Using an Annotated L2 Hungarian Corpus to Study Vowel Harmony Development

Markus Dickinson & Scott Ledbetter
LCR 2013, Bergen, Norway

29 September 2013

Introduction

The situation:

- Learner corpora have been useful for studying various aspects of the interlanguage of second language learners (L2ers)
- However, much of the work in learner corpora has been focused on Western European languages

Introduction

The situation:

- Learner corpora have been useful for studying various aspects of the interlanguage of second language learners (L2ers)
- However, much of the work in learner corpora has been focused on Western European languages

We present an error-annotated corpus of learner Hungarian for research in second language acquisition (SLA)

- We use the corpus & annotation to start an investigation of vowel harmony

Motivation and Goals

- We aim to test the utility of the annotation scheme with an analysis of learner interlanguage (IL), focusing on a single phenomenon in the L2: **Vowel harmony**

Motivation and Goals

- We aim to test the utility of the annotation scheme with an analysis of learner interlanguage (IL), focusing on a single phenomenon in the L2: **Vowel harmony**
- Guided by these questions:
  - **Theoretically:** What is the process underlying the acquisition of vowel harmony in Hungarian?
  - **Methodologically:** How do the corpus & annotation help us address the question of vowel harmony acquisition?

Motivation and Goals

- We aim to test the utility of the annotation scheme with an analysis of learner interlanguage (IL), focusing on a single phenomenon in the L2: **Vowel harmony**
- Guided by these questions:
  - **Theoretically:** What is the process underlying the acquisition of vowel harmony in Hungarian?
  - **Methodologically:** How do the corpus & annotation help us address the question of vowel harmony acquisition?
- The work presented today is preliminary & based on a small sample, but shows promise for answering these questions
  - Our hope: the corpus & annotation design can be extended as needed and used to study other phenomena & languages

Outline

Background

Hungarian

- Hungarian possesses rich inflectional & derivational morphology
- It also exhibits an extensive case system (20 cases)

Background

Hungarian

- Hungarian possesses rich inflectional & derivational morphology
- It also exhibits an extensive case system (20 cases)
- Most morphemes alternate according to vowel harmony, e.g. the inessive in (1)

(1) a. ház **-ban**
house -INESSIVE[bk]
'in (a) house'

b. könyv **-ben**
book -INESSIVE[fr]
'in (a) book'

Background

Vowel Harmony

- Vowel harmony determines the selection of allomorphs based on assimilation of features between stem and affix (Hayes et al., 2009)
- Vowels are characterized by frontness or backness in the vowel space and, in the case of front vowels, also roundedness:

Front		Back	
Rounded	Unrounded	Rounded	Unrounded
ü /y/	i /i/	u /u/	
ű /y:/	í /i:/	ú /u:/	
ő /ø:/	é /e:/	ó /o:/	
ö /ø/	e /ɛ/	o /o/	
		a /ɒ/	á /a:/

Background

Vowel Harmony

- The general rule is: stems with only back vowels select a suffix with back vowels (2a,3a) and stems with front vowels select a suffix with front vowels (2b, 3b)

(2)
a. ház -hoz
house -ALL[bk]
'toward a house'

b. szék -hez
chair -ALL[fr]
'toward a chair'

(3)
a. ház -ban
house -INESS[bk]
'in a house'

b. szék -ben
chair -INESS[fr]
'in a chair'

Background

Vowel Harmony

- The general rule is: stems with only back vowels select a suffix with back vowels (2a,3a) and stems with front vowels select a suffix with front vowels (2b, 3b)
- Within front vowels, there is a further distinction of rounded and unrounded, though with many cases, this is neutralized (compare allative (2c) and inessive (3c) case)

(2)
a. ház -hoz
house -ALL[bk]
'toward a house'

b. szék -hez
chair -ALL[fr]
'toward a chair'

c. könyv -höz
book -ALL[fr.rd]
'toward a book'

(3)
a. ház -ban
house -INESS[bk]
'in a house'

b. szék -ben
chair -INESS[fr]
'in a chair'

c. könyv -ben
book -INESS[fr]
'in a book'

Vowel harmony

We picked vowel harmony to study partly because its simplicity makes for a nice initial foray into using the annotation

Vowel harmony

We picked vowel harmony to study partly because its simplicity makes for a nice initial foray into using the annotation

