Bulletin of the *Transilvania* University of Braşov
Series IV: Philology and Cultural Studies • Vol. 14(63) No. 2 – 2021
https://doi.org/10.31926/but.pcs.2021.63.14.2.3

Automatic generation of language exercises based on a universal methodology: An analysis of possibilities

Vanja SLAVUJ[1], Lucia NACINOVIC PRSKALO[2], Marija BRKIC BAKARIC[3]

The aim of the paper is to examine the possibilities for automatic generation of language learning exercises and compare them to those manually compiled by language instructors. The paper first presents a universal methodology applied in manually created exercises for learning language for specific purposes, elaborated with examples in the field of academic English. Next, the automation of the procedure is explored through a series of steps which include creating the corpus, analysing each exercise type and the possibility of its automatic generation, automatically generating the exercise, and evaluating the end result. The results of the evaluation suggest that automatic generation of exercises can serve as a preliminary step of a two-stage process of exercises development in which each exercise, however, needs additional approval from the language expert.

Keywords: *language for specific purposes, universal methodology, automatic exercise generation, language learning*

1. Introduction

This paper first presents the methodology behind the creation of language learning exercises within the LanGuide language learning mobile application, currently under development as part of the EU-funded project titled LanGuide, and next explores the possibilities of automatic generation of language exercises taking the methodology as its starting point.

The LanGuide project gathers experts from the fields of linguistics, first (L1) and second (L2) language teaching, and computer science from five European countries and six universities, and aims at creating an open access language learning tool. Such a tool is specifically designed and organised as a distance

[1] University of Rijeka, Croatia, vslavuj@uniri.hr
[2] University of Rijeka, Croatia, lnacinovic@uniri.hr
[3] University of Rijeka, Croatia, mbrkic@uniri.hr

learning tool for improving language skills of different stakeholders at the tertiary level of education (Kompara Lukančič and Fabijanić 2020, 37). In addition to covering English for specific purposes (ESP), the tool is planned to include the possibility of learning at least the basics of another five languages of the project partners, namely Croatian, Romanian, Slovene, Spanish and Swedish.

The LanGuide guidance tool generates a series of pre-prepared language exercises to a language learner while targeting a specific language skill or category. These are similar to exercises and tasks one often finds on language tests. Language lexical tests, among others, have a crucial role in the process of learning languages for specific purposes (LSP). Creating language exercises manually is extremely time-consuming and expensive. The main motivation for this work is the question of how to provide enough exercises at each of the three levels of language proficiency and various learning personas singled out within the project LanGuide, in order to ensure that learners are provided with a different set of examples whenever they use the language guidance mobile application. Existing applications for language learning, such as Duolingo, are based on a human generated list of sentences and/or texts. Therefore, the main question explored within this research is whether it is possible to use a theoretically unlimited source of real examples of language use for creating exercises.

Section 2 of this paper presents related work with details on various approaches to automatically creating language exercises. Details of the LanGuide approach to language learning using the mobile application and, more specifically, details of creating language learning exercises for it are given in the first part of Section 3. The second part of Section 3 introduces and presents the methodology applied in the process of generating exercises automatically. Evaluation of automatically created exercises is given in Section 4. The main findings and concluding remarks are briefly summarized in the last section of the paper.

2. Related work

As previously suggested, the LanGuide tool is a mobile application (or an m-learning application) that utilizes the latest developments in mobile phone and/or smartphone technology, namely larger screen size with higher resolution, stronger processing power, multimedia opportunities, and ease of access to the global network (Bateson and Daniels 2012, 137), to deliver a distance learning experience. When learning at a distance, the majority of the learning process is done outside traditional classroom environments and with the lack of immediate presence of the language teacher by employing the capabilities of different digital technologies

(Lamy 2013, 144). Furthermore, mobile-assisted learning or m-learning brings additional flexibility regarding the place, time and access opportunities of language learning (Glenn Stockwell 2013, 202; Taki and Amini 2017, 61) as well as its almost seamless integration into our daily lives (Bax 2003, 25). Taki and Amini (2017, 59) suggest that such applications may represent an effective way of language learning as they allow for a personal and learner-centred way for language learning.

However, creating language exercises manually is extremely time-consuming, demanding, and expensive. Therefore, various methodologies for creating exercises automatically have been presented throughout the last two decades. The resulting systems can be categorized by the languages they support, targeted aspects of learning (e.g. grammar-oriented, vocabulary-oriented, etc.), types of exercises implemented, external linguistic resources they use (e.g. WordNet, word lists, dictionaries), different natural language processing (NLP) methods that are implemented, etc.

The following subsections describe some of the most commonly used methods from the field of natural language processing, examples of their implementation in automatic exercise generation, and some additional resources that can be used for this purpose.

2.1. Corpora

According to Bennett (2010, 2), a corpus is “a large, principled collection of naturally occurring examples of language stored electronically”. Since corpora provide rich models of language in terms of lexical, grammatical, and morphological features, collocation patterns, semantic features, etc., they are used by various groups such as linguists, social scientists, humanities scholars, lexicographers, natural language processing experts, and so on. In the recent years, corpora have also been used in language teaching (Volodina 2008, 31-32), as they allow customization according to learners' needs or course requirements and offer the possibility of generating teaching materials and exercises automatically.

