International Journal of TESOL Studies (2021)
Vol. 3 (2) 55-70 https://doi.org/10.46451/ijts.2021.06.05

Article

Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data

Arianna Zanetti*
University of Gothenburg, Sweden

Elena Volodina
University of Gothenburg, Sweden

Johannes Graën
Pompeu Fabra University, Spain

Abstract
Creating language learning exercises is a time-consuming task and made-up sample sentences frequently lack authenticity. Authentic samples can be obtained from corpora, but it is necessary to identify material that is suitable for language learners. Parallel corpora of written text consist of translated material. Comparing the text in one language with its translation into another (known) language makes the structure accessible to the learner. However, the correspondence of words between the two languages is more important. By carefully selecting well-suited parallel sentences, a learner can explore the target language in a guided way. We present an approach to generate a novel type of language learning exercise from a large parallel corpus based on movie subtitles. The size of the corpus allows for defining selective criteria, favoring precision over recall. It is a non-trivial task to give reliable feedback to automatically generated exercises. ICALL literature often deals with fill-in-the-blanks exercises or multiple-choice questions, which allow for very limited answer options. Our proposed exercise is a special case of sentence reconstruction on bilingual sentence pairs. It combines two elements which have proven to be effective for language learning: a gamified approach, to awaken the students' competitive desire, and the identification of syntactic structures and vocabulary use, to improve language sensitivity. This article presents the methods used to select example pairs and to implement a prototype.

Keywords
ICALL, exercise generation, parallel corpora

1 Introduction

Natural Language Processing (NLP) and Language Technology (LT) deal with the analysis of natural

***Corresponding Author**
Address: Arianna Zanetti, via Aristotele 14 - 42048, Rubiera (RE), Italy
Email: arianna.zanetti4@studio.unibo.it

languages in written and spoken form and are becoming ubiquitous in the modern (digitized) society. NLP and LT have the potential to influence people's everyday life, just as Google has already proven, without people even knowing that they owe that to the development of Language Technology. The areas where we see LT's (omni) presence daily include translation (*https://translate.google.com*), effective information search (*https://www.google.com*), ticket reservation, lexicon services, e.g. Oxford English dictionary online (Simpson & Weiner, 1989), to name just a few examples. Duolingo (*https://www.duolingo.com*; Ahn, 2013), a mobile app that provides several types of exercises to assess language knowledge, both in spoken and written form, and Rosetta Stone (*https://www.rosettastone.eu*), a speech engine to practice conversation in a chosen language, among others, have proven that Language Technology can also boost language learning.

Anyone involved in teaching languages can confirm that there is always a need for new teaching materials. Wilson (1997), for instance, writes about the problem of addressing students of different levels and creating materials for that:

> *In language course design there are two major problems:*
>
> - *How to provide a range of materials to meet the needs of students with different abilities.*
> - *How to provide at every ability level enough exercises to ensure that a student is confronted by a different set of examples whenever he or she uses the language learning program.*

Creating language learning exercises manually is a tedious and perhaps thankless task since the novelty of each particular item is soon worn out and there is a perpetual need for new unseen items. Due to the recent advances in NLP techniques, automatic methods are more and more frequently employed for the exercise generation task (Alfter, Borin, & al., 2018; Meurers, Ziai, & al., 2010; Meyer & al., 2016) and in studies in which technology is incorporated into real-life classroom settings (Burstein & al., 2017; Meurers, De Kuthy, & al., 2019) which is a token that technology is mature enough to be accepted by teachers (cf. Volodina, 2020). Nonetheless, exercise generation is a nontrivial task and multiple issues need to be addressed and solved before their adoption to language learning curricula. In this paper we focus on two major challenges in exercise generation:

- the problem of *ambiguity*, which we try to reduce using parallel corpora, and
- a follow-up/derivative problem of *selection* of *appropriate parallel sentences* for which we experiment with reusing and assembling existing tools in a new scenario.

To exemplify the problem with *ambiguity*, we can consider the following gap example (with or without the first letter clue):

(1) He works as a [d].....

In this example, a learner can submit multiple legitimate variants (target items) into the gap, e.g. doctor, driver, dentist, etc. which makes the problem of automatic scoring very difficult.

Various ways have been explored to address the issue of ambiguity in exercise generation. Horsmann and Zesch (2014) have examined whether utilizing the context and structural information of the sentence can help selecting only non-ambiguous sentences for the exercises, the results being inconclusive. Meyer et al. (2016) have suggested to use a new version of gapped exercise, bundled gaps, where up to four sentences with the same target item are used to demonstrate the item in various syntactic patterns and various lexical distributions, thus self-disambiguating the potential target item candidate. Burstein et al. (2017) avoid the problem of ambiguity by generating suggestions for exercises that teachers need to approve, thus eliciting manual disambiguation. Katinskaia et al. (2017) avoid using exercises where text-wide context cannot be used to disambiguate items. Where this is impossible, they use online dictionaries to prompt a hint (i.e., translation of the word in question) in a nontarget language, excluding other potential candidates.

Using a hint in another language seems to be a reliable approach, as long as exercises are focused on lexical knowledge. When grammatical knowledge is being trained, however, dictionary lookup might not

be enough. Consider another example for an exercise involving word order restructuring:

(2) If you need to come in time, take a bus.

If a student gets the sentence shuffled by tokens [a, bus, come, if, in, need, take, time, to, you], multiple legitimate solutions can be provided, e.g.

(2a) If you need to take a bus, come in time.

(2b) Come in time, if you need to take a bus.

(2c) If you need to come in time, take a bus.

(2d) Take a bus, if you need to come in time.

(2e) If you need, come in time to take a bus.

(2f) If you need, take a bus to come in time.

The ambiguity of this type will not be solved by a dictionary lookup. We propose to use parallel sentences from two or more languages as a basis for exercise generation, so that the translations can be used for disambiguation purposes. The sentence in the target language (the language a learner is studying) can be complemented by an equivalent sentence in another language to narrow down the number of correct answers.

