The Language or the Task Design?

Re-Evaluating Morphological Inflection Tasks

Jordan Kodner
Salam Khalifa
Sarah Payne
Zoey Liu (UFL)

ATLaC
April 28, 2023

Morphological Inflection

Patterns of word formation to express grammatical categories

English *walk*+PAST → *walked*

Mandarin 3+PL → *tāmen* 'they'

Shona *bik*+1SG.SUBJ+6CL.OBJ+PAST+CAUS+PASS → *ndakachibikiswa* 'I was made to cook it'

Hebrew √*ĦTL*+DIM+SG+DEF → *ha-ħataltúl* the kitty'

Latin *amic*+FEM+SG+GEN → *amīcae* 'the friend's'

- Roots/stems are modified by many processes
  {suf,pref,in,circum}fixation, stem mutations, reduplication…
- Express number, tense, mood, voice, aspect, evidentiality, possession, case…
- Common across world languages
  But vary dramatically along many dimensions of complexity
- Poses a learning challenge for both machines and humans

Morphological Inflection as a Task

Training Time (lemma, inflected form, feature set) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time (lemma, feature set) pairs → predict the inflected forms

swim	?	V;PRS;3;SG
box	?	N;PL
cat	?	N;SG
…	…	…

Morphological Inflection as a Task

Training Time (lemma, inflected form, feature set) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time (lemma, feature set) pairs → predict the inflected forms

swim	?	V;PRS;3;SG	→	swims
box	?	N;PL		
cat	?	N;SG		
…	…	…		

Morphological Inflection as a Task

Training Time

(lemma, inflected form, feature set) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time

(lemma, feature set) pairs → predict the inflected forms

swim	?	V;PRS;3;SG	→	swims
box	?	N;PL	→	boxes
cat	?	N;SG		
…	…	…		

Morphological Inflection as a Task

Training Time

(**lemma**, **inflected form**, **feature set**) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time

(**lemma**, **feature set**) pairs → predict the **inflected forms**

swim	?	V;PRS;3;SG	→	swims
box	?	N;PL	→	boxes
cat	?	N;SG	→	cat
…	…	…		…

Why Do NLP and Comp Ling Researchers Study This?

1. Traditionally taken to be useful in downstream tasks
   - At least in settings where pipelining is still a thing → low-resource settings?
   - Particularly for languages with lots of inflectional morphology
2. May provide insight into the behavior of NN architectures
3. May elucidate aspects of linguistic typology
4. May elucidate aspects of language acquisition

Why Do NLP and Comp Ling Researchers Study This?

1. Traditionally taken to be useful in downstream tasks
   - At least in settings where pipelining is still a thing → low-resource settings?
   - Particularly for languages with lots of inflectional morphology
2. May provide insight into the behavior of NN architectures
   - A particular kind of string-to-string mapping problem
   - Varying performance ideally reflects divergent properties of different architectures
3. May elucidate aspects of linguistic typology
4. May elucidate aspects of language acquisition

Why Do NLP and Comp Ling Researchers Study This?

1. Traditionally taken to be useful in downstream tasks
    - At least in settings where pipelining is still a thing → low-resource settings?
    - Particularly for languages with lots of inflectional morphology
2. May provide insight into the behavior of NN architectures
    - A particular kind of string-to-string mapping problem
    - Varying performance ideally reflects divergent properties of different architectures
3. May elucidate aspects of linguistic typology
    - Typology - systematically characterizes how languages are the same/different
    - Differing performance across languages ideally identifies typological differences
4. May elucidate aspects of language acquisition

Why Do NLP and Comp Ling Researchers Study This?

1. Traditionally taken to be useful in downstream tasks
   - At least in settings where pipelining is still a thing → low-resource settings?
   - Particularly for languages with lots of inflectional morphology

A typological issue!

2. May provide insight into the behavior of NN architectures
   - A particular kind of string-to-string mapping problem
   - Varying performance ideally reflects divergent properties of different architectures
3. May elucidate aspects of linguistic typology
   - Typology - systematically characterizes how languages are the same/different
   - Differing performance across languages ideally identifies typological differences
4. May elucidate aspects of language acquisition

Why Do NLP and Comp Ling Researchers Study This?

1. Traditionally taken to be useful in downstream tasks
   - At least in settings where pipelining is still a thing → low-resource settings?
   - Particularly for languages with lots of inflectional morphology
2. May provide insight into the behavior of NN architectures
   - A particular kind of string-to-string mapping problem
   - Varying performance ideally reflects divergent properties of different architectures
3. May elucidate aspects of linguistic typology
   - Typology - systematically characterizes how languages are the same/different
   - Differing performance across languages ideally identifies typological differences
4. May elucidate aspects of language acquisition
   - Acquisition - formal study of how children initially learn their native languages
   - Computational learners ideally point towards feasible models for human learning

Why Do NLP and Comp Ling Researchers Study This?

1. **Traditionally taken to be useful in downstream tasks**
    - At least in settings where pipelining is still a thing → low-re
    - Particularly for languages with lots of inflectional morpho
2. **May provide insight into the behavior of NN arc**
    - A particular kind of string-to-string mapping problem
    - Varying performance ideally reflects divergent properties of different architectures
3. **May elucidate aspects of linguistic typology**
    - Typology - systematically characterizes how languages are
    - Differing performance across languages ideally identifies t
4. **May elucidate aspects of language acquisition**
    - Acquisition - formal study of how children initially learn th
    - Computational learners ideally point towards feasible models for human learning

Linguistics informing specific questions in NLP
(we're cautiously optimistic for this particular task)

NLP informing specific questions in linguistics
(we're skeptical for this particular task)

Is this task already solved?

Reported on inflection shared tasks is often near-ceiling

Accuracy of the best system on a subset of the 2018 CoNLL-SIGMORPHON shared task languages

Variable across systems, but really good overall on on medium and high training!

	High (10,000)	Medium (1,000)	Low (100)
Adyghe	100.00(uzh-2)	94.40(uzh-1)	90.60(ua-8)
Albanian	98.90(bme-2)	88.80(iitbhu-iiith-2)	36.40(uzh-1)
Arabic	93.70(uzh-1)	79.40(uzh-1)	45.20(uzh-1)
Armenian	96.90(bme-2)	92.80(uzh-1)	64.90(uzh-1)
Asturian	98.70(uzh-1)	92.40(iitbhu-iiith-2)	74.60(uzh-2)
Azeri	100.00(axsemantics-2)	96.00(iitbhu-iiith-2)	65.00(iitbhu-iiith-2)
Bashkir	99.90(uzh-2)	97.30(uzh-2)	77.80(iitbhu-iiith-1)
Basque	98.90(bme-2)	88.10(iitbhu-iiith-2)	13.30(uzh-1)
Belarusian	94.90(uzh-1)	70.40(uzh-1)	33.40(ua-8)
Bengali	99.00(bme-3)	99.00(uzh-2)	72.00(uzh-2)
Breton	100.00(waseda-1)	96.00(uzh-2)	72.00(uzh-1)
Bulgarian	98.30(uzh-2)	83.80(uzh-2)	62.90(ua-8)
Catalan	98.90(uzh-2)	92.80(waseda-1)	72.50(ua-8)
Classical-syriac	100.00(axsemantics-1)	100.00(axsemantics-2)	96.00(uzh-2)
Cornish	—	70.00(uzh-1)	40.00(ua-4)
Crimean-tatar	100.00(iit-varanasi-1)	98.00(uzh-2)	91.00(iitbhu-iiith-2)
Czech	94.70(uzh-1)	87.20(uzh-1)	46.50(uzh-2)
Danish	95.50(uzh-1)	80.40(uzh-1)	87.70(ua-6)
Dutch	97.90(uzh-1)	85.70(uzh-1)	69.30(ua-6)
English	97.10(uzh-2)	94.50(uzh-1)	91.80(ua-8)

Is this task already solved?

But performance on closely related languages is highly variable…

Azeri	100.00(axsemantics-2)	96.00(iitbhu-iiith-2)	65.00(iitbhu-iiith-2)
Turkish	98.50(uzh-2)	90.70(uzh-1)	39.50(iitbhu-iiith-2)
Turkmen	—	98.00(iitbhu-iiith-1)	90.00(uzh-2)

Belarusian	94.90(uzh-1)	70.40(uzh-1)	33.40(ua-8)
Russian	94.40(uzh-2)	86.90(uzh-1)	53.50(uzh-1)
Ukrainian	96.20(uzh-2)	81.40(uzh-1)	57.10(ua-6)

Finnish	95.40(uzh-1)	82.80(uzh-1)	25.70(uzh-1)
Ingrian	—	92.00(uzh-2)	46.00(iitbhu-iiith-2)
Karelian	—	100.00(uzh-2)	94.00(ua-5)

Kashubian	—	88.00(bme-2)	68.00(ua-5)
Lower-sorbian	97.80(uzh-1)	85.10(uzh-1)	54.30(ua-6)
Polish	93.40(uzh-2)	82.40(uzh-2)	49.40(ua-6)

Danish	95.50(uzh-1)	80.40(uzh-1)	87.70(ua-6)
Norwegian-bokmaal	92.10(uzh-2)	84.10(uzh-1)	90.10(ua-6)
Swedish	93.30(uzh-1)	79.80(uzh-1)	79.00(ua-8)

Czech	94.70(uzh-1)	87.20(uzh-1)	46.50(uzh-2)
Slovak	97.10(uzh-1)	78.60(uzh-1)	51.80(uzh-2)

Galician	99.50(uzh-1)	90.80(uzh-1)	61.10(uzh-2)
Portuguese	98.60(uzh-2)	94.80(uzh-2)	75.80(uzh-2)

Irish	91.50(uzh-2)	77.10(uzh-1)	37.70(uzh-1)
Scottish-gaelic	—	94.00(iitbhu-iiith-1)	74.00(iitbhu-iiith-2)

Is this task already solved?