There are different types and categories of corpora. For example, a monolingual corpus contains text in only one language. It can be used for various tasks, such as checking the correct usage of a word, identifying common patterns, finding the most natural word combinations, etc. Fenogenova and Kuzmenko (2016, 22) use it in combination with the Pearson’s Academic collocation list for creating five different types of lexical exercises aimed at learning academic collocations. The authors conclude that the quality of generated exercises is heavily dependent on the corpora used for their creation. The evaluation of the generated exercises, suggested by Fenogenova and Kuzmenko (2016, 25), consists of analysing

the percentage of correct answers and maximum scores per each exercise type. Moreover, the distribution of answers in multiple choice exercises reveals which choices are too easy, which are not appropriate, and which are possibly interchangeable. Bick (2005) uses corpora in different languages for automatic exercise generation in grammar.

A parallel corpus consists of two or more monolingual corpora of different languages. The languages have to be aligned and the translations of the corresponding segments have to be matched. The most obvious application of parallel corpora is in the field of machine translation, but they can also be used for automatic generation of language exercises. For example, Zanetti, Volodina, and Graën (2020, 62) apply methods for selecting example pairs from a large parallel corpus of movie subtitles in order to generate exercises which involve unscrambling sentences. Since this type of exercise can result in multiple correct sentences, the authors suggest complementing each sentence by the equivalent sentence in another language, thus narrowing down the number of correct answers. The manual evaluation is conducted by assessing whether the sentence is appropriate for the purpose, whether it contains sensitive vocabulary, whether it is sufficiently context independent to be used for an exercise, and, finally, whether the sentence pair is a good translation.

Depending on the subject area, domain, and topics they cover, corpora can be categorized as general or specialized. While general corpora, such as the 'British National Corpus' (BNC Consortium 2007), consist of general texts, specialized corpora contain texts restricted to a specific field, domain or topic. An example of a specialized corpus is the 'Michigan Corpus of Academic Spoken English' (Simpson et al. 2002), which contains spoken language focusing on contemporary university speech.

Most systems for automatic generation of exercises use different types of corpora. While Fenogenova and Kuzmenko (2016, 22) use well-known existing corpora, others allow uploading user-created material such as text segments (e.g. Perez and Cuadros (2017, 49) and Malafeev (2015, 442)).

2.2. Part-of-speech tagging

Part-of-speech (POS) tagging "refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context" (Pykes 2020). An example of a tagged sentence is given in (1). The corresponding lexical term and its tag are given under each token in the sentence.

(1)

It	is	a	beautiful	day	.
pronoun	Verb	Determiner	Adjective	Noun	Punctuation mark
(PRP)	(VBZ)	(DT)	(JJ)	(NN)	

A set of all POS tags forms a tagset. They differ for different languages. Tagsets can contain different levels of detail: they may contain only basic tags for the most common parts of speech (e.g. N for noun, V for verb, etc.) or they may contain tags that reveal more detail and distinguish between nouns in singular and plural, verbal conjugations, tenses, aspect, and so on.

Since the size of modern corpora is typically very large, automatic annotation of POS tags is usually performed. The automatic systems for annotating POS tags are called POS taggers. The availability of POS taggers for different languages varies. The accuracy of a tagger usually depends on the level of detail of the POS tags in a tagset. Also, taggers that annotate only the most common word types usually have high accuracy. For example, the accuracy of a well-known POS tagger for English is over 97% (Manning 2011, 1).

POS tagging can be useful in various linguistic tasks, e.g. word sense disambiguation, Named Entity Recognition (NER), sentiment analysis, question answering, etc. The application of POS tagging in automatic exercise generation is less obvious, but can be quite useful.

One of the possible applications of POS tagging in automatic exercise generation is the generation of appropriate distractors in multiple choice exercises. For example, in the work of Knoop and Wilske (2013, 41), POS tagging is used to determine appropriate distractors in fill-in-the-gap exercises with multiple possible answers. In Perez and Cuadros (2017, 49), POS tagging is used to determine the 'pedagogical target', i.e. which word category the user wants to focus on (e.g. nouns, verbs, modals, prepositions, etc.).

2.3. WordNet

"WordNet is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept" (Princeton University 2010). Synsets are interlinked by means of conceptual-semantic and lexical relations. Lexical relations include synonymy (words that have similar meanings, e.g. sofa-couch), polysemy (words that have more than one meaning, e.g. mouse as an animal and mouse as a computer input

device), hyponymy/hypernymy (hypernyms are more general synsets and hyponyms are more specific, e.g. bird-robin), meronymy/holonymy (part-whole relation, e.g. table-leg), antonymy (lexical opposites, e.g. black and white), etc.

WordNet is used for numerous tasks, such as word sense disambiguation, automatic text classification, automatic text summarization, information retrieval, machine translation, etc. It can also be used in the automatic generation of language exercises. For example, Knoop and Wilske (2013, 41) use WordNet to find appropriate distractors in multiple-choice exercises. They use antonyms or false synonyms of the target word as distractors. In the work of Brown, Frishkoff, and Eskenazi (2005), WordNet is used to generate six types of vocabulary exercises, including definition, synonym, antonym, hypernym, hyponym, and cloze questions. The definition item requires a definition of the word available in WordNet. The synonym, antonym, hypernym, and hyponym items require the user to match two corresponding words in the specified lexical relation. The cloze item requires the use of the target word in a specific context, either in a complete sentence or in a phrase. The sample sentence or phrase is retrieved from WordNet.

