji.

Gajek. E. (2010). Social and cognitive constructivism in practice on the basis of eTwinning project in science. In Z. C. Zacharia, C. P. Constantinou & M. Papaevipidou (Eds.), *Computer Based Learning in Science* (pp. 41–47). Cyprus: University of Cyprus.

Gajek, E. (2012). Constructionism in action. European eTwinning Projects. In F. Zhang (Ed.), *Computer-Enhanced and Mobile-Assisted Language Learning: Emerging Issues and Trends* IGI-Global Australia (pp. 116–136).

Gajek, E. (2017). Curriculum integration in distance learning at primary and secondary educational levels on the example of eTwinning projects; Challenges and Future Trends of Distance Learning. *Education Sciences, 8*(1), 1–15. doi:10.3390/educsci8010001

Gajek, E. (2021). Cooperative Blended Learning and Teaching – on the Example of eTwinning. A. Palalas, C Lazou (eds.) *Blended Language Learning: Evidence-Based Trends and Applications*, 94-116. International Association for Blended Learning (IABL) & Power Learning Solutions.

Garcia, I. (2013). Learning a language for free while translating the web. Does Duolingo work? *International Journal of English Linguistics, 3*(1), 19–25. doi: 10.5539/ijel.v3n1p19

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–4.

Jiang, Y., Schlagwein, D., & Benatallah, B. (2018). A review of crowdsourcing for education: State of the art of literature and practice. *PACIS 2018 Proceedings*, 180. Retrieved from https://aisel.aisnet.org/pacis2018/180

Johnstone, S. M., & Soares, L. (2014). Principles for developing competency-based education programs, *Change: The Magazine of Higher Learning, 46*(2), 12–19.

Key Competences for Lifelong Learning. (2019). Luxembourg: Publication Office of the European Union. Retrieved from https://www.fi.uu.nl/publicaties/literatuur/2018_eu_key_competences.pdf.

Kurcz, I. (2011). Charakterystyka kompetencji językowej – reprezentacje umysłowe. In I. Kurcz & H. Okuniewska (Eds.), *Język jako przedmiot badań psychologicznych*. Psycholingwistyka ogólna i neurolingwistyka. Warszawa: Academica.

McLeod, S. (2019). Constructivism as a theory for teaching and learning. *Simply Psychology.* Retrieved from https://www.simplypsychology.org/constructivism.html.

Sallis. E. (2014). *Total quality management in education*. Routledge. Retrieved from https://herearmenia.files.wordpress.com/2011/09/ebook-sclub-org__total_quality_management_in_education.pdf

Skaržauskaitė, M. (2012). The application of crowd sourcing in educational activities. *Social technologies, 2*(1), 67–76.

Zdravkova, K. (2020). Ethical issues of crowdsourcing in education, *Journal of Responsible Technology, Vol. 2–3*. doi: 10.1016/j.jrt.2020.100004

Uporaba množičenja v izobraževanju na primeru eTwinning: izkušnje s Poljske

Projekt eTwinning je usmerjen predvsem v spodbujanje sodelovanja med šolami v Evropi in v drugih državah, obenem pa program zaradi sodelovanja več kot milijona učiteljev iz 44 držav predstavlja tudi obsežno dejavnost množičenja v izobraževanju. V članku bo projekt, ki strukturira povezane pedagoške pristope in prakse, analiziran in obravnavan z vidika načel množičenja. Učitelji in učenci v programu sodelujejo prostovoljno. Vse dejavnosti sodelovanja, priprave gradiva in objave rezultatov, ki potekajo na spletu in poudarjajo učenje jezikov, izpolnjujejo značilnosti učinkovite uporabe množičenja v izobraževanju. Opravljamo dve vrsti analiz, in sicer splošno analizo značilnosti programa in lokalno analizo izbranih projektov. Splošna analiza povezuje prakse množičenja z dejavnostmi projekta eTwinning. Lokalna analiza temelji na izstopajočih projektih, ki so bili dani v ocenjevanje za državno nagrado na Poljskem, kar je ponazorjeno tudi z dejavnostmi in sklici na javna spletna mesta projektov. Namen besedila je pokazati, da lahko učitelji učinkovito uporabljajo množičenje v izobraževanju, tudi kadar niso ciljno usmerjeni v njegovo uporabo.

Ključne besede: množičenje, eTwinning, učenje jezikov

EnetCollect – European Network for Combining Language Learning with Crowdsourcing Techniques (COST Action CA16105): a review of the project's vision, organization, progress, and achievements

Lionel NICOLAS
Eurac Research, Institute for Applied Linguistics, Bolzano

Verena LYDING
Eurac Research, Institute for Applied Linguistics, Bolzano

This article reviews the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect), an extensive network project created to foster research and innovation (R&I) on the combination of crowdsourcing and language learning. Accordingly, we explain how it began, introduce its overall logic and organization, and discuss its achievements in terms of both (1) creating a new R&I community through a concluded large network project, and (2) fostering R&I on a high-potential and mostly unexplored subject.

Nicolas, L., Lyding, V.: EnetCollect – European Network for Combining Language Learning with Crowdsourcing Techniques (COST Action CA16105): a review of the project's vision, organization, progress, and achievements. Slovenščina 2.0, 10(2): 202–225.

1.03 Drugi znanstveni članki / Other Scientific Articles

DOI: https://doi.org/10.4312/slo2.0.2022.2.202-225

We also discuss the challenges involved and lessons learned, whether in orchestrating and leading a new R&I community or the challenges we faced and generally observed in the efforts of enetCollect members, as they explored the many facets of such a versatile enterprise.

Keywords: enetCollect, COST Action, Crowdsourcing, Language Learning

1 Introduction

EnetCollect, the European Network for Combining Language Learning with Crowdsourcing Techniques, was a network project running as a COST Action funded by the COST (European Cooperation in Science and Technology) Association through the Horizon 2020 Framework Programme of the European Union. As explained on the COST website,[1] "A COST Action is an interdisciplinary research network that brings researchers and innovators together to investigate a topic of their choice for four years. COST Actions are typically made up of researchers from academia, SMEs, public institutions and other relevant organizations or interested parties." EnetCollect started in March 2017 and ended in September 2021 after an extension of six months was granted to partly remediate the challenges posed by the COVID-19 pandemic. Over its 4.5 years of existence, enetCollect involved more than 200 stakeholders from 41 different countries. It served as a starting point or catalyzer for more than 50 scientific publications, as well as several project proposal submissions and related funded research projects.

