About the Applicability of Combining Implicit Crowdsourcing and Language Learning for the Collection of NLP Datasets.

Verena Lyding[1], Lionel Nicolas[1], Alexander König[2]
[1]Eurac Research, Bolzano, Italy, [2]CLARIN ERIC, Utrecht, Netherlands
verena.lyding,lionel.nicolas@eurac.edu, alex@clarin.eu

Abstract

In this article, we present a recent trend of approaches, hereafter referred to as Collect4NLP, and discuss its applicability. Collect4NLP-based approaches collect inputs from language learners through learning exercises and aggregate the collected data to derive linguistic knowledge of expert quality. The primary purpose of these approaches is to improve NLP resources, however sincere concern with the needs of learners is crucial for making Collect4NLP work. We discuss the applicability of Collect4NLP approaches in relation to two perspectives. On the one hand, we compare Collect4NLP approaches to the two crowdsourcing trends currently most prevalent in NLP, namely Crowdsourcing Platforms (CPs) and Games-With-A-Purpose (GWAPs), and identify strengths and weaknesses of each trend. By doing so we aim to highlight particularities of each trend and to identify in which kind of settings one trend should be favored over the other two. On the other hand, we analyze the applicability of Collect4NLP approaches to the production of different types of NLP resources. We first list the types of NLP resources most used within its community and second propose a set of blueprints for mapping these resources to well-established language learning exercises as found in standard language learning textbooks.

Keywords: Crowdsourcing, Language Learning, Natural Language Processing, Language Resources

1. Introduction

The lack of NLP resources or the quality and/or coverage issues of existing ones is a long-standing obstacle that has slowed down the research in NLP for all languages in general, especially for lower-resourced ones. As most NLP resources cannot be obtained in a purely automatic fashion, creating and/or curating them requires human intervention and, accordingly, a key obstacle for the creation of such datasets is the high cost, both temporal and economic. As a result, most efforts to build NLP resources have focused on a limited set of NLP resources for a handful of languages such as English and other widely used languages. In order to tackle this challenge, some efforts have relied on *crowdsourcing* (Howe and others, 2006) to increase the amount of manpower and/or reduce costs.

Indeed, as reCAPTCHA and the Wikipedia initiative have proven, crowdsourcing is a versatile approach that can be successfully applied to overcome challenging tasks that, in most cases, cannot be solved by automatic means and/or require an excessive amount of cost-intensive expert manpower. Crowdsourcing can be applied in many fields, provided that the tasks tackled can be solved by a crowd of people with a compatible skill set. This aspect makes NLP a very apt field of application since, depending on how the task is presented, it can rely on the language skills of any language speaker as a potential crowd to tackle the collection of NLP datasets for the languages spoken nowadays.[1] Efforts aiming at crowdsourcing NLP resources thus started soon after the rise of crowdsourcing back in 2006 (Howe and others, 2006) and have been followed up by numerous efforts over the past 1.5 decades.

In this article, we discuss the recent trend of Collect4NLP-based approaches which collect the inputs provided by language learners to exercises automatically generated from NLP resources and aggregate them in order to derive linguistic knowledge of expert quality (Nicolas et al., 2021) that can be used to update and/or extend NLP resources. In other words, they consider language learners as linguistic experts through a controlled setting designed in the form of language learning exercises and use a large quantity of their inputs to make up for their lower reliability.

First we discuss the applicability of the Collect4NLP approach by comparing it to Crowdsourcing Platform (CP) and Games-With-A-Purpose (GWAPs) based approaches in Section 4, and then present a range of NLP resource types that are compatible with Collect4NLP-based approaches by discussing how the NLP resources could be mapped to exercises in language learning textbooks in Section 5. Before, in Section 2, we overview the state of the art and briefly introduce in Section 3 the key aspects of the Collect4NLP approaches. We discuss future work and conclude in Section 6.

2. Related Works

The related works include the different trends of crowdsourcing approaches used to collect NLP datasets. As such, the relevant state of the art is composed of the three aforementioned trends (CP-based, GWAP-based and Collect4NLP-based approaches) and single efforts that do not fit in any of the three trends.

CP-based crowdsourcing approaches are the ones most commonly explored since crowdsourcing came into the NLP landscape. They rely on dedicated platforms in

[1] Even though linguistic skills can vary among people.

which users perform tasks and are rewarded for it, such as the Amazon Mechanical Turk[2] (AMT), Clickworker[3] or CrowdFactory[4]. In general, the reward on crowdsourcing platforms is a financial compensation of some sorts. In addition, a few crowdsourcing platforms exist, which base their work on a purely altruistic or educational motivation of their volunteers, in the spirit of Citizen Science, such as e.g. Zooniverse[5] or Distributed Proofreaders [6]. Relevant examples of efforts of this trend are, among many others, efforts to collect and transcribe speech corpora (Callison-Burch and Dredze, 2010; Evanini et al., 2010), to carry out word-sense disambiguation (Biemann, 2013) and named entity annotation (Finin et al., 2010; Lawson et al., 2010; Ritter et al., 2011), to create parallel corpora (Zaidan and Callison-Burch, 2011; Post et al., 2012) or to translate WordNets (Ganbold et al., 2018).

GWAP-based approaches mostly started after 2010 and became a trend sufficiently developed and specific to motivate the organization of a regular series of dedicated 'Games and NLP' workshops collocated with NLP conferences.[7] Some of the most well-known GWAP-efforts concern the annotation of anaphoras (Chamberlain et al., 2008; Poesio et al., 2012; Poesio et al., 2013), lexico-semantic associations between words (Lafourcade, 2007), knowledge rules (Rodosthenous and Michael, 2016), syntactic dependency relations (Fort et al., 2014; Guillaume et al., 2016a) or annotation of text-segmentation (Madge et al., 2017), but GWAPs have also been used for other specific subjects such as e.g. the labelling of speech data for language recognition tasks (Cieri et al., 2021).

