Morphological Inflection: A Reality Check

Sarah Payne & Salam Khalifa

Stony Brook University

first.last@stonybrook.edu

Outline

- What is morphological inflection?
- Evaluation complexities
- Setup
- Findings & Implications

What is Morphological Inflection?

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **English:** *walk*+PAST → *walked*
  - **Mandarin:** 3+PL → *tāmen* 'they'
  - **Hebrew:** √*ħtl*+DIM+SG+DEF → *haħataltul* 'the kitty'
  - **Latin:** *amic*+FEM+SG+GEN → *amīcae* 'the friend's'
  - **Shona:** *bik*+1SG.SUBJ+6CL.OBJ+PAST+CAUS+PASS→ *ndakachibikiswa* 'I was made to cook it'

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **Roots/stems** modified by many processes
    - **Suffixation/prefixation/circumfixation, stem mutations, reduplication**

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **Roots/stems** modified by many processes
    - **Suffixation/prefixation/circumfixation, stem mutations, reduplication**
  - Express **number, tense, mood, voice, aspect, evidentiality, …**

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **Roots/stems** modified by many processes
    - **Suffixation/prefixation/circumfixation,** **stem mutations,** **reduplication**
  - Express **number, tense, mood, voice, aspect, evidentiality, …**
  - Common across the world's languages
    - Vary dramatically in terms of **complexity** or **"richness"**

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **Roots/stems** modified by many processes
    - **Suffixation/prefixation/circumfixation, stem mutations, reduplication**
  - Express **number, tense, mood, voice, aspect, evidentiality, …**
  - Common across the world's languages
    - Vary dramatically in terms of **complexity** or **"richness"**
  - Poses a learning challenge for both **machines** and **humans**

Morphological Inflection as an NLP task

- **Training:** (`lemma`, `inflected form`, `feature set`)

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

Morphological Inflection as an NLP task

- **Training:** (`lemma`, `inflected form`, `feature set`)

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

- **Testing:** (`lemma`, `feature set`) → `inflected form`

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

Morphological Inflection as an NLP task

- **Training:** (`lemma`, `inflected form`, `feature set`)

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

- **Testing:** (`lemma`, `feature set`) → `inflected form`

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

Morphological Inflection: Applications

Cognitive Modeling

- Insight into the **cognitive computations** underlying morphological learning
- **Past Tense Debate**
  - Early **connectionist** account (Rumelhart & McClelland 1986)
  - Several shortcomings
- Recent advances in **ANN architectures**
  - Renewed interest in the plausibility of ANNs as **cognitive models**

Natural Language Processing

- **Traditionally:** downstream tasks
  - In settings where **pipelining** is still common (e.g., **low-resource**)
  - Particularly for languages with **lots of inflectional morphology**
- May provide insight into the behavior of **ANN architectures**
  - A particular kind of **string-to-string mapping** problem
  - Varying performance may reflect **divergent properties** of different architectures

Morphological Inflection: Solved?

- **Kirov & Cotterell (2018):** encoder-decoder network can **overcome practical limitations** of older ANNs
  - Near **100% test accuracy**
  - Learn **several inflectional classes** at once

Morphological Inflection: Solved?

- **Kirov & Cotterell (2018):** encoder-decoder network can **overcome practical limitations** of older ANNs
  - Near **100% test accuracy**
  - Learn **several inflectional classes** at once
- **Corkerey et al. (2019):** K&C model still fails empirically
  - Predictions **don’t match well** with human nonce word judgments: **over-irregularizes!**
  - Massive **variability in model rankings & correlation with human rankings** between seeds

Morphological Inflection: Solved?

Best systems on a subset of the 2018 **CoNLL-SIGMORPHON** shared task

	High	Medium	Low
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)

Very good performance on medium and high training

Morphological Inflection: Solved?