EnetCollect sought to create a new Research and Innovation (R&I) trend combining the well-established domain of language learning with recent and successful crowdsourcing approaches to leverage for all languages, in the medium- to long-term, the crowdsourcing potential of an ever-growing number of language learners and teachers. Such potential would fuel an innovation breakthrough for producing two types of cost-intensive materials: language learning materials, such as lesson or exercise content, and language-related datasets, such as Natural Language Processing (NLP) resources.

[1] https://www.cost.eu/cost-actions/what-are-cost-actions/

In Section 2, we describe the premise and existing work that enetCollect built upon and, in Section 3, detail its objectives and organization in working groups. In Section 4, we discuss the achievements that could be reached with respect to its objectives, as well as the challenges we faced and the lessons we learned from it. Finally, Section 5 concludes the article and introduces D4Collect,[2] a DARIAH Working Group created as a follow-up to enetCollect.

2 Background and related work

In terms of international groups of stakeholders exploring the combination of crowdsourcing and language learning, enetCollect was a first-of-a-kind project. Indeed, unlike many other network projects that grew out of smaller initiatives (e.g. specialized workshops or task forces), most of the stakeholders had yet to work on the subject, and only a few had interacted with one another before participating in enetCollect. The majority thus spontaneously joined because of enetCollect's appeal to their respective interests, be it in terms of language learning, crowdsourcing or the creation of language-related datasets. Likewise, they started exploring the subject almost from scratch, as very few past works were directly relevant to it.

Indeed, before 2017, only a few initiatives had combined both language learning and crowdsourcing. Notable among them is the well-known online language learning platform Duolingo (von Ahn, 2013), which originally followed an enetCollect-compatible logic to crowdsource translations and, to this day, offers some loyal users the possibility of participating in the creation of language learning material.[3] Moreover, only a limited number of research efforts have combined language learning and crowdsourcing to produce part-of-speech corpora (Sangati et al., 2015) or syntactic knowledge (Hladká et al., 2014).

Nonetheless, the individual states of the art for crowdsourcing, language-related datasets creation and language learning are extensive. While their characteristics and similarities with enetCollect's objectives fall far beyond the scope of this article, we can, however,

[2] https://www.dariah.eu/activities/working-groups/combining-language-learning-with-crowdsourcing-techniques-d4collect/

[3] As do other platforms, like Memrise (https://www.memrise.com/)

point out that they include – with respect to crowdsourcing and NLP dataset creation[4] – numerous efforts implementing approaches such as Wisdom-of-the-Crowd[5] (WoC), Human-based Computation[6] (HC) or Games-With-A-Purpose[7] (GWAP), and – with respect to language learning – an even wider number of efforts related to, among others, the various different Computer-assisted Language Learning (CALL) communities[8].

3 Objectives and organizational structure

3.1 Objectives

As previously mentioned, the overall objective of enetCollect was to kickstart a new R&I trend on the combination of language learning and crowdsourcing in order to trigger an innovation breakthrough for the production of both language learning material and language-related datasets, such as NLP resources. Integrating crowdsourcing approaches into the language-learning material-creation workflow promises to facilitate the production of even more diversified language learning materials and language-related datasets at reduced cost by outsourcing part of the cost-intensive manual work to crowds of teachers and learners (see Sections 4.2 and 4.3 for more details and references). This would contribute to addressing the two challenges of (a) fostering the language skills of all citizens in a globalizing world regardless of their diverse social, educational, and linguistic backgrounds, and (b) solving the longstanding challenge of creating extensive language-related datasets for all languages taught, not only those that receive the most financial and research support.

In order to foster this new R&I trend, enetCollect tackled several network- and research-oriented subgoals.

[4] Relevant references can be found by searching for the term “crowdsourcing” in the ACL Anthology (https://aclanthology.org/)

[5] https://en.wikipedia.org/wiki/Wisdom_of_the_crowd

[6] https://en.wikipedia.org/wiki/Human-based_computation

[7] https://en.wikipedia.org/wiki/Human-based_computation_game

[8] E.g. the European Association of Computer Assisted Language Learning (http://www.eurocall-languages.org/)

Network-oriented goals

(1) To bring together relevant stakeholders from different domains (language learning, crowdsourcing, language-related domains, especially NLP, and computer science in a wider sense) interested in exploring the combination of language learning with crowdsourcing techniques to reach their respective objectives.
(2) To establish and consolidate communication channels and dissemination procedures.
(3) To foster complementary and follow-up project building and funding acquisition.

Research-oriented goals

(4) To create a shared understanding and theoretical framework to approach the combination of language learning and crowdsourcing by revising the state of the art, analyzing directly and indirectly related approaches, and establishing a shared terminology.
(5) To research use cases, work on prototypes combining language learning and crowdsourcing and gather evaluation data.

3.2 Network structure

EnetCollect was organized around five distinct yet interconnected working groups whose efforts directly tackled the two aforementioned research-oriented goals:

- WG1, R&I on explicit crowdsourcing for language learning material production,
- WG2, R&I on implicit crowdsourcing for language learning material production,
- WG3, user-oriented design strategies for a competitive solution,
- WG4, technology-oriented specifications for a flexible and robust solution,
- WG5, application-oriented specifications for an ethical, legal and profitable solution.

Working Groups (WGs) 1 and 2 were created to tackle the core objectives of enetCollect, namely researching how crowdsourcing techniques could be applied to language learning. A practical distinction was made between works focused on *explicit* crowdsourcing (WG1) and works focused on *implicit* crowdsourcing (WG2), where *explicit* crowdsourcing refers to activities where the crowd is aware of their participation in a crowdsourcing effort and intentionally participates. In contrast, in *implicit* crowdsourcing activities the crowd is not necessarily aware of their contribution to a crowdsourcing effort, or the act of contributing is not the primary motivation for their participation. Such a distinction was made pragmatically, as we expected WG1 activities to be mostly targeted at crowds of teachers and WG2 activities at crowds of learners. Accordingly, we expected WG1 members to be less interested in WG2-related activities, and vice versa, and wanted to ensure an effective use of the participants' time and effort.