2.4. Other linguistic resources

In addition to the natural language processing techniques and resources mentioned above, some automatic language exercise generation systems also use other linguistic resources such as various specific word lists, dictionaries, collocation lists, etc.

As mentioned above, Fenogenova and Kuzmenko (2016, 22) use two well-known corpora (the British Academic Written English Corpus (BAWE) and the British National Corpus (BNC)) and the Academic Collocation List for automatic generation of collocation-based exercises.

Some systems use manually or automatically generated resources. For example, Malafeev (2015, 444-445) developed a system called ‘Exercise Maker’ for automatic generation of language exercises, which includes seven different types of exercises: word formation, error correction, open cloze, word bank, missing words, text fragments and verb forms. The author compiled a number of linguistic resources for exercise generation, including lists of the most common English word forms, a list of rules that allow realistic spelling, a list of adverbs used in the verb forms exercise, a list of verb forms, some manually written shorter lists of articles, conjunctions, prepositions, pronouns, etc.

3. Context, datasets and methods

3.1. Universal methodology for creating LSP learning exercises

The LanGuide tool takes as its starting point the Common European Framework of Reference for Languages (or CEFR) and assumes the action-oriented approach to language learning described therein. Following this approach, language learning occurs as part of learners' engagement in language activities, which involve dealing with spoken or written texts related to different themes and belonging to different domains of everyday life, in order to accomplish different tasks (Council of Europe 2001, 9). In the process, the learners employ their linguistic competences, general ones as well as communicative language competences, which are modified or reinforced with time. Such language activities in the LanGuide tool were prepared by the linguists and language teachers involved in the project and are based on the analysis of learner needs and the resulting syllabus created at the beginning of the project.

Further in line with the CEFR, the LanGuide tool caters for learners at three proficiency levels: (1) basic, (2) intermediate, and (3) advanced. The proficiency bands, however, are not as granulated as in the CEFR (where there are 6 proficiency bands altogether), as such detail was not deemed necessary taking into consideration the basic aim and target audience of the tool. Instead, the A1 and A2 levels from the CEFR were taken to make up the basic level, B1 and B2 the intermediate level, and C1 and C2 the advanced level of proficiency. At each level, there are language exercises or tasks created for productive language skills (speaking and writing), receptive language skills (listening and reading), and grammatical exercises and vocabulary items, following the CEFR's descriptors appropriate for each of the included levels.

There are three categories of target users of the LanGuide m-learning application: (1) university students, (2) university teachers, and (3) administrative staff. For each category of users, the tool is able to provide language exercises appropriate to their proficiency level and the selected language skill.

Finally, as stated in the LanGuide project plan, the tool does not support learning general English, but focuses on ESP. Thus, there are four broad areas defined to achieve this, namely (1) English for academic purposes (EAP), (2) administrative or secretarial English, (3) English for mobility purposes and (4) English for IT purposes. In this paper, the focus will remain only on the first area – EAP – and all the examples provided will pertain to it.

Given the complexity of the approach described above, the overall approach taken in the development of the LanGuide m-learning tool can be summarised as shown in Figure 1.