But performance on closely related languages is highly variable…

Azeri	100.00(axsemantics-2)	96.00(iitbhu-iiith-2)	65.00(iitbhu-iiith-2)
Turkish	98.50(uzh-2)	90.70(uzh-1)	39.50(iitbhu-iiith-2)
Turkmen	—	98.00(iitbhu-iiith-1)	90.00(uzh-2)

Belarusian	94.90(uzh-1)	70.40(uzh-1)	33.40(ua-8)
Russian	94.40(uzh-2)	86.90(uzh-1)	53.50(uzh-1)
Ukrainian	96.20(uzh-2)	81.40(uzh-1)	57.10(ua-6)

Finnish	95.40(uzh-1)	82.80(uzh-1)	25.70(uzh-1)
Ingrian	—	92.00(uzh-2)	46.00(iitbhu-iiith-2)
Karelian	—	100.00(uzh-2)	94.00(ua-5)

Kashubian	—	88.00(bme-2)	68.00(ua-5)
Lower-sorbian	97.80(uzh-1)	85.10(uzh-1)	54.30(ua-6)
Polish	93.40(uzh-2)	82.40(uzh-2)	49.40(ua-6)

Danish	95.50(uzh-1)	80.40(uzh-1)	87.70(ua-6)
Norwegian-bokmaal	92.10(uzh-2)	84.10(uzh-1)	90.10(ua-6)
Swedish	93.30(uzh-1)	79.80(uzh-1)	79.00(ua-8)

Czech	94.70(uzh-1)	87.20(uzh-1)	46.50(uzh-2)
Slovak	97.10(uzh-1)	78.60(uzh-1)	51.80(uzh-2)

Galician	99.50(uzh-1)	90.80(uzh-1)	61.10(uzh-2)
Portuguese	98.60(uzh-2)	94.80(uzh-2)	75.80(uzh-2)

Irish	91.50(uzh-2)	77.10(uzh-1)	37.70(uzh-1)
Scottish-gaelic	—	94.00(iitbhu-iiith-1)	74.00(iitbhu-iiith-2)

Unsurprising in ML when different samples yield different performance, but what in particular is going on here?

Revisiting Train-Test Overlap

- Of course, no train triples appeared in test
- But what about lemmas or feature sets individually?

  Conceptually, test items have four possible licit relationships with train

Illustrative Train Set

eat	eating	V;V.PTCP;PRS
run	ran	V;PST

Illustrative Test Set

eat	V;PST	← No OOV, not attested together
run	V;NFIN	← Only feature set is OOV
see	V;PST	← Only lemma is OOV
go	V;PRS;3;SG	← Lemma and feature set are OOV
run	V;PST	← Train-on-test (not present)

Revisiting Train-Test Overlap

- Of course, no train triples appeared in test
- But what about lemmas or feature sets individually?

  Conceptually, test items have four possible licit relationships with train

Illustrative Train Set

```
eat   eating   V;V.PTCP;PRS
run   ran      V;PST
```

Illustrative Test Set

```
eat   V;PST        ← No OOV, not attested together
run   V;NFIN       ← Only feature set is OOV
see   V;PST        ← Only lemma is OOV
go    V;PRS;3;SG   ← Lemma and feature set are OOV
run   V;PST        ← Train-on-test (not present)
```

Do lemma and/or feature set overlap predict performance?

Overlaps as Performance Ceilings

Lemma Overlap — % of test items with lemmas attested in train

Feature Set Overlap — % of test items with feat sets attested in train

% Overlap defines the performance ceiling for a hypothetical system with zero ability to generalize along a given dimension

Overlaps as Performance Ceilings

Lemma Overlap % of test items with lemmas attested in train

Feature Set Overlap % of test items with feat sets attested in train

% Overlap defines the performance ceiling for a hypothetical system with zero ability to generalize along a given dimension

Training Size	Best Acc	Feat Set Overlap	Δ
Low (100)	39.5%	39.6%	-0.1%
Medium (1,000)	90.7	94.1	-3.4
High (10,000)	98.5	100	-1.5

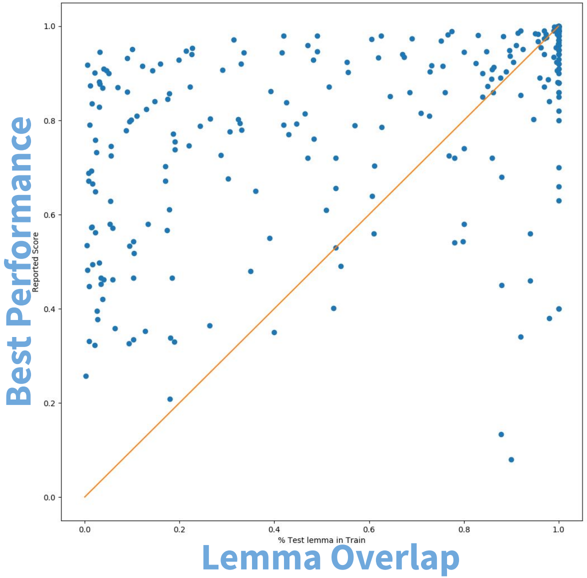
Very suspicious ceiling-like results for Turkish…
Inflectional category generalization should be possible!

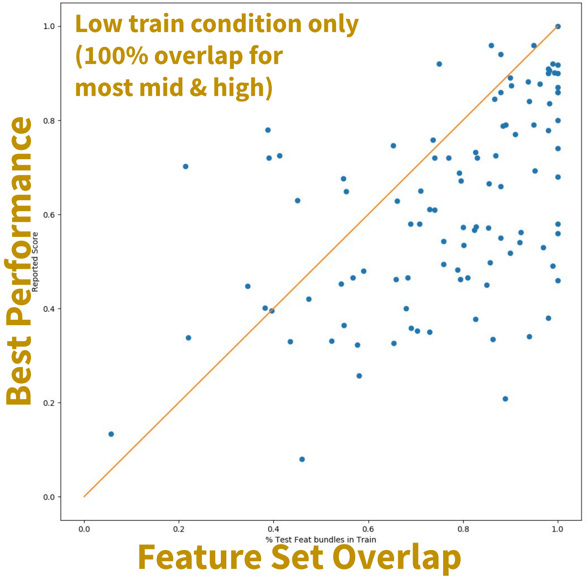
Overlaps as Performance Ceilings

Lemma overlap is not a ceiling; Feature set overlap is a soft ceiling

Many points above the ceiling suggests good lemma generalization ability

Few points above the ceiling suggests poor feature set generalization





Our Motivating Suspicions

- Cross-linguistic differences are actually primarily driven by sampling effects
  → We don't know how typology relates to performance
- Train-test overlaps, especially feature set overlap leads these sampling effects
- High reported performance is due to artificially high feature set overlap
  → Systems may not actually be generalizing like they appear too

Two Research Areas

1. **Uncontrolled data biases → inflated/variable performance**

   Must/how to control for lemma and feature set overlap (2022, *SIGMORPHON*)

   Must/how to also control for sampling strategy (*under review*)

   Must/how to also control for original corpus size (*in prep*)

2. **Inflated/variable performance → linguistic claims unmotivated**

   Behavior is not acquisition-like (2022, *SIGMORPHON*; 2023, *CogSci*; *in prep*)

   - **Alternative models (w/ Belth & Yang): (2021, *SCiL*; 2021, *CogSci*; *in prep*)**

   Behavior doesn't reflect typology (2022, *SIGMORPHON*; *under review*; *in prep*)

Two Research Areas

1. **Uncontrolled data biases → inflated/variable performance**

   Must/how to control for lemma and feature set overlap (2022, *SIGMORPHON*)

   Must/how to also control for sampling strategy (*under review*)

   Must/how to also control for original corpus size (*in prep*)

2. **Inflated/variable performance → linguistic claims unmotivated**

   Behavior is not acquisition-like (2022, *SIGMORPHON*; 2023, *CogSci*; *in prep*)

   Alternative models (w/ Belth & Yang): (2021, *SCiL*; 2021, *CogSci*; *in prep*)

   Behavior doesn't reflect typology (2022, *SIGMORPHON*; *under review*; *in prep*)

Kodner, Khalifa, et *xviii* al. (SIGMORPHON 2022)

2022 SIGMORPHON Typologically Diverse Inflection Shared Task

33 languages from 10 families

Afro-Asiatic:
Semitic:
- Arabic
- Hebrew

Uralic:
Ugric:
- Hungarian

Finnic:
- Karelian
- Ludian
- Veps

Turkic:
Kipchak:
- Kazakh

Oghuz:
- Turkish

Austronesian:
Malayo-Polynesian:
- Lamahalot

Chutko-Kamchatkan:
North:
- Chukchi

South:
- Itelmen

Tungusic:
North:
- Evenki

South:
- Xibe

Yeniseian:
- Ket

Koreanic:
- Korean

Kartvelian:
- Georgian

Indo-European:
Armenian:
- E. Armenian

Germanic:
- Gothic
- Low German
- Old English
- Middle Low German
- Old Norse
- Old High German