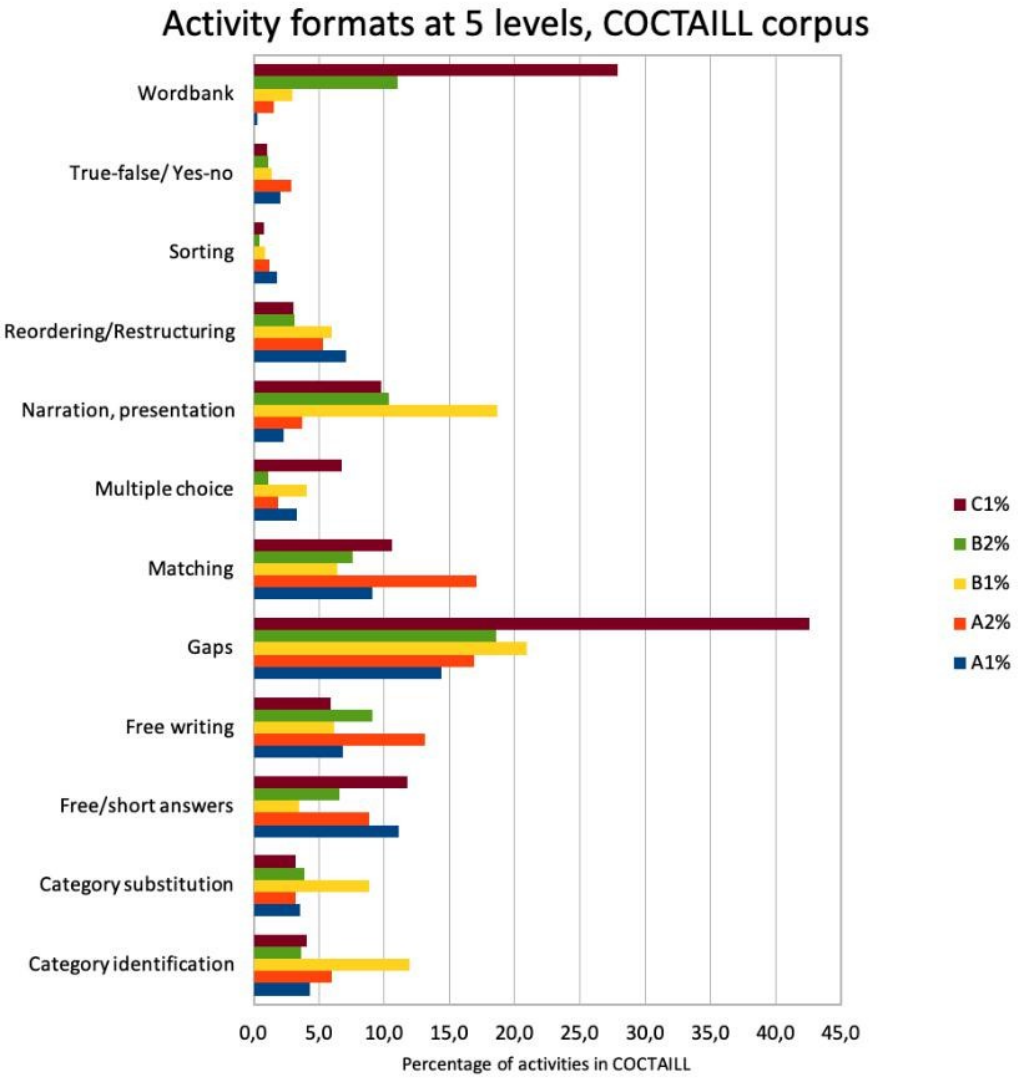


Figure 1. Distribution of exercise types in the COCTAILL corpus

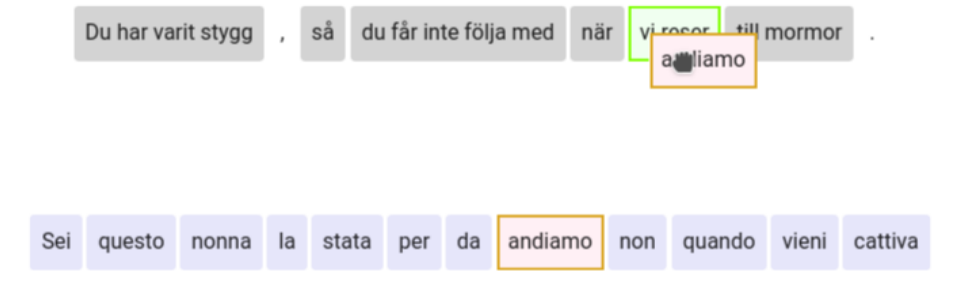
The idea was inspired by Alfter and Graën (2019) where the authors used a similar approach employing hints in a non-target language from aligned parallel corpora for training knowledge of particle verbs (single item level).

To our knowledge, parallel corpora are rarely used in automatic exercise generation. There may be several reasons for this, one of the most realistic ones being that parallel corpora have not been built with language learning in mind, and the problem of selecting appropriate sentence pairs for language learners at a given level of linguistic competence becomes very prominent. To investigate the feasibility and the validity of the suggested approach, we focus on one exercise type, namely reordering exercise. The selection of this exercise is primarily motivated by the fact that very few automatic approaches target grammar exercises involving word order syntactic level, whereas an overview of textbook corpora shows that this exercise type is used at all proficiency levels. As an example, Figure 1 shows the exercise types found in the COCTAILL textbook corpus (Volodina, & al. 2014).

