LingBaW *Linguistics Beyond And Within 9 (2023), 23-35*

Automatically generated language learning exercises for Finno-Ugric languages

Zsanett Ferenczi
Pázmány Péter Catholic University, Budapest, Hungary

Abstract
Morphologically rich languages always constitute a great challenge for language learners. The learner must be able to understand the information encoded in different word forms of the same root and to generate the correct word form to express certain syntactic functions and grammatical relations by conjugating a verb or declining a noun, an adjective or a pronoun. One way to improve one's language skills is through exercises that focus on certain aspects of grammar. In this paper, a language learning application is presented that is intended to help learners of Finnish and Hungarian (with Hungarian and Finnish L1, respectively) acquire new vocabulary items, as well as practice some grammar aspects that according to surveys are considered difficult by learners of these languages with the other Finno-Ugric language being the learner's native tongue, while alleviating the need to create these exercises manually. This application is a result of an on-going research project. In this research project, bilingual translation pairs and additional monolingual data were collected that can be utilized to build language learning exercises and an online bilingual dictionary with the help of automatic methods. Several linguistic patterns and rules were defined in order to automatically select example sentences that focus on a given part of the target language. These sentences were automatically annotated with the help of language processing tools. Due to the large size of the previously collected data sets, to date, only a subset of the analyzed sentences and the bilingual translation pairs has been manually evaluated. The results of this evaluation are discussed in this paper in order to estimate the precision of the methodology presented here. To ensure the precision of the information and the reliability of the application, only manually validated data sets are displayed. In this project, continuous data validation is planned, since it leads to more and more examples and vocabulary items that learners can benefit from.

Keywords: natural language processing, computer-assisted language learning, virtual flashcards, Finno-Ugric languages

1. Introduction

Finno-Ugric languages constitute a subfamily of the Uralic language family, Hungarian and Finnish being the ones with the highest number of speakers. Most Finno-Ugric languages, particularly Finnish and Hungarian, have rich morphology, they have extensive case systems with numerous word forms belonging to a certain verbal or prenominal root. Learning the correct usage of each grammatical case and the correct inflection or declension in the case of each root is tedious and time consuming, even for speakers of the other Finno-Ugric language.

Finnish and Hungarian possess different linguistic and grammatical characteristics that make language learning even more difficult for non-native speakers (Finnish as a foreign language (FFL) learners and Hungarian as a foreign language (HFL) learners). In the case of Finnish, for example, such characteristics include consonant gradation and the three cases that the grammatical object of a sentence can appear in. To illustrate this latter, consider the sentences in Table 1. As shown in the table, the case of the underlined objects differ.

Table 1: *Different cases that a grammatical object can take in a Finnish sentence.*

Finnish sentence	English translation	Case of the object
Minun täytyy ostaa uusi televisio.	I have to buy a new television.	nominative
Me ostamme kaksi kirjaa.	We are buying two books.	partitive
Mies maalasi talon.	The man painted the house.	accusative

According to Karlsson and Chesterman's survey (2008), the difference of Finnish vocabulary from that of any Indo-European language is one of the many surprises that FFL learners have to face when trying to master Finnish. Although Hungarian and Finnish belong to the same language family, the difference in their vocabularies also constitutes great challenges for learners. The possible cases that a Finnish object can appear in also causes some confusion. Since the rules, that define which case is used in a certain situation, are quite complex, it takes some time and practice to really understand their correct usage. Differentiating between the three past tenses (imperfekti, perfekti, pluskvamperfekti) also requires attention from speakers of Hungarian when learning Finnish, since there is only one past tense that is used in standard Hungarian. Understanding how the passive construction is formed and when it is used is also difficult, even more so, because it is often used in Finnish, but is not present in Hungarian, and it diverges from the "prototypical" passives that learners with Hungarian L1 might have encountered when learning other languages like English or German.

Máté (1999) conducted a survey and observed that learners of Hungarian often experience difficulties, when they learn about the definite and indefinite verb conjugation in Hungarian. Some of the learners also struggle while trying to learn the proper usage and meaning of verbal prefixes and when trying to understand the possessive construction.

