Annotating Errors in a Hungarian Learner Corpus

Markus Dickinson, Scott Ledbetter

Indiana University
{md7,saledbet}@indiana.edu

Abstract

We are developing and annotating a learner corpus of Hungarian, composed of student journals from three different proficiency levels written at Indiana University. Our annotation marks learner errors that are of different linguistic categories, including phonology, morphology, and syntax, but defining the annotation for an agglutinative language presents several issues. First, we must adapt an analysis that is centered on the morpheme rather than the word. Second, and more importantly, we see a need to distinguish errors from secondary corrections. We argue that although certain learner errors require a series of corrections to reach a target form, these secondary corrections, conditioned on those that come before, are our own adjustments that link the learner's productions to the target form and are not representative of the learner's internal grammar. In this paper, we report the annotation scheme and the principles that guide it, as well as examples illustrating its functionality and directions for expansion.

Keywords: Hungarian, learner language, error annotation

1. Introduction and Motivation

Learner corpora have become increasingly popular in the last twenty years, as illustrated by the 2011 conference on *Learner Corpus Research* at the Université catholique de Louvain. Such corpora, especially when annotated with errors, have been useful for studying various aspects of the interlanguage of language learners (e.g., Myles, 2005; Granger, 2008), for developing pedagogical materials and lexicographic resources (e.g., Nicholls, 2003), and for developing and evaluating error detection systems (e.g., Meurers et al., 2010; Granger, 2003). One limitation has been that most of the work has focused on Western European languages. Determining how to annotate learner corpora of more morphologically-rich languages has had very little work (though, see Lee et al., 2011; Hana et al., 2010). As morphologically-rich languages have different types of errors and error patterns, we focus in this paper on defining an error annotation scheme for corpora containing data from learners of Hungarian.

Our goal is to define the scheme such that, after applying it to a corpus, error detection systems can be developed and evaluated for learner Hungarian—specifically, systems capable of providing feedback. Part of that goal is thus to describe learner interlanguage (see, e.g., Ellis, 2008): by analyzing non-target forms and their features, we can begin to understand learner grammars and how they differ from those of native speakers. This is a crucial step toward providing individualized, relevant feedback to the language learner.

To achieve this, we must define an annotation scheme. This consists of: defining the units of analysis (i.e., morphemes), defining the annotation categories—some of which need to be tailored to Hungarian—and accounting for interactions between layers of linguistic analysis. Our proposal incorporates a distinction between error annotation and a secondary layer of annotation which does not represent errors, but maps to a target form. While we focus on the basic principles for annotating learner data in an agglutinative language, the distinction between errors and secondary adjustments is relevant for any language.

In section 2., we discuss the properties of Hungarian relevant for annotating learner data and some background on error annotation in learner corpora. Our data and general annotation framework are described in section 3. before turning to the bulk of the work, in section 4., where the annotation scheme is described, and heuristics are outlined for handling multiple possible analyses. In section 5., extended examples are examined, highlighting some of difficulties and solutions. The annotation scheme is relatively simple, clearly distinguishes errors from secondary emendations, and covers a range of phenomena in Hungarian.

2. Background

2.1. Hungarian

Hungarian belongs to the Finno-Ugric language family, known for its agglutinative morphology: words are formed by concatenating morphemes, resulting in rich inflectional and derivational systems, as in (1).

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

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

c. ház -ban
house -INESSIVE
'in (a) house'

In Hungarian, 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). There are 20 cases (e.g. inessive in (1b)), many of which roughly correspond to adpositions in other languages. For both verbs and nouns, specific ordering of suffixes must be observed (Törkenczy, 2008).

Another characteristic of Hungarian, and other Finno-Ugric languages, is vowel harmony. Stems select allomorphs based on distinctions between back and front vowels, and within front vowels between rounded and unrounded vowels. For example, the inessive case suffix is realized with a front vowel, *-ben*, in (1b) because the root *könyv* contains only a front vowel, whereas in (1c) it appears with a back vowel, *-ban*, because of the back vowel in the root *ház*.

2.2. Error-annotated learner corpora

We follow a line of work on annotating learner corpora with error information (e.g., Nicholls, 2003; Lüdeling et al., 2005; Rozovskaya and Roth, 2010). Some annotation schemes make use of multi-layered annotation, such as with the FALKO corpus of German (Lüdeling et al., 2005). This allows for multiple interpretations and makes it easy to annotate errors spanning more than one word. Multiple layers of annotation also allow one to treat error annotation as an incremental process, e.g., building from smaller constituents to larger ones (Boyd, 2010; Hana et al., 2010), a key point we build from in section 4. In particular, Hana et al. (2010) use a "two-stage annotation design, based on three levels." Level 0 is the transcribed input; Level 1 contains orthographic and morphological corrections, with only individual forms treated; and Level 2 handles all other corrections, including, e.g., changes in word order. Our scheme shares much in spirit with this approach.

