Automatic morphological analysis of learner Hungarian

Scott Ledbetter
Indiana University
Bloomington, IN, USA
saledbet@indiana.edu

Markus Dickinson
Indiana University
Bloomington, IN, USA
md7@indiana.edu

Abstract

In this paper, we describe a morphological analyzer for learner Hungarian, built upon limited grammatical knowledge of Hungarian. The rule-based analyzer requires very few resources and is flexible enough to do both morphological analysis and error detection, in addition to some unknown word handling. As this is work-in-progress, we demonstrate its current capabilities, some areas where analysis needs to be improved, and an initial foray into how the system output can support the analysis of interlanguage grammars.

1 Introduction and Motivation

While much recent research has gone into grammatical error detection and correction (Leacock et al., 2014), this work has a few (admitted) limitations: 1) it has largely focused on a few error types (e.g., prepositions, articles, collocations); 2) it has largely been for English, with only a few explorations into other languages (e.g., Basque (de Ilarraza et al., 2008), Korean (Israel et al., 2013)); and 3) it has often focused on errors to the exclusion of broader patterns of learner productions—a crucial link if one wants to develop intelligent computer-assisted language learning (ICALL) (Heift and Schulze, 2007) or proficiency classification (Vajjala and Loo, 2013; Hawkins and Buttery, 2010) applications or connect to second language acquisition (SLA) research (Ragheb, 2014). We focus on Hungarian morphological analysis for learner language, attempting to build a system that: 1) works for a variety of morphological errors, providing detailed information for each; 2) is feasible for low-resource languages; and 3) provides analyses for correct and incorrect forms, i.e., is both a morphological analyzer and an error detector. Perhaps unsurprisingly, we find that the best way to accomplish these goals is to hearken back to the *parsing ill-formed input* literature (see Heift and Schulze, 2007, ch. 2) and develop a rule-based system, underscoring the point that different kinds of linguistic properties require different kinds of systems (see Leacock et al., 2014, ch. 7).

We hope to make the analysis of Hungarian morphology maximally useful. Consider ICALL system development, for example: successful systems not only provide meaningful feedback for learners but also model learner behavior (e.g., Amaral and Meurers, 2008). To do this requires tracking correct and incorrect use of different linguistic phenomena (e.g., case). Furthermore, one likely wants to keep track of individual differences between learners as well as to track general developmental trends—a point relevant to SLA research more generally (Dörnyei, 2010; Gass and Selinker, 2008).

In addition to providing a platform for ICALL development and SLA research, another long-term goal of our project is to develop an annotated corpus of learner Hungarian, including both linguistic and error annotation. The exact delineation between the two kinds of annotation is an open question (Ragheb and Dickinson, 2014), and building an analyzer which does both can show the link for at least certain types of errors. Additionally, the link between corpus data and automatic analysis is part

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 31–41,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

of an important feedback loop: if one views error detection as the relaxation of grammatical constraints (Reuer, 2003; Schwind, 1995), it is important to determine which constraints may be relaxed—given the huge space of possible variation (e.g., reordering affixes)—and this work is a step in that direction.

One further point is worth mentioning: the analyzer we describe makes use of a limited amount of grammatical knowledge in a rule-based system, allowing for potential application to other languages with minimal effort and resources. Our hope is that this can provide a basis for research into other lesser-resourced languages and some less-investigated error types. The system is also flexible and adaptable, designed to allow for the variation and inconsistencies expected of early learner language.

The paper is organized as follows. In Section 2 we discuss facts about Hungarian relevant for building an analyzer, as well as previous research in relevant areas, and in Section 3 we describe the data used for analysis. We turn to the actual analyzer in Section 4, employing a simple chart-parsing strategy that allows for feature clashes and crucially relies on a handful of handwritten affixes, which essentially encode the "rules" of the grammar (i.e., the approach is fairly lexicalized). The evaluation in Section 5 is tripartite, reflecting our different goals: evaluating the quality of assigned morphological tags (Section 5.1), the error detection capabilities (Section 5.2), and the ability to extract information for learner modeling (Section 5.3). The work is still in progress, and thus the evaluation also points to ways in which the system can be improved.

2 Background and Previous Work

2.1 Hungarian

Hungarian is an agglutinative language belonging to the Finno-Ugric family. It has a rich inflectional and derivational morphological system, as illustrated in (1). Verbs take suffixes to indicate number, person, tense, and definiteness, as in (1a), in addition to suffixes which alter aspectual quality or modality. Nouns, meanwhile, take suffixes for number, internal and external possession, and case (1b), of which there are 20 (e.g. inessive in (1b)), many of which roughly correspond to adpositions in other languages. Allomorphs of most suffixes are selected based on vowel harmony, for which features (e.g. +BK) must match, as with the inessive case in (1b) and (1c). For both verbs and nouns, the ordering of grammatical suffixes is fixed (Törkenczy, 2008).

