Apples – Journal of Applied Language Studies
Vol. 8, 3, 2014, 39-54

Using Automatic Morphological Tools to Process Data from a Learner Corpus of Hungarian

Péter Durst, University of Szeged; Martina Katalin Szabó, University of Szeged; Veronika Vincze, MTA-SZTE Research Group on Artificial Intelligence & University of Szeged; & János Zsibrita, University of Szeged

The aim of this article is to show how automatic morphological tools originally used to analyze native speaker data can be applied to process data from a learner corpus of Hungarian. We collected written data from 35 students majoring in Hungarian studies at the University of Zagreb, Croatia. The data were analyzed by magyarlanc, a sentence splitter, morphological analyzer, POS-tagger and dependency parser, which found 667 unknown word forms. We investigated the recommendations made by the Hungarian spellchecker hunspell for these unknown words and the correct forms were manually chosen. It was found that if the first suggestion made by hunspell was automatically accepted, an accuracy score of 82% could be attained. We also introduce our automatic error tagger, which makes use of our annotation scheme developed on the basis of the special characteristics of Hungarian morphology and learner language, and which is able to reliably locate and label morphological errors.

Keywords: Hungarian language, learner corpus, natural language processing, morphological parsing, automatic error tagging

1 Introduction

Currently, the widely accepted idea of *interlanguage* (Selinker 1972) suggests that errors constitute a natural phenomenon in the process of foreign language learning and use. A systematic analysis of errors in a learner corpus may shed light on certain features of the learner language, which may be interpreted by linguists. Possible new findings may be conceived on the one hand as novelties but on the other hand they may simply contribute to already known facts because, as Granger (2002: 4) noted, corpora generally complement rather than replace earlier data sources.

As explained below, automatic tools can be used to find and group errors in a large amount of text, which permits the statistical analysis of learner language data. Although experience gained with some target languages is definitely a good basis for developing a method of analysis in another language, the

Corresponding author's email: durst.peter@gmail.com
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies, University of Jyväskylä

http://apples.jyu.fi

identification and the classification of errors has to be language specific. Morphologically rich languages have a number of elements (morphemes) and several rules of word formation (assimilation, vowel harmony, etc.) that are important sources of errors and they have to be analyzed one by one, hence setting up an error tagging scheme designed specifically for the given language is essential for the task.

Compiling and analyzing a learner corpus of Hungarian may produce benefits for language learning and teaching: authors of coursebooks and pedagogical grammar reference books may find it useful to have information on the frequency of certain error types, for example. De Cock and Granger (2005) suggested that results from learner corpora research could probably be best used in developing learners' dictionaries, which is a relevant issue for teaching and learning Hungarian as a foreign language since up until now there is no such dictionary available for learners of Hungarian.

Apart from the direct benefits in language teaching and learning, other areas of linguistic research might also apply relevant findings obtained from the analysis of learner corpora since the experience gained with Hungarian, which is a morphologically rich language, may shed light on new aspects of more general issues, which go far beyond the limitations of one particular language. According to Suni (2013), research carried out on the acquisition of Finno-Ugric languages may have significant implications for the study of second language acquisition in general. However, investigations based on learner corpora of Hungarian have just started. Although there is a wide range of questions which call for the specific investigation of Hungarian learner corpora, such as the peculiarities of its verbal system and the complex possessive structures, to date there has only been one paper published in this area (Dickinson & Ledbetter 2012), which outlines a general scheme for error tagging in a Hungarian learner corpus. Apart from the obvious need for more research to be done, it may also be of interest that recently a natural language processing (NLP) toolkit called *magyarlanc* (consisting of a sentence splitter, a morphological analyzer, a POS-tagger and a dependency parser) has been developed to analyse L1 Hungarian data (Zsibrita et al. 2013), which may be applied in this new area as well.

This paper presents the first steps of a rather complex project to compile and analyse Hungarian learner corpus data. As for the whole ongoing project, we have set ourselves the following goals: 1) to start building a learner corpus of Hungarian and determine the most effective and feasible methods of collecting data; 2) to develop the methods of text processing using the already existing automatic NLP tools with just some minor modifications; 3) to set up an error tagging scheme which meets the special features of Hungarian and is capable of producing data ready to be interpreted and used in other areas (e.g. second language acquisition (SLA) research and developing teaching materials). A big restricting factor in our study was the relatively low budget, which ruled out the possibility of reliance on costly methods like manual annotation, the transcription of oral data and processing of handwritten texts.

