Mining a parallel corpus for automatic generation of Estonian grammar exercises

Antoine Chalvin, Egle Eensoo, François Stuck

Mining a parallel corpus for automatic generation of Estonian grammar exercises

Antoine Chalvin, Egle Eensoo, François Stuck

Institut national des langues et civilisations orientales (INALCO)
65 rue des Grands-Moulins, 75013 Paris, France
E-mail: antoine.chalvin@inalco.fr, egle.eensoo@inalco.fr, francois.stuck@inalco.fr

Abstract

The aim of our research is to develop a system to generate Estonian grammar exercises for French-speaking learners, based on a large lemmatised parallel corpus (http://corpus.estfra.ee) and on the data of the Comprehensive French–Estonian Dictionary (http://www.estfra.ee). We concentrate on exercises on nominal and verbal morphology. Although the corpus is not syntactically tagged, we also explore the possibilities of generating some types of syntax exercises. The system generates on demand exercises consisting of a specified number of Estonian sentences, in which relevant word forms are replaced by their lemmas. The learner has to construct the right form and can check his or her answers. Sentences are accompanied by their French translation. In this article, we concentrate on the problems related to the definition and tuning of sentence selection criteria. Exercises can be generated at three levels of difficulty. Relevant sentences are picked up in the corpus according to their length and the "frequency" of the lemmas they contain, i.e. the presence of the lemmas in one of the four subsets of headwords specified in the data of the dictionary: basic vocabulary (4000 words), small dictionary (10 000 words), lower-medium dictionary (15 000 words), and upper-medium dictionary (40 000 words).

Keywords: parallel corpora; readability; e-learning; Estonian as a foreign language; grammar exercises

1. Background and objectives

Since the 1990s there has been a growing interest in using corpora for language learning purposes (see Boulton, 2008; Huang, 2011). One of the key approaches in this field is 'data-driven learning' (DDL), which has been described as an "attempt to cut out the middleman" and to give the learners "direct access to the data" (Johns 1994: 297). In practice, the DDL, which focuses on the use of corpus concordances in the classroom, still supposes the guidance of a teacher. A more effective way to really "cut out the middleman" is to develop systems that use corpora as a source to generate self-correcting tests. An impressive number of test generation systems have been developed in the field of EFL (English as a Foreign Language), mainly to generate vocabulary tests in multiple-choice format (e.g. Coniam, 1997; Gao, 2000; Mitkov & Ha, 2003; Hoshino & Nakagawa, 2005; Brown et al., 2005; Liu et al., 2005; Sumita et al., 2005; Kilgarriff et al., 2010), and more rarely grammar tests (Chen et al., 2006; Lee & Seneff, 2007; Hoshino & Nakagawa, 2008). For French, the GramEx system developed by Beltrachini, Gardent & Kriszewski (2012) is not based on corpora, but on a grammar-based sentence generation process.

The aim of our project is to develop a system to automatically generate fill-in-the-blank Estonian grammar exercises consisting of authentic sentences. Fill-in-the-blank exercises are widely used in foreign language learning to help build grammar proficiency. One of their drawbacks is that they usually consist of specially designed sentences, which do not necessarily reflect real language use. The other drawback of manually designed exercises is that, since their creation is very time-consuming, textbooks and learning environments usually propose a limited number of them, which is not sufficient for the learner to acquire full proficiency on the specific points dealt with in the exercises. Our idea is that the automatic generation of exercises from a corpus of authentic language material could remedy these drawbacks and offer the learner the possibility to continue building his/her grammatical proficiency after he/she has completed all the exercises in his/her textbook. The system we want to develop is thus conceived as complementary to traditional language learning materials. It may address the needs of elementary, intermediate or advanced learners, but probably not those of complete beginners. Its implementation is complicated by a number of difficulties related to the quality of the corpus and the definition of complexity (readability) criteria for sentence selection. Our main concern, in the first stage of the project, is not so much pedagogical as computational: we want to determine how to process a large corpus of real unmodified texts in order to make it a suitable source for generating L2 grammar exercises. In other words: how to extract from a general language corpus a specific subcorpus more fitted to the needs of foreign language learning? And what kind of grammar exercises is it possible to create on the basis of a morphologically tagged corpus?