To help learners of these languages practice these particular aspects of Finnish and Hungarian, a computer-assisted language learning (CALL) tool is presented. This tool consists of two modules: a virtual flashcard module, that helps learners acquire new vocabulary items, and a cloze exercise module, which facilitates grammar learning and focuses on the challenges observed by Máté (1999) and Karlsson and Chesterman (2008). Examples of the different grammar exercise types are selected automatically from previously obtained Finnish and Hungarian data sets (for more details, see Ferenczi (2021a)); it consists of bilingual translation pairs, synonyms, example sentences and definitions. This data set had been extracted automatically from different sources, such as Wiktionary, WordNet and OPUS with the help of one already existing and several newly created tools during this research project.

This paper is structured as follows. The Extracted Data section briefly presents the automatic methods applied to obtain data and the results of data collection. The Language Learning Application section then discusses how virtual flashcards and grammar exercises were

generated. To determine the precision and ensure the high quality of such a language learning tool, a subset of the generated examples has been manually checked by speakers of these languages. The preliminary results of this evaluation are presented in the Intermediate Results section. The paper concludes with a discussion of the obtained results, future work and briefly discusses the possibilities for further development.

2. Extracted data

As mentioned earlier, during one of the previous steps of the research project, automatic methods were applied in order to provide data for an online bilingual dictionary and language learning application. The bilingual word pairs and definitions – besides providing the contents of the dictionary – can be used to create a virtual flashcard module, while the obtained example sentences can be utilized when automatically generating cloze exercises for Finnish and Hungarian. In order to automatically collect this kind of data, several resources were used. In this section, these resources, as well as the methods that have been applied to extract data for the languages in question are briefly introduced.

2.1. Wiktionary

One of the resources used for data collection was an online dictionary project called Wiktionary, that has several language editions. The Wiktionary edition defines the target language of the dictionary, which means that the translations, explanations and definitions are given in this language. The source language, on the other hand, can be any language, and the headword of a Wiktionary article can be in several (source) languages, when it is part of the vocabulary of several languages (as can be seen in Figure 1, where the headword is *kuka*, and it is defined in both Finnish and Hungarian).

The Finnish Wiktionary edition contains 416,295 articles (which equals to the number of headwords this Wiktionary incorporates), while the Hungarian one contains 369,292. A Wiktionary article has different sections, that are either obligatory or optional to include when editing a page. To edit a page, one must accustom himself with the markup language used in the Wiktionary articles, which can differ in different language editions. In order to parse a certain Wiktionary edition, one can download a Wiktionary dump file, which has an XML structured format. Each page is tagged with the <page> node and the headword is given within the <title> tags. The content of the page appears in the <text> node, with the same markup as can be seen in Figure 1. The first section that is marked with a double equals sign (==) is the language of the headword (==*Finnish*== and ==*Hungarian*==). After this, three equals signs denote the part-of-speech of the word, such as ===*Noun*===. The translation (or definition if the language of the headword is the same as the language of the Wiktionary) appears after a # sign and a space. Another important section is the translation table that only appears when the language of the headword is equal to the language of the Wiktionary edition. This table contains equivalents of the headword in other languages, and the information is separated into different tables in case the headword has many senses (see Figure 2).

```
==Finnish==
[...]
===Pronoun===
{{fi-pron}}

# {{lb|fi|interrogative}} [[who]]
#: {{ux|fi|'''Kuka''' on ovella?|'''Who''' is at the door?}}
# {{lb|fi|relative}} [[who]] {{gloss|as an 'independent' relative pronoun; see the usage notes}}
#: {{ux|fi|En tiedä '''kuka''' sen teki.|I don't know '''who''' did it.}}
# {{lb|fi|relative|dialectal}} [[who]]
#: {{syn|fi|joka|mikä}}
[...]

==Hungarian==
[...]
====Adjective====
{{head|hu|adjective}}

# {{senseid|hu|dumb}} [[dumb]] (as a fish), [[tongue-tied]] (not saying a word)
[...]
====Noun====
{{hu-noun|pl=kukák}}

# {{senseid|hu|garbage can}} [[garbage can]], [[trash can]], [[refuse]] [[bin]] {{gloss|especially an outdoor container}}
#: {{cot|hu|szemetes|szemetesvödör|szemétkosár|szemetesláda|szemétláda}}
[...]
```