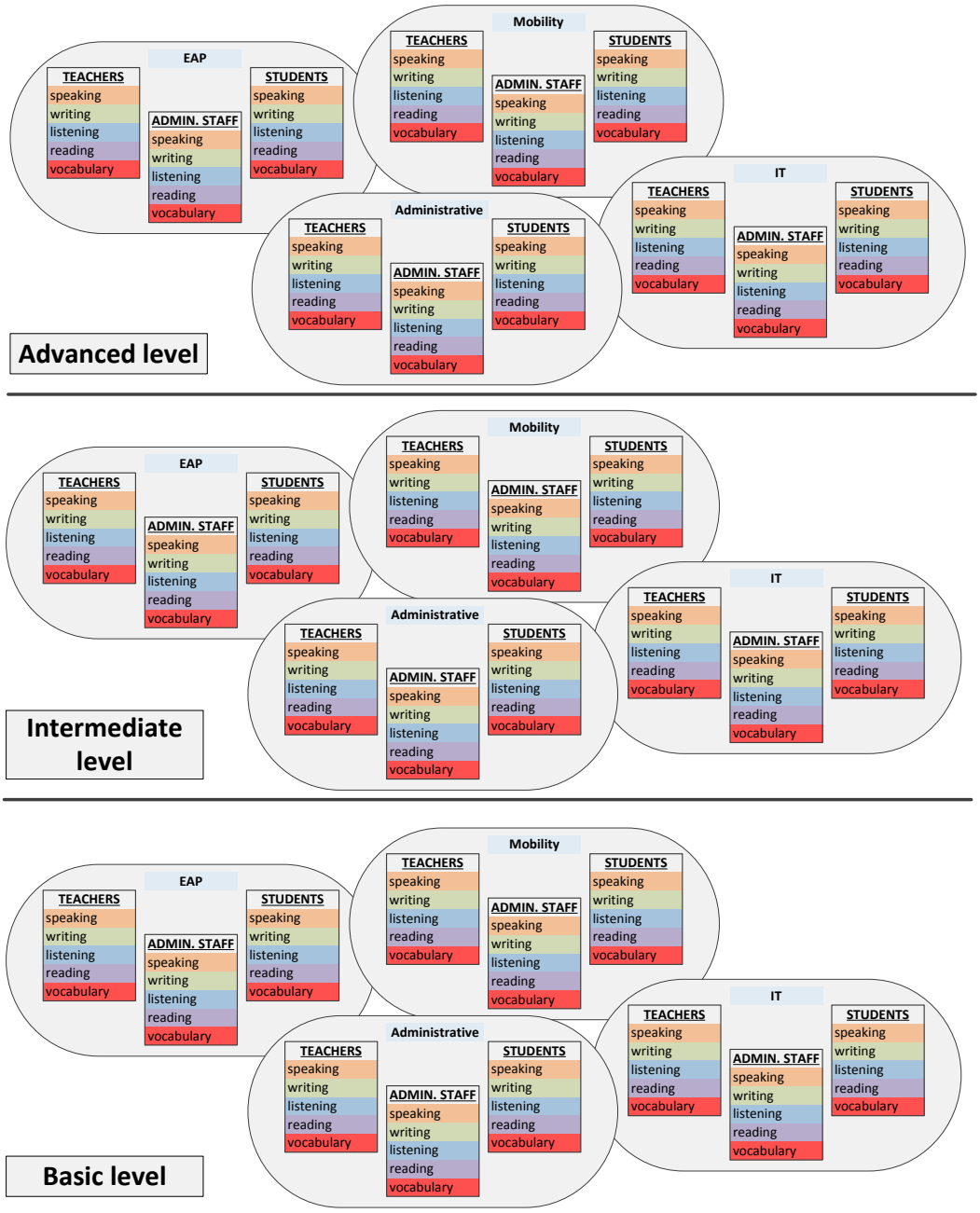

Figure 1. Overview of the LanGuide approach

3.1.1. Characteristics of exercises created within the LanGuide approach

When designing materials for the LanGuide tool following the previously described approach, a somewhat adapted version of the well-known ADDIE model was employed, complemented with the approach described by Klimova (2015, 634). The ADDIE approach is a five-stage process comprised of analysis, design, development, implementation and evaluation of materials, and has proven well-suited to organising the creation of online learning materials for a course, including language ones (Cuesta 2010, 183). The development of materials for the LanGuide tool includes one more stage – internal evaluation of course materials – that precedes the implementation of the materials into the tool. The sequence of stages in the development of language exercises for the LanGuide tool is given in Figure 2.

The first two stages, Analysis and Design, are preparatory stages, during which the needs of the various learners who will use the tool are analysed and determined. Additionally, during these stages it is imperative to establish instructional goals, define instructional content, and contemplate delivery options and restrictions posed by the technology (Cuesta 2010, 183-84). These are rather comprehensive procedures and may involve a variety of approaches. During this stage, material creators, in collaboration with the IT team, decided on the appropriate task types to be included into the LanGuide tool: given the context of distance learning, only those types of tasks for which there is a possibility of automatic evaluation by the tool were deemed as appropriate (namely, multiple choice, fill-in-the-gaps, and matching tasks, or their slight varieties). Prior to that, the LanGuide language team agreed on the appropriate communicative activities, learning outcomes and language content to be included.

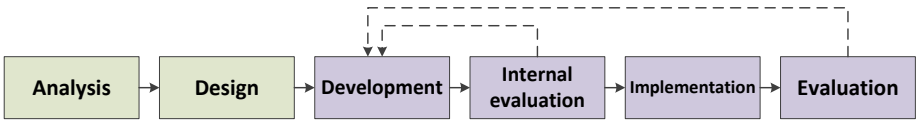

Figure 2. Stages in the development of learning materials in LanGuide

Following the preparation phase, it is necessary to start the creation of learning materials, keeping in mind and following the guidelines set in the previous two stages. During the development phase in the LanGuide approach, suitable texts were found and, if necessary, adapted to the needs of the defined approach. The selection of language learning materials, or, better yet, language texts necessary for carrying out linguistic tasks, included several criteria for doing so, following the work of Schader and Waibel (2016, 113). These, for example, included

considerations of whether materials are appropriate for the age of learners, whether they promote independent thinking in learners, how they fit the background of the defined learning context, and how easy it is to adapt them to suit the instructional needs.

Once the materials are created, the LanGuide approach to materials development requires them to undergo initial evaluation. Each exercise is thus evaluated by a selected member of the project's linguistic team (other than the creator of the material). After receiving the initial feedback on their materials, their creators redo them as suggested and prepare them for implementation. The goal of this stage is to increase the quality of the created materials even before they are introduced for use by the target learners of language.

Evaluated materials are then implemented into the LanGuide tool and available for use by the language learners. After being used for a particular period of time, the learners are able to give their evaluation and feedback on all the materials they used during learning. If the evaluation feedback is positive and does not suggest any changes, the materials remain in the tool. Otherwise, they are adapted by their creators and implemented into the tool once again.

Given the described context of materials use, as well as their type and complexity, there exists a real opportunity to automatically generate these language exercises, thus allowing for a speedier and easier creation of a large number of exercises (e.g., Perez and Cuadros 2017).

