Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection

Basilio Calderone, Nabil Hathout, Olivier Bonami

Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection

Basilio Calderone[a] **Nabil Hathout**[a] **Olivier Bonami**[b]
[a]CNRS, CLLE, Université de Toulouse [b]Université de Paris, LLF, CNRS
basilio.calderone@univ-tlse2.fr
nabil.hathout@univ-tlse2.fr
olivier.bonami@u-paris.fr

Abstract

The paper presents four models submitted to Part 2 of the SIGMORPHON 2021 Shared Task 0, which aims at replicating human judgements on the inflection of nonce lexemes. Our goal is to explore the usefulness of combining pre-compiled analogical patterns with an encoder-decoder architecture. Two models are designed using such patterns either in the input or the output of the network. Two extra models controlled for the role of raw similarity of nonce inflected forms to existing inflected forms in the same paradigm cell, and the role of the type frequency of analogical patterns. Our strategy is entirely endogenous in the sense that the models appealing solely to the data provided by the SIGMORPHON organisers, without using external resources. Our model 2 ranks second among all submitted systems, suggesting that the inclusion of analogical patterns in the network architecture is useful in mimicking speakers' predictions.

1 Introduction

Psycho-computational experiments deal with the capability of the computational models of language to capture and mimic the behavioural responses of speakers exposed to the same data stimuli. In phonology, and in morphology, a large number of computational models (both symbolic and subsymbolic) have been tested on various experimental data in order to evaluate their capacity to simulate linguistic behavior (Rumelhart and McClelland, 1986; Nosofsky, 1990; Albright and Hayes, 2003; Hahn and Bailey, 2005; Hay et al., 2004; Albright, 2009).

More recently, Kirov and Cotterell (2018) argue that a neural encoder-decoder network (ED) can perform morphological inflection tasks in a cognitively valid manner. In particular, the authors claim that, in a *wug* test protocol, ED's outputs significantly correlate with human judgements for nonce verbs supporting the assumption that the model learns representations of specific knowledge and involves cognitive mechanisms that are known to underlie language processing in the speakers. However, whether and how such models are able to mimic the human behaviour of subjects exposed to the same stimuli is still an open question (Corkery et al., 2019). Part 2 of Shared Task 0 addresses this same issue. More specifically, it adopts the experimental approach of Albright and Hayes (2003). The task is to design models which predict the inflected forms of a set of nonce verbs in a given language and that have output scores of the predictions that correlate with human judgements.

In this paper we report on a series of experiments that address the shared task by exploring whether pre-learned formal analogies between strings can be usefully combined with an ED architecture to alleviate some limitations of the application of an ED architecture to raw forms. We present two types of models using analogical patterns in the input (M1) vs. in the output (M2) of an ED network, and compare their performance to that of two baselines that focus respectively on the phonotactic typicality of outputs (M3) and the type frequency of alternations (M4). We report good performance for the M2 architecture and quite poor performance for M1.

In Section 2, we shortly present the data provided by the organisers. Section 3 focuses on the analogical patterns we use in our models. The models architecture, the training parameters and the results are reported in section 4. We then discuss the results in Section 5 and draw conclusions (Section 6).

2 Data and goal

The linguistic data provided by the organisers are inflected verb forms in English (ENG), German

(DEU) and Dutch (NLD). For each language, four datasets are released: (a) a training dataset, (b) a development dataset of attested verb forms, (c) a development dataset of *wug* forms which includes human judgements, and (d) a test dataset of *wug* forms without the human judgements. The goal of the Shared Task is to assign a score to each *wug* form in (d) that correlates as closely as possible with the human judgements (see Section 4.2 for more details on the evaluation process).

The entries of the datasets include lemma/form pair and a UniMorph (Kirov et al., 2018) tag (UT) specifying the part of speech and the paradigm cell of the form. The pairs are provided as written forms (orthographically) in datasets (a) and (b) and in IPA (phonologically) in all four datasets.

Table 1 summarizes the size of the training data (a), its phonological make-up, the number of morphosyntactic tags and the proportion (%) of syncretism, i.e. of forms that fill two or more paradigm cells of the same lexeme. Although the number of phonemes is substantially similar in the three languages, the datasets differ in the number of entries (twice as many entries in DEU as in ENG) and number of cells. The three datasets have a comparable amount of syncretism.

