Automatic Morphological Analysis of Learner Hungarian

Scott Ledbetter & Markus Dickinson
Indiana University

4 June, 2015
Introduction and Motivation

Background

Methods

Evaluation & Preliminary Results

Summary & Outlook
Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:
Introduction

Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:

1. Largely focused on a few error types (e.g., prepositions)
Introduction

Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:

1. Largely focused on a few error types (e.g., prepositions)
2. Largely for English
Introduction

Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:

1. Largely focused on a few error types (e.g., prepositions)
2. Largely for English
3. Often focused on errors to the exclusion of broader patterns of learner productions,
Introduction

Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:

1. Largely focused on a few error types (e.g., prepositions)
2. Largely for English
3. Often focused on errors to the exclusion of broader patterns of learner productions, which are important for:
   - Intelligent computer-assisted language learning (ICALL) (e.g., Heift and Schulze, 2007)
   - Proficiency classification (e.g., Vajjala and Loo, 2013)
   - Second language acquisition research (e.g., Ragheb, 2014)
Recent research on grammatical error detection & correction (Leacock et al., 2014) has its limitations:

1. Largely focused on a few error types (e.g., prepositions)
2. Largely for English
3. Often focused on errors to the exclusion of broader patterns of learner productions, which are important for:
   - Intelligent computer-assisted language learning (ICALL) (e.g., Heift and Schulze, 2007)
   - Proficiency classification (e.g., Vajjala and Loo, 2013)
   - Second language acquisition research (e.g., Ragheb, 2014)

Our focus: Hungarian morphological analysis for learner language
Hungarian morphological analysis for learner language

We attempt to build a system that:

1. Works for a variety of morphological errors, providing detailed information about each one
   - e.g., supports learner modeling (Amaral and Meurers, 2008)
Hungarian morphological analysis for learner language

We attempt to build a system that:

1. Works for a variety of morphological errors, providing detailed information about each one
   - e.g., supports learner modeling (Amaral and Meurers, 2008)
2. Is feasible for low-resource languages
Hungarian morphological analysis for learner language

We attempt to build a system that:

1. Works for a variety of morphological errors, providing detailed information about each one
   - e.g., supports learner modeling (Amaral and Meurers, 2008)
2. Is feasible for low-resource languages
3. Provides analyses for correct and incorrect forms, i.e., is both a morphological analyzer & an error detector
Hungarian morphological analysis for learner language

We attempt to build a system that:

1. Works for a variety of morphological errors, providing detailed information about each one
   - e.g., supports learner modeling (Amaral and Meurers, 2008)
2. Is feasible for low-resource languages
3. Provides analyses for correct and incorrect forms, i.e., is both a morphological analyzer & an error detector

Best way to accomplish goals: build a rule-based system
Hungarian morphological analysis for learner language

We attempt to build a system that:

1. Works for a variety of morphological errors, providing detailed information about each one
   - e.g., supports learner modeling (Amaral and Meurers, 2008)

2. Is feasible for low-resource languages

3. Provides analyses for correct and incorrect forms, i.e., is both a morphological analyzer & an error detector

Best way to accomplish goals: build a rule-based system

- Hearkens back to the parsing ill-formed input literature (see Heift and Schulze, 2007, ch. 2)
- Illustrates that different linguistic properties need different kinds of systems (see Leacock et al., 2014, ch. 7)
Keeping it simple

The analyzer employs a simple chart-parsing strategy
  ▶ Allows for feature clashes
    ▶ Step towards determining which constraints (of a huge space of possible variations) may be relaxed
Keeping it simple

The analyzer employs a simple chart-parsing strategy

- Allows for feature clashes
  - Step towards determining which constraints (of a huge space of possible variations) may be relaxed
- Relies on a handful of handwritten affixes, which essentially encode the “rules” of the grammar
  - Step towards developing analyzers for lesser-resourced situations
Keeping it simple

