Exploring Linguistic Probes for Morphological Generalization

Jordan Kodner and Salam Khalifa* and Sarah Payne*
Department of Linguistics & Institute for Advanced Computational Science
Stony Brook University, Stony Brook, NY, USA
{first.last}@stonybrook.edu

Abstract

Modern work on the cross-linguistic computational modeling of morphological inflection has typically employed language-independent data splitting algorithms. In this paper, we supplement that approach with language-specific probes designed to test aspects of morphological generalization. Testing these probes on three morphologically distinct languages, English, Spanish, and Swahili, we find evidence that three leading morphological inflection systems employ distinct generalization strategies over conjugational classes and feature sets on both orthographic and phonologically transcribed inputs.

1 Introduction

The current practice in the evaluation of computational morphological inflection models, such as that employed in the SIGMORPHON, CoNLL-SIGMORPHON and SIGMORPHON-UniMorph shared tasks (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022) as well as in more targeted studies focused on specific languages or the generalization behavior of computational models (Goldman et al., 2022; Wiemerslage et al., 2022; Kodner et al., 2023b; Guriel et al., 2023; Kodner et al., 2023a), is to train on (lemma, inflection, features) triples and predict inflected forms from held-out (lemma, features) pairs. The algorithm for generating train-test splits is both random and language-independent, which has proven successful in distinguishing morphological inflection models at the gross quantitative level. Models differ in their performance across languages and in their ability to generalize across lemmas or feature sets.

In this paper, we both replicate this type of analysis and contrast it with new language-specific probes for testing models’ generalization abilities in a more controlled fashion. We examine 13 probes in three languages – English, Spanish, and Swahili – chosen for data availability and their distinct morphological characteristics. In addition, we investigate the effect of presentation style on performance: that is, whether the choice between phonological transcription or orthography has a substantial effect on outcomes. We report on three systems which differ widely in their behavior. They often – but not always – make reasonable linguistic generalizations, even in their incorrect predictions. In addition, we find no statistical effect for presentation style, even on English. This has implications for research that attempts to use neural networks for cognitive modeling and evaluates on orthographic data rather than more domain-appropriate phonological transcriptions.

2 Languages

Three languages were chosen whose inflectional morphologies range from entirely fusional (English), to mixed (Spanish), to mostly agglutinative (Swahili). In highly agglutinative languages, individual features in a set tend to correspond to distinct morphological patterns, so a model may generalize to unseen feature sets by mapping component features to their corresponding patterns. This is exemplified by the Swahili example (1), in which most features correspond to individual morphemes; only the person/number prefix maps to more than one feature. On the other hand, highly fusional languages map entire feature sets to single patterns. This is shown by the Spanish example (2), in which all features map together onto a single suffix.

(1) **Swahili ulipika** ‘you (singular) cooked’

\[ \text{u-} - \text{l-} - \text{pik-} - \text{a} \]

2.SG-PST-cook-IND

(2) **Spanish cocinaste** ‘you (singular) cooked’

\[ \text{cocina-} - \text{ste} \]

cook- 2.SG.PST.IND
All data was adapted from Kodner et al. (2023b), which was in turn extracted from UniMorph 3 and 4 (McCarthy et al., 2020; Batsuren et al., 2022). The data was subjected to additional processing as described below. For each language, only verbs were extracted, and multi-word expressions were excluded. Importantly, no morpheme segmentation is provided in the UniMorph data unlike our illustrative examples, so agglutinativity must be discovered by each system.

**English (Germanic):** English triples were transcribed using an IPA translation of the CMU Pronouncing Dictionary. Triples without available transliterations were discarded. When the dictionary provided multiple transcriptions, the lemma and inflection transcriptions which minimized Levenshtein distance between them were chosen.

**Spanish (Romance):** Spanish triples were transcribed using the Epitran package (Mortensen et al., 2018), which does not include stress information. UniMorph treats the infinitive as the lemma and includes the 2nd person plural in its paradigms.