	ENG	DEU	NLD
entries	41 658	100 011	74 176
phonemes	43	44	39
UTs	11	29	7
syncretism %	53	50	42

Table 1: Training data

3 Analogical Patterns

Three of our model architectures rely on alternation patterns (APs) describing the formal relationship between two word-forms. An AP is a pairing of two word-forms patterns (WPs) with shared variables over substrings which represent word parts that are common between the two forms. For example, the two German word-forms *anʃpiːlən* 'to allude to' and *anʃpiːlənt* 'alluded to' are related by the pattern (an$X$ən, an$X$ənt) where the variable X represents ʃpiːl.

The number of different APs satisfied by a pair of forms is typically large. For instance, there are 256 (2^8) distinct patterns relating *anʃpiːlən* and *anʃpiːlənt*, some of which are shown in (1), where italic capital letters represent variables over strings.

(1)

WP_1	WP_2	
anʃpiːlən	anʃpiːlənt	← TAP
Xnʃpiːlən	Xnʃpiːlənt	
aXʃpiːlən	aXʃpiːlənt	
an$X$piːlənt	an$X$piːlən	
aXʃYiːZəT	aXʃYiːZəTt	
XnYpZlTn	XnYpZlTnt	
Xən	Xənt	← FAP
aX	aXt	
XnY	XnYt	
Xn	Xnt	
X	Xt	← BAP

Henceforth we will note patterns using the '+' symbol as a general notation for variables, and rely implicitly on order to match variables in alternations. Hence e.g. the AP (XnYpZlTn, XnYpZlTnt) will be noted +n+p+l+n/+n+p+l+nt.

Most of the patterns in (1) are of little morphological interest. This is in particular the case of the trivial alternation pattern (TAP) which just records the two strings without making any generalization over common elements. A broad alternation pattern (BAP) is an optimal pattern that can be inferred by pairwise alignment of the two forms, without taking into account the situation in the rest of the paradigm. In principle there can be more than one BAP for a pair of form, although that rarely happens in practice. This type of pattern is of crucial interest to the study of the implicative structure of paradigms (Ackerman et al., 2009; Ackerman and Malouf, 2013) and the induction of inflection classes (Beniamine et al., 2017), but does not lead to the identification of affixes familiar from typical grammatical descriptions. For that purpose, multiple alignments across the paradigms are necessary (Beniamine and Guzmán Naranjo, 2021), and lead to what we call a fine alternation pattern (FAP): here *ən* is the infinitive suffix and *ənt* is the present participle suffix.

The crucial intuition behind the determination of FAPs is that they identify recurrent partials (Hockett, 1947) across both paradigms and lexemes. For instance, the FAP in (1) is motivated by the fact that the substrings *an* and *ʃpiːl* are shared across all pairs of paradigm cells of *anspielen* (2), while the substrings *ən* and *ənt* recur in many (infinitive, present participle) pairs across lexemes (3).

(2)

anʃpiːlən	ʃpiːlət+an	V;IND;PST;2;PL
anʃpiːlən	ʃpiːltə+an	V;SBJV;PST;1;SG
anʃpiːlən	ʃpiːlt+an	V;IMP;2;PL
anʃpiːlən	ʃpiːlə+an	V;IMP;2;SG
anʃpiːlən	anʃpiːləst	V;SBJV;PST;2;SG
anʃpiːlən	anʃpiːlst	V;IND;PRS;2;SG

(3)

tariːfiːrən	tariːfiːrənt	V.PTCP;PRS
‘to tariff’		
tariːfiːrən	tariːfiːrtən	V;SBJV;PST;3;PL
astən	astənt	V.PTCP;PRS
‘to lug’		
astən	astətət	V;SBJV;PST;2;PL
vaiˆnən	vaiˆnənt	V.PTCP;PRS
‘to cry’		
vaiˆnən	vaiˆnt	V;IND;PRS;3;SG
tsɛrʃtraiˆtən	tsɛrʃtraiˆtənt	V.PTCP;PRS
‘to disagree’		
tsɛrʃtraiˆtən	tsɛrʃtraiˆtəst	V;SBJV;PST;2;SG

In this paper, we rely on an algorithm for inferring BAPs and FAPs initially developed to create Glawinette (Hathout et al., 2020). Glawinette is a French derivational lexicon created from the definitions of the GLAWI machine readable dictionary (Sajous and Hathout, 2015; Hathout and Sajous, 2016). Glawinette provides a description of derivational morphology by means of morphological families and derivational series; it is part of an effort aiming at the design of derivational paradigms. BAPs and FAPs have been adapted to the datasets of the current task, analogizing inflectional paradigms to derivational families and pairs of inflectional paradigm cells to derivational series. For instance, (4) presents an excerpt of an inflectional series in the inflectional paradigm of the verb *anspielen* that realizes the features `V;NFIN` and `V.PTCP;PST`. In turn, this series yields two series of word-forms, the ones in the left column and the ones in the right column.