In this article, we intend to show how some already existing NLP tools can be used for analyzing a learner corpus of Hungarian and to lay down the principles of an error tagging scheme. As for the specific analysis of our learner corpus, we decided to concentrate only on nominal errors because it was easier to determine their scope than that of verbal morphology. A special feature of Hungarian verb conjugation is that the definite object is marked on the verbs and therefore examining errors of Hungarian verbal morphology is a more complex task,

which should be partly based on the experience gained from the analysis of nouns. It should be added that our aim is to find and categorize non-target word forms but we will not deal with the possible explanations of errors.

This article is structured as follows: first, in sections 2 and 3, we will describe our data and the possible morphological errors in nouns as well as the error tagging method. This is followed by reporting and discussing the results derived from the presently available corpus in Section 4. In Section 5, we round off with a summary and suggestions for future work.

2 Data collection and analysis

Data was collected from 35 students majoring in Hungarian studies at the University of Zagreb in Croatia in the form of written assignments, which included a written composition of some 1500 characters on a given theme. 18 compositions were written about the difficulties of the Hungarian language (*Difficulties*) as an assignment for a seminar in linguistics, 11 on Hungarian immigrants in England (*England*) and 6 on "a person I like" (*person I like*) as home assignments in two language courses (cf. Table 1). The length of the texts varied from 4467 to 1002 characters with an average length of 2120 characters excluding spaces. The assignments were compulsory as was the choice of topic. All the compositions were written and submitted in a Word document format using a standard Hungarian keyboard to rule out problems arising from ambiguities during the process of transcribing handwritten data and to ensure faster and cost-free data collection. Students were not allowed to use any dictionaries or on-line translation sites and the spellchecker function of MS Word had to be switched off. All learners had received at least one year of language instruction and the majority had a B1 level of proficiency according to the definition of the Common European Framework of Reference for Languages. (The proficiency levels of students were based on information got from their language instructors, and not on formal test results.)

Table 1. Data on the HunLearner corpus

	Difficulties	*England*	*person I like*	Total
Number of compositions	18	6	11	35
Number of sentences	559	134	258	951
Number of tokens	10433	1930	3936	16299

Demographic data along with additional information about students' language learning backgrounds were also recorded and will be analysed in some later study. The size of the HunLearner corpus is constantly growing but due to practical limitations, we do not intend to include any handwritten or oral data.

For the automatic analysis of the data collected, we first employed *magyarlanc* 2.0, a language processing toolkit developed for Hungarian (Zsibrita et al. 2013). It split the texts into sentences and tokens, then the tokens were morphologically tagged and sentences were syntactically parsed. For erroneous tokens which were labelled as unknown by the morphological analyzer, the spellchecker *hunspell* (Trón et al. 2005) provided possible correct versions, out of which the correct one was manually selected. As we only concentrated on nominal errors, we selected the corrected forms for nouns. Based on a comparison of the erroneous form with its corrected counterpart, a rule-based algorithm assigned

an error tag to each erroneous token, which made it possible to get some statistical data on the frequency of different error categories. A detailed classification of errors, the error coding system, statistical data and a discussion of results will be presented later on.

3 Error tagging reflecting the characteristics of Hungarian morphology

When analyzing a learner corpus, it seems appropriate to mention that the term *error* will be used here to mean all non-target forms, without any distinction being made between errors and mistakes. We are aware of the differences between an occasionally produced non-target form and a systematic deviation from the target language that reflects the structure of the learner's interlanguage, which is different in certain ways from the target language, but it is not our aim here to explain the origin of the erroneous forms. Also, we should add that the attitude to investigating learners' errors has changed over the past few decades and although this research method was neglected for some time on the grounds of it being unreliable, the possible benefits of error analysis are being reevaluated as part of the computer-aided analysis of learner language (Granger 2003). It was indeed the theoretical background behind the interpretation of erroneous forms (cf. the Contrastive Analysis Hypothesis in Gass & Selinker 2001: 72–79) that triggered a general refusal, but error analysis in its present form simply provides a quantitative summary of erroneous forms and it is really up to the reader to decide how the results should be interpreted.

3.1 Main principles

The main principles of error tagging in this study can, for example, be found in Granger (2003: 467) and we will comment on the ideas presented therein. The annotation scheme of our study seems to agree with all of Granger's (ibid.) criteria except for one. Setting up an *informative* but *manageable* system was an important aspect and we sought to include all the necessary information on Hungarian nominal morphology, bearing in mind the fact that its complexity should be kept within reasonable limits even if it is an automated process. This annotation scheme is also *flexible*, since it permits the addition or deletion of elements in its coding system. During the evaluation process, data can be listed according to each element of the error codes, thus enabling researchers to compile lists of error types after taking into account different aspects of the error tagging scheme in any combination. Since we use only automatic tools here for error tagging, *consistency* between annotators is not a relevant issue. The only criterion that is not fulfilled is *re-usability*. Due to the special morphological properties it is not possible to devise categories that are sufficiently general for use in other languages. More general categories, which would be applicable to other languages, would result in insufficient information for a proper evaluation of the data and it is quite likely that little of substance could be recommended for the SLA or foreign language teaching (FLT) areas. This is why we devised a category system for language error analysis.