2. The Estonian-French parallel corpus

Our system is based on the Estonian-French parallel corpus (CoPEF: http://corpus.estfra.ee) compiled by the French-Estonian Lexicography Association (Prantsuse-eesti leksikograafiaühing, Tallinn). The corpus was designed primarily to address the needs of lexicographers compiling a comprehensive Estonian-French dictionary of 110 000 entries (GDEF: http://www.estfra.ee). Considering this specific purpose and the relatively limited number of available bilingual texts, the main principle followed in the compilation of the corpus was to attain the critical mass needed for lexicographical work, and not to produce a balanced corpus. The whole corpus contains 65 million words and is subdivided into seven subcorpora:

- Estonian literature (3.85 million words),
- French literature (4.09 million words),
- Estonian non-fiction (132 000 words),

- French non-fiction (990 000 words),
- European Union legislative texts (26.3 million words),
- Debates of the European Parliament (28.2 million words),
- Bible (1.4 million words).

The corpus is lemmatised and morphologically tagged. Estonian texts were tagged with Estmorf (cf. Kaalep 1996, 1998) and disambiguated with Tahmm (Tahmm, 1998). But the result is not 100% reliable. Tahmm does not always choose the right variant. In some cases it is not able to disambiguate and results in several variants. This occurs especially when the variants refer to the same grammatical form and differ only in their lemmas (Tahmm, 1998). Potential mistakes in morphological analysis will have to be taken into account when designing the exercises. In order to reduce their impact, it is necessary to avoid exercises based exclusively on specific forms that Tahmm has difficulty identifying. For example, we will not propose specific exercises on the formation of singular genitive, because some of the "genitive" forms that the learner would have to build could be in fact the singular partitive or singular nominative of the same word (homography between these three forms is quite frequent). We can propose instead more global exercises on nominal morphology, including genitive and partitive forms, but without specifying which of these cases is concerned in each question.

Sentence-level alignment of the corpus was made at different periods with different tools, either automatically (for EU texts) or semi-automatically (for other subcorpora). In the latter case, alignments with a low probability index were controlled and corrected manually. A few literary texts were aligned fully manually. The reliability of alignments was not precisely estimated, but there are obviously mistakes, which might cause problems in the exercises by giving wrong French translations to Estonian sentences.

For exercise generation purposes, we decided to exclude the EU legislative subcorpus, which contains a high proportion of long sentences, repetitive formulae and technical vocabulary. We also excluded the Bible, from which the Estonian and French translations included in the corpus are stylistically marked and do not represent standard contemporary language. However, the remaining subcorpora also contain many sentences which could be difficult to understand for language learners. Generating "good" grammar exercises thus implies selecting sentences fitted to the proficiency level of the learner, which means evaluating the readability of the sentences.

3. Selection of sentences, readability criteria

3.1 Previous work

Works on readability started in the early 40's (Dale & Chall, 1948; Flesch, 1948), mainly to improve native learners' reading skills. They used surface textual features, such as the average number of words or sentences, or the proportion of words not belonging to the basic vocabulary, combined through a linear regression model to set out simple readability formulae. Although this approach gave some acceptable results, it was criticised for its simplicity. Later works (Kintsch & Vipond, 1979; Redish & Selzer, 1985; Meyer, 1982) introduced more complex features, such as text cohesion, information density or macrostructure, but in fact for little gain. During the last fifteen years, with the progress and spread of corpus and NLP techniques, such as automatic classification, works on readability have been renewed (Collins-Thompson & Callan, 2004; Feng et al., 2010; François & Fairon, 2012). More and more complex features covering various linguistic fields (lexical, syntactic, semantic, discursive) are now implemented and evaluated for various languages. As for Estonian, work has been done since the 70's on the readability of textbooks for native speakers. A readability formula was proposed by Mikk (1980, 1991), based on two criteria: average length of independent sentences and abstraction level of repeated nouns.

Beyond its technical aspect we should not forget that the very notion of readability has several meanings, and most of them concern whole texts. For example, one can assess the readability of a text by testing its global understanding through the ability of writing an abstract or answering questions.

Moreover, the works on readability often differ when targeting the mother tongue (L1) or a foreign language (L2). Some works deal with French as a second language (Henry, 1975; Richaudeau, 1979; Daoust et al., 1996; François & Fairon, 2012). We are not aware of any similar work dealing with readability of Estonian as a second language.

Being concerned more, in this study, by short text segments or sentences than whole texts, our point of view on readability will follow that of Kilgarriff: "intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context. We call this its 'readability'" (Kilgarriff et al., 2008).

So we will define readability as the ability for a learner to understand the constituents and the structure of a sentence, sufficiently to modify or complete it.

It is known that cultural knowledge and familiarity with the domain facilitate the comprehension process. Nevertheless, as we are working with a bilingual corpus of general language and can provide the translation of any text segment, we assume, in this study, that the impact of world knowledge on readability, as we defined it above,

is largely neutralised and that the readability of a sentence, for a foreign language learner, depends mainly on two characteristics: its syntactical complexity and its lexical complexity.