(4)

apzuːxən	apgəzuːxt	‘to search’
aplɔxən	apgəlɔxt	‘to punch off’
aprʏkən	apgərʏkt	‘to disengage’
apgʊkən	apgəgʊkt	‘to peek’

Basic preprocessing. The forms in the test set (d) being in IPA, we only computed phonological BAPs and FAPs. BAPs and FAPs have been computed for all the entries of all four datasets. In addition, two basic modifications were performed. First, particles were reorder so as to appear in the same position in the infinitival and inflected word-forms. For instance, *wechsele über*, *vɛksələ yːbər* ‘to switch over’ is reordered as *yːbərvɛksələ*. Second, all phonemes represented by digraphs and trigraphs in IPA were replaced with arbitrary unigraphs (capital letters; eg. `S` is substituted for *yː*).

Broad alternation patterns. Each entry in the datasets consists of the infinitive and another form of some lexeme, accompanied by the UT of the second form. The BAP of a pair of forms is computed through an alignment of the two word-forms and the identification of their common parts and their differences. The alignment is computed by means of the `SequenceMatcher` method of the python `difflib` library; we then go through the sequences provided by the method and create the word-form patterns by replacing the common parts by + and copying the differences in their respective patterns. For example, `SequenceMatcher` aligns the forms *aptaiˆlən* and *apgətaiˆlt* as in (5) which yields the `++en/+ge+t` BAP. BAPs are therefore calculated separately for each entry considering only the two forms.

(5)

word-form1	ap		taiˆl	ən
word-form2	ap	gə	taiˆl	t
BAP_1	+		+	ən
BAP_2	+	gə	+	t

Note that a BAP can also be seen as a characterization of an analogical series. For instance, the pairs of forms in (4) can all be aligned in exactly the same way as in (5), they all have the same BAP `++ən/+gə+t` and they form formal analogies (Lepage, 1998, 2004b,a; Stroppa and Yvon, 2005; Langlais and Yvon, 2008). More specifically, if two pairs of forms (F_1, F_2) and (F_3, F_4) have the same BAP, then $F_1 : F_2 :: F_3 : F_4$. BAPs could also be computed for entire inflectional paradigms as proposed by Hulden (2014). Also note that BAPs are not specific to an inflection class, as two classes may exhibit common behavior in one part of their paradigm but not in another. For instance, the BAP `+/+s` describes the formal relation that connects the infinitive and the `V;PRS;3;SG` form of both regular (*work*) and irregular English verbs (*eat*).

Fine alternation patterns. Unlike BAPs which are derived solely from the examination of pairwise alternations, FAPs rely on the place of the two word-forms in the overall morphological system to identify more stable recurrent partials corresponding to traditional exponents. For instance, the BAP relating the German weak verbs like *anspielen* to

its present participle *anspielend*, relying on the optimal alignment between the two forms, does not identify the infinitive and past participle exponents *-en* and *-end*. These cannot be deduced from an isolate pair of word-forms, and require considering, across the paradigm, all the pairs of word-forms that include infinitives or present participles and finding out the pair of endings that best characterizes, across lexemes, the infinitives "similar" to *anspielen* and the present participles "similar" to *anspielend*. The main challenges in the identification of the FAPs are then (*i*) that they involve the entire dataset and cannot be computed locally for a single pair of word-forms; (*ii*) that we need to formally define what "similar" means; (*iii*) that we potentially need to consider all the APs of all the pairs of words included in the dataset; (*iv*) that we need a reasonable operational approximation of what could considered as linguistically relevant.