The annotation scheme outlined by Dickinson and Ledbetter (2012) labels learners' errors of different linguistic categories, including phonology, morphology and syntax observed from their Hungarian learner corpus. Taking

into account the complexity of Hungarian morphology and our objective to get statistically analyzable results using the available tools, we decided to tag just the non-target forms of nouns. Although marking morphological errors on nouns poses several difficulties, it is a well-defined task while verbal morphology overlaps considerably with syntax, and hence constitutes a complex issue due to such peculiarities as the context-dependent word order and the verb-object agreement in Hungarian. A general analysis of valency frames (that is, the complement structures of verbs) is feasible with the available tools and it is an important preprocessing stage before a complex syntactic analysis and error tagging. We examined valency frames in our corpus and compared the results with L1 data, but error tagging has not yet been devised for errors other than those involving just nouns. Most likely, it is a good idea to first evaluate the experiences gained with nouns and then broaden the focus of error tagging to other linguistic levels.

The error tagging scheme used in the present study is based mainly on Granger's criteria presented above and on a doctoral dissertation written on the acquisition of Hungarian as a foreign language (Durst 2010). In addition, this scheme is corpus-driven, i. e. based on the data of our corpus, and based on the practical experiences provided by teachers of Hungarian as a foreign language (HFL).

As for the most important aspects of Hungarian noun morphology which have to appear in the end results and thus in the error tagging scheme, there are four basic factors. These are: 1) Several nouns have two stem allomorphs; 2) Suffixation is often accompanied by phonological changes of consonants including assimilation; 3) The choice among the different vowel variations of a suffix is determined by the rules of vowel harmony; 4) More than one suffix may be added to a stem at the same time but their order is not optional. The relevant morphological rules will be elaborated on in Section 3.2, where we present the error codes used here.

3.2 Typical morphological errors and their coding

Each erroneous word is given a code of four characters which indicate the evaluation of the given word form from four different aspects. The first character (A, B or C) tells us whether the appropriate stem allomorph is used and whether it is spelt correctly. The second character of the error code is a number between 1 and 5 and it covers problems with morphological assimilation. The third character is for the vowel harmony features and errors connected with the possessive *j* suffix. The fourth character of the code may be 0 to indicate that no suffix is added to the word stem, and 1 or 2 to indicate that one suffix or at least two suffixes are added to it, respectively.

3.2.1 Word stems

The first character (A, B or C) tells us whether the appropriate stem allomorph is used and whether it is spelt correctly. Hungarian nouns have either one stem or two stem allomorphs. Around 30% of all Hungarian nouns have more than one stem allomorph, but they include the most frequently used words and therefore constitute a major source of errors. In the case of words that have more stem allomorphs, one of them is used in the nominative case and with certain suffixes, while the other allomorph appears in the plural and in the accusative forms as

well as in possessive forms and with certain suffixes, so they are quite frequent. The four main types are v-stems, as given in Table 2, which insert a *v* sound in the second stem allomorph, which may be accompanied by vowel changes too; vowel shortening stems, as shown in Table 3, whose last syllable shortens if certain suffixes are added to them; vowel deleting stems, as shown in Table 4, where the vowel before the stem-final consonant is deleted in certain cases; and vowel lengthening stems, as given in Table 5, where the word-final *-a* and *-e* sounds become *-á-* and *-é-*, respectively when a suffix is added to the word (except for a few new suffixes). It should be mentioned that although the group is called "vowel lengthening stems", the difference between the phonemes represented by *a* and *á* or *e* and *é* is not only their length but their phonetic values are also different.

Possible non-target forms include simple typological errors and wrong choices of stem and therefore the following codes are used: A – correct, B – spelling error in stem, C – wrong form of stem. Where a word form with erroneous spelling coincides with an existing stem allomorph, it is marked as C – wrong form of stem.