(1) a. fut -ott -ál
run -PST -2SG.INDEF
'you [2sg.] ran'

b. könyv -eim -ben
book[-BK] -1SG.PL[-BK] -INESSIVE[-BK]
'in my books'

c. ház -ban
house[+BK] -INESSIVE[+BK]
'in (a) house'

The rich morphology of Hungarian necessitates taking the morpheme as the basic unit of analysis. A single morpheme can convey a wealth of information (e.g. person, number, definiteness on verb suffixes), and a sufficiently extensive set of phonological and morphological features must be used, particularly if one is to capture individual variation.

2.2 Morphological analysis for Hungarian

Morphological analysis for agglutinative languages tends to be based on finite-state transducers (Koskenniemi, 1983; Oflazer, 1994; Özlem Çetinoğlu and Kuhn, 2013; Aduriz et al., 2000). These are robust, but the process is not quickly adaptable to other languages, as every rule is language-specific, and there is no clear way to handle learner innovations.

For Hungarian, HuMor (High-speed Unification Morphology) (Prószéky and Kis, 1999) uses a bank of pre-encoded knowledge in the form of a dictionary and feature-based rules. Megyesi (1999) extends the Brill tagger (Brill, 1992), a rule-based tagger, with simple lexical templates. Tron et al. (2005) derive a morphological analyzer, Hunmorph, from a language-independent spelling corrector, using a recursive affix-stripping algorithm that relies on a dictionary to remove affixes one by one until a root morpheme is found. The dictionary is customizable to other languages, and the idea of using affix-removal to identify stems is similar to our technique

(Section 4). Morphdb (Trón et al., 2006), a lexical database for Hungarian, encodes only irregularities and uses features on the appropriate lexical items to apply the proper phonological and morphological processes during analysis. These various tools have been incorporated into a variety of other Hungarian systems (Halácsy et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013). For approaches like Hunmorph and Morphdb that rely on a dictionary, unknown words are the main problem—also a crucial issue for innovative learner forms.

2.3 Grammatical error detection

There is some work exploring morphological derivations in learner language. Dickinson (2011) looks for stem-suffix mismatches to identify potential errors (for Russian) and uses heuristics to sort through multiple analyses. There is, however, no evaluation on learner data. We focus on building a small grammar to explicitly license combinations and provide a variety of evaluations on real learner data. Prior work in L2 Hungarian uses the HunLearner corpus (Durst et al., 2014; Vincze et al., 2014) to develop systems to automatically identify errors. Our work explores similar directions, focusing not only on the identification of non-target forms but also systematically describing them and making that information available in the form of morphological annotation.

The work presented here is related to the idea of constraint relaxation and constraint ranking (e.g., Menzel, 2006; Schwind, 1995), wherein grammatical constraints are defeasible (see Leacock et al., 2014, ch. 2). In the case of morphology, the primary process of relaxing constraints is in allowing stems and affixes to combine which are generally not allowed to do so (see also Section 4).

There is a wealth of research on statistical error detection and correction of grammatical errors for language learners (Leacock et al., 2014), including for Hungarian (Durst et al., 2014; Vincze et al., 2014). As has been argued before (e.g., Chodorow et al., 2007; Tetreault and Chodorow, 2008), statistical methods are ideal for parts of the linguistic system difficult to encode via rules. Since Hungarian morphology is a highly rule-governed domain of the language and since we want detailed linguistic information for feedback, we do not focus on statistical methods here. We hope, however, to eventually obtain an appropriate distribution of errors in order to incorporate probabilities into the analysis.

The emphasis on rule-based error detection allows one to connect the work to broader techniques for modeling learner behavior, in the context of ICALL exercises (Thouësny and Blin, 2011; Heift, 2007) or in mapping and understanding development (cf. Vajjala and Loo, 2013; Vyatkina, 2013; Yannakoudakis et al., 2012). Our evaluation thus focuses on multiple facets of the output and its use (Section 5).

3 Data and Annotation

3.1 Corpus

The corpus was collected from L1 English students of Hungarian at Indiana University and is divided into three levels of proficiency (Beginner, Intermediate, Advanced) as determined by course placement in one of three two-semester sequences. The corpus consists of journal entries, each a minimum ten sentences in length on a topic selected by the student.