(*i*) The regularities that determine the FAPs are holistic properties of the dataset, i.e. of the union of the datasets (a), (b), (c) and (d). The consequence is that each FAP depends on the entire dataset, and FAPs have to be recomputed each time any of the datasets (a), (b), (c) or (d) is modified.

(*ii*) The pairs in (4) are good examples of what similar may mean, from an inflectional point of view. This type of similarity can be defined in terms of analogy. We first assume that pairs of forms that satisfy the same BAP constitute an analogical series (as they satisfy the formal analogy encoded by the BAP). Word-forms belonging to the same column in an analogical series are then considered as similar. In our example, the word-forms in each column in (4) count as similar.

(*iii*) We limit the number of patterns to be considered by looking only at the ones that are involved in the characterization of similar word-forms. In other words, once the sets of similar word-forms are created, we only consider the similarities that exist between the word-forms that belong to each set, since only these ones may be part of an FAP.

Let Φ be the set of pairs of word-forms satisfying some BAP, and Φ_1 (resp. Φ_2) the set of word-forms that are the first (resp. second) element of a pair in Φ. What we are looking for are the patterns that characterize a large enough subset of the word-forms in Φ_1 that are in correspondence with patterns that characterize a large enough subset of the word-forms in Φ_2, i.e. such that the pair of patterns characterize a large enough subset of the pairs in Φ. These APs are obtained as follows.

We first collect the WPs that possibly characterize the word-forms in Φ_1 by computing a pattern of word-forms for each pair of word-forms made up of two word-forms from Φ_1. These patterns are dual of the ones illustrated in (5) as we need to represent what the word-forms have in common and not their differences. For instance, in the second column of (1), the pattern that describes the common part of *apgəzuːxt* and *apgəlɔxt* is `apgə+xt` and the one for the common part of *apgəlɔxt* and *apgərʏkt* is `apgə+t`. If the number of word-forms in the column is large and varied enough, all the relevant WPs that characterize a part of the word-forms will be collected. We then align the patterns obtained for the two column. This is done by considering the WPs as if they were word-forms and computing their analogical signature, i.e. their BAP. For instance, we have *aplɔxən : aprʏkən :: apgəlɔxt : apgərʏkt*. The BAP for the first pair is `++ən/+gə+t` and the BAP for the second is identical; the pattern that characterizes aplɔxən:aprʏkən is `ap+ən` and the one that characterizes the second is `apge+t`. These two WPs are aligned because their BAP is `++ən/+gə+t`, i.e. the same as the BAP of the two pairs of word-forms.

By doing the same computation for all pairs of word-forms and matching them with respect to their BAP, we end up with a number of FAP candidates that we first screen in order to exclude those that describe only a small part of Φ, or that are made up of WPs that describe a small part of Φ_1 or Φ_2. Another feature that helps select valuable FAPs is the number of variable parts (+) it contains. For our models, we only used FAPs that contain exactly one variable part, but this number could be increased for languages with templatic morphology or that make use of infixes.

(*iv*) We assume that optimal FAPs are pairs of WPs that recur both within the paradigm and across the lexicon, as we illustrated in (2) and (3). For instance, the FAP of a pair of word-forms *anʃpiːlən* and *anʃpiːlənt* consists of the aligned patterns describing the largest number of word-forms similar to *anʃpiːlən* on the one hand, and the largest number of word-forms similar to *anʃpiːlənt* on the other hand. It turns out that this is the pattern `+ən/+ənt`. More precisely, let (F_1, F_2) be a pair of word-forms and let $\{(P_1, Q_1), (P_2, Q_2), ..., (P_n, Q_n)\}$

be the FAP candidates connecting F_1 and F_2 (i.e. the set of the aligned WPs of F_1 and F_2). Let $|X|$ be the number of form pairs that satisfy an alternation that contain the WP X. The FAP of (F_1, F_2) is then the WP pair (P_i, Q_i) such that $|P_i| + |Q_i| = \max_{j=1}^{n}([P_j| + |Q_j|)$. FAPs are therefore selected separately for each pair of word-forms.

The models M1 and M2 presented in Section 4 use FAPs computed from the union of the datasets (a), (b), (c) and (d) for each of the three languages of the task.

Discussion. BAPs and FAPs give different types of information: BAPs capture relations between pairs of forms independently of the rest of the system, and are hence crucial to addressing the implicative structure of paradigms (Wurzel, 1989). FAPs on the other hand characterize a pair of forms taking into account their place in the rest of the system; this typically leads to more specific patterns that are satisfied by a smaller set of pairs.