Table 2. V-stems

Word	Gloss	English	Possible erroneous form	1st character of error code
ló	horse-NOM	horse		A
lóval	horse-INSTR/COM	with (the) horse	*loval *lovval	B C
lovak	horse-PL	horses	*lók	C
lovat	horse- ACC	horse	*lót	C
lovon	horse-SUP	on (the) horse	*lón	C

Table 3. Vowel shortening stems

Word	Gloss	English	Possible erroneous form	1st character of error code
kéz	kéz-NOM	hand		A
kezek	hand-PL	hands	*kézek	C
kezet	hand-ACC	hand	*kézet	C
kézben	hand-INES	in (the) hand	*kezben	C
kézen	hand-SUP	on (the) hand	*kezen	C

Table 4. Vowel deleting stems

Word	Gloss	English	Possible erroneous form	1st character of error code
bokor	bush-NOM	bush		A
bokrok	bush-PL	bushes	*bokorok	C
bokrot	bush-ACC	bush	*bokorot	C
bokorban	bush-INES	in (the) bush	*bokrban	C
bokron	bush-SUP	on (the) bush	*bokoron	C

Table 5. Vowel lengthening stems

Word	Gloss	English	Possible erroneous form	1st character of error code
macska	cat-NOM	cat		A
macskák	cat-PL	cats	*macskak	C
macskát	cat-ACC	cat	*macskat	C
macskában	cat-INES	in (the) cat	*macskaban	C
macskán	cat-SUP	on (the) cat	*macskan	C

3.2.2 Assimilation of suffixes

The second character of the error code is a number between 1 and 5, and it denotes problems associated with morphological assimilation. Various types of phonological assimilation can be observed in Hungarian (Ács & Siptár 1994: 568–574), but there are only two suffixes which are of interest in our present study. As for nouns, full morpho-phonological assimilation, which is reflected in spelling, occurs only with the instrumental-comitative and the translative-factive cases indicated by the appropriate variation of the *-val / -vel* and the *-vá / -vé* suffixes, respectively. This is also a common source of mistakes among learners as they may both underuse and overuse assimilation. Ignoring the need for assimilation (i.e. doubling the last consonant of the noun stem) occurs more frequently here, which produces erroneous forms like **cukorval* (*cukor-val*, sugar-INSTR/COM, meaning 'with sugar') instead of the correct *cukorral* (cf. Table 6). However, performing assimilation where it is not needed appears less frequently. Error codes include the possible combinations of necessary/unnecessary, performed/unperformed assimilations and their correctness.

Table 6. Shortening stems

Correct TL form	Gloss	English	Possible erroneous form and description	2nd character of error code
cukorban	sugar-INES	in (the) sugar	correct No assimilation (no assimilation needed)	1
cukorral	sugar-INST/COM	with (the) sugar	*cukorrel Assimilation correct (assimilation needed) Other problem in suffix	2
cukornak	sugar-DAT	for (the) sugar	*cukorrak Assimilation present but not necessary	3
cukorral	sugar-INST/COM	with (the) sugar	*cukorval Assimilation not present but necessary	4
cukorral	sugar-INST/COM	with (the) sugar	*cukornal Assimilation present but incorrect	5

Based on common experience in Hungarian language teaching, the error type **cukorval* (4) occurs by far the most frequently, while types **cukorrak* (3) and **cukornal* (5) appear only occasionally. Hence the reason why the latter two should be included here in the code system is based on theoretical grounds.

3.2.3 Vowel harmony

A more general and more frequent phonological assimilation is that of vowel harmony, which is a characteristic feature of Hungarian but it is also present in genetically non-related Asian and Turkic languages. The fundamental rules based on vowel harmony govern the system of suffixation and help language learners select the appropriate variation of a suffix. Nouns and suffix variations can be categorized according to the vowels that they contain. Hungarian suffixes usually have two or more variations which carry exactly the same meaning, but differ phonologically from each other. Word stems that contain only front vowels take the suffix variation which also contains a front vowel with some exceptions for historical linguistic reasons, while the suffix variation that contains a back vowel is added to word stems that contain only back vowels or both back vowels and front vowels. A nice example of this is the Hungarian *-ban / -ben* suffix, whose meaning corresponds to the English preposition *in*.

The phonemic orthography of Hungarian allows us to rely only on simple written forms instead of IPA symbols, so we may say that the *a, á, o, ó, u* and *ú* letters represent back vowels, while *e, é, i, í, ö, ő, ü,* and *ű* stand for front vowels, however, a subgroup of rounded front vowels (*ö, ő, ü, ű*) may also be distinguished. Words like *autó* 'car' and *ház* 'house' contain only back vowels and *mozi* 'cinema' contains both front and back vowels, so the *-ban* variation is added to them to create word forms like *autóban* (autó-ban, car-INES, 'in car'), *házban* (ház-ban, house-INES, 'in house') and *moziban* (mozi-ban, cinema-INES, 'in cinema'). The front vowel suffix variation *-ben* is added to words such as *leves* 'soup' and then in the word form *levesben* (leves-ben, soup-INES, 'in soup') the suffix phonologically fits the word stem.