- Vowel harmony is easily studied with our annotation, as we mark up the data by morpheme rather than by word

TXT	lengyelul	
SEG	lengyel	ul
CHA		**CV**
TGT		ül

- In the text (TXT) tier, the word *lengyelul* 'in Polish' is segmented (SEG) into stem and affix

Vowel harmony

We picked vowel harmony to study partly because its simplicity makes for a nice initial foray into using the annotation

- Vowel harmony is easily studied with our annotation, as we mark up the data by morpheme rather than by word

TXT	lengyelul	
SEG	lengyel	ul
CHA		**CV**
TGT		ül

- In the text (TXT) tier, the word *lengyelul* 'in Polish' is segmented (SEG) into stem and affix
- Then the error is marked in the character (CHA) tier as CV (vowel harmony) and associated with a target (TGT) form
- Error codes are easily searchable within the corpus for closer inspection

Why vowel harmony?

- For the learners in our corpus (L1 English), the phenomenon is a new one and they can be expected to need time to acquire this new type of system

Why vowel harmony?

- For the learners in our corpus (L1 English), the phenomenon is a new one and they can be expected to need time to acquire this new type of system
- While vowel harmony is usually straightforward, numerous exceptions can make selection unpredictable and thus present difficulties for learning (e.g. stem changes, homophony)
  - Compare *Ír* -**ek** 'the Irish' and *Ír* -**ok** 'I write'

Why vowel harmony?

- For the learners in our corpus (L1 English), the phenomenon is a new one and they can be expected to need time to acquire this new type of system
- While vowel harmony is usually straightforward, numerous exceptions can make selection unpredictable and thus present difficulties for learning (e.g. stem changes, homophony)
  - Compare *Ír* -**ek** 'the Irish' and *Ír* -**ok** 'I write'
- Our analysis can shed light on the troubles learners have and possibly inform instructors and researchers as to the most likely problem areas for targeted instruction.

Background

Error Annotation

Long line of work on error annotation in learner corpora

- Suri and McCoy (1993); Granger (2003); Nicholls (2003); Lüdeling et al. (2005); Boyd (2010); Hana et al. (2010); Rozovskaya and Roth (2010); ...

Background

Error Annotation

Long line of work on error annotation in learner corpora

- Suri and McCoy (1993); Granger (2003); Nicholls (2003); Lüdeling et al. (2005); Boyd (2010); Hana et al. (2010); Rozovskaya and Roth (2010); ...

Multi-layered annotation (cf. Lüdeling et al., 2005):

- Allows for multiple interpretations
- Allows for errors spanning more than one word
- Allows error annotation to be an incremental process (Boyd, 2010; Hana et al., 2010)

Data and Annotation

The corpus was collected from students of Hungarian at IU across three levels: beginning, intermediate, & advanced

Data and Annotation

The corpus was collected from students of Hungarian at IU across three levels: beginning, intermediate, & advanced

- The texts are journals, composed of entries on various topics chosen by the students, each 10–15 sentences in length
- There are currently 14 journals in the corpus (9 beginning, 1 intermediate, and 4 advanced): approx. 2400 sentences

Data and Annotation

The corpus was collected from students of Hungarian at IU across three levels: beginning, intermediate, & advanced

- The texts are journals, composed of entries on various topics chosen by the students, each 10–15 sentences in length
- There are currently 14 journals in the corpus (9 beginning, 1 intermediate, and 4 advanced): approx. 2400 sentences
- We transcribe each journal and annotate errors with EXMARaLDA in stages:
  - Segment the text on morpheme boundaries
  - Identify errors in each of our four tiers
  - Adjust productions to match a target form

Data and Annotation

The corpus was collected from students of Hungarian at IU across three levels: beginning, intermediate, & advanced

- The texts are journals, composed of entries on various topics chosen by the students, each 10–15 sentences in length
- There are currently 14 journals in the corpus (9 beginning, 1 intermediate, and 4 advanced): approx. 2400 sentences
- We transcribe each journal and annotate errors with EXMARaLDA in stages:
  - Segment the text on morpheme boundaries
  - Identify errors in each of our four tiers
  - Adjust productions to match a target form
- We analyzed data to assign *features* to individual morphemes (e.g., back vowel stem)

Data and Annotation

The corpus was collected from students of Hungarian at IU across three levels: beginning, intermediate, & advanced