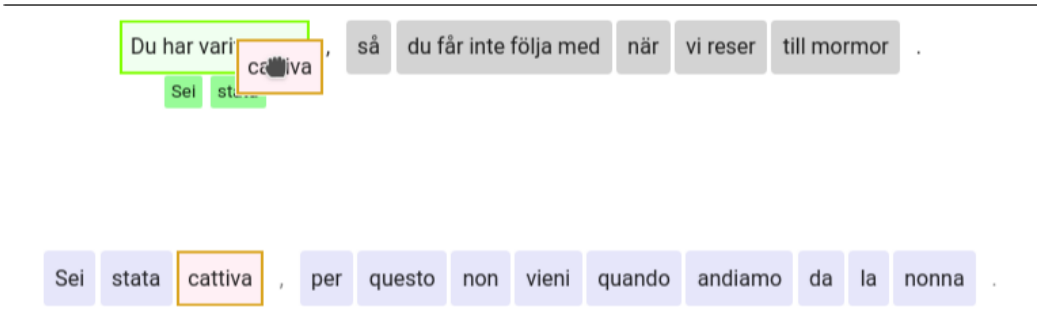
The proposed exercise works as follows: each target sentence is split into tokens, which are shuffled and presented along with the source language sentence in its original order split into larger units; the learner is asked to assign each target language token to one of the source language units, thus reconstructing partly the grammatical structure and identifying lexical correspondences. The idea to use larger syntactic units is based on the intuition that content words and their respective functional dependents grouped together can be helpful in encouraging the identification of syntactic structures and vocabulary use in the target language. The groups of words are also beneficial to the alignment of the sentences in the different languages at the word level, which is fundamental especially for beginner learners who do not have enough knowledge of the language to make associations on their own. Semantic correspondence between two natural languages is usually to be found on units bigger than a single word; especially functional words frequently do not correspond to a meaningful counterpart in the other language. Figure 2 shows an exemplification of the dynamic of the exercise.

In the first image (Figure 2a), the sentence pair is shown, with a Swedish sentence (source language) separated in larger units of meaning, along with the alignment and the corresponding Italian sentence (target language). The following two images show the dynamic of the exercise, in which the learner drags the tokens to the correct group. The last image (Figure 2d) shows what happens in case an association is not correct.

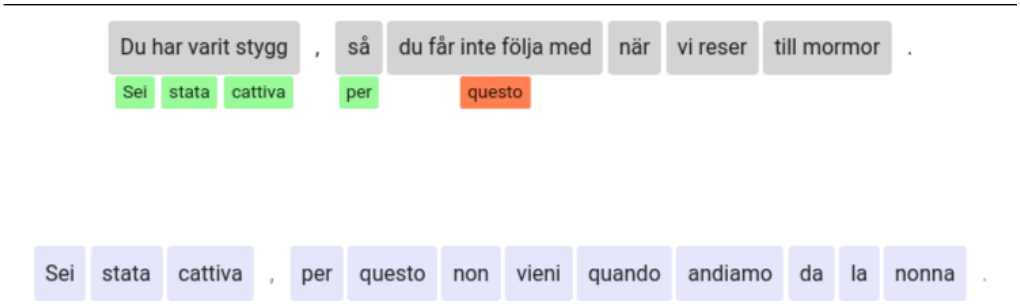
(a) The aligned Italian-Swedish sentence pair used for the exercise.



(b) Exemplification of the solution of the exercise [1/2].



(c) Exemplification of the solution of the exercise [2/2].



(d) Incorrect association.

Figure 2. Various stages of our proposed exercise

The puzzle-like character of the exercise is an example of a gamified approach, which has proven to be effective for language learning (Klimova & Kacet, 2017).

The article gives an overview of the related work in the following section, then describes the sentence selection and pre-processing pipeline that was implemented and the experiments that were run on three different language pairs: Italian-English, English-Swedish and Swedish-Italian.

2 Related Work

Teubert (1996) and Lawson (2001) state that a parallel corpus of reasonable size contains more knowledge for a learner than any bilingual dictionary. Carefully selected parallel sentences allow a person to learn from comparing the target language (L2) with a known language, e.g. the learner's mother tongue (L1), to explore the structure of the L2 in a guided way that is, at the same time, both unbiased and unmediated. It also gives them a chance to practice without the limitations of a classroom environment, self-pacing the learning activity and frees the teachers from the time-consuming task of generating exercises to provide the class with, and the comparison with a known language helps dealing with the ambiguity issues that rise from exercise generation in a monolingual environment. The problem is how to use a parallel corpus to automatically generate exercises and how to obtain all the necessary information to give meaningful feedback on the users' solutions.

At the present day, there is still a lack of systems which automatically generate exercises and provide feedback on the submitted solution. The most common applications for language learning, like Duolingo, are based on a human generated list of sentences. The reason may be found in language models being not completely reliable. Every NLP tool comes with a small but nonzero error rate. Some doubts have also been presented by researchers in pedagogy, saying the view of language learning promoted by technicians is limited and focuses more on the words and grammar used than on the creativity and communication, as would do a human teacher (Mishan & Strunz, 2003; Potter,

2004). Considering the improvements in research in NLP and the increasing availability of parallel data, though, the questions arise if it is possible to remove this need for a human intervention and offer the user a theoretically unlimited source of real examples of language use. For many people, the possibility to test their knowledge in a more flexible way and training as many times as they want on a subject without the risk of being embarrassed in front of a colleague or teacher would result in quicker and better improvements in their skills. Moreover, the system could keep track of the user's progress to provide more specific feedback (Rudzewitz & al., 2017), taking into account, for example, their mother tongue to predict positive or negative L1 transfer, or previous answers they entered, to tailor the teaching activities to the specific needs and guide the learning process. Cobb and Boulton (2015) found 116 empirical studies to support the benefits of data-driven learning. According to them, the exposure to authentic input supports intuition and helps the learners gradually reproduce the underlying lexical, grammatical, pragmatic, and other patterns implicit in the language they encounter. The interest in NLP for language learning is shown also by the different conferences organized on the topic, the first one being Applied Natural Language Processing in 1997 which devoted a session to CALL (Computer Assisted Language Learning), and by the increasing proportion of articles in journals dedicated to CALL (e.g., CALL, ReCALL, CALICO). Even if a computer cannot completely substitute the teacher in what is called a "tutor" application, it is possible to develop ICALL (Intelligent Computer Assisted Language Learning) systems at different levels, realizing "tools" for specific tasks (Levy, 1997), like analysing syntactic and/or semantic correctness of the user inputs.

Not every sentence from a corpus can be used for exercise generation, it is necessary to select appropriate examples, depending on the type of exercise and the proficiency level of the learner. The sentence must be well-formed, comprehensible in style, register and vocabulary, and sufficiently context independent, but there are no agreed characteristics that make a "good sentence". Too many constraints could result in oversimplified sentences, while out-of-the-ordinary examples could be more interesting or relate to the learners' world of experience and be better from a pedagogical point of view (Segler, 2007). A large part of the research in this area comes from the selection of good dictionary examples (Kilgarriff & al., 2008), because in both cases the sentence must be understandable out of context, semantically clear, and well-structured. Pilán et al. (2014) differentiate two general methods: machine learning and rule-based NLP.