In many cases where a suffix beginning with a consonant is added to a word that ends in a consonant, a vowel is added to facilitate pronunciation. Although historically this vowel belonged to the archaic stem of the word, in many cases today they are generally referred to as linking vowels. The *-t* ending, which marks the accusative case is either added to the word stem on its own or is preceded by a linking vowel which is chosen according to the phonological properties of the word. Therefore apart from choosing the appropriate vowel the process of suffixation includes a decision on whether it is needed at all, which is, of course, a common source of error. Learners of Hungarian may learn some basic rules and then will know that no linking vowel precedes the accusative *-t* in the case of words that end in sibilants; for instance the accusative case of *busz* 'bus' is *buszt* (busz-t, bus-ACC). Also, in most cases it is not difficult to choose the linking vowel from the four variations (*-e-, -ö-, -o-, -a-*): if the word contains only front unrounded vowels (e.g. *szék* 'chair') the *-e-* linking vowel is used to form the accusative case (*széket,* szék-et, chair-ACC), while a word that contains both front and back vowels such as *templom* 'church', or only back vowels such as *kabát* 'coat' generally needs an *-o-*linking vowel (*templomot*, templom-ot, church-ACC; *kabátot*, kabát-ot, coat-ACC). The *-a-* linking vowel is not used productively: it is added only to a limited number of words of ancient Finno-Ugrian origin.

As can be seen from the above examples, besides the regular process of suffixation which obeys the rules of vowel harmony, learners may be confronted by several irregularities which may be accounted for diachronically but are difficult to remember and hence constitute another common source of errors. The rules of suffixation and vowel harmony mentioned here are, of course, only the basic ones as a systematic and complete description of this phenomenon falls outside the scope of our present study. However, these vague outlines should be sufficient to demonstrate how the most typical sources of errors arise and thus should form the basis of error coding. The error codes are summarized in Table 7:

Table 7. The most common error types

Correct TL form	Gloss	English	Possible erroneous form and description	3rd character of error code
templomot	church-ACC	church	Correct	A
házban	house-INES	in (the) house	*házben Vowel harmony problem	B
széket	chair-ACC	chair	*széköt Wrong linking vowel	C
buszt	bus-ACC	bus	*buszot Unnecessary linking vowel	D
templomot	church-ACC	church	*templomt Lack of linking vowel	E
éjfélkor	midnight-TEMP	at midnight	*éjfélker Unnecessary vowel harmony	H
tanulása	studying-3SGPOSS	its studying	*tanulásá Other suffix error	X

Possessive suffixes pose a special morphological challenge for the learners in the third person. In the third person singular possessive form the rules of vowel harmony help the learners decide if they should use a front vowel variation (*-e / -je*) or a back vowel variation (*-a / -ja*), but it is a big challenge to pick the right form with or without the *-j-* and the rules for it are quite complex, for the average language learner, it is seemingly without rhyme or reason. All the more so, because in L1 speech there are vacillations where both variations are acceptable (for example, *virága* and *virágja* are both correct for *his/her/its flower*). Using a separate character in the error code to indicate these types of errors would have probably been unnecessary so we decided instead to include the issue of the possessive *-j* in the third character of the code. The following Table 8 lists some typical cases:

Table 8. Errors associated with the possessive *-j* suffix

Correct TL form	Gloss	English	Possible erroneous form and description	3rd character of error code
könyve	könyv-3SGPOSS	his/her/its book	*könyvje Unnecessary possessive *j*	F
kabátja	kabát-3SGPOSS	his/her/its coat	*kabáta Lack of possessive *j*	G

3.2.4 Multiple suffixation

Agglutinative languages like Hungarian may use more than one suffix in the same word form. Although this feature appears in the code, errors are actually recorded in the previous three characters and this code only permits us to group the results according to the number of suffixes to see if this factor influences the performance of language learners. Hence the fourth character of the code may be 0 to indicate that no suffix is added to the word stem, and 1 or 2 to indicate that one suffix or at least two suffixes have been added to it, respectively. Currently, errors originating specifically from multiple suffixation (e.g. wrong order of suffixes) are not coded and it is also the subject of further development of how we should assign two or more erroneous suffixes in the present system.

3.2.5 Some examples

Table 9 provides a typical sample of the automatically tagged word forms.

Table 9. A sample of the automatically tagged word forms

Correct TL form	Gloss	English	Erroneous form found in the corpus and description	Error code
viszonyt	relation-ACC	relation	viszony<u>o</u>t unnecessary linking vowel	A1D1
hídjai	bridge-PL-3SGPOSS	bridges	h<u>i</u>djai wrong form of stem	C1A2
ragozást	inflexion-ACC	inflexion	r<u>á</u>gozást spelling error in stem	B1A1
tanszéken	department-SUP	at (the) department	tanszék<u>o</u>n wrong linking vowel	A1C1
gyakorlatokon	practise-PL-SUP	in practice sessions	gy<u>o</u>korlatokon spelling error in the stem	B1A2

The above examples contain the following errors: The first example contains an unnecessary linking vowel so it is marked with a label "D" as a third element of the error code. In the case of the second sample the wrong form of stem was chosen by the student, thus the word form got a label "C" in the first place of the error code. The third sample contains a spelling error in the stem as it is given by the label "B" in the first position of the error code. The error in the fourth data is the wrong linking vowel hence it has a label "C" as the third element of the error code. The last sample contains the same error as the third one; it has a spelling error in the stem so it is given the label "B" in the first place of the error code.

