Ye Olde French: Effect of Old and Middle French on SIGMORPHON-UniMorph Shared Task Data

William Kezerian, Lam An Wyner, Sandro Ansari and **Kristine M. Yu**
Department of Linguistics
University of Massachusetts Amherst
{wkezerian, lamanwyner, alexanderans, kmyu} @umass.edu

Abstract

We offer one explanation for the historically low performance of French in the SIGMORPHON-UniMorph shared tasks. We conducted experiments replicating the 2023 task on French with the non-neural and neural baselines, first using the original task splits, and then using splits that excluded Old and Middle French lemmas. We applied a taxonomy on our errors using a framework based on Gorman et al. (2019)'s annotation scheme, finding that a high portion of the French errors produced with the original splits were due to the inclusion of Old French forms, which was resolved with cleaned data.

1 Introduction

The annual SIGMORPHON-UniMorph shared task on morphological (re)-inflection has been a locus for developing language resources, algorithms, and tasks in the computational morphology community since Cotterell et al. (2016). While the details of the task settings have varied over the years, the basic task has been: given training data consisting of triples ⟨lemma, inflection features, inflected form⟩, (e.g., ⟨*désarmer*, 2.SG.SUBJ, *désarmerais*⟩ for French 'disarm',) train a model to infer inflected forms given pairs ⟨lemma, inflection features⟩.

Already in 2016, Cotterell et al. (2016) noted the remarkable average accuracy achieved in the basic task (95.56% averaged across languages in the top system) and the "surprising" huge performance gap between neural and non-neural approaches (e.g., with the best performing neural approach exceeding the best non-neural one by as much as 60% in accuracy within a language). Seven years and three versions of UniMorph later, the most recent SIGMORPHON-UniMorph shared task, Goldman et al. (2023, p. 120) similarly remarks that performance over individual languages was "quite impressive" and that "all neural systems outperformed the non-neural systems" on average across languages.

However, Goldman et al. (2023) also notes the mysteriously poor performance of systems on French in two senses. First, no system achieved higher than 77.7% accuracy on French. The other languages for which models' accuracies peaked at 80% were Navajo, Ancient Greek, Sanskrit, Belarusian, and Sami. Goldman et al. (2023, p. 121) implicitly note the oddness of French (a suffixing, fusional, high-resource language) in this group: "While there is no one characteristic shared between all of these languages, it is worth noting that this list includes the only two extinct languages tested in this task, and the only mostly prefixing language. Perhaps further development of tailored models could help fill this gap."

Second, neural systems did not outperform non-neural systems on French—while the non-neural baseline achieved 77.7% accuracy, the best performing neural system achieved 74.7% accuracy. In fact, the non-neural baseline was the best performing system in English, Danish, and French. Goldman et al. (2023, p. 120) again point out the oddness of French in this group: "Partial explanation may be the small size of the inflection tables in Danish and English that necessitated inclusion of many lemmas in the training set and may facilitated better generalization ability of the non-neural baseline. **Admittedly, this explanation is not valid for French,[1] but this language was proven difficult in previous shared tasks (Cotterell et al., 2017, 2018) and in other works (Silfverberg and Hulden, 2018; Goldman and Tsarfaty, 2021).**"

Why has French been a particularly challenging language for inflection tasks since it was first added to UniMorph in 2017? In this paper, we show that Old and Middle French lemmas/forms have been erroneously included in French UniMorph data in all SIGMORPHON-UniMorph shared tasks involving

[1] Splits were sampled from 500 lemmas for French, but 2000 for Danish and 3000 for English (Goldman et al., 2023, Table 2).

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 39–50
June 20, 2024 ©2024 Association for Computational Linguistics

French, as well as Silfverberg and Hulden (2018); Goldman and Tsarfaty (2021). We also provide evidence that including these Old and Middle French forms has caused anomalously poor performance via three replication experiments of the 2023 shared task for French—two excluding Old and Middle French lemmas—and an error analysis of the results.

2 Background

To contextualize our claim that Old and Middle French forms have resulted in poor performance, we will first provide background information on Old and Middle French and explain its presence in Wiktionary, as well as a brief history of poor performance on French in past inflection tasks. Hereafter "Old French" will be used as shorthand to encompass both Old French and Middle French.

2.1 Old French

Old French evolved into Middle French in the 14th century, then to modern French in the 17th century. Old French conjugation tables have been extensively documented in the English edition of Wiktionary (the source of data for `fra`, the French UniMorph data file). The only cited source for these tables is *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle* (Godefroy, 1881), which outlines all of the possible conjugations for Old French verbs.

2.1.1 Old French lemmas and suffixes

Most suffixes used in Old French verb conjugations are not licit verb inflection suffixes in modern French. These include *-ois*, *-oit*, and *oient* in the past imperfect and *â* in past perfect suffixes such as *-astes* and *-asmes*. For more examples, see Table A2 in the Appendix.