Machine learning techniques are based on human-annotated sentences and estimate the parameters of a model using multiple linguistics dimensions at the same time. It is an important step to determine which features are the most predictive and help label the complexity of a sentence because, most of the times, the criteria teachers use to select them cannot be verbalized, they derive from intuition. Rule-based NLP techniques, on the other hand, work with predefined rules written by specialists and are thus more customizable and can be suitable for a wider range of applications. In reading comprehension tasks, it is more important to consider vocabulary knowledge. Second Language Acquisition (SLA) research has established that the student should know at least 95-98% of the words in a text to comprehend it (Laufer & Ravenhorst-Kalovski, 2010). When the example's goal is to demonstrate the use of a word, it is important to reduce syntactic complexity and to use semantically related words and frequent co-occurrences (Segler, 2007), but also to differentiate the use of the word in different contexts (Minack & al., 2011).

The idea to extract larger units from a sentence is not new, but there is no agreed-upon solution because researchers approach this intuition from different perspectives. According to Boulton (2016), psycholinguistic research on chunking supports the idea that the mind works with exemplars beyond the level of words. They suggest to choose clusters/n-grams with the same number of words in each string to see how they group together, even if these clusters may not carry much meaning. Other works based on groups of words can be found in Byrd and Coxhead (2010), where four-word bundles are extracted from a corpus to help learners find specific formulations and improve their academic writing. Cobb and

Boulton (2015) describe a three months experiment that showed the words met through concordances are retained in the 75.9% of the cases, against the words met through simple definitions, which are retained only in the 63.9%.

"Lexical bundles", first introduced by Biber et al. (1999), then implemented in an automated scheme for language learning by Wu et al. (2014), are multiword units with distinctive syntactic patterns and discourse functions that are frequently used in academic prose, for example verb + noun, noun + of + noun, etc.

"Reconstruction" exercises have been used in language teaching in a monolingual context, because they target a variety of aspects of language, e.g. word order, word phrase knowledge, syntactic patterns, cases, function word use. The problem is they introduce ambiguity, as is shown in the introduction, because the words of a sentence can be combined in multiple ways. This makes it difficult to automate meaningful feedback.

3 Data

For exercise generation, we use corpora from the OPUS project. The OPUS project (*http://opus.nlpl.eu*) made available the OpenSubtitles corpus (Lison & Tiedemann, 2016), the biggest online multi-language subtitles database in more than 60 languages, aligned at the sentence level between different languages and associated to different types of useful annotations, including the PoS (Part of Speech) tag of every token, and the dependency-parsing relations.

Subtitles are, according to Lison and Tiedemann (ibid.), the world's largest open collection of parallel data. Their conversational nature makes them ideal for exploring language phenomena and properties of everyday language (Paetzold & Specia, 2016). The main disadvantage is that subtitles must obey time and space constraints, and often try to lip-sync, whenever talking people are shown in the videos, therefore the translation is freer than in other domains, especially for some language pairs. Sometimes, they summarize the audio instead of completely transcribing it. How they "compress" the conversation differs due to structural divergences between the languages, cultural divergences and disparities in subtitling tradition and conventions. The entries in OpenSubtitles consist of user uploads of movie and TV subtitles and the association of the sentences to time codes makes it possible to align them efficiently across languages, or between the same language to find alternative translations (Tiedemann, 2012). An overlap score determines how likely the two sentences are to be aligned. Two sentences are considered the best candidates when the overlap is large. To create a parallel corpus from the subtitles' database, it was necessary to clean up and filter the original sources, to avoid incorrect language or title tags added by the users and make character encoding and format uniform. Since the subtitles are submitted voluntarily by users, with a crowdsourcing method, a possible issue is the reliability of the data. If the same users produce more subtitles, they probably can be trusted, and it is always possible for other people to update existing documents or to report errors, but the quality is not guaranteed as the providers of the subtitles' translation platform cannot check every document that is uploaded. Despite this, the corpus has already been successfully used by many applications, like Reverso (*https://context.reverso.net*), a translator of sentences in context, or Sketch Engine (*https://www.sketchengine.eu*), a corpus management system with concordancing and text analysis functionalities for linguists and lexicographers. To the author's knowledge, this is a first attempt at creating a tool for language learning with it.

4 Development Process

4.1 Pipeline design

In order to generate the reconstruction exercise, we identified three fundamental pre-processing operations: (1) align the sentence pairs at the word level; (2) identify syntactically associated tokens in the source language; (3) estimate the target proficiency level of the sentence pair.

To select the sentence pairs, we use several filtering processes. The filters produce fewer examples that can be used for the exercises, but they assure the sentences are appropriate for the language learning task with an acceptable level of accuracy, without the need for a human intervention. First, we look at the sentence alignment provided by OpenSubtitles and consider only sentence pairs in which the sentences have a 1:1 correspondence - sentences in an aligned parallel corpus can have a 1:1 correspondence but also 1:0 if they do not have a counterpart in the other text or 1:n if the meaning is separated into multiple sentences - and an overlap score of more than 50%. Then, since free translations with very different syntactic structures would not be useful for the purposes of the exercise, we compare the PoS tags of the words in each sentence and exclude the pairs in which the content words do not match. Finally, we exclude the sentences if they have fewer than five tokens or a finite verb is missing, because they are likely to be elliptic or too context dependent.