Figure 1: *Structure of the Wiktionary article `kuka` in the English Wiktionary database (Note: parts of the article have been eliminated for illustration purposes.)*

Figure 2: *Translation tables in the English Wiktionary article umbrella on the public interface*

Wikt2dict (Ács et al., 2013) is an already existing tool that extracts bilingual word pairs from Wiktionary. This tool has two functions: the first (*extract*) can collect data using the translation tables present in some of the articles, where there is any target language equivalent in the

translation tables. Another function of this tool takes advantage of the transitive property of translation. If a word in language 1 translates into a word in language 2, and the same word in language 2 translates into a word in language 3, then hypothetically, the translation relation between words in language 1 and 3 also holds. This assumption, however, must be treated with caution: homonymous and polysemous words can lead to wrong translation candidates. The language through which the connection is established is called pivot language (here language 2 is pivot). Based on this hypothesis, the *triangulate* method extracts data when it is given three languages. It looks for translations present in the translation tables where both languages 1 and 3 have equivalents in a pivot language Wiktionary article. Since English is the Wiktionary edition that contains the most number of articles, it was used as pivot to create a connection between Finnish and Hungarian. Using the *extract* method 12,731 bilingual word pairs were collected, while the *triangulate* method resulted in 294,757 Finnish–Hungarian translation pairs in total.

WiktionaryParser (Ferenczi, 2021b) is an algorithm that was created in this project to collect information present in the main body of a Wiktionary article. While translation tables are only created when the headword is a meaningful word in the language of the Wiktionary edition (i.e. in case of English headwords in the English Wiktionary) and the *wikt2dict* method only parses this section of the articles, this new algorithm parses the Finnish and Hungarian Wiktionary dumps and extracts Hungarian and Finnish headwords present in them, respectively. Since Wiktionary contains example sentences and other valuable information, too, the script extracts these from the articles and stores them. This method resulted in 9,544 Finnish–Hungarian translation pairs, 29,221 Finnish and 1,157 Hungarian example sentences, as well as 111,555 Finnish and 30,423 Hungarian definitions. The algorithm is freely available.

2.2. WordNet

WordNet is an ontology which was first created for English (Miller, 1995). It has been translated into several languages since then, for example, into Finnish by Lindén and Carlson (2010), and into Hungarian by Miháltz et al. (2008). The basic unit of this resource is a synonym set (synset) which contains lexical items belonging to the same concept. Therefore, a synset can contain one or more words or expressions. 56% of all Finnish synsets contain more than one word, and 26% of all Hungarian synsets consist of more than one word. All synsets have a synset offset (an eight digit long unique identifier) in these databases, which can be used to link the concepts from two different language editions. Based on this information, an algorithm was created during the research project (called *WordNet Connector* (Ferenczi, 2021c)) to link Finnish and Hungarian synsets. It first links the different synsets using the identifier and then extracts bilingual translation pairs by combining each element of these two synonym sets with each other. The algorithm also extracts two separate lists of synonym pairs for Finnish and Hungarian, and a list of Hungarian example sentences, since the Hungarian WordNet contains this type of data, as well, unlike the Finnish WordNet. With the help of this script, 25,419 synsets were connected, producing 98,883 Finnish–Hungarian translation pairs. 54,535 Finnish and 28,197 Hungarian synonym pairs were extracted, as well as 36,484 example sentences were collected from the Hungarian edition. *WordNet Connector* is a freely available tool.

2.3. OPUS

OPUS (Tiedemann and Nygaard, 2004) is a collection of automatically extracted bilingual data, including subtitles, documents of the European Commission, software localization and many more. This data set contains sentence alignments, as well as word alignments. Bilingual translation pairs can be extracted with the help of these Finnish–Hungarian word alignments. *OPUS Extractor* was created (Ferenczi, 2021d) to collect these pairs and clean the data with the help of a pre-defined pattern. This is done by a regular expression which filters out word pairs where at least one of the words contain characters that do not belong to the Finnish or Hungarian alphabets. This algorithm is freely available. This method resulted in more than a million translation pairs, but as expected, different forms of the same root in one language are marked as equivalents of many forms of the same root in the other language. To reduce the number of incorrect word pairs, each word was lemmatized and different occurrences of the same word pair were united. This caused an approximately 60% decrease in the number of word pairs, producing 391,136 translation pairs in total.

