Exploring Linguistic Probes for Morphological Inflection

Jordan Kodner
Salam Khalifa*
Sarah Payne*
Stony Brook University

EMNLP 2023
Singapore

Morphological Inflection

Patterns of word formation to express grammatical categories

English *walk*+PAST → *walked*

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

Mandarin 3+PL → *tāmen* ‘they’

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

English *walk*+PAST → *walked*

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

Mandarin 3+PL → *tāmen* ‘they’

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’

- 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 an NLP 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`
…	…	…		…

Traditional Data Splitting

Traditional Language-Independent Random Splitting

e.g., SIGMORPHON shared task pre-2022

+ The same algorithm can be used across languages
+ Results are in some way more comparable across languages
− But offers next to no control over which phenomena appear in which splits

Traditional Data Splitting

Traditional Language-Independent Random Splitting

e.g., SIGMORPHON shared task pre-2022

- \+ The same algorithm can be used across languages
- \+ Results are in some way more comparable across languages
- − But offers next to no control over which phenomena appear in which splits

Overlap-Aware Language-Independent Random Splitting

e.g., SIGMORPHON 2022 and 2023 shared tasks

- \+ The proportion of triples with lemmas or feature sets overlapping in test and train is controlled → Holds this variable constant across languages/splits
- − But still no control over which phenomena appear in which splits

Language-Dependent Data Splitting

Data splits to test specific pieces of morphological generalization

- Tests specific pieces of the paradigm of a specific language
  → Much more control over what is being tested
- Can select patterns to tests specific kinds of generalization
  Over lemmas, over features, pre/in/suffixation, fusional vs agglutinative…
- Requires a “quantity over quality” approach, because morphological patterns need to be identified individually

Language-Dependent Data Splitting

Data splits to test specific pieces of morphological generalization

- Tests specific pieces of the paradigm of a specific language
  → Much more control over what is being tested
- Can select patterns to tests specific kinds of generalization
  Over lemmas, over features, pre/in/suffixation, fusional vs agglutinative…
- Requires a “quantity over quality” approach, because morphological patterns need to be identified individually

Some of these probes may be practically impossible but still provide useful information about how the model ‘thinks’

Experimental Setup: Data Sets

Verbs from three languages extracted from UniMorph 3+4

- **English**, **Spanish**, and **Swahili** are typologically distinct
- UniMorph is frequently used as a data source for morphological inflection
- Combining and normalizing UniMorph 3 and 4 maximizes the available data
- Transcribed data sets were created in parallel to UniMorph's orthography

  → All splits were created with parallel orthographic and transcribed versions

	# Lemmas	# Feature Sets	# Triples	
English (Germanic)	9,118	5	27,836	Highly fusional
Spanish (Romance)	7,326	152	1,077,655	Mixed
Swahili (Bantu)	131	169	10,925	Highly agglutinative

Experimental Setup: Data Format

Basic Format

- TRAIN consisted of 1600 training triples and 400 fine-tuning triples
- TEST consisted of up to 1000 test pairs (lemma, feature set)
- All random splits were performed five times with distinct randoms seeds

Experimental Setup: Data Format

Basic Format

- TRAIN consisted of 1600 training triples and 400 fine-tuning triples
- TEST consisted of up to 1000 test pairs (lemma, feature set)
- All random splits were performed five times with distinct randoms seeds

Orthography vs Transcriptions

- Parallel IPA transcriptions were produced for each language
  cmudict-ipa[1] for English, Epitran[2] for Spanish and Swahili
- All data splits were created with parallel transcription and orthography versions in order to test the effect of presentation style

[1]https://github.com/menelik3/cmudict-ipa, [2]Mortensen et al. 2018

Experimental Setup: Systems

Three systems were evaluated

CLUZH

Char transducer (Clematide et al 2022) — SIGMORPHON 2022 best performer w/ code

CHR-TRM

Char transformer (Wu et al 2021) — Commonly used baseline

ENC-DEC

Bidir LSTM (Kirov & Cotterell 2018) — Treated as cognitively plausible model

Experimental Setup: List of Probes

BLIND: Language-independent random sampling (Kodner et al, 2023, *ACL*)

Verbs: English (`en`; highly fusional) ←→ Spanish (`es`) ←→ Swahili (`sw`; highly agglutinative)