4 Combining analogical patterns and encoder-decoder networks

Early work on connectionist models of the acquisition of morphology involved pattern associators that learn relationships between a base lexical form (i.e. the lemma) and a target form (i.e. the inflected form). For example, Rumelhart and McClelland (1986) propose a simulation of how English past tense is learned. They focus on pairs of verb forms like *go-went* and *walk-walked* and consider that morphological learning is a gradual process which includes an intermediate phases of "over-regularization" (where the past form of *go* is *goed* instead of *went*). This yields the well-known"U-shape" curves observed in the developmental phases of morphological competence in children.

More recently, models based on deep learning architectures have been used (Malouf, 2017) and in particular sequence-to-sequence models able to predict one form of a lexeme from another (Faruqui et al., 2016; Kirov and Cotterell, 2018).

These approaches are based on the assumption that the morphological learning can reduce to a simple mapping between a base form and an inflected one. Generalizations over similar mappings (e.g. *love-loved*,*walk-walked* vs. *sing-sang*, *ring-rang*) are learned from the dependences between the phonemes in sequences. The APs presented in Section 3 provide a description complementary to the lemma-form mapping in which analogical regularities may be locally to a single pair of forms (BAPs) or globally from the entire lexicon (FAPs). These paradigmatic analogies emerge when the forms are contrasted with all other forms of their lexeme and the other forms that occupy the same cell in the paradigm (Bonami and Beniamine, 2016; Ahlberg et al., 2015; Albright, 2002).

The models we designed for the task combine the capacity of the sequence-to-sequence models to learn the regularities present in strings of phonemes with the alternation patterns acquired from the paradigms, in order to predict native speaker responses in a *wug* test.

4.1 Models

We designed four models for the shared task.

4.1.1 Model 1

In the first model, M1, we consider morphological inflection as a mapping over sequences. The mapping is implemented by bidirectional LSTMs with dropouts (Hochreiter and Schmidhuber, 1997; Gal and Ghahramani, 2016). The hidden states of the encoder are used to initialize the decoder states. The model adopts a teacher forcing strategy to compute the decoder's state in the next time-step. M1 takes as input four sequences: the lemma (IPA-encoded), the UT, the BAP and the FAP patterns. The output sequence is the inflected form (IPA-encoded). The output layer uses a sigmoid to produce a probability distribution over the output phonemes.

M1 Input: {lemma + UT + BAP + FAP}
Output: {inflected form}

The probability score assigned to the *wug* forms is the joint probability of the its phonemes. The model M1 addresses the task directly. We expected the prediction of a model that uses all the available information including BAPs and FAPs would be accurate and highly correlated with the judgments of the speakers.

4.1.2 Model 2

The second model, M2, relies on FAPs to identify the crucial thing to be predicted in a wug task, namely the inflectional pattern of the output form. Hence the model is trained to predict, instead of the raw output form, the word pattern that constitutes the second part of the FAP (FAP_2) and identifies

its place in the inflection system while abstracting away from what is common between the input and output forms.

M2 Input: {lemma + UT}
Output: {FAP_2}

Computationally M2 is similar to M1 except for the input/output sequences involved. In particular, the probability score of a *wug* form is the joint probability of the output symbols.

4.1.3 Model 3

Our third model, M3, estimates a possible wordlikeness effects due to phonological similarity of the inflected forms that have the same UT. Wordlikeness is the extent to which a sound sequence composing a form is phonologically acceptable in a given language. It mostly depends on the phonotactic knowledge of the speakers (Vitevitch and Luce, 2005; Hahn and Bailey, 2005) and on the existence of phonologically similar words in the mental lexicon (Albright, 2007). For example, a *wug* past form like *saɪndɪd* included in the English test dataset could trigger wordlikeness effect because it is similar to an attested past form *saɪdɪd* (*sided*). For each of the three languages, we designed a classifier which predicts whether an inflected form is assigned to a specific UT in the train set. The target UTs are the ones of the inflected forms in the three test sets (d), namely`V;PST;1;SG` for ENG, `V;PST;PL` for NLD and `V.PTCP;PST` for DEU. Technically, for each language, the M3 model is an LSTM-based binary classifier which takes the inflected form as input and outputs whether it is assigned to the target UT (value 1) or not (value 0). At training time, the forms which are assigned to the target UT and to another UT, are only kept with their target UT.