The analyzer employs a simple chart-parsing strategy

- Allows for feature clashes
  - Step towards determining which constraints (of a huge space of possible variations) may be relaxed
- Relies on a handful of handwritten affixes, which essentially encode the “rules” of the grammar
  - Step towards developing analyzers for lesser-resourced situations
- Allows for flexibility & adaptability in, e.g., the positing of valid forms
  - Step towards different analyses for different kinds of learners
Different goals, different evaluations

The evaluation is tripartite, reflecting our different goals:
Different goals, different evaluations

The evaluation is tripartite, reflecting our different goals:

1. Quality of assigned morphological tags
Different goals, different evaluations

The evaluation is tripartite, reflecting our different goals:

1. Quality of assigned morphological tags
2. Error detection capabilities
The evaluation is tripartite, reflecting our different goals:

1. Quality of assigned morphological tags
2. Error detection capabilities
3. Ability to extract information for learner modeling
Different goals, different evaluations

The evaluation is tripartite, reflecting our different goals:

1. Quality of assigned morphological tags
2. Error detection capabilities
3. Ability to extract information for learner modeling

The work is still in progress, and thus the evaluation also points to ways in which the system can be improved.
Introduction and Motivation

Background

Methods

Evaluation & Preliminary Results

Summary & Outlook

References
Hungarian (Magyar) is an agglutinative language in the Finno-Ugric family. It possesses rich inflectional and derivational morphology, an extensive case system, free word order, and vowel harmony.
Hungarian

- Hungarian (Magyar) is an agglutinative language in the Finno-Ugric family
- It possesses rich inflectional and derivational morphology, an extensive case system, free word order, and vowel harmony

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

b. könyv -et olvas
book -ACC read
‘he/she reads a book’
Hungarian

- Hungarian (Magyar) is an agglutinative language in the Finno-Ugric family
- It possesses rich inflectional and derivational morphology, an extensive case system, free word order, and vowel harmony

(1) a. fut -ott -ál
run -PST -2SG.INDEF
‘you [2sg.] ran’
b. könyv -et olvas
book -ACC read
‘he/she reads a book’

(2) a. ház -ban
house -INESS
‘in (a) house’
b. könyv -eim -ben
book -1SG.PL -INESS
‘in my books’
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- HuMor (High-speed Unification Morphology) (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules

- Hunmorph (Trón et al., 2005; Halácsy et al., 2006)
  - Recursive affix-stripping
  - Our work: similar technique

- Morphdb (Trón et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013)
  - Lexical database, encoding for irregularities
  - Our work: borrow a lexicon, but no meta-information

Compared to Durst et al. (2014) & Vincze et al. (2014), we focus on descriptions of target and non-target-like forms
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- HuMor (High-speed Unification Morphology) (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules

- Hunmorph (Trón et al., 2005; Halácsy et al., 2006)
  - Recursive affix-stripping
  - Our work: similar technique

- Morphdb (Trón et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013)
  - Lexical database, encoding for irregularities
  - Our work: borrow a lexicon, but no meta-information

Compared to Durst et al. (2014) & Vincze et al. (2014), we focus on descriptions of target and non-target-like forms
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- **HuMor** (High-speed Unification Morphology) (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules
- **Hunmorph** (Trón et al., 2005; Halácsy et al., 2006)
  - Recursive affix-stripping

Compared to Durst et al. (2014) & Vincze et al. (2014), we focus on descriptions of target and non-target-like forms.
A few tools available for Hungarian:

- **HuMor (High-speed Unification Morphology) (Prószéky and Kis, 1999; Laki et al., 2013)**
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules
- **Hunmorph (Trón et al., 2005; Halácsy et al., 2006)**
  - Recursive affix-stripping
  - Our work: similar technique

