MORPHOLOGICAL INFLECTION: A REALITY CHECK

JORDAN KODNER[1]
SARAH PAYNE[1*]
SALAM KHALIFA[1*]
ZOEY LIU[2]

[1]{first.last}@stonybrook.edu
[2]liu.ying@ufl.edu

Stony Brook University
UF UNIVERSITY of FLORIDA

MORPHOLOGICAL INFLECTION

Patterns of word formation which express grammatical categories

- Processes vary dramatically across languages: pre/in/circum/suffixation, stem mutation, reduplication…
- So do which grammatical categories are marked: number, tense, mood, voice, aspect, evidentiality, possession, case…

INFLECTION AS AN NLP TASK

TRAIN: given (lemma, infl. form, feat. set) triples

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

TEST: predict inflected forms from (lemma, feat. set) pairs

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

THREE OVERSIGHTS IN PRIOR WORK

1) **UNIFORM SAMPLING** creates an unnatural bias towards "easier" low-frequency regular types. **We propose naturalistic frequency WEIGHTED sampling or controlled OVERLAPAWARE sampling to balance OOV lemmas and feature sets in the evaluation data.**
2) **SINGLE DATA SPLITS** hide variability intrinsic to sampling from corpora and assumes the generalizability and informativity of test results. **We propose sampling with several random seeds and measuring variability.**
3) **UNCONTROLLED OVERLAPS** between lemmas and feature sets independently in train and test obscure the contributions of the language, model, and corpus on performance. **We propose controlling for lemma and feature set overlap.**

TYPES OF TRAIN–TEST OVERLAP

FOUR LICIT TYPES OF OVERLAP

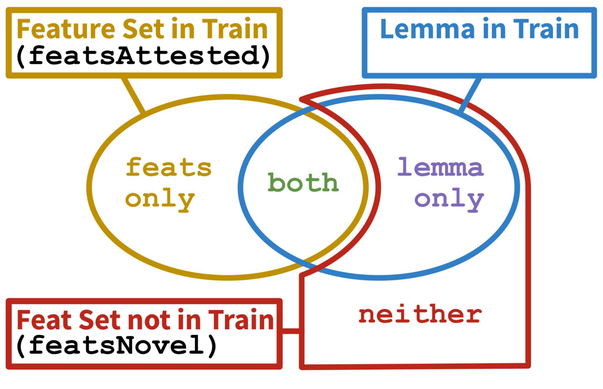
Since lemmas and feature sets can be combined, there are four distinct types of licit test item.

ILLUSTRATIVE TRAINING SET

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

ILLUSTRATIVE TEST SET

```
eat   V;PST          ← No OOV
run   V;NFIN         ← Only feature set is OOV
see   V;PST          ← Only lemma is OOV
go    V;PRS;3;SG     ← Both are OOV
```



Visualization of overlap types. We predicted that `featsNovel` would prove more challenging than `featsAttested`.

CONSEQUENCES OF THE DATA SAMPLING STRATEGY

SAMPLING STRATEGIES

- **UNIFORM** - random sampling
- **WEIGHTED** - frequency-weighted random sampling
- **OVERLAPAWARE** - makes sure that ~50% of test items have OOV feature sets

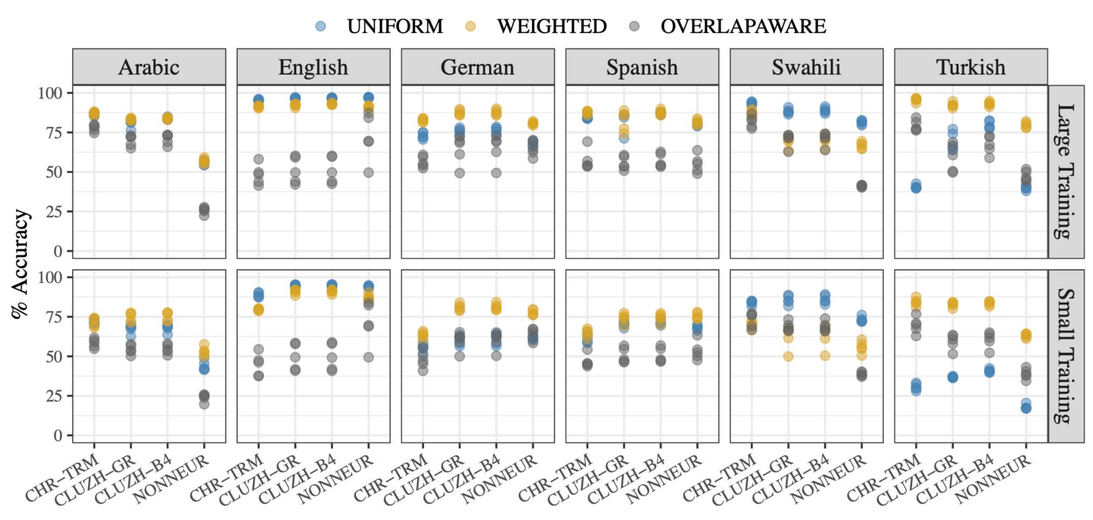
Test vs S Train	μ %featsAttested	σ
UNIFORM	80.33%	19.50%
WEIGHTED	90.44	11.13
OVERLAPAWARE	48.81	0.98
Test vs L Train	**μ %featsAttested**	**σ**
UNIFORM	96.17%	5.55%
WEIGHTED	95.36	7.28
OVERLAPAWARE	49.92	0.17