Experimental Setup: List of Probes

BLIND: Language-independent random sampling (Kodner et al, 2023, *ACL*)

Verbs: English (`en`; highly fusional) ←→ Spanish (`es`) ←→ Swahili (`sw`; highly agglutinative)

PROBE: Random sampling testing specific morphological patterns

Agglutinative feature generalization probes

`es-FUT`	suffixation
`es-AGGL`	suffixation (harder)
`sw-1PL`	prefixation
`sw-NON3`	prefixation (harder)
`sw-FUT`	string infixation
`sw-PST`	str infix w/ distractor

Conjugational class generalization probes

`es-IR`	suffixation
`es-IRAR`	suffixation (harder)

Fusional feature generalization probes

`en-NFIN`	suffixation
`en-PRS`	suffixation
`en-PRS3SG`	suffixation
`es-PSTPFV`	suffixation
`sw-PSTPFV`	str infix w/ distractor

Agglutinativity and Generalization

Agglutinative Patterns - Feasible

- Roughly 1-to-1 mapping between features in a set to morphological patterns
- Generalize across feature sets with overlapping features should be possible
- Swahili is overwhelmingly agglutinative

Approx. one afffix per feature
Swahili *ulipika* “you cooked”

u-	*li-*	*pik-*	*a*
2.SG-	PST-	cook-	IND

Agglutinativity and Generalization

Agglutinative Patterns - Feasible

- Roughly 1-to-1 mapping between features in a set to morphological patterns
- Generalize across feature sets with overlapping features should be possible
- Swahili is overwhelmingly agglutinative

Approx. one afffix per feature
Swahili *ulipika* "you cooked"

u-	*li-*	*pik-*	*a*
2.SG-	PST-	cook-	IND

Fusional Patterns - Infeasible

- Whole feature sets roughly correspond to non-decomposable patterns
- Correct generalization can be impossible, but errors are potentially informative
- English inflection is fusional
  Spanish is mixed

One unitary suffix
Spanish *cocinaste* "you cooked"

cocina-	*ste*
cook-	2.SG.PST.IND

Example Probe: es-FUT

	SG	PL
1	INF+é	INF+ámos
2;INFM	INF+ás	INF+áis
2;FORM	INF+á	—
3	INF+á	INF+án

The Spanish future is agglutinative: Infinitive + person/number marking similar to most other tense/moods.

UniMorph-specific: The infinitive is the lemma. There is no 2;FORM;PL

Example Probe: `es-FUT`

For 5 random seeds:

- 5 of 7 person/number combinations containing V;IND;FUT are randomly withheld for TEST
- TRAIN sampling proceeds as normal except for these 5 feature sets 1600 training + 400 fine-tuning
- TEST sampling then proceeds as normal
- All triples except for those with the 5 withheld feature sets are discarded.

All PROBE splits follow similar logic

	SG	PL
1	INF+é	INF+ámos
2;INFM	INF+ás	INF+áis
2;FORM	INF+á	—
3	INF+á	INF+án

The Spanish future is agglutinative: Infinitive + person/number marking similar to most other tense/moods.