**Swahili (Bantu):** Swahili triples were transcribed using Epitran. UniMorph treats the bare stem as the lemma. The Swahili data set is much smaller than the others and does not contain any forms with negative marking. We found many inconsistencies in the UniMorph 4 feature tags which we normalized. Most importantly, PFV and PRF, which both indicate perfect aspect, were mapped to PFV and tag order was made consistent. Triples in swc.sm with tags which could not be clearly mapped to tag sets in swc were excluded.

### 3 Data Splits

We created several random data splits to study dimensions of morphological generalization. As in prior work, BLIND splits were made without regard to specific linguistic properties of the triples. PROBE splits were sensitive to the properties of the triples: they were divided into relevant and irrelevant sets according to properties of the feature sets or lemmas. The irrelevant triples were split as in BLIND in order to pad training to the same length as BLIND, but irrelevant test triples were discarded. The relevant triples were split in a way specific to each probe, controlling which occurred in train+fine-tuning. For both BLIND and PROBE, each split was performed five times with unique random seeds, and each seed was used to produce parallel orthographic and transcription versions of the splits for evaluating presentation style, yielding ten samples in total. Data sets contained 1600 training items + 400 for fine-tuning. BLIND splits contained 1000 test items.

**BLIND:** Following Kodner et al. (2023b), this splitting strategy ensures that approximately 50% of test items contain OOV feature sets but is otherwise blind to the identity of those feature sets. This was shown in the 2022 SIGMORPHON-UniMorph shared task to create more opportunities for testing dimensions of generalization across feature sets than more traditional uniform random sampling (Kodner et al., 2022).

**English Probe Splits**

**en-NFIN:** This probe is designed to test what the system does when it knows nothing about the relevant tag. The NFIN tag, which maps to no change, only appears in the infinitive and non-3rd singular present. Triples with the NFIN tag were excluded from training and presented during test. No system can know what inflection the NFIN corresponds to, so it can only succeed if it defaults to the lemma.

**en-PRS:** This probe tests the implications of UniMorph’s design choice of annotating the non-3rd singular present with NFIN rather than PRS. We used identical splits to en-NFIN, but replaced the NFIN tag with PRS. Since this is shared with the present 3rd singular (PRS; 3; SG) and participle (V.PTCP; PRS), a system should be reasonably expected to generalize either the -s or -ing endings to PRS items in test.

**es-PRS3SG:** This probe also replaces NFIN with PRS but instead withholds PRS; 3; SG from training and presents it during test. Success on this probe is impossible, since systems cannot learn to map this tag set to -s during training, but a system should generalize either the bare lemma or the -ing of the present participle.

**Spanish Probe Splits**

**es-FUT:** This probe tests the ability of systems to learn a basic agglutinative suffixing pattern: the future tense (IND; FUT) is formed by suffixing regular person/number marking onto the infinitive (correr ‘to run,’ correr-ás ‘you will run’). There are six possible person/number combinations attested in the future indicative paradigm. For each split/seed, two feature sets with IND; FUT were randomly cho-

---

1https://github.com/jkodner05/ACL2023_RealityCheck
2https://github.com/jkodner05/EMNLP2023_LingProbes with summaries provided in the Appendix.
sen, and then triples with these features were randomly sampled to appear in training for models to learn the pattern. All other feature sets containing IND;FUT were withheld from training, and triples containing them were sampled to appear in test.4

**es-AGGL:** The conditional (COND) and imperfect (IND;PST;IPFV) are also agglutinative. Two each of conditional, imperfect, and future feature sets were independently selected to appear in the training split, and the rest were withheld for evaluation. This is a more challenging version of the es-FUT probe because a system has to learn three agglutinative patterns at once.