3.1.2. Exercise example

An example of the exercise created by a language expert, following the approach described above, is given in Figure 3. It is a multiple-choice activity (implemented as a drop-down menu in the application itself) in which the language learner has to choose the most appropriate option/word from the ones offered so that the text makes sense. The text was slightly adapted from the original source to suit the needs of language learning whereas all the options of a given item (blank that needs to be filled) were carefully chosen by the task creator and feature words belonging to different word classes, thus requiring the learner to think about word formation.

The task shown below is intended for the intermediate level administrators to evaluate their vocabulary skills (i.e. word formation).

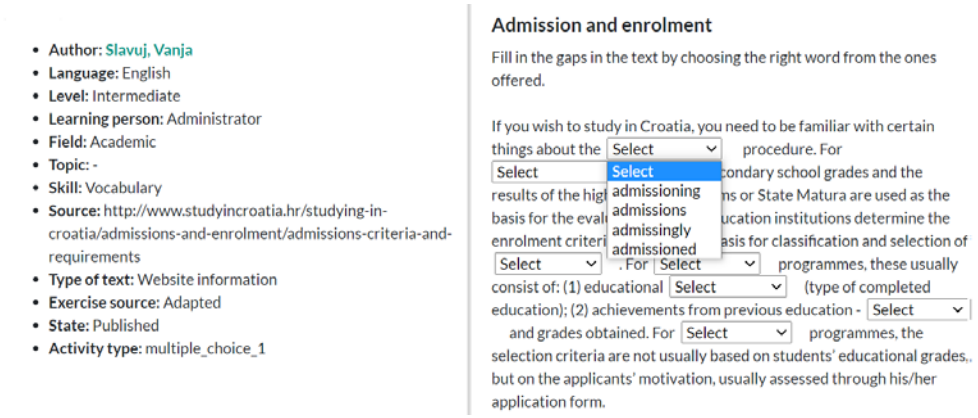

Figure 3. Example of a vocabulary exercise created using the LanGuide methodology (right) and its metadata (left)

3.2. Methodology for automatic generation of LSP learning exercises

One of the aims of this work is to make the automation procedure accessible to language teachers, non-experts in NLP, in order to involve them in the process and to affect their attitudes toward automation, which are usually negative due to poor understanding of NLP methods. Since the Content Management System of the LanGuide app currently supports three types of exercises, i.e. fill-in-the-gap, matching, and multiple choice, we will restrict our considerations to these types of items.

The experiment described herein includes grammatical, lexical, reading, and writing exercises. The tasks involving the four basic language skills (reading, listening, writing, and speaking) are usually abundant with grammar and vocabulary components, which are often very important, if not crucial, for creating understanding. However, out of the four already mentioned basic skills, we take into account only two, namely reading and writing. Since speaking is problematic in itself regarding automatic assessment, even in the case of manually created exercises, we exclude it from this research. Additionally, in order to automatically create listening exercises, a selection of suitable spoken corpora has to be created prior to the generation of exercises. Listening, therefore, remains out of scope of this work as well.

We propose a two-phase approach to automatic generation of LSP exercises which makes use of the Sketch Engine (Kilgarriff et al. 2014, 7), a tool for creating and manipulating corpora, available at http://www.sketchengine.eu, or any similar tool.

The first phase implies compiling or selecting a suitable corpus. The second phase consists of three steps – in the first step of the phase a Cassandra Query Language (CQL) query is formed in order to obtain a list of appropriate sentences. In the second step, the words which satisfy the created CQL condition are scanned and their suitability for a particular proficiency level is determined. In the third and final step, the sentences which belong to the same proficiency level are grouped.

Vocabulary items across all three levels of proficiency (basic, intermediate and advanced) can be selected in different ways. For example, one could use the English Vocabulary Profile (EVP), part of the English Profile – the CEFR for English ('English Vocabulary Profile' 2021), which is offered by the Cambridge University Press free of charge, thus allowing educators, materials developers, test creators, syllabus designers and other practitioners to obtain reliable information regarding words, phrases and their meanings and to map them to a particular level of the CEFR. In this research we use similar academic vocabulary lists available at https://www.academicvocabulary.info/download.asp (Gardner and Davies 2014, 305).

In line with the methodology developed for the manual creation of exercises within the LanGuide project and with regard to the exercise types currently supported by the accompanying content manager, we build three exercises per category or skill included in the research. The only exception is the writing skill for which only fill-in-the-gap and matching tasks are considered appropriate.

3.2.1. Corpus

The task of generating language learning exercises automatically implies using a wide range of NLP methods and techniques. In order to make our methodology transferrable to languages other than English, we compile a bilingual English-Croatian mobility corpus from the selected documents that can be retrieved at https://op.europa.eu/, as there are parallel documents available also for other partner languages involved in the project. The post-alignment editor used for correcting automatically obtained sentence alignments in our approach is the InterText Editor (Vondricka 2014, 1875).

Our final corpus from which exercises are automatically generated is composed of 6 documents and contains around 174,000 words on the English side.

3.2.2. Automatically generated exercises

Using the described approach, a total of 11 language exercises are generated automatically. Examples of three different exercise types are given in Figure 4, Figure 5, and Figure 6.