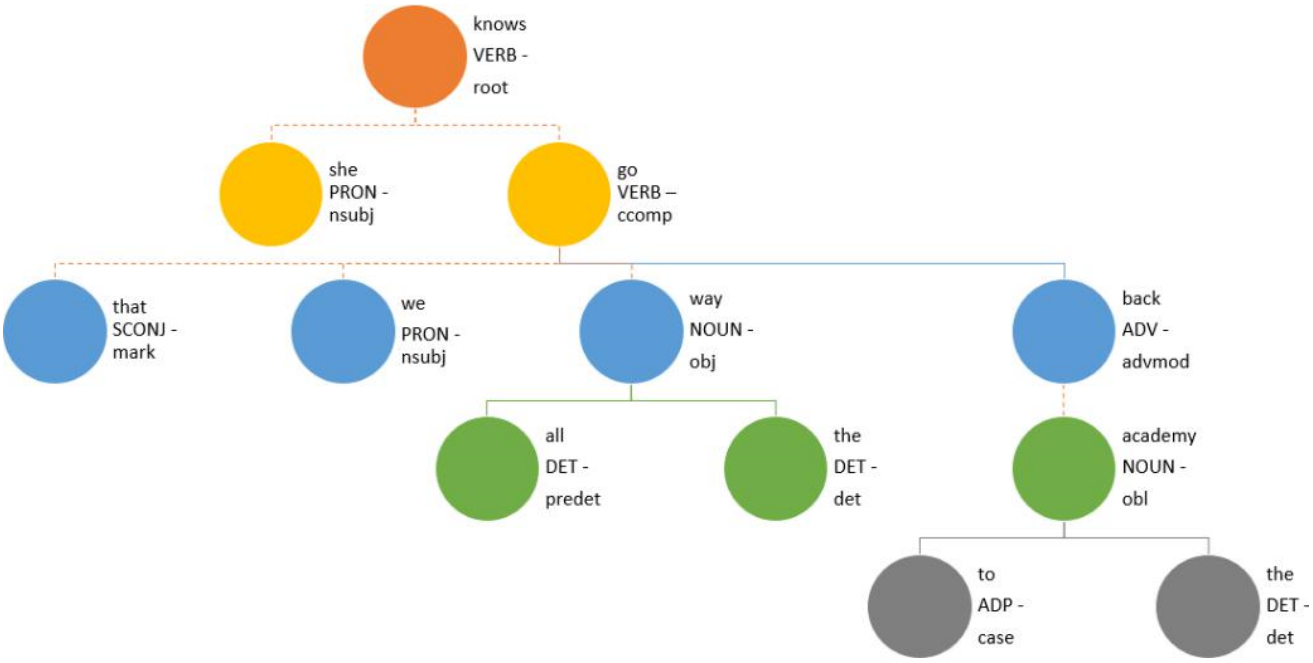


Figure 3. Exemplification of the clustering algorithm

4.2 Alignment

The word alignment is performed using eflomal (Östling and Tiedemann 2016; *https://github.com/robertostling/eflomal*), an unsupervised, low-memory alignment tool built on top of the IBM models (Brown & al., 1993). It builds a statistical model from the input data, to describe how the source language generates the target language through a set of alignment variables and estimates the word alignments one sentence at a time in a given document. It has proven to be accurate and computationally efficient; however, using an unsupervised approach with no prior knowledge of any of the two languages used (i.e., it is only based on parallel sentences as input), it is not always possible to obtain high levels of accuracy. Especially shorter parallel corpora provide insufficient data from which the algorithm could deduce correspondences. To overcome this issue, eflomal provides the option to initialize the algorithm with a previously obtained probability distribution, trained on a larger parallel corpus, and saved to a file that can be loaded for subsequent iterations. In our case, we used approximately 10,000,000 aligned sentences to train the probability distribution.

In this work, larger token groups are obtained from syntactic dependency relations using a divisive hierarchical clustering algorithm based on the definition of two concepts: core tokens and dependents. Every cluster is allowed to contain only one core element: when a second core element is found, the branch of the parsing tree is cut, creating a new cluster. *Core elements* are identified using two different rules, based on the Universal Dependencies tag-set and label-set used in the OpenSubtitles corpus:

1. PoS tag is noun (NOUN), proper noun (PROPN) or verb (VERB) and dependency relation is compound, name, multi-word-expression, goes-with, auxiliary, passive auxiliary, case marking or PoS is coordinating conjunction (CCONJ), subordinating conjunction (SCONJ) or interjection (INTJ).
2. dependency relation is one of nominal subject, passive nominal subject, clausal subject, passive clausal subject, clausal complement, open clausal complement, object, indirect object, oblique nominal.

An exemplification of this algorithm is shown in Figure 3, where the branches that have been cut are identified with a dashed line. The complete Universal Dependency tag-set for PoS tags and dependency relations is available at *https://universaldependencies.org/guidelines.html.*

4.3 Proficiency classification

The final step is to estimate the target proficiency level for each sentence. The Council of Europe defined a standard framework to provide guidelines for language teaching and assessment (Council of Europe, 2001), identified by the acronym CEFR (Common European Framework of Reference). It comprises 6 levels: A1, A2 basic; B1, B2 independent, and C1, C2 proficient. It would be ideal to assign the CEFR level to a sentence, but the competence and skills needed at any level are not clearly defined and there is room for interpretations in different languages and target groups (Volodina, Pijetlovic & al., 2013). Also, the description of the levels is based more on teachers' perceptions than on empirical evidence from learner data. There are ready-to-use tools available to calculate the complexity of a sentence only for some languages; many of the readability tools in literature only deal with long documents so the results are not accurate when applied to a sentence. In this work we use Tint (Aprosio & Moretti, 2018) for Italian and English, and HitEx (Pilán & al., 2013) for Swedish.

Tint is an open-source NLP suite built on Stanford CoreNLP (Manning & al., 2014) meant to (1) organize in a single framework most of the standard NLP modules, from tokenization to PoS tagging, morphological analysis and parsing, and (2) optimize the pipeline for the Italian language. Its readability module to estimate the complexity of a sentence is inspired by two earlier works: READ-IT, which is only available in the form of an online demo (http://www.italiannlp.it/demo/read-it); and Tonelli et al. (2012), a system based on CohMetrix, which estimates the readability of a document at three levels of proficiency: elementary, middle and highschool, using a large number of features, among which the familiarity of content words, calculated as their frequency in a reference corpus.

The metric proposed by Tint is calculated from a series of indices:

- Number of content words, hyphens, and distribution of tokens based on PoS in the sentence.
- Type-token ratio (TTR) between the number of different lemmas and the number of tokens.
- Lexical density, number of content words divided by the total number of words.
- Amount of coordinate and subordinate clauses, along with the ratio between them.
- Average and max depth of the parse tree.
- Gulpease formula (Lucisano & Piemontese, 1988) to measure the readability at the document level, index created specifically for the Italian language based on the number of characters contained in each word.