Turning to agglutinative languages, Lee et al. (2011) annotate post-positional particle errors in Korean. While useful for developing particle error detection systems, they explicitly ignore other error types. The agglutination forces them to add a segmented layer of annotation before any other layers, a practice we follow.

3. Data and Annotation

The initial corpus has been collected from students of Hungarian at Indiana University from three different levels: beginning, intermediate, and advanced. The texts are journals, composed of entries on various topics (chosen by the student), and each one is ten to fifteen sentences in length. Currently, data from fourteen journals is included (9 beginning, 1 intermediate, and 4 advanced). More data is being collected, but this is sufficient to begin defining the annotation.

Annotation was carried out using EXMARaLDA (Extensible Markup Language for Discourse Annotation), a freely available tool (Schmidt, 2010).[1] EXMARaLDA allows for multiple simultaneous tiers of annotation; can be exported in XML format; and provides a corpus management tool and concordancer for ease of maintenance and analysis.

To define the annotation, the second author, an advanced student of Hungarian, marked the annotation, consulting native Hungarian speakers throughout the process to check the work. Preliminary data from each level (beginner, intermediate, and advanced) was consulted to define the first iterations of the scheme.

Each text was first transcribed and then segmented manually by morpheme in EXMARaLDA, keeping both versions in the annotation file. Next, the learner errors were marked in the segmented data according to the annotation scheme detailed below. After a number of files were annotated, the results were checked by a native speaker instructor of Hungarian for accuracy—a process which has been continually repeated. During annotation, we kept a record of decisions, which now comprise the annotation manual for the project.

4. Annotation scheme

4.1. Anotation scheme overview

Similar to Hana et al. (2010), the annotation scheme is a compromise between annotating every relevant property and focusing on what can be annotated reliably within a small project. We turn now to an overview.

First, as Hungarian is agglutinative, we recognize the need to segment the data into morphemes before annotating. To mark case errors and corrections, for example, we need the case marker to be an individual unit of analysis.

Secondly, there are different types of errors reflecting different levels of linguistic analysis. For instance, for (2), the annotation is as in Figure 1, marking 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.

(2) Ajanl -om bor -t , nem sör -t .
recommend 1SG.DEF 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) (some layers not shown)

We distinguish annotation *categories* from annotation *layers*, where layers reflect an ordering between corrections and categories are unordered within a layer. We define four basic error annotation categories, reflecting **character** (CHA), **morphological** (MOR), **grammatical relation** (REL), and **sentence** (SNT) errors. Conceptually, we treat them all as one layer; i.e., different categories of errors can be annotated for the same word, and error spans can overlap, if necessary.[2] The collection of error categories (see section 4.3.) and the target (TGT) sentence make up this **error layer**.

In principle, each category could be annotated with its own target form, allowing for a series of ordered, explicit corrections. This would capture the fact that, e.g., a spelling error could precede a word order error (cf. Hana et al., 2010).

[1] EXMARaLDA and related tools are available online at http://www.exmaralda.org/en_index.html

[2] Practically, each category has its own tier in EXMARaLDA.

However, as we are limited in annotators, we keep the annotation simple, annotating only one target form for all categories in one layer. This has a possible benefit: an annotator can mark different error types without having to specify the exact contribution of each error to the target form.

4.2. Secondary adjustments

While annotating each local error is useful, this is insufficient for deriving a final target form. In the annotation for (3) in Figure 2, for instance, after adding a correction for case (MSC) from the unmarked nominative to the accusative *-t*, the sentence is still not well-formed. The case marker triggers a phonological rule of Hungarian, and the final vowel of the stem now needs to be lengthened, resulting in *teá*. Without the case suffix, the word was originally well-formed, so it is difficult to refer to there being a second error. Rather, we refer simply to an *adjustment*, conditioned on a previous correction. These changes are made in the second layer of annotation, the **adjustment layer**.

(3) Szeret -ek kávé -t és tea .
love 1SG.INDEF coffee ACC and tea .
'I love coffee and tea.'

	TXT	Szeretek		kávét		és	tea		.
	SEG	Szeret	ek	kávé	t	és	tea		.
Layer 1	CHA								
	MOR								
	REL							MSC	
	SNT								
	TGT	Szeret	ek	kávé	t	és	tea	**t**	.
Layer 2	CHA						CL		
	MOR								
	REL								
	SNT								
	TGT	Szeret	ek	kávé	t	és	**teá**	t	.