Indic:
- Assamese
- Braj
- Kholosi
- Magahi
- Gujarati

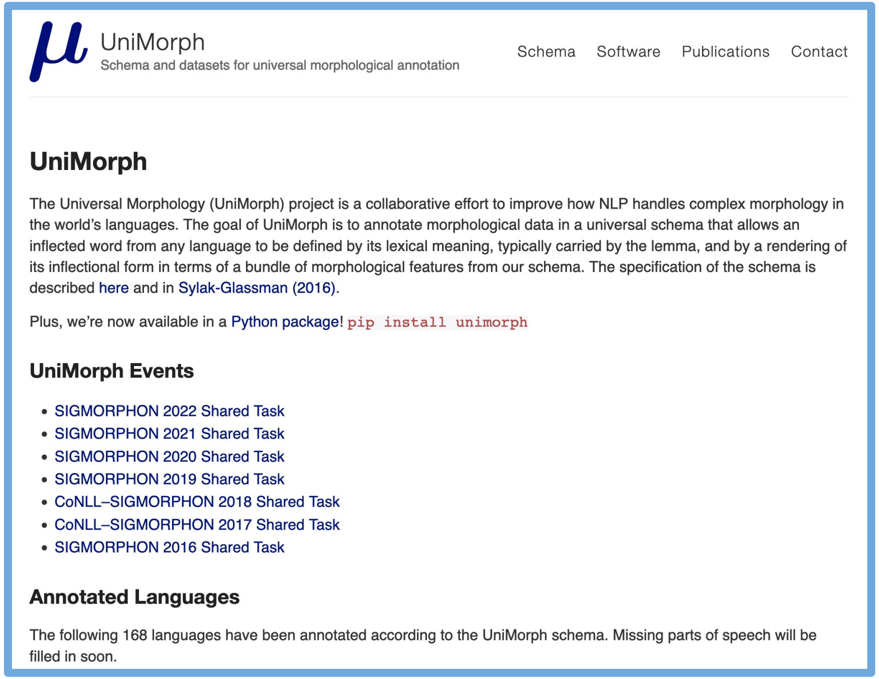
Slavic:
- Polish
- Pomak
- Slovak
- Upper Sorbian

Kodner, Khalifa, et *xviii* al. (SIGMORPHON 2022)

2022 SIGMORPHON Typologically Diverse Inflection Shared Task[1]

- **33 languages from 10 families**
- **Data from UniMorph 3/4 collection of morphological corpora[2]**

All corpora contain `(lemma,infl,feats)` triples with no frequency information



UniMorph
Schema and datasets for universal morphological annotation

Schema Software Publications Contact

UniMorph

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema. The specification of the schema is described here and in Sylak-Glassman (2016).

Plus, we're now available in a Python package! `pip install unimorph`

UniMorph Events

- SIGMORPHON 2022 Shared Task
- SIGMORPHON 2021 Shared Task
- SIGMORPHON 2020 Shared Task
- SIGMORPHON 2019 Shared Task
- CoNLL–SIGMORPHON 2018 Shared Task
- CoNLL–SIGMORPHON 2017 Shared Task
- SIGMORPHON 2016 Shared Task

Annotated Languages

The following 168 languages have been annotated according to the UniMorph schema. Missing parts of speech will be filled in soon.

[1]Code available at: https://github.com/sigmorphon/2022InflectionST, [2]McCarthy et al (2020)

Kodner, Khalifa, et *xviii* al. (SIGMORPHON 2022)

2022 SIGMORPHON Typologically Diverse Inflection Shared Task[1]

- 33 languages from 10 families
- Data from UniMorph 3/4 collection of morphological corpora[2]
- Train-Dev-Test splits were made with overlaps in mind
- Small Train ⊂ Large Train
- Small Train-Test feature set overlap ≤50% and as close to 50% as possible

  Large Train-Test feature set overlap naturally approached 100%

  Lemma overlap was naturally lower when feature set overlap was controlled

Split	Size
Small Train	700
Large Train	7000
Dev	1000
Test	2000

[1]Code available at: https://github.com/sigmorphon/2022InflectionST, [2]McCarthy et al (2020), Batsuren et al (2022)

Submitted Systems

CLUZH Clematide, Wehrli, & Makarov

Character-level neural transducer with teacher-forcing, individual embeddings for each feature

Flexica Scherbakov & Vylomova

Extension of non-neural baseline

OSU Elsner & Court

Character-level transformer augmented with exemplar model

TüMorph-FST Merzhevich, Gbadegoye, Girrbach, Li, & Shim

Hand-built FSTs for Chukchi, Kholosi, and Upper Sorbian

TüMorph-Main " " " " & "

Modification of Wu et al (2021) which predicts distributions over FST states

UBC Yang, Yang, Nicolai, & Silfverberg

Modification of Wu et al (2021) with hallucination

NeurBase Wu et al (2021)

Character-level transformer

NonNeurBase same as 2021

Finds common prefixes/suffixes in lemma-inflection pairs

Submitted Systems

CLUZH Clematide, Wehrli, & Makarov

Character-level neural transducer with teacher-forcing, individual embeddings for each feature

Flexica Scherbakov & Vylomova

Extension of non-neural baseline

OSU Elsner & Court

Character-level transformer augmented with exemplar model

TüMorph-FST Merzhevich, Gbadegoye, Girrbach, Li, & Shim

Hand-built FSTs for Chukchi, Kholosi, and Upper Sorbian

TüMorph-Main " " " " & "

Modification of Wu et al (2021) which predicts distributions over FST states

UBC Yang, Yang, Nicolai, & Silfverberg

Modification of Wu et al (2021) with hallucination

NeurBase Wu et al (2021)

Character-level transformer

NonNeurBase same as 2021

Finds common prefixes/suffixes in lemma-inflection pairs

Baselines

Submitted Systems

CLUZH Clematide, Wehrli, & Makarov

Character-level neural transducer with teacher-forcing, individual embeddings for each feature

Flexica **Scherbakov & Vylomova**

Extension of non-neural baseline

OSU Elsner & Court

Character-level transformer augmented with exemplar model

TüMorph-FST **Merzhevich, Gbadegoye, Girrbach, Li, & Shim**

Hand-built FSTs for Chukchi, Kholosi, and Upper Sorbian

TüMorph-Main " " " " & "

Modification of Wu et al (2021) which predicts distributions over FST states

UBC Yang, Yang, Nicolai, & Silfverberg

Modification of Wu et al (2021) with hallucination

NeurBase Wu et al (2021)

Character-level transformer

NonNeurBase **same as 2021**

Finds common prefixes/suffixes in lemma-inflection pairs

Non-neural

Summary Results

System	Small Training Condition					Large Training Condition				
	Overall	Both	Feats	Lemma	Neither	Overall	Both	Feats	Lemma	Neither
CLUZH	56.871	77.308	77.966	31.269	43.255	67.853	90.991	87.171	41.425	60.300
Flexica	34.406	59.503	61.616	6.390	14.562	38.243	66.846	73.007	4.985	21.337
OSU	*47.688**	*79.310**	*82.308**	*8.565**	*44.133**	46.734	89.565	85.308	4.843	16.768
TüM-FST	*67.308**	*100.00**	*75.000**	*55.319**	*72.115**	—	—	—	—	—
TüM-M	*41.591**	*58.907**	*62.469**	*18.597**	*27.613**	57.627	77.995	76.009	34.916	48.720
UBC	57.234	75.963	74.201	35.519	46.060	71.259	89.503	85.063	50.583	66.224
NeurBase	47.626	65.027	66.539	24.929	35.601	62.391	80.462	77.627	42.166	55.563
NonNeur	33.321	58.475	59.969	5.566	14.431	37.583	67.434	72.283	4.843	16.768

*OSU, TüMorph-FST, and TüMorph-Main were only run on some languages in small (italicized)

TüMorph-FST, was not run on large training

Summary Results

System	Small Training Condition					Large Training Condition				
	Overall	Both	Feats	Lemma	Neither	Overall	Both	Feats	Lemma	Neither
CLUZH	56.871	77.308	77.966	31.269	43.255	67.853	90.991	87.171	41.425	60.300
Flexica	34.406	59.503	61.616	6.390	14.562	38.243	66.846	73.007	4.985	21.337
OSU	*47.688**	*79.310**	*82.308**	*8.565**	*44.133**	46.734	89.565	85.308	4.843	16.768
TüM-FST	*67.308**	*100.00**	*75.000**	*55.319**	*72.115**	—	—	—	—	—
TüM-M	*41.591**	*58.907**	*62.469**	*18.597**	*27.613**	57.627	77.995	76.009	34.916	48.720
UBC	57.234	75.963	74.201	35.519	46.060	71.259	89.503	85.063	50.583	66.224
NeurBase	47.626	65.027	66.539	24.929	35.601	62.391	80.462	77.627	42.166	55.563
NonNeur	33.321	58.475	59.969	5.566	14.431	37.583	67.434	72.283	4.843	16.768

- **All systems perform much better**
  when test item feature sets are seen (Both, Feats Only)
  than when they are novel (Lemma Only, Neither)
- **Overall performance on Large Training is lower than in previous years**

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity

Paradigm Size - Are unseen feature sets a real problem?

- Feature sets (= inflectional categories = paradigm cells) follow sparse long-tailed frequency distributions

\+ For languages with paradigms with 10^2 or 10^3 items, not all will be attested in even millions of training tokens

− For languages with small paradigms, most/all feature sets should be attested

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes Small paradigm → maybe not

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes	Small paradigm → maybe not

Agglutinativity - Are feature set → form mappings predictable?

\+ A perfectly agglutinative language would express each feature as its own affix (each feature maps to a morphological form)

→ Can predict the form of the feature set from its members

− A perfectly fusional language would express each feature set as its own morphological operation (each feature set maps to a morphological form)

→ Cannot predict the form of the feature set from its members

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes Small paradigm → maybe not
- Highly agglutinative → yes Highly fusional → no

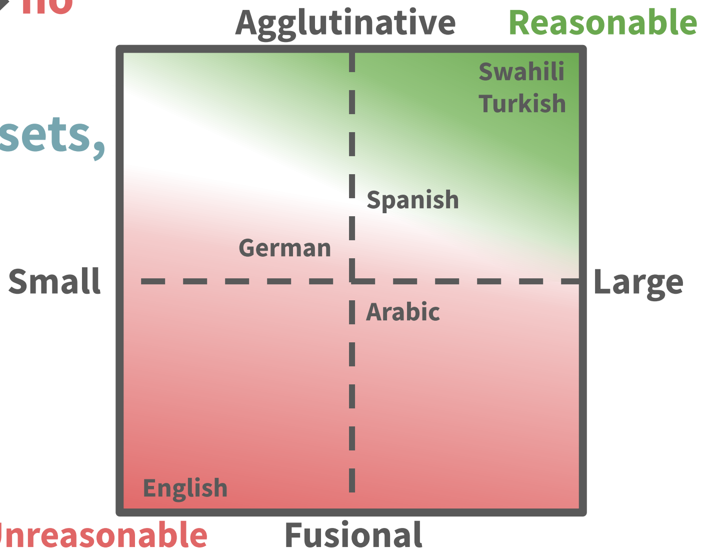
Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes Small paradigm → maybe not
- Highly agglutinative → yes Highly fusional → no

If systems can generalize to unseen feature sets,

we should see a much smaller performance hit on the most agglutinative languages



Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes Small paradigm → maybe not
- Highly agglutinative → yes Highly fusional → no

“Could an undergrad do it?”

