What transfers in morphological inflection? Experiments with analogical models

Micha Elsner
Department of Linguistics
The Ohio State University
melsner0@gmail.com

Abstract

This paper investigates how abstract processes like suffixation can be learned from morphological inflection task data using an analogical memory-based framework. In this framework, the inflection target form is specified by providing an example inflection of another word in the language. This model is capable of near-baseline performance on the SigMorphon 2020 inflection challenge. Such a model can make predictions for unseen languages, allowing one-shot inflection for natural languages and the investigation of morphological transfer with synthetic probes. Accuracy for one-shot transfer can be unexpectedly high for some target languages (88% in Shona) and language families (53% across Romance). Probe experiments show that the model learns partially generalizable representations of prefixation, suffixation and reduplication, aiding its ability to transfer. The paper argues that the degree of generality of these process representations also helps to explain transfer results from previous research.

1 Introduction

Morphological transfer learning has proven to be a powerful and effective technique for improving the performance of inflection models on under-resourced languages. The beneficial effects of transfer between source and target languages are known to be higher when the two are closely related (Anastasopoulos and Neubig, 2019) or typologically similar (Lin et al., 2019), mediated by the effect of script (Murikinati et al., 2020). But these effects are not always consistent; a variety of researchers report failure of transfer between closely related languages, or surprising successes with rather dissimilar ones (Sec 2). Pushing forward our understanding of these cases requires a more nuanced understanding of what is transferred by morphological transfer learning— that is, what abstract representational concepts do inflection networks acquire and how are these shared across languages?

This is a difficult question to address in the standard framework for inflection (Kann and Schütze, 2016), in which morphosyntactic properties are closely tied to their specific exponents in a particular language as well as to the more abstract processes by which these exponents are applied. In such a network, it is difficult to test whether a generic suffixation operation has been learned, without reference to a particular form/feature mapping, for instance between the Maori passive feature PASS and the spelling of a particular passive suffix *-tia*. Suffixing as a generic operation is much more likely to be useful in another language than the individual suffix. This work decouples these representational pieces by performing inflection in an analogical, memory-based framework.[1] In this framework, inflection instances do not have tags; rather, they include an instance of the desired mapping with respect to a different lemma (Figure 1). For example, to produce a passive Maori verb, the system takes an example verb with its passive and completes the four-part analogy: *lemma : target :: exemplar lemma : exemplar target*. The advantage of this redefinition of the task is that, in principle, the system does not need to learn anything about the individual affixes of a particular language, since these can be copied from the exemplar. Thus, it is possible to investigate how well such a system has learned a particular morphological process such as suffixation, which is expected to be present in a variety of languages.[2]

[1] "Memory-based" has been used in the literature to refer to models with dynamic read-write memory (Graves et al., 2016), as well as KNN-like exemplar models which store a large number of examples in a static memory (van den Bosch and Daelemans, 1999). The current work is of the latter type.

[2] Code available at: https://github.com/melsner/transformerbyexample.

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 154–166

Section 5 shows that this analogical framework for inflection can predict inflections across a variety of languages, demonstrating reasonable performance on the Sigmorphon 2020 multilingual benchmark (Vylomova et al., 2020). Section 6 describes one-shot learning experiments, performing language transfer without fine-tuning, and shows that for languages with concatenative affixes, one-shot transfer can be more effective than previously thought. Section 7 studies the system's ability to apply different types of morphological processes using constructed stimuli, showing that some configurations are capable of learning generic and transferable representations of processes including prefixing, suffixing and reduplication.

2 Related work

The overall positive effect of transfer learning is well established (McCarthy et al., 2019). Previous research has also evaluated how the choice of source language affects the performance in the target. While there is a robust trend for related languages to perform better, there are also many reports of exceptions. Kann (2020) finds that Hungarian is a better source for English than German and a better source for Spanish than Italian. She concludes that matching the target language's default affix placement (prefixing/suffixing) is important, and that agglutinative languages might be beneficial to transfer learning in general, but that genetic relatedness is not always a necessary or sufficient for effective transfer. Lin et al. (2019) also find that Hungarian and Turkish are good source languages for a surprising variety of unrelated targets. Rather than attribute this to agglutination, they propose that these languages lead to good transfer because of their large datasets and difficulty as tasks. Further puzzling results come from Anastasopoulos and Neubig (2019), who find that Italian data does not improve performance in closely related Ladin or Neapolitan[3] once monolingual hallucinated data is available, and that Latvian is as good a source for Scots Gaelic as its relative Irish.