Figure 2: Two-layer error annotation for (3)

There are two things to note about these secondary changes. First, they contain the same categories as the primary error layer. Secondly, and crucially, these changes are not the same as errors; they are secondary emendations which only emerge consequent to correcting the specific errors learners make. For learning, knowing both what is erroneous (first layer) and what is a correct sentence (second layer) is useful. The distinction is important so that the learner is not penalized in the annotation for these corrections. While errors may be evidence of a systematic departure from the target language, note that adjustments make no assumptions about the learner's grammar.

An extended example of the annotation is given in Figure 3, for (4). We note a few points here. First, the verb suffix *-nak* (3rd person plural) is wrong, as it should be 1st person plural, *-unk*. Changing the suffix again leads to changing the stem form, as is reflected in the MSR (root selection) adjustment from *van* to *vagy*. However, there is also a word order error, requiring the tokens to move earlier in the sentence. We treat word order errors as combinations of insertions and deletions, linking them with numbers. Note that we are able to treat the shift for *van* and *-unk* as the same error (SS) by combining the cells for SNT. To some extent, information is lost here: the MAP error is where *-nak* becomes *-unk*, but this is not directly indicated, only after the ordering change. This is the cost of annotating only a single target form for all error categories; however, the MAP category at least makes clear the nature of the issue.

(4) de mi nem barat -om -ok van -nak .
but we not friend 1SG PLURAL be 3PL.INDEF .
'But we aren't friends'

4.3. Annotation scheme in detail

The current annotation scheme with error codes is given in Table 1, where subcategories of the four main categories are distinguished. While a few error categories are somewhat Hungarian-specific (e.g., vowel harmony), most reflect linguistic properties found across languages.

Category	**Character (CHA)**	
Subcategory	Phonology	Spelling
Type	Vowel length (CL) Vowel harmony (CV) Phon. confusion (CP) Doubling (CD)	Compounding(CC) Typo (CT)

Category	**Morpheme (MOR)**	
Subcategory	Agreement	Derivation
Type	Person (MAP) Number (MAN) Case (MAC) Definiteness (MAD)	Omission (MDO) Insertion (MDI) Ordering (MDS)

Category	**Relation (REL)**
Type	Case (MSC) Root (MSR) Copula (COP) Generic (MS)

Category	**Sentence (SNT)**
Type	Omission (SO) Insertion (SI) Ordering (SS)

Table 1: The annotation scheme

We make a distinction between errors in four categories: character (CHA), morpheme (MOR), relation (REL), and sentence (SNT). The categories in that order follow a bottom-up analysis of language, but each one is independent in the annotation.

Errors at the character level are divided into two subcategories, spelling and phonology. Spelling covers aspects of orthography that do not generally cause a change in meaning: compounding errors and typos. Phonology, meanwhile, covers errors that are important for distinctions in meaning in Hungarian. These include vowel length, consonant doubling, vowel harmony, and phonemic confusions. Though all of these can be considered spelling errors given the medium of the data (especially vowel length and con-

	TXT	de	mi	nem			baratomok			vannak		.
	SEG	de	mi	nem			barat	om	ok	van	nak	.
Layer 1	CHA						CL					
	MOR							MDI			MAP	
	REL											
	SNT				SS-1					SS-1		
	TGT	de	mi	nem	van	unk	barát		ok			.
Layer 2	CHA											
	MOR											
	REL				MSR							
	SNT											
	TGT	de	mi	nem	vagy	unk	barát		ok			.

Figure 3: An extended example of error annotation, for (4)

sonant doubling), we draw a distinction because they are contrastive features in the language.

The morpheme level concerns agreement and derivation (i.e. inflectional and derivational morphology). Agreement errors cover the inflection of nouns and verbs, as seen in (1). Verbs are analyzed for person, number, and definiteness, while nouns are analyzed for person, number, and case. Derivation errors concern the omission, insertion, and ordering of morphemes.

Relation errors concern the interaction of different tokens and largely represent errors in selection. This category includes case (as determined by the verb), the copula (the use of which depends on context and person), and choice of morpheme (root or affix) when the choice depends on other elements in the sentence.

Finally, the sentence category represents errors which go beyond selection and may interact with semantics or pragmatics. Just as with morphemes, words can be inserted, omitted, or ordered in non-target-like ways. Because word order is flexible in Hungarian and dependent on specific contexts, only fairly strict ordering rules are observed in the analysis. In other words, we follow the Principle of Minimal Interaction (Hana et al., 2010): if a construction can be grammatical, it is kept as correct, even if a better version exists (see also section 4.4.).