Although it is fairly easy to identify Old French verb suffixes, there are no universal patterns that make it clear whether a lemma itself is Old French or modern French. This task requires French linguistic knowledge and investigation into the documentation on the verb.

2.2 Poor performance on French verb inflection in SIGMORPHON-UniMorph and related inflection tasks

SIGMORPHON-UniMorph shared tasks French verbal paradigms were first included in the SIGMORPHON-UniMorph shared inflection task in 2017.[2] In that task (subtask 1), the best-performing system (UE-LMU, neural) scored 89.50% by-form accuracy on French in the high resource setting, cf. 81.50% from the non-neural baseline (Cotterell et al., 2017, Table 12). Among the 52 languages in the task, only 4 had comparably poor performance (Cotterell et al., 2017, Table 9).

In the 2018 SIGMORPHON shared task (task 1), French appeared as a surprise language. The highest accuracy on French in the high resource setting was 90.40% (uzh-2, neural), cf. 82.80% from the non-neural baseline. Only 8 out of 103 languages had comparable or lower performance. (Cotterell et al., 2018, Tables 9, 10, 14). French was also included as part of the French-Occitan pair in the 2019 SIGMORPHON-UniMorph shared Task 1 involving training on high-resource languages to infer inflection on genetically related low-resource languages, but inferring something about French from performance is difficult since performance varied highly by how closely the two languages in the pair were related. After 2019, French was not included in SIGMORPHON-UniMorph shared tasks again until 2023.

While Romanian, Hungarian, and Latin yielded poorer performance than French in both the 2017 and 2018 shared tasks, Gorman et al. (2019, p. 143; Table 4)'s error analysis of the 2017 shared task discovered that all three of these languages suffered from a preponderance of **extraction errors** in how UniMorph parsed Wiktionary's inflectional paradigms that would have impacted performance in both 2017 and 2018. Gorman et al. (2019) did not perform an error analysis of French.

Minimal supervision inflection tasks Goldman et al. (2023, p. 120) also pointed to poor performance on French in Silfverberg and Hulden (2018) and Goldman and Tsarfaty (2021). However, Silfverberg and Hulden (2018) did not report *uniformly* low performance for French verbs across tasks. They trained an encoder-decoder model on 1 to 3 forms randomly sampled from: (i) 1000 randomly sampled inflection tables from UniMorph, or (ii) 1,131 inflection tables from UniMorph that contained items among the 10,000 most frequent word tokens from Al-Rfou' et al. (2013)'s dump of the French edition of Wikipedia. The task was then to generate the remaining missing forms in each

[2] While UniMorph 4.0 (Batsuren et al., 2022) added adjectives and nouns in `fra.segmentation`, all inflection tasks for French discussed in this paper have been only for verbs.

inflection table. When the inflection tables were randomly sampled, accuracy on missing forms in French verbs was the lowest of all 8 languages/part of speech data sets for 1, 2, and 3 forms, e.g., 83.64% for 3 forms, cf. 74.07% for the baseline model, a new implementation of Malouf (2017)'s LSTM model.

But when the inflection tables for training were sampled to contain *the most frequent* forms, accuracy for French verbs was 31.34% for French verbs (cf. 14.34% for the baseline)——in the middle of the pack among the 8 data sets, and higher than for Spanish verbs or Finnish verbs. Moreover, Silfverberg and Hulden (2018) reported one instance of near perfect accuracy for French: 99.50% accuracy in validating their implementation of Malouf (2017)'s LSTM in a replication of Malouf (2017)'s experiments using their original data from Flexique (Bonami et al., 2013). Flexique is an open source database for studying French inflection that builds on Lexique version 3.70 (New et al., 2001, 2004), an open source lexical database of French annotated with phonological, morphological, and frequency information. Lexique data is drawn from texts published after 1950 and subtitle files of French films available on the web and thus would not be expected to contain Old French.

In Malouf (2017)'s experiments (as well as Silfverberg and Hulden (2018)'s replications thereof) accuracy isn't only near-perfect for French (99.92%), but also highest for French out of 7 languages for both the LSTM system and the non-neural baseline from the 2017 SIGMORPHON shared task (99.06%) (Malouf, 2017, Table 2), described in §3.2.2. High accuracy on French is not due to the particular LSTM system, since the non-neural baseline did as well, and since the LSTM system did not perform well as the baseline in Silfverberg and Hulden (2018).