3. Database

During the research project, a language independent MySQL database had been developed. The collected data are saved in this database which provides the basis for the online dictionary and language learning application. To store data in a non-redundant way, one of the basic concepts of the database is that each and every linguistic data is treated as an *entity*. The Entity table stores information about lemmas, multi-word expressions (MWE) and even sentences, such as their type (lemma, MWE or sentence), their identifier, language, and in some cases the part-of-speech. Entities can have certain relations with other entities. Such a relationship is described by the identifiers of the two entities participating in it and the type of the relationship, for example, two entities can be the translations of each other, or a lemma can be connected to an example sentence that demonstrates the way the lemma is used. These relations are stored in the Relation table. The language learning application is based on the data stored in this database, although only those word pairs and sentences can be displayed in the application which have been manually checked to ensure that learners are presented with correct language data. The database is developed in a way that the learners' answers are also saved in a separate table. This way it is possible to collect useful information about the challenges and difficulties faced by learners of Finnish and learners of Hungarian, which can be used in the future by researchers who want to validate their hypotheses, without having to gather data from learners.

4. Language learning application

4.1. Virtual flashcards

A popular way to acquire new vocabulary items or memorize difficult ones is to use so-called flashcards. On one side of these cards, there is a new, so far not known item (e.g. a word in the

target language that the learner does not know yet). On the other side, there is either the translation in the source language or a target language definition of that item. Therefore, depending on the language of the explanation, the flashcards can be bilingual or monolingual. These cards can be created on paper that the learners can take in their hands and turn over to learn the target word or validate their answers, or it can be done in a digital format. There have been many studies that found that CALL has a positive effect on the learning of vocabulary items (Basoglu and Akdemir, 2010, Kilickaya and Krajka 2010). Elgort (2013) used Vocabulary Size Test which is a multiple-choice test and showed that intermediate proficiency learners of English perform significantly better when the vocabulary items are presented with their translation equivalents in the L1 (Russian). However, this observation is not so significant in the case of more advanced learners. Jo (2018) conducted an extensive experiment regarding vocabulary and also noticed that learners achieved higher scores on posttests when the L1 definitions were used instead of L2 definitions. Another useful feature that might be added to flashcards is pronunciation. Hungarian and Finnish, however, have very similar rules of pronunciation (Weöres, n.d.), e. g. the main stress is always on the first syllable. Another phonetical similarity between the two languages is vowel harmony. According to Korhonen (2012), among the surveyed Hungarians who learn Finnish, the majority agreed that pronunciation is not a difficult part of Finnish language learning. Since this application is intended for Hungarians learning Finnish and Finns learning Hungarian, pronunciation is not included in the flashcards in this phase of the project.

In the web application presented in this paper, learners are given the options to use the L1 equivalent or L2 definition of the new vocabulary items they are about to learn. After choosing an option, and selecting the target language the learners want to practice, the flashcard module is divided into two phases: in the first phase (practice phase), the learners need to get acquainted with the new items, they can turn over the cards as many times as needed, while in the second phase (test phase), the learners' active–productive knowledge (Laufer et al., 2004) is tested. They need to recall the newly acquired items by typing in the expression they just learnt, when a given L1 equivalent or L2 definition appears on the screen.

In Figure 3, the two sides of a monolingual Finnish virtual flashcard are shown: the target word (*lapsi* = *child*) is given on one side of the card, and its L2 definition is given on its other side. Note, that while on the figure, the sides of the flashcard are shown side by side, on the web interface, the learners can only see one side at a time, and turn it over (make its other side appear) by clicking on it.

Figure 3. *Example for a Finnish flashcard*

In the test phase, when the learners submit their answers, feedback is given immediately by the system. It is important to note that the application only accepts the original word or expression as correct answer. Synonyms are not accepted, although they might be correct translations of a

given L1 expression. Nevertheless, this limitation can be addressed in the future phases of the project.