Compared to Durst et al. (2014) & Vincze et al. (2014), we focus on descriptions of target and non-target-like forms.
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- **HuMor (High-speed Unification Morphology)** (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules
- **Hunmorph** (Trón et al., 2005; Halácsy et al., 2006)
  - Recursive affix-stripping
  - Our work: similar technique
- **Morphdb** (Trón et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013)
  - Lexical database, encoding for irregularities
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- HuMor (High-speed Unification Morphology) (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules
- Hunmorph (Trón et al., 2005; Halácsy et al., 2006)
  - Recursive affix-stripping
  - Our work: similar technique
- Morphdb (Trón et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013)
  - Lexical database, encoding for irregularities
  - Our work: borrow a lexicon, but no meta-information
Automatic morphological analysis for Hungarian

A few tools available for Hungarian:

- **HuMor (High-speed Unification Morphology)** (Prószéky and Kis, 1999; Laki et al., 2013)
  - Pre-encoded dictionary and feature-based rules
  - Our work: small dictionary of affixes/rules
- **Hunmorph (Trón et al., 2005; Halácsy et al., 2006)**
  - Recursive affix-stripping
  - Our work: similar technique
- **Morphdb (Trón et al., 2006; Bohnet et al., 2013; Farkas et al., 2012; Zsibrita et al., 2013)**
  - Lexical database, encoding for irregularities
  - Our work: borrow a lexicon, but no meta-information

Compared to Durst et al. (2014) & Vincze et al. (2014), we focus on descriptions of target and non-target-like forms.
Introduction and Motivation

Background

Methods

Evaluation & Preliminary Results

Summary & Outlook
Corpus data

- Daily journals from students of Hungarian (n=14) at 3 levels of proficiency at IU (9391 sentences)
  - Topics self-selected, each entry 10–15 sentences in length
  - More descriptive language than typically found in exercises
Corpus data

- Daily journals from students of Hungarian (n=14) at 3 levels of proficiency at IU (9391 sentences)
  - Topics self-selected, each entry 10–15 sentences in length
  - More descriptive language than typically found in exercises
- For all learners, data spans at least one semester
  - For 4 learners, texts are available from multiple semesters
Error annotation

Dickinson and Ledbetter (2012)

- Unit of analysis = the morpheme (SEGmentation layer)
Error annotation
Dickinson and Ledbetter (2012)

- Unit of analysis = the morpheme (SEGmentation layer)
- Errors annotated by linguistic categories:
  - Characters (CHA, phonology & spelling)
  - Morphemes (MOR, agreement & derivation)
  - Relations between morphemes (REL, selection)
  - Sentences (SNT, syntax)
Error annotation
Dickinson and Ledbetter (2012)

- Unit of analysis = the morpheme (SEGmentation layer)
- Errors annotated by linguistic categories:
  - Characters (CHA, phonology & spelling)
  - Morphemes (MOR, agreement & derivation)
  - Relations between morphemes (REL, selection)
  - Sentences (SNT, syntax)
- Target (TGT) layer provided
Example annotation

(3) Ajanl -om bor -t , nem sör -t
recommend 1SG.DF wine ACC , not beer ACC
‘I recommend wine, not beer.’

<table>
<thead>
<tr>
<th>TXT</th>
<th>Ajanlom</th>
<th>bort</th>
<th>,</th>
<th>nem</th>
<th>sört</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEG</td>
<td>Ajanl</td>
<td>om</td>
<td>bor</td>
<td>t</td>
<td>,</td>
<td>nem</td>
</tr>
<tr>
<td>CHA</td>
<td>CL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOR</td>
<td>MAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGT</td>
<td>Ajánl</td>
<td>ok</td>
<td>bor</td>
<td>t</td>
<td>,</td>
<td>nem</td>
</tr>
</tbody>
</table>
Morphological analysis

Analyzer uses:

- A freely available Hungarian dictionary (≈ wordlist)
- A knowledge base of derivational & inflectional affixes
Morphological analysis

Analyzer uses:

- A freely available Hungarian dictionary (≈ wordlist)
- A knowledge base of derivational & inflectional affixes

Analysis:

- Segmentation & morphological parsing inspired by (C)CG (Steedman and Baldridge, 2011)
  - Use CKY chart-parsing algorithm (Cocke and Schwartz, 1970)
Morphological analysis

Analyzer uses:

- A freely available Hungarian dictionary (≈ wordlist)
- A knowledge base of derivational & inflectional affixes

Analysis:

- Segmentation & morphological parsing inspired by (C)CG (Steedman and Baldridge, 2011)
  - Use CKY chart-parsing algorithm (Cocke and Schwartz, 1970)
- System posits possible stems implied by attested affixes
Knowledge base

Bulk of grammatical knowledge encoded here:

- Hand-wrote 205 affixes
  - Inflection (e.g. noun case, verb conjugation, possession)
  - Derivation (e.g. nominalizing suffixes, verbalizing suffixes)
Knowledge base

Bulk of grammatical knowledge encoded here:

- Hand-wrote 205 affixes
  - Inflection (e.g. noun case, verb conjugation, possession)
  - Derivation (e.g. nominalizing suffixes, verbalizing suffixes)
- An affix corresponds to a set of possible categories, encoding combinatory possibilities
  - Affix categories also contain features describing relevant linguistic properties

(4)  -ot: KP[k:acc]\N[vh:bk]
Building an analysis

(5) lát-hat-atlan-ság-od-at
see-“be able”-NEG-“ness”-2SG-ACC
‘your [2sg] invisibility’
Building an analysis

\[(5) \quad \text{lát -hat -atlan -ság -od -at}
\]
\[
\text{see - “be able” -NEG - “ness” -2SG -ACC}
\]

‘your [2sg] invisibility’
Unknown stems

Affix-driven system:

- If no root is found in the dictionary, system can posit a possible stem for the word based on the found affixes.
Unknown stems

Affix-driven system:

- If no root is found in the dictionary, system can posit a possible stem for the word based on the found affixes

(6) *₅ h₄ á₃ z₂ o₁ t₀

‘house+ACC’
Unknown stems

Affix-driven system:

- If no root is found in the dictionary, system can posit a possible stem for the word based on the found affixes

\[(6) \quad *_5 h_4 á_3 z_2 o_1 t_0 \]

‘house+ACC’

- \( ház_N + ot_{KN\backslash N} \)
Unknown stems

Affix-driven system:

- If no root is found in the dictionary, system can posit a possible stem for the word based on the found affixes

(6) \(*_5 \, h_4 \, á_3 \, z_2 \, o_1 \, t_0\)

‘house+ACC’

- \(ház_N + ot_{KN\backslash N}\)
- \(házo_{N_{hyp}} + t_{KN\backslash N}\)
Unknown stems

Affix-driven system:

- If no root is found in the dictionary, system can posit a possible stem for the word based on the found affixes

\[(6) \quad *_5 \, h \, 4 \, á \, 3 \, z \, 2 \, o \, 1 \, t \, 0\]
‘house+ACC’

- \(ház_N + ot_{KN/N}\)
- \(házo_{N_hyp} + t_{KN/N}\)

- Allows for root morphemes that are nonstandard, misspelled, proper nouns, etc.
Constraint relaxation

During derivation, features of affixes and stems are compared.
Constraint relaxation

During derivation, features of affixes and stems are compared

- Inconsistencies & clashes are marked

\[
\begin{align*}
\text{ház} & \quad \text{o}t \\
\text{N[+LOW]} & \quad \text{KN\N[-LOW]} \\
\text{KN[!LOW]} & \quad \text{KN[!LOW]}
\end{align*}
\]
Constraint relaxation

During derivation, features of affixes and stems are compared

- Inconsistencies & clashes are marked

\[
\begin{array}{c}
\text{h} & \text{á} & \text{z} \\
\text{N} [+\text{LOW}] & \text{o} & \text{t} \\
\text{KN} \backslash \text{N} [-\text{LOW}] \\
\text{KN} [+!\text{LOW}] \\
\end{array}
\]

- Stem requires lowered allomorph (-at) of accusative suffix, but unlowered allomorph is provided

- Clash of features ([−LOW] / [+LOW]) indicates learner’s current understanding
**Constraint relaxation**