The corpus at present contains data for 14 learners (9 Beginner, 1 Intermediate, 4 Advanced), 9391 sentences total, with 10 annotated journals. The corpus represents both cross-sectional and longitudinal data. Productions from multiple learners can be compared across or (for beginners) within proficiency levels, and a single learner's data over time can also be analyzed. Additionally, passages are often longer and feature more descriptive language than those produced for grammatical exercises.

3.2 Annotation

Each journal has been transcribed manually and annotated for errors with EXMARaLDA (Schmidt, 2010).[1] The text is segmented on morpheme boundaries, and errors are identified in four different tiers, matched to a target form. The annotation scheme is specifically for Hungarian, but the principles behind it can be extended to other morphologically rich languages (Dickinson and Ledbetter, 2012).

The annotation marks different types of errors re-

[1] http://www.exmaralda.org/en_index.html

flecting different levels of linguistic analysis. For instance, for (2), the annotation shows a CL (vowel length) error on the verb stem and an MAD (definiteness) error on the verb suffix—i.e. the definite suffix does not agree with the indefinite noun complements—as shown in Figure 1.

(2) **Ajanl** **-om** bor -t , nem sör -t
recommend 1SG.DF wine ACC , not beer ACC
'I recommend wine, not beer.'

TXT	Ajanlom		bort		,	nem	sört		.
SEG	Ajanl	om	bor	t	,	nem	sör	t	.
CHA	CL								
MOR		MAD							
TGT	Ajánl	ok	bor	t	,	nem	sör	t	.

Figure 1: Error annotation for (2)

There are four basic error annotation categories, reflecting *character* (CHA, e.g., vowel harmony, phonological confusion), *morphological* (MOR, e.g., agreement in person, case), *grammatical relation* (REL, e.g., case, root selection), and *sentence* (SNT, e.g., insertion, ordering) errors. A full list of categories can be found in Dickinson and Ledbetter (2012). Different categories of errors can be annotated for the same word, and error spans can overlap if necessary. A target (TGT) sentence is also provided. The morphological analyzer discussed in section 4 is designed to recognize errors within the morphological (MOR) and character (CHA) tiers.

4 Morphological Analysis

Our goal for analyzing a word is to provide its derivation, in order to support morphological analysis, error detection, and learner modeling. A derivation here refers to a breakdown of a word's internal structure into individual morphemes, i.e., a root morpheme plus affixes, and we want to provide as much of a derivation as we can even when: a) the root is unknown, or b) the learner has misapplied an affix (e.g., it is inappropriate for the rest of the word). We discuss the knowledge base (Section 4.1), the basic algorithm (Section 4.2), and our first pass at making the analyzer more robust (Section 4.3).

4.1 Knowledge base

There are two parts to the knowledge base, a handcrafted suffix base and a dictionary obtained from another project. The dictionary is obtained from *A Magyar Elektronikus Könyvtár*.[2] To model lesser-resourced situations, one can experiment with differing sizes of this lexicon; in general, this type of resource does not have to contain much information.

The suffix base, on the other hand, is where we encode the rules for morphological combination, and it thus must be developed with more care. We use 205 affixes, including those for noun case, plurals, verb conjugation, and possession. An affix corresponds to a set of possible categories, the encoding inspired by the Combinatory Categorial Grammar (CCG) framework (Steedman and Baldridge, 2011). For example, the accusative case marker *-t* has one possible category KN\N, indicating that it would create a new category KN (cased noun phrase) if it was combined with a noun (N) on the left.

Each affix category contains features describing relevant linguistic properties. For example, features for the entry for the affix *-ot* indicates that: a) it contains back vowels and b) it is accusative case when combined with a noun stem. As another example, the plural noun suffix *-ok* also contains back vowels, but its features furthermore indicate a stem-lowering effect—i.e. successive affixes must adhere to a restricted subset of allomorphs based on vowel harmony. The suffix base represents our grammar engineering, but, as noted, it is quite small.

4.2 Building an analysis

To efficiently determine the correct combinations of root and affixes, we use a basic CYK chart parsing algorithm (Cocke and Schwartz, 1970), treating each letter as a unit of analysis; as suffixes drive the analysis, we process from right to left. At each possible interval of starting and ending sequences within a word, the system verifies if the sequence is either attested in the affix base or in the dictionary of attested language forms. If the sequence is found, a corresponding category is placed into

[2] http://www.mek.iif.hu/porta/szint/egyeb/szotar/ssa-dic/

the chart. While finite-state techniques are the standard for morphological analyzers (section 2.2), chart parsing is easy to implement and makes the processing architecture extendible to syntactic phenomena.

Consider *házot* ('house+ACC'), indexed in (3) and with a corresponding chart in Figure 2. Here, both *-t* and *-ot* can be suffixes, but as only *ház*—and not *házo*—is a verified noun (the N in cell 2–5), the segmentation *ház+ot* provides the correct analysis.