4.2. Grammar exercises

CALL methods can have a positive impact on grammar learning when compared to non-CALL methods (Aslani and Tabrizi, 2015). One type of task that can be implemented with computers is cloze exercises (otherwise known as fill-in-the-blank exercises) which aims to help with grammar learning. The task is to reconstruct the missing parts of a sentence or paragraph of an L2 text. In some cases, the lemma of the missing word is given and the only task is to fill in the blank with the correct form, in other cases, it is completely up to the learners to guess what lexical item fits in the context. It is also possible to evaluate the performance of the learners automatically after the answers are submitted.

In this CALL application, several cloze exercises are included for different aspects of Finnish and Hungarian grammar. For Finnish, the following three grammar issues were considered: the different cases of the object, the three past tenses, and the passive verb conjugation. For Hungarian, exercises were built to help practice the definite and indefinite verb conjugation, the usage of verbal prefixes, and the possessive construction. In this section, the linguistic patterns, that were used to select an appropriate subset of sentences from the database, and the structure of the exercises are described.

To automatically build cloze exercises, some kind of target language text or corpus is needed. As mentioned before, numerous Finnish and Hungarian example sentences were collected from different sources. They were extracted using the tools presented above in the previous phase of the project. These sentences can be utilized in an application where one of the words is hidden and the learners are asked to reconstruct the original sentence. After observing the data, it was noticed that the sentences vary in length and quality. Since only grammatical, complete sentences should be used in these tasks, a condition is applied to filter out unwanted data. Only those example sentences are considered complete which contain at least 3 words, which start with a capital letter and end with a punctuation (full stop, exclamation mark, question mark, etc.), and which do not contain special characters, such as <, >, = or $.

These conditions decrease the number of sentences that can be used in the exercises, in the case of Finnish to 18,043, and in the case of Hungarian to 17,450. Manually selecting which sentences can be utilized in a certain grammar exercise is time-consuming. However, it is possible to automatize this process with the use of SQL queries that look for certain linguistic patterns in the database, and select a subset of sentences to include them in specific tasks. For this, the sentences need to be tokenized and lemmatized, as well as morphologically analyzed and dependency parsed. To achieve this, three tools were used: the Hungarian *emtsv* pipeline (Indig et al., 2019), *omorfi* (Pirinen, 2015) for tokenization, lemmatization and morphologic analysis of Finnish sentences, while the dependency parsing was conducted with the help of *uralicNLP* (Hämäläinen, 2019). The output data from the analyzers are stored in the same database where the linguistic data can be found, and this database is queried when the data need to be loaded for a certain exercise type.

Finnish exercises. In the CALL application, three types of cloze exercises were implemented that help language learners practice different aspects of Finnish grammar.

Finnish objects can appear in 4 cases: nominative, partitive, accusative and genitive. To create a fill-in-the-blank exercise to practice which one of the four cases to use in a particular sentence, first a subset of the data needs to be queried from the database, and then, the object needs to be removed. Sentences containing a noun, adjective, pronoun or numeral with the *DOBJ* dependency tag are selected, when one of the four possible cases (*Case=Nom*, *Case=Par*, *Case=Acc*, *Case=Gen*) can be found among morphological codes of the object. This word is replaced by an input field, and the learners are asked to put the given lemma in the correct case. The lemma of the missing word is given in parentheses after the input field, as can be seen in Figure 4.

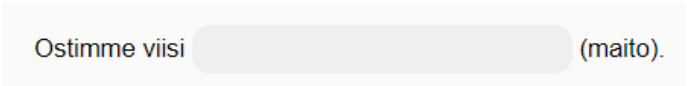

Figure 4. *Example task to reconstruct the missing word form (the correct case for a Finnish object)*