During derivation, features of affixes and stems are compared

- Inconsistencies & clashes are marked

  ![h á z]
  
  \[N[+LOW]\]

  ![o t]
  
  \[KN\\N[-LOW]\]

  \[KN[!LOW]\]

- Stem requires lowered allomorph (-at) of accusative suffix, but unlowered allomorph is provided

- Clash of features \([-LOW] / [+LOW]\) indicates learner’s current understanding

**Importance of grammar-writer:** put relaxable constraints into features & non-relaxable constraints into main categories
Introduction and Motivation

Background

Methods

Evaluation & Preliminary Results

Summary & Outlook
Morphological analysis

- Evaluate accuracy for native (L1) & learner (L2) data

Example:

a. lát-t-2sg.indef ’you saw’

b. V0m1i2s324s5-6-7-8n9

Indicates: Main verb (m), indicative mood (i), past tense (s), second person (2), singular (s), indefinite (n)
Morphological analysis

- Evaluate accuracy for native (L1) & learner (L2) data
- System returns one or more derivations with morph. features
  - Same scheme as for Szeged Corpus (Csendes et al., 2004)
Morphological analysis

- Evaluate accuracy for native (L1) & learner (L2) data
- System returns one or more derivations with morph. features
  - Same scheme as for Szeged Corpus (Csendes et al., 2004)

(7) a. lát -t -ál
    see -PST -2SG.INDEF
    ‘you saw’

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

- Indicates: Main verb (m), indicative mood (i), past tense (s), second person (2), singular (s), indefinite (n)
Native language data
Evaluation set-up

We hand-verified the annotations of 1000 tokens from Szeged Corpus (Csendes et al., 2004)
Native language data
Evaluation set-up

We hand-verified the annotations of 1000 tokens from Szeged Corpus (Csendes et al., 2004)

- **Precision & Recall** defined w.r.t. context-independent list of appropriate tags;
Native language data

Evaluation set-up

We hand-verified the annotations of 1000 tokens from Szeged Corpus (Csendes et al., 2004)

- **Precision & Recall** defined w.r.t. context-independent list of appropriate tags;
  - **Accuracy** defined w.r.t. context-specific tags
Native language data

Evaluation set-up

We hand-verified the annotations of 1000 tokens from Szeged Corpus (Csendes et al., 2004)

- **Precision & Recall** defined w.r.t. context-independent list of appropriate tags;
  - **Accuracy** defined w.r.t. context-specific tags
- **Unknown POS**: system recognizes a word but no tag for it
  - Analyzer doesn’t have access to POS data in its dictionary
Native language data
Evaluation set-up

We hand-verified the annotations of 1000 tokens from Szeged Corpus (Csendes et al., 2004)

- **Precision & Recall** defined w.r.t. context-independent list of appropriate tags;
  - **Accuracy** defined w.r.t. context-specific tags
- **Unknown POS**: system recognizes a word but no tag for it
  - Analyzer doesn’t have access to POS data in its dictionary
- **Unknown word**: system cannot produce a derivation
Native language data

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>POS</th>
<th>+N</th>
<th>POS+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.308</td>
<td>—</td>
<td>0.307</td>
<td>—</td>
</tr>
<tr>
<td>Recall</td>
<td>0.262</td>
<td>—</td>
<td>0.315</td>
<td>—</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.467</td>
<td>0.568</td>
<td>0.505</td>
<td>0.592</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
</tr>
</tbody>
</table>

- POS: take into account only main POS
- +N: posit additional noun tag for unknown POS cases
- Major issue: monomorphemic nouns, pronouns, adjectives, or adverbs without any affix to guide guessing
- Limitation of dictionary (unknown POS) is a major problem
**Native language data**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>POS</th>
<th>+N</th>
<th>POS+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.308</td>
<td>—</td>
<td>0.307</td>
<td>—</td>
</tr>
<tr>
<td>Recall</td>
<td>0.262</td>
<td>—</td>
<td>0.315</td>
<td>—</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.467</td>
<td>0.568</td>
<td>0.505</td>
<td>0.592</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
</tr>
</tbody>
</table>