UniMorph-specific: The infinitive is the lemma. There is no 2;FORM;PL

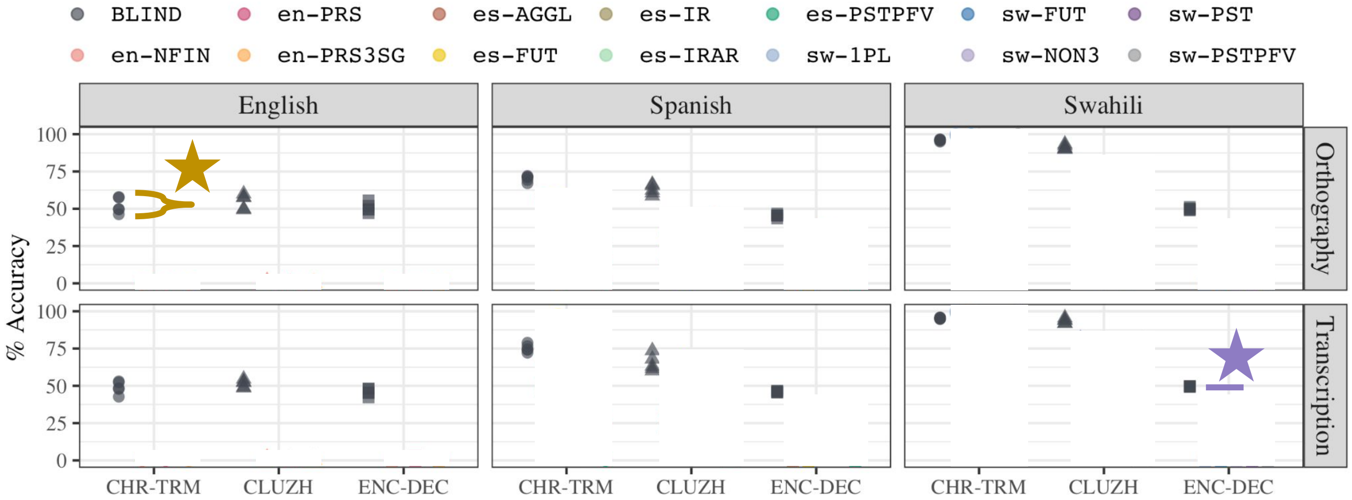
Orthography vs Transcription

The effect of presentation style is small and inconsistent

- Orthography +4.07 for English, -0.45 for Swahili, -2.80 for Spanish
- In an ANOVA analysis, only system and language are significant predictors

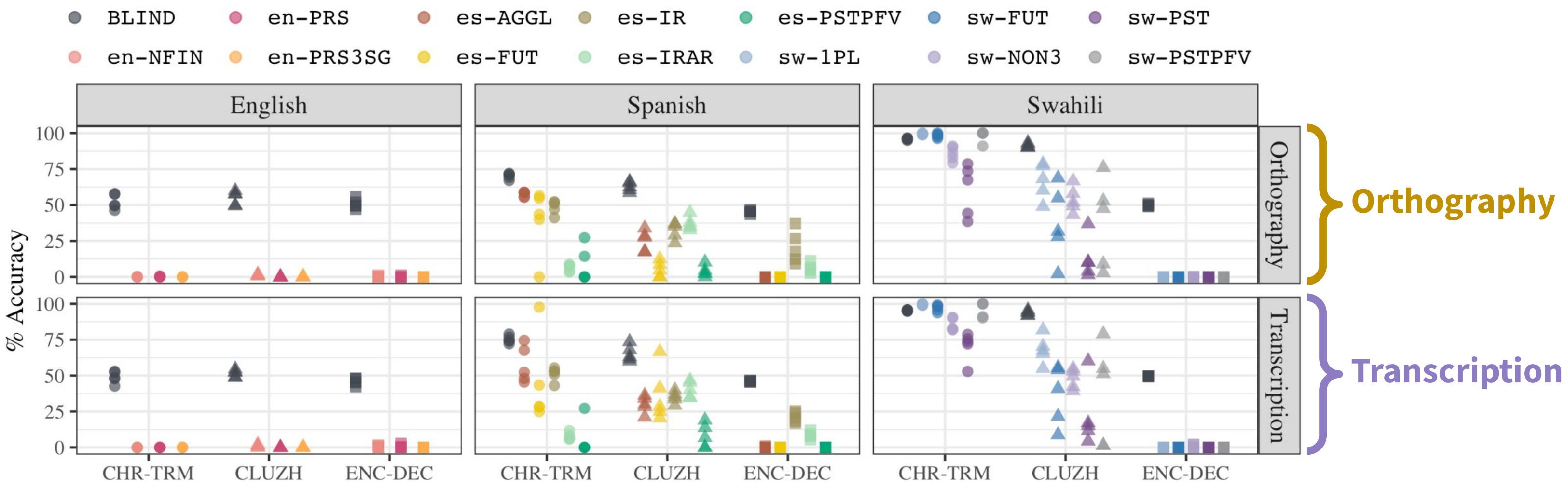
Variable	F-Statistic	*p*-Value
System	68.093	<2e-16
Seed	0.223	0.925
Presentation style	0.014	0.906
Language	76.588	<2e-16
Language * Presentation	1.061	0.351