- The texts are journals, composed of entries on various topics chosen by the students, each 10–15 sentences in length
- There are currently 14 journals in the corpus (9 beginning, 1 intermediate, and 4 advanced): approx. 2400 sentences
- We transcribe each journal and annotate errors with EXMARaLDA in stages:
  - Segment the text on morpheme boundaries
  - Identify errors in each of our four tiers
  - Adjust productions to match a target form
- We analyzed data to assign *features* to individual morphemes (e.g., back vowel stem)
  - Need to develop automatic analysis for this step, integrated with segmentation & error analysis (in-progress)

Error Annotation Scheme: General Approach

Dickinson and Ledbetter (2012)

- We take the morpheme as the basic unit of analysis, though errors can span multiple morphemes

Error Annotation Scheme: General Approach

Dickinson and Ledbetter (2012)

- We take the morpheme as the basic unit of analysis, though errors can span multiple morphemes
- A single morpheme can reflect different types of errors from different levels of linguistic analysis
  - CHA: Characters or phonemes (e.g., vowel harmony)
  - MOR: Morphemes (e.g., agreement in person)
  - REL: Relations between morphemes (e.g., case)
  - SNT: The sentence (e.g., ordering)

Error Annotation Scheme: General Approach

Dickinson and Ledbetter (2012)

- We take the morpheme as the basic unit of analysis, though errors can span multiple morphemes
- A single morpheme can reflect different types of errors from different levels of linguistic analysis
  - CHA: Characters or phonemes (e.g., vowel harmony)
  - MOR: Morphemes (e.g., agreement in person)
  - REL: Relations between morphemes (e.g., case)
  - SNT: The sentence (e.g., ordering)
- We maintain a distinction between *errors* and *adjustments*
  - **Errors:** elements that differ from target language usage
  - **Adjustments:** secondary changes to derive a target form

Error Annotation Scheme: General Approach

Dickinson and Ledbetter (2012)

- We take the morpheme as the basic unit of analysis, though errors can span multiple morphemes
- A single morpheme can reflect different types of errors from different levels of linguistic analysis
  - CHA: Characters or phonemes (e.g., vowel harmony)
  - MOR: Morphemes (e.g., agreement in person)
  - REL: Relations between morphemes (e.g., case)
  - SNT: The sentence (e.g., ordering)
- We maintain a distinction between *errors* and *adjustments*
  - **Errors:** elements that differ from target language usage
  - **Adjustments:** secondary changes to derive a target form
- While errors may be evidence of a systematic departure from the target language, note that adjustments make no assumptions about the learner's grammar

Example Annotation

(4) 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		.
Error	CHA								
	MOR								
	REL							MSC	
	SNT								
	TGT	Szeret	ek	kávé	t	és	tea	**t**	.
Adjust.	CHA						CL		
	MOR								
	REL								
	SNT								
	TGT	Szeret	ek	kávé	t	és	**teá**	t	.

Example Annotation

Vowel Harmony

(5) ő magyar és ő nem beszél német -ul
3SG Hungarian and 3SG NEG speak German[fr] ADV[bk]
'she is Hungarian and she doesn't speak German'

TXT	ő	magyar	és	ő	nem	beszél	németul	
SEG	ő	magyar	és	ő	nem	beszél	német	ul
CHA								**CV**
TGT								ül

- The learner has produced the back vowel allomorph of the adverbial suffix *-ul* when the stem *német* contains only front vowels, and this is notated with the error code CV

Example Annotation

Vowel Harmony

- In addition to error annotation, we posit features for individual root morphemes based on the affixes they combine with

Example Annotation

Vowel Harmony

- In addition to error annotation, we posit features for individual root morphemes based on the affixes they combine with

  (5) ő magyar és ő nem beszél német -ul
  3SG Hungarian and 3SG NEG speak German[fr] ADV[bk]
  'she is Hungarian and she doesn't speak German'

- In (5) features are attributed to *német* in the learner's lexicon, reflecting its combination with a back vowel suffix

  (6) német {vh: bk}

Initial Analysis

Method

- In our case study of the utility of the annotation, we track the development of two morphemes for several learners
  - We take the inessive case ending (*-ban/-ben*) and the adverbial derivational suffix used with language names (*-ul/-ül*)