Collect4NLP-based approaches started to be more intensively explored in the context of a European network project called enetCollect COST Action (European Network for Combining Language Learning with Crowdsourcing Techniques) started in 2017 and completed in 2021 (Nicolas et al., 2020). This project fostered the development of numerous efforts to combine language learning and crowdsourcing to create lexical knowledge or semantic relations between words (Rodosthenous et al., 2019; Lyding et al., 2019; Rodosthenous et al., 2020; Millour et al., 2019; Smrz, 2019; Araneta et al., 2020; Arhar Holdt et al., 2021) or knowledge about idioms (Eryiğit et al., 2022). To our knowledge related research prior to enetCollect is very limited and just includes a few efforts in order to collect translations (von Ahn, 2013), part-of-speech annotations (Sangati et al., 2015) and syntactic dependencies (Hladká et al., 2014) or is related only to the exercise generation part of the paradigm, such as works by (Greene et al., 2004) or (Pilán and Johansson, 2013). The state of the art also includes crowdsourcing efforts that do not fit well in any of the three trends. With respect to this "varia" group, the state of the art includes, among others, efforts to collect sentiment annotations (Funk et al., 2018), spelling errors (Tachibana and Komachi, 2016) and speech data (Mollberg et al., 2020).

3. Collect4NLP in a Nutshell

The umbrella term Collect4NLP stands for *Combining Language Learning with Crowdsourcing Techniques for NLP dataset collection* and includes all approaches implementing an implicit crowdsourcing paradigm (Nicolas et al., 2020). This paradigm states that ***IF** an NLP dataset can be used to generate language learning exercises **THEN** the answers to these exercises can be used to enhance the NLP dataset.*

The paradigm frames a synergy between NLP stakeholders and language learners, resulting from the fact that, on an abstract level, both groups perform similar types of actions: creating and curating a language model. Indeed, while the former create, curate and use a language model in the form of a digital NLP dataset that "teaches" a computer program how to process and produce language content, the latter create, curate and use a language model in the form of personal knowledge allowing them to process and produce language data. By channeling through crowdsourcing the learners' efforts to complete exercises that are automatically-generated from NLP resources, the learners formulate, as a "side-effect" of the learning activity, linguistically-motivated choices and decisions. Those can be used as a (potentially noisy) source of data for the enhancement of NLP resources. In other words, this paradigm considers learners as linguists of lower reliability. Instead of consulting expert linguists on a linguistic question, the paradigm suggests to combine the implicit "judgements" of several learners to answer the same linguistic question.

As demonstrated in Nicolas et al. (2021) for the use-case of enhancing a lexical network for Romanian with synonyms, aggregation mechanisms can make up for the lower reliability of learner data by combining a larger quantity of data. This way by aggregating the inputs of multiple learners to a same set of questions linguistic knowledge of expert quality can be created.

Such aggregation mechanisms work best if the inputs crowdsourced from the learners are as simple as possible. If an exercise allows one to, directly or indirectly, deduce a yes/no judgement from the learner (e.g. *Is 'food' a common noun?*) then we would assume that in most cases[8] the reliability of the learners' answers will range from 50% (random answers) to 100% (correct answer). This implies that each single learner answer, even the weakest ones with 51% reliability, will contribute to reaching statistical certainty. In other words, provided that a sufficient number of answers to the

[2] `https://www.mturk.com/`

[3] `https://www.clickworker.com`

[4] `https://www.cloudfactory.com/`

[5] `https://www.zooniverse.org/`

[6] `https://www.pgdp.net`

[7] `https://gamesandnlp.com/past-workshops/`

[8] effects of language interference like 'false friends' aside

same yes/no question can be collected, deriving the correct answer by cross-matching multiple learner judgements is statistically achievable.

4. Weaknesses and Strengths across the Three Trends of Approaches

In this section, we discuss how the three trends of crowdsourcing approaches compare to one another with respect to the following partly interrelated aspects: crowd motivation, crowd size, crowd involvement, crowdsourcing rate, crowdsourcing quality, and crowdsourcing costs. The crowd size and involvement as well as the crowdsourcing rate and quality are the key variables influencing the amount of data that can be crowdsourced for each trend, if successfully applied. Indeed, the larger the crowd involved and the higher the crowdsourcing quality and resulting crowdsourcing rate the greater will be the amount of data that can be crowdsourced in a certain amount of time.

The crowd motivation and crowdsourcing costs describe the core conditions that have to be met to set up approaches of each crowdsourcing trend and keep them running. The crowd motiviation describes the preconditions and incentives for a crowdsourcing trend to work, while the crowdsourcing costs discuss the technical and pragmatical requirements that have to be fulfilled to put a crowdsourcing trend into practice.

The detailed comparison relies on the practical experience we accumulated in researching Collect4NLP approaches while also keeping track of the state of the art of the two other trends.

We conclude this section by an overall discussion of the comparable aspects and individual strengths and weaknesses of each of the trends in relation to the others.

4.1. Crowd Motivation

The major factor for any crowdsourcing initiative to be successful is the incentive it provides for a crowd to participate. The three trends we are looking at provide substantially different incentives for participation.

CP. Crowdsourcing platforms attract their crowdworkers by a financial award for each crowdsourcing action that is a small amount of money for each completed HIT (Human Intelligence Task). Poesio et al. (2017) report that rewards are usually fairly small in the range from 0.01 - 0.20 US $ per HIT, depending on the complexity of the task. The higher the award the higher the motivation to participate.

GWAP. GWAP approaches aim to attract a crowd by offering some fun or interesting game-like interaction which at the time of game-playing is collecting data. GWAPs use different gamification features like interaction with other players, leaderboards, speed, level progression and badge systems. This way they aim to attract different types of game players, like *socializers*, *achievers* or *players* (Tondello et al., 2016). The more satisfying or addictive the game experience the higher the motivation to participate.