Rule of thumb for if a system
can be expected to do it

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes   Small paradigm → maybe not
- Highly agglutinative → yes   Highly fusional → no

"Could an undergrad do it?"

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish
guakamole 'guacamole'

Feature Set	Inflected Form
N;ACC;SG	?
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	?
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
...	...

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes   Small paradigm → maybe not
- Highly agglutinative → yes   Highly fusional → no

"Could an undergrad do it?"

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish
guakamole 'guacamole'

Feature Set	Inflected Form
N;ACC;SG	?
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	?
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
…	…

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes   Small paradigm → maybe not
- Highly agglutinative → yes   Highly fusional → no

"Could an undergrad do it?"

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish
guakamole 'guacamole'

Feature Set	Inflected Form
N;ACC;SG	?
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	?
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
...	...

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes   Small paradigm → maybe not
- Highly agglutinative → yes   Highly fusional → no

“Could an undergrad do it?”

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish *guakamole* ‘guacamole’

Feature Set	Inflected Form
N;ACC;SG	?
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	?
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
…	…

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes   Small paradigm → maybe not
- Highly agglutinative → yes   Highly fusional → no

"Could an undergrad do it?"

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish *guakamole* 'guacamole'

Feature Set	Inflected Form
N;ACC;SG	*?*
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	*?*
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
…	…

Typological Expectations

Is generalization to unseen feature sets a reasonable expectation?

- Two linguistic dimensions at play: paradigm size and agglutinativity
- Large paradigm → yes Small paradigm → maybe not
- Highly agglutinative → yes Highly fusional → no

"Could an undergrad do it?"

Rule of thumb for if a system can be expected to do it

e.g., partial paradigm for Turkish *guakamole* 'guacamole'

Feature Set	Inflected Form
N;ACC;SG	*guakamoleyi*
N;ACC;PL	*guakamoleleri*
N;DAT;SG	*guakamoleye*
N;DAT;PL	*guakamolelere*
N;ACC;PL;PSS3S	*guakamolelerini*
N;DAT;PL;PSS3S	*guakamolelerine*
…	…

Performance on the Most Agglutinative Languages

The Agglutinative Languages:

Chukchi, Evenki, Georgian, Hungarian, Itelmen, Karelian, Kazakh, Ket, Korean, Ludic, Mongolian, Turkish, Veps, Xibe

No system generalizes well to unseen feature sets even when they technically should be able to

Features	Small Training		Large Training	
System	Seen	Novel	Seen	Novel
CLUZH	78.837	34.118	90.198	40.657
Flexica	60.885	11.386	69.173	10.094
OSU	*77.800**	*30.376**	88.497	13.456
TüM-FST	*100.00**	*17.778**	—	—
TüM-Main	*61.730**	*14.816**	74.667	29.433
UBC	75.994	39.232	89.213	49.799

*OSU, TüMorph-FST, and TüMorph-Main were only run on some languages in small (italicized)

Kodner, Khalifa, et *xviii* al. (SIGMORPHON 2022)

Conclusions

- Systems tend to generalize well to unseen lemmas, poorly to feature sets
  → Overlaps must be controlled for or reported separately
  → Previous results are probably task- rather than language-dependent
- Poor feature set generalization even when the task is feasible
  → Previously unrecognized aspect of NNs linguistic generalization abilities
  → A practical concern for languages with large paradigms

Kodner, Payne, Khalifa, & Liu (*under review*)

How does train-test sampling affect model behavior?

- **Quality over quantity: 5 languages that we could analyze more deeply**
  German, English, Spanish, Swahili and Turkish verbs
  Swahili and Turkish are highly regular and agglutinative
- UniMorph 3+4 intersected with text for frequency information
- Uniform vs frequency-weighted vs overlap-aware sampling
- Resplitting/reevaluating on 5 random seeds
- Evaluated 4 systems from SIGMORPHON 2022

Kodner, Payne, Khalifa, & Liu (*under review*)

How does train-test sampling affect model behavior?

- Quality over quantity: 5 languages that we could analyze more deeply
- UniMorph 3+4 intersected with text for frequency information

  CHILDES for German, English, and Spanish

  Wikipedia for Swahili and Turkish

  This step also filters out some errors from UniMorph
- Uniform vs frequency-weighted vs overlap-aware sampling
- Resplitting/reevaluating on 5 random seeds
- Evaluated 4 systems from SIGMORPHON 2022

Kodner, Payne, Khalifa, & Liu (*under review*)

How does train-test sampling affect model behavior?

- Quality over quantity: 5 languages that we could analyze more deeply
- UniMorph 3+4 intersected with text for frequency information
- Uniform vs frequency-weighted vs overlap-aware sampling

UNIFORM	doable on raw UniMorph
WEIGHTED	more naturalistic; weighted by corpus frequency
OVERLAPAWARE	balances test items with seen and unseen feature sets

- Resplitting/reevaluating on 5 random seeds
- Evaluated 4 systems from SIGMORPHON 2022

Kodner, Payne, Khalifa, & Liu (*under review*)

How does train-test sampling affect model behavior?

- **Quality over quantity: 5 languages that we could analyze more deeply**
- **UniMorph 3+4 intersected with text for frequency information**
- **Uniform vs frequency-weighted vs overlap-aware sampling**
- **Resplitting/reevaluating on 5 random seeds**

  A way to assess how typical a given evaluation's results are

  Previously applied to morphological segmentation[1]
- **Evaluated 4 systems from SIGMORPHON 2022**

Split	Size
Small Train	400 + 100 finetune
Large Train	1600 + 400 finetune
Dev	500
Test	1000

[1] Liu & Prud'hommeaux (2022)

Kodner, Payne, Khalifa, & Liu (*under review*)

How does train-test sampling affect model behavior?

- Quality over quantity: 5 languages that we could analyze more deeply
- UniMorph 3+4 intersected with text for frequency information
- Uniform vs frequency-weighted vs overlap-aware sampling
- Resplitting/reevaluating on 5 random seeds
- Evaluated 4 systems from SIGMORPHON 2022

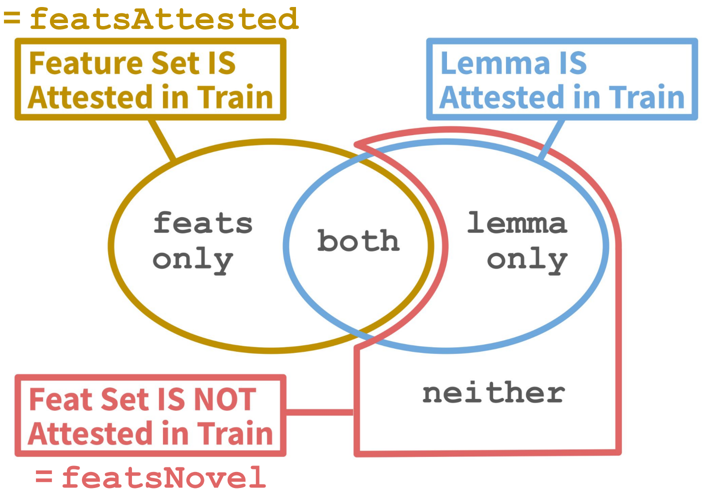
  Clematide et al (2022) with beam decoding ← best performer with available code

  Clematide et al (2022) with greedy decoding

  Wu et al (2021)

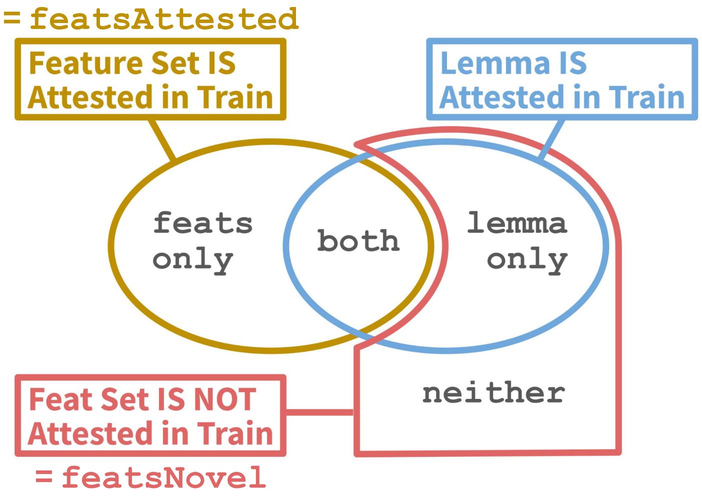
  Non-Neural Baseline

Effect of Sampling Strategy on Overlaps



Small Train	featsAttested	featsNovel	σ
UNIFORM	80.33%	19.67%	19.50
WEIGHTED	90.44	9.56	11.13
OVERLAPAWARE	48.81	51.19	0.98
Large Train	**featsAttested**	**featsNovel**	**σ**
UNIFORM	96.17%	3.83%	5.55
WEIGHTED	95.36	4.64	7.28
OVERLAPAWARE	49.92	50.08	0.17

Effect of Sampling Strategy on Overlaps



Small Train	featsAttested	featsNovel	σ
UNIFORM	80.33%	19.67%	19.50
WEIGHTED	90.44	9.56	11.13
OVERLAPAWARE	48.81	51.19	0.98
Large Train	**featsAttested**	**featsNovel**	**σ**
UNIFORM	96.17%	3.83%	5.55
WEIGHTED	95.36	4.64	7.28
OVERLAPAWARE	49.92	50.08	0.17