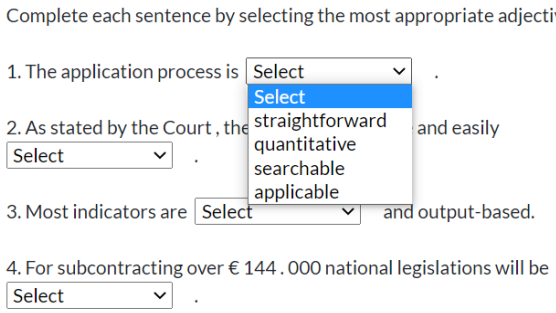

Figure 4. Example of an automatically generated exercise – multiple choice task

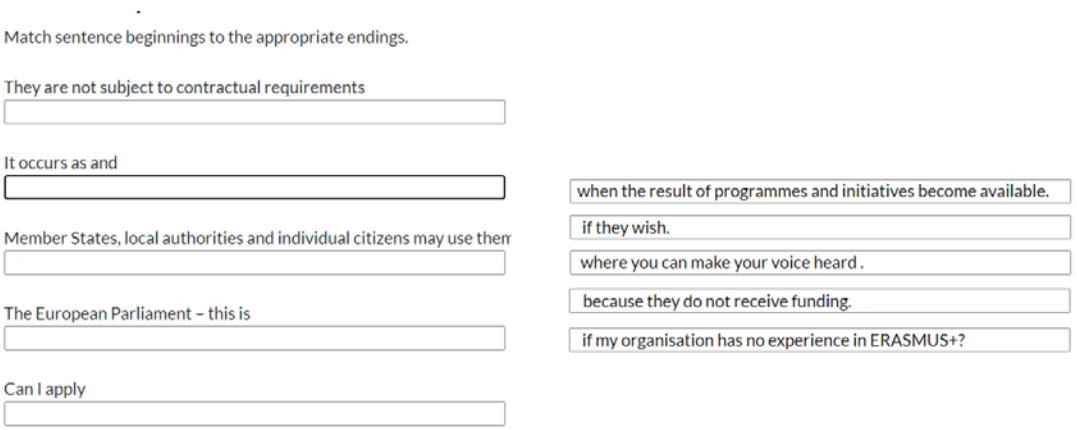

Figure 5. Example of an automatically generated exercise – matching

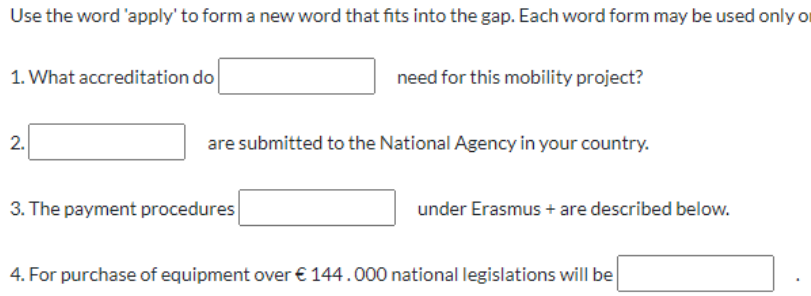

Figure 6. Example of an automatically generated exercise – cloze task

In order to build exercises with collocations (implemented as fill-in-the-gap type of activity), we first generate a word list consisting of words of a specified part of speech, then create word sketches and finally extract the desired number of items from a specified relation (such as the 'objects of' type of relation). The collocates can be extracted based on the descending score of word frequency or randomly from a defined top list. A C-test type of task (fill in the missing words letter by letter; several letters at the beginning of the word are given and the number of letters is indicated) can be generated in the same manner. Of course, one needs to make sure that there is at least one distinct collocate per each selected word. In addition, cloze tasks with rational deletion (filling in specified words from a particular word class, such as conjunctions) can be extracted from the concordance tool by specifying a suitable CQL query, e.g. as in (2):

(2) <s/> containing []{*a*,} "*X*|*Y*|*Z*" []{*b*,} within *span*,

where *X, Y,* and *Z* stand for the specified words, *a* for the minimum number of tokens before the conjunction, *b* for the minimum number of tokens after the conjunction, and *span* to the sentence length. Obtaining sentences which contain words with the same root can be done in a similar fashion, e.g. as in (3):

(3) <s/> containing [lemma="*root*.*"] within *span*

Matching exercises (implemented as drag-and-drop type of activity) are corpus-specific as they can be generated by exploiting the corpus structure and, with respect to that, specifying a suitable CQL query, e.g. as in (4):

(4) <s/> containing <s> [][tag="N.*"][word==":"] within *span*

Multiple choice tasks (e.g. sentence completion with appropriate words, word categories, or phrases) can be extracted by specifying CQL queries such as in (5):

(5) <s/> containing [tag="N.*"] within *span*

where *tag* refers to a desired part of speech, i.e. nouns in this case. Since a sentence often contains multiple words of the same part of speech, identical sentences are grouped and treated as a single instance during the selection process. Also, multiple choice items offered should not be synonymous. Therefore, an additional step of checking the top 10 thesaurus list of each word is introduced to make sure that other words which are also selected do not appear in it.

