Automatic Morphological Analysis of Learner Hungarian

Scott Ledbetter & Markus Dickinson
Indiana University

4 June, 2015

Introduction and Motivation

Introduction

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)

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)

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 flexbility & 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

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

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

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

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

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

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

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

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

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.'

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

Morphological analysis

Analyzer uses:

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

Morphological analysis

Analyzer uses:

- A freely available Hungarian dictionary ($\approx$ 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 ($\approx$ 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) lát -hat -atlan -ság -od -at
see -“be able” -NEG -“ness” -2SG -ACC
‘your [2sg] invisibility’

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\text{lát}}{\text{V}} \quad \dfrac{\text{hat}}{\text{V}\backslash\text{V}}}{\text{V}}< \quad \dfrac{\text{atlan}}{\text{A}\backslash\text{V}}}{\text{A}}< \quad \dfrac{\text{ság}}{\text{N}\backslash\text{A}}}{\text{N}}< \quad \dfrac{\text{od}}{\text{PN}\backslash\text{N}}}{\text{PN}}< \quad \dfrac{\text{at}}{\text{KN}\backslash\text{PN}}}{\text{KN}}<$$

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) $*_5$ h $_4$ á $_3$ z $_2$ o $_1$ t $_0$
'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) $*_5$ 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) $*_5$ h $_4$ á $_3$ z $_2$ o $_1$ t $_0$
  'house+ACC'

  - $ház_N + ot_{KN\backslash N}$
  - $ázo_{N_{hyp}} + t_{KN\backslash N}$

- 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

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

Constraint relaxation

During derivation, features of affixes and stems are compared

- Inconsistencies & clashes are marked

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

- 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

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

- 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

Morphological analysis

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

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

	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

Native language data

	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

- *POS*: take into account only main POS

Native language data

	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

- *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

Native language data

	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

- *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

	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

- *Strict*: no constraint relaxation; *Free*: constraint relaxation
- Compare to Magyarlanc (ML, Zsibrita et al., 2013)

Raw learner data

- Same 1021 tokens, now with no corrections

	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

Raw learner data

- Same 1021 tokens, now with no corrections

	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

- Higher unknown word rate

Raw learner data

- Same 1021 tokens, now with no corrections

	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

- Higher unknown word rate
- *Magyarlanc*: accuracy falls by about 10%

Raw learner data

- Same 1021 tokens, now with no corrections

	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

- Higher unknown word rate
- *Magyarlanc*: 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:

	Score
Precision	0.380
Recall	0.789
F_1	0.513
$F_{0.5}$	0.424

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:

	Score
Precision	0.380
Recall	0.789
F_1	0.513
$F_{0.5}$	0.424

With hypothesis:

	Score
Precision	0.400
Recall	0.043
F_1	0.078
$F_{0.5}$	0.152

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

Grammar extraction

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)

	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

Complexity

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

	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

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

Complexity

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

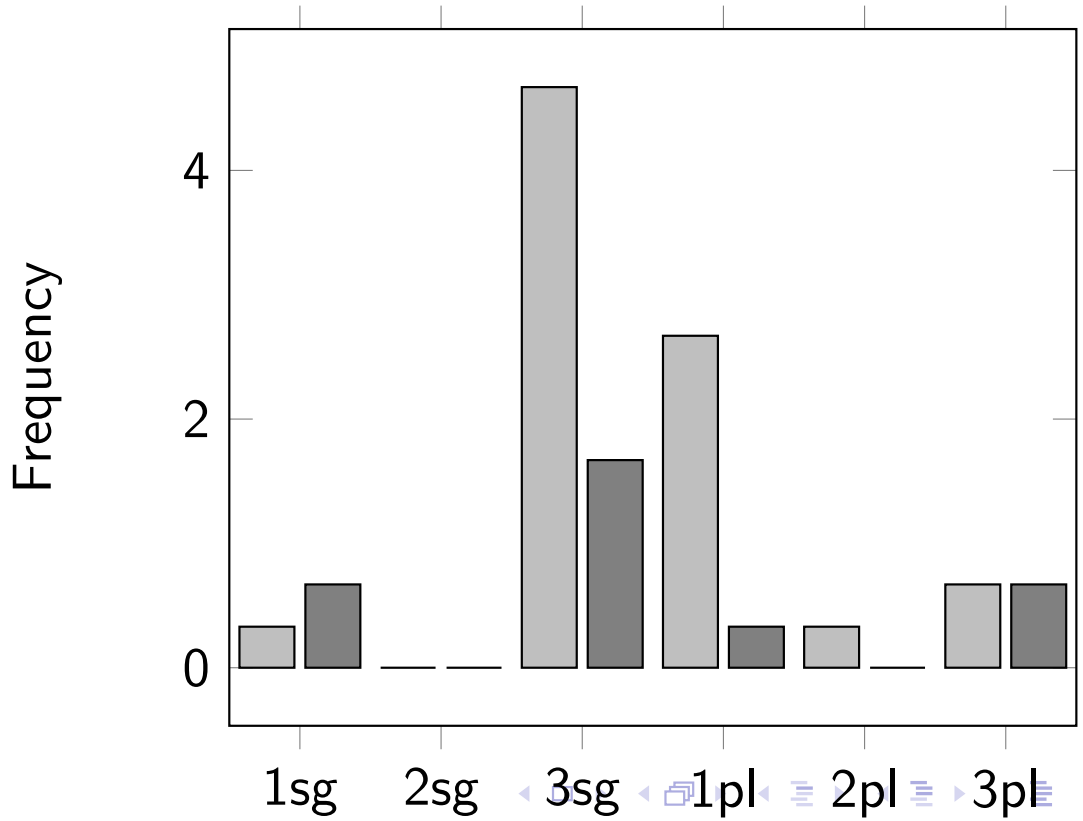
	MPW	WPS
Beg01	1.38	5.79
Beg02	1.40	4.37
Beg03	1.52	3.84
Beg04	1.31	5.43
Beg06	1.52	5.75
Beg08	1.44	2.81
Beg09	1.58	3.28
Int01	1.51	6.40
Adv01	1.60	15.73
Adv02	1.66	10.90