(3) $*_5$ h $_4$ á $_3$ z $_2$ o $_1$ t $_0$

					4
					3
N					2
N_{hyp}					1
KN			KN\N	KN\N	0
5	4	3	2	1	

Figure 2: Chart for (3)

As the system is affix-driven, if no root is found matching an item in the dictionary, the system can posit a possible stem for the word based on the affixes that *were* found. This possible stem is then added to the chart like an attested root, with the information noted that it is hypothesized, indicated here as N_{hyp} in cell 1–5. This ability to hypothesize is an important feature of the analyzer, as it allows for "erroneous" or "nonstandard" root morphemes, crucial to analyzing learner language.

4.3 Constraint relaxation

When general categories are combined in the chart (Section 4.2), features of affixes and stems are also compared. Any inconsistencies violating the grammar of Hungarian are marked. A sample derivation obtained from the chart in Figure 2 is given in Figure 3, here with one feature shown. The stem requires a lowered allomorph (*-at*) of the accusative suffix, but the unlowered allomorph is provided.

$$\dfrac{\dfrac{\text{h}\quad\text{á}\quad\text{z}}{\text{N[+LOW]}}\qquad\dfrac{\text{o}\quad\text{t}}{\text{KN}\backslash\text{N[-LOW]}}}{\text{KN[!LOW]}}$$

Figure 3: Feature clash during derivation

The feature clash here indicates a learner innovation, providing some analysis of the their current understanding of the language. Importantly for processing, we currently require: a) equivalence of main categories (e.g., KN\N must combine with N), and b) proper ordering of affixes. Neither of these relaxations seemed to be required for our data, though future analysis may prove otherwise. In that light, we can note the importance of the grammar-writer to put relaxable constraints (e.g., sub-category information) into features and non-relaxable constraints into the main categories.

5 Evaluation

As mentioned earlier, we evaluate the system in three different ways. First, we treat the system as a straight morphological analyzer and evaluate the quality of assigned morphological tags (Section 5.1). Secondly, employing some constraint relaxation abilities, we evaluate the system's capabilities in performing error detection (Section 5.2). Finally, we illustrate the ability of the system to provide information on interlanguage grammars, namely the ability to help distinguish between individual learners and levels of learners (Section 5.3).

5.1 Morphological analysis

The system is first evaluated in terms of accuracy of morphological analysis, both on native (L1) and learner (L2) data. For every word, the system returns one or more derivations, representing the internal structure of the word, and the associated morphological features, here represented as a morphological code. Take, for example, the verb in (4a).

(4) a. lát -t -ál
see -PST -2SG.INDEF
'you saw'

b. V m i s 3 s - - - n
0 1 2 3 4 5 6 7 8 9

The morphological code in (4b) for the verb follows the scheme used to annotate the Szeged Corpus (Csendes et al., 2004), applicable to multiple languages. Each numbered field corresponds to a feature, and different letters or numbers give the values.

After the initial verb indicator (V), the code in (4b) indicates: main verb (m), indicative mood (i), past tense (s), third person (3), singular (s), indefinite (n). Three fields are unused (e.g., one for grammatical gender, not found in Hungarian).

As the system is fairly resource-light (Section 4.1), we do not expect state-of-the-art accuracy, but we do need to gauge whether it is effective enough for our purposes and to know how to improve for the future. We start by investigating its general accuracy on L1 data, presenting the analyzer with a selection of native Hungarian data from the Szeged Corpus (Csendes et al., 2004), taking the first 1000 tokens from a section of compositions (in order to verify results by hand and to compare to the 1021 tokens of learner data discussed below). The results are in the *Total* column of Table 1.

	Total	POS	+N	POS+N
Precision	0.308	—	0.307	—
Recall	0.262	—	0.315	—
Accuracy	0.467	0.568	0.505	0.592
Unk. POS	0.425	0.425	0.425	0.425
Unk. Word	0.067	0.067	0.067	0.067

Table 1: Morphological analysis on L1 Hungarian data

The corpus provides both a single, context-specific tag and a list of all appropriate tags, and we use a set of measures to reflect this situation. **Precision** is calculated as the number of codes produced by the analyzer that appear in the gold standard *list* divided by the total number of codes produced, and **recall** is the number of codes produced by the analyzer that appear in the gold standard *list* divided by the total number of codes in the gold standard. **Accuracy** is the percentage of cases where the analyzer produces, among its output, the correct *context-specific* gold tag. As the analyzer doesn't have access to part of speech data in its dictionary, it may recognize a word but have no tag for it, in which case it produces an **unknown POS** tag. Finally, when the analyzer cannot produce a derivation, it returns an **unknown word** tag.