**es-PSTPFV:** This probe tests what a system does when it is forced to predict missing parts of a fusional paradigm. As shown in (2), preterite (IND;PST;PFV) person/number forms are fusional, because they manifest distinctly from the marking for all other tense/aspects (except for the 1st person plural) and are not decomposable into a preterite part and a person/number part. Since a system cannot predict these forms from their component person/number features matched with other tenses, it should perform poorly even if it succeeds at es-FUT and es-AGGL. A system should generalize person/number marking from the other tenses.

**es-IR:** This probe tests generalization across conjugational classes rather than across feature sets. Spanish verbs fall into three classes clearly indicated by their infinitive suffix: -ar, -er, and -ir. -ir shares many, but not all, of its inflections with -er. For this probe, all but 50 randomly chosen -ir lemmas are banned from train sampling, which results in 10-18 -ir training triples sampled per seed. A system should predict the appropriate -er form or one with -i- for each -ir evaluation item or overapply the majority -ar ending.

**es-IRAR:** This probe is similar to es-IR but much more challenging because all -ar and -ir verbs are withheld from training. We predict that a system should produce the -er inflected form or replace -e- in forms with the appropriate -a- or -i-.

**Swahili PROBE Splits**

**sw-1PL:** As in (1), person/number and tense/aspect marking is marked agglutinatively before the stem in Swahili. The 1st person plural (1;PL) is marked with a tu- prefix. Two feature sets containing 1;PL were randomly selected by seed and allowed to appear in train, and the rest were withheld for eval-

---

4Illustrations for two seeds are provided in the Appendix.
3rd singular present -(e)s which is indicated in transcription. However, it is also possible that the difference is due to the choice of transcription dictionary or how we processed it. For Spanish, a transcription scheme that retained stress may have proven more challenging than Epitran’s which lacks it.

Table 1: ANOVA analysis on BLIND showing significant effect for system and language but not presentation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>68.093</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>seed</td>
<td>0.223</td>
<td>0.925</td>
</tr>
<tr>
<td>presentation</td>
<td>0.014</td>
<td>0.906</td>
</tr>
<tr>
<td>language</td>
<td>76.588</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>presentation * lang</td>
<td>1.061</td>
<td>0.351</td>
</tr>
</tbody>
</table>

We follow this with an ANOVA analysis of five variables: the system, seed, presentation style, language, and the interaction between presentation and language, to determine which differences in mean accuracy are unlikely to be due to chance. Summarized in Table 1, we find that only the choice of system and the language are significant, not presentation style. This is consistent with visual inspection. The conclusion is the same on BLIND, PROBE, and both combined.

5.2 Generalization and Linguistic Analysis

Feature Set Generalization in BLIND

A breakdown of BLIND test triple by type in Table 2 replicates prior work (Kodner et al., 2022, 2023b) demonstrating that generalization to unseen feature sets is particularly challenging. All systems showed lower accuracy on OOV feature sets (fsOOV & bothOOV) than on other triples. ENC-DEC shows virtually no ability to do this.5

Table 2: Average performance on BLIND orthography across seeds and languages. noOOV = triples where both lemma and feat. set were observed; lmOOV = lemma is OOV; fsOOV = feature set is OOV. bothOOV = lemma and feature set are OOV.

<table>
<thead>
<tr>
<th>System</th>
<th>noOOV</th>
<th>lmOOV</th>
<th>fsOOV</th>
<th>bothOOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHR-TRM</td>
<td>94.40%</td>
<td>82.68%</td>
<td>52.90%</td>
<td>36.46%</td>
</tr>
<tr>
<td>CLUZH</td>
<td>93.93</td>
<td>95.43</td>
<td>17.43</td>
<td>48.93</td>
</tr>
<tr>
<td>ENC-DEC</td>
<td>93.79</td>
<td>86.01</td>
<td>2.53</td>
<td>1.43</td>
</tr>
</tbody>
</table>

For CHR-TRM and CLUZH, we observe a much smaller performance gap on OOV feature sets for the more agglutinative languages (Swahili -9.5 points < Spanish -49.76 < English -81.38), which indicates that these systems can perform some degree of generalization across feature sets. This contradicts the prior work, which uncovered no substantial difference between fusional and agglutinative languages, indicating an inability to perform this kind of generalization.