- Overlap rate is high but not 100% when not controlled for
- Overlap rate is highly variable across seeds/languages when not controlled for
- UNIFORM and WEIGHTED are similar
- OVERLAPAWARE succeeds at its goal

Average Performance - OVERLAPAWARE

	Small Training				Large Training			
Language	featsAttested	featsNovel	µ %Δ	Overall	featsAttested	featsNovel	µ %Δ	Overall
Arabic	66.14%	31.11%	*-52.96*	47.81%	76.09%	46.09%	*-39.43*	61.06%
English	88.45	18.99	*-78.53*	53.72	91.95	19.32	*-78.99*	55.63
German	74.12	41.60	*-43.87*	57.81	81.84	43.24	*-47.17*	62.54
Spanish	79.90	21.92	*-72.57*	50.35	87.92	24.83	*-71.76*	56.37
Swahili	84.79	41.75	*-50.76*	62.28	88.56	44.01	*-50.30*	66.14
Turkish	84.18	31.43	*-62.66*	57.03	90.94	35.59	*-60.86*	63.23

agglutinative

Average Performance - OVERLAPAWARE

Language	Small Training				Large Training			
	featsAttested	featsNovel	μ %Δ	Overall	featsAttested	featsNovel	μ %Δ	Overall
Arabic	66.14%	31.11%	*-52.96*	47.81%	76.09%	46.09%	*-39.43*	61.06%
English	88.45	18.99	*-78.53*	53.72	91.95	19.32	*-78.99*	55.63
German	74.12	41.60	*-43.87*	57.81	81.84	43.24	*-47.17*	62.54
Spanish	79.90	21.92	*-72.57*	50.35	87.92	24.83	*-71.76*	56.37
Swahili	84.79	41.75	*-50.76*	62.28	88.56	44.01	*-50.30*	66.14
Turkish	84.18	31.43	*-62.66*	57.03	90.94	35.59	*-60.86*	63.23

agglutinative

- Performance is strictly better on Large Train than Small Train
- Language ranking by average performance is consistent on both training sizes
- But performance gap between featsAttested vs feats Novel does not improve
- Performance hit on featsNovel is not smaller for the agglutinative languages

Score Range and Standard Dev across Random Seeds

- Score ranges are large
  → Results on a single split are likely not representative
- Range and standard deviation
  OverlapAware > Weighted > Uniform

Small Train	Score Range	σ
Uniform	4.51%	1.84
Weighted	6.33	2.57
OverlapAware	12.13	5.01
Large Train	**Score Range**	**σ**
Uniform	3.99%	1.68
Weighted	4.08	1.66
OverlapAware	13.06	5.50

Kodner, Payne, Khalifa, & Liu (*under review*)

Main Conclusions

- **UNIFORM and WEIGHTED sampling are similar, OVERLAPAWARE is adversarial**
  Some FeatsNovel test items do appear in UNIFORM and WEIGHTED
  Performance is lowest on OVERLAPAWARE
- **Score ranges are quite high across randoms seeds**
  Performance on one random sample unlikely to reflect true performance
  High variability for OVERLAPAWARE → it matters which feature sets are in train

Ongoing Follow-Up (*in prep*)

How does the size of the original corpus affect sampling?

- All training-test splits are subsampled from pre-existing corpora
  → Larger corpora are more downsampled that smaller corpora
  This will change expected overlaps?
- **Intuition:** smaller initial corpus should yield higher expected overlaps?
  → If overlaps are uncontrolled in sampling, performance should be systematically higher for languages with smaller initial corpora

Analytic and Empirical Analyses

What is the expected overlap for a given subsample?

- **Same reasoning for feature set and lemma overlap**
- **Depends on** initial corpus size
  train and test size
  class size (# of items w/ given feature set or lemma)
  number of classes

Analytic and Empirical Analyses

What is the expected overlap for a given subsample?

- Same reasoning for feature set and lemma overlap
- Depends on
  - initial corpus size
  - train and test size
  - class size (# of items w/ given feature set or lemma)
  - number of classes
- As initial corpus size grows
  - # lemmas grows, but class size is constant
  - # feature sets is constant, but class size grows

Analytic and Empirical Analyses

What is the expected overlap for a given subsample?

- Same reasoning for feature set and lemma overlap
- Depends on
  - initial corpus size
  - train and test size
  - class size (# of items w/ given feature set or lemma)
  - number of classes
- As initial corpus size grows
  - # lemmas grows, but class size is constant
  - # feature sets is constant, but class size grows

This can be investigated empirically and analytically

Analytic Calculation

A two-part calculation

- "How many lemmas (or feat sets) will be sampled at least once in training?"
  $P_{\#train}(m)$= from pmf for #classes drawn from a multivar. hypergeometric distr.[1]

m = # of classes sampled

k = # class in data set

n = # of items sampled

N = # of items in data

N = vector of class sizes

$$P(m|\mathbf{N},n) = \frac{\sum_{i=1}^{m}\binom{k-i}{k-m}(-1)^{m-i}U(n,i,\mathbf{N})}{\binom{N}{n}}$$

where $$U(n,j,\mathcal{P}(\mathbf{N})) = \sum_{X\in C_j(\mathcal{P}(\mathbf{N}))}\binom{X}{n}$$

and $$C_j(\mathcal{P}(\mathbf{N})) = \{\sum N_i^* \in \gamma : \gamma \subseteq \mathcal{P}(\mathbf{N}) \wedge |\gamma| = j\}$$

[1]Walton (1986)

Analytic Calculation

A two-part calculation

- “How many lemmas (or feat sets) will be sampled at least once in training?”
  $P_{\#train}(m)$= from pmf for #classes drawn from a multivar. hypergeometric distr.[1]

m = # of classes sampled

k = # class in data set

n = # of items sampled

N = # of items in data

N = vector of class sizes

$$P(m|\mathbf{N},n) = \frac{\sum_{i=1}^{m} \binom{k-i}{k-m}(-1)^{m-i}U(n,i,\mathbf{N})}{\binom{N}{n}}$$

where $$U(n,j,\mathcal{P}(\mathbf{N})) = \sum_{X \in C_j(\mathcal{P}(\mathbf{N}))} \binom{X}{n}$$

and $$C_j(\mathcal{P}(\mathbf{N})) = \{\sum N_i^* \in \gamma : \gamma \subseteq \mathcal{P}(\mathbf{N}) \wedge |\gamma| = j\}$$

This becomes impractical to calculate for large *N*, *n*, and *k*

[1]Walton (1986)

Analytic Calculation

A two-part calculation

- **"How many lemmas (or feat sets) will be sampled at least once in training?"**
  $P_{\#train}(m)$= from pmf for #classes drawn from a multivar. hypergeometric distr.[1]
- **"How many items in test have lemma (or feat set) overlap with train?"**
  Partition test items into lemmaAttested/lemmaUnattested by some *m* → Expected proportion of test items with overlap follows bivariate hypergeometric distr.

[Final formula TBD]

[1]Walton (1986)

Empirical Investigation

Given a corpus, train size, and test size, what is the overlap?

- **Perform many UNIFORM train-test splits and calculate average overlaps**
- Simulate smaller corpora by randomly removing lemmas
- Run some systems on some of these train-test splits and report performance

Empirical Investigation

Given a corpus, train size, and test size, what is the overlap?

- Perform many UNIFORM train-test splits and calculate average overlaps
- Simulate smaller corpora by randomly removing lemmas

    For a set of languages with different paradigm sizes,

        vary corpus size, train size, test size

        Perform many train-test splits

        Report overlaps as a function of these and make some 3D plots

- Run some systems on some of these train-test splits and report performance

Empirical Investigation

Given a corpus, train size, and test size, what is the overlap?

- Perform many UNIFORM train-test splits and calculate average overlaps
- Simulate smaller corpora by randomly removing lemmas
- Run some systems on some of these train-test splits and report performance

Does overlap correlate with corpus size holding train/test size constant?

Does performance correlate with corpus size holding train/test size constant?

Two Research Areas

1. **Uncontrolled data biases → inflated/variable performance**

   Must/how to control for lemma and feature set overlap (2022, *SIGMORPHON*)

   Must/how to also control for sampling strategy (*under review*)

   Must/how to also control for original corpus size (*in prep*)

2. **Inflated/variable performance → Linguistic claims unmotivated**

   Behavior is not acquisition-like (2022, *SIGMORPHON*; 2023, *CogSci*; *in prep*)

   - **Alternative models (w/ Belth & Yang): (2021, *SCiL*; 2021, *CogSci*; *in prep*)**

   Behavior doesn't reflect typology (2022, *SIGMORPHON*; *under review*; *in prep*)

Two Research Areas

1. Uncontrolled data biases → inflated/variable performance

   Must/how to control for lemma and feature set overlap (2022, *SIGMORPHON*)

   Must/how to also control for sampling strategy (*under review*)

   Must/how to also control for original corpus size (*in prep*)

2. Inflated/variable performance → Linguistic claims unmotivated

   Behavior is not acquisition-like (2022, *SIGMORPHON*; 2023, *CogSci*; *in prep*)

   - Alternative models (w/ Belth & Yang): (2021, *SCiL*; 2021, *CogSci*; *in prep*)

   Behavior doesn't reflect typology (2022, *SIGMORPHON*; *under review*; *in prep*)

Kodner and Khalifa (SIGMORPHON 2022)

2022 SIGMORPHON Acquisition-Inspired Inflection Shared Task[1]

To what extent do systems show learning trajectories similar to children on child-like input?