We can see in Table 1 that the analyzer provides the correct tag in only 47% of the 1000 test cases. Yet the frequency of the *unknown POS* tag indicates that nearly half of the time, the analyzer recognizes the word but cannot determine its internal structure—i.e., we are not positing incorrect codes so much as positing nothing. The majority of these words are monomorphemic nouns, pronouns, adjectives, or adverbs: without the overt morphology indicated by the affixes in the knowledge base, the analyzer relies only on the dictionary, which contains no information about part of speech. Precision and Recall seem fairly low, but a closer inspection of the data reveals that a number of codes are mostly correct, differing from the gold standard by only one or two fields. Taking into account only part of speech (*POS*), accuracy increases to nearly 57%.

Because nouns were one of the most common parts of speech for which the analyzer could determine no structure, a second evaluation was performed, positing an additional noun tag in each case where the *unknown POS* tag was returned (*+N*). Precision fell by a slim margin (due to the increase in proposed tags), while Recall rose by about 5% and Accuracy by 4%. Taking into account only part of speech (*POS+N*, Accuracy reaches 59%.

Our second analysis targets learner data. In this analysis, the corrected forms for 1021 words produced by L2 Hungarian learners were manually annotated with morphological codes from the Szeged Corpus scheme. These gold standard codes were compared to those returned by the analyzer, as above with the native data. The design of the analyzer emphasizes flexibility, and we compare stricter and more permissive derivations, ignoring feature clashes that would otherwise result in an incomplete parse of a given word (Section 4.3). Results are in Table 2, where $Total_{Strict}$ reflects the performance of the analyzer when run with strict settings, i.e., no feature clashes allowed, and $Total_{Free}$ reflects performance when feature clashes are allowed (and recorded) during derivation. The same tokens were also analyzed by the *magyarlanc* tool (Zsibrita et al., 2013), developed for analyzing the standard language, as a benchmark (*ML*). *Magyarlanc* returns only one analysis per word, and thus accuracy was the principal measure for comparison.

Accuracy is on a par with the native L1 data when the system is used with strict settings, and approxi-

	Total$_{Strict}$	Total$_{Free}$	ML
Accuracy	0.499	0.509	0.846
Unk. POS	0.499	0.499	—
Unk. Word	0.109	0.097	0.027

Table 2: Morph. analysis on corrected L2 Hungarian data

mately half of the test cases were recognized by the analyzer. With flexibility, accuracy increases by 1% and the unknown word rate decreases by about the same margin. *Magyarlanc* outperforms the system, but even on corrected learner data, accuracy is 85%.

The final analysis is on raw learner data (the same 1021 words with no corrections) to test the analyzer's flexibility with the idiosyncracies in authentic learner language. Results are in Table 3.

	Total$_{Strict}$	Total$_{Free}$	ML
Accuracy	0.464	0.478	0.753
Unk. POS	0.456	0.456	—
Unk. Word	0.137	0.119	0.074

Table 3: Morph. analysis on raw L2 Hungarian data

Accuracy is still fairly low, with a slim increase in performance with the more permissive settings. With *magyarlanc*, accuracy falls by about 10%. For both, the unknown word rate is higher than with corrected data. Again, a large proportion of the test cases involve monomorphemic words for which the analyzer recognizes no internal structure. Access to POS data, as with *magyarlanc*, would greatly improve performance. In general, however, an emphasis on flexibility and adaptability seems to have benefits for describing learner language, decreasing unknown word rate and maintaining accuracy.

5.2 Error detection

The next evaluation assesses the system's ability to automatically detect errors in learner data. As discussed in Section 4.3, an error occurs when features clash (cf. Figure 3). Feature clashes also arise from unknown words, as the category of a word not in the dictionary is unspecified. Evaluation of the system as a whole is given in the *Total* column of Table 4.

	Total	Morph	Char
Precision	0.380	0.380	0.380
Recall	0.625	0.789	0.938
F$_1$	0.472	0.513	0.541
F$_{0.5}$	0.412	0.424	0.431

Table 4: Error detection using only dictionary stems

Precision is the number of correctly identified errors divided by the number of errors suggested by the analyzer. **Recall** is the number of correctly identified errors divided by the number of errors in the gold annotation. The **F$_1$ score** is the harmonic mean of precision and recall; because precision is critical when providing feedback to learners, **F$_{0.5}$** is also given, weighing precision more heavily. Precision in Table 4 is very low, below 40%; i.e., 60% of the "errors" identified by the morphological analyzer are false positives. Recall is better, at over 60%.