Previous analyses of transfer learning have attempted to differentiate the contributions of various parts of the model through factored vocabularies or ciphering (Kann et al., 2017b; Jin and Kann, 2017). These methods give disjoint representations to characters and tags in the source and target languages, or disrupt the mapping between them. Low-level correspondence between character sets is the most important factor for successful transfer in very low-resource settings, but models with disjoint character representations still succeed at transfer once at least 200 target examples are available, indicating that higher-level information is also transferred and contributes to performance.

Kann et al. (2017b) also represents a prior one-shot morphological learning experiment. Their setting is not quite the same as the one here; they assume access to a single inflected form in half the paradigm cells in their target language (Spanish) which are used to fine-tune a pretrained system. Because their system uses the conventional tag-based framework, they are capable of filling cells for which no example is available (zero-shot learning), while the memory-based system presented here is not. On the other hand, the current work does not use fine-tuning or require target-language data at training time. They evaluate inflection on both seen and unseen cells as a function of five source languages, four of which are in the Romance family. The best one-shot transfer within Romance scores 44% exact match, the worst 13%. Transfer from unrelated Arabic scores 0%. One-shot learning experiments in this work use a much larger set of languages, and although performance in the typical case is similar, the best results are substantially better.

The memory-based design of the current work is rooted in cognitive theories of morphological processing. The widely accepted dual route model of morphological processing postulates that the mind retrieves familiar inflected forms from memory as well as synthesizing forms from scratch (Milin et al., 2017; Alegre and Gordon, 1998; Butterworth, 1983). It has often been claimed that memorized forms of specific words are central to the structure of inflection classes (Bybee and Moder, 1983; Bybee, 2006; Jackendoff and Audring, 2020). In such a theory, production of a form of a rare lemma is guided by the memory of the appropriate forms of common ones. Additional evidence for this view comes from historical changes in which one word's paradigm is analogically remodeled on another's (Krott et al., 2001; Hock and Joseph, 1996, ch.5). Liu and Hulden (2020) evaluate a model very similar to this one (a transformer in which target forms of other words, which they term "cross-table" examples, are provided as part of the input). They

[3] Regional Romance languages spoken in Northern and Southern Italy respectively.

	Lemma	Target specification	→	Target
Standard inflection generation	waiata	V;PASS		waiatatia
Memory-based	waiata	karanga : karangatia		waiatatia
	waiata	kaukau : kaukauria		waiatatia

Figure 1: Differing inputs for inflection models, eliciting the passive of the Maori verb *waiata* "sing". The memory-based system relies on an exemplar verb as the target specifier; shown here are *karanga* "call", which takes a matching suffix, and *kaukau* "swim", which mismatches.

find that such examples are complementary to data hallucination and yield improved results in data-sparse settings. Some earlier non-neural models also rely on stored word forms (Skousen, 1989; Albright and Hayes, 2002).

3 Exemplar selection

The system uses instances generated as described in Figure 1, separating the lemma, exemplar lemma and exemplar form with punctuation characters. Each instance also contains two features indicating the language and language family of the example (e.g. LANG_MAO, FAM_AUSTRONESIAN).

The selection of the exemplar is critical to the model's performance. Ideally, the lemma and the exemplar inflect in the same way, reducing the inflection task to copying. But this is not always the case. For example, Maori verbs fall into inflection classes, as shown in Figure 1; when the exemplar comes from a different class than the lemma, copying will yield an invalid output, so the model has to guess which class the input belongs to.[4]

This paper presents experiments using two settings: In **random selection**, the exemplar lemma is chosen arbitrarily from the set of training lemma/form pairs for the appropriate language and cell. This makes the task difficult, but allows the model to learn to cope with the distribution of inputs it will face at test time. In **similarity-based selection**, each source lemma is paired with an exemplar for which the transductions are highly similar. This makes the task easy, but since it relies on access to the true target form, it can be used only for model training, not for testing.[5] All models are evaluated using instances generated using random selection.

To perform similarity-based selection, each lemma is aligned with its target form in the training data in order to extract an edit rule (Durrett and DeNero, 2013; Nicolai et al., 2016). (For the first memory-based example in Figure 1, both words have the same edit rule -+*tia*.) The selected exemplar/form pair uses the same edit rule, if possible. During training, a lemma is allowed to act as its own exemplar, so that there is always at least one candidate. However, words in the test set must be given exemplars from the training set. If a cell in the test set does not appear in the training set, no prediction can be made; in this case, the system outputs the lemma. Extending the model to cover this case is discussed below as future work.[6]

4 Model design

The system uses the character-based transformer (Wu et al., 2020) as its learning model; this is a sequence-to-sequence transformer (Vaswani et al., 2017) tuned for morphological tasks, and serves as a strong official baseline for the Sigmorphon 2020 task. Moreover, transformers are known to perform well in the few-shot setting (Brown et al., 2020). All default hyperparameters[7] match those of Wu et al. (2020).