🚨 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)

Morphological Inflection isn't solved! 🚨🚨🚨

Evaluation Complexities

Morphological Inflection: Outstanding Issues

- NNs are trained on **unrealistically large/saturated data**
- NNs are rarely evaluated against **child learning trajectories** and **error patterns**

Belth, Payne et al. (2021, Cogsci)
Kodner, Payne et al. (2023, ACL)
Kodner, Khalifa, Payne, & Liu (2023, Cogsci)

- Current evaluation metrics fail to control for:
  - **Overlap** between train and test
  - Performance **variation** across multiple splits
  - **Frequency effects** in uniform sampling

Kodner, Payne et al. (2023, ACL)
Kodner, Khalifa & Payne (2023, EMNLP)

Common Evaluation Practices

- **Uniform** sampling & **large training** sets
    - Training and evaluation sets sampled uniformly by type from a corpus. Usually large corpora.

Common Evaluation Practices

- **Uniform** sampling & **large training** sets
  - Training and evaluation sets sampled uniformly by type from a corpus. Usually large corpora.
  - Practical, **but** creates an unnatural bias towards low-frequency types which are mostly regular.

Common Evaluation Practices

- **Uniform** sampling & **large training** sets
  - Training and evaluation sets sampled uniformly by type from a corpus. Usually large corpora.
  - Practical, **but** creates an unnatural bias towards low-frequency types which are mostly regular.

Contribution

- A frequency weighted sampling to match practical use-cases during child language acquisition (aka true low-resource).
- A sampling strategy that balances OOV lemmas and features to evaluate models’ generalizability.

Common Evaluation Practices

- Evaluation on **single splits**:

Common Evaluation Practices

- Evaluation on **single splits**: a single set of Train+Dev+Test splits.

Common Evaluation Practices

- Evaluation on **single splits**: a single set of Train+Dev+Test splits.
  - Problematic because:
  - Results based on a single set aren’t informative in low-res settings.
  - Even in high-res, a set sampled using a different seed results in a different performance.

Common Evaluation Practices

- Evaluation on **single splits**: a single set of Train+Dev+Test splits.
  - Problematic because:
  - Results based on a single set aren't informative in low-res settings.
  - Even in high-res, a set sampled using a different seed results in a different performance.

Contribution

- Use multiple sets of splits
- Use variable data sizes.

Common Evaluation Practices

Uncontrolled overlap between train & test components

- Two types: lemma overlap and feature overlap.
- It hinders the evaluation of generalizability due to the uncontrolled OOV rates.

Common Evaluation Practices

Uncontrolled overlap between train & test components

- Two types: lemma overlap and feature overlap.
- It hinders the evaluation of generalizability due to the uncontrolled OOV rates.

Contribution

- Control for both types of overlap regardless of the split and sampling technique.