Collect4NLP. Collect4NLP approaches aim to integrate crowdsourcing activities with a language learning service. Thus the incentive for a crowd of people to participate is their desire or need to improve their language skills. The more effective and engaging the language learning experience the higher the motivation to participate.

4.2. Crowd Size

Concerning crowd size we have to distinguish between the overall size of the crowd that can be targeted by a set of approaches and the effective subset of the target crowd that we might be able to reach and involve for each trend (see Section 4.3 on crowd involvement).

CP. Due to being a paid service and related legal and tax regulations the crowd targeted by crowd platforms is limited to legal adults. For the same reasons, some platforms such as Amazon Mechanical Turk (AMT) require their users to be tax payers of a specific country. Finally, depending on the crowdsourcing task a certain level of language skills or a limitation to speakers of a specific mother tongue might be imposed by the task provider. Overall, CP-based approaches tend to apply stronger selection criteria on their crowd than GWAPs and Collect4NLP which limits the size of the crowd but safeguards crowdsourcing quality (see Section 4.5).

GWAP. The target group for GWAPs comprises theoretically everyone who has access to a computer device. GWAPs particularly target a public that is interested in playing games which is known to be huge and rapidly growing.[9] In 2008, von Ahn and Dabbish (2008) stated that according to a report of the Entertainment Software Association 'more than 200 million hours are spent each day playing computer and video games in the U.S.'. However, it has to be considered that most GWAPs are not comparable to modern video games in terms of their user experience but rather offer a user-friendly task design with some gamification features.[10] As with CPs, for language-related GWAPs the size of the target crowd is also limited by the required language skills of its users though the pre-selection of the crowd is less strict than for CPs.

Two of the most successful GWAPs for creating NLP resources are Jeux-de-mots (Lafourcade and Nathalie, 2020) and Phrase Detectives (Chamberlain et al., 2016). Over six years, more than 2700 active users have created an annotated corpus of 302,224 tokens in Phrase Detectives, and over a period of 13 years around 1.47 million games of Jeux-de-Mots have been played.

Collect4NLP. Similarly to GWAPs, the size of the crowds that can potentially be involved in a Collect4NLP-based effort are enormous as the target

[9] `https://www.researchandmarkets.com/reports/5546908/gaming-market-global-industry-trends-share`

[10] Jurgens and Navigli (2014) observe that 'current games are largely text-based and closely resemble traditional annotation tasks' (see also Section 4.3)

group, in principle, extends to all people interested in learning a language. Indeed, a report from the European Commission (2012) states that 21% of the Europeans aged over 14 years, which amounts to about 90 million people, are actively learning a language. Given that a good share of them should have access to online tools the target group theoretically amounts to several million people.

4.3. Crowd Involvement

The crowd involvement depends on the outreach to recruit a crowd and the duration of their participation to the crowdsourcing effort, also called user retention.

CP. For CP-based approaches, the outreach to a crowd is managed by the platforms themselves. CPs usually have large user bases and effective mechanisms to promote tasks. While the average participation time per week can considerably vary among crowdworkers, studies have proven that the involvement in crowdsourcing activities on CPs is mid-to-long term (more than several months or years) for more than half of the crowdworkers.[11] Given that participation is financially driven, the size of the participating crowd can be adjusted by increasing the provided budget.

GWAP. With respect to the outreach to participants, GWAP-based approaches usually rely on specialized channels of dissemination (e.g. specialized mailing lists) and social media campaigns. The duration of the crowd's participation to GWAP approaches is limited by the time the game offered remains attractive to the users and some GWAPs managed to find some very loyal users (e.g. Poesio et al. (2012)). This attractiveness aspect can be a fairly challenging one to tackle as it requires researchers to formulate and provide linguistic tasks in a joyful manner while competing with immense amounts of games devised primarily for the purpose of entertainment and often with a much higher development and promotional budget. We therefore often observe that even highly elaborated and promoted GWAPs are no longer used or available after some time.

Collect4NLP. With respect to the outreach to a crowd of language learners we have to distinguish two scenarios: outreach to learners for participating in prototypical Collect4NLP learning applications and outreach to learners for using a fully-fledged learning solution that integrates Collect4NLP approaches. The first case relies on promotion campaigns and is comparable to the outreach efforts for GWAPs. The second case would be more comparable to the promotion of CPs as stable programs that provide a specific service to its users. While several efforts of the first type have been created in the past years, until now no large-scale application of Collect4NLP approaches in full-grown or commercial language learning applications exists. As prototypical efforts have shown (cf. e.g., Lyding et al. (2019; Nicolas et al. (2021)) a crowd of users can be successfully attracted for an experiment but this does not guarantee that a substantial part of the crowd will continue being active beyond the duration of an experimental period of a few weeks. This is likely due to the limited learning value any prototypical application can offer. It logically follows from this observation that the involvement of a bigger crowd of learners presupposes the systematic integration of Collect4NLP approaches into a full-grown (and possibly established) language learning platform. Once this can be achieved (see Section 4.6) the outreach to learners and their retention should become very feasible, as the growing business of online language learning solutions shows (see e.g. the growth of DuoLingo[12], Babbel[13] and Busuu[14] over the past years).

4.4. Crowdsourcing Rate

In addition to the size of the crowd that can be reached, the crowdsourcing rate, that is the rate of return for the different crowdsourcing trends, depends on two factors: (1) the ratio between the user's time investment and the data crowdsourced, and (2) the aggregation factor to derive reliable results from crowdsourced data. Accordingly, one hour of activity of a crowd of 20 people can have a greatly different crowdsourcing revenue for each of the different crowdsourcing trends.

CP. CP-based approaches allocate almost all of the user's time to crowdsourcing tasks. Excluded are only some training tasks to prepare the user or occasionally testing to evaluate the reliability of the users or to select a subset of them. Given that training sessions or selection tests are usually unpaid[15] the ratio of crowdsourcing is close to 100%. Also, crowdworkers are selected and evaluated (see Section 4.5), therefore the aggregation factor is expected to be rather low to derive a meaningful result, though task-dependent. The rate of return for CP-based approaches is very high.