2.3 Hypotheses

In sum, in all but one of the inflection tasks reviewed in this section where UniMorph was the source of the French data, French accuracy was anomalously low relative to other languages. The one exception is Silfverberg and Hulden (2018)'s task, where the French UniMorph data was filtered to include only high frequency forms. When the source of French data was Flexique rather than UniMorph, accuracy was near perfect for both neural and non-neural models. In addition, unlike in the 2023 shared task, neural models outperformed the non-neural baseline on French in 2017 and 2018.

We hypothesized that: (i) Old French forms were prevalent in the UniMorph task splits when French yielded poor performance, i.e., in the 2017, 2018, and 2023 SIGMORPHON-UniMorph shared tasks, as well as Silfverberg and Hulden (2018)'s experiment that randomly sampled 1,000 inflection tables from French UniMorph, and (ii) Old French forms were not as prevalent or even absent in the task splits where French yielded better or near-perfect performance, i.e., in Silfverberg and Hulden (2018)'s experiment that filtered UniMorph inflection tables for high frequency forms, and in Malouf (2017)'s tasks splits from Lexique.

We also hypothesized that (iii) the prevalence of Old French forms in task splits was what was causing the anomalously poor performance on French. To support this hypothesis, we conducted three experiments replicating the 2023 SIGMORPHON-UniMorph task on French with the non-neural and neural baselines, first using the original task splits, and then re-sampling the task splits to exclude Old French lemmas in two different ways. Our prediction was that removing the Old French verbs from the task would lead to improvement in accuracy across both baseline models due to the elimination of errors related to Old French. We did not have a hypothesis about why the neural models failed to outperform the non-neural baseline on French in the 2023 shared task, but hoped that conducting an error analysis would reveal some insights.

3 Materials and methods

All source data, scripts for processing data, output files, and the error analysis spreadsheet can be found at `https://github.com/ProphecyOak/TIGRE-2023sigmorphon`, which includes a README explaining how to run the scripts. The script `reproduce.sh` can be run to repeat the steps used to produce the output data from the SIGMORPHON-UniMorph 2023 shared task replication experiments used for error analysis.

3.1 Identifying Old French lemmas

3.1.1 Data

We checked the prevalence of of Old French lemmas in the French UniMorph 4.0 files `fra`, `fra.args` and `fra.segmentations`.[3] We also checked for Old French in

[3] `https://github.com/unimorph/fra`, accessed March 4, 2024

the following train/dev/test splits from past SIGMORPHON-UniMorph shared tasks: **2017/2018**: `french-train-high`, `french-dev`, `french-covered-test`[4]; **2019**: `french-train-high` from the french-occitan training data[5]; (iv) **2023**: `fra.trn`, `fra.dev`, `fra.test`[6].

The French UniMorph data from 2017 (UniMorph 1.0, (Kirov et al., 2016)) and 2018 (UniMorph 2.0 (Kirov et al., 2018)) included 7,535 lemmas and 367,732 forms. 12,000 triples were sampled without replacement for the splits. We only included the high-resource training data set from 2017 and 2018 (5,592 lemmas / 10,000 forms), since the medium and low resource training data were proper subsets of the high resource training data (Cotterell et al., 2017, Table 3; Cotterell et al., 2018, Table 2) .

In addition, we also checked for Old French lemmas in Malouf (2017)'s French data extracted from Flexique (`french.dat`[7]) and Silfverberg and Hulden (2018)'s random sample of 1000 lemmas from French UniMorph 2.0 (`fr.um.V.txt`) and sample of 1,131 lemmas filtered to contain high frequency forms (`fr.um.V.top.txt`).[8] Goldman and Tsarfaty (2021) reported using training/testing splits from Silfverberg and Hulden (2018).

3.1.2 Detecting Old French Lemmas

We determined which lemmas were Old French by writing a script `lang-stats.py` that checked the entry of each lemma in the English edition of Wiktionary for Old French.

Because this script relies entirely on Wiktionary entries, the definition of Old French may not be entirely accurate in all cases. While most pages cite *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle* (Godefroy, 1881) as the source for Old French definitions, not all pages had corresponding entries in said dictionary, so the reliability of Wiktionary for performing this task is questionable; however, given that we had no expertise in Old French, we chose to use the available Wiktionary entries in order to automate the process of checking the UniMorph data.

3.2 2023 SIGMORPHON-UniMorph replication experiments

3.2.1 Generating splits without Old French lemmas

To generate two new sets of splits without Old French lemmas from the original splits for the 2023 shared task, we filtered `fra.trn`, `fra.dev`, and `fra.tst` from the original splits using `fra.segmentations` in the current French UniMorph 4.0 repository. The `fra.segmentations` file contains morpheme segmentations developed for UniMorph 4.0 (Batsuren et al., 2022). We first confirmed that `fra.segmentations` contained no Old French lemmas using the procedure specified in §3.1.2. We then wrote a script `formatSegmentations.py` that converted `fra.segmentations` from the old feature schema from UniMorph 3.0 to the new hierarchical feature schema of UniMorph 4.0 used in the 2023 task splits. This new file `fra.total` was then sampled to create two sets of splits.