Unlike the first two WGs, Working Group 3 (WG3) was focused on language learning only and aimed at reviewing and exploring user-oriented design strategies for online language learning applications, with the ultimate intent of fostering know-how with regard to attracting and retaining a crowd of teachers and learners.

Finally, Working Groups 4 and 5 (WG4 and WG5) were focused on the technical aspects (WG4) and the ethical, legal, or business-related aspects (WG5) of applications for language learning and crowdsourcing. They were established to account for and study the transversal challenges met by the efforts undertaken in WG1, WG2 and WG3.

Besides the five working groups, three additional coordinating groups called the Outreach coordination, Dissemination coordination and Exploitation coordination, were created to better address WG-transversal needs and ensure, whenever possible and relevant, homogeneous approaches in doing so. Such coordination groups were thus designed to better monitor and support the efforts tackling the three aforementioned network-oriented goals.

4 Achievements, failures and lessons learned

In this section, we discuss the extent to which enetCollect succeeded in pursuing the five goals.

4.1 Network-oriented objectives

4.1.1 Bringing together relevant stakeholders from different domains

EnetCollect was originally designed to involve stakeholders fitting four profiles: (1) content-creation experts, ranging from teachers to researchers; (2) content-usage experts, primarily teachers, who would provide end-user perspectives for the creation of crowdsourced material; (3) crowdsourcing experts, mostly researchers, concerned with crowdsourcing strategies and methods; and (4) Content Management System (CMS) developers, especially Learning Management System (LMS) developers, who would provide expert knowledge to study the technical conditions needed to devise an adequate online environment.

As the participants often matched more than one target profile, we are unable to provide precise statistics regarding the composition of the enetCollect network. Nonetheless, we can attest that all four targeted groups were represented, with university stakeholders (researchers and language teachers) making up the greatest part. In contrast, content-creation experts, CMS developers and, in general, non-academic and commercial stakeholders, took much more effort to engage (even though some did participate, especially through meetings). This can be explained by enetCollect's research-oriented nature (like most COST Actions) and by its funding scheme, which does not cover human resources but networking activities. EnetCollect's topic fits into the agenda of researchers, especially young ones, rather than those of language learning teachers, textbook creators or online providers, who usually follow output-oriented, well-defined and established procedures with little room for exploration, even more so when the cost of human resources is not covered.

Overall, enetCollect brought together an interdisciplinary consortium of more than 120 Management Committee (MC) members, 200 associated members registered on the intranet (including MC members) and more than 275 people signing up to the main mailing list (including associated members). As shown in Figure 1, the growth in intranet and mailing list registration was constant until the beginning of the COVID-19 pandemic, while the number of MC members almost reached the maximum possible after one year. As the network grew in a rather fast and organic fashion, no transversal need was identified, and the Outreach coordination group that was originally appointed to tackle any such related need quickly became inactive.

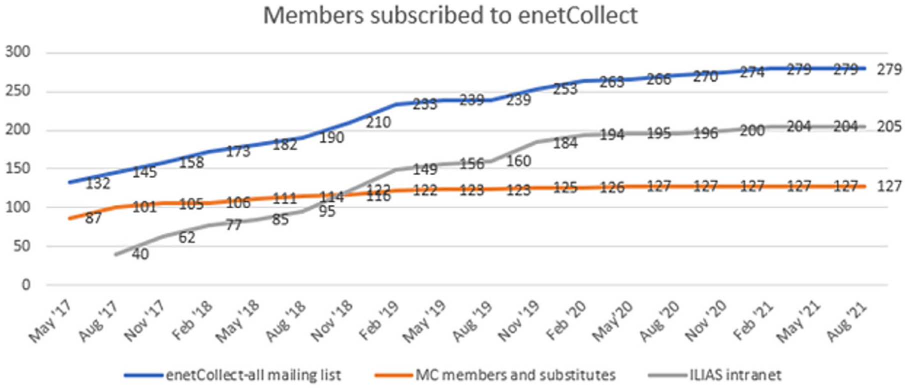


Figure 1: EnetCollect member statistics over the lifetime of the Action.

In terms of direct in-person interactions and collaborations, the Action funded 54 scientific exchanges between pairs of members and organized two hackathon-like events that allowed members (74 participants overall) to intensively collaborate and start new shared efforts over the span of a few days, and nine meetings allowing members to present and discuss their results (519 participants). Numerous collaborative efforts took place online, especially after the COVID-19 pandemic had started and in-person interactions were not recommended.

We would have considered this goal entirely fulfilled had we managed to involve more of the less-represented profiles noted earlier. Nonetheless, our achievements were satisfying, especially regarding the relatively high number of interactions and participations, as it is comparable with some small- and medium-sized well-established research communities we know of and participate in.

4.1.2 Establishing and consolidating communication channels and dissemination procedures

In terms of communication, we set up a website, an intranet, three social media accounts (Twitter, Facebook and ResearchGate), a video channel (Videolectures), a Zotero repository for scientific publications, 19 different mailing lists, as well as branding materials (logo, flyers, a Microsoft PowerPoint template, etc.). So as to disseminate enetCollect's achievements outside the network, the Action funded ten participations at scientific events (mostly conferences).

The website's primary use was to share enetCollect's objectives and achievements with a wider audience. Before the COVID-19 pandemic, the website averaged 3,000 visits per month. The use of the intranet was minimal, but allowed us to make available documents and obtain basic yet practical information about the members at registration. The social media accounts, video channels, and publication repository allowed for better internal dissemination among members. In contrast, the numerous mailing lists allowed for better targeting of the relevant set of members for every communication channel. The Dissemination coordination group greatly facilitated the organization of the dissemination efforts in a systematic fashion. Overall, the main communication channels showed a steady increase in use, which subsequently and suddenly dropped with the onset of the pandemic.

While not all communication efforts were fruitful (e.g. some mailing lists remained inactive), we consider dissemination an achieved goal. As a positive side-effect, it also allowed the members contributing to it to gain practical experience and skills, which will certainly be of interest to future network initiatives.