Another grammar issue that FFL learners face is the existence of three past tenses in Finnish. These are: simple past (imperfekti), present perfect (perfekti) and past perfect (pluskvamperfekti). The two latter is composed of the auxiliary verb *olla* (in either present or simple past) and the past participle of the second verb. Knowing which past tense to use in a certain environment requires a lot of practice. For this, an exercise was created where sentences containing any of the three past tenses appear, while the verb gets replaced by a text input field. The learners need to decide which is the correct past tense and conjugate the given verb into the correct form. One of the following two conditions must be met by a Finnish sentence to be part of this exercise: it either needs to contain a verb in simple past form (*Tense=Past*), or its main verb has to be *olla* (in either present or simple past tense) and there must be an active past participle form (marked by the *Connegative=Yes* and *Tense=Past* morphological codes) in the same sentence. These words are replaced by text fields on the interface, so that learners can reconstruct the correct verb tense. It has been observed that in some cases, two text fields appear, because the main verb *olla* and the past participle can separate from each other in the compound tenses. Therefore, another rule is necessary to eliminate these sentences from the query results, since one out of three past tenses (i.e. the simple past) can immediately be excluded, and the aim of the task is to let the learners decide which past tense is the most suitable in the given context. It must be ensured that the two verbs are adjacent in the compound tenses, so that each sentence contains only one text field, not providing any implications about which tense may be correct.

Passive voice is used quite often in Finnish. One of the reasons is that the first person plural form of the present tense indicative is replaced by the passive form in colloquial Finnish. Correctly conjugating the verbs in this form is therefore essential for language learners. To build cloze exercises that help FFL learners practice this construction, sentences containing a verb in passive form are selected from the database. This appears as *Voice=Pass* among the morphological codes of verbs. The passive verb is then replaced by a text input field and the first infinitive form of the verb is given in parentheses.

Hungarian exercises. In the case of Hungarian, three grammar issues were processed and observed in order to build exercises, which can help HFL learners master these topics. Formulating linguistic patterns for these led to the automatic extraction of thousands of example sentences that can be used in these exercises.

In Hungarian, transitive verbs have two paradigms: a definite and an indefinite conjugation. This increases the number of endings in the case of some verbs, and further complicates the process to choose the correct inflection in the given context. To practice the difference between these two conjugations, example sentences which contain a transitive verb are queried from the database. A transitive verb appears in at least one sentence with the *Definite=Def* pattern. After listing the transitive verbs with the help of this rule, sentences are selected that contain any of these verbs. There is no restriction regarding the paradigm that the verbs appear in in these instances; they can have a definite (*Definite=Def*) or indefinite (*Definite=Ind*) conjugation, since the task of the learner is to decide which one of the two paradigms is the correct one. After the extraction of sentences, the verb is replaced with a text input field, and the lemma of the verb is given in parentheses. Since Hungarian is a pro-drop language, the pronoun of the subject is also given italicized within the parentheses after the lemma of the verb, see Figure 5.

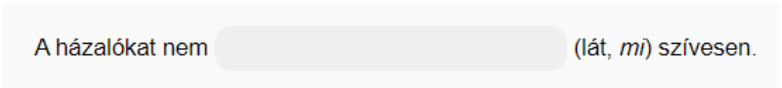

Figure 5. *Example for the Hungarian definite and indefinite conjugation task*

The correct use of verbal prefixes (or preverbs) can also be challenging. Preverbs in Hungarian can appear both preverbally and postverbally, depending on the structure and word order of the sentence. It is a debated topic exactly which words belong to this category (Kalivoda, 2021), but in this application 13 words are defined as preverbs: *be, ki, le, fel, meg, el, át, bele, ide, oda, szét, össze, vissza*. This list can be augmented or reduced as necessary in the future. The condition, that defines which Hungarian sentences are suitable for this exercise, is the following: the sentence must contain any of these 13 words either as an isolated word, or attached to the beginning of a verb. The manual validation of 300 sentences that were obtained with this initial condition led to the observation that the beginning of some verbs were incorrectly marked as preverbs, although they were part of the root. The exclusion of such sentences was implemented by adding the list of these verbs to the condition. Examples for such verbs include e.g. *kiabál* (shout), *felel* (respond), and *megy* (go).