GWAP. GWAP-based approaches are comparable to CP-based approaches with respect to both the crowdsourcing ratio and the aggregation factor. The reliability of crowdplayers is difficult to estimate a priori. On the one hand, other than paid crowdworkers (cf. Eickhoff and de Vries (2013)) crowdplayers have no reason to cheat as they do not earn money with the activity. On the other hand, nothing might restrain them from cheating, as players have less to lose in case they would be expelled from the activity. The rate of return for GWAP-based approaches is high.

[11] International Labour Organization (2018) report that '56 per cent of survey respondents had performed crowdwork for more than a year; 29 per cent had crowdworked for more than three years'.

[12] https://www.duolingo.com/

[13] https://www.babbel.com/

[14] https://www.busuu.com/

[15] International Labour Organization (2018) report that 'On average, workers spent 20 minutes on unpaid activities for every hour of paid work, searching for tasks, taking unpaid qualification tests, researching clients to mitigate fraud and writing reviews.

Collect4NLP. Collect4NLP-based approaches need to work on top of a proper language learning service as the language learning offer is the incentive for the crowd of users to participate. As such, learning services need to ensure a reliable feedback for most tasks they send to their users while they can only crowdsource data from users for a smaller fraction of the tasks. This only allows for a very low crowdsourcing ratio of ideally less than 10% which could be increased by intelligent strategies for deriving meaningful feedback from less reliable source data. Also, language learners are expected to be less reliable than both (mother tongue) crowdworkers and crowdplayers which requires a higher aggregation factor and leads to a lower crowdsourcing rate for Collect4NLP-based approaches as compared to CP and GWAP.

4.5. Crowdsourcing Quality

The quality of NLP data collected by any crowdsourcing trend depends on the linguistic expertise of its crowd as well as on the performance profiling of each member of the crowd. The estimated proficiency and performance of the crowd will determine the aggregation factor and thus in return strongly impact the crowdsourcing rate as described above.

CP. CP-based approaches usually target L1 speakers or proficient L2 speakers. In addition, often pre-tests or intermediate testing is performed to identify and exclude low-performing crowdworkers.

GWAP. GWAP-based approaches also usually target proficient L1 or L2 speakers. They sometimes request an initial training phase to learn how the GWAP works (cf. e.g. Fort et al. (2020), Chamberlain et al. (2016)) but rarely exclude participants through pre-testing .

Collect4NLP. Collect4NLP-based approaches target language learners which are typically composed of lesser proficient L1 speakers improving their mother tongues and mostly L2 speakers learning foreign languages. As such, the overall linguistic expertise of the crowds targeted by Collect4NLP approaches can greatly vary, and requires the continuous profiling of the performances of their participants in order to give different weight to the answers of different learners when aggregating the data crowdsourced from them.

4.6. Crowdsourcing Costs

One of the major advantages of crowdsourcing for data collection is the expectedly lower cost as compared to traditional contractual work. Still costs occur for any crowdsourcing initiative to be set up and kept running.

CPs. On the one hand, CP-based approaches come with costs for paying each HIT performed by the crowdworkers. On the other hand, infrastructure costs for setting up a CP-based crowdsourcing activity are very low. Crowdsourcing platforms have been around for almost two decades now (e.g. the AMT was launched in 2005) and have received a great deal of attention ever since. Accordingly, a rather diversified set of platforms with varying characteristics have been developed, tested and used. Therefore, there exist clear solutions to define tasks and process the data crowdsourced that anybody can rely on.

GWAP. In GWAP-based approaches no costs have to be foreseen for paying the crowd, however compared to CP-based approaches less ready-to-use infrastructure is available which leads to higher development costs. GWAP-approaches can rely on a number of freely available code repositories and libraries such as PythonGameLibraries[16] in order to implement the gaming aspects of their approaches. Also, a growing number of previous efforts such as Fort et al. (2020) and Guillaume et al. (2016b) have made their code freely available. However, as different GWAPs usually target very specific and varying crowdsourcing tasks and require different solutions for the processing of the data crowdsourced, the creation of any new GWAP comes with considerable development costs[17].

Collect4NLP. As a recent research trend, Collect4NLP has no generic reference code repositories or libraries to rely on at present, even though code repositories of several prototypes such as in Lyding et al. (2019) and Araneta et al. (2020) are made freely available. This means that the development of Collect4NLP approaches is open-ended and challenging and corresponding development costs are currently still high. In addition, even though the automatic generation of language learning exercises has been researched in numerous CALL efforts, most past efforts have primarily focused on textual data in which part of a textual content is removed and learners are asked to fill a gap (cf. e.g. Knoop and Wilske (2013; Lee et al. (2019)) but not on a wider variety of different types of content provided by NLP resources (e.g. lexical networks). As such, also the generation of exercises is an open research challenge that comes with considerable costs.

4.7. Which Trend to Favor over the Others?

The detailed discussion of the three trends of approaches demonstrates that each trend has its particular strengths and none of them truly dominates the other two. When undertaking a new initiative to crowdsource NLP datasets, and more generally speaking when envisioning the future of crowdsourcing NLP datasets we should therefore carefully consider what each trend has to offer, and which investments it requires.

CP-based approaches are the most established and most reliable solution, be it in terms of crowd involvement, crowdsourcing rate or crowdsourcing quality. They are the most seamless way to start a new crowdsourcing project as they rely on established service infrastructures (e.g. AMT) and the crowd size is easily scalable as long as financial means are available to pay for the

[16] `https://wiki.python.org/moin/PythonGameLibraries`

[17] Poesio et al. (2013) indicate a total of around 100,000 US $ of development costs for Phrase Detectives which allowed to annoatate 162,000 complete tokens in three years.

crowdworkers. Leaving ethical issues aside (Fort et al. (2011), Schmidt (2013)), the major drawback of CP-based approaches are their continuous costs.