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

Illustrative Train Set

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

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

Illustrative Train Set

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

Illustrative Test Set

eat	V;PST	← **No OOV,** not attested together

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

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

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

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

Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

Illustrative Train Set

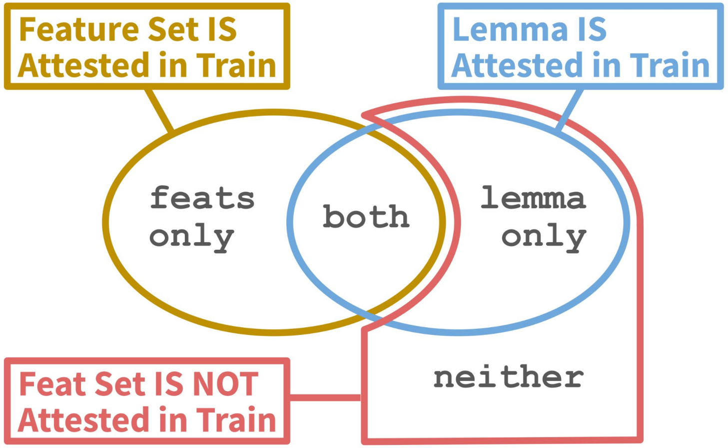
eat	eating	V;V.PTCP;PRS
run	ran	V;PST

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

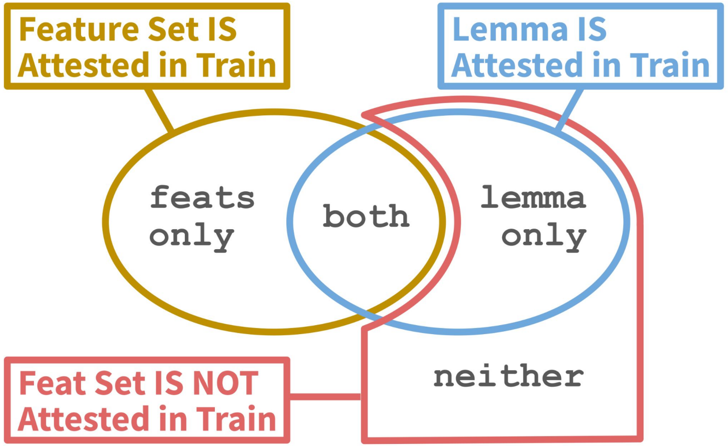
Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:



Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:



eat	V;PST	← both
run	V;NFIN	← lemmaOnly
see	V;PST	← featsOnly
go	V;PRS;3;SG	← neither

Setup

Setup

- 6 Languages
- 3 Split Types
- 4 Systems
- 5 Random Seeds

Setup: Languages

- **6 Languages:** English, Arabic, German, Spanish, Swahili, Turkish

  increasingly agglutinative →

  - **UniMorph 3 + 4** intersected with **frequency information** for weighted sampling

CHILDES	**Wikipedia**	**PATB**
German, English, Spanish	Swahili & Turkish	Arabic

- 3 Split Types
- 4 Systems
- 5 Random Seeds

Setup: Split Types

- **6 Languages:** English, Arabic, German, Spanish, Swahili, Turkish
- **3 Split Types:**
  - **UNIFORM:** partition UniMorph **uniformly at random**
  - **WEIGHTED:** partition at random weighted by **type frequency**
  - **OVERLAPAWARE:** enforce a maximum **50% proportion of FEATSATTESTED**
- 4 Systems
- 5 Random Seeds

Setup: Systems

- **6 Languages:** English, Arabic, German, Spanish, Swahili, Turkish
- **3 Split Types:** Uniform, Weighted, and OverlapAware
- **4 Systems:**
  - **CLUZH-B4:** character-level **transducer** that significantly outperformed the 2022 SIGMORPHON baseline, with **beam decoding**
  - **CLUZH-GR:** character-level **transducer** with **greedy decoding**
  - **CHR-TRM:** character-level **transformer** that was used as a baseline in 2021 and 2022 SIGMORPHON shared tasks
  - **NonNeur:** non-neural baseline using a **majority classifier**
- 5 Random Seeds

Setup: Random Seeds

- **6 Languages:** English, Arabic, German, Spanish, Swahili, Turkish
- **3 Split Types:** Uniform, Weighted, and OverlapAware
- **4 Systems:** CLUZH-B4, CLUZH-GR, CHR-TRM, NonNeur
- **5 Random Seeds** for re-splitting and re-evaluation

Feature Overlap in Training

	SmallTrain	featsAttested	featsNovel	σ
400 train 100 ftune 1000 test	Uniform	80.33	19.67	19.5
	Weighted	90.44	9.56	11.1
	OverlapAware	48.81	51.19	0.98
	LargeTrain	**featsAttested**	**featsNovel**	**σ**
1600 train 400 ftune 1000 test	Uniform	96.17	3.83	5.55
	Weighted	95.36	4.64	7.28
	OverlapAware	49.92	50.08	0.17

Feature Overlap in Training

SmallTrain	featsAttested	featsNovel	σ
Uniform	80.33	19.67	19.5
Weighted	90.44	9.56	11.1
OverlapAware	48.81	51.19	0.98