3.2 Syntactical complexity

The intuitive meaning of the notion of syntactical complexity at sentence level can be defined in formal terms as the number of nodes in the parse tree of the sentence. In practice, this criterion is not applicable to large corpora, because identifying and counting nodes generally requires manual coding (Szmrecsányi, 2004: 1033).

A more automatable approach could consist in counting certain types of surface units which qualify as good indicators of structural complexity, such as subordinating conjunctions and relative pronouns, or commas in languages where they function mainly as clause separators (for Estonian, see e.g. Kerge, 2002). The drawback of this method is that it is language-specific: subordinating units are different in each language, and this type of units might not be pertinent for languages in which subordination is not materialised by specific words or in which complexity can be achieved by means other than subordination.

Another criterion of complexity which has been widely used is sentence length (i.e. the number of words of the sentence). It has the advantage of being language-independent and very easy to implement. It seems also quite pertinent. A comparison conducted on 50 English sentences suggests that counting words gives almost the same complexity rankings as counting the nodes or calculating a complexity index based on the number of subordinating units, verbal forms and noun phrases (Szmrecsányi, 2004). It seems indeed quite logical that long sentences are structurally more complex than shorter ones, even if there may be exceptions. Since counting words is the most economical method and gives very consistent results, we decided to adopt this criterion to evaluate the syntactical complexity of the sentences. We intuitively defined three length ranges: up to 10 words, from 11 to 15 words, and from 16 to 29 words. For a language such as Estonian, which uses fewer function words than English or French (it has no article and 14 declension cases which notably reduce the use of pre- or postpositions), adding five words to a sentence generally results in a significant increase in syntactical complexity.

If excessively long sentences are difficult to understand by language learners, sentences that are too short can also cause problems, because they are understandable only within a larger context. Three words seemed to be a minimum for an Estonian sentence to constitute a sufficiently clear and autonomous message. We thus excluded sentences shorter than three words.

3.3 Lexical complexity

Since the corpus is not balanced, we could not take as a criterion for evaluating lexical

complexity, the frequency of the lemmas in the corpus. Neither did we find reliable external data on the frequency of Estonian words. The first frequency dictionary of contemporary Estonian (Kaalep & Muischnek, 2002) is not fully satisfying, as it was made from a very small corpus (1 million words) and contains only 10 000 words. A newer frequency list, based on a larger corpus (15 million words), was recently released (http://www.cl.ut.ee/ressursid/sagedused1/). Although much more comprehensive (40 000 lemmas), it still contains some oddities (from a pedagogical point of view), such as the presence of very specific terms among the most frequent words, or very different rankings of words belonging to the same semantic series. We thus decided to evaluate the lexical complexity of sentences on the basis of manually compiled or checked word lists, i.e. the subsets of the GDEF.

The GDEF is divided into four subsets of entries: basic vocabulary (4000 words), small dictionary (10 000 words), lower-medium dictionary (15 000 words), and upper-medium dictionary (40 000 words). These headword lists have been established by GDEF lexicographers, who used as a basis the above mentioned frequency dictionary as well as entry lists compiled by the Institute of Estonian Language for an Estonian Fundamental Dictionary (Eesti keele põhisõnastik) and for a general bilingual dictionary base with Estonian as a source language (Eesti-X sõnastikupõhi). These lists compiled for lexicographical purposes appeared more consistent and better suited to pedagogical purposes than automatically calculated frequency lists. A reason for that is probably the fact that entry selection principles followed by lexicographers compiling small or medium dictionaries are somewhat similar to those followed by authors of language textbooks (priority given to concrete notions and words of everyday life, consistency of semantic series, etc.). The four subsets of the GDEF give us four levels of lexical complexity.

3.4 Global sentence complexity and its relationship with language proficiency

Combined with the three levels of syntactical complexity, the four levels of lexical complexity give us 12 categories. This classification is obviously too complex to be understandable by the learner. It has to be reduced to a limited number of proficiency levels. One has to determine which combinations of lexical and syntactical complexity give sentences that can be understood without too much effort (and with the help of the translation) by learners of each level. A quick evaluation led us to the following table of equivalences, which remains a working hypothesis and needs to be confirmed by a more comprehensive assessment. Proficiency levels are expressed according to the categories of the Common European Framework of Reference for Languages.