GWAPs and Collect4NLP-based approaches have the potential to greatly reduce costs in the long term, as the crowd is participating due to motivations other than financial ones. However, the challenge of these approaches is to satisfy the expectations of the crowd, be it in terms of fun in playing games or progress in learning a language. For GWAPs this means creating games that live up to today's gaming standards, while for Collect4NLP-based approaches an integration with existing effective language learning solutions would be desirable. Creating effective solutions requires considerable research and development efforts. As mentioned above, little programming frameworks and tools for the creation of GWAP-based and Collect4NLP-based approaches exist, and for Collect4NLP also the mechanisms for generating exercises and aggregating potentially flawed learner responses still have to be explored and defined. If these challenges can be overcome, for both trends large crowds could be involved. This would also allow to make up for the expected lower crowdsourcing rate and lower crowdsourcing quality, in particular in relation to Collect4NLP-based approaches.

We conclude that for achieving short-term results of reasonable scope CP-based approaches are the safest and most economical choice. At the same time, we see a strong need for advancing research and development efforts on GWAP-based and Collect4NLP-based approaches in order to work towards sustainable solutions in the long term, provided that such approaches could rely on an immense crowd of unpaid contributors and thus bear a much greater and ethically less problematic potential to advance NLP resource creation.

5. Applicability of the Approach for Different Types of NLP Resources

In order to demonstrate how the Collect4NLP approach could be applied to different types of NLP resources we started by looking for a reference set of common language resource types. After some searching we ended up at the CLARIN Resource Families (CRF) (Fišer et al., 2018), which we decided to be a suitable reference set. The CRF are a manually curated set of collections of linguistic resources (and tools, but those are not relevant for our approach) grouped into so-called families. They provide an overview of the resources available in the CLARIN infrastructure and beyond and thus constitute a de facto standard of the current state-of-the-art of NLP resources in Europe. They have been very popular with researchers, because they adhere to certain quality standards and come with brief descriptions and the most important metadata, such as resource size, text sources, time periods, annotations and licences as well as links to download pages or concordancers.

The CRF distinguish three coarse groups of resources: corpora, lexical resources and tools, of which the first two are relevant for our case. The groups are subdivided into more fine grained categories as displayed for corpora and lexical resources in Table 1

Corpora	Lexical resources
Computer-mediated communication corpora	Lexica
Corpora of academic texts	Dictionaries
Historical corpora	Conceptual Resources
L2 learner corpora	Glossaries
Literary corpora	Wordlists
Manually annotated corpora	
Multimodal corpora	
Newspaper corpora	
Parallel corpora	
Parliamentary corpora	
Reference corpora	
Spoken corpora	

Table 1: NLP resources in the CRF

5.1. Tasks in NLP Resource Collection

The corpora and lexical resources listed among the CRF cover a wide range of datasets. They differ both in terms of their type of content as well as, and partly related to it, in their basic data characteristics and the annotation layers applied to them.

For corpora the most prevailing characteristics are contemporary and written data and the most prevalent annotations are basic processing including:

- tokenisation
- lemmatisation
- PoS/MSD-tagging
- syntactic parsing (partly).

For lexical resources the following data entries and annotations are most common:

- lemmas
- word forms
- basic morphological information
- semantic relations
- usage examples.

These characteristics translate into a set of tasks for creating or curating NLP resources, such as *'detecting word boundaries'*, *'assigning grammatical categories to words'*, *'linking words by semantic relations'*, *'creating word definitions'*, etc.

These tasks are usually carried out intentionally by experts or instructed laymen in order to create NLP resources. However, we claim that the Collect4NLP approach allows to shift part of these tasks to language learners. This requires to provide them with a meaningful learning exercise, whose completion produces the required type of data as a side effect.

Exercise	Annotation
’Odd-one out’	semantic relation
Synonyms	semantic relation
Antonyms	semantic relation
Forming word groups	semantic relation
Identify words	headword selection
Assign grammar category	part-of-speech tagging
Filling the gap	part-of-speech tagging

Table 2: Language exercises and related annotations

In the following subsections we first look into common language learning exercises, and second outline a number of blueprints for how language learning exercises can be combined with NLP resource creation tasks.

5.2. Relevant Exercises for Collect4NLP

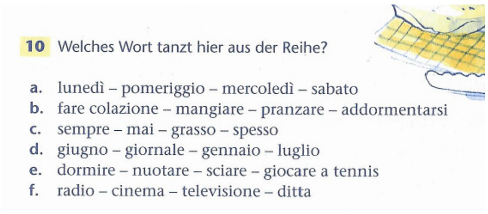
To get an overview of the types of exercises that are commonly used in language learning we investigated a number of language learning books. Assuming that established exercise types are effective and meaningful for language learning we created a list of those exercises that could serve for crowdsourcing purposes. The resulting collection (see Annex A) is, obviously, not an exhaustive list, but we tried to get a more diverse sample by looking at books from various publishers, courses for various languages and various types of books (e.g. full language course, exercise book).
While looking at the exercises we kept in mind which annotation tasks we aim to complete and grouped the exercises accordingly (see Table 2).
Hereafter, we explain for a number of exercises how they can be used to generate or correct NLP resources.

5.3. Blueprints

The purpose of the following blueprints is to provide examples of pairs of NLP resources and language learning exercises and to show how they can be combined to both serve a language learning need and to crowdsource NLP data. The full list of blueprints is found in a related technical report[18].

5.3.1. ’Odd-One Out’ and ’Find the Companion’

One classic exercise type consists of a list of words of which the learner has to select the ’odd-one out’. The exercise is designed such that all words apart from one have the same semantic property (e.g. *’days of the week’*). In order to correctly identify the one word that is different the learner has to understand the meaning of the different words. Figure 1 gives an example of this type of exercise for learners of Italian.
The NLP resources that such an exercise could be linked to are conceptual lexica or wordnets (see CRF[19]) that encode hyponymy relations. For example, ’bulldog’, ’labrador’ and ’poodle’ are all hyponyms of the hypernym ’dog breed’, while ’sparrow’ is not.