4.1.3 Strategies for related project building and funding acquisition

Three levels of funding acquisition were actively considered and fostered by sharing information on the enetCollect website and mailing lists, motivating members at meetings and via email, and by offering information and specialized sessions at enetCollect events.

The first level consisted of smaller project funding that could accompany enetCollect as soon as possible and contribute to achieving its objectives throughout the lifetime of the Action and thereafter. For this line of initiatives, we identified several options: national COST-related funding (as found in Switzerland, Turkey, etc.), PhD scholarships associated with enetCollect member institutions, Marie Curie Individual Fellowship grants and small-scale national funding schemes. EnetCollect members were successful in acquiring some national COST-related funding, PhD scholarships and small-scale national funding, while only one Marie Curie Individual Fellowship was obtained.

The second level of funding corresponded to funding to acquire mid-way through the Action period in order to follow up on specific aspects of enetCollect after its end. For this line of initiatives, the Erasmus+ Key Action 2 scheme, European-funded joint projects across two countries and medium scale national funding schemes were identified as relevant. Related efforts led to a few project applications, which were unfortunately unsuccessful. This was mainly due to the pandemic-related cancellation of meetings and of intensive network interaction during the second half of the Action, leading enetCollect consortia to discontinue the preparation of new proposals and/or the improvement of rejected ones.

The third level of funding was sought towards the second half of the Action to further develop enetCollect with the objective of creating a long-term stable research and application context. This funding effort was expected to be piloted by a consortia of enetCollect leaders. For this line of initiatives, we considered the Horizon 2020/Europe research and innovation program and ICT training networks. Similarly to the second level of funding, the work on preparing such a large-scale proposal was discontinued when the COVID-19 pandemic hit, as the consortia of enetCollect leaders focused on keeping their respective parts of enetCollect as active as possible.

For all of the above reasons, we would consider this objective as mostly unfulfilled. As lessons learned, we believe our efforts should have been more narrowly targeted at only a few well-identified funding schemes accessible to most members. In that respect, we believe the Marie Curie Individual Fellowships and the Erasmus+ Key Action 2 schemes to be the most relevant.

4.2 Research-oriented objectives

4.2.1 Transversal challenges faced

Research-wise, we could observe three major transversal challenges.

Regarding the first challenge, network project schemes such as COST Actions do not typically cover human resource costs but primarily rely on a stakeholder's willingness to invest time in the short-term for a medium- to long-term return on investment in the form of scientific publications and/or funded projects. At the same time, because these schemes are open by nature, they rely on meetings and scientific exchanges to define milestones and make progress. While the fact of not covering human resource costs naturally limits the participation of non-publicly funded stakeholders while fostering the participation of publicly funded ones, both this aspect and the need for scientific exchanges and meetings are impossible to fulfill if the participating stakeholders are unable to allocate time or meet in person, which is what happened during the chaotic COVID-19 period. Our experience tells us that network project schemes should cover some minimal human resource costs, especially for the leaders of the project, and should factor in the possibility of being put on hold if extraordinary circumstances require it.

With respect to the second challenge, and as recorded in the Zotero repository, enetCollect members published more than fifty scientific journal, workshop and conference articles, thus creating the groundwork for a previously largely unexplored topic. Nonetheless, enetCollect's own interdisciplinary nature (linguistics, lexicography, language studies, language pedagogy, Computer-Assisted Language Learning, NLP, etc.) proved challenging for the publication of scientific articles for various reasons connected to the emerging and interdisciplinary research subject tackled by enetCollect. As such, its publications indirectly relate to several research areas without having its own venue and audience outside of the project itself. Therefore, publishing works that are related to, but do not fully match the expectations of a research community, has proven to be challenging on various occasions for various reasons. First, reviewers naturally have specialized knowledge and expectations on only part of the interdisciplinary subject discussed (i.e. the language learning or the computational/crowdsourcing side). As a result, scientific publications need to be tailored to the specific interests of the targeted research community, thus forcing authors to prioritize some research aspects over others. Second, and unexpectedly, a notable number of reviewers also had inadequate expectations regarding aspects they knew little about (e.g. NLP reviewers with respect to CALL-related evaluation procedures), thus compelling authors to spend time addressing their concerns. Third, the research outputs of enetCollect are rather exploratory and were often considered too vague or preliminary. Fourth, since very few related works exist, reference values for evaluation are often missing, thus potentially undermining the credibility of the work. For all these reasons, the need to establish new publication venues for this emerging field seems inevitable.

Regarding the third challenge, stakeholders participating in such networks usually have little time for them, and their innovative nature rarely aligns with their short-term interests, nor are they fully covered by their expertise. As such, most efforts can only be conducted by a group of participants and, in order to further enhance achievements, proper strategies to foster such collaboration are needed. Our experience with enetCollect allowed us to identify one very suitable strategy which fostered a large number of the collaborations that led to the scientific achievements discussed in the upcoming sections. This is the organization of hackathon-like events where some members first answer an open call for topics they would like to tackle collaboratively, and are then asked to lead a taskforce. The topics are later disclosed to the remaining members, who can then ask to participate in one or more of the taskforces. The candidate participants of the topics that received enough attention are then invited to the hackathon-like event to kickstart the task forces by working intensively over the span of a few days and perform the groundwork needed for their collaborations to develop after the end of the event itself.

4.3 Creating a shared understanding and a theoretical framework for approaching the combination of language learning and crowdsourcing

With respect to this theory-oriented goal, a literature review conducted in 2017/2018 by WG1 members revealed that there were very few examples of past crowdsourcing efforts in the field of language learning at that time. They also gathered the opinions of relevant stakeholders in three ways. Firstly, they conducted a short survey among themselves to identify the aspects of language learning with the most potential for crowdsourcing. Secondly, they circulated a survey among teachers to assess their familiarity with crowdsourcing methods, and find possible use cases in teaching practice (Arhar Holdt et al., 2020). Finally, they circulated another survey among learners to determine their familiarity with crowdsourcing, and their attitudes towards materials (potentially) produced in this way (Hatipoglu et al., 2020; Miloshevska et al., 2021). Understanding stakeholders' perspectives was crucial in setting up a suitable theoretical framework, and to identify the areas with the most potential for crowdsourcing. WG1 members also explored specific subjects, including how to develop an open dictionary for the contemporary Serbian language using crowdsourcing techniques (Lazić Konjik and Milenković, 2021), how to develop pedagogically appropriate language corpora through crowdsourcing and gamification (Zviel-Girshin et al., 2021) to crowdsource linguistic knowledge regarding Dutch blends, neologisms and language variation (Dekker and Schoonheim, 2018), and how to crowdsource second language learning material, with a focus on vocabulary lists, in order to reduce dependency on costly expert manpower (Alfter et al., 2020).