- **POS**: take into account only main POS
Native language data

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>POS</th>
<th>+N</th>
<th>POS+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.308</td>
<td>—</td>
<td>0.307</td>
<td>—</td>
</tr>
<tr>
<td>Recall</td>
<td>0.262</td>
<td>—</td>
<td>0.315</td>
<td>—</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.467</td>
<td>0.568</td>
<td>0.505</td>
<td>0.592</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
</tr>
</tbody>
</table>

- **POS**: take into account only main POS
- **+N**: posit additional noun tag for unknown POS cases
  - Major issue: monomorphemic nouns, pronouns, adjectives, or adverbs without any affix to guide guessing

Limitation of dictionary (unknown POS) is a major problem
Native language data

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>POS</th>
<th>+N</th>
<th>POS+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.308</td>
<td>—</td>
<td>0.307</td>
<td>—</td>
</tr>
<tr>
<td>Recall</td>
<td>0.262</td>
<td>—</td>
<td>0.315</td>
<td>—</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.467</td>
<td>0.568</td>
<td>0.505</td>
<td>0.592</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
<td>0.067</td>
</tr>
</tbody>
</table>

- **POS**: take into account only main POS
- **+N**: posit additional noun tag for unknown POS cases
  - Major issue: monomorphemic nouns, pronouns, adjectives, or adverbs without any affix to guide guessing

Limitation of dictionary (unknown POS) is a major problem
Corrected learner data

- Corrected forms for 1021 tokens from L2 Hungarian learners

<table>
<thead>
<tr>
<th></th>
<th>Total\textsubscript{Strict}</th>
<th>Total\textsubscript{Free}</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.499</td>
<td>0.509</td>
<td>0.846</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.499</td>
<td>0.499</td>
<td>—</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.109</td>
<td>0.097</td>
<td>0.027</td>
</tr>
</tbody>
</table>

- \textit{Strict}: no constraint relaxation; \textit{Free}: constraint relaxation
- Compare to Magyarlanc (ML, Zsibrita et al., 2013)
Raw learner data

- Same 1021 tokens, now with no corrections

<table>
<thead>
<tr>
<th></th>
<th>Total\textsubscript{Strict}</th>
<th>Total\textsubscript{Free}</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.464</td>
<td>0.478</td>
<td>0.753</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.456</td>
<td>0.456</td>
<td>—</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.137</td>
<td>0.119</td>
<td>0.074</td>
</tr>
</tbody>
</table>
Raw learner data

- Same 1021 tokens, now with no corrections

<table>
<thead>
<tr>
<th></th>
<th>Total_{Strict}</th>
<th>Total_{Free}</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.464</td>
<td>0.478</td>
<td>0.753</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.456</td>
<td>0.456</td>
<td>—</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.137</td>
<td>0.119</td>
<td>0.074</td>
</tr>
</tbody>
</table>

- Higher unknown word rate
Raw learner data

- Same 1021 tokens, now with no corrections

<table>
<thead>
<tr>
<th></th>
<th>Total$_{\text{Strict}}$</th>
<th>Total$_{\text{Free}}$</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.464</td>
<td>0.478</td>
<td>0.753</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.456</td>
<td>0.456</td>
<td>—</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.137</td>
<td>0.119</td>
<td>0.074</td>
</tr>
</tbody>
</table>

- Higher unknown word rate

- Magyarlanc: accuracy falls by about 10%
Raw learner data

- Same 1021 tokens, now with no corrections

<table>
<thead>
<tr>
<th></th>
<th>Total_{Strict}</th>
<th>Total_{Free}</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.464</td>
<td>0.478</td>
<td>0.753</td>
</tr>
<tr>
<td>Unk. POS</td>
<td>0.456</td>
<td>0.456</td>
<td>—</td>
</tr>
<tr>
<td>Unk. Word</td>
<td>0.137</td>
<td>0.119</td>
<td>0.074</td>
</tr>
</tbody>
</table>