As noted above, these four categories are used for both the error layer and the adjustment annotation layer.

4.4. Multiple analyses

As there can be multiple possible ways of interpreting an error and positing a correction (cf., e.g., Lüdeling et al., 2005), we employ a few heuristics to narrow the scope of possible annotations.

First, we use context as much as possible. Given that we are annotating journals, the meaning of a sentence is often straightforward and derivable from the overall narrative. Secondly, we try to give the learner the benefit of the doubt and posit as few errors as possible (Dickinson and Ragheb, 2009). If two analyses seem equally likely, based upon the context, we posit the one which leads to fewer corrections. Thirdly, we bias towards more informative annotation over less informative, in the interest of making searching for linguistic properties easier. For example, if a learner has an error which could be posited as a vowel harmony error or a spelling error, we annotate it as a vowel harmony error, all other things being equal. In this way, someone searching for vowel harmony can find it (and call it a spelling error for themselves, if they wish). Such a property is less easily findable if annotated with the less informative category.

Finally, if truly necessary, we can employ the original intention of the multi-layered annotation and annotate multiple possible interpretations. Although our effort is still young, we have yet to need this option.

We will see how these heuristics play out in some of the examples in the next section.

5. Annotation examples

We now present several specific cases that informed our decisions and show the extensibility of the annotation scheme. The first two cases deal with label definitions, while the last three deal with handling multiple analyses.

Looking at the annotation for sentence (5) in Figure 4, the single error MDO (derivation: omission) reflects an incomplete possessive structure. The second noun *oldal* should be suffixed with *a*, as shown in the first layer target form. This triggers two phonological rules in the adjustment layer—the selection of the allomorph *n* due to the suffix vowel and a lengthening of the *a*—the former of which was not accounted for in the original scheme.

(5) a hegy oldal -on
the mountain side SUPERESSIVE
'on the mountain side'

TXT	a	hegy	oldalon		
SEG	a	hegy	oldal		on
MOR				MDO	
TGT	a	hegy	oldal	a	on
CHA				CL	**CI**
TGT	a	hegy	oldal	á	n

Figure 4: Annotation for (5), showing a new label, CI

One solution is to use an additional error code, CI (character: insertion). More generally, if we find a production that cannot be annotated with the current scheme and requires a new code, there is almost no change to the annotation workflow. The code simply needs to be added to the XML definitions of the error codes, and from that point on, it is available for use in annotation. The alternative is to redefine the set of error codes we already have, and such decisions have to be done on a case-by-case basis, balancing informativeness with the scheme's compactness.

For (6), annotated in Figure 5, the learner produces a suffix with a front rounded vowel (FRD), matching the quality of the stem's vowel. Though the features match, the correct form is actually an irregular allomorph (with a front vowel).

(6) könyv -ök
book.FRD PL.FRD
'books'

TXT	könyvök	
SEG	könyv	ök
CHA		**CV**
TGT		ek

Figure 5: Annotation for (6), showing the CV label being used for an irregular (non-harmonic) ending

While CV (vowel harmony) is the correct label, the scheme does not recognize that the learner correctly followed a phonological rule. We would need to include features linked to the error in order to capture the fact that the learner selected a correctly-harmonized, though incorrect, variant.

Turning to cases where multiple analyses may be applicable, in (7) there is an ambiguous error concerning either vowel length or morphological case. For the meaning of 'I went home', the long *á* should be short, written without an accent. However, the accent would be correct if the learner meant 'I went [(in)to] his/her house'. To achieve the second meaning, the learner would have to select the inessive or allative case for *ház* rather than the unmarked nominative that was used. This would be indicated by an MSC error—and also an additional article before *ház*.

(7) Ment -em ház -a
Go[PST] 1SG house 3SG
'I went home'

TXT	Mentem		háza	
SEG	Ment	em	ház	a
CHA			CL	
TGT	Ment	em	haza	

Figure 6: Annotation for (7), annotating the simpler vowel length (CL) error over a case error

As can be seen in in Figure 6, we opt to annotate a vowel length error, in accordance with our heuristics. This is the simplest correction and results in the fewest changes to the learner's production, i.e., the fewest morphemes to change. It also more appropriately matches the context in the journal before this sentence.

In (8), the verb *szeret* is in the base form, which is identical to the third person singular form, but the context requires a first person suffix.