Hungarian expresses possession by adding the possessive suffix to the possessed object. This suffix depends on the person of the possessor, vowel harmony, whether the possessed object is singular or plural, and whether the possessed noun ends with a vowel or a consonant. This explains the number of allomorphs the possessive suffix has. For instance, in case of a third person singular possessor the suffix can be *-a, -e, -ja,* or *-je*. Furthermore, the suffix on the possessed object may cause the ending of the root to change, such as in this example: *kutya* (dog), *kutyá-ja* (his/her dog), where the final vowel (initially *-a*) becomes *-á*. The possessive suffix is marked by the morphological analyzer with the *Number[psor]* and *Person[psor]* features on the possessed noun, adjective or numeral. If a sentence contains a word that has these

features among its morphological codes, it can serve as an example sentence in this task. After retrieving the compatible sentences from the database, the words which express the possessed object are replaced with text input fields and the lemma of these words are given. The person of the possessor is also given within parentheses, because the pronominal possessors are usually pro-dropped. See Figure 6.

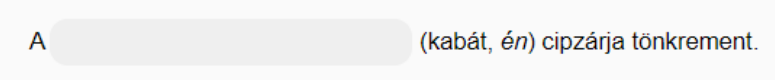

Figure 6. *Example task for the Hungarian possessive constructions*

5. Intermediate results

Monolingual flashcards introducing the target words with their definition were created for both Finnish and Hungarian. The number of monolingual Finnish flashcards is 111,593, while Hungarian data resulted in 70,198 flashcards. Bilingual flashcards, which present target words with their translation in the other language, were also developed in the CALL application. More than 800,000 such flashcards can be generated from the collected data set. Since automatic data extraction does not always lead to perfectly accurate data, the information that is presented to the language learners must be checked and validated. The flashcards therefore will be manually evaluated, before making the application freely available for the public. This evaluation has already started, and 273 bilingual flashcards were approved for use by validators, while 114 translation pairs were marked as incorrect. Hence, automatic data extraction methods presented in this paper have 70,54% precision on average. The validation of monolingual flashcards has not yet started.

Using the data set that was automatically extracted with different methods, and the patterns that were defined in order to generate cloze exercises automatically, the retrieval of thousands of example sentences was possible for each exercise type. The exact number of sentences can be seen in Table 2. These sentences and the generated exercises will also be manually evaluated, before learners can practice the different grammar aspects with their help.

To date, a subset of sentences and their analysis has been evaluated. The results are presented in Table 3. 50 sentences had been randomly selected for Finnish and 50 for Hungarian, and the output of the applied language processing tools (*emtsv*, *omorfi* and *uralicNLP*) were analyzed. The most erroneous output was given by the lemmatizer in the *omorfi* tool. More than half of the evaluated sentences has been analyzed incorrectly by that submodule. The Hungarian natural language processing tool (*emtsv*) gives better overall performance than the Finnish one (68% vs. 30% precision). The least precision was reached by the dependency parser module (*emDep*) of *emtsv*, that produced incorrect output for 14% of the evaluated sentences.

It is of high importance that language learners only encounter correct linguistic data and feedback in this application, since erroneous input may negatively affect the learning process. This is why the results of manual evaluation are stored in the database, and only accurate sentences will be found in the exercises.

Table 2: *Number of sentences obtained for each exercise type.*

Exercise type	Number of sentences matched
Finnish objects	7,088
Finnish past tenses	5,133
Finnish passive construction	2,092
Hungarian definitive and indefinitive conjugation	5,830
Hungarian verbal prefixes	5,227
Hungarian possessive construction	4,896

Table 3: *Details of manual validation.*

	Finnish	Hungarian
Not well-formed sentences	2 (4%)	1 (2%)
Erroneous lemmatization	28 (56%)	5 (10%)
Erroneous morphological features	3 (6%)	3 (6%)
Erroneous dependency analysis	2 (4%)	7 (14%)
Precision	15 (30%)	34 (68%)
Total number of validated sentences	50 (100%)	50 (100%)

6. Conclusion

In this paper, several language learning exercises were presented for Finnish and Hungarian. The automatically extracted bilingual translation pairs showed a 70,54% precision based on the manual evaluation of 387 data points, while the natural language processing tools (*emtsv*, *omorfi*, *uralicNLP*) proved to require further work to improve their precision when isolated sentences are provided as their input.