The morphological analyzer is not currently designed to handle syntax errors or many agreement errors, as it considers only one word at a time. Thus, additional scores are calculated for errors below the tier of syntax (see Section 3.2). In the *Morph* column, only those errors from the morphological tier and below are considered (i.e., *Morph* and *Char*). For *Char*, only those errors from the character tier are considered. Recall improves considerably by this restricted focus, up to nearly 94% for *Char*.

Considering the importance of precision, the analyzer needs much improvement. A closer analysis illustrates some of the problems with the algorithm and with the test data. The vast majority of false positives (˜40%) are for proper names. Most named entities are obviously not in the dictionary (excepting, e.g., *Magyarország* 'Hungary'), and the system cannot recognize them. As described in Section 4, the analyzer can posit hypothetical stems to complete a derivation, estimating words as they exist in the learner's vocabulary. A second evaluation was performed, allowing the system to hypothesize that any unknown word may be a valid item in the learner's vocabulary. Results are in Table 5.

Precision sees a modest increase to 40%, while recall falls to less than 10%. Limiting the scope of analysis once more increases recall (to nearly 7%), but the F-scores remain less than half of those in the

	Total	Morph.	Char.
Precision	0.400	0.400	0.400
Recall	0.038	0.043	0.067
F_1	0.070	0.078	0.114
$F_{0.5}$	0.139	0.152	0.200

Table 5: Error detection including hypothesized stems

previous evaluation. Investigating the system's performance more closely once again reveals a problem with unknown words and proper names. While the analyzer is able to posit hypothetical lexical entries, including nouns, it is impractical to allow any unknown word to be a potential noun. One of the most frequent errors, especially for beginners, is vowel length. Allowing any word to be hypothesized allows any number of these errors to go unnoticed. A possible solution for vowel length errors is to run a spelling corrector as part of the pipeline (Durst et al., 2014), and more generally a short list of common Hungarian names could improve performance.

Another problem for the analyzer is the appearance of irregular stems in the derivation. For example, the analyzer correctly produces a derivation for *megyek* ('I go', dictionary form *megy*) but not for *mennek* ('they go'). The derived base form *men* must be deemed a new word and potential error. One way to combat this problem is to encode irregular lexical items into the knowledge base of the system.

One final issue is the limited scope of the system. The most frequent source of errors is due to Hungarian's extensive case system. The analyzer can identify accusative or nominative case on nouns, for example, but because it considers each word individually, it cannot determine whether there is an error. Performance improves when excluding such types, but adding context-sensitivity is a crucial future step.

5.3 Grammar extraction

The final evaluation is the most exploratory, involving the extraction of properties which might be useful for comparing different learners. The space of possibly relevant metrics is quite large (Lu, 2010, 2012; Vajjala and Loo, 2013; Vyatkina, 2013; Yannakoudakis et al., 2012), and in this exploratory study we focus on a small number of metrics surrounding: a) complexity, and b) paradigm coverage. An overall goal is to sort out features which are good at distinguishing learner level from those which characterize individual learner differences.

Complexity Complexity is often used to describe the syntax of learners and the structure of their sentences. We consider the average number of **morphemes per word (MPW)** and of **words per sentence (WPS)**. Tokenization and segmentation are performed by the analyzer (and checked for accuracy). The last five journal entries for each learner are analyzed, to avoid masking change over time, as interlanguage is always changing.

	MPW	WPS
Beg01	1.38	5.79
Beg02	1.40	4.37
Beg03	1.52	3.84
Beg04	1.31	5.43
Beg06	1.52	5.75
Beg08	1.44	2.81
Beg09	1.58	3.28
Int01	1.51	6.40
Adv01	1.60	15.73
Adv02	1.66	10.90

Table 6: Complexity measures for learners of Hungarian

The beginning learners produce a range of morphemes per word, with some even approaching the production of the advanced learners. Even the least morphologically productive learner (Beg04) attains 1.31 morphemes per word. This particular aspect of morphological complexity, while it increases with greater proficiency, seems to be a largely individual feature of learner language, making it a potential candidate for classification tasks to identify specific learners or to characterize individual differences. Sentence length, while it has individual variation, seems to increase over the course of acquisition and thus may be an indicator of proficiency.