The **form-sampled** (seg-minimal) splits included only the *forms* that were contained in both the original splits and `fra.segmentations`; these splits were thus smaller than the original ones. The **lemma-sampled** (seg) set included only *lemmas* that were contained in both the original 2023 splits and `fra.segmentations` but all forms for those lemmas contained in `fra.segmentations`. Since `fra.segmentations` included many more forms than `fra`, the lemma-sampled splits were larger than the original splits. We based our splits on the original splits to preserve the original demographics, but wanted to account for both larger training and lower training amounts without adding in too much of our own biases.

3.2.2 Algorithms

Since one anomalous aspect of performance on French in the 2023 shared task was that the non-neural baseline outperformed neural models, we included both the non-neural baseline[9] and the neural baseline (Wu et al., 2021)[10] in our experiments. The non-neural baseline has been used

[4] https://github.com/sigmorphon/conll2017/tree/master/all/task1, https://github.com/sigmorphon/conll2018/tree/master/task1/surprise

[5] https://github.com/sigmorphon/2019/blob/master/task1/french--occitan/french-train-high

[6] https://github.com/sigmorphon/2023InflectionST/tree/main/part1/data

[7] https://github.com/rmalouf/abstractive/blob/master/data/french.dat.gz

[8] https://github.com/mpsilfve/pcfp-data/tree/master/data

[9] Accessed from https://github.com/sigmorphon/2023InflectionST/blob/main/part1/baselines/nonneural.py

[10] Accessed from https://github.com/omagolda/neural-transducer/tree/master/example/sigmorphon2023-shared-tasks

for SIGMORPHON-UniMorph shared tasks since Cotterell et al. (2017), and the neural baseline, a character-level transformer, since Pimentel et al. (2021). The non-neural records prefixing and suffixing rules, and then uses a matching heuristic to decide which rules to apply given the set of features.

We did not test systems other than the baselines, since the focus of this paper is issues with the gold data independent of algorithm choice. Also, code was not yet available for Canby and Hockenmaier (2023)'s top-performing neural systems; the other neural system was outperformed by the neural baseline anyway, and the submitted finite state transducer systems performed comparably to the non-neural baseline.

3.2.3 Error taxonomy

The error taxonomy we used is an extension of Gorman et al. (2019)'s annotation scheme for the 2017 SIGMORPHON-UniMorph shared task. Gorman et al. (2019, p. 142) split errors into four major categories, three of which are cited below and used in an identical fashion. We omitted spelling errors due to a lack of errors that differed only in spelling.

Since we were trying to account for the influence of Old French verbs in a given error, we added a superordinate category **Old French errors**.

Old French errors This category includes all errors that can be attributed to the presence of outdated verbs in training, development, and test data. We specified two sub-categories: (i) **Old French Lemma errors**, for Old French verb lemmas that are not used in modern French, i.e., extraction errors, and (ii) **Old French affix overapplication errors**, which involve applying Old French inflecting patterns learned from muddied data to modern verbs. These were not considered allomorphy errors because the Old French affixes did not constitute "existing allomorphic patterns in the target language" (Gorman et al., 2019), i.e. French.

Free Variation This category was the same as Gorman et al. (2019)'s free variation category and included verbs which have "free variation" in French, but where only one form was available due to the UniMorph scraping procedure. In these cases, the error was a grammatical form but not included as a correct form in the gold data.

Allomorphy Errors These were divided into two subcategories used in Gorman et al. (2019) but not reported in the paper[11]: (i) **Affix overregularization errors**: errors where the target irregular affix was replaced with one that is regular. (ii) **Affix overirregularization errors**: errors where regular affixes were replaced by irregular affixes.

Silly Errors This category was the same as Gorman et al. (2019)'s silly error category and encompassed cases where the model's prediction was extremely dissimilar to the gold data. This dissimilar form was not present elsewhere in the given inflectional category for the language. Silly errors included completely strange and random inflectional forms that differed greatly from the lemma and were primarily seen in inflection errors made by the neural model, see §4.2.1.

3.2.4 Annotation procedures

The annotation conducted for Gorman et al. (2019)'s experiment was annotated by native speakers and some by second-language speakers with expertise in computational linguistics. Our annotators fall into this second category, thus, annotation of the error data was carried out both as researchers with backgrounds in linguistics and as advanced French speakers.

Our error categorization used an order of priority similar to Gorman et al. (2019)'s, though starting with **Old French errors** and proceeding thereafter through **Free Variation**, **Allomorphy**, and **Silly**. However, the first step of this priority order was only applied very conservatively. The **Old French Lemma** error category was only applicable to those lemmas which, according to Wiktionary, were not modern French verbs. The **Old French affix overapplication** error category was only applied when we found the model had used a suffixing rule which exists nowhere in modern French but had been scraped from an Old French Wiktionary entry.