LC \ SC	1	2	3
1	A2	B1	B2
2	B1	B1	B2
3	B1	B1	B2
4	B2	B2	B2

Table 1: Sentence complexity and language proficiency
(LC: lexical complexity; SC: syntactical complexity)

3.5 Sentence selection process and results

The bitexts of the CoPEF corpus are aligned at a so-called segment level. A segment is usually a sentence, but not always. It can also be a set of sentences or a sentence chunk (see Table 2 below).

Before applying any complexity selection on the corpus segments, a filtering is made to keep only the valid ones. The segment validation process follows the rules here below.

	multi-sentence	**single sentence**	**sentence chunk**
Estonian literature	4,980	80,006	40,296
French literature	4,297	115,021	46,019
Estonian non-fiction	85	1,906	301
French non-fiction	973	16,573	4,264
European Parliament	26,506	532,630	63,279
TOTAL	**11,279**	**497,511**	**561,116**

Table 2: Types of segments and their number per subcorpus

1. The segment must not be a sentence chunk, but a set of one or more “well-formed” sentences, i.e. it must start with an upper-case letter and end with a strong punctua-

tion; it must contain at least one finite verb; it must contain more than two words but fewer than thirty.

2. The segment must contain only acceptable words, i.e. words which are either a supposed proper nouns or an entry in one of the four subsets of the GDEF dictionary.

The resultant set of valid segments is then broken up into twelve subsets combining the four lexical and the three syntactic complexity levels (Table 3).

A final step reduces them to three segment sets according to the patterns of Table 1. They correspond to the three desired proficiency levels.

The numbers of segments for each level are as follows: A2: 22 558; B1: 21 758; B2: 10 862. As can be seen from the table below, the percentage of selected segments is quite low (5.9% of the total). It is significantly lower for the European Parliament subcorpus than for the other subcorpora, and, among the latter, significantly higher for French literary texts. This reflects, on the one hand, the higher lexical complexity of European Parliament debates (more technical terms) and, on the other hand, the lesser complexity of Estonian literary translations, as compared with Estonian original texts.

		Estonian literature	French literature	Estonian non-fiction	French non-fiction	European Parliament	TOTAL
Corpus total size		125 282	165 337	2 292	21 810	622 415	937 136
LC1	**SC1**	3 247	6 454	25	443	12 389	22 558
	SC2	304	308	4	43	1 725	2 384
	SC3	52	45	5	13	413	528
LC2	**SC1**	1 793	3 662	35	376	5 039	10 905
	SC2	422	486	22	95	1 851	2 876
	SC3	134	128	3	35	751	1 051
LC3	**SC1**	639	1 287	14	150	2 099	4 189
	SC2	174	228	6	42	954	1 404
	SC3	60	77	7	30	546	720
LC4	**SC1**	843	1615	14	180	2 759	5 411
	SC2	288	371	17	83	1 334	2 093
	SC3	109	145	8	38	759	1 059
Total number of selected segments		8 065	14 806	160	1 528	30 619	55 178
% of selected segments		6,4	9,0	7,0	7,0	4,9	5,9

Table 3: Number of segments at different complexity levels in the corpus
(LC: lexical complexity; SC: syntactical complexity)

4. Converting sentences into exercises

4.1 Types of exercises

Taking into account the main difficulties of learners of Estonian as a foreign language, we generate two types of exercises, aimed at developing two types of language competence: 1) morphological competence (constructing forms), and 2) syntactical competence (choosing the appropriate form in a given context).

Morphological exercises present the user with sentences in which one inflected verb or substantive has been replaced by a textbox containing the corresponding lemma. Each exercise deals only with one type of form (e.g. partitive plural or indicative present), so the user knows which case and number or tense and mood has to be used and his/her task consists only of constructing the form and typing it in the text box. We generate this type of exercise for all declension cases (except singular nominative) and for the main verbal forms (present indicative, simple past indicative, present conditional, present imperative). For verbal forms, we give an additional hint after the lemma that tells the user which person has to be used, because there are many sentences in which the person cannot be predicted from the context. The French translation can help the user to disambiguate in many, but not all, cases. Performing separate exercises on each person would be too monotonous for the learner.

Syntax exercises are more difficult to generate, because the corpus is tagged only morphologically. It is still possible to imagine some types of syntax exercises relying only on morphological tags. The most obvious topic that can be dealt with is the use of declension cases: the user is presented sentences in which various case forms are replaced by textboxes with the corresponding lemmas. He/she must find which case has to be used in the context and construct the inflected form. Exercises can either mix all cases indifferently or concentrate on a certain subset of cases which can be used for similar syntactic purposes (e.g. nominative, genitive and partitive, which in Estonian can all be used to mark the object, depending on the context, or the so-called local cases, which are used to form adverbials of place or direction). For successfully performing this type of exercise, the learner needs to see the translation, otherwise many forms are impossible to predict unequivocally. An alternative possibility is to provide at the beginning the list of all inflected forms which have to be placed in the different sentences.