- Text difficulty based on word lists from DeMauro's Dictionary of Basic Italian (*https://bit.ly/nuovo-demauro*).

The lower the calculated value, the higher the proficiency level needed to understand the sentence. The Tint pipeline is released under the GNU General Public Licence, version 3 and it is written following the Stanford CoreNLP paradigm, so it is possible to integrate it into an application and to extend or replace its modules. The current version of the readability module supports four languages: English, Spanish, Galician, and Italian. For the languages other than Italian, the pipeline leading to the computation of the metric is the one from the Standard CoreNLP library, included in Tint.

HitEx (Hitta Exempel, "Find Examples") is a hybrid system which uses a combination of machine learning methods and heuristic rules to assign the estimated CEFR level on a 5-level scale (from A1 to C1) to a sentence. The supervised machine learning approach exploits available data to assess the complexity of sentences, considering multiple linguistic dimensions at the same time, while the rules make the selection customizable to task-specific needs. Some of the criteria associated with the sentence goodness are used as filters, as they target negative aspects like incompleteness, non-alphabetical tokens, and anaphora, others as rankers to compute a goodness score. Other than considering the presence or absence of linguistic elements, to make the system more pedagogically aware the latest version of HitEx filters out sensitive vocabulary; adds vocabulary profiling using SVALex (François & al., 2016), a learner-oriented vocabulary list based on coursebooks text with associated frequencies and CEFR levels, and checks for the presence of lemmas associated with a higher proficiency level in the KELLY list (Volodina, Kokkinakis & al., 2012).

Pilán et al. (2017) evaluated this classification method obtaining an accuracy of 63.4% for exact matches and 92% within one CEFR level of distance. The sentence selection algorithm is integrated into the learning platform Lärka (*https://spraakbanken.gu.se/larkalabb/hitex*) and can be accessed through the online graphical interface or as a web service.

5 Results

The sentence selection pipeline was manually evaluated processing 10 random documents extracted from the OpenSubtitles corpus for each language pair and annotating them. The Italian-English documents contained 5575 sentence pairs in total. Out of these, considering the strict filtering algorithm only 430 were chosen. As we said in the beginning, at the moment we are not interested in the recall as we want to be sure the precision of the proposed sentences is sufficiently high so that human intervention is not needed. All 430 sentences were checked and 392 of them were considered appropriate by the authors for the purposes of this work, because they were both good translations, did not contain sensitive vocabulary and were sufficiently context independent to be used for an exercise. The calculated value of precision is 91.16%.

For the English-Swedish pairs the results were not as good, especially because the Swedish sentences were, in the majority of the cases, a summary of the English sentence instead of literal translations. Some examples can be accepted, because the additional token is one that does not influence the meaning, like "well" or "yes", but many pairs had to still be excluded because the syntactic structure differed too much to be used for language learning purposes. In the examined documents, many sentences were also opposite in terms of negative wording/positive wording or active form/passive form between the two languages. For example, the English sentence "You have to give me an injection." is translated with the Swedish "Jag måste få en spruta" (literally: "I need a syringe").

The algorithm selected 598 out of the 5571 pairs present in the documents, with an estimated precision of 72.74%.

Not all the pairs were tagged as "bad" because they were not matching; some were also excluded because the topics discussed were not appropriate for a classroom environment. These kinds of sentences will probably be excluded by HitEx, which contains a sensitive vocabulary filter.

The situation is similar for the Italian-Swedish pairs: some of the sentences used in the test were selected because their structures looked similar, even if the meaning was different. For example, the Italian sentence: "La (DET) tua (DET) proposta (NOUN) è (AUX) una (DET) follia (NOUN)." was aligned with the Swedish "Tiden (NOUN) är (AUX) knapp (ADJ) broder (NOUN)", even if the meaning is completely different – to increase the tolerance of the filter, the non matching PoS tags are ignored if they are auxiliary verbs (AUX), articles (DET), or punctuation marks (PUNCT) in either of the sentences, as the first tests showed these elements do not affect the accuracy of the selection and they exclude many promising sentences. Like with the Swedish-English pairs, some sentences were also tagged negatively for appropriateness reasons.

The documents in total contained 5124 pairs and the algorithm selected 306 of them, with an estimated precision of 69.28%.

Italian ↓ / Swedish →	Basic	Medium	Proficient
Basic	148	45	4
Medium	7	7	0
Proficient	0	1	1

English ↓ / Swedish →	Basic	Medium	Proficient
Basic	251	70	4
Medium	54	27	6
Proficient	13	5	0

English ↓ / Italian →	Basic	Medium	Proficient
Basic	252	32	0
Medium	60	15	0
Proficient	18	6	1

Figure 4. Complexity distribution over the examples

A manual annotation of the grouping of words was not possible because there was no feasible way to avoid bias. To check the usefulness of the separation of words into larger units, we run an experiment using the clusters to further filter the examples and remove the ones in which one sentence is shorter than the other, checking that every cluster has at least one lexical word matching with a word in the other language. This reduces the recall, even if the value is still acceptable, compared to what we would obtain keeping only the sentences in which every lexical word has a match, but the precision improves considerably: 96.10% for the Italian-English pairs, 90.82% for English-Swedish and 83.51% for Swedish-Italian. This confirms that the clusters, especially in the case of function words, can be used to improve the alignment at the group level.

6 Discussion

The sentences are in general short, out of 2024 "good" sentences in the original documents, 690 were

discarded because they had fewer than 5 tokens; and both Tint and HitEx found the majority of them simple: between A1 and A2 with few exceptions for the Swedish sentences, over 60 for Italian and English (Figure 4).

Despite this, the accuracy found for the sentence selection in the annotated pairs was between 70%–90%, in particular: 91.16% for the Italian-English pairs, 72.74% for English-Swedish and 69.28% for Swedish-Italian. Making the selection even more strict using the matching between larger units of words, we reached an accuracy higher than 80% for all the language pairs. This suggests that avoiding human intervention is possible.

Also, the elaborated groups that were found by the clustering algorithm resemble the phrases, even if it was not the intention of this work to assign a meaning to them or classify them.