- For NNs to be useful in studying language acquisition, they should be reasonable models of language acquisition
- One desideratum for reasonable computational cognitive models is the ability to simulate human behavior

[1] https://github.com/sigmorphon/2022InflectionST

Kodner and Khalifa (SIGMORPHON 2022)

2022 SIGMORPHON Acquisition-Inspired Inflection Shared Task[1]

- **Three languages with substantial literature on morphology acquisition**
  English past tense, German noun plurals, Arabic noun plurals
- English and German data drawn from CHILDES collection of child-directed speech corpora[2] and intersected with UniMorph
- Arabic drawn from the Penn Arabic Treebank[3] then intersected w/ UniMorph
- Train-Dev-Test splits were made with WEIGHTED sampling
- Nested train sets increase in increments of 100
  to simulate developmental trajectories

[1] https://github.com/sigmorphon/2022InflectionST

Kodner and Khalifa (SIGMORPHON 2022)

2022 SIGMORPHON Acquisition-Inspired Inflection Shared Task[1]

- **Three languages with substantial literature on morphology acquisition**
- **English and German data drawn from CHILDES collection of child-directed speech corpora[2] and intersected with UniMorph**
- **Arabic drawn from the Penn Arabic Treebank[3] then intersected w/ UniMorph**
- **Train-Dev-Test splits were made with WEIGHTED sampling**
- **Nested train sets increase in increments of 100 to simulate developmental trajectories**

[1]https://github.com/sigmorphon/2022InflectionST, [2]https://childes.talkbank.org/, [3]Diab et al (2013)

Kodner and Khalifa (SIGMORPHON 2022)

2022 SIGMORPHON Acquisition-Inspired Inflection Shared Task[1]

- Three languages with substantial literature on morphology acquisition
- English and German data drawn from CHILDES collection of child-directed speech corpora[2] and intersected with UniMorph
- Arabic drawn from the Penn Arabic Treebank[3] then intersected w/ UniMorph
- Train-Dev-Test splits were made with WEIGHTED sampling
- Nested train sets increase in increments of 100 to simulate developmental trajectories

Split	Ara	Deu	Eng
Max Train	1000	600	1000
Dev	343	500	454
Test	600	600	600

[1]https://github.com/sigmorphon/2022InflectionST, [2]https://childes.talkbank.org/, [3]Diab et al (2013)

Kodner, Khalifa, Payne, & Liu (CogSci 2023)

Follow-Up on Acquisition-Inspired Shared Task

- **Same three languages and acquisition phenomena**
  Identical data for Arabic and German
  Used all of NA-English CHILDES
- UNIFORM vs WEIGHTED sampling
- Evaluated with 5 random seeds
- Same systems as the paper under review

Kodner, Khalifa, Payne, & Liu (CogSci 2023)

Follow-Up on Acquisition-Inspired Shared Task

- Same three languages and acquisition phenomena
- UNIFORM vs WEIGHTED sampling

  WEIGHTED frequency-weighted sampling better reflects acquisition setting

  More frequent words are more likely to be acquired earlier[1]
- Evaluated with 5 random seeds
- Same systems as the paper under review

[1]Goodman (2008)

Kodner, Khalifa, Payne, & Liu (CogSci 2023)

Follow-Up on Acquisition-Inspired Shared Task

- Same three languages and acquisition phenomena
- UNIFORM vs WEIGHTED sampling
- Evaluated with 5 random seeds

Similar analyses to the paper under review

- Same systems as the paper under review

Kodner, Khalifa, Payne, & Liu (CogSci 2023)

Follow-Up on Acquisition-Inspired Shared Task

- Same three languages and acquisition phenomena
- UNIFORM vs WEIGHTED sampling
- Evaluated with 5 random seeds
- Same systems as the paper under review

CLUZH	Clematide et al (2022) /w beam and greedy decoding
CHR-TRM	Wu et al (2021)
Non-neural baseline	

Submitted Systems (SIGMORPHON, 2022)

CLUZH	Clematide, Wehrli, & Makarov
HeiMorph	Ramarao, Zinova, Tang & van de Vijver
OSU	Elsner & Court
CHR-TRM	Wu et al (2021)
NonNeurBase	same as 2021

Submitted Systems (SIGMORPHON, 2022)

CLUZH	Clematide, Wehrli, & Makarov
HeiMorph	Ramarao, Zinova, Tang & van de Vijver
OSU	Elsner & Court
CHR-TRM	Wu et al (2021)
NonNeurBase	same as 2021

Character transformer with bigram-aware halluciation

Submitted Systems (SIGMORPHON, 2022)

CLUZH	Clematide, Wehrli, & Makarov
HeiMorph	Ramarao, Zinova, Tang & van de Vijver
OSU	Elsner & Court
CHR-TRM	Wu et al (2021)
NonNeurBase	same as 2021

Same system as Subtask 1

Submitted Systems (SIGMORPHON, 2022)

CLUZH	Clematide, Wehrli, & Makarov
HeiMorph	Ramarao, Zinova, Tang & van de Vijver
OSU	Elsner & Court
CHR-TRM	Wu et al (2021)
NonNeurBase	same as 2021

Ran these for CogSci 2023

Patterns in the Acquisition of English Past Tense

- Productive/Default *-ed* acquired around age 3 on a few hundred verb types[1]
- Over-regularization - Children apply -ed where it should not apply

  ******What dat feeled?*[2]
- Over-irregularization - Order of magnitude less common

  ******fry-frew* by analogy with *fly-flew*

  Consistent asymmetry cross-linguistically[3]

[1]Brown (1973), Marcus et al. (1992), [2]Brown (1973), [3]Clahsen et al. (1992), Xu & Pinker (1995), Mayol et al. (2007)

Patterns in the Acquisition of English Past Tense

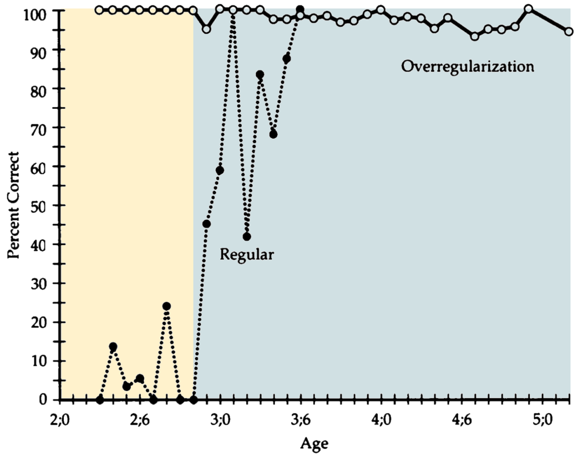
- Productive/Default -*ed* acquired around age 3 on a few hundred verb types[1]
- Over-regularization - Children apply -ed where it should not apply
- Over-irregularization - Order of magnitude less common
- *U*-shaped learning[4]

  Performance improves, worsens, improves

  Suggestions three phases in learning

  1. Memorization
  2. Learn productive -*ed*
  3. Relearn exceptions to -*ed*

Adam from the Brown Corpus



[1]Brown (1973), Marcus et al. (1992), [2]Brown (1973), [3]Clahsen et al. (1992), Xu & Pinker (1995), Mayol et al. (2007), [4]Marcus et al. (1992), Prasada & Prince (1993)

Patterns in the Acquisition of German Noun Plurals

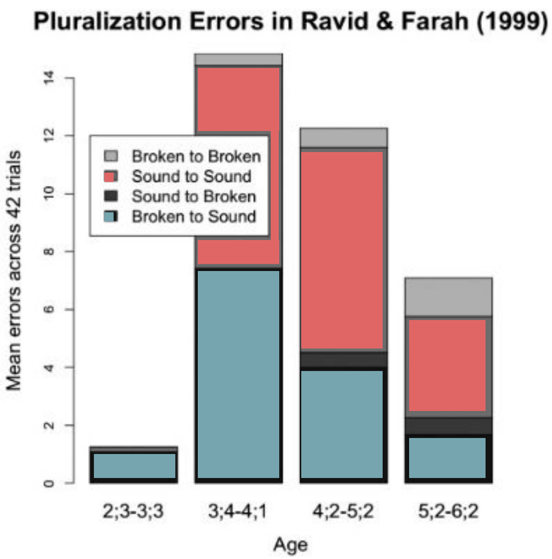
- Confound in English verbs - the productive ending is by far the most frequent
- German nouns take one of five endings[1]
  -s is the least frequent and the productive "ending of last resort"[1]
- *-e* and *-∅* are acquired before *-er* and *-s*[2]
- Productive use of *-s* appears late[1]
- Endings partially conditioned on gender and stem-final segments[3]
- Interacts with Umlaut (a kind of stem change)

Suffix*	% of all	% of NEUT
-(e)*n*	37.3%	3.2%
-*e*	34.4%	51.9%
-∅	19.2%	21.5%
-*er*	2.0%	10.6%
-*s*	4.0%	7.7%
other	2.1%	5.1%

[1]Elsen (2002), [2]Kopcke (1998), Szagun (2001), [4]Sonnenstuhl & Huth, 2002, *Numbers from Corkery et al. (2019)

Patterns in the Acquisition of Arabic Noun Plurals

- Arabic has two plural types
  Sound plurals take a suffix: MASC *-ūn*, FEM *-āt*
  Broken plurals undergo a stem change: dozens of patterns
- Errors are overwhelmingly
  (MASC) sound → (FEM) sound
  Broken → (FEM) sound
  Example of the over-regularization asymmetry
- Arabic-learning children show *u*-shaped learning[1]



[1]Ravid & Farah (1999)

Summary Results at Max Training Size (SIGMORPHON'22)

	at N=1000		at N=600			at N=1000	
System	English	Ortho	German	Suffix	Umlaut	Arabic	SfSmB
CLUZH	88.67%	91.17%	80.17%	89.00%	90.67%	65.83%	75.50%
HeiMorph	77.33	82.0	73.33	85.83	88.83	59.33	71.00
OSU	88.67	90.67	75.00	85.67	90.17	65.33	76.00

Summary Results at Max Training Size (SIGMORPHON'22)

	at N=1000		at N=600			at N=1000	
System	English	Ortho	German	Suffix	Umlaut	Arabic	SfSmB
CLUZH	88.67%	91.17%	80.17%	89.00%	90.67%	65.83%	75.50%
HeiMorph	77.33	82.0	73.33	85.83	88.83	59.33	71.00
OSU	88.67	90.67	75.00	85.67	90.17	65.33	76.00

