SIGMORPHON-UniMorph 2022 Shared Task 0: Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation

Jordan Kodner, Salam Khalifa *et xxviii al.*

https://aclanthology.org/2022.sigmorphon-1.18/
https://aclanthology.org/2022.sigmorphon-1.19/

SIGMORPHON 2022
Seattle, July 14, 2022

Two Subtasks

Generalization and Typologically Diverse Morphological Inflection

- 33 languages from 10 families
- Large and small training sets
- Iteration on the “classic” inflection task
- Focused on two dimensions of generalizations:
  1) Over lemmas
  2) Over feature sets

Modeling Inflection in Language Acquisition

- How do learning trajectories for automatic systems compare to children’s learning trajectories?
- Three classic languages/patterns
  1) English past tense
  2) German noun plurals
  3) Arabic noun plurals

Subtask 1: Languages

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
  - Gujarati
  - Kholosi
  - Magahi
- Slavic
  - Polish
  - Pomak
  - Slovak
  - Upper Sorbian

Subtask 1: Four types of test (lemma, features) pairs

Sample training

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

Sample test

eat	V;PST	(both)
run	V;NFIN	(lemma)
see	V;PST	(features)
go	V;PRS;3;SG	(neither)

Both	lemma and feature set attested in training (not together)
Lemma	only lemma in training
Features	only feature set in training
Neither	neither lemma nor feature set in training

Subtask 1: Four types of test (lemma, features) pairs

Sample training

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

Sample test

eat	V;PST	(both)
run	V;NFIN	(lemma)
see	V;PST	(features)
go	V;PRS;3;SG	(neither)

Both	lemma and feature set attested in training (not together)
Lemma	only lemma in training
Features	only feature set in training
Neither	neither lemma nor feature set in training

Not controlled for in previous iterations

Subtask 1: Systems

CLUZH	Clematide, Wehrli, & Makarov
Flexica*	Scherbakov & Vylomova
OSU	Elsner & Court
TüMorph-FST	Merzhevich, Gbadegoye, Girrbach, Li, & Shim
TüMorph-Main	" " " " & "
UBC*	Yang, Yang, Nicolai, & Silfverberg
NeurBase	same as 2021
NonNeurBase	same as 2021

*Submitted after deadline

Subtask 1: Systems

CLUZH	Clematide, Wehrli, & Makarov
Flexica*	Scherbakov & Vylomova
OSU	Elsner & Court
TüMorph-FST	Merzhevich, Gbadegoye, Girrbach, Li, & Shim
TüMorph-Main	" " " " & "
UBC*	Yang, Yang, Nicolai, & Silfverberg
NeurBase	same as 2021
NonNeurBase	same as 2021

Baselines

*Submitted after deadline

Subtask 1: Systems

CLUZH	Clematide, Wehrli, & Makarov
Flexica*	Scherbakov & Vylomova
OSU	Elsner & Court
TüMorph-FST	Merzhevich, Gbadegoye, Girrbach, Li, & Shim
TüMorph-Main	" " " " & "
UBC*	Yang, Yang, Nicolai, & Silfverberg
NeurBase	same as 2021
NonNeurBase	same as 2021

Non-neural

*Submitted after deadline

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

Subtask 1: Summary Results

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

All systems perform much better when test item feature sets are seen than when they are novel

Subtask 1: Summary Results

True even for agglutinative languages

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

Different strengths?
CLUZH outperforms when feat sets are seen
but UBC outperforms when they are novel

Subtask 1: Summary Results

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

Subtask 1: Seen vs Unseen on Agglutinative Langs

- Exponence of a feature set is (at least largely) predictable from individual features
  → Generalization should be possible
  "Could an undergrad do it?"
- Chukchi, Evenki, Georgian, Hungarian, Itelmen, Karelian, Kazakh, Ket, Korean, Ludic, Mongolian, Turkish, Veps, and Xibe

Features	Small		Large	
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)

Subtask 1: Conclusions

- Systems consistently generalize to new lemmas
  better than to unseen feature sets,
  even when generalization to unseen feature sets should be feasible
- Systems vary in their relative ability to perform each generalization

→ Reported performance (and rankings) are sensitive to these overlaps in data splits

→ Gains are yet to be had for languages with large paradigms

Subtask 2: Human-like?