The efforts of WG2 members mostly focused on learners as the most relevant crowd to perform implicit crowdsourcing, and can for the most part be related to an overarching paradigm that pairs up a type of exercise with a specific type of language-related dataset, which can be used to generate exercise content (Nicolas et al., 2020, 2021). More specifically, in order to understand how such a paradigm could be implemented, these efforts studied its context: some efforts studied the exercises compatible with the paradigm (i.e. which content could be automatically generated from specific language-related datasets, Lyding et al., 2022), other efforts studied the type of language-related datasets most commonly crowdsourced,[9] or aimed at mapping the existing language learning platforms where such a paradigm could be integrated. Other specific efforts researched how to adequately apply

[9] The results of these efforts have yet to be published.

this paradigm to crowdsource semantic relations between English or Romanian words (Lyding et al., 2019; Nicolas et al., 2021; Rodosthenous et al., 2019, 2020), defined new workflows to include teachers and crowdsource linguistic knowledge about English verb-particle constructions (Grace Araneta et al., 2020), as well as crowds of relatives of learners to crowdsource Alsatian lexical knowledge (Millour et al., 2019). While they did not specifically target a crowd of learners or teachers, others studied how to crowdsource recordings of Italian Dialects (Sangati et al., 2018) and complex associations among words by means of a board game workflow (Smrz, 2019).

WG3 objectives were largely pursued together with WG1, WG2 and WG5 objectives. Some related efforts made it possible to map and study a rather large number of existing language learning solutions (Bączkowska, 2021; Bodorík and Bédi, 2018; Grygo and Gajek, 2018). Other work allowed us to better understand teachers' and learners' perception of crowdsourcing as a concept (Arhar Holdt et al., 2020; Hatipoglu et al., 2020), while other studies presented enetCollect from a language learning perspective (Gajek, 2020; Lyding et al., 2018). In a number of publications (Cornillie, 2018; Gajek, 2018; Murray and Giralt, 2018), WG3 members also discussed important design choices when creating a language learning application, especially one that made use of crowdsourcing, and in others (Cucchiarini and Strik, 2018; Ostanina-Olszewska, 2019; Pereira et al., 2018) they discussed recent language learning technologies.

WG4 efforts, which aimed at defining technology-oriented specifications, allowed us to draw two conclusions. Firstly, the developments made in the context of enetCollect were still too heterogeneous and prototypical to define any transversal technical solution they could share and rely on. Indeed, even though many approaches undertaken shared some common needs (e.g. aggregation methods to cross-check the linguistic inputs that were crowdsourced), it was too early to establish technical solutions encoding sophisticated and standardized methods. Secondly, no open-source solution was readily available to implement a language learning platform and, regarding the closest solutions that could have been adapted (i.e. the Learning Management Systems, also known as LMS), the related communities at the time had little interest in language learning but were more focused on other subjects (e.g. mathematics or physics) posing fewer subject-specific technical challenges. Indeed, because of its nature, language learning requires specific technologies, such as automatic speech recognition. This explained the absence of readily available solutions for Language Learning and the difficulty in involving LMS-oriented stakeholders.

Finally, WG5 aimed at devising application-oriented specifications for an ethical, legal and profitable solution. Similarly to WG4, the efforts pursued in the context of enetCollect were still too recent to define any transversal specifications on these aspects. Nonetheless, several WG5 members managed to tackle relevant issues in a prospective fashion and discussed aspects such as the ownership of the data, the need for private or open-source code, third-party dependencies or privacy (Chua et al., 2018; Chua and Rayner, 2018); how to balance a collaboration between teachers and academics (Chua and Rayner, 2019); how to implement gamification strategies in an ethical fashion (Murray and Giralt, 2018); as well as legal issues with respect to European regulations of online learning platforms such as LMS, Massive Open Online Courses (MOOCS) or Open Educational Resources (OERs) (Zdravkova, 2018, 2019). Finally, a framework to address ethical issues affecting three groups of stakeholders (collaborative content creators, prospective users, and the institutions intending to implement the approach for educational purposes) was proposed (Zdravkova, 2020). As no direct collaboration with business-oriented stakeholders could be established, the question of defining business guidelines was not explored.

Overall, with respect to this objective, the versatility of the relevant aspects discussed is far greater than we originally expected and without much overlap in terms of main focus. Such versatility allows us to draw two further conclusions. Firstly, a dedicated R&I community is needed to adequately take on the topic. Second, the lack of convergence in terms of main focus (most of the aforementioned publications could hardly cite one another in their state of the art as a directly comparable work), might lead one to think that the paths followed still have many interesting results to yield, while others are still waiting to be explored. In other words, we believe that the research on the combination of crowdsourcing and language learning has progressed and gained notable results, but is still in its early stages.

4.4 To research use cases and work on prototypes

With respect to this output-oriented objective, most achievements were obtained in the context of WG1 and WG2, often in collaboration with WG3 and WG4.

In the context of WG1, experiments were performed to crowdsource linguistic knowledge regarding Dutch blends, neologisms and language variation (Dekker and Schoonheim, 2018), and to crowdsource vocabulary lists to be used as L2 learning material (Alfter et al., 2020). Other efforts fostered the development of a mobile application for the gamified improvement of two automatically compiled dictionaries for Slovene (Arhar Holdt et al., 2021; Arhar Holdt and Čibej, 2020; Čibej and Arhar Holdt, 2019), and the development of LARA, a learning and reading assistant with explicit crowdsourcing abilities aimed at teachers (Akhlaghi et al., 2019b, 2019a, 2020; Bédi, Bernharðsson, et al., 2020; Bédi, Butterweck, et al., 2020; Bédi et al., 2019; Butterweck et al., 2019; Chua and Rayner, 2019; Habibi, 2019).