Ignoring minor orthographic errors

Only evaluated suffix
Random baseline: 20%

Only evaluated Umlaut
Random baseline: 50%

Ignoring broken-to-broken errors
Random baseline: 33.3%

Summary Results at Max Training Size (SIGMORPHON'22)

	at N=1000		at N=600			at N=1000	
System	English	Ortho	German	Suffix	Umlaut	Arabic	SfSmB
CLUZH	88.67%	91.17%	80.17%	89.00%	90.67%	65.83%	75.50%
HeiMorph	77.33	82.0	73.33	85.83	88.83	59.33	71.00
OSU	88.67	90.67	75.00	85.67	90.17	65.33	76.00

Ortho: Ignoring minor orthographic errors

Suffix: Only evaluated suffix
Random baseline: 20%

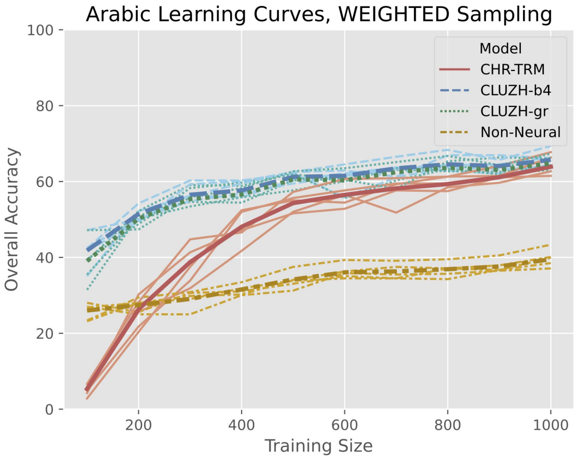
Umlaut: Only evaluated Umlaut
Random baseline: 50%

SfSmB: Ignoring broken-to-broken errors
Random baseline: 33.3%

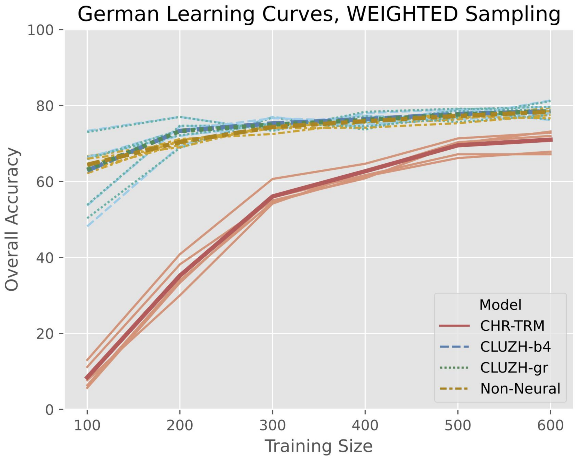
→ Performance decreases as pattern complexity increases →

Learning Curves (CogSci’23)

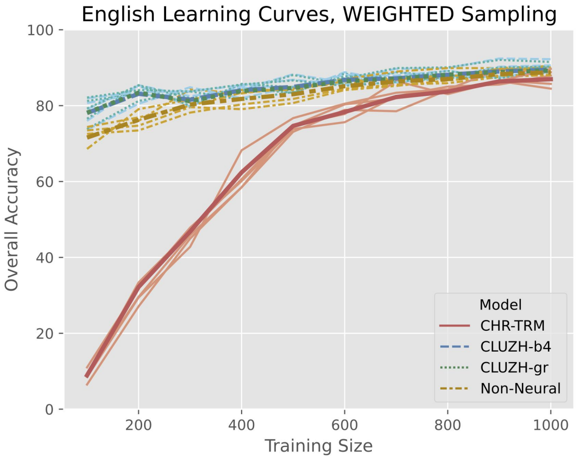
Arabic



German



English



Thin/light lines = individual seeds Bold/dark lines = averages across seeds

- Non-Neural underperforms on Arabic
- CHR-TRM underperforms on small data
- Noticeable but minor variability across seeds

Evaluating English Over-Regularization (SIGMORPHON'22)

What do systems do with the large-ish class of verbs ending in *-ing*?

- **The goal here is not to make correct predictions, but human-like predictions**
- **Do they over-regularize (→ *-ed*)**
- **Or over-irregularize (analogy with irregulars)**

In the training set

```
swing-swung
sing-sang
thing-thinged
ding-dinged
sling-slung
cling-clung
```

In the gold test set

```
sting-stung      fling-flung
ring-rang        ping-pinged
bring-brought    king-kinged
spring-sprang    string-strung
```

Evaluating English Over-Regularization (SIGMORPHON'22)

What do systems do with the large-ish class of verbs ending in *-ing*?

- **The goal here is not to make correct predictions, but human-like predictions**
- **Do they over-regularize (→ *-ed*)**
- **Or over-irregularize (analogy with irregulars)**

System	*-ed*	*-ang*	*-ung*	Other
(Gold)	2	2	3	1
CLUZH				
HeiMorph				
OSU				

Evaluating English Over-Regularization (SIGMORPHON'22)

What do systems do with the large-ish class of verbs ending in *-ing*?

- The goal here is not to make correct predictions, but human-like predictions
- Do they over-regularize (→ *-ed*)
- Or over-irregularize (analogy with irregulars)

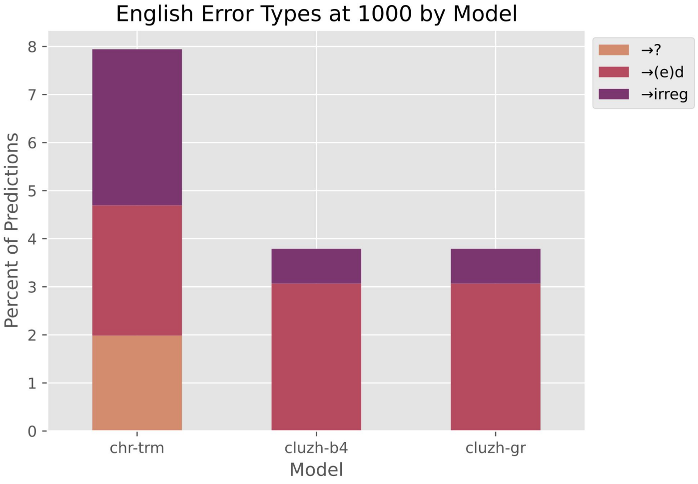
System	*-ed*	*-ang*	*-ung*	Other
(Gold)	2	2	3	1
CLUZH	4	1	3	0
HeiMorph	8	0	0	0
OSU	8	0	0	0

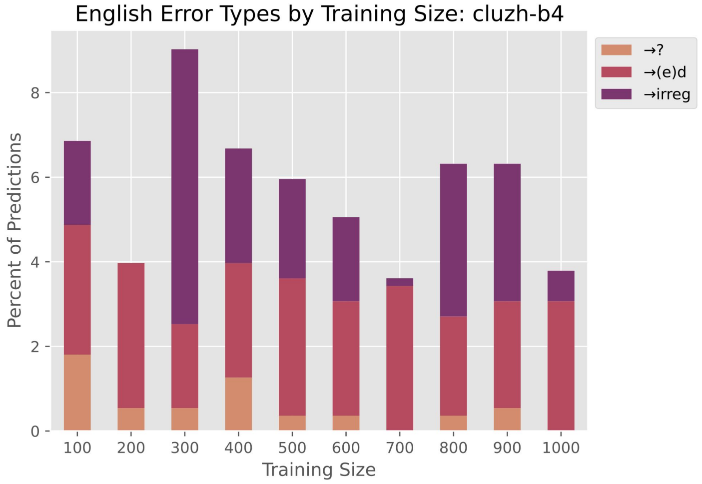
Over-regularization dominates, but CLUZH also over-irregularizes

Evaluating English Over-Regularization (CogSci'23)

What do systems do more broadly?

- **Evaluated on manually annotated gold and prediction data**
- **All systems over-irregularize proportionately far more than child learners**
- **No system shows a *u*-shaped learning pattern**





Evaluating Productivity in German (SIGMORPHON'22)

Distribution of plural suffixes is similar in train and test

- Both overall and by-gender

Set	%-*e*	%-(*e*)*n*	%-*er*	%-∅	%-*s*	#
Train	27.8%	38.5%	3.0%	26.7%	4.6%	600
Train F	2.8	96.2	0.0	0.5	0.5	212
Train M	45.4	7.3	1.5	41.2	4.5	262
Train N	33.3	4.0	11.1	40.5	11.1	126
Test	30.5%	36.7%	2.8%	24.8%	5.2%	600
Test F	3.5	95.0	0.0	0.0	1.5	201
Test M	48.9	9.2	0.3	35.9	5.6	284
Test N	32.2	2.6	13.9	40.9	10.4	115

Evaluating Productivity in German (SIGMORPHON'22)

Systems probability match

- **Gold (G)** - Prediction (P) confusion matrices by model
- All systems probability match but slightly prefer -∅
- ? indicates nonsense predictions

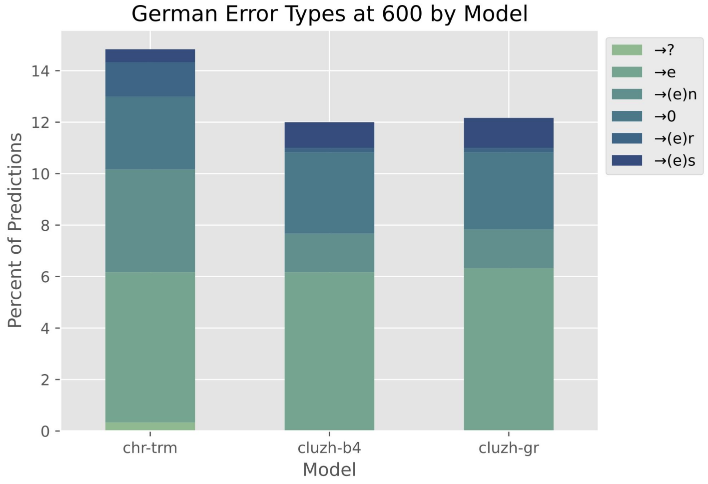
CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	168	16	13	0	18	215
P -(*e*)*n*	6	198	0	1	2	207
P -*er*	0	0	3	0	0	3
P -∅	8	5	0	148	0	161
P -*s*	1	1	1	0	11	14
P ?	0	0	0	0	0	0
Sum	183	220	17	149	31	600