4 Results

4.1 Prevalence of Old French in past inflection tasks and UniMorph files

4.1.1 UniMorph 4.0 files

The `fra.args` file[12]—which seems to the source file for 2023 SIGMORPHON-UniMorph shared task splits—contained 20.8% (1564/7535) lemmas from Old French, and we confirmed that there were no old lemmas in `fra.segmentations`.

[11] Thanks to Kyle Gorman sharing full annotation scheme.

[12] and also the `fra` file, which is different from `fra.args` only in using the UniMorph 3.0 feature scheme

Year	Frequency of Old lemmas			Freq. of inflected forms from Old lemmas		
	Train	Dev	Test	Train (/10,000)	Dev (/1,000)	Test (/1,000)
2017	1,146 (20.5%)	182 (19.4%)	193 (20.5%)	2,045 (20.4%)	194 (19.4%)	206 (20.6%)
2018	1,165 (20.8%)	214 (22.6%)	203 (21.6%)	2,108 (21.1%)	221 (22.1%)	215 (21.5%)
2019	1,139 (20.5%)	N/A	N/A	2,052 (20.5%)	N/A	N/A
2023	86 (21.5%)	5 (10%)	9 (18%)	2,142 (21.4%)	100 (10%)	180 (18%)

Table 1: Raw and relative frequencies of Old French lemmas and forms inflected from Old French lemmas in SIGMORPHON-UniMorph shared task splits. Only high-resource training sets were included, see §3.1.1 for details.

4.1.2 SIGMORPHON-UniMorph shared task splits

We determined that Old French lemmas typically occurred in approximately 20% of each of the train/test/dev splits in past SIGMORPHON-UniMorph shared tasks involving French, as summarized in Table 1.

4.1.3 Minimal supervision inflection tasks

We found that Malouf (2017)'s French data extracted from Flexique contained a very small number (7 out of 5220) of Old French lemmas (see §5). This was unexpected since Flexique is based on post-1950s texts and subtitles from French movies. Additionally, Silfverberg and Hulden (2018)'s random sample of 1000 lemmas from French UniMorph 2.0 (`fr.um.V.txt`) contained 216 Old French lemmas (21.6%) while the high-frequency filtered sample (`fr.um.V.top.txt`) included only 114 historical lemmas (10.1%).

4.2 2023 shared task replication experiments

The number of distinct lemmas and forms for each split for all three experiments is given in Table 2. Filtering the original splits via `fra.segmentation` resulted in the loss of 22%, 10%, and 18% of lemmas in each split respectively. The form-based split ended up having 3 fewer lemmas because the matching performed in our script does not account for the inconsistencies in `fra.segmentations` and `fra` in representing reflexive forms.

Removing Old French lemmas improved the accuracy for the non-neural baseline ("RU") by 10.32-11.72% and for the neural ("NN") algorithm by 12.32-13.34% (Table 3). Whether the original splits were re-sampled by-lemma ("Seg") or by-form ("Seg-Minimal") made only about a 1% difference. Even with the re-sampled splits excluding Old French lemmas, the non-neural baseline still outperformed the neural baseline by about 10%.

Split	Original	Seg	Seg-Minimal
Train	400:10,000	312:15,890	309:6,407
Dev	50:1,000	45:2,265	45:754
Test	50:1,000	41:2,101	41:701

Table 2: Number of distinct lemmas:forms in each split for each experiment.

System	Original	Seg	Seg-Minimal
RU	83.15%	94.87%	93.47%
NN	71.40%	83.72%	84.74%

Table 3: By-form accuracy for the non-neural (RU) and neural (NN) models, aggregated across test and dev sets.

4.2.1 Error analysis

The distribution of error types (defined in §3.2.3) for each of the three experiments (original, Seg by lemma, Seg-Min by form) and algorithms is shown in Figure 1, combining errors across dev and test sets. Table A1 shows the raw counts for error types in dev and test sets separately. The abbreviations in the figure and table correspond to the error taxonomy categories as ordered in §3.2.3.

Old Lemma and Old French affix overapplication errors ("**Old Rule**") occurred only for the original splits, comprising 56.6% and 37.8% of errors for the non-neural and neural models, respectively. Example overapplication errors are in Table A2. For both models, the other major error type was affix overregularization ("**Over Reg**") allomorphy errors, comprising 33-37% of the errors for the original splits. Overregularization was the most frequent error in the Seg and Seg-Min resampling experiments—about 90% of errors for the nonneural model and 56-64% of errors for the neural model. The neural model differed from the nonneural model in having many silly errors—20% of the errors for the original splits and 32-40% for the Seg-resampled splits.