4 Results

To get data on the frequency of errors, we first used *magyarlanc 2.0*, a toolkit developed for the automatic linguistic processing of Hungarian texts (Zsibrita et al. 2013), which assigns a morphological code to each word in the text. Words that were tagged as unknown by the morphological parser – that is, they were treated as non-standard Hungarian words – were further investigated since we

sought to categorize morphological errors found in the corpus. For unknown words, the spellchecker *hunspell* (Trón et al. 2005) made certain recommendations, for instance, for the erroneous word *szlá*, we got *szláv* 'Slav, Slavic' and *szál* 'thread'. In cases where multiple corrections were available, the one that best fitted into the context was chosen manually. With this methodology, 60% of the unknown words could be corrected (see Row 4 in Table 10). The majority of the other cases turned out to be either proper names or foreign words, which were not included in the dictionaries of either *magyarlanc 2.0* or *hunspell*.

Focusing on the correction of nominal errors, we filtered the nouns from the data (45% of the words corrected, see Row 5 in Table 10), and we selected those that contained a morphological error, i.e. we disregarded cases where the lack of proper morphological analysis was due to segmentation errors (e.g. two words were spelt as one). Thus, 157 erroneous nouns were found in our investigations, which accounts for about 40% of the corrected words (Row 6 in Table 10). Table 10 lists the number and rate of the unknown and corrected words in the corpus and subcorpora.

Table 10. The number and rate of the unknown and corrected words

	Difficulties	England	Person I like	Total
1. Number of words	8692	3271	1622	13585
2. Number (rate) of unknown words	393 (4.52%)	146 (4.46%)	128 (7.89%)	667 (4.91%)
3. Number of corrections offered by the spellchecker	2328	614	679	3621
4. Number (rate) of accepted corrections	237 (60.31%)	110 (75.34%)	50 (39.06%)	397 (59.52%)
5. Number (rate) of corrected nouns	100 (42.19%)	58 (52.73%)	24 (48%)	182 (44.84%)
6. Number (rate) of filtered nouns	80 (33.76%)	56 (50.91%)	21 (42%)	157 (39.55%)

It should be mentioned that the rate of unknown words turned out to be much higher in the *Person I like* subcorpus than that in the other two. Moreover, here the rate of accepted words is also lower than in the other subcorpora. This may be related to the fact that due to the topic of the compositions, there are lots of proper names – primarily person and location names in the texts – which the automatic tools could not properly analyze.

4.1 Automatic error tagging

For the automatic tagging of morphological errors, we developed a rule-based system which assigns error tags to each morphologically erroneous noun. The system compares the original uncorrected form and its corrected counterpart, and automatically suggests an error tag for each case. We manually checked the quality of automatic tags in the *Difficulties* subcorpus and found that only two cases out of 80 were mistagged, which demonstrates that our system's accuracy meets our original expectations.

Automatic error tagging also made it possible for us to get statistical data on each error type. Hence, we were able to gather data on the relative frequency of stem and suffix errors on the one hand and the frequency of assimilation and

vowel harmony errors on the other. Although not directly related to morphological errors, we also checked the rate of accent errors. In the spelling system of Hungarian vowels, each grapheme has a counterpart which differs only in an accent (such as *u* and *ú*). The difference between these vowels may be their length and their phonetic values too, but here the term *accent* is used strictly in its diacritical sense. Our own preliminary analyses of the texts produced by learners of Hungarian suggested that the placement of accents could be a crucial issue for non-native speakers of Hungarian, so we also automatically counted errors related to the wrong usage of an accent. Statistical findings are presented in Table 11.

Table 11. Number of morphological errors found in the corpus

Stem errors	**134**
Spelling error in the stem	122
Wrong form of stem	12
Suffix errors	**27**
Vowel harmony problem	5
Wrong linking vowel	8
Unnecessary linking vowel	3
Lack of linking vowel	1
Lack of possessive *j*	2
Other suffix error	8
Stem + suffix errors total	**161**
Accent errors	**40**

Among stem and suffix errors in total, the most frequent error was a spelling error in the stem (76%), quite often with the improper use of accents (28% of spelling errors within the stem were related to improper accent use). Among suffix errors, selecting the wrong linking vowel was the most frequent one (29% of all suffix errors).