- Higher unknown word rate
- *Magyarlan*: accuracy falls by about 10%
- Large proportion of test cases involve monomorphemic words for which the analyzer recognizes no internal structure
Error detection

- Error: unanalyzed word or clash between features
Error detection

- Error: unanalyzed word or clash between features
- Precision, recall, & F-score, with/without hypothesized roots
  - Here evaluated only against CHA & MOR layers
Error detection

- Error: unanalyzed word or clash between features
- Precision, recall, & F-score, with/without hypothesized roots
  - Here evaluated only against CHA & MOR layers

Without hypothesis:

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.380</td>
</tr>
<tr>
<td>Recall</td>
<td>0.789</td>
</tr>
<tr>
<td>F₁</td>
<td>0.513</td>
</tr>
<tr>
<td>F₀.₅</td>
<td>0.424</td>
</tr>
</tbody>
</table>
Error detection

- Error: unanalyzed word or clash between features
- Precision, recall, & F-score, with/without hypothesized roots
  - Here evaluated only against CHA & MOR layers

Without hypothesis:

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.380</td>
</tr>
<tr>
<td>Recall</td>
<td>0.789</td>
</tr>
<tr>
<td>F₁</td>
<td>0.513</td>
</tr>
<tr>
<td>F₀.⁵</td>
<td>0.424</td>
</tr>
</tbody>
</table>

With hypothesis:

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.400</td>
</tr>
<tr>
<td>Recall</td>
<td>0.043</td>
</tr>
<tr>
<td>F₁</td>
<td>0.078</td>
</tr>
<tr>
<td>F₀.⁵</td>
<td>0.152</td>
</tr>
</tbody>
</table>
Error detection

Major issue: unknown words & proper names
Error detection

Major issue: unknown words & proper names

- Without hypothesis: 40% of false positives are proper names
Error detection

Major issue: unknown words & proper names

- Without hypothesis: 40% of false positives are proper names
- With hypothesis: any unknown word or stem could now be a potentially correct form
  - ... including previously-analyzed stem errors
Error detection

Major issue: unknown words & proper names

- Without hypothesis: 40% of false positives are proper names
- With hypothesis: any unknown word or stem could now be a potentially correct form
  - ... including previously-analyzed stem errors

Some possible solutions:

- Spelling corrector as part of the pipeline (Durst et al., 2014)
- Short list of common Hungarian names
- Determine preference ordering of analyses
Grammar extraction

Exploratory analysis of approximating learner’s interlanguage grammar
Exploratory analysis of approximating learner’s interlanguage grammar

- **Goal:** Extract as much information as possible from learner productions to infer features of the interlanguage
- Sort out features which are good at distinguishing learner level from those which characterize individual learner differences
## Complexity

- **Morphemes Per Word (MPW)**
- **Words Per Sentence (WPS)**

<table>
<thead>
<tr>
<th></th>
<th>MPW</th>
<th>WPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beg01</td>
<td>1.38</td>
<td>5.79</td>
</tr>
<tr>
<td>Beg02</td>
<td>1.40</td>
<td>4.37</td>
</tr>
<tr>
<td>Beg03</td>
<td>1.52</td>
<td>3.84</td>
</tr>
<tr>
<td>Beg04</td>
<td>1.31</td>
<td>5.43</td>
</tr>
<tr>
<td>Beg06</td>
<td>1.52</td>
<td>5.75</td>
</tr>
<tr>
<td>Beg08</td>
<td>1.44</td>
<td>2.81</td>
</tr>
<tr>
<td>Beg09</td>
<td>1.58</td>
<td>3.28</td>
</tr>
<tr>
<td>Int01</td>
<td>1.51</td>
<td>6.40</td>
</tr>
<tr>
<td>Adv01</td>
<td>1.60</td>
<td>15.73</td>
</tr>
<tr>
<td>Adv02</td>
<td>1.66</td>
<td>10.90</td>
</tr>
</tbody>
</table>

MPW seems to be a largely individual feature of learner language. WPS has individual variation, but seems to increase over the course of acquisition: may indicate proficiency.
Complexity