4. Evaluation of automatically generated language exercises

An experiment is conducted to examine how English teachers cope with automatically created exercises. Four sets of exercises are created as outlined in the section on methodology. To simplify the evaluation procedure, which should be neither too tedious nor too time-consuming, one example per each supported exercise type and per each supported skill type is generated. Due to a small sample of exercises, three evaluators are considered sufficient to assess the generated exercises. In addition to a quantitative evaluation of exercises, which is based on the scores obtained by evaluators when solving the exercises, a subjective evaluation is also performed. It gives evaluators the chance to express their opinion on the suitability of the exercises regarding the type and level of language proficiency, and to warn about possible ambiguity which is not necessarily reflected in the achieved quantitative scores.

Six out of eleven exercises (two exercises of each type) were assessed as suitable regarding both the type and the intended proficiency level (Table 1). The only comment on these six exercises concerns instructions of one of the fill-in-the-gap exercises which should explicitly state that each word form should be used only once in order to make it clearer for the learners and to avoid ambiguity.

Table 1. Evaluation of automatically generated exercises

Task ID	Type	Skill/Category	Proficiency level	Maximum score/Total	Average score	Remark
AC001GI	SELECT	Grammar	intermediate	5/5	5	-
AC007GB	FILL-IN	Grammar	basic	3/6	3	Difficulty
AC008GB	MATCH	Grammar	basic	5/5	5	-
AC003VI	SELECT	Vocabulary	intermediate	4/4	4	-
AC004VI	FILL-IN	Vocabulary	intermediate	4/4	3	Instructions
AC010VB	MATCH	Vocabulary	basic	6/6	5	Difficulty
AC009RI	SELECT	Reading	intermediate	5/5	5	Level
AC011RB	FILL-IN	Reading	basic	3/3	3	-
AC005RB	MATCH	Reading	basic	3/3	3	-
AC006WB	FILL-IN	Writing	basic	3/6	2.33	Difficulty
AC002WI	MATCH	Writing	intermediate	6/6	6	Level

The greatest issue was detected in the case of two exercises with collocations because they were assessed as too difficult and lacking appropriate context, which contributed to the increased difficulty of the task. One of these writing exercises was of the fill-in-the-gap type with the first letter of the base given (a C-test type of task), while the second one was listed under vocabulary and was of the matching type.

Another major issue was identified concerning the fill-in-the-gap exercise including the use of modal verbs, again as a result of multiple possible answers that fitted each gap. However, even manually created tasks on modal verbs are notoriously tricky to solve if not provided with enough context or definite indicators to guide verb selection.

Both the reading exercise of the multiple-choice type and the writing exercise of the type match, generated by exploiting the corpus structure, were assessed as too difficult for the intended level and a suggestion was made to redefine them as appropriate for the higher proficiency level.

Overall, however, the evaluation revealed that exercises of the type fill-in-the-gap are the least suitable for automatic generation, since two out of four exercises failed in the manual evaluation task. Another issue detected during the evaluation procedure is that multiple possible answers in the tasks with collocations and modal verbs made none of the exercise types suitable for automatic generation, at least not in the context-free form. Therefore, in our future work we intend to generate a set of context-dependent collocation exercises.

All in all, the evaluation reveals that adequate grammatical and lexical exercises, as well as those covering reading and writing skills, which are suitable for all the three exercise types supported can be automatically generated, the only limitations being a careful selection of the grammatical field covered and ensuring enough context for the exercises with collocations to narrow down the number of correct/possible answers.

5. Conclusion

The task of creating language exercises manually is largely a time-consuming and expensive one. Additionally, sentences and texts within tasks created in that way might be seen as lacking in authenticity, as they have been specifically intended for didactic purposes.

One such approach is taken in the creation of the language learning tool named LanGuide in which the task of creating language exercises for three language proficiency levels, three learning personas, and four language skills (plus vocabulary and grammar) was given to language teachers and other language experts. Based on the LanGuide methodology of exercise creation, outlined in this paper, it is noticeable that exercise creators, in addition to text selection and adaptation, need to consider a large number of (learner- and context-specific) variables in order to create valid language exercises, which often proves a very time-consuming endeavour. Moreover, once the exercises have been created, they

have to undergo a scrutinous evaluation of other language experts, as well as language learners, in order to make sure the exercises adhere to the set standards. In order to make the approach speedier and easier for language teachers and other non-ICT-experts, automatic procedures for generating exercises might be considered.

Within this research, a two-phase approach to automatic generation of LSP exercises is suggested. The first phase implies compiling or selecting a suitable corpus, and the second phase is concerned with querying the corpus and processing results.

The selection and creation of the parallel corpus is guided by the topics defined within the framework of the LanGuide project for the field of academic English and by the availability of the documents in all the project partners' languages to set grounds for a multilingual corpus creation. Although the research presented in this paper exploits only one side of the parallel corpus, namely that for English, by enabling learners to compare a text in one language with its translation in their mother tongue and vice versa, they can explore the target language in a guided way. Therefore, the compiled corpus can be used for expanding the supported exercise types.

The task of multilingual corpus creation could be further simplified by using a corpus of subtitles given that subtitles are available in all target languages. In that case, the alignment procedure could be completely automatic and, conditionally said, error-free, due to the association of the sentences to time codes. However, due to space and time restrictions, translation in this field is freer than in other domains, which could have a negative impact on the automatic generation of exercises.

The analysis of the compiled corpus in the preparation phase reveals that the quality of the corpus is of great importance for the diversity of the automatically created exercises. We, therefore, opt for a guided approach to corpora creation. For example, in the corpus compiled within this research, there are segments that contain both a question and its answer, or a subtitle and the respective description, which proves to be convenient for generating reading and writing exercises.