This discrepancy may be explained by how we processed the Swahili data. The published UniMorph data contains several inconsistencies, summarized in Section 2, in its feature tags which made the task of generalization unfairly challenging in prior work. We corrected this by normalizing the feature sets and removing triples which we could not adequately fix. However, this would only explain the middling reported performance on other agglutinative languages if their UniMorph data sets turn out to be highly inconsistent as well. This question is left for future work.

Generalization on Swahili PROBE Splits

CHR-TRM is very successful at generalizing the prefix in sw-1PL and string infix in sw-FUT. CLUZH sometimes applies the 1st singular prefix instead

---

5A complete breakdown is provided in the Appendix.
of the plural in sw-1PL, but is less consistent for sw-FUT, where it produces many infixes belonging to other tenses in addition to nonsense outputs. sw-NON3 was developed as a more challenging version of sw-1PL, and performance was indeed lower. However, instead of “near-miss” errors suggesting linguistic generalization, both systems produced a range of person/number, tense, and nonsense errors not clearly related to the probe.

sw-PST was designed to be similar to sw-FUT but with a distractor fusional string infix. As expected, performance was lower, especially for CLUZH, which often substituted the distractor past perfect, other tense marking, or omitted tense altogether similar to its errors on sw-FUT. Its errors in sw-PSTPFV were overwhelmingly application of the simple past, which can be explained as a generalization of PST while ignoring the PFV tag. These are mostly reasonable errors that point to some degree of generalization. ENC-DEC showed no ability to generalize according to component morphological features even this highly agglutinative setting.

**Generalization on Spanish PROBE Splits**

For es-FUT, CHR-TRM and CLUZH identified the correct pattern of infinitive + person/number marking but often produced slightly incorrect forms. In the orthographic tests, many of these errors involved incorrect stress marking. These could be considered reasonable “near-misses” that point to generalization. However, for the more challenging es-AGGL, both produced less interpretable errors. es-PSTPFV was designed to be impossible but insightful, since preterite person/number marking cannot be predicted from the person/number marking of other tenses. Indeed, CLUZH generalized the person/number marking from the other tenses as well as the few preterite person/number markings presented during training, indicating that it can employ the relevant generalization. However, CHR-TRM and ENC-DEC were less interpretable.

es-IR and es-IRAR differ from the other probes in that they require generalization across conjugational classes indicated by the form of the lemma rather than generalization across feature sets. This proved appropriately challenging for all systems, which all produced many nonsense errors and some near-miss overuse of -er endings or combined -a-ver and -i-ver endings. Notably, these were also the only probes for which ENC-DEC showed some success. The tasks that the system was asked to perform in (Kirov and Cotterell, 2018) rewarded analogy across lemma forms rather than unseen feature sets. This is similar to the es-IR and es-IRAR probes but distinct from all the other probes which were focused on feature generalization.

**Generalization on English PROBE Splits**

Accuracy was near-zero on every English PROBE. This was expected, because the tasks were effectively impossible, however, qualitative analysis is insightful. For en-NFIN and en-PRS, a system would succeed if it defaulted to the bare lemma as an inflection of last resort. However, no system took that approach. For en-NFIN, CHR-TRM consistently outputted -ing forms with other errors, CLUZH produced -(e)d forms, and ENC-DEC produced mostly -(e)s forms with other errors.

The replacement of UniMorph’s NFIN with PRS did indeed have an effect, indicating that some behavior on English can be attributed to this corpus design choice: Both CHR-TRM and CLUZH now consistently produced present participle -ing forms, while ENC-DEC instead produced -(e)d or -(e)s. This indicates generalization over the PRS feature. For en-PR35S, CLUZH always produced the bare lemma form, showing clear generalization of PRS. CHR-TRM only produced nonsense errors, while ENC-DEC produced nonsense errors suffixed with -(e)d or -ing.