4.2 Automatic error correction

The manual annotation of the corrected forms also allowed us to examine the possibilities of automatic error correction. Here, we tested several simple methods for correcting the morphological errors. When we selected the first word form offered by *hunspell*, we got an accuracy score of 81.86% for the total number of corrected words, which constituted 49% of the total number of words unknown to the morphological analyzer.

In addition, we applied another method, namely, we examined which word forms offered by *hunspell* occur in the Szeged Treebank (Csendes et al. 2005), which is the largest Hungarian treebank that was manually POS-tagged and syntactically parsed. If the treebank contained more than one of the *hunspell* suggestions, we selected the one with the highest frequency value. This method resulted in an accuracy of 83%, but it could be applied only in the case of 318 words as there were corrections which did not occur in the Szeged Treebank, hence no frequency value could be obtained for them.

Next, the above two methods were combined: first, we selected from the corrections the one that occurred most frequently in the treebank. Second, for

words where no frequency data could be obtained, i.e. they did not occur in the treebank, the first option offered by *hunspell* was used. This method yielded an accuracy score of 82.62%, so it outperformed our first method.

Our results indicate that the number of erroneous word forms can be significantly reduced in Hungarian texts written by non-native speakers: about half of the errors can be eliminated by using simple methods based on spellcheckers and frequency data, which is a promising line of research for the automatic processing of non-standard texts.

4.3 Syntactic errors

There may be cases in the corpus where the word form is morphologically correct, but it does not fit into the syntactic context as the valency frame of the verb requires the presence of another case suffix. The automatic detection of such cases can only be carried out with the help of syntactic information because morphological analysis itself does not suffice. Hence, we analyzed the corpus with the dependency parser integrated into *magyarlanc 2.0*, and then we gathered valency frames from the corpus.

At the time of our investigation, altogether there were 953 valency frames in HunLearner, which were compared with those gathered from the short business news subcorpus (Vincze 2014) of the Szeged Dependency Treebank (Vincze et al. 2010). Those that were not included in the dependency treebank were examined in detail (306 valency frames, 32.11% of the total number of frames). As a first step, we filtered verbs with an empty valency frame since pronominal subjects and objects can remain phonologically covert in a Hungarian sentence and with automatic methods, missing but otherwise required arguments cannot be clearly distinguished from those omitted for grammatical reasons. Afterwards, we got 278 valency frames (29.17%). In 37 cases, one of the arguments was labeled as an unknown word by the POS-tagger, and this incorrect morphological tagging resulted in an incorrect syntactic parse of the sentence. To sum up, there are 241 valency frames (25.29%) in HunLearner which should be further examined. As our preliminary studies show, some of these problematic valency frames are indeed erroneous (e.g. *nekem nem érdekel* (I-DAT not be.interested-3Sg) instead of *engem nem érdekel* (I-ACC not be.interested-3Sg) "I am not interested"). In other cases, the dependency parser yields an incorrect parse. Furthermore, there are valency frames which are grammatically sound but they just do not happen to occur in the Szeged Dependency Treebank so they ended up in this category (e.g. *felvág valamivel* (up.cut something-INS) "to show off with something"). Later, we would like to check the frequency of the above error types associated with valency frames and also see how the number of erroneous valency frames can be reduced by using automatic methods.

5 Summary and suggestions for future work

In this paper we presented a method for processing a corpus of Hungarian language students. Our method utilized a variety of automatic and morphological tools in order to examine texts for errors.

We collected Hungarian texts written by 35 students majoring in Hungarian Studies at the University of Zagreb in Croatia and put them through our processing method, using the automatic NLP tools *magyarlanc* and *hunspell*. These automatic tools were originally developed to process data from Hungarian native speakers. The collected datasets were then analyzed by *magyarlanc*, which consists of a sentence splitter, a morphological analyzer, a POS-tagger and a dependency parser. During this processing phase, 667 unknown word forms were detected. The correct forms of these words were then manually selected by annotators using recommendations offered by the Hungarian spellchecker *hunspell.* It was found that if the first suggestion offered by *hunspell* was automatically accepted, an accuracy rate of 82% could be achieved. This shows that relatively simple methods can significantly reduce the number of incorrect word forms in a non-standard text. This is a promising result and it appears to show that the automatic processing of non-standard Hungarian texts can be quite effective.

Our next task was to tag the morphological errors in the corpus and this was carried out by our automatic error tagger, which we developed after taking into account the special characteristics of Hungarian morphology and the morphological errors commonly made by learners of Hungarian. This automatic error tagger proved to be efficient. We manually checked it on a small sub-sample and found that the quality met our expectations; only two cases out of eighty were errors incorrectly tagged.