In the context of WG2, the V-trel vocabulary trainer with implicit crowdsourcing abilities is geared toward learners and teaches them – through a Telegram bot[10] – English and Romanian semantic relations between words, while crowdsourcing their linguistic judgements (Lyding et al., 2019; Nicolas et al., 2021; Rodosthenous et al., 2019, 2020). Two other prototypes were also implemented – again through Telegram bots – a new learning and teaching workflow to generate exercises and crowdsource linguistic knowledge about English verb-particle constructions (Grace Araneta et al., 2020), as well as a crowdsourcing mechanism to obtain recordings of Italian dialects (Sangati et al., 2018). The prototype described in Millour et al. (2019), however, relied on a role-playing game framework to crowdsource Alsatian lexical knowledge from learners and their relatives. The prototype discussed in Smrz (2019) fully reimplemented a popular board game in order to crowdsource complex associations among words.

[10] https://telegram.org/

Overall, the research on use cases and the work on prototypes has been more limited than we originally hoped, as attested by the limited number of outputs to crowdsource linguistic knowledge other than lexical knowledge. We identified two main reasons for this. Firstly, the minimal involvement of non-academic stakeholders had noticeable implications for devising and testing prototypes. Indeed, enetCollect was somehow lacking direct evaluation and feedback from those stakeholders who use and create language learning solutions daily. It also prevented enetCollect from accessing existing exercise content or the involvement of large crowds of students and online learners needed to more extensively test the prototypes that were devised. Secondly, the efforts to tackle this goal were mostly planned for the second half of the action, while the efforts planned for the first half would for most part focus on building the network and researching the theoretical framework. As such, the bulk of efforts with regard to research on use cases and work on prototypes only began some months before the COVID-19 pandemic itself started, which obviously limited many of these efforts and completely halted others.

5 Conclusions and future steps

While the achievements of enetCollect were rated by the COST agency as "excellent" and "very good" in their mid-way and final formal evaluations, our overall assessment is more modest.

Regarding the network-oriented goals, we believe that enetCollect mostly fulfilled its role. Indeed, given the rather large number of stakeholders that participated and collaborated, we believe it is fair to say that a new research community was created. We also believe that enetCollect could have achieved even greater results had it been supported by the COST agency with more readily available tools, procedures or guidelines to tackle various transversal aspects, such as dissemination or funding acquisition.

Regarding research-oriented objectives, the high-potential of the language learning and crowdsourcing combination was more widely acknowledged than we had originally imagined, as attested by the large participation of an international audience of stakeholders, who deemed enetCollect worth their time. In terms of outputs, we believe the number of publications achieved and the prototypes devised to be fair considering the innovative nature of enetCollect and the disruption caused by COVID-19. Nonetheless, for the reasons discussed in Sections 4.4 and 4.5, we believe that the research on this topic is still in its early stages.

As a follow-up to enetCollect, we have established the DARIAH Working Group *Combining Language Learning with Crowdsourcing Techniques* (D4COLLECT), which will serve as a flexible and dynamic bottom-up institutional framework for knowledge exchange, research coordination and capacity building. Following in enetCollect's footsteps, D4COLLECT aims to bring together language teachers and experts in linguistics, computational linguistics, educational sciences, software engineering and the digital humanities to explore digital workflows, tools, and solutions for deploying implicit and explicit crowdsourcing methods in the creation of language-learning materials and the collection of language datasets. Our first efforts will target the organization of hackathon-like events (see Section 4.2.1). D4COLLECT will also serve as a practical context to promote the submission of project proposals that, if funded, would allow to speed up and better shape the efforts of the Working Group's members.

Acknowledgements

While this article was authored by two people, it reviews and describes the efforts and achievements of many enetCollect members, including many whose names can be found in the references below and whom we would like to thank. We are particularly thankful to the members of the Core Group who led enetCollect with us for over 4.5 years.

Finally, we would like to thank Greta H. Franzini and Egon W. Stemle for their support in polishing this article.

References

Akhlaghi, E., Bédi, B., Bektaş, F., Berthelsen, H., Butterweck, M., Chua, C., ..., & Strik, H. (2020). Constructing Multimodal Language Learner Texts Using LARA: Experiences with Nine Languages. *Proceedings of LREC 2020, 12th Language Resources and Evaluation Conference* (pp. 323–331).

Akhlaghi, E., Bédi, B., Butterweck, M., Chua, C., Gerlach, J., Habibi, H., ..., & Zuckermann, G. (2019a). Overview of LARA: A Learning and Reading Assistant. *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education* (pp. 99–103). doi: 10.21437/SLaTE.2019-19

Akhlaghi, E., Bédi, B., Butterweck, M., Chua, C., Gerlach, J., Habibi, H., ..., & Zuckermann, G. (2019b). Demonstration of LARA: A Learning and Reading Assistant. *Proceedings of SLaTE of the 8th ISCA Workshop on Speech and Language Technology in Education* (pp. 37–38). doi: 10.21437/SLaTE.2019-19

Alfter, D., Lindström Tiedemann, T., & Volodina, E. (2020). *Expert judgments versus crowdsourcing in ordering multi-word expressions*. Paper presented at Swedish Language Technology Conference, Gothenburg, Sweden. Retrieved from https://gubox.app.box.com/v/SLTC-2020-paper-16

Arhar Holdt, Š., & Čibej, J. (2020). Rezultati projekta Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost. *Proceedings of the Language Technologies and Digital Humanities Conference* (pp. 3–9). Retrieved from http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Rezultati-projekta_Slovar-sopomenk-sodobne-slovenscine.pdf

Arhar Holdt, Š., Logar, N., Pori, E., & Kosem, I. (2021). Game of Words: Play the Game, Clean the Database. *Proceedings of the 14th Congress of the European Association for Lexicography, EURALEX 2021* (pp. 41–49).

Arhar Holdt, Š., Zviel-Girshin, R., Gajek, E., Durán-Muñoz, I., Bago, P., Fort, K., ..., & Zanasi, L. (2020). Language Teachers and Crowdsourcing: Insights from a Cross-European Survey. *Journal of the Institute of Croatian Language and Linguistics // Časopis Instituta Za Hrvatski Jezik i Jezikoslovlje, 46*(1), 1–28. doi: 10.31724/rihjj.46.1.1

Bączkowska, A. (2021). An overview of popular website platforms and mobile apps for language learning. *Forum Filologiczne Ateneum* (pp. 9–35).