M3 For inflected UT in the test set (d),
Input: {inflected form}
Output: {0,1}

The score assigned to the *wug* form is simply the probability outputted by the system.

4.1.4 Model 4

Our fourth model, M4, simply uses the type frequency of the BAP relating the *wug* lemma and the *wug* form as a score for the test dataset.

M4 Raw type frequency of the BAP relating the *wug* lemma and *wug* form

This is meant as a very simple baseline, capturing in a very crude fashion the intuition that speakers judge as more natural wugs that fit into a more frequent pattern.

4.2 Results

The submissions to the task are evaluated using the AIC scores from a mixed-effects beta regression model (Magnusson et al., 2017) where the scaled human ratings (DV) were predicted from the submitted model's ratings (IV). The regression implements a random intercept for lemma type. Table 2 reports the AIC scores of the test data for the three languages.

Models	ENG	NLD	DEU
M1	−33.4	−60.0	−16.1
M2	**−43.0**	**−66.0**	**−98.8**
M3	−37.5	−64.9	−12.9
M4	−40.7	−36.8	−72.9

Table 2: AIC scores calculated on the basis of the final test data. Lower scores are better.

5 Discussion

The performance of our four models suggest the following observations. First, M2 outperforms our three other models for all three languages, and also ranked second of all systems submitted to the shared task. Second, there is a striking difference in performance between M2 and M1, which had an similar architecture, but performed very poorly—worse than our baseline M4 model, and second to last of all systems submitted to the shared task. Although more experiments are needed to conclude on this point, we conjecture that the better performance of M2 might be due to the fact that it abstracts away from the question of predicting the shape of the stem in the output, but focuses instead on that part of the inflected form that is not to be found in the input. This seems to match intuitions about human behavior: when dealing with inflections, speakers may have a hard time applying the right pattern, but they never have a hard time remembering what the stem looks like, even if it is phonotactically unusual (see Virpioja et al. 2018 for a psycho-computational study).

The other surprising result is that M4, which was intended as a crude baseline, did surprisingly well on the English and German data, although it performed very poorly on Dutch. This is interest-

ingly complementary to the performance of M3, which did surprisingly well on Dutch but poorly on German. As M3 is entirely focused on phonotactic similarity while M4 is focused on the frequency of alternations, this suggests that the three inflection systems (to the extent that they are faithfully represented by the datasets) raise different kinds of challenges to speakers.

To better assess the quality of M2, we examined how well it statistically correlates with human performance in Albright and Hayes's (2003) experiments on islands of reliability (IOR) in regular and irregular inflection in English. Albright and Hayes are trying to establish that speakers rely on structured linguistic knowledge as encoded in their Minimal Generalization Learner (MGL, Albright and Hayes, 2006) rather than pure analogy when inflecting novel words. To establish this, they collected both productions of human participants asked to inflect a novel word, and judgments on pairs of word-forms. They show that the MGL leads to a better correlation with human results than a purely analogical system based on Nosofsky (1990) (NOS in the table below). As Table 3 shows, our M2 performs at a level comparable to the MGL. More precisley, it clearly outperforms it on irregular verbs while trailing on regulars. Importantly, M2 does that without relying on any structured knowledge of the kind found in the MGL, although it does rely on a more complete view of the morphological system. This suggests that the conclusions of Albright and Hayes should be reconsidered.

Models	Ratings		Production probabilities	
	reg.	irr.	reg.	irr.
MGL	0.745	0.570	0.678	0.333
NOS	0.448	0.488	0.446	0.517
M2	0.583	0.595	0.611	0.560

Table 3: Pearson correlations (r) of participant responses to models. Core IOR verbs ($n = 41$). See Albright and Hayes (2003) for the list of nonce verbs exploited in the experiment

6 Conclusion

At the time of writing, we do not have the descriptions of the other systems that were submitted to the shared task. As a result, we are not able to identify the reasons for the good and not so good performance of the four systems we submitted. The objective of our participation was to test different hypotheses. The main one is the relatively low importance of stems when predicting the acceptability of wug forms, as evidenced by the good performance of the M2 model, which which only predicts the FAP of the inflected form. Therefore, M2 is output-oriented in the sense that the properties that characterize the input, i.e. the lemma, are not used during training.