We also applied several simple methods for automatic error correction, the results of which indicate that about half of the originally unknown words could be assigned a proper correction. Furthermore, we took some preliminary steps in the direction of automatically correcting syntactic errors and we concluded that by relying on a valency database, some of the syntactic errors can be easily identified.

In the future, we would like to further extend our corpus with new texts and to carry out more complex investigations into other types of errors made by learners of Hungarian. We also intend to extend the database with Hungarian texts written by learners with various L1 backgrounds. This study will provide an opportunity for us to carry out comparative studies on Hungarian language products of speakers with different mother tongues.

Later on, we plan to develop new automatic methods to process the errors in syntax and word usage and then apply them in different areas. We also think that dictionaries and gazetteers of named entities can be integrated into the morphological parser, with special regard to the nationality and geographical background of the creators and the topics of the texts. For instance, in the case of HunLearner, dictionaries of Croatian names and locations could prove beneficial for our future research.

The corpus is freely available on the homepage [http://rgai.inf.u-szeged.hu/hunlearner].

Acknowledgement

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013). Veronika Vincze was partially funded by the National Excellence Program TÁMOP-4.2.4.A/2-11/1-2012-0001 of the State of Hungary, co-financed by the European Social Fund.

Abbreviations

ACC = accusative
DAT = dative
INES = inessive
INST/COM = Instrumental / Comitative
NOM = nominative
PL= plural
3SGPOSS = 3rd person singular possessive
SUP = superessive
TEMP = tempora

References

Ács, P. & Siptár, P. 1994. Túl a gondozott beszéden. [Beyond careful speech]. In K. Ferenc (ed.), *Strukturális magyar nyelvtan. Fonológia.* Budapest: Akadémiai Kiadó, 550–578.

Csendes, D., Csirik, J., Gyimóthy, T. & Kocsor, A. 2005. The Szeged Treebank. In V. Matousek, P. Mautner & T. Pavelka (eds.), *Text, Speech and Dialogue, 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005, Proceedings*. Berlin - Heidelberg, Germany: Springer, 123–131.

De Cock, S. & Granger, S. 2005. Computer learner corpora and monolingual learners' dictionaries: the perfect match. *Lexicographica* 20, 72–86.

Dickinson, M. & Ledbetter, S. 2012. Annotating Errors in a Hungarian Learner Corpus. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.), *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey, May 23–25, 2012. European Language Resources Association (ELRA), 1659–1664. [Retrieved April 27, 2012]. Available at http://www.lrec-conf.org/proceedings/lrec2012/pdf/758_Paper.pdf

Durst, P. 2010. *A magyar mint idegen nyelv elsajátításának vizsgálata – különös tekintettel a főnévi és igei szótövekre, valamint a határozott tárgyas ragozásra.* [The acquisition of Hungarian as a foreign language – with special attention to nominal and verbal stems and the definite verb conjugation]. Unpublished PhD thesis, University of Pécs, Hungary.

Gass, S. & Selinker, L. 2001. *Second Language Acquisition: An Introductory Course.* Second Edition. Hillsdale: Lawrence Erlbaum Associates.

Granger, S. 2002. A Bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins, 3–33.

Granger, S. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20, 465–480.

Selinker L. 1972: Interlanguage. *IRAL* 10, 209-230.

Suni, M. 2013. The impact of Finno-Ugric languages in second language research: looking back and setting goals. *Lähivõrdlusi. Lähivertailuja* 22, 407–435. [Retrieved February 8, 2013]. Available at http://www.rakenduslingvistika.ee/ajakirjad/index.php/lahivordlusi/article/view/LV22.14/203

Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. & Varga, D. 2005. Hunmorph: open source word analysis. In M. Jansche (ed.), *Proceedings of ACL Workshop on Software*. June 27, 2005. Michigan, USA: Association for Computational Linguistics, 77–85. [Retrieved May 17, 2013]. Available at http://clair.eecs.umich.edu/aan/paper.php?paper_id=W05-1106#pdf

Vincze, V. 2014. Valency frames in a Hungarian corpus. *Journal of Quantitative Linguistics*, 21 (2), 153–176.

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z. & Csirik, J. 2010. Hungarian Dependency Treebank. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, May 17–23, 2010. Paris, France: European Language Resources Association (ELRA), 1855–1862. [Retrieved May 17, 2013]. Available at http://www.lrec-conf.org/proceedings/lrec2010/pdf/465_Paper.pdf

Zsibrita J., Vincze V. & Farkas, R. 2013. magyarlanc 2.0: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: G. Angelova, K. Bontcheva & R. Mitkov (eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, Bulgaria, September 9–11, 2013. Hissar, Bulgaria: Incoma Ltd Shoumen, 763–771.

Received December 9, 2013
Revision received November 10, 2014
Accepted December 5, 2014