Bédi, B., Bernharðsson, H., Chua, C., Björg Guðmarsdóttir, B., Habibi, H., & Rayner, M. (2020, August 20). Constructing an interactive Old Norse text with LARA. *Proceedings of EUROCALL 2020, European Association for Computer Assisted Language Learning, Copenhagen.*

Bédi, B., Butterweck, M., Chua, C., Gerlach, J., Björg Guðmarsdóttir, B., Habibi, H., ..., & Vigfússon, S. (2020). LARA: An extensible open source platform for learning languages by reading. *Proceedings of EUROCALL 2020, European Association for Computer Assisted Language Learning, Copenhagen.*

Bédi, B., Chua, C., Habibi, H., Martinez-Lopez, R., & Rayner, M. (2019). Using LARA for language learning: A pilot study for Icelandic. *CALL and Complexity – Short Papers from EUROCALL 2019*, *Louvain-la-Neuve, Belgium*. Retrieved from https://books.google.fr/books?id=EHnCDwAAQBAJ&lpg=PA33&lr&hl=fr&pg=PA33#v=onepage&q&f=false

Bodorík, M., & Bédi, B. (2018). In Search of the State of Language Learning Online in Europe. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands*. Retrieved from http://ceur-ws.org/Vol-2390/PaperA1.pdf

Butterweck, M., Chua, C., Habibi, H., Rayner, M., & Zuckermann, G. (2019). Easy Construction of Multimedia Online Language Textbooks And Linguistics Papers with LARA. *Proceedings of the 12th Annual International Conference of Education, Research and Innovation* (pp. 7302–7310). doi: 10.21125/iceri.2019.1737

Chua, C., Habibi, H., Rayner, M., & Tsourakis, N. (2018). Decentralising Power: How we are Trying to Keep CALLector Ethical. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands*. Retrieved from http://ceur-ws.org/Vol-2390/PaperC3.pdf

Chua, C., & Rayner, M. (2018). What do the Founders of Online Communities Owe to their Users? *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands.* Retrieved from http://ceur-ws.org/Vol-2390/PaperD1.pdf

Chua, C., & Rayner, M. (2019). Vegetarians Vampires: Why the CALL Technology Provider Doesn’t Have to Suck the Teacher’s Blood. *Proceedings of the 12th Annual International Conference of Education, Research and Innovation* (pp. 7860–7869). doi: 10.21125/iceri.2019.1863

Čibej, J., & Arhar Holdt, Š. (2019). Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene. *Proceedings of the ELex 2019 Conference*: *Electronic lexicography in the 21st century, Sintra, Portugal*. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_19.pdf

Cucchiarini, C., & Strik, H. (2018, October 24). Crowdsourcing for Research on Automatic Speech Recognition-enabled CALL. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018, Leiden, The Netherlands.* Retrieved from http://ceur-ws.org/Vol-2390/PaperB2.pdf

Dekker, P., & Schoonheim, T. (2018, October 24). Crowdsourcing for Dutch using PYBOSSA: Case studies on Blends, Neologisms and Language Variation. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands*. Retrieved from http://ceur-ws.org/Vol-2390/PaperB3.pdf

Gajek, E. (2020). Crowdsourcing in language learning as a continuation of CALL in varied technological, social, and ethical contexts. *Proceedings of EUROCALL 2020, European Association for Computer Assisted Language Learning* (pp. 75–80). doi: 10.14705/rpnet.2020.48.1168

Grace Araneta, M., Eryigit, G., König, A., Lee, J.-U., Luís, A., Lyding, V., ..., & Sangati, F. (2020). Substituto—A Synchronous Educational Language Game for Simultaneous Teaching and Crowdsourcing. *Proceedings of the 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)* (pp. 1–9). doi: 10.3384/ecp201759

Grygo, A., & Gajek, E. (2018). Risks of Using Duolingo by Polish Learners at Primary Level. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018, Leiden, The Netherlands.* Retrieved from http://ceur-ws.org/Vol-2390/PaperA4.pdf

Habibi, H. (2019). LARA Portal: A Tool for Teachers to Develop Interactive Text Content, an Environment for Students to improve Reading Skill. *Proceedings of the 12th Annual International Conference of Education, Research and Innovation* (pp. 8221–8229). doi: 10.21125/iceri.2019.1954

Hatipoglu, C., Gajek, E., Miloshevska, L., & Delibegovic Dzanic, N. (2020). Crowdsourcing for widening participation and learning opportunities: A view from language learners' window. *Proceedings of EUROCALL 2020, European Association for Computer Assisted Language Learning* (pp. 81–87). doi: 10.14705/rpnet.2020.48.1169

Hladká, B., Hana, J., & Lukšová, I. (2014). Crowdsourcing in language classes can help natural language processing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *Vol. 2* (pp. 71–72).

Lazić Konjik, I., & Milenković, A. (2021). The Development of the Open Dictionary of Contemporary Serbian Language Using Crowdsourcing Techniques. *Proceedings of the 14th Congress of the European Association for Lexicography, EURALEX 2021* (pp. 479–484).

Lyding, V., Nicolas, L., Bédi, B., & Fort, K. (2018). Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing Techniques (enetCollect). *Future-Proof CALL: Language Learning as Exploration and Encounters – Short Papers* (pp. 176–181). doi: 10.14705/rpnet.2018.26.833

Lyding, V., Nicolas, L., & König, A. (2022, June). About the Applicability of Combining Implicit Crowdsourcing and Language Learning for the Collection of NLP Datasets. *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022* (pp. 46–57).

Lyding, V., Rodosthenous, C. T., Sangati, F., ul Hassan, U., Nicolas, L., König, A., Horbacauskiene, J., & Katinskaia, A. (2019). V-trel: Vocabulary Trainer for Tracing Word Relations—An Implicit Crowdsourcing Approach. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria.* Retrieved from http://lml.bas.bg/ranlp2019/proceedings-ranlp-2019.pdf

Millour, A., Araneta, M. G., Lazić Konjik, I., Raffone, A., Pilatte, Y.-A., & Fort, K. (2019). Katana and Grand Guru: A Game of the Lost Words (DEMO). *Proceedings of the Ninth Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'19), Poznan, Poland.*

Miloshevska, L., Delibegović Džanić, N., Hatipoğlu, Ç., & Gajek, E. (2021). Crowdsourcing for language learning in Turkey, Bosnia and Herzegovina, Republic of North Macedonia and Poland. *Journal of Narrative and Language Studies*, *9*, 106–121.

