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.

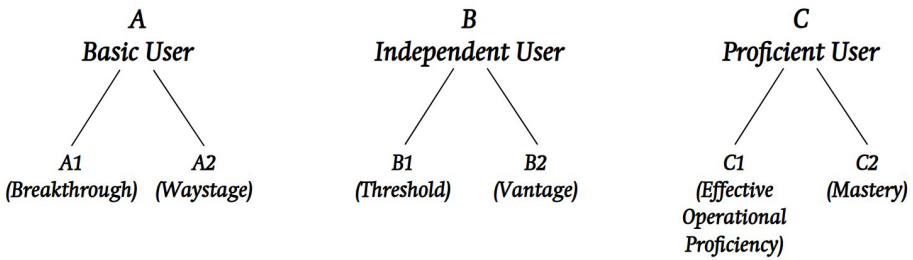

Figure 2: The "Common Reference Levels" as defined by (Council of Europe, 2001).

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

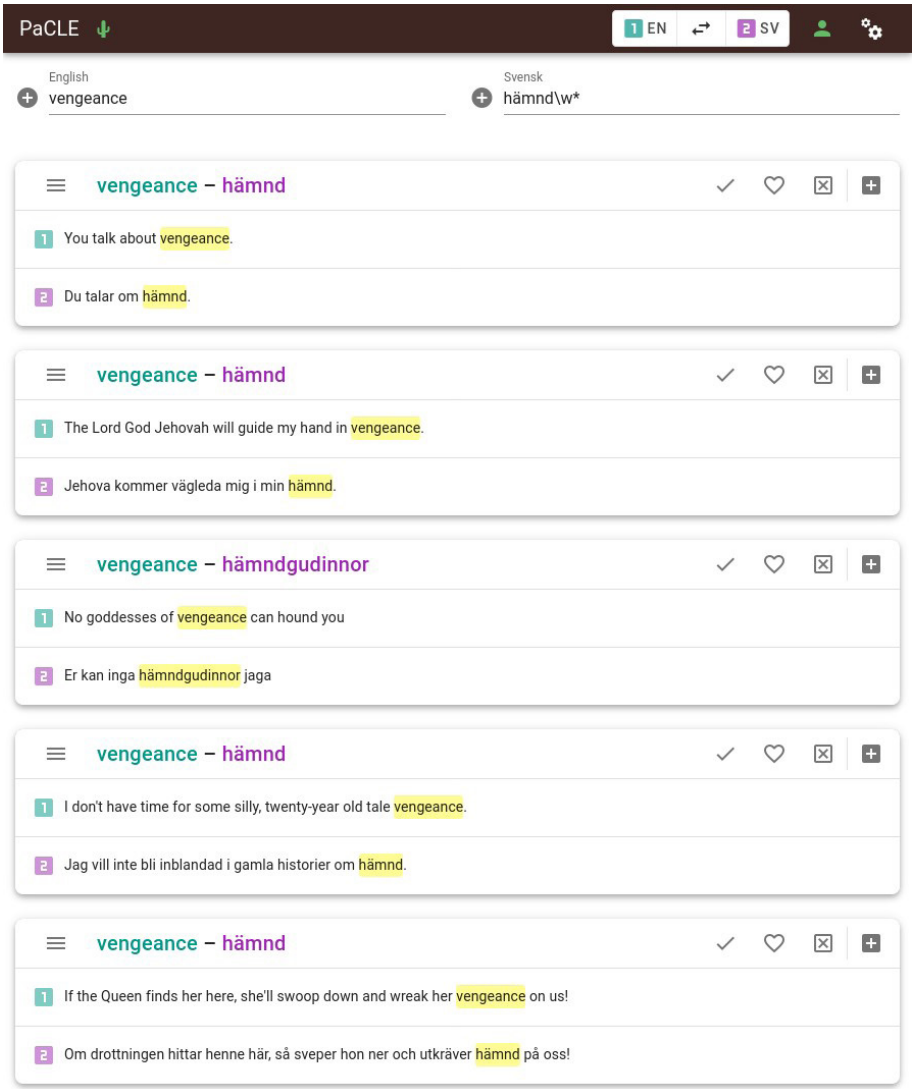

Figure 3: The PaCLE application showing five examples for a parallel corpus search in the English-Swedish part of OpenSubtitles. Matching parts are highlighted. The use of advanced regular expressions is supported.

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

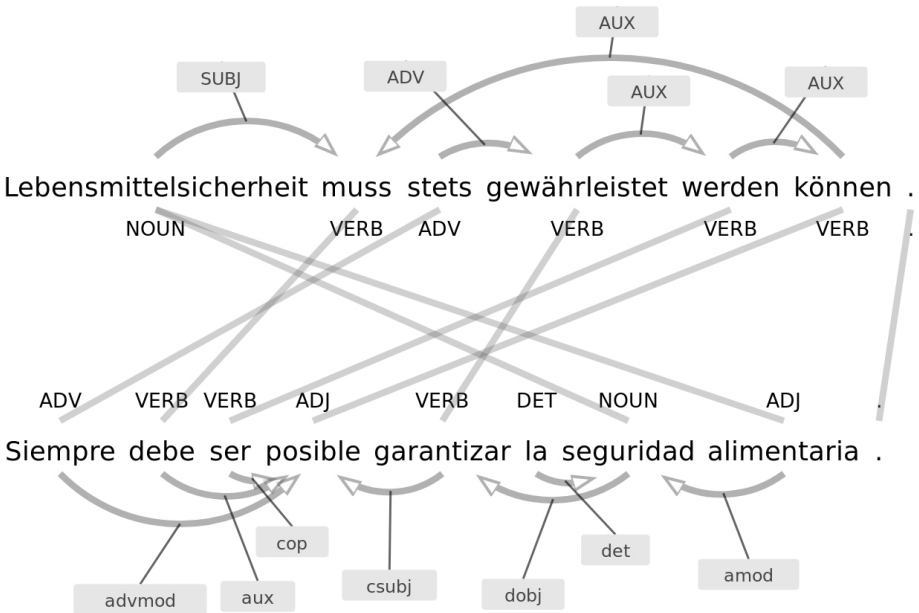

Figure 4: Sentence pair in German and English with different syntactic structures, which is highlighted by the heavily crossing alignment links. Here, language-dependent label sets have been used instead of Universal Dependencies.

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