- 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

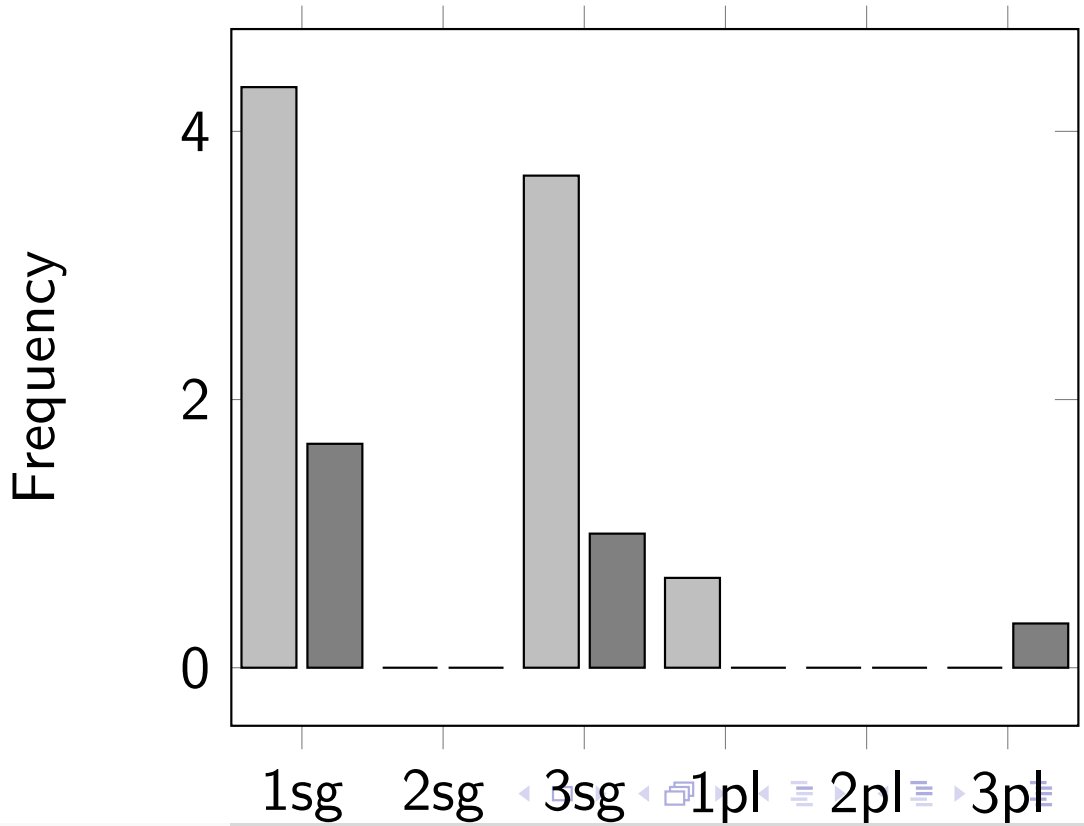
- Beg. learner
- Favors 3rd person & 1st person plural, indefinite