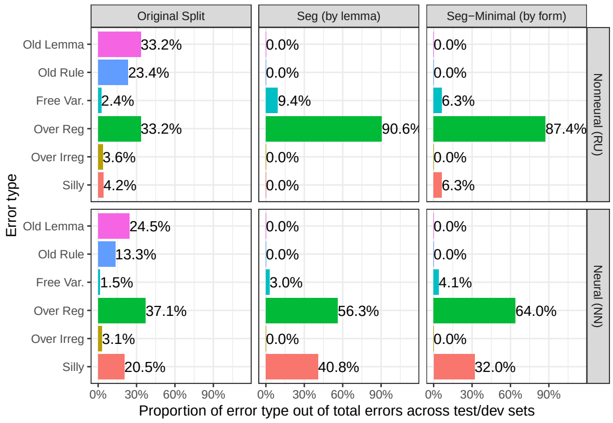

Figure 1: Proportion of error type out of all errors across test/dev sets for neural (NN) and non-neural (RU) baselines for each experiment.

5 Discussion

We determined that Old French comprised around 20% of the lemmas in the SIGMORPHON-UniMorph 2017, 2018, and 2023 shared task splits, as well as Silfverberg and Hulden (2018)'s random sample of 1,000 inflection tables from French UniMorph. These were all cases when French yielded anomalously poor performance relative to other languages. However, Silfverberg and Hulden (2018)'s sample of French UniMorph filtered for high frequency forms, which yielded better performance, only had 10% Old French lemmas, and Malouf (2017)'s French data from Lexique that yielded near-perfect accuracy had only 0.13% Old French lemmas.

In short, performance on French inflection tasks was inversely proportional to the proportion of Old French lemmas present in the task data. Furthermore, errors related to the presence of Old French in the data were prevalent in our replication of the 2023 shared task with the original task splits for both non-neural and neural models. Removing Old French from the splits eliminated these errors. Interestingly, the non-neural baseline still outperformed the neural baseline even when Old French lemmas were removed. Thus, the presence of Old French in the original 2023 task splits doesn't seem to be the cause of the the non-neural baseline outperforming the neural model.

These improvements suggest that correctly separating modern, Old, and Middle French into separate datasets is important for computational morphology tasks. UniMorph itself has separate repositories for Old (`unimorph/fro`) and Middle (`unimorph/frm`) French, so the erroneous inclusion of lemmas from both Old and Middle French creates confusing inconsistencies for potential future projects which may want to work on all three languages.

This bug was noted and addressed in the UniMorph 3.0 revision (McCarthy et al., 2020): "Finally, a bug in the previous extraction process caused languages' data to be read into other languages' files whose names are their suffixes. For instance, 'Greek' contained data from 'Ancient Greek', and 'French' contained data from 'Middle French'. Filtering and rerunning our extraction process eliminated these erroneously grouped paradigms" (McCarthy et al., 2020, p. 3924). However, the issue persists in the `fra` and `fra.args` data files in UniMorph 3.0 and UniMorph 4.0.

It is worth noting that the task of distinguishing Old and modern French involves a degree of nuance. The sampled Flexique data contained 7 lemmas which were classified as Old French by our script, but according to the French edition of Wiktionary, these words have been repurposed as either idiomatic expressions or legal terms in modern French, now using modern inflection patterns. While future projects should be sure that their data makes this distinction, simple scraping of the English edition of Wiktionary may present issues for obtaining truly representative lexical data.

5.1 Data Inconsistencies

Elsner et al. (2019, p. 78-79) notes that none of the SIGMORPHON datasets provide an adequate lexical set to account for the Zipfian distribution of words in natural language. For example, "spotty coverage of high frequency words for German appears to be typical of the UniMorph datasets." Similarly, we found that our splits lacked highly frequent verbs such as *être* ('to be'), *faire* ('to do'), and *pouvoir* ('to be able to'), which were included in the more exhaustive `fra.segmentations`. Despite the limited size of the training data, we nevertheless noticed some further data inconsistencies that would have caused more issues if they had been included in the dataset to the extent that they are represented in the language. This includes (i) inconsistency in the documentation of French reflexives in Wiktionary, and (ii) the presence of multiple possible grammatical inflections for verbs such as *-eler* and *-eter* verbs.

5.1.1 Reflexive inconsistencies

Reflexive verbs in French include a reflexive pronoun *se* (onesself) that is the object of the verb, e.g. *il se regarde* ('he looks at himself'). Despite there being no reflexive verbs in the test or development splits, reflexives verb forms are quite common in the French language. The three lemmas that did appear with reflexive pronouns in the training split were inconsistently recorded (two had reflexives pronouns in the inflected forms but not in the lemma, while the third had a reflexive pronoun in the lemma as well). These inclusions were enough to cause the neural model to erroneously identify *génuflexionner* ('to bend the knee') as a reflexive verb, though this verb takes no object.