Another syntax topic on which we can generate exercises is the use of adpositions (postpositions and prepositions). In each sentence an adposition is replaced by a textbox. The user has to find the adposition fitting to the context (adpositional reaction of a verb or a nominal) and/or to the meaning of the sentence (here also translation is necessary). The list of adpositions which have to be placed in the blanks can be given or not in the beginning of the exercise.

We also consider the possibility of generating exercises on particle verbs, taking as a basis the list of verbs identified as such in the GDEF (1411 particle verbs combining one of 460 simple verbs with one of 67 adverbial particles). The user would be asked to identify in a list the appropriate particle (or the appropriate couple verb-particle) to fill the blank(s) in a sentence. A specific problem for generating that type of exercise is the fact that the particle can be placed either in the left context of the verb (with infinitives and participles) or in the right context (with finite forms). In the latter case, it is often separated from the verb by other constituents. Furthermore, many particles can also be used as adverbs, in which case they do not form a lexical unit with the verb. On the whole, automatically identifying particle verb constituents in order to create exercises seems possible, but rather tricky. We identified possible solutions, but left their implementation as a direction for further work.

4.2 Generation process

4.2.1 Exercise definition and configuration

Through an HTML form (Fig. 1), the user is asked to define the type of the desired exercise, i.e.:

- its class (e.g. nominal or verbal morphology, use of cases, adpositions, particle verbs);
- its precise content (e.g. case and number for nominal morphology, mood and tense for verbal morphology).
- The user must then specify the source of segments from which the exercise items are to be generated. He or she will define:
- the set of subcorpora to be used,
- the proficiency level.

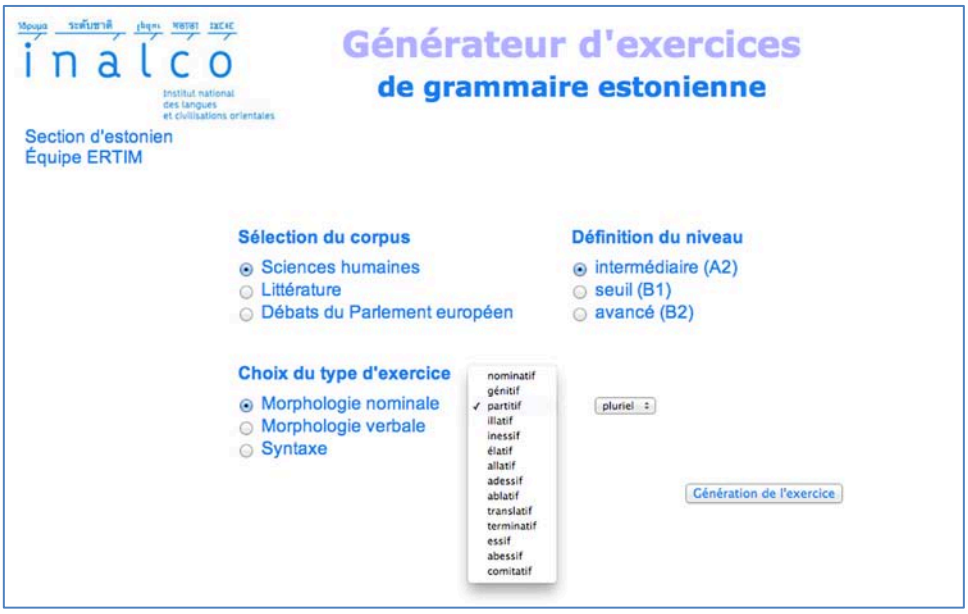

Figure 1: Screenshot of the exercise generator

Some hidden parameters, automatically set, help control item generation and exercise layout.

4.2.2 Exercise generation and display

The generation process first selects candidate-items. To do so, it obtains the list of tagged Estonian segments of the desired level from the chosen subcorpora. Then it parses them at both morphological and syntactical level to filter out any segments that do not fit the specified type of exercise, or that would lead to some identified ambiguities (e.g. we filter out verbal forms ending with the emphatic particle *-gi/-ki*, which is not tagged).

Among the candidate items, a very limited number are selected to be 'blanked out' and become part of the exercise, according to the following principles:

- one blank per item (or more than one for the advanced level, if the sentence length allows it);
- a similar lemma will never be reused as a blank within the current exercise (this is necessary to avoid over-representation of very frequent words, such as the verb *olema* 'to be' in verbal morphology exercises);
- items are chosen randomly.