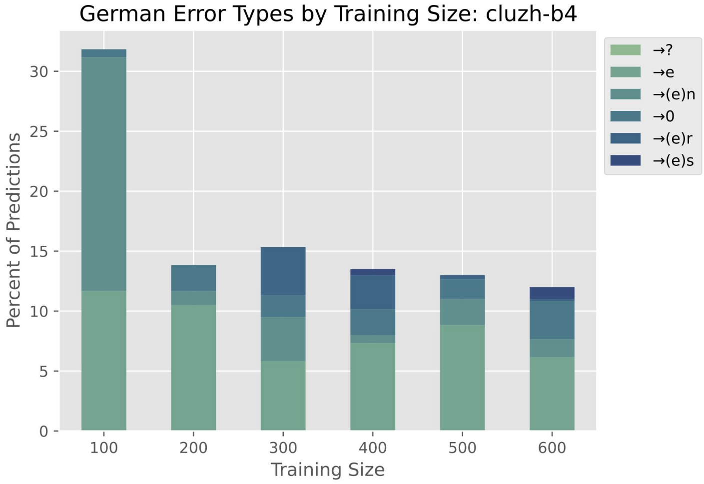
HeiMor	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	154	12	12	4	16	199
P -(*e*)*n*	14	194	0	0	4	212
P -*er*	4	0	4	1	4	13
P -∅	9	10	0	142	1	162
P -*s*	1	1	1	0	3	6
P ?	1	2	0	2	3	8
Sum	183	220	17	149	31	600

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	155	19	13	1	18	206
P -(*e*)*n*	7	184	0	0	2	193
P -*er*	2	0	3	1	0	6
P -∅	11	10	1	142	1	165
P -*s*	2	1	0	1	8	12
P ?	6	6	0	4	2	18
Sum	183	220	17	149	31	600

Evaluating Productivity in German (CogSci'23)

- Half of errors were over-application of -*e* for all systems
- Some over-application of -*s* is present for all systems on the full training set
- Other than -*e*, error distribution is unstable over time for CLUZH-b4
- Early over-application of -*e* is encouraging





Evaluating Productivity in Arabic (SIGMORPHON'22)

Distribution of plural patterns differs in train and test

- Broken down by gender and rationality

Set	SFem	SMasc	Brokn	Sum
Train	424	140	140	998
Train F	222	0	85	307
Train M	202	140	349	691
Train H	24	129	84	237
Train NH	400	11	350	761
Test	257	62	281	600
Test F	156	0	73	229
Test M	101	62	208	371
Test H	15	50	43	108
Test NH	242	12	238	492

Evaluating Productivity in Arabic (SIGMORPHON’22)

Systems prefer Sound Feminines

- Gold (G) - Prediction (P) confusion matrices by model
- Preference for sound feminine matches developmental findings
- ? indicates nonsense productions

CLUZH	G SF	G *SM*	G B	Sum
P SF	213	5	52	270
P SM	2	51	16	69
P B	38	4	206	248
P ?	4	2	7	13
Sum	257	62	281	600

HeiMor	G SF	G *SM*	G B	Sum
P SF	227	7	72	306
P SM	3	43	15	61
P B	18	5	177	200
P ?	9	7	17	33
Sum	257	62	281	600

OSU	G SF	G *SM*	G B	Sum
P SF	218	8	49	275
P SM	5	50	15	70
P B	29	2	202	233
P ?	5	2	15	22
Sum	257	62	281	600

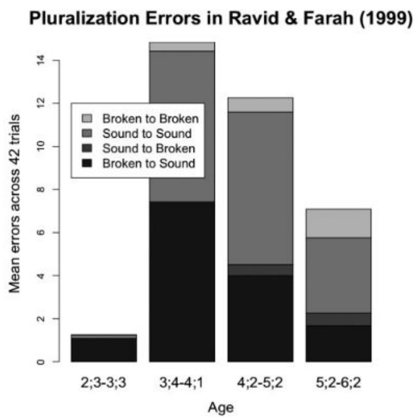
Evaluating Productivity in Arabic (SIGMORPHON'22)

Comparison with Developmental Literature

- Sound→Sound and Broken→Sound errors dominate developmentally
- But each system prefers Broken→Sound and Broken→Broken
- →Broken are over-irregularizations

Consistent with other "single-route" systems that rely on analogy

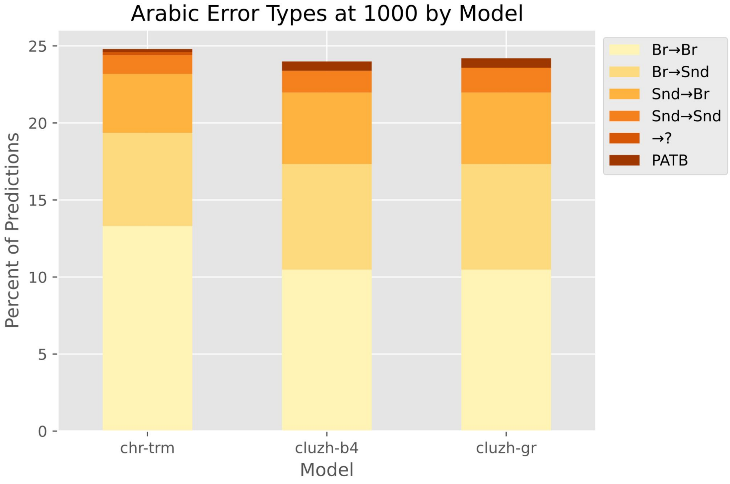
Set	So→So	So→Br	Br→So	Br→Br
CLUZH	7	42	68	52
HeiMor	10	23	87	65
OSU	13	31	64	57

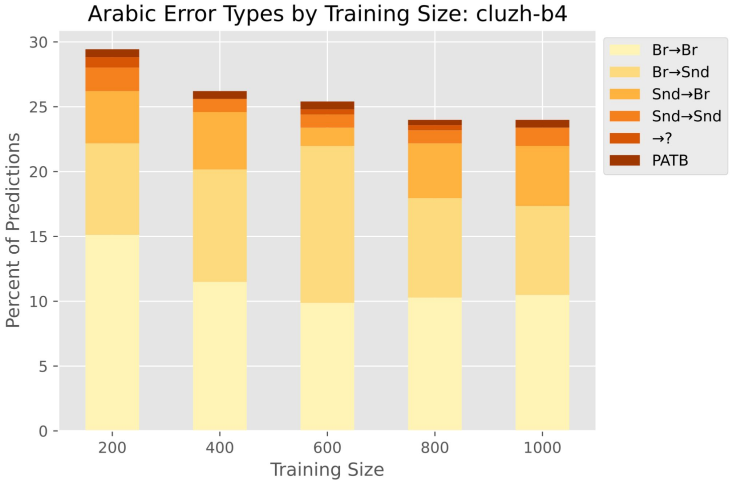


Evaluating Productivity in Arabic (CogSci'23)

Consistent with analysis from SIGMORPHON'22

- Sound→Sound and Broken→Sound errors dominate developmentally
- But each system prefers Broken→Sound and Broken→Broken
- No clear *u*-shaped learning





SIGMORPHON’22 and CogSci’23

Main Conclusions

- Performance on English > German > Arabic reflects pattern complexity
- Overall accuracy is pretty good!
  Especially considering the very low training sizes
- But error patterns are not human-like
  Heavily biased toward probability matching
  Far too much over-irregularization
  No *u*-shaped learning in English or Arabic

Such models are clearly not human-like

→ unlikely to be informative about language acquisition

Final Conclusions

1. Traditionally taken to be useful in downstream tasks
   - Maybe, but generalization to OOV feature sets is a weakness, particularly for the languages that inflection would be useful for
2. May provide insight into the behavior of NN architectures
3. May elucidate aspects of linguistic typology
4. May elucidate aspects of language acquisition

Final Conclusions

1. Traditionally taken to be useful in downstream tasks
   - Maybe, but generalization to OOV feature sets is a weakness, particularly for the languages that inflection would be useful for
2. May provide insight into the behavior of NN architectures
   - Yes, but care needs to be taken to differentiate impact of data design decisions from the systems being investigated
3. May elucidate aspects of linguistic typology
4. May elucidate aspects of language acquisition

Final Conclusions

1. Traditionally taken to be useful in downstream tasks
   - **Maybe,** but generalization to OOV feature sets is a weakness, particularly for the languages that inflection would be useful for
2. May provide insight into the behavior of NN architectures
   - **Yes,** but care needs to be taken to differentiate impact of data design decisions from the systems being investigated
3. May elucidate aspects of linguistic typology
   - **Probably not.** We find that current leading systems are hardly impacted by typology
4. May elucidate aspects of language acquisition

Final Conclusions

1. **Traditionally taken to be useful in downstream tasks**
   - **Maybe,** but generalization to OOV feature sets is a weakness, particularly for the languages that inflection would be useful for
2. **May provide insight into the behavior of NN architectures**
   - **Yes,** but care needs to be taken to differentiate impact of data design decisions from the systems being investigated
3. **May elucidate aspects of linguistic typology**
   - **Probably not.** We find that current leading systems are hardly impacted by typology
4. **May elucidate aspects of language acquisition**
   - **Probably not.** We find that current leading systems do not behave like humans.
     → **They are unlikely to be good models for acquisition.**

Special Thanks to:
Jeff Heinz, Charles Yang,
Contributors and participants
in the SIGMORPHON 2022 shared tasks

Thank you!