The collected sentences and translation pairs will only appear in the CALL application once their precision is ensured by manual evaluators. Data validation is ongoing in this project with the help of speakers of both Finnish and Hungarian.

One of the shortcomings of the flashcard module is the limited feedback it can produce to date, since only the original target words are considered to be correct, although one definition or source word might have more than one corresponding words in the target language.

Another possible future improvement of this application may be the usage of learners' data: to collect data anonymously from language learners that can enable researchers to further investigate the difficulties that learners face when learning Finno-Ugric languages and empirically support or counter a hypothesis or theory about these morphologically rich languages.

References

Ács, J., Pajkossy, K., and Kornai, A. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the sixth workshop on building and using comparable corpora*, 52–58.

Aslani, M., and Tabrizi, H. H. 2015. Teaching grammar to Iranian EFL learners through blended learning using multimedia softwares. Journal of Applied Linguistics and Language Research 2(8): 76–87.

Basoglu, E. B., and Akdemir, O. 2010. A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flash cards. Turkish Online Journal of Educational Technology-TOJET 9(3): 1–7.

Elgort, I. 2013. Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing* 30(2): 253–272.

Ferenczi, Zs. 2021a. Finn–magyar fordítási párok kinyerése automatikus módszerekkel. In Gráczi Tekla Etelka és Ludányi Zsófia (Eds.), *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből,* 131–150. Nyelvtudományi Kutatóközpont, Budapest.

Ferenczi, Zs. 2021b. *Wiktionary Parser*. https://github.com/ferenczizsani/wiktionary_parser

Ferenczi, Zs. 2021c. *WordNet Connector*. https://github.com/ferenczizsani/connect_wordnets

Ferenczi, Zs. 2021d. *OPUS Extractor*. https://github.com/ferenczizsani/opus_extractor

Hämäläinen, M. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software* 4(37).

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., and Makrai, M. 2019. One format to rule them all–The emtsv pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop,* 155–165. Association for Computational Linguistics, Florence.

Jo, G. 2018. English Vocabulary Learning with Wordlists vs. Flashcards; L1 Definitions vs. L2 Definitions; Abstract Words vs. Concrete Words. *Culminating Projects in English*. 132.

Kalivoda, Á. 2021. *Igekötős szerkezetek a magyarban*. [Preverb Constructions in Hungarian.] Ph.D. thesis, Pázmány Péter Catholic University, Budapest.

Karlsson, F., and Chesterman, A. 2008. *Finnish: an essential grammar.* Routledge.

Kilickaya, F., and Krajka, J. 2010. Comparative usefulness of online and traditional vocabulary learning. Turkish Online Journal of Educational Technology-TOJET 9(2): 55–63.

Korhonen, S. 2012. *Oppijoiden suomi. Koulutettujen aikuisten käsitykset ja kompetenssit* [Perceptions and competences of adult learners of Finnish]. Helsinki: Helsingin yliopisto.

Laufer, B., Elder, C., Hill, K., and Congdon, P. 2004. Size and strength: Do we need both to measure vocabulary knowledge? *Language testing* 21(2): 202–226.

Lindén, K., and Carlson, L. 2010. FinnWordNet – Finnish WordNet by translation. *LexicoNordica–Nordic Journal of Lexicography* 17: 119–140.

Máté, J. 1999. A magyar nyelv elsajátításának nehézségei a finn anyanyelvű tanulók szempontjából [Difficulties to learn Hungarian for Finnish learners]. *Hungarologische Beiträge* 12: 91–112.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., and Váradi, T. 2008. Methods and results of the Hungarian WordNet project. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (Eds.), *Proceedings of The Fourth Global WordNet Conference,* 311–321. University of Szeged.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11): 39-41.

Pirinen, T. A. 2015. Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. *SKY Journal of Linguistics* 28: 381-393.

Tiedemann, J., and Nygaard, L. 2004. The OPUS Corpus – Parallel and Free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal. European Language Resources Association (ELRA).

Weöres, Gy. n.d.. *The Relationship between the Finnish and the Hungarian Languages.* https://histdoc.net/sounds/hungary.html