The French translation is then retrieved and associated to the item. A complementary feature could consist of linking each lemma of the item to the corresponding article of the GDEF. This would assist the learner in developing his/her lexical knowledge and overcoming possible comprehension difficulties due to loose translation of the segment (quite frequent in literary texts). The implementation of this feature will become relevant when at least one subset of the GDEF is fully available, which is not yet the case.

The requested exercise is generated as an XML document describing, on one hand, the different items (Estonian blanked out text, French translation, answer), and, on the other hand, the various generation and layout parameters. An XSL style-sheet transforms it into a dynamic HTML document.

The exercise generator provides the user with an HTML fill-in-the-blank exercise (Figure 2) with classical functionalities, like "answer evaluation", "reset", "answers" and various help modes (lemma in the blank, list of possible answers, no help at all).

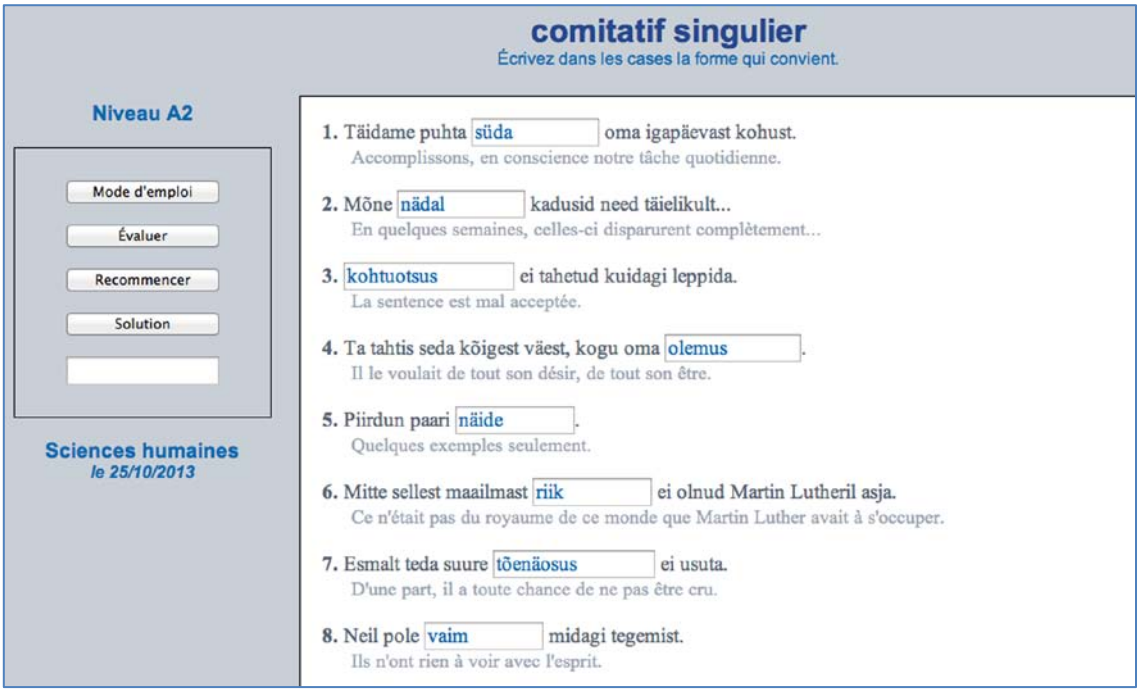

Figure 2: Screenshot of an exercise on comitative singular

4.3 Results and evaluation

In the last stage of the project, it will of course be necessary to have all types of exercises evaluated by learners of Estonian as a foreign language at different proficiency levels. At the present stage, we evaluated the linguistic and pedagogical relevance of 991 automatically generated exercise items, selected randomly among the 6454 A2-level (LC1-SC1) segments of the French literature subcorpus (and also for adposition exercises in the LC2-SC2 and LC3-SC3 segments of the same subcorpus). This preliminary evaluation was made by Antoine Chalvin, in the light of his 15 years' experience of teaching Estonian grammar to French students. It appeared that the overwhelming majority of items were linguistically pertinent (the form in the blank corresponded to the topic of the exercises) and pedagogically appropriate (blanks were possible to fill with the help of hints, the context and/or the translation). Exercises on verbal morphology had the highest reliability rate (97%), followed by exercises on case forms other than genitive and partitive singular (91%). Exercise on these last two forms contained, as expected, a significant number of errors (only 77% of the items were adequate). Exercises on adpositions were the least reliable (67%).