For example, considering the Italian-English sentence pair: "We have turned this place upside down" / "Abbiamo ribaltato questo posto". The word alignment fails, because the words "upside down" do not have a counterpart in the Italian sentence. Even so, the translation is literal, when we group [have turned upside down], the meaning perfectly matches the Italian verb [ribaltato]. The same thing happens with verb tenses, for example "I have been looking" matches the Swedish "Jag har letat", even if the word "been" on its own is not aligned to any token; or particles, for example in the sentences "It is the murder weapon" / "É l'arma del delitto", the particle "del" in Italian is used to express the same function as "murder weapon" in English. It is easier for a learner of the language to remember the vocabulary and grammatical structure when it is associated with an expression they are familiar with in their native language, even if the form is different, rather than trying to associate a meaning to a specific function word.

At the same time, the presence of the sentence in the other language offers the possibility to generate an exercise in which the solution is not ambiguous and it shows a real example of usage of the word in context in the target language.

Our pipeline extracts the majority of the features directly from the corpus, including sentence alignment, PoS tags, lemmatization and syntactic dependencies for every sentence. The pre-processing operations needed are mostly used to filter the sentences, in order to obtain a high quality in the candidate selection. Other than that, external components are used to align the sentences at the word level and to estimate the target proficiency level for the exercise. This means the system is almost entirely language independent, and every component can be substituted without modifications to other parts of the system. By adding a complexity estimator component for another language, the implementation can theoretically be used with any language pair.

This pre-selection of sentence pairs, associated to a target proficiency level, can be used for different exercises, the sentence reconstruction is only one example but there can be many other situations in language learning in which the translation of the example in another (known) language is useful to reduce the ambiguity and, thus, make it possible to give reliable feedback on the student's solution.

7 Conclusion and Future Work

This work showed an example of how parallel corpora can be used to select a theoretically infinite number of sentences for exercise generation without human intervention. The proposed format was one example but following the same pipeline to pre-process the aligned sentences, many others could be created, for different language pairs. We showed that grouping words into larger units can be used to improve word alignment and shift the focus to syntactic patterns, which have proved to be beneficial for language learners.

The corpus of translated subtitles is interesting in terms of the number of sentences it offers and in their wide variety of genres, from colloquial or slang to narrative – in the case of documentaries. The

problem is that the majority of the sentences are simple and short. This suggests that in a future work it might be interesting to add other ways to make the exercise more difficult for learners of higher proficiency levels, for example giving the base form of the word and asking the learner to change it into the correct one. Other corpora than OpenSubtitles can also be explored to cover the needs of more advanced level students.

Since an issue was also found in the sensitivity of the vocabulary used, the introduction of the subtitles corpus in a language learning environment requires a more elaborate filter, for example excluding categories of movies or programs and making a sub-corpus with only specific film classifications.

In future work, other than making the exercise more elaborate, it would be important to run more accurate tests, with real users, to prove the usefulness of this approach from a pedagogical point of view. This is fundamental because while the annotations can tell us if the selected sentences are "good" translations or if the alignments are correct, it is the users' interactions which provide valuable feedback to decide whether or not a system can be applied in real-world contexts.

Tests with real users will also allow the creation of a profile of the learners in order to give feedback appropriate to their level and follow their progresses, suggesting increasingly complex exercises to solve.

The sentence selection process could be improved to consider the semantics of the tokens, which has been ignored in this work, for example to check if the meaning of two aligned tokens is in the same semantic space. This would allow to automatically exclude all those sentences in which the structure is similar but the meaning is different.

The entire pipeline, apart from the readability estimator, is language independent, so it could be easily extended to other language pairs, even if it would be necessary to test if the accuracy does not decrease, and to add a different complexity measure, if a tool for the new language is not available.

Acknowledgements

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".

References

Ahn, L. von (2013). *Duolingo: Learn a language for free while helping to translate the web*. Proceedings of the 2013 international conference on Intelligent user interfaces. ACM, pp. 1–2.

Alfter, D., Borin, L., Pilán, I., Tiedemann, T. L. & Volodina E. (2018). *From language learning platform to infrastructure for research on language learning*. CLARIN Annual Conference 2018, p. 53.

Alfter, D. & Graën, J. (2019). *Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?* Proceedings of the 22nd Nordic Conference on Computational Linguistics, pp. 321–326.

Aprosio, A. & Moretti G. (2018). *Tint 2.0: An allinclusive suite for NLP in Italian*. Academia University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Boulton, A. (2016). *Integrating corpus tools and techniques in ESP courses*. ASP [Online], 69, pp. 113–137.

Brown, P., Pietra, S. D., Pietra, V. D. & Mercer, R. (1993). *The mathematics of statistical machine translation: Parameter estimation*. Computational Linguistics 19, pp. 263–311.

Burstein, J., Madnani, N., Sabatini, J., McCaffrey, D., Biggers, K. & Dreier, K. (2017). *Generating language activities in realtime for English learners using language muse*. Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, pp. 213–215.

Byrd, P. & Coxhead, A. (2010). *On the other hand: Lexical bundles in academic writing and in the teaching of EAP*. University of Sydney Papers in TESOL 5, pp. 31–64.

Cobb, T. & Boulton, A. (2015). *Classroom applications of corpus analysis*. D. Biber & R. Reppen (eds), Cambridge handbook of English corpus linguistics (pp. 478–497). Cambridge University Press.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, Cambridge University Press.

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). *Svalex: a CEFRgraded lexical resource for Swedish foreign and second language learners.* Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 213–219.

Horsmann, T. & Zesch, T. (2014). *Towards automatic scoring of cloze items by selecting low-ambiguity contexts*. Proceedings of the third workshop on NLP for computer-assisted language learning, pp. 33–42.

Katinskaia, A., Nouri, J. & Yangarber, R. (2017). *Revita: a system for language learning and supporting endangered languages*. Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, pp. 27–35.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). *GDEX: Automatically finding good dictionary examples in a corpus*. Proceedings of the 13th EURALEX International Congress. Spain, July 2008.

Klimova, B. & Kacet J. (2017). *Efficacy of Computer Games on Language Learning*. TOJET: The Turkish Online Journal of Educational Technology 16, issue 4 October 2017.