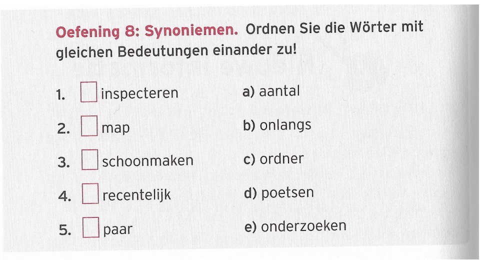


10 Welches Wort tanzt hier aus der Reihe?

a. lunedì – pomeriggio – mercoledì – sabato
b. fare colazione – mangiare – pranzare – addormentarsi
c. sempre – mai – grasso – spesso
d. giugno – giornale – gennaio – luglio
e. dormire – nuotare – sciare – giocare a tennis
f. radio – cinema – televisione – ditta

Figure 1: Find the ’odd-one out’

Blueprint: A language learning exercise could automatically be generated from a conceptual network by extracting several hyponyms of any relevant hypernym[20] and by putting any word that is not among the hyponyms in the middle. When used in language learning the answers given by a number of learners can be used to verify or discard some of the hyponym relations encoded in the resource.
A similar, but slightly different exercise type is the ’find the companion’ scenario. It provides a list of words to the learner among which they have to match those pairs of words that have the same meaning. Figure 2 shows an example of this type of exercise for Dutch.



Oefening 8: Synoniemen. Ordnen Sie die Wörter mit gleichen Bedeutungen einander zu!

1.	□ inspecteren	a) aantal
2.	□ map	b) onlangs
3.	□ schoonmaken	c) ordner
4.	□ recentelijk	d) poetsen
5.	□ paar	e) onderzoeken

Figure 2: Find the ’companion’

Also here, the NLP resources that such an exercise could be linked to are conceptual lexica or wordnets given that they encode synonymy relations.
Blueprint: A number of synonyms are taken from the wordnet together with words that are suspected to have a similar meaning. Students are presented with the words and have to match the synonyms. If the suspected synonyms are matched a lot of the time this can be taken to mean they are indeed synonyms.

5.3.2. Identify Words in String

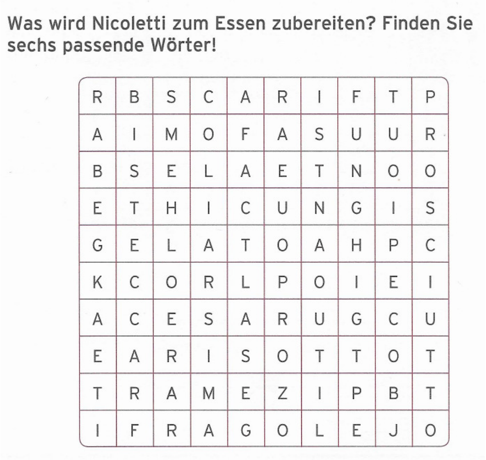
Another common type of exercise asks the learners to identify existing words within the target language. Commonly this exercise takes the form of a word grid where the student has to find a certain number of words, see Figure 3 for an example for Italian. Another way is to present the learner with just a very long string that

[18] see `https://enetcollect.net/ilias/goto.php?target=file_1214_download&client_id=enetcollect` for a more comprehensive list

[19] `https://clarin.eu/resource-families/lexical-resources-conceptual-resources`

[20] Relevance will be determined in relation to the learners’ level and their learning target, e.g. vocabulary acquisition related to *’food and cooking’*.

contains the words to be identified, see Figure 4 for an example for Dutch.

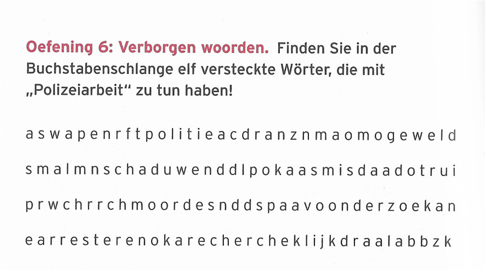


Was wird Nicoletti zum Essen zubereiten? Finden Sie sechs passende Wörter!

R	B	S	C	A	R	I	F	T	P
A	I	M	O	F	A	S	U	U	R
B	S	E	L	A	E	T	N	O	O
E	T	H	I	C	U	N	G	I	S
G	E	L	A	T	O	A	H	P	C
K	C	O	R	L	P	O	I	E	I
A	C	E	S	A	R	U	G	C	U
E	A	R	I	S	O	T	T	O	T
T	R	A	M	E	Z	I	P	B	T
I	F	R	A	G	O	L	E	J	O

Figure 3: Find words in a grid

As can be seen from the two examples provided, often these exercises constrain the possible words by explicitly stating a semantic domain that they should belong to as in Figure 3, but that is not always the case.



Oefening 6: Verborgen woorden. Finden Sie in der Buchstabenschlange elf versteckte Wörter, die mit „Polizeiarbeit" zu tun haben!

aswapenrftpolitieacdranznmaomogeweld

smalmnschaduwenddlpokaasmisdaadotrui

prwchrrchmoordesnddspaavoonderzoekan

earresterenokarecherchekl ijkdraalabbzk

Figure 4: Find words in a string

Such a type of exercise could be used on an annotated text corpus that has been tokenized/lemmatized automatically. The learner input can confirm the results produced by the NLP pipeline by ensuring that it has detected the right words or word forms. For such a setup, one would need the more generic exercises like in figure 4 and not the domain-specific ones. Likewise this exercise type could also be linked again to ”word-based” NLP resources like dictionaries and lexica and could help, for example, to confirm possible neologisms that have been pre-identified by an NLP pipeline. If a certain number of learners find these words in the exercise, it can be assumed that they are actual words.