Coverage Taking Hungarian's morphological richness into account, we propose **paradigm coverage** to represent the frequency of different verb forms within the same tense and mood (here, present indicative), thus showcasing how much of the paradigm space a learner is using. Any occurrence of the appropriate verbal affix on any verb is counted, and the sum of the affix frequencies is normalized by dividing by the number of journal entries. Given space constraints, only one beginning and one advanced learner are presented in Figures 4 and 5. Average frequencies for the indefinite form are in light gray and for the definite in dark gray.[3]

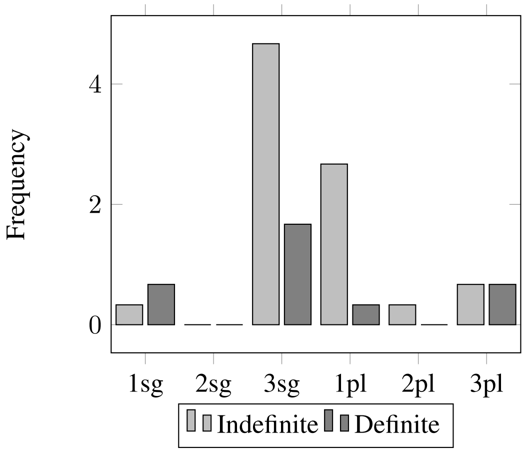

Figure 4: Affix coverage for learner Beg01

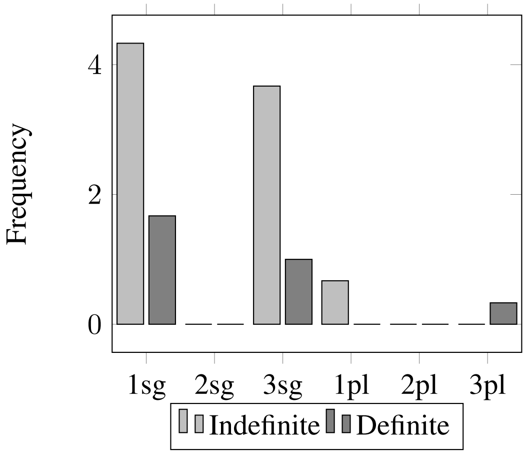

Figure 5: Affix coverage for learner Adv02

While there are definitely genre effects (e.g., lack of second person), the individual differences here may help form a more complete picture of a learner's interlanguage. Learner Beg01 appears to have some of the most complete knowledge of the present indicative paradigm among beginners, with representation in the first and third person singular and plural, definite and indefinite. Learner Adv02 exhibits many instances of the first person, characteristic of narrative description. This metric seems to be unique to individual learners (and their choice of topic), as some beginning learners exhibit more complete paradigms than the advanced learners.

To return to the theme of the whole paper: regardless of the conclusions drawn exactly from such paradigms, it is only by automatic morphological analysis that one is able to investigate differences in morphological complexity and paradigm coverage.

6 Summary and Outlook

We have presented a rule-based morphological analysis system for learner Hungarian, employing constraint relaxation, and have performed three different evaluations to illustrate its utility for linguistic analysis, error analysis, or downstream applications. We have used very little in the way of handbuilt resources, and, while the system still needs improvement, the information captured by the analyzer already shows promise for describing the interlanguage of learners of Hungarian.

There are a number of ways to improve the system. Named entities in particular have been a problem for other approaches (Durst et al., 2014), and we intend to use similar methods to increase accuracy, including lists of common names. While syntactic context is presently unavailable to the analyzer for disambiguation, we hope to extend the methodology to syntax in the future. We also intend to explore how a record of language use may aid in disambiguation: if an ambiguous stem has only ever occurred previously with verbal morphology, for example, there is a good chance that its current use is as a verb. Finally, given a desire to be resource-light and applicable to other languages, one may investigate iterative bootstrapping methods to allow for the reduction of the initial size of the knowledge base, instead building a gradual inventory through analyzing a set of learner data itself.

[3] Definiteness is decided by the object of the verb, i.e., *a cat* (indefinite) or *the cat* (definite).

Acknowledgments

We would like to thank the participants of the IU CL discussion group, as well as the three anonymous reviewers, for their many helpful comments.

References

Itziar Aduriz, Eneko Agirre, Izaskun Aldezabal, Iñaki Alegria, Xabier Arregi, Jose Maria Arriola, Xabier Artola, Koldo Gojenola, Aitor Maritxalar, Kepa Sarasola, and Miriam Urkia. 2000. A word-grammar based morphological analyzer for agglutinative languages. In *Proceedings of the 18th conference on Computational linguistics (COLING 2000), vol. 1*, pages 1–7.

Luiz Amaral and Detmar Meurers. 2008. From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for intelligent computer-assisted language learning. *Computer-Assisted Language Learning*, 21(4):323–338.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajic. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1(1):415–428.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 112–116.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30.

John Cocke and Jacob T. Schwartz. 1970. *Programming languages and their compilers: Preliminary notes*. CIMS, NYU, second edition.

Dóra Csendes, János Csirik, and Tibor Gyimóthy. 2004. The szeged corpus: A pos tagged and syntactically annotated hungarian natural language corpus. In *Text, Speech and Dialogue: 7th International Conference, TSD*, pages 41–47.