Murray, L., & Giralt, M. (2018, October 24). Motivational, Ethical and Gamification Issues in Crowdsourcing. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*. enetCollect WG3 & WG5 Meeting 2018, *Leiden, The Netherlands*. Retrieved from http://ceur-ws.org/Vol-2390/PaperC1.pdf

Nicolas, L., Aparaschivei, L., Lyding, V., Rodosthenous, C., Sangati, F., König, A., & Forascu, C. (2021). An Experiment on Implicitly Crowdsourcing Expert Knowledge about Romanian Synonyms from L1 Language Learners. *Proceedings of 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)* (pp. 1–14). Retrieved from https://ep.liu.se/konferensartikel.aspx?series=ecp&issue=177&Article_No=1

Nicolas, L., Lyding, V., Borg, C., Forascu, C., Fort, K., Zdravkova, K., Kosem, I., ..., & HaCohen-Kerner, Y. (2020). Creating Expert Knowledge by Relying on Language Learners: A Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning. *Proceedings of LREC 2020, 12th Language Resources and Evaluation Conference* (pp. 268–278). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.34

Nicolas, L., Ostanina-Olszewska, J., Arhar Holdt, Š., Čibej, J., Borg, C., Lyding, V., & Barreiro, A. (2021). Introducing an implicit crowdsourcing opportunity to teachers. *CALL for Background* (pp. 115–136). Peter Lang. Retrieved from https://www.peterlang.com/view/9783631849200/html/ch14.xhtml

Ostanina-Olszewska, J. (2019). Modern technology in language learning and teaching. *Linguodidactica*, *22*, 153–164. doi: 10.15290/lingdid.2018.22.10

Pereira, M. J., Fialho, P., Coheur, L., & Ribeiro, R. (2018). Chatbots' Greetings to Human-Computer Communication. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands.* Retrieved from http://ceur-ws.org/Vol-2390/PaperD2.pdf

Rodosthenous, C., Lyding, V., König, A., Horbacauskiene, J., Katinskaia, A., ul Hassan, U., Isaak, N., Sangati, F., & Nicolas, L. (2019). Designing a Prototype Architecture for Crowdsourcing Language Resources. *Poster Session of the 2nd Conference on Language, Data and Knowledge 2019*. Retrieved from http://ceur-ws.org/Vol-2402/paper4.pdf

Rodosthenous, C., Lyding, V., Sangati, F., König, A., ul Hassan, U., Nicolas, L., Horbacauskiene, J., Katinskaia, A., & Aparaschivei, L. (2020). Using Crowdsourced Exercises for Vocabulary Training to Expand ConceptNet. *Proceedings of LREC 2020, 12th Language Resources and Evaluation Conference* (pp. 307–316).

Sangati, F., Abramova, E., & Monti, J. (2018). DialettiBot: A Telegram Bot for Crowdsourcing Recordings of Italian Dialects. *Proceedings of the Fifth Italian Conference on Computational Linguistics*. CLIC-it 2018, Torino.

Sangati, F., Merlo, S., & Moretti, G. (2015). School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE* (pp. 16–19).

Smrz, P. (2019). Crowdsourcing Complex Associations among Words by Means of A Game. *Proceedings of CSTY 2019, 5th International Conference on Computer Science and Information Technology* (p. 9). Retrieved from https://aircconline.com/csit/abstract/v9n14/csit91407.html

Von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 1–2).

Zdravkova, K. (2018). Privacy of Crowdsourcing Educational Platforms in the Light of New EU Regulation. *Proceedings of the EnetCollect WG3 & WG5 Meeting 2018*, *Leiden, The Netherlands.* Retrieved from http://ceur-ws.org/Vol-2390/PaperC2.pdf

Zdravkova, K. (2019). Compliance of MOOCs and OERs with the new privacy and security EU regulations. *Proceedings of the 5th International Conference on Higher Education Advances (HEAd'19)*, *Vol. 1* (pp. 159–167). Retrieved from http://www.headconf.org/wp-content/uploads/pdfs/9063.pdf

Zdravkova, K. (2020). Ethical issues of crowdsourcing in education. *Journal of Responsible Technology*, *2–3*, 100004. doi: 10.1016/j.jrt.2020.100004

Zviel-Girshin, R., Zingano Kuhn, T., R. Luis, A., Koppel, K., Šandrih Todorović, B., Arhar Holdt, Š., Tiberius, C., & Kosem, I. (2021). Developing pedagogically appropriate language corpora through crowdsourcing and gamification. *Proceedings of EUROCALL 2021* (pp. 312–317). doi: 10.14705/rpnet.2021.54.1352

EnetCollect – Evropska mreža za združevanje jezikovnega izobraževanja s tehnikami množičenja (COST Action CA16105): pregled projektne vizije, ureditve, napredka in dosežkov

V tem članku predstavljamo pregled Evropske mreže za združevanje jezikovnega izobraževanja s tehnikami množičenja (enetCollect), obseženega projekta za spodbujanje raziskav in inovacij (R&I) na področju združevanja množičenja in učenja jezikov. Opisujemo začetke projekta, predstavljamo njegovo splošno zasnovo in ureditev ter razpravljamo o dosežkih v smislu (1) ustvarjanja nove skupnosti za raziskave in inovacije z zaključenim obsežnim mrežnim projektom in (2) spodbujanja raziskav in inovacij na večinoma neraziskanem področju z velikim potencialom.

Razpravljamo tudi o povezanih izzivih in pridobljenih izkušnjah pri oblikovanju in vodenju nove skupnosti R&I ter izzivih, ki smo jih opazili pri delu članic mreže enetCollect med spoznavanjem številnih plati tako raznolikega projekta.

Ključne besede: enetCollect, COST Action, množičenje, jezikovno izobraževanje