Had more reflexives been included in the train and test splits, the effect of inconsistent data on the models' accuracies would have been much greater. These inconsistencies include duplicate pages, transitive verb pages with "reflexive" usage shown in the definitions but not in the conjugations, and those listed as transitive but conjugated using reflexive pronouns. Many of the most common reflexive verbs are entirely missing or have been deleted due to differing opinions on the necessity of the reflexive form having separate documentation. Had they been included in our splits, we predict that the inconsistencies would have posed issues for properly measuring each model's performance on French.

5.1.2 Multiple grammatical inflections

There exists a prescriptivist body in the French government, l'Académie Française, which is tasked with publishing the French dictionary as well as setting official orthography changes in the language over time. This has resulted in a degree of free variation in the inflection of French verbs. In accordance with the Académie's prescriptions, Wiktionary has a number of French verb charts that have multiple options mapped to a single morphosyntactic tag, where UniMorph only scrapes one option per lemma/feature pair. The most common of these are *-eler* and *-eter* verbs, which can now be conjugated by either doubling the consonant or adding an è before said consonant, except for those derived from *appeler* ('to call') or *jeter* ('to throw').

Since UniMorph only scrapes one option, when models predict one of the other permitted conjugations, they are marked incorrect. There was only one of the aforementioned *-eler* and *-eter* verbs in our data, *craqueler* ('to crack') in the dev split. The errors that resulted from this free variation were noted in the annotation scheme, but such errors would be much greater in number if the data had been more exhaustive. By performing a more inclusive scrape of Wiktionary that grabs all of the grammatical inflected versions of a lemma with a given morphosyntactic tag,[13] we predict there would be an increase in accuracy since this would mark inflections correct that would previously have been erroneously marked as a mistake in the predicted form.

5.2 Proposed fixes for French

We propose that future shared inflection tasks use `fra.segmentations` rather than `fra/fra.args`, which would eliminate all errors that fell under the **Old French error** category in our taxonomy. The improvements to accuracy as a result of this change are reflected in our results. Using `fra.segmentations` instead would also allow more comprehensive inclusion of common French verbs, which would generate results that are more reflective of how these models handle the French lexicon.

Additionally, we advise caution in scraping French reflexive verbs from the English edition of Wiktionary, as well as verbs with free variation, as described in §5.1. Wiktionary is subject to inconsistencies as well as disagreement between Wiktionary entry authors despite its richness in linguistic data.

Finally, as French is a very well-documented language, there are several other resources for linguistic data which may be more consistent and reliable than the English edition of Wiktionary. These could help circumvent data consistency issues in future computational linguistics tasks. For example, the Morphalou3 lexicon takes into account purely orthographic variations on individual words, including those allowed by the additional rules prescribed by the French Academy in 1990, and is a consolidation of Morphalou with 4 other French lexicons (DELA, Dicollecte, LGLex/LGLexLefff, and Lefff) (ATILF, 2023). The GLÀFF lexicon (Hathout et al., 2014) is specifically based on the French edition of Wiktionary and thus does not have the same consistency issues as the English edition. It also includes the overall frequency of each lexeme (per million

[13] Malouf et al. (2020, §3.4) has an alternative suggestion: to remove paradigms with multiple grammatical inflections from the data.

words across various large French corpora), which would be helpful in selecting for more common words when designing train/test/dev sets.

5.3 Beyond French

It was only because of our in-depth attention to French results and our particular linguistic knowledge of French that we were able to spot the erroneous inclusion of Old French in the UniMorph data and then perform the qualitative error analyses in this paper. There are, no doubt, other UniMorph languages which could benefit from similar language-specific studies. Yet, while the cross-linguistic coverage of UniMorph and SIGMORPHON-UniMorph shared tasks has rapidly expanded across the past decade, detailed, language-specific analyses of UniMorph data and/or SIGMORPHON-UniMorph results remain few in number.

Studies that have examined particular languages in detail have found issues with Wiktionary data and/or extraction errors—for instance, in Romanian, Hungarian, and Latin (3 of the 12 languages examined in the error analyses of Gorman et al. (2019)), as well as Navajo (Malouf et al., 2020). In an examination of UniMorph data, Malouf et al. (2020) raises some of the same issues that we discussed in §5: limited size of data sets and the availability of multiple grammatical inflectional forms for a single paradigm cell. Malouf et al. (2020) points out that there are several inconsistencies in choices made by Wiktionary editors for Navajo entries which negatively affect the overall performance of morphological inflection models when using Navajo data from UniMorph. For instance, Wiktionary provides separate entries for bare nouns and their possessed forms for some but not all Navajo lemmas. While the possessed forms should certainly be included, the decision to keep the entries separate for certain nouns is confusing and causes some inflected forms to be treated as lemmas in their own right.