M1 and M2 models are able to predict inflected forms and FAP_2 patterns for any UT in the training set while M3 models are specialized on a single UT. In future work, we plan to develop specialized versions of M1 and M2 in order to estimate the importance of the tested inflectional series (i.e. of the set of form pairs with the same UTs as the entries in test set) with respect to the entire training set. We further plan to test our models on more complete datasets in which the inflected forms could be predicted from other forms than the lemma, but also jointly from several forms of the lexeme (Bonami and Beniamine, 2016).

Acknowledgement

Experiments presented in this paper were carried out using the OSIRIM platform, that is administered by IRIT and supported by CNRS, the Region Occitanie, the French Government and ERDF.

References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar*, pages 54–82. Oxford University Press, Oxford.

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language*, 89:429–464.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1029, Denver, Colorado. Association for Computational Linguistics.

Adam Albright. 2002. Islands of reliability for regular morphology: Evidence from italian. *Language*, 78:684–709.

Adam Albright. 2007. Gradient phonological acceptability as a grammatical effect. https://www.mit.edu/~albright/papers/Albright-GrammaticalGradience.pdf.

Adam Albright. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 8:9–41.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Adam Albright and Bruce Hayes. 2006. Modeling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations. In Gisbert Fanselow, Féry Caroline, Vogel Ralf, and Schlesewsky Matthias, editors, *Gradience in Grammar: Generative Perspectives*, page 185—204. Oxford University Press, Oxford.

Sacha Beniamine, Olivier Bonami, and Benoît Sagot. 2017. Inferring inflection classes with description length. *Journal of Language Modelling*, 5(3):465–525.

Sacha Beniamine and Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. In *Proceedings of the Society for Computation in Linguistics*, volume 4.

Olivier Bonami and Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9:156–182.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ulrike Hahn and Todd M. Bailey. 2005. What makes words sound similar? *Cognition*, 97:227–267.

Nabil Hathout and Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a workable French machine-readable dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Nabil Hathout, Franck Sajous, Basilio Calderone, and Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3870–3878, Marseille.

Jennifer Hay, Janet Pierrehumbert, and Mary E. Beckman. 2004. Speech perception, well-formedness and the statistics of the lexicon. In John Local, Richard Ogden, and Rosalind Temple, editors, *Phonetic Interpretation: Papers in Laboratory Phonology VI*, Papers in Laboratory Phonology, pages 58–74. Cambridge University Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Charles F. Hockett. 1947. Problems of morphemic analysis. *Language*, 23:321–343.

Mans Hulden. 2014. Generalizing inflection tables into paradigms with finite state operations. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 29–36, Baltimore, Maryland.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

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).

Philippe Langlais and François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, page 51–54, Manchester.

Yves Lepage. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, volume 2, pages 728–735, Montréal.

Yves Lepage. 2004a. Analogy and formal languages. *Electronic notes in theoretical computer science*, 53:180–191.

Yves Lepage. 2004b. Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on Computational Linguistics (COLING-2004)*, pages 736–742, Genève.

Arni Magnusson, Hans J. Skaug, Anders Nielsen, Casper W. Berg, Kasper Kristensen, Martin Maechler, Koen J. van Bentham, Benjamin M. Bolker,

and Mollie E. Brooks. 2017. *glmmTMB: Generalized Linear Mixed Models using Template Model Builder*.

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

Robert M. Nosofsky. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34:393–418.

David. E. Rumelhart and James. L. McClelland. 1986. On learning the past tense of English verbs. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 2, pages 216–271. MIT Press.

Franck Sajous and Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, pages 405–426, Herstmonceux, England.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 120–127, Ann Arbor, MI. ACL.

Sami Petteri Virpioja, Minna Lehtonen, Annika Hultén, Henna Kivikari, Riitta Salmelin, and Krista Lagus. 2018. Using statistical models of morphology in the search for optimal units of representation in the human mental lexicon. *Cognitive Science*, 42(3):939–973.

Michael S. Vitevitch and Paul A. Luce. 2005. Increases in phonotactic probability facilitate spoken-nonword repetition. *Journal of Memory and Language*, 52:93–204.

Wolfgang Ulrich Wurzel. 1989. *Inflectional Morphology and Naturalness*. Kluwer, Dordrecht.