Laufer, B. & RavenhorstKalovski G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22, 15–30.

Lawson, A. (2001). *Collecting, aligning and analysing parallel corpora*. In Ghadessy, M., Henry, A., & Roseberry, R.L. (eds.). Small corpus studies and ELT: Theory and practice (pp. 279–310). John Benjamins Publishing Company.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Lison, P. & Tiedemann, J. (2016). *OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

Lucisano, P. & Piemontese, M. (1988). *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana*. Scuola e Città, a. XXXIX.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D (2014). *The Stanford CoreNLP natural language processing toolkit*. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B. & Ziai, R. (2019). Scaling up intervention studies to investigate reallife foreign language learning in school. *Annual Review of Applied Linguistics*, 39, 161–188.

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V. & Ott, N. (2010). *Enhancing authentic web pages for language learners*. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications.

Meyer, N., Wojatzki, M. & Zesch, T. (2016). *Validating bundled gap filling - Empirical evidence for ambiguity reduction and language proficiency testing capabilities*. Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition.

Minack, E., Siberski, W. & Nejdl W. (2011). *Incremental diversification for very large sets: a streaming-based approach*. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, New York, NY, USA.

Mishan, F. & Strunz, B. (2003). *An application of XML to the creation of an interactive resource for authentic language learning tasks*. Cambridge University Press.

Östling, R. & Tiedemann, J. (2016). *Efficient Word Alignment with Markov Chain Monte Carlo.* The Prague Bulletin of Mathematical Linguistics N. 106, 2016, 125–146. doi: 10.1515/pralin-2016-0013.

Paetzold, G. & Specia, L. (2016). *Collecting and Exploring Everyday Language for Predicting Psycholinguistic Properties of Words*. Proceedings of COLING 2016, 26th International Conference on Computational Linguistics: Technical Papers, 1669-1679.

Pilán, I., Volodina, E. & Johansson, R. (2013). *Automatic Selection of Suitable Sentences for Language Learning Exercises*. 10.14705/rpnet.2013.000164.

Pilán, I., Volodina, E. & Johansson, R. (2014). *Rule-based and machine learning approaches for second language sentence-level readability*. Proceedings of the ninth workshop on innovative use of NLP for building educational applications, 174-184.

Pilán, I., Volodina, E. & Zesch, T. (2016). *Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks*. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.

Pilán, I., Volodina, E. & Borin, L. (2017). *Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation*. Traitement Automatique des Langues (TAL), special issue on NLP for learning and teaching. 57.

Potter, A. (2004). Interactive rhetoric for online learning environments. *The Internet and Higher Education*, 7, 183–198.

Rudzewitz, B., Ziai, R., Kuthy, K. D. & Meurers, D. (2017). *Developing a webbased workbook for English supporting the interaction of students and teachers*. Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017. Linköping Electronic Conference Proceedings.

Segler, T. (2007). *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Institute for Communicating and Collaborative Systems, University of Edinburgh.

Simpson, J. & Weiner, E. S. (1989). *Oxford English dictionary online*. Oxford: Clarendon Press.

Teubert, W. (1996). Comparable or parallel corpora? *International Journal of Lexicography, 9*(3), 238-264.

Tiedemann, J. (2012). *Parallel Data, Tools and Interfaces in OPUS*. Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, European Language Resources Association (ELRA).

Tonelli, S., Manh, K., & Pianta, E. (2012). *Making readability indices readable*. Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Montr´eal, Canada, June. Association for Computational Linguistics.

Volodina, E., Kokkinakis, S. & Johansson, R. (2012). *Semiautomatic selection of best corpus examples for Swedish: Initial algorithm evaluation*. Proceedings of the SLTC 2012 workshop on NLP for CALL, Linköping Electronic Conference.

Volodina, E., Pijetlovic, D., Pilán, I. & Johansson, S. (2013). *Towards a gold standard for Swedish*

CEFRbased ICALL. Proceedings of the Second Workshop on NLP for ComputerAssisted Language Learning, NEALT Proceedings Series. Vol. 17.

Volodina, E., Pilán, I., Borin, L. & Tiedemann, T. (2014). *A flexible language learning platform based on language resources and web services*. Proceedings of LREC 2014, Reykjavik, Iceland.

Volodina, E., Pilán, I., Eide, S. R. & Heidarsson, H. (2014). *You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language*. Proceedings of the third workshop on NLP for computer-assisted language learning.

Volodina, E. (2020). *Common Pitfalls in the Development of ICALL Applications*. https://spraakbanken.gu.se/blogg/index.php/2020/04/30/common-pitfalls-in-the-development-of-icall-applications. Online; accessed November 2020.

Wilson, E. (1997). The automatic generation of CALL exercises from general corpora. In Anne Wichmann, Steven Fligelstone, Tony McEnery, & Gerry Knowles (Eds.). Teaching and language corpora (pp. 116–130). London: Routledge.

Wu, S., Fitzgerald, A. & Witten, I. (2014). *Second Language Learning in the Context of MOOCs*. Proceedings of the 6th International Conference on Computer Supported Education.

Arianna Zanetti graduated from the Alma Mater Studiorum University of Bologna (Italy) in Computer Science Engineering, choosing as the main focus of her career the fields of Artificial Intelligence and Natural Language Processing. She then obtained a master's degree in Language Technology from the University of Gothenburg, where she worked on a project for Intelligent Computer-Assisted Language Learning for her thesis.

Elena Volodina (PhD, Docent) is a Researcher at Språkbanken, the Department of Swedish at the University of Gothenburg, Sweden. She has been active with development of electronic resources and applications for second language learning, with focus on Intelligent Computer-Assisted Language Learning, exercise generation and corpus-based text research. She is currently engaged in building national infrastructure for research on Swedish as a Second Language, SweLL, and in creating and analyzing lexical and grammatical profiles for immigrant Swedish.

Johannes Graën did his PhD on alignment methods for parallel corpora. One of his main interests is the utilization of those corpora for building language learning applications. He received a grant from the Swiss National Science Foundation for carrying on his research on corpus-based Computer-Assisted Language Learning applications with the help of Språkbanken (University of Gothenburg, Sweden) and Grael (Pompeu Fabra University, Spain).