Arantza Díaz de Ilarraza, Koldo Gojenola, and Maite Oronoz. 2008. Detecting erroneous uses of complex postpositions in an agglutinative language. In *Proceedings of COLING-08*. Manchester.

Markus Dickinson. 2011. On morphological analysis for learner language, focusing on russian. *Research on Language and Computation*, 8(4):273–298.

Markus Dickinson and Scott Ledbetter. 2012. Annotating errors in a hungarian learner corpus. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*.

Zoltán Dörnyei. 2010. *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Routledge.

Péter Durst, Martina Katalin Szabó, Veronika Vincze, and János Zsibrita. 2014. Using automatic morphological tools to process data from a learner corpus of hungarian. *Apples Journal of Applied Language Studies*, 8(3):39–54.

Richárd Farkas, Veronika Vincze, and Helmut Schmid. 2012. Dependency parsing of hungarian: Baseline results and challenges. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 55–65.

Susan M. Gass and Larry Selinker. 2008. *Second Language Acquisition: An Introductory Course*. Routledge, third edition.

Péter Halácsy, András Kornai, Csaba Oravecz, Viktor Trón, and Dániel Varga. 2006. Using a morphological analyzer in high precision POS tagging of Hungarian. In *Proceedings of LREC*, pages 2245–2248.

John A. Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1):1–23.

Trude Heift. 2007. Learner personas in call. *CALICO Journal*, 25(1):1–10.

Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Ross Israel, Markus Dickinson, and Sun-Hee Lee. 2013. Detecting and correcting learner korean particle omission errors. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 1419–1427. Nagoya, Japan.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Ph.D. thesis, University of Helsinki, Helsinki.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, second edition.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Beáta Megyesi. 1999. Improving Brill's POS tagger for an agglutinative language. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 275–284.

Wolfgang Menzel. 2006. Detecting mistakes or finding misconceptions? diagnosing morpho-syntactic errors in language learning. In Galia Angelova, Kiril Simov, and Milena Slavcheva, editors, *Readings in Multilinguality*, pages 71–77. Incoma Ltd., Shoumen, Bulgaria.

Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Özlem Çetinoğlu and Jonas Kuhn. 2013. Towards joint morphological analysis and dependency parsing of turkish. In *Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing 2013)*, pages 23–32.

Gabor Prószéky and Balazs Kis. 1999. A unification-based approach to morpho-syntactic parsing agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 261–268.

Marwa Ragheb. 2014. *Building a Syntactically-Annotated Corpus of Learner English*. Ph.D. thesis, Indiana University, Bloomington, IN.

Marwa Ragheb and Markus Dickinson. 2014. Developing a corpus of syntactically-annotated learner language for English. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13), Poster Session*. Tübingen, Germany.

Veit Reuer. 2003. Error recognition and feedback with lexical functional grammar. *CALICO Journal*, 20(3):497–512.

Thomas Schmidt. 2010. Linguistic tool development between community practices and technology standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards*. Malta.

Camilla B. Schwind. 1995. Error analysis and explanation in knowledge based language tutoring. *Computer Assisted Language Learning*, 8(4):295–324.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. In Robert Borsley and Kersti Börjars, editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.

Joel Tetreault and Martin Chodorow. 2008. The ups and downs of prepositions error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872.

Sylvie Thouësny and Françoise Blin. 2011. Modeling language learners' knowledge: What information can be inferred from learners' free written texts? In M. Levy, F. Blin, C. Bradin Siskin, and O. Takeuchi, editors, *WorldCALL: International Perspectives on Computer-Assisted Language Learning*, pages 114–127. Routledge, New York.

Miklós Törkenczy. 2008. *Hungarian Verbs and Essentials of Grammar, 2nd ed.* McGraw-Hill, New York.

Viktor Tron, Gyögy Gyepesi, Péter Halácsky, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: Open source word analysis. In *Proceedings of the Workshop on Software*, pages 77–85. Association for Computational Linguistics.

Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1670–1673.

Sowmya Vajjala and Kaidi Loo. 2013. Role of morpho-syntactic features in estonian proficiency classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Veronika Vincze, János Zsibrita, Péter Durst, and Martina Katalin Szabó. 2014. Automatic error detection concerning the definite and indefinite conjugation in the hunlearner corpus. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*.

Nina Vyatkina. 2013. Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal*, 97:11–30.

Helen Yannakoudakis, Ted Briscoe, and Theodora Alexopoulou. 2012. Automating second language acquisition research: Integrating information visualisation and machine learning. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 35–43.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. Magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In *Proceedings of RANLP*, pages 763–771.