6 Conclusions

In this work, we present language-specific probes to evaluate the ability of computational systems to perform morphological generalizations as a complement to prior work relying on large-scale language-independent data-splitting. Systems differ substantially both in their ability to perform generalizations and in their problem-solving strategies. Of the two systems showing generalization ability, both perform well on the simplest probes but struggle on more complex but feasible probes. On probes designed to be impossible but insightful, both systems show some degree of reasonable generalization. In addition to this, we find no significant effect of presentation style on the behavior of the three systems. While we maintain that the choice of presentation style should be driven by application when possible (usually orthography for NLP or transcription for cognitive modeling), our results suggest that it can be reasonable to use orthography when transcription is impractical.

8937
Limitations
The main limitation of the the language-specific splitting approach compared to the traditional language-independent splitting approach is the language-specific and domain-specific expertise needed to design the probes themselves. Nevertheless, this work successfully demonstrates the utility of such probes. Future work with a larger group of experts in a wider range of languages could extend this approach to more test cases. We believe that this could be implemented on a larger scale such as for a complementary shared task in the future.

Ethics Statement
To the best of our knowledge, all results published in this paper are accurate. All data sources are free, publicly available, and cited in the article. No sensitive data was used which could violate individuals’ privacy or confidentiality. Authorship and acknowledgements fairly reflect contributions.

Acknowledgements
We thank Zoei Liu for motivating discussion. Experiments were performed on the SeaWulf HPC cluster maintained by RCC and the Institute for Advanced Computational Science (IACS) at Stony Brook University and made possible by National Science Foundation (NSF) grant No. 1531492. The third author gratefully acknowledges funding through the IACS Graduate Research Fellowship and the NSF Graduate Research Fellowship Program under NSF Grant No. 2234683.

References


A Appendix

Table 3: Type frequencies for lemmas, feature sets, and triples for each language data set before splitting.

<table>
<thead>
<tr>
<th># Lemmas</th>
<th># Feature Sets</th>
<th># Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>9188</td>
<td>5</td>
</tr>
<tr>
<td>Spanish</td>
<td>7326</td>
<td>152</td>
</tr>
<tr>
<td>Swahili</td>
<td>131</td>
<td>169</td>
</tr>
</tbody>
</table>

Table 4: Average percent accuracy and (accuracy range) on BLIND English across seeds for each system and presentation style. noOOV = triples where both lemma and feature set were observed; lmOOV = lemma is OOV; fsOOV = feature set is OOV. bothOOV = lemma and feature set are OOV.

Table 5: Average percent accuracy and (accuracy range) on BLIND Spanish across seeds for each system and presentation style. noOOV = triples where both lemma and feature set were observed; lmOOV = lemma is OOV; fsOOV = feature set is OOV. bothOOV = lemma and feature set are OOV.

Table 6: Average percent accuracy and (accuracy range) on BLIND Swahili across seeds for each system and presentation style. noOOV = triples where both lemma and feature set were observed; lmOOV = lemma is OOV; fsOOV = feature set is OOV. bothOOV = lemma and feature set are OOV.

Table 7: Percent accuracy and (accuracy range) on English PROBE splits by system and presentation style.

Table 8: Percent accuracy and (accuracy range) on Spanish PROBE splits by system and presentation style.

Table 9: Percent accuracy and (accuracy range) on Swahili PROBE splits by system and presentation style.
Table 10: Description of es-FUT train and test feature sets for seeds 0 and 4. Seeds 1-3 provide similar results, and similar splitting logic applies to the other PROBE splits. Triples with feature sets containing FUT are partitioned at random by seed into those that can only be sampled for train+finetune and those that can only be sampled for test. Otherwise, data is split uniformly at random as in most SIGMORPHON shared tasks. No relevant PROBE feature set appears in both train and test, but all other feature sets may. Evaluation is only performed on sampled triples with the relevant PROBE feature set. This approach allows us to test the specific impact of the PROBE in an otherwise typical setting.