Initial Analysis

Method

- In our case study of the utility of the annotation, we track the development of two morphemes for several learners
  - We take the inessive case ending (*-ban/-ben*) and the adverbial derivational suffix used with language names (*-ul/-ül*)
- These are among the first and most frequent harmonizing morphemes the learners are expected to encounter
- For both morphemes, the distinction is made between front and back vowels but not between rounded and unrounded

Initial Analysis

Method

- We consider many aspects of production, including accuracy, consistency in allomorph distribution, & innovation

Initial Analysis

Method

- We consider many aspects of production, including accuracy, consistency in allomorph distribution, & innovation
- Though the annotation highlights errors, we consider all instances of a given morpheme

Initial Analysis

Method

- We consider many aspects of production, including accuracy, consistency in allomorph distribution, & innovation
- Though the annotation highlights errors, we consider all instances of a given morpheme
  - More complete picture of the underlying morphemes in learner's IL (what they do right + what they do wrong)

Initial Analysis

Method

- We consider many aspects of production, including accuracy, consistency in allomorph distribution, & innovation
- Though the annotation highlights errors, we consider all instances of a given morpheme
  - More complete picture of the underlying morphemes in learner's IL (what they do right + what they do wrong)
  - Segmentation makes this step relatively easy

Initial Analysis

Accuracy

- We can measure accuracy of allomorph selection using frequency of the CV error code among total occurrences of the inessive case suffix and adverbial derivational suffix

Initial Analysis

Accuracy

- We can measure accuracy of allomorph selection using frequency of the CV error code among total occurrences of the inessive case suffix and adverbial derivational suffix

	Inessive			Adverbial		
Learner	Errors	Total	Accuracy	Errors	Total	Accuracy
Beg01	3	140	0.979	2	52	0.962
Beg02	8	118	0.932	3	36	0.917
Beg03	11	92	**0.880**	0	13	1.000
Beg04	1	36	0.972	0	11	1.000
Int01	0	85	1.000	0	31	1.000
Adv03	0	109	1.000	1	17	0.941

Initial Analysis

Accuracy

- We can measure accuracy of allomorph selection using frequency of the CV error code among total occurrences of the inessive case suffix and adverbial derivational suffix

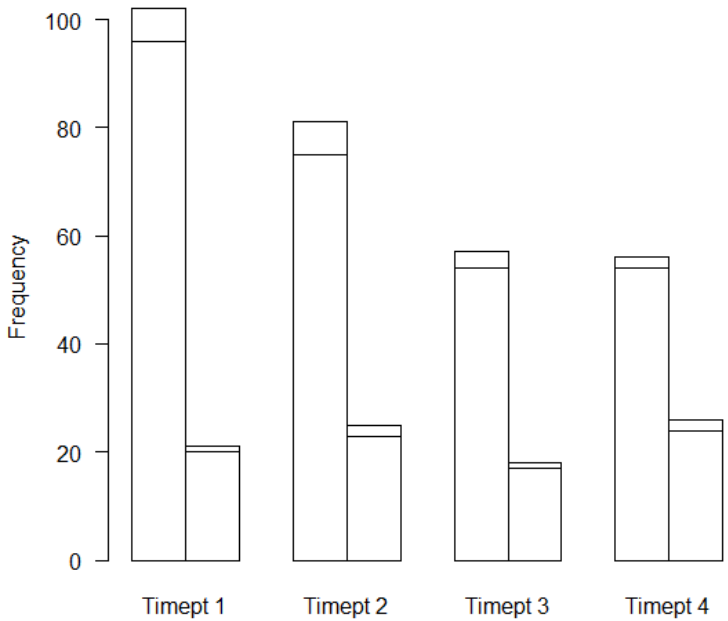
	Inessive			Adverbial		
Learner	Errors	Total	Accuracy	Errors	Total	Accuracy
Beg01	3	140	0.979	2	52	0.962
Beg02	8	118	0.932	3	36	0.917
Beg03	11	92	**0.880**	0	13	1.000
Beg04	1	36	0.972	0	11	1.000
Int01	0	85	1.000	0	31	1.000
Adv03	0	109	1.000	1	17	0.941

$\Rightarrow$ Though the phenomenon is not present in the L1, learners are fairly accurate