(8) én szeret kávé
I love.3SG coffee
'I love coffee'

TXT	én	szeret		kávé	
SEG	én	szeret		kávé	
MOR			**MAP**		
REL					MSC
TGT	én	szeret	ek	kávé	t

Figure 7: Annotation for (8), where an informative tag (MAP) was chosen over a less informative one (MDO)

The corresponding error MAP (agreement: person) in Figure 7 could also be MDO (derivation: omission) to indicate a lack of inflection. In line with our heuristic of being more informative when context cannot disambigaute, we choose the more informative tag MAP. Users of the corpus can thus search for and evaluate person agreement errors.

The issue of the amount of information present in the annotation can become complicated. In (9), the learner has supplied two cases: accusative and inessive (where superessive would be correct). At most one case is permitted on any noun in Hungarian.

(9) ez szép nyar -at -ban
this beautiful summer ACCUSATIVE INESSIVE .
'this is beautiful in summer'

TXT	ez	szép	nyaratban		
SEG	ez	szép	nyar	at	ban
CHA			CL		
MOR				**MDI**	
REL					MSC
TGT	ez	szép	nyár		on

Figure 8: Annotation for (9), illustrating the annotation for erroneous use of two case endings

In the annotation scheme, a case error (MSC) reflects the selection of an inappropriate case, as shown with the annotation for the second case used in Figure 8, as the inessive is corrected to the superessive. We seem to have at least three options here: 1) treat *at* and *ban* as a unit on the REL and TGT tiers, as a single MSC error; 2) treat them as two separate MSC errors, correcting one to *on* and the other to nothing; or 3) treat one (*ban*) as an MSC error and the other (*at*) as an insertion (MDI).[3]

[3]We treat *ban* as the MSC error since the functions of inessive and superessive are more similar than either to nominative.

The first option is appealing, but note that it conflates what really are two errors: multiple cases, as well as the wrong case (i.e., an insertion and a substitution). The second option is informative, but also redundant, as both errors are treated equally (MSC), even though the errors are of a different nature. The third option is the one we pursue, as it treats the different errors uniquely. While the MDI (derivation: insertion) label seems to result in a loss of information, by not marking this as a case error, it importantly shows that the error is not a problem with selection—as the relation (REL) errors imply—but simply in the constraints on case usage in Hungarian. In the future, new error codes or features attributed to specific errors could capture more fine-grained information for further analysis.

6. Summary and Outlook

Our annotation scheme allows for the description of an agglutinative and morphologically-rich language. We take morphemes as the base unit of analysis in order to capture the errors learners make at all levels of linguistic structure and provide error codes for each linguistic category. We also distinguish errors from adjustments, which allows for ordered corrections and avoids penalizing the learner for additional changes conditioned on error corrections.

The annotation scheme is rather coarse-grained, but is feasible for the scope of the project. We have a single annotator, and the scheme is extensible to account for new phenomena with minimal changes to the annotation workflow.

In the future, the most important step is to continue collecting and annotating more data. As far as modifications to the annotation scheme, we hope to add annotation linking errors to a "conditioning" element, e.g., linking a case error to the verb which licenses that case (cf., e.g., Hana et al., 2010). In EXMARaLDA, this can be accomplished with references to the IDs corresponding to each morpheme in the segmentation tier. We also plan to add features to the final annotation, in order to better analyze a learner's interlanguage. By comparing the features of learner productions to those of target language examples, we will be able to identify aspects of the interlanguage to target with specific feedback.

Acknowledgments

We would like to thank Valéria Varga for her help in checking annotations, the IU CL discussion group for feedback, and the anonymous reviewers for their comments.

References

Adriane Boyd. 2010. EAGLE: an error-annotated corpus of beginning learner German. In *Proceedings of LREC-10*. Malta.

Markus Dickinson and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70. Milan, Italy.

Rod Ellis. 2008. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, second edition.

Sylviane Granger. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3):465–480.

Sylviane Granger. 2008. Learner corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An international handout*, volume 1, pages 259–275. Mouton de Gruyter.

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. Error-tagged learner corpus of Czech. In *Proceedings of LAW-10*, pages 11–19. Uppsala, Sweden.

Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2011. Challenges in annotating korean particle errors. Talk given at Learner Corpus Resarch 2011. Louvain-la-Neuve, Belgium. September 16, 2011.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*. Birmingham.

Detmar Meurers, Niels Ott, and Ramon Ziai. 2010. Compiling a task-based corpus for the analysis of learner language in context. In *Proceedings of Linguistic Evidence 2010*, pages 214–217. Tübingen.

Florence Myles. 2005. Review article: Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4):373–391.

Diane Nicholls. 2003. The cambridge learner corpus - error coding and analysis for lexicography and ELT. In *Proceedings of Corpus Linguistics 2003*, pages 572–581. Lancaster University.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of NLP-BEA*, pages 28–36. Los Angeles.

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.

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