Average Performance Summary



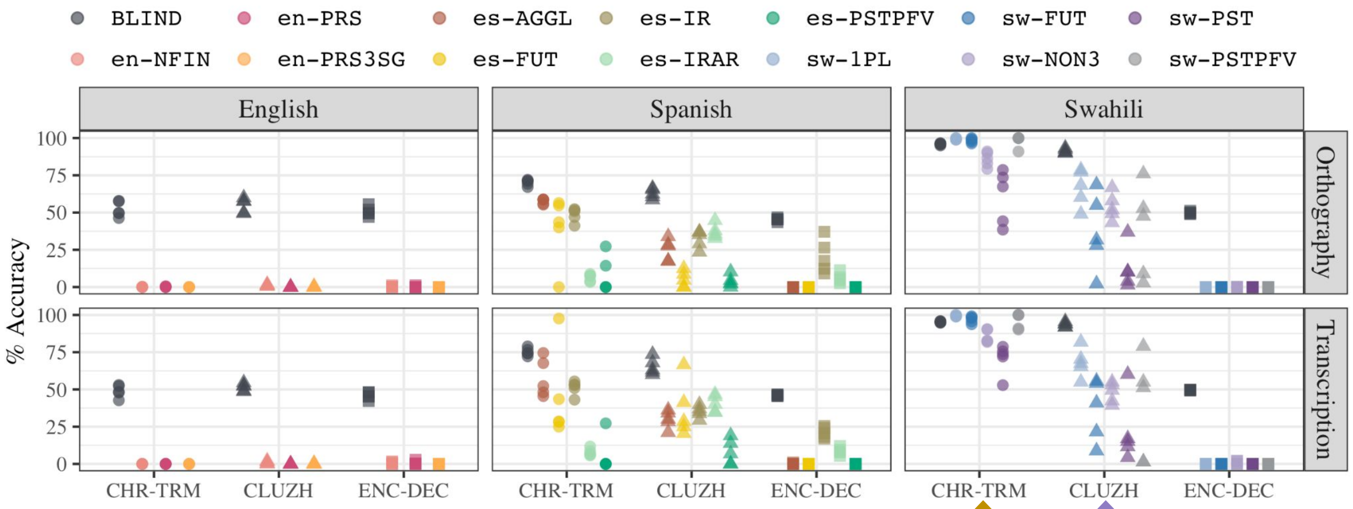
- **Scores ranges across seeds on BLIND**
  from 11.60 (CHR-TRM English Ortho)
  to 0.60 (ENC-DEC Swahili Transcr)

Average Performance Summary



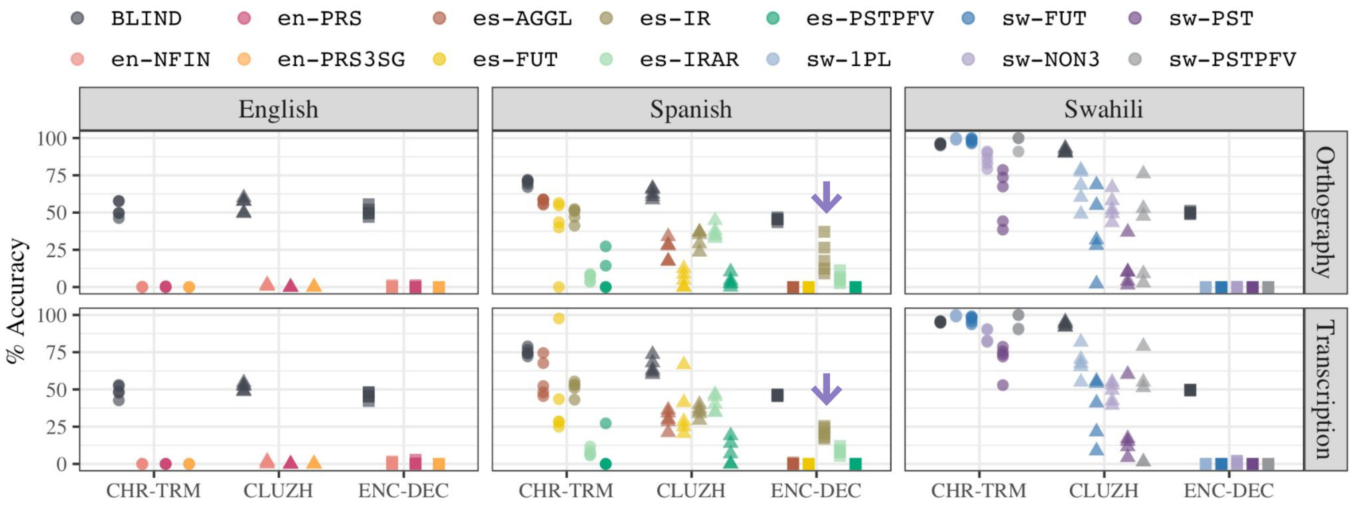
- Scores ranges across seeds on BLIND from 11.60 (CHR-TRM English Ortho) to 0.60 (ENC-DEC Swahili Transcr)
- Orthography vs Transcription are visually similar on all BLIND and PROBE splits