Blueprint: A number of potential words are taken from a) an annotated corpus, which has been tokenized/lemmatized or b) a dictionary where they have been added as neologisms by an NLP pipeline. These possible words are then inserted together with a larger number of ”confirmed” words into a word grid or a long string. If most learners also pick out the ”new” words they can be considered as actual words.

6. Conclusions and Future Work

We discussed in this article the applicability of the recent trend of Collect4NLP-based crowdsourcing approaches by comparing it to CP-based and GWAP-based approaches with respect to several key aspects and by outlining a first set of blueprints for combining NLP resources with language learning exercises. Our conclusion regarding the relevance and viability of Collect4NLP-based approaches is that they have noticeable advantages over the other two trends with respect to the crowd motivation and accordingly crowd size and crowdsourcing costs. Also, the reported efforts to match language learning exercises with types of NLP datasets suggest that Collect4NLP-based approaches are indeed applicable to several popular types of datasets registered within the CRF. Both analyses indicate the high potential of Collect4NLP-approaches for large-scale and sustainable crowdsourcing efforts, both concerning the crowdsourcing potential as well as concerning the needs of NLP stakeholders curating datasets. At the same time, this new trend also comes with demanding challenges related to researching its mechanisms and integrating them into language learning solutions. In addition to its overall novelty, this may explain why Collect4NLP-approaches have been less researched so far compared to the other two trends.

In terms of future works, as next steps we will discuss our current conclusions on the comparisons discussed in Section 4 with experts in CP- and GWAP-based approaches. Indeed, as our main research expertise lies with Collect4NLP-based approaches our overall vision of the three trends might be biased to some extent and deserves continuous exchange and confrontation with experts of the related communities.

With respect to advancing Collect4NLP-based approaches, our next steps will focus on extending the list of blueprints matching language learning exercises with types of datasets that could be crowdsourced as discussed in Section 5. We foresee to study more textbooks for a wider set of source and target languages, possibly also extending over non-European languages with the intuition that we will encounter other exercises that could be linked to a type of dataset we have not considered yet. With a similar reasoning in mind, we will also explore the types of exercises provided by language learning apps. Last but not least, we intend to perform a finer grained comparison between the types of datasets targeted by previous efforts implementing CP- and GWAP-based approaches in order to evaluate how Collect4NLP-based approaches compare to the other two approaches in relation to dataset coverage.

7. Acknowledgements

This work is based on a virtual mobility grant of COST Action enetCollect (CA16105), supported by COST (European Cooperation in Science and Technology).

8. Bibliographical References

Araneta, M. G., Eryiğit, G., König, A., Lee, J.-U., Luís, A., Lyding, V., Nicolas, L., Rodosthenous, C., and Sangati, F. (2020). Substituto – a synchronous educational language game for simultaneous teaching and crowdsourcing. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Gothenburg, Sweden, November. LiU Electronic Press.

Arhar Holdt, Š., Logar, N., Pori, E., and Kosem, I. (2021). "Game of Words": Play the Game, Clean the Database. In *Proceedings of the 14th Congress of the European Association for Lexicography (EURALEX 2021)*, pages 41–49, Alexandroupolis, Greece, July.

Biemann, C. (2013). Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122, Mar.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics 08)*, pages 42–49.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2039–2046, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Cieri, C., Fiumara, J., and Wright, J. (2021). Using games to augment corpora for language recognition and confusability. In *Proc. Interspeech 2021*, pages 1887–1891.

Eickhoff, C. and de Vries, A. P. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16:121–137.

Eryiğit, G., Şentaş, A., and Monti, J. (2022). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, page 1–33.

European Commission, D.-G. f. C. (2012). Europeans and their languages. Special eurobarometer 386 report, Survey conducted by TNS Opino & Social, and co-ordinated by the European Commission.

Evanini, K., Higgins, D., and Zechner, K. (2010). Using amazon mechanical turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 53–56. Association for Computational Linguistics.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

Fišer, D., Lenardič, J., and Erjavec, T. (2018). CLARIN's Key Resource Families. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Fort, K., Adda, G., and Cohen, K. B. (2011). Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, June.

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6. ACM.

Fort, K., Guillaume, B., Pilatte, Y.-A., Constant, M., and Lefèbvre, N. (2020). Rigor mortis: Annotating MWEs with a gamified platform. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4395–4401, Marseille, France, May. European Language Resources Association.

Funk, C., Tseng, M., Rajakumar, R., and Ha, L. (2018). Community-driven crowdsourcing: Data collection with local developers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ganbold, A., Chagnaa, A., and Bella, G. (2018). Using crowd agreement for wordnet localization. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*.

Greene, C., Keogh, K., Koller, T., Wagner, J., Ward, M., and Genabith, J. (2004). Using nlp technology in call. In *Proceedings of the InSTIL/ICALL Symposium on Computer Assisted Learning 2004*.

Guillaume, B., Fort, K., and Lefebvre, N. (2016a). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Guillaume, B., Fort, K., and Lefebvre, N. (2016b). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3041–3052, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Hladká, B., Hana, J., and Lukšová, I. (2014). Crowdsourcing in language classes can help natural language processing. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Howe, J. et al. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

International Labour Organization, I. (2018). Digital labour platforms and the future of work: Towards decent work in the online world. Executive summary, Survey of working conditions conducted by the International Labour Organization, Geneva, Switzerland.

Jurgens, D. and Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

Knoop, S. and Wilske, S. (2013). Wordgap - automatic generation of gap-filling vocabulary exercises for mobile learning. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, pages 39–47, Oslo, Norway, May. Linköping Electronic Conference Proceedings, NEALT Proceedings Series.

Lafourcade, M. and Nathalie, L. B. (2020). Game design evaluation of GWAPs for collecting word associations. In *Workshop on Games and Natural Language Processing*, pages 26–33, Marseille, France, May. European Language Resources Association.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 71–79. Association for Computational Linguistics.

Lee, J.-U., Schwan, E., and Meyer, C. M. (2019). Manipulating the difficulty of C-tests. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 360–370, Florence, Italy, July. Association for Computational Linguistics.