To evaluate the proposed methodology, four sets of exercises are automatically generated, i.e. one example per each supported exercise type and per each supported language skill or category. Over 70% of the created exercises are assessed positively by three evaluators. The conducted manual analysis shows that the most problematic exercise type is fill-in-the-gap, mostly due to the possibility of multiple correct answers for each gap, which could not be induced automatically. In summary, we would like to point out that the automatic generation of exercises can serve at least as the first phase of a two-step process in which each exercise thus created requires approval by the instructor or language expert.

Acknowledgement

This research was supported by the Erasmus+ grant number 19-203-060377 - KA2-HE-01/19.

References

Bateson, Gordon and Paul Daniels. 2012. “Diversity in Technologies.” In *Computer-Assisted Language Learning: Diversity in Research and Practice*, ed. by Glenn Stockwell, 127–146. Cambridge: Cambridge University Press.

Bax, Stephen. 2003. “CALL - Past, Present and Future.” *System* 31 (1): 13–28.

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom, Corpus Linguistics for Teachers*. University of Michigan Press ELT.

Bick, Eckhard. 2005. “Live Use of Corpus Data and Corpus Annotation Tools in CALL: Some New Developments in VISL.” In *CALL Conference at CBS*. Copenhaguen.

BNC Consortium. 2007. “The British National Corpus.” *Version 3 (BNC XML Edition)*. Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Brown, Jonathan C., Gwen A. Frishkoff, and Maxine Eskenazi. 2005. “Automatic Question Generation for Vocabulary Assessment.” In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT ’05*, 819-826. Morristown, NJ, USA: Association for Computational Linguistics.

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

Cuesta, Liliana. 2010. “The Design and Development of Online Course Materials: Some Features and Recommendations.” *Profile Issues in Teachers’ Professional Development* 12 (1): 181–201.

“English Vocabulary Profile.” 2021. Cambridge University Press. 2021. https://www.englishprofile.org/wordlists.

Fenogenova, Alena, and Elizaveta Kuzmenko. 2016. “Automatic Generation of Lexical Exercises.” *CEUR Workshop Proceedings* 1886: 20–27.

Gardner, Dee, and Mark Davies. 2014. “A New Academic Vocabulary List.” *Applied Linguistics* 35 (3): 305–27.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine : Ten Years On.” *Lexicography* 1 (1): 7–36.

Klimova, Blanka Frydrychova. 2015. “Designing an EAP Course.” *Procedia - Social and Behavioral Sciences* 191: 634–38.

Knoop, Susanne and Sabrina Wilske. 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*: 39–47.

Kompara Lukančič, Mojca and Ivo Fabijanić. 2020. “LanGuide - A Tool for Learning English.” In *English and Italian in the Frame of Genre-Based Research and Foreign Language Learning*, ed. by Jasna Potočnik Topler, 33–73. Maribor: University of Maribor Press.

Lamy, Marie-Noelle. 2013. “Distance CALL Online.” In *Contemporary Computer-Assisted Language Learning*, ed. by Michael Thomas, Hayo Reinders, and Mark Warschauer, 141–58. London: Bloomsbury Academic.

Malafeev, Alexey 2015. “Exercise Maker: Automatic Language Exercise Generation.” In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2015)*, ed. by Vladimir Selegey, 14(21), 441–52. Russian State University for the Humanities.

Manning, Christopher D. 2011. “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?” In *Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science*, Vol. 6608. Berlin, Heidelberg: Springer.

Perez, Naiara and Montse Cuadros. 2017. “Multilingual CALL Framework for Automatic Language Exercise Generation from Free Text.” In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of the Software Demonstrations*, 49–52.

Princeton University. 2010. “About WordNet.”

Pykes, Kurtis. 2020. *Part Of Speech Tagging for Beginners*. last modified November 25, 2020, https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba.

Schader, Basil and Saskia Waibel. 2016. “Finding and Selecting Suitable Materials.” In *Foundations and Backgrounds*, ed. by Basil Schader, 113–117. Zurich: Center for IPE (International Projects in Education) Zurich University of Teacher Education.

Simpson, Rita, Sarah Briggs, Janine Ovens, and John Malcolm Swales. 2002. *The Michigan Corpus of Academic Spoken English*. Michigan: University of Michigan.

Stockwell, Glenn. 2013. “Mobile-Assisted Language Learning.” In *Contemporary Computer-Assisted Language Learning*, ed. by Mark Thomas, Michael; Reinders, Hayo; Warschauer, 201–216. London: Bloomsbury Academic.

Taki, Saeed and Neda Amini. 2017. “Evaluating ELT Materials: A Comparison between Traditional Materials and Mobile Apps.” *International Journal of Foreign Language Teaching and Research* 5 (20): 59–78.

Volodina, Elena. 2008. *From Corpus to Language Classroom: Reusing Stockholm Umeå Corpus in a Vocabulary Exercise Generator SCORVEX*. Master thesis, University of Gothenburg.

Vondricka, Pavel. 2014. “Aligning Parallel Texts with InterText.” In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*: 1875–79.

Zanetti, Arianna, Elena Volodina, and Johannes Graën. 2020. “NLP Methods for the Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data”. *International Journal of TESOL Studies* 3: 55–70.