Coverage

Verbal paradigm in the present tense indicative

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



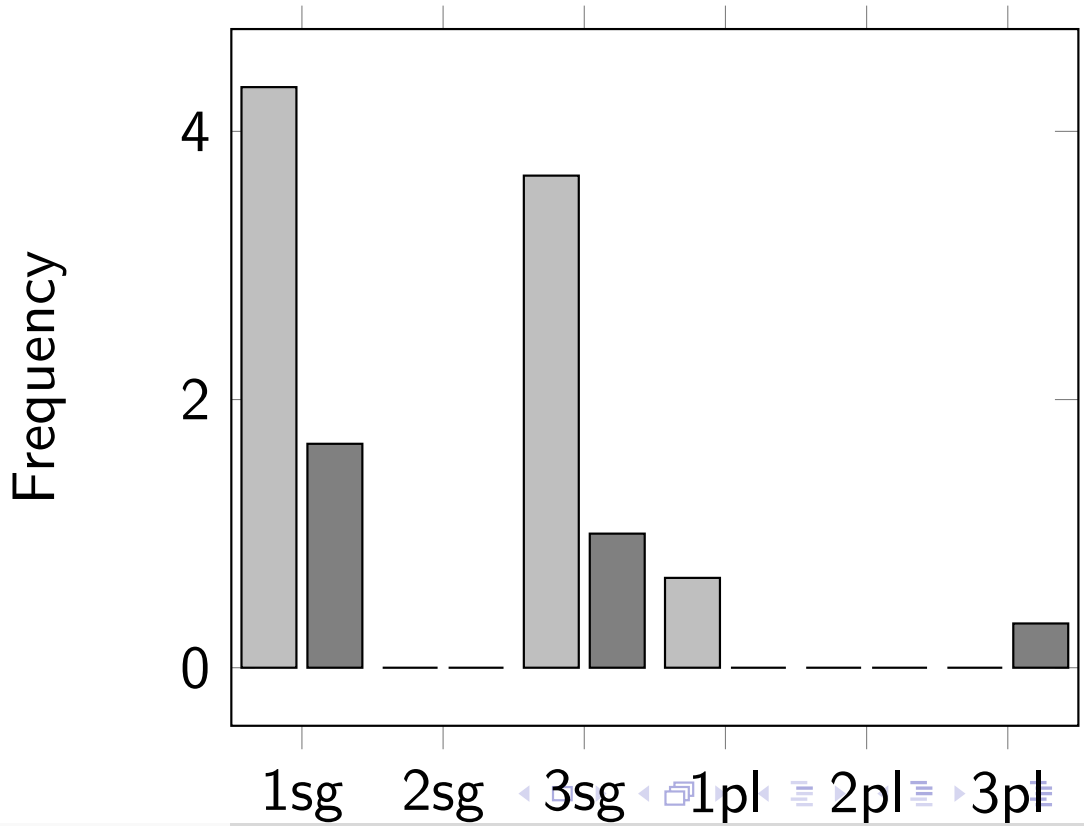
Coverage

Verbal paradigm in the present tense indicative

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

⇒ Different usage patterns

- Investigation possible by auto. analysis



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

Outlook

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!

References

Amaral, L. and Meurers, D. (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.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajic, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1(1):415–428.

Cocke, J. and Schwartz, J. T. (1970). *Programming languages and their compilers: Preliminary notes*. CIMS, NYU, 2nd rev. version edition.

Csendes, D., Csirik, J., and Gyimóthy, T. (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.

Dickinson, M. and Ledbetter, S. (2012). Annotating errors in a hungarian learner corpus. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey.

Durst, P., Szabó, M. K., Vincze, V., and Zsibrita, J. (2014). Using automatic morphological tools to process data from a learner corpus of hungarian. *Apples Journal of Applied Language Studies*, 8(3):39–54.

Farkas, R., Vincze, V., and Schmid, H. (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.

Halácsy, P., Kornai, A., Oravecz, C., Trón, V., and Varga, D. (2006). Using a morphological analyzer in high precision pos tagging of hungarian. In *Proceedings of LREC*, pages 2245–2248.

Heift, T. and Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Laki, L. J., Novák, A., and Siklósi, B. (2013). An english-to-hungarian morpheme-based statistical machine translation system with reordering rules. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation (HyTra)*, pages 42–50.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2014). *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, second edition.

Prószéky, G. and Kis, B. (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.

Ragheb, M. (2014). *Building a Syntactically-Annotated Corpus of Learner English*. PhD thesis, Indiana University, Bloomington, IN.

Steedman, M. and Baldridge, J. (2011). Combinatory categorial grammar. In Borsley, R. and Börjars, K., editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.

Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., and Varga, D. (2005). Hunmorph: Open source word analysis. In *Proceedings of the Workshop on Software*, pages 77–85. Association for Computational Linguistics.

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

Vajjala, S. and Loo, K. (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.

Vincze, V., Zsibrita, J., Durst, P., and Szabó, M. K. (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)*.

Zsibrita, J., Vincze, V., and Farkas, R. (2013). Magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In *Proceedings of RANLP*, pages 763–771.