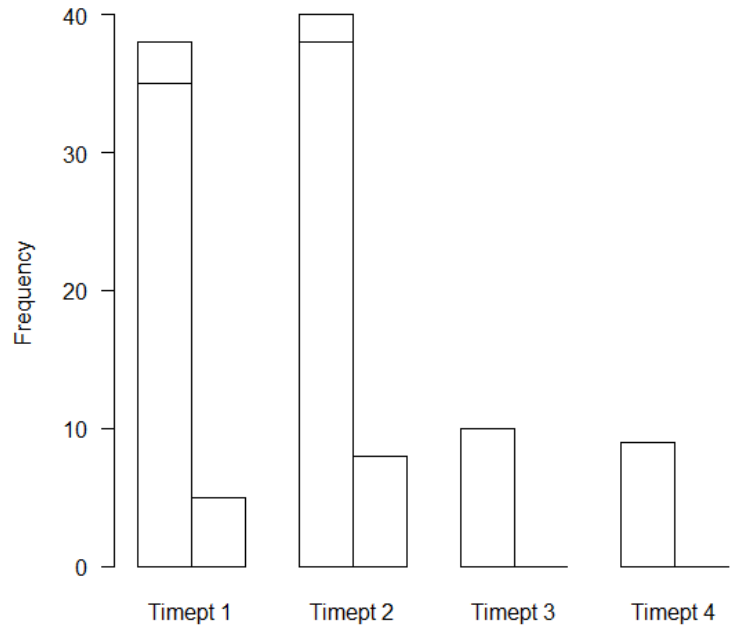
Initial Analysis

Consistency (usage over time)

Inessive morphemes



Adverbial morphemes

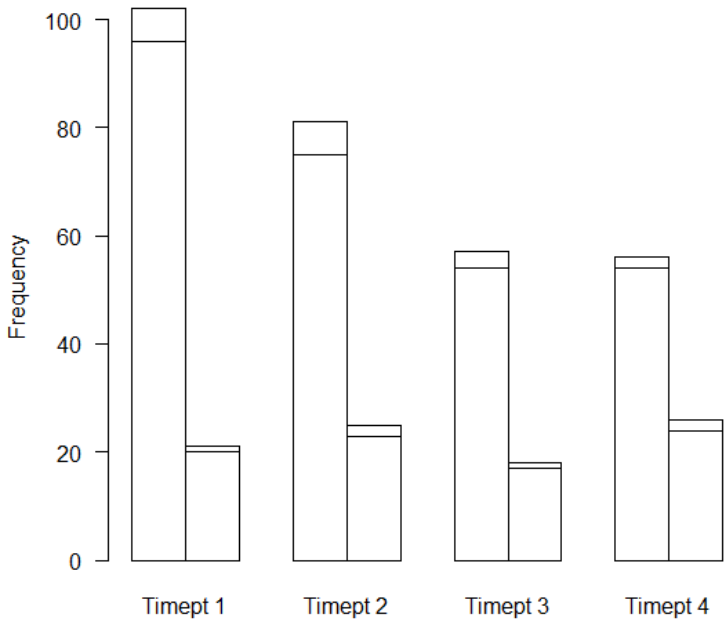


- For each pair of bars: back vowel on left, front vowel on right
  - whole bar = frequency of occurrence
  - top of bar = errors within occurrence

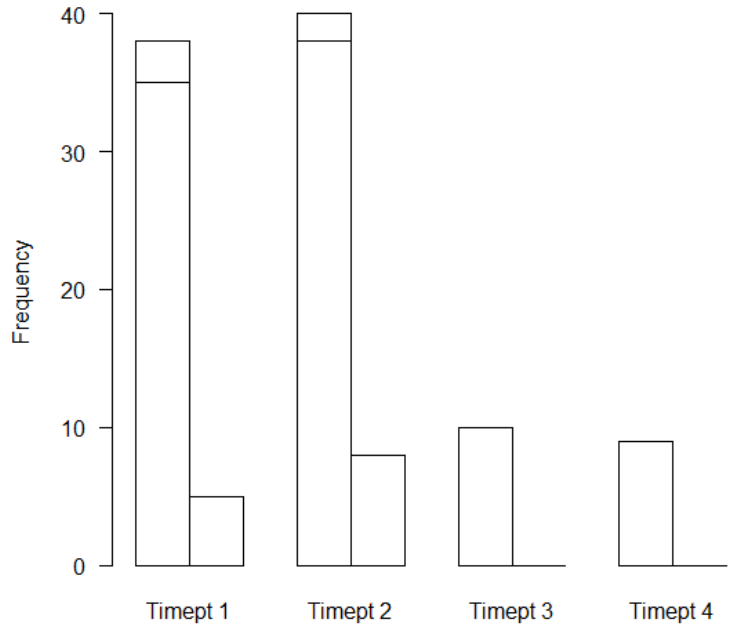
Initial Analysis

Consistency (usage over time)