SYSTEMS

Drawn from SIGMORPHON 2022 Shared Task

- **CHR-TRM** (Wu et al., 2021): a character transformer
- **CLUZH** (Wehrli et al., 2022): a character transducer. **GR** = greedy, **B4** = beam size 4 decoding
- **NONNEUR:** non-neural baseline

Average distribution of featsAttested items in the test set



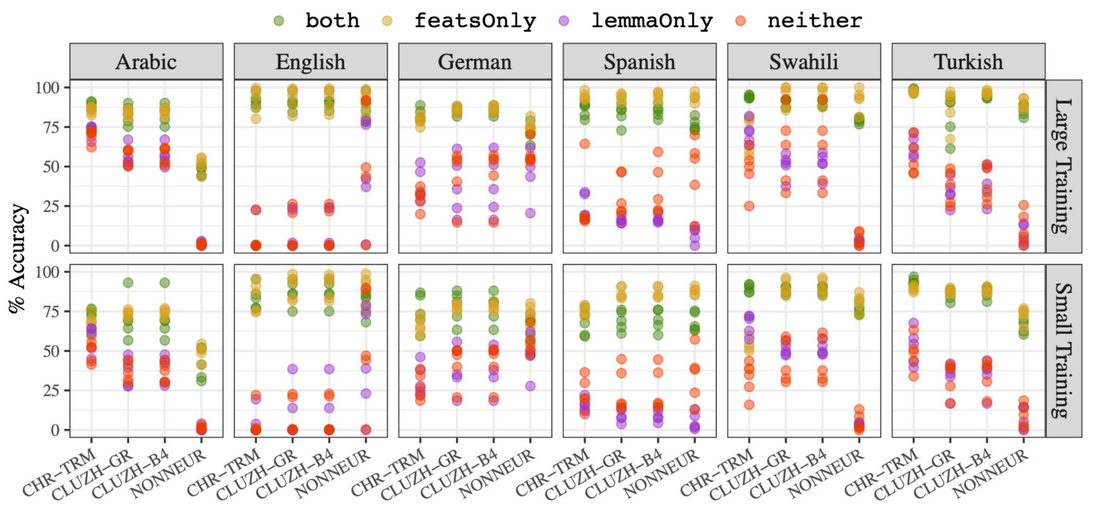
RESULTS

- Some `featsNovel` items are present in test regardless of sampling strategy, but **OVERLAPAWARE** yields the most `featsNovel` and most consistent rate across languages and seeds
- Performance is generally lowest on **OVERLAPAWARE** (due to the large number of `featsNovel` items)
- Ranking of **UNIFORM** and **WEIGHTED** performance depends more on language than model or training size
- However, variability across seeds is highest for **OVERLAPAWARE**. This suggests that it matters which feature sets are in `featsNovel` vs `featsAttested`

CONSEQUENCES OF TEST ITEM OVERLAP TYPES

RESULTS

- **Performance is >50% lower** on `featsNovel` (●●) vs `featsAttested` (●●), irrespective of training size and for each system
- No consistent drop for OOV lemmas (●●) vs attested lemmas (●●)
- Wide variability across seeds



TYPOLOGY AND GENERALIZATION

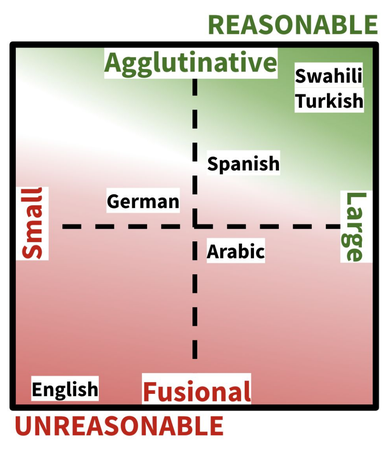
IS GENERALIZATION TO UNSEEN FEATURE SETS A REASONABLE EXPECTATION?

PARADIGM SIZE

+ **Large paradigms** → OOV feature sets likely
- **Small paradigms** → OOV feature sets unlikely

AGGLUTINATIVITY

+ **Agglutinative** → inflection of feature set derivable from inflections of individual features
- **Fusional** → inflection of feature set not derivable from individual features



Train Size	Language Strategy	Avg. Score Difference
Small	Arabic	33.00%
	Swahili	40.04
	German	40.35
	Turkish	41.96
	Spanish	52.60
	English	74.10
Large	Arabic	35.79%
	German	36.19
	Swahili	39.26
	Turkish	52.14
	Spanish	61.01
	English	80.17

If systems effectively generalized to novel feature sets, Avg. Score Difference between `featsNovel` and `featsAttested` subsets would be lowest for agglutinative Swahili and Turkish

RESULTS

- For all systems, **generalization to unseen feature sets proves challenging even for agglutinative languages** (Swahili and Turkish) where this should be possible
- Suggests unresponsiveness to morphological typology
- And identifies an area of work for future improvement

LINK TO PAPER