400 train
100 ftune
1000 test

LargeTrain	featsAttested	featsNovel	σ
Uniform	96.17	3.83	5.55
Weighted	95.36	4.64	7.28
OverlapAware	49.92	50.08	0.17

1600 train
400 ftune
1000 test

Overlap rate is high but not 100% when not controlled for
UNIFORM & WEIGHTED are similar for large training size

Feature Overlap in Training

	SmallTrain	featsAttested	featsNovel	σ
400 train 100 ftune 1000 test	Uniform	80.33	19.67	19.5
	Weighted	90.44	9.56	11.1
	OverlapAware	48.81	51.19	0.98
	LargeTrain	**featsAttested**	**featsNovel**	**σ**
1600 train 400 ftune 1000 test	Uniform	96.17	3.83	5.55
	Weighted	95.36	4.64	7.28
	OverlapAware	49.92	50.08	0.17

Overlap rate is highly variable across seed/language when not controlled for

Findings & Implications

Evaluation

We evaluated across several dimensions:

- Training set size
- Sampling Strategy
- Overlap awareness

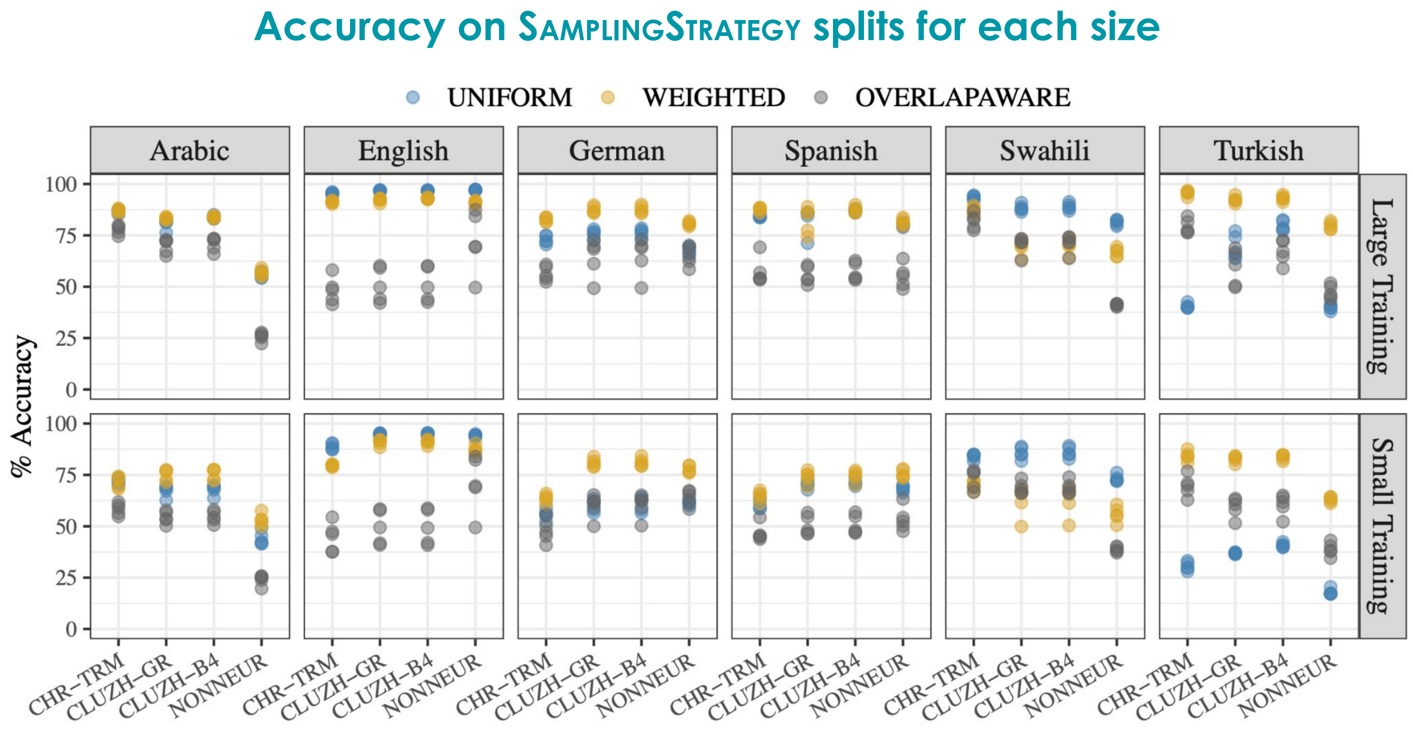
Evaluation

We evaluated across several dimensions:

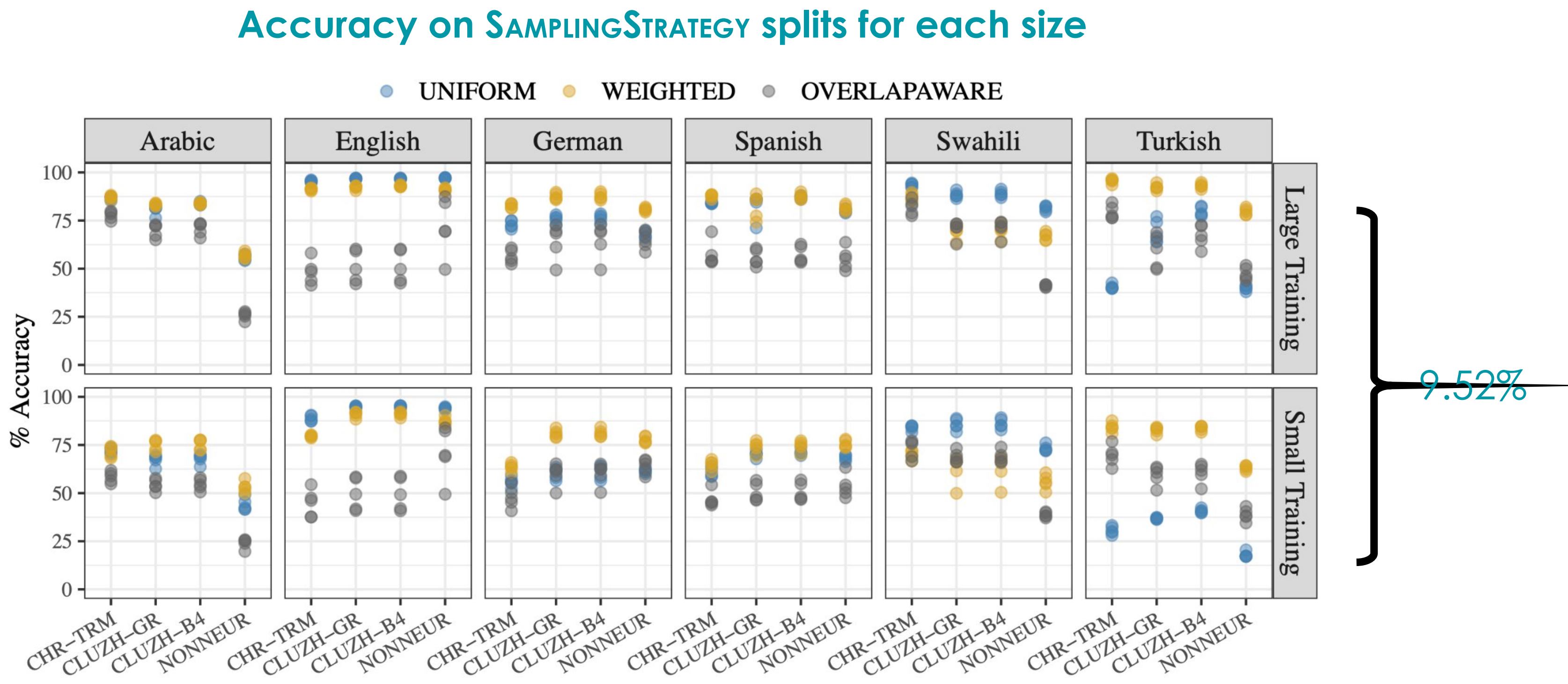
- Training set size
- Sampling Strategy
- Overlap awareness

All reported accuracies are averaged across 5 splitting seeds per language.