The detailed analysis of exercises revealed several types of problems, which made some items difficult or disconcerting for the learner.

A first category of problems was caused by errors in lemmatisation or morphological analysis. At this stage, we were unable to solve this problem, because identifying and correcting errors in the corpus would have been very time consuming. In the

exercises we generated, we discovered a few recurrent errors which could be searched and corrected semi-automatically in the corpus. For example, several verb forms ending in *-ta* (factitive derivational suffix or infinitive ending) were wrongly analysed as nouns in the abessive case (the abessive suffix is *-ta*), several active past participles (in *-nud*) were analysed as plural nominative of substantives in *-nu* (which is a far less common form), several postpositions or adverbs ending in *-l* were analysed as adessive forms of substantives (suffix *-l*), etc. If correcting errors in the corpus proves too difficult, another way to solve the problem would be to generate a list of ambiguous forms and exclude them from exercises in which a confusion is possible (e.g. in an exercise on the translative case, never create a blank on the form *peaks*, which, though analysed as the translative singular of *pea* 'head', could in fact be the conditional present of the verb *pidama* 'have to').

A pedagogical problem which affected mainly exercises on adpositions was the possibility of multiple correct answers, either because the translation was not sufficient to specify the meaning of the sentence, or because, although the meaning was clear, several synonym adpositions could be used, but only one of them being recognised as correct by the automatic correction system. This could be frustrating and disconcerting for the learner. A possible way to reduce the impact of this problem could be to make a list of synonym adpositions (such as *saadik* and *peale* 'since', *seas* and *hulgas* 'among') and instruct the system to accept them as correct variants.

The problem of multiple answers also affects exercises dealing with plural forms of substantives, because Estonian has two plural paradigms. The so-called *i*-plural, usually very rare, nonetheless occurs rather frequently for certain words as a variant of the more common *de*-plural (*aastail* vs. *aastatel* 'in the years'; *päevil* vs. *päevadel* 'in the days'). The morphological tags in the corpus do not distinguish these variants. However, in the 991 items analysed, we found very few *i*-plural forms.

A third problem affects morphology exercises combining several forms (e.g. several persons in verb exercises, or several cases in multi-case exercises), namely, the excessive predominance of certain forms in the questions. One of the forms dealt with in a given exercise could be much more frequent in the corpus than the other forms. If exercise items are picked up randomly in the corpus, this particular form has chances to be more present also in the exercise, leaving little space for the others. This is the case, for example, in our conjugation exercises, where the third person singular concerns at least 60% of the items. To reduce monotony and maximise the usefulness of these exercises, it will be necessary to find a way to balance the representation of the forms.

The last (minor) problem is excessive easiness. In exercises on nominal and verbal morphology, many forms are very easy to construct, because the stem serving as a basis (singular genitive for substantives, indicative present stem for verbs) is easily predictable from the lemma. In order to make exercises more interesting and more

useful for the learner, we should find a way to over-represent problem words, i.e. words whose radical is not predictable from the lemma. Lists of such words could be easily generated with the aid of morphological data included in the GDEF.

5. Conclusion

By applying sentence readability criteria to a large real language corpus of around 940 000 segments, we generated a 'readable' corpus of 55 000 segments. We showed that, on the basis of such a corpus, it is possible to generate a very high number of fill-in-the-blank grammar exercises that can serve as a useful training material for learners of Estonian, without it being necessary to submit these exercises to prior manual control and filtering by a language teacher. On the whole, generated exercises have a surprisingly high degree of pertinence and reliability. Residual problems, such as lemmatisation errors, possibility of multiple answers, monotony of questions and excessive predictability of answers, do not seem insurmountable and will be addressed in a second stage of the project. Once operational, the system will be made freely available on the Internet.

A possible further development, on the basis of the same corpus, could be a French grammar exercise generator for Estonian learners. This would probably be even easier to implement, due to the lower frequency of morphological homography in French as compared with Estonian.

The general methodology of our project and large parts of the program could also be applied to other language pairs for which a reliable morphologically tagged parallel corpus of general language is available.

6. References

Beltrachini, L., Gardent, C. & Kruszewski, G. (2012). Generating Grammar Exercises. In *The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT Worskhop 2012*. Montreal, Canada.

Boulton, A. (2008). Esprit de corpus: promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et Corpus*, 3, pp. 37-46.

Brown, J. C., Frishkoff, G. A. & Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Pp. 819-826.

Chen, C.-Y., Liou, H.-C. & Chang J. S. (2006). FAST – An Automatic Generation System for Grammar Tests. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.* Sydney: Association for Computational Linguistics.