Average Performance Summary



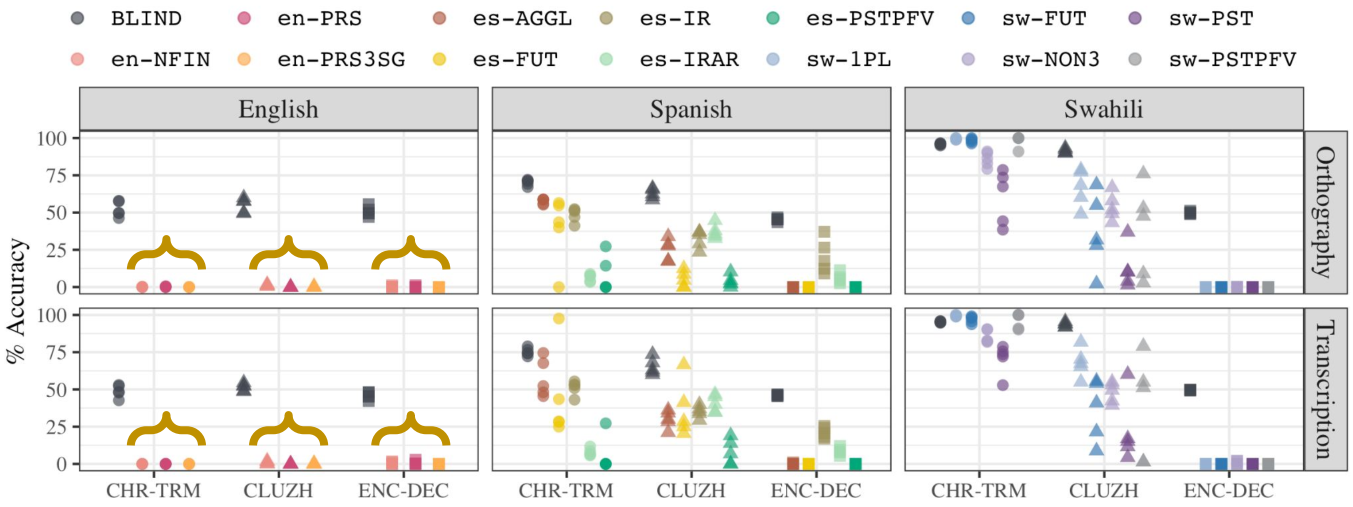
- CHR-TRM performs especially well on Swahili PROBE splits
- CLUZH shows very high variability across seeds on Swahili PROBE splits

Average Performance Summary



- **ENC-DEC only achieves meaningful performance on `es-IR` and `es-IRAR`**
  → No ability to generalize across feature sets

Average Performance Summary



- English PROBE splits are impossible
- No system performed well, but errors are insightful →
- No model outputs the bare lemma

- All output primarily -*ing*, -*(e)d*, or -*(e)s* forms
- When NFIN is replaced with PRS, CHR-TRM and CLUZH output primarily -*ing* or -*(e)s*, showing generalization of PRS feature from PRS;3;SG and/or PRS;PRS.PTCP

Main Conclusions

- Orthography vs Transcriptions makes no major difference for these languages
  Even for English, average performance only differs by 4 points
- Score ranges are high across randoms seeds
  Performance on one random sample unlikely to reflect true performance
- Language-specific probes reveal systems achieve generalization differently
  Systems succeed and fail on different probes
  The types of errors that they make reveal generalization strategies

Thank you!

Stony Brook University

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

NSF