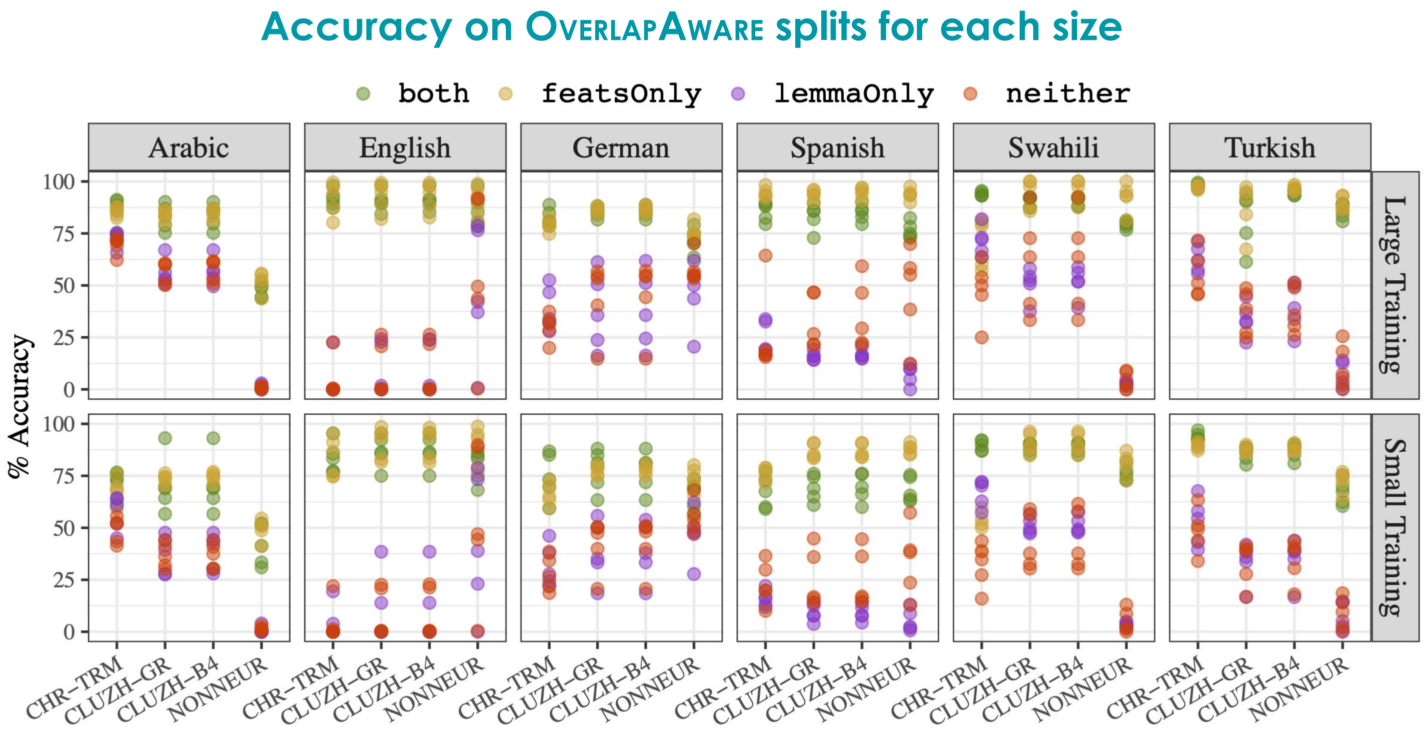
Results: Effect of Sampling Strategy



Results: Effect of Sampling Strategy

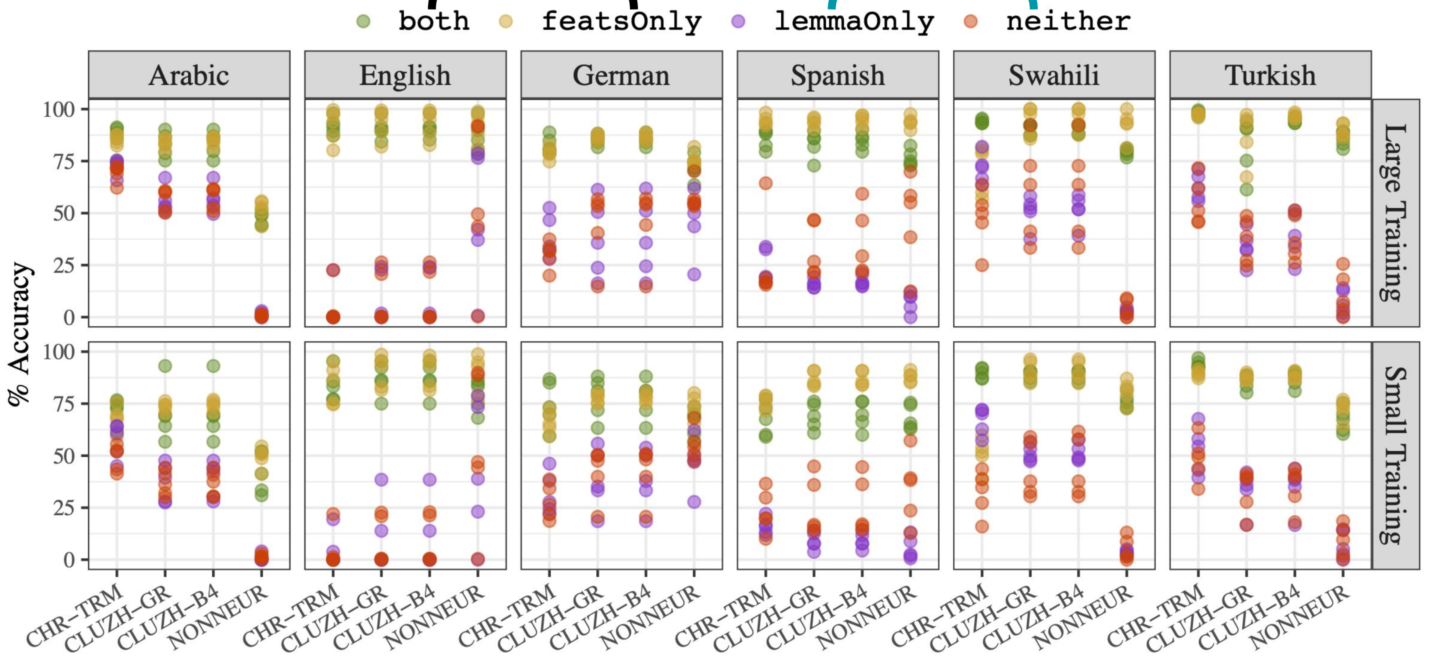


Results: Effect of Overlap



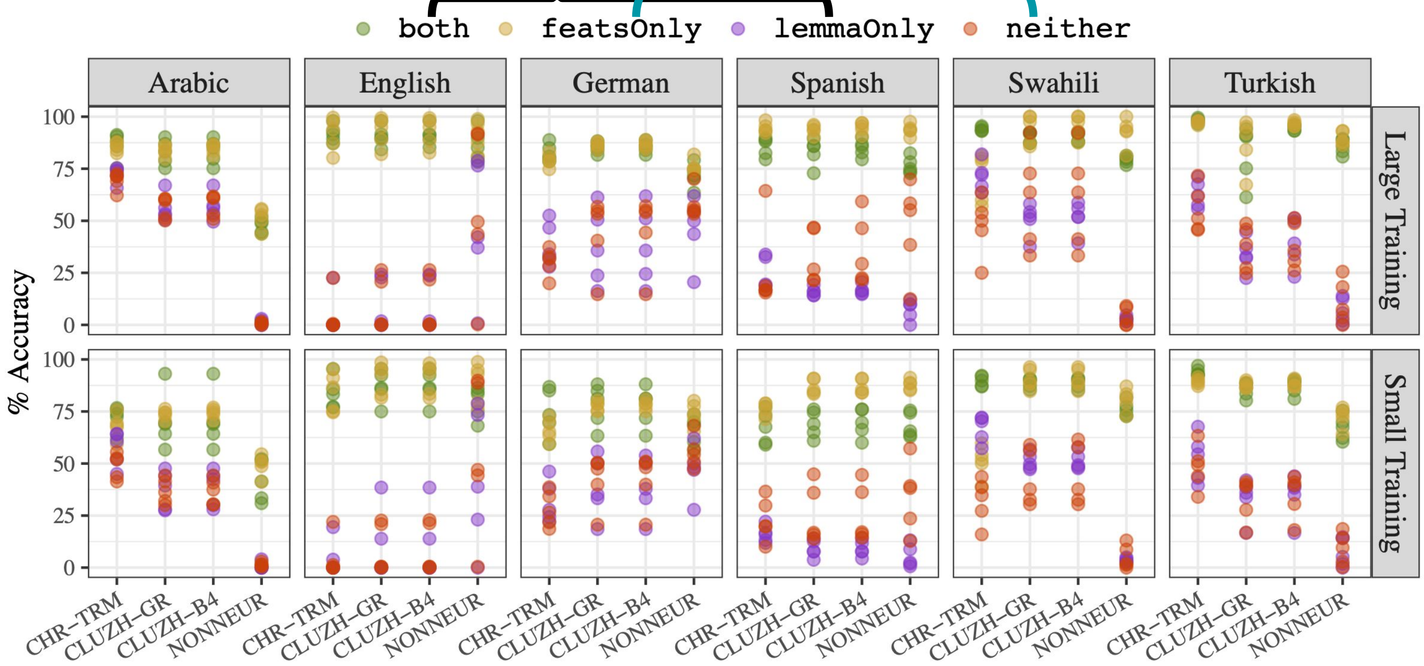
Results: Effect of Feature Overlap

FEATSATTESTED - FEATSNOVEL = **49.75%** (large) **48.02%** (small)



Results: Effect of Lemma Overlap

LEMMAATTESTED - LEMMANOVEL < 2% (small & large)



Results: Correlation with Overlaps

Correlations between Accuracy and overlap partition
Uncontrolled: WEIGHTED and UNIFORM, Controlled: OVERLAPAWARE

Overlap Partition	Uncontrolled ρ	Controlled ρ
featsAttested	0.97	0.45
featsNovel	0.16	0.93
lemmaAttested	0.84	0.88
lemmaNovel	0.78	0.82
both	0.89	0.49
featsOnly	0.73	0.21
lemmaOnly	0.24	0.89
neither	-0.04	0.85

Results: Correlation with Overlaps

Accuracy difference between FEATSATTESTED AND FEATSNOVEL and correlation with FEATSATTESTED for each language

Train Size	Language Strategy	Avg. Score Difference	featsAttested $\sim$Accuracy ρ
Small	Arabic	33.00%	0.57
	Swahili	40.04	0.63
	German	40.35	0.23
	Turkish	41.96	0.83
	Spanish	52.60	0.75
	English	74.10	0.66
Large	Arabic	35.79%	0.44
	German	36.19	0.73
	Swahili	39.26	0.64
	Turkish	52.14	0.59
	Spanish	61.01	0.64
	English	80.17	0.82

Thank you!