6 Conclusion

When shared tasks include dozens and dozens of languages, it is hard to interpret results when each individual language could be affected by data issues like those we have discussed in this paper. Such problems underscore the need for shared tasks to include qualitative, language-by-language analysis of data and results in addition to reporting accuracy. It is admittedly a tall order to do analyses like the one in this paper for each of the over two dozen languages from the SIGMORPHON-UniMorph 2023 shared task, but perhaps shared tasks could explicitly focus on probing and improving data quality and otherwise emphasize language-by-language error analysis as an essential step of analyzing results. This kind of work would naturally encourage collaboration between language experts/linguists and modelers, as suggested in Malouf et al. (2020)'s statement of best practices for computational modeling of cross-linguistic morphology. By closely examining the distribution of errors produced, future projects can concentrate on eliminating prevalent error categories that have previously hindered model performance, enabling focused improvements in shared tasks.

Acknowledgements

We thank Kyle Gorman for discussion of Gorman et al. (2019), UniMorph mailing list (https://groups.google.com/g/unimorph) members for responding to questions about French UniMorph, and members of the UMass Amherst LINGUIST 409 Introduction to Computational Linguistics Fall 2023 course for helpful feedback.

References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

ATILF. 2023. Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóğa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Olivier Bonami, G. Caron, and C. Plancq. 2013. Flexique : An inflectional lexicon for spoken French.

Marc Canby and Julia Hockenmaier. 2023. A Framework for Bidirectional Decoding: Case Study in Morphological Inflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4485–4507, Singapore. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L. King, Luana Lamberti Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):53–98.

Frédéric Godefroy. 1881. *Dictionnaire de l'ancienne Langue Française et de Tous Ses Dialectes Du IXe Au XVe Siècle*. F. Vieweg, Paris.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2021. Minimal Supervision for Morphological Inflection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2088, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird Inflects but OK: Making

Sense of Morphological Generation Errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1007–1012, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Boris New, C. Pallier, Ludovic Ferrand, and Rafael Matos. 2001. Une base de données lexicales du français contemporain sur internet : LEXIQUE™//A lexical database for contemporary french : LEXIQUE™. *L'année psychologique*, 101(3):447–462.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Miikka Silfverberg and Mans Hulden. 2018. An Encoder-Decoder Approach to the Paradigm Cell Filling Problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Appendix

	Old Lemma	Old Rule	Free Var.	Over Reg	Over Irreg	Silly	Total
Orig/RU/dev	33	31	8	34	8	0	114
Orig/NN/dev	30	30	7	64	8	50	189
Orig/RU/test	79	48	0	78	4	14	223
Orig/NN/test	82	31	0	106	6	44	269
Seg/RU/dev	0	0	21	41	0	0	62
Seg/NN/dev	0	0	21	123	0	125	269
Seg/RU/test	0	0	0	162	0	0	162
Seg/NN/test	0	0	0	277	0	165	442
Min/RU/dev	0	0	6	18	0	1	25
Min/NN/dev	0	0	9	46	0	52	107
Min/RU/test	0	0	0	65	0	5	70
Min/NN/test	0	0	0	96	0	19	115

Table A1: Frequency of errors types for dev and test splits for experiment (original, Seg, Seg-Minimal) and algorithm (RU vs. NN). The errors are listed left to right in the order of taxonomy priority.

lemma	features	gold	model prediction
absoudre	COND.3SG IND.PST.PFV.2PL	absoudrait absolûtes	absoud**roit** absou**istes**
désarmer	COND.2SG IND.PST.IPFV.3SG IND.PST.PFV.1PL	désarmerais désarmait désarmâmes	désarmer**ois** désarm**oit** désarm**asmes**
délayer	IND.PST.IPFV.3PL IND.PST.IPFV.1SG	délayaient délayais	délay**oient** délay**ois**
alanguir	IND.PST.PFV.1SG	alanguis	alangu**a**
abonder	IND.PST.PFV.2PL	abondâtes	abond**astes**
mendier	SUBJ.PST.3SG	mendiât	mend**ast**
tuner	SUBJ.PST.2PL	tuneriez	tuniss**oiz**
objectiver	SUBJ.PRES.2PL	objectiviez	objectiv**ez**

Table A2: Examples of modern French verbs erroneously inflected with Old French suffixes. Triples mentioned in §1 in first three columns, fourth column is an error that falls into the **Old French affix overapplication** (**Old Rule**) category. Refer to yellow-highlighted data in `ErrorAnnotations.xlsx` in the GitHub repository.