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

- Data was extracted from child-directed corpora within CHILDES when possible
- Small training sets of high frequency items were provided
  in line with computational literature on language acquisition
- Three heavily studied morphological patterns were chosen

Subtask 2: Morphological Patterns

Three well-studied patterns in the (computational-)acquisition literature

English Past Tense

- Default *-ed*
  overwhelming majority
- Plenty of high freq
  irregular verbs
  sing-sang
  sting-stung
  go-went…

German Noun Plurals

- Several regular patterns
- Phonological and gender conditioning
- "Minority default" *-s*
  "Pattern of last resort"
- Frequency-matching won't work well

Arabic Noun Plurals

- Two types
  1) Suffixed "sound" plurals
     Masc *-ūn*, Fem *-āt*
  2) Stem changing "broken" pl
     Dozens of patterns

Subtask 2: Systems

CLUZH	Clematide, Wehrli, & Makarov
HeiMorph	Ramarao, Zinova, Tang & van de Vijver
OSU	Elsner & Court
NeurBase	same as 2021
NonNeurBase	same as 2021

Subtask 2: Systems

CLUZH	Clematide, Wehrli, & Makarov	Same system as Subtask 1
HeiMorph	Ramarao, Zinova, Tang & van de Vijver	
OSU	Elsner & Court	Same system as Subtask 1
NeurBase	same as 2021	Same system as Subtask 1
NonNeurBase	same as 2021	Same system as Subtask 1

Subtask 2: Summary Results

	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

Subtask 2: Summary Results

	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

Subtask 2: Summary Results

	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

→ Performance decreases as pattern complexity increases →

Subtask 2: English *-ing* Verbs

In natural child speech, over-reguarlization errors (→ *-ed*) are overwhelmingly more common than over-irregularization errors (analogy with irregulars)

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

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
```

Subtask 2: English *-ing* Verbs

In natural child speech, over-reguarlization errors (→ *-ed*) are overwhelmingly more common than over-irregularization errors (analogy with irregulars)

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

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

Subtask 2: English *-ing* Verbs

In natural child speech, over-reguarlization errors (→ *-ed*) are overwhelmingly more common than over-irregularization errors (analogy with irregulars)

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

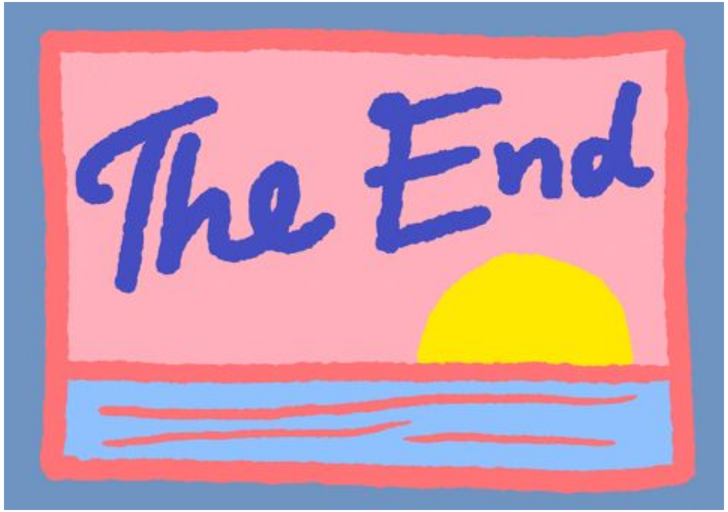
System	*-ed*	*-ang*	*-ung*	Other
(Gold)	2	2	3	1
CLUZH	4	1	3	0
HeiMorph	8	0	0	0
OSU	8	0	0	0

Over-regularization dominates, but CLUZH also over-irregularizes

The situation is not as rosy for German or Arabic. See the paper

Subtask 2: Conclusions

- Performance is generally good in quantitative terms, but there is room for improvement
- Errors are not particularly human-like but share some commonalities



Now, the system presentations

Co-Authors:

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkuş, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, Ekaterina Vylomova

And Additional Thanks:

Jeff Heinz, Charles Yang, Ellen Broselow, Garrett Nicolai, Maria Ryskina, Ben Ambridge