Collins-Thompson, K. & Callan, J. (2004). A language modeling approach to

predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. Boston, pp. 193-200.

Coniam, D. (1997). A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Cloze Tests. *CALICO Journal*, 2-4, pp. 15-33.

Dale, E. & Chall, J.S. (1948). A formula for predicting readability. *Educational research bulletin*, 27(1) pp. 11-28

Daoust, F., Laroche, L. & Ouellet, L. (1996). SATOCALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1), pp. 205-234.

Feng, L., Martin Jansche, M., Huenerfauth, M., Elhadad, N. (2010). Comparison of Features for Automatic Readability Assessment. In *Proceedings of Coling 2010 (Poster Volume)*, Beijing, pp. 276-284.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3) pp. 221-233.

François, T. & Fairon, C. (2012). An AI readability Formula for French as a Foreign Language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, Jeju, South-Korea, pp. 466-477.

Henry, G. (1975). *Comment mesurer la lisibilité ?* Bruxelles: Labor.

Hoshino, A. & Nakagawa, H. (2005). A Real- Time Multiple-Choice Question Generation for Language Testing: A Preliminary Study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Ann Arbor, Michigan, pp. 1-8.

Hoshino, A. & Nakagawa, H. (2008). A Cloze Test Authoring System and Its Automation. Advances in Web Based Learning. In *ICWL 2007 : 6th International Conference Edinburgh, UK, August 15-17, 2007*. Berlin/Heidelberg: Springer, pp. 252-263.

Huang, L.-S. (2011). Corpus-aided language learning. *ELT Journal*, 65(4), pp. 481-484.

Johns, T. (1994). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (ed.). *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press, pp. 293-313.

Kaalep, H.-J. (1996). ESTMORF, a Morphological Analyzer for Estonian. In H. Õim (ed.) *Estonian in the Changing World*. Tartu, pp. 43-98.

Kaalep, H.-J. (1998). Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. *Keel ja Kirjandus*, 1/1998, pp. 22-29.

Kaalep, H-J., Muischnek, K. (2002). *Eesti kirjakeele sagedussõnastik*. Tartu: TÜ kirjastus.

Kerge, K. (2002). *Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX*

sajandil. TPÜ eesti keele osakonna veebitoimetised, Lingvistika 1. http://digar.nlib.ee/digar/contentpdf?key=637f6b16470041ae9d0f91a60fde1410&group=2 Accessed 27 August 2013.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (eds), Proceedings of the XIII EURALEX International Congress, Barcelona: Universitat Pompeu Fabra, pp. 425-433.

Kilgarriff, A., Smith, S. & Avinesh, P.V.S. (2010). Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In Proceedings of ICON-2010: 8th International Conference on Natural Language Processing.

Kintsch, W. & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson (ed.) *Perspectives on Memory Research*. Hillsdale NJ: Lawrence Erlbaum, pp. 329-365.

Lee, J. & Seneff, S. (2007). Automatic Generation of Cloze Items for Prepositions. In *Interspeech 2007*, vol. 3, pp. 2173-2176.

Liu, C.L., Wang, C.H., Gao, Z.M., & Huang, S.M. 2005. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items, In Proceedings of the Second Workshop on Building Educational Applications Using NLP, pp. 1-8, Ann Arbor, Michigan, 2005.

Meyer, B.J.F. (1982). Reading research and the composition teacher: The importance of plans. *College composition and communication*, 33(1), pp. 37-49.

Mikk, J. (1980). *Teksti mõistmine*, Tallinn: Valgus.

Mikk, J. (1991). Studies on teaching material readability. In *Papers on education II: Problems of textbook effectivity*, Tartu, pp. 34-50.

Mitkov, R. & Ha, L.A. (2003). Computer-Aided Generation of Multiple-Choice Tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, Edmonton, Canada, May, pp. 17-22.

Redish, J.C. & Selzer, J. (1985). The place of readability formulas in technical communication. *Technical communication*, 32(4), pp. 46-52.

Richaudeau, F. (1979). Une nouvelle formule de lisibilité. *Communication et Langages*, 44, pp. 5-26.

Szmrecsányi, Benedikt M. 2004. On Operationalizing Syntactic Complexity. In : *JADT 2004 : 7es Journées internationales d'Analyse statistique des données textuelles*, pp. 1031-1038.

Tahmm (1998) = Morfoloogiline ühestaja (beetaversioon). http://www.eki.ee/keeletehnoloogia/projektid/tahmm/tahmm.html. Accessed 10 April 2013.