Lyding, V., Rodosthenous, C., Sangati, F., ul Hassan, U., Nicolas, L., König, A., Horbacauskiene, J., and Katinskaia, A. (2019). v-trel: Vocabulary trainer for tracing word relations - an implicit crowdsourcing approach. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 674–683, Varna, Bulgaria. INCOMA Ltd.

Madge, C., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2017). Experiment-driven development of a gwap for marking segments in text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, pages 397–404. ACM.

Millour, A., Araneta, M. G., Lazić Konjik, I., Raffone, A., Pilatte, Y.-A., and Fort, K. (2019). Katana and Grand Guru: a Game of the Lost Words (DEMO). In *Proceedings of the ninth Language & Technology Conference*, Poznan, Poland, May.

Mollberg, D. E., Jónsson, Ó. H., orsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gunason, J. (2020). Samrómur: Crowd-sourcing data collection for icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Nicolas, L., Lyding, V., Borg, C., Forăscu, C., Fort, K., Zdravkova, K., Kosem, I., Čibej, J., Holdt, Š. A., Millour, A., et al. (2020). Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 268–278.

Nicolas, L., Aparaschivei, L. N., Lyding, V., Rodosthenous, C., Sangati, F., König, A., and Forascu, C. (2021). An experiment on implicitly crowdsourcing expert knowledge about Romanian synonyms from language learners. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 1–14, Online, May. LiU Electronic Press.

Pilán, Ildikó, E. V. and Johansson, R. (2013). Automatic selection of suitable sentences for language learning exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*. Dublin: Research-publishing.net.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2012). The phrase detective multilingual corpus, release 0.1. In *Collaborative Resource Development and Delivery Workshop Programme*, page 34.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April.

Poesio, M., Chamberlain, J., and Kruschwitz, U., (2017). *Crowdsourcing*, pages 277–295. Springer, Dordrecht, 06.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Rodosthenous, C. and Michael, L. (2016). A Hybrid

Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium (STAIRS 2016)*, volume 284, pages 111–122. IOS Press, August.

Rodosthenous, C. T., Lyding, V., König, A., Horbacauskiene, J., Katinskaia, A., ul Hassan, U., Isaak, N., Sangati, F., and Nicolas, L. (2019). Designing a prototype architecture for crowdsourcing language resources. In Thierry Declerck et al., editors, *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK 2019), Leipzig, Germany, May 21, 2019*, volume 2402 of *CEUR Workshop Proceedings*, pages 17–23. CEUR-WS.org.

Rodosthenous, C., Lyding, V., Sangati, F., König, A., ul Hassan, U., Nicolas, L., Horbacauskiene, J., Katinskaia, A., and Aparaschivei, L. (2020). Using crowdsourced exercises for vocabulary training to expand conceptnet. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 307–316.

Sangati, F., Merlo, S., and Moretti, G. (2015). School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE*, pages 16–19.

Schmidt, F. A. (2013). The good, the bad and the ugly: Why crowdsourcing needs ethics. In *2013 International Conference on Cloud and Green Computing*, pages 531–535.

Smrz, P. (2019). Crowdsourcing Complex Associations among Words by Means of A Game. In *Proceedings of CSTY 2019, 5th International Conference on Computer Science and Information Technology.*, volume 9, Dubai, UAE, December.

Tachibana, R. and Komachi, M. (2016). Analysis of english spelling errors in a word-typing game. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 385–390.

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., and Nacke, L. E. (2016). The gamification user types hexad scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, page 229–243, New York, NY, USA. Association for Computing Machinery.

von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM*, 51(8):58–67, aug.

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*, pages 1–2. ACM.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics.

A. Appendix

Here we will provide a short list of all the exercise types we identified. For more context, including the accompanying blueprints and example pictures we refer to the related technical report[21].

"Odd one out"
Exercise: Students are presented with a list of words with the same semantic property (e.g. days of the week). They have to pick the one word that does not belong with the others, the "odd one out".

Relation: At location
Exercise: The student is presented with pictures of a number of things that belong to a certain location. For example *"Which of these things can be bought in which kind of store?"* The student has to match the product to the store. Or: *"Which of these types of furniture can be found in which room?"* The student has to match the items of furniture to the rooms.

Labelling, text-retrieval
Exercise: *"What are these ads about?"* Students are presented with short texts that they need to connect to a term that is most likely the topic of the text.

Definitions
Exercise: *"Write down the terms for these definitions."* Students are shown short definitions and have to provide the term that is described.

Collocations
Exercise: *"Connect these fragments to form collocations."* Students are presented with a number of typical collocations (e.g. "a flock of sheep"), but they are broken apart and shuffled. The students need to connect the parts to form real collocations.

Gender
Exercise: *"Fill in the correct adjective in the correct form."* Students are presented with a sentence missing an adjective. The adjectives are provided in their base forms. Students have to match the adjective to the right sentence and make sure that it has the right form to agree with the corresponding noun.

Antonyms
Exercise: *"Write down the opposite."* Students are presented with a number of words and have to provide the opposite.

Generic Relations
Exercise: *"Which words belong together?"* Students are presented with a number of words and have to match them into pairs.

[21] see `https://enetcollect.net/ilias/goto.php?target=file_1214_download&client_id=enetcollect` for a more comprehensive list

Synonyms or ”find the companion”
Exercise: *”Match the words that mean the same.”* Students are presented with a number of words and have to match the ones that have the same meaning.

Identify words
Exercise: *”Find all the words.”* Students are presented with a long string of letters or a word grid in which they have to identify a number of words all related to a specific topic.

Orthography
Exercise: *”Read the text and mark all the orthographic mistakes.”* Students are presented with a text and have to mark all orthographic mistakes they can spot.

Grammar
Exercise: *”Check the sentences that contain grammatical errors.”* Students are presented with a number of sentences and have to mark the ones that contain a grammatical error.