Inessive morphemes



Adverbial morphemes



- For each pair of bars: back vowel on left, front vowel on right
  - whole bar = frequency of occurrence
  - top of bar = errors within occurrence

$\Rightarrow$ Usage decreases while precision (slightly) increases

Initial Analysis

Innovation

- In other cases, searching for other types of errors (e.g. character insertion, CI, or phonological confusion, CP) can be helpful for pinpointing instances of vowel harmony application

(7) a. él -esz
live[fr] -2SG.INDEF[fr]
'you live' (cf. él -**sz**)

b. büd -üs
stink[fr.rd] -ADJ[fr.rd]
'stinky' (cf. büd -**ö**s)

Initial Analysis

Innovation

- In other cases, searching for other types of errors (e.g. character insertion, CI, or phonological confusion, CP) can be helpful for pinpointing instances of vowel harmony application

(7) a. él -esz
live[fr] -2SG.INDEF[fr]
'you live' (cf. él -**sz**)

b. büd -üs
stink[fr.rd] -ADJ[fr.rd]
'stinky' (cf. büd -**ö**s)

- In (7a), the learner has an epenthetical vowel in the suffix, a correct match to the harmonizing features in the root verb

Initial Analysis

Innovation

- In other cases, searching for other types of errors (e.g. character insertion, CI, or phonological confusion, CP) can be helpful for pinpointing instances of vowel harmony application

(7) a. él -esz
live[fr] -2SG.INDEF[fr]
'you live' (cf. él -**sz**)

b. büd -üs
stink[fr.rd] -ADJ[fr.rd]
'stinky' (cf. büd -**ö**s)

- In (7a), the learner has an epenthetical vowel in the suffix, a correct match to the harmonizing features in the root verb
- In (7b), the adjectival suffix repeats the stem's high front rounded *ü* in place of the mid *ö*

Initial Analysis

Innovation

- In other cases, searching for other types of errors (e.g. character insertion, CI, or phonological confusion, CP) can be helpful for pinpointing instances of vowel harmony application

(7) a. él -esz
live[fr] -2SG.INDEF[fr]
'you live' (cf. él -**sz**)

b. büd -üs
stink[fr.rd] -ADJ[fr.rd]
'stinky' (cf. büd -**ö**s)

- In (7a), the learner has an epenthetical vowel in the suffix, a correct match to the harmonizing features in the root verb
- In (7b), the adjectival suffix repeats the stem's high front rounded *ü* in place of the mid *ö*

⇒ Important to study the interaction of errors/lingusitic properties

Summary and Outlook

Summary:

- The corpus and its annotation allow for an analysis of learner Hungarian at the level of individual morphemes
- Searchable error codes pinpoint specific instances of a given phenomenon, and additional features can be used to further investigate individual forms
- The longitudinal nature of the corpus gives insight into the development of the learner's grammar over time

Summary and Outlook

Summary:

- The corpus and its annotation allow for an analysis of learner Hungarian at the level of individual morphemes
- Searchable error codes pinpoint specific instances of a given phenomenon, and additional features can be used to further investigate individual forms
- The longitudinal nature of the corpus gives insight into the development of the learner's grammar over time

Outlook:

- Collect & annotate more data
- Expand the study to all learners in the data set
- Expand to developmental patterns for other features
- Finish developing automatic system for speeding up analysis

Köszönöm szépen!

References

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

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)*. Istanbul, Turkey.

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

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.

B. Hayes, P. Siptár, K. Zuraw, and Z. Londe. 2009. Natural and unnatural constraints in hungarian vowel harmony. *Language*, 85(4):822–863.

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

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.

Linda Z. Suri and Kathleen F. McCoy. 1993. A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners. Technical Report 93–16, Department of Computer and Information Sciences, University of Delaware, Newark, DE.