- Morphemes Per Word (MPW)
- Words Per Sentence (WPS)

<table>
<thead>
<tr>
<th></th>
<th>MPW</th>
<th>WPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beg01</td>
<td>1.38</td>
<td>5.79</td>
</tr>
<tr>
<td>Beg02</td>
<td>1.40</td>
<td>4.37</td>
</tr>
<tr>
<td>Beg03</td>
<td>1.52</td>
<td>3.84</td>
</tr>
<tr>
<td>Beg04</td>
<td>1.31</td>
<td>5.43</td>
</tr>
<tr>
<td>Beg06</td>
<td>1.52</td>
<td>5.75</td>
</tr>
<tr>
<td>Beg08</td>
<td>1.44</td>
<td>2.81</td>
</tr>
<tr>
<td>Beg09</td>
<td>1.58</td>
<td>3.28</td>
</tr>
<tr>
<td>Int01</td>
<td>1.51</td>
<td>6.40</td>
</tr>
<tr>
<td>Adv01</td>
<td>1.60</td>
<td>15.73</td>
</tr>
<tr>
<td>Adv02</td>
<td>1.66</td>
<td>10.90</td>
</tr>
</tbody>
</table>

- MPW seems to be a largely individual feature of learner language
Complexity

- Morphemes Per Word (MPW)
- Words Per Sentence (WPS)

<table>
<thead>
<tr>
<th></th>
<th>MPW</th>
<th>WPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beg01</td>
<td>1.38</td>
<td>5.79</td>
</tr>
<tr>
<td>Beg02</td>
<td>1.40</td>
<td>4.37</td>
</tr>
<tr>
<td>Beg03</td>
<td>1.52</td>
<td>3.84</td>
</tr>
<tr>
<td>Beg04</td>
<td>1.31</td>
<td>5.43</td>
</tr>
<tr>
<td>Beg06</td>
<td>1.52</td>
<td>5.75</td>
</tr>
<tr>
<td>Beg08</td>
<td>1.44</td>
<td>2.81</td>
</tr>
<tr>
<td>Beg09</td>
<td>1.58</td>
<td>3.28</td>
</tr>
<tr>
<td>Int01</td>
<td>1.51</td>
<td>6.40</td>
</tr>
<tr>
<td>Adv01</td>
<td>1.60</td>
<td>15.73</td>
</tr>
<tr>
<td>Adv02</td>
<td>1.66</td>
<td>10.90</td>
</tr>
</tbody>
</table>

- MPW seems to be a largely individual feature of learner language
- WPS has individual variation, but seems to increase over course of acquisition: may indicate proficiency
Coverage

Verbal paradigm in the present tense indicative

- Beg. learner
- Favors 3rd person & 1st person plural, indefinite
Coverage

Verbal paradigm in the present tense indicative

- Adv. learner
- Favors 1st & 3rd person singular, indefinite

![Bar chart showing frequency of verbal paradigms](image)
Coverage

Verbal paradigm in the present tense indicative

- Adv. learner
- Favors 1st & 3rd person singular, indefinite
- Different usage patterns
- Investigation possible by auto. analysis

![Graph showing frequency of different verb forms]
Automatic Morphological Analysis of Learner Hungarian
Summary

We have:

- Presented a rule-based morphological analysis system for learner Hungarian, employing constraint relaxation
  - We have used little in the way of hand-built resources
- Performed three different evaluations to illustrate its utility for linguistic analysis, error analysis, or downstream applications
  - Information captured by the analyzer shows promise for describing the interlanguage of learners of Hungarian
Future directions:

▶ Handle named entities (Durst et al., 2014), e.g., lists of common names
▶ Extend methodology to syntax
▶ Explore record of language use to aid in disambiguation:
  ▶ e.g., if ambiguous stem only ever occurred previously with verbal morphology, its current use may be verbal
▶ Investigate iterative bootstrapping methods to allow for reduction of initial knowledge base
Thank you!

Köszönöm!