As discussed in prior work (Anastasopoulos and Neubig, 2019; Kann and Schütze, 2017), it is important to pretrain the model to predispose it to copy strings. To ensure this, the system is trained on a synthetic dataset. Each synthetic instance is generated within a random character set. The instance consists of a random pseudo-lemma and pseudo-exemplar created by sampling word

[4] In cases of class-dependent syncretism, the model must also guess which cell is being specified. For instance, German feminine nouns do not inflect for case, but some masculine nouns do, so the combination of a masculine lemma and a feminine exemplar can yield an unsolvable prediction problem.

[5] Within the training set, the same lemma/inflected form pair can appear as both an exemplar and a target instance; a reviewer speculates that this might allow the model to memorize lexically-specific outputs within the training set even when using random selection. To avoid this issue, no training scores are reported in this paper.

[6] In the SigMorphon 2020 datasets, this rarely occurs in practice. $\geq 99\%$ of target cells are covered in all languages except Ingrian (98%), Evenki (96%), and notably Ludic (61%).

[7] Including 4 layers, batches of 64, and the learning rate schedule.

lengths from the training word length distribution and then filling each one with random characters. With probability 50% the example is given a prefix; independently with probability 50% a suffix; independently with probability 10% an infix at a random character position. Prefixes and suffixes are random strings between 2-5 characters long and infixes are 1-2 characters long. (This means that, in some cases, no affix is added and the transformation is the identity, as occurs in cases of morphological syncretism.) An example such instance is *mpieňjmel:rbeaikkea::**zlü**rbeaikkea**üe*** with output ***zlü**mpieňjmel**üe***. The language tags for these examples indicate the kinds of affixation operations which were performed, for example LANG_PREFIX_SUFFIX; the family tag identifies them as SYNTHETIC. While this synthetic dataset is inspired by hallucination techniques (Anastasopoulos and Neubig, 2019; Silfverberg et al., 2017), note that these synthetic instances are not presented to the model as part of any natural language.

The Sigmorphon 2020 data is divided into "development languages" (45 languages in 5 families: Austronesian, Germanic, Niger-Congo, Oto-Manguean and Uralic) and "surprise languages" (45 more languages, including some members of development families as well as unseen families). Data from all the "development languages", plus the synthetic examples from the previous stage, is used to train a multilingual model, which is fine-tuned family. Finally the family models are fine-tuned by language. During multilingual training and per-family tuning, the dataset is balanced to contain 20,000 instances per language; languages with more training instances than this are subsampled, while languages with fewer are upsampled by sampling multiple exemplars (with replacement) for each lemma/target pair. For the final language-specific fine-tuning stage, all instances from the specific language are used.

5 Fine-tuned results

This section shows the test results for fully fine-tuned models on the development languages. Table 1 shows the average exact match and standard deviation by language family. Full results are given in Appendix A. Tables also show the results of the official competition baseline which is closest to the current work, the character transformer (Wu et al., 2020) fine-tuned by language, TRM-SINGLE.

Because the results of exemplar-based models can vary based on the choice of exemplar, the system applies a simple post-process to compensate for unlucky choices: it runs each lemma with five randomly-selected exemplars and chooses the majority output.

Family	Random	Similarity	Base
Austronesian (4)	83 (13)	67 (21)	81 (18)
Germanic (10)	87 (10)	51 (16)	90 (9)
Niger-Congo (9)	98 (4)	94 (9)	97 (3)
Oto-Manguean (10)	82 (16)	39 (23)	86 (12)3
Uralic (11)	92 (6)	46 (14)	93 (0.05)
Overall	89 (12)	57 (26)	90 (11)

Table 1: Fine-tuned accuracy scores for models trained with random and similarity-based selection, compared to the baseline. Num languages in family and score standard deviation across languages in parentheses.

Neither model achieves the same performance as the baseline (90%), although the random-exemplar model (89%) comes quite close. The similar-exemplar model (57%) is clearly inferior due to its severe mismatch between training and test settings. Performance varies across language families. All models perform well in Niger-Congo, although the conference organizers state that data from these languages may have been biased toward regular forms in an unrepresentative way.[8] The random-exemplar model is at or near baseline performance in Austronesian and Uralic, but falls further below baseline in Germanic and Oto-Manguean. Both of these families are characterized by complex inflection class structure in which randomly chosen exemplars are less likely to resemble the target for a given word.

The similar-exemplar model also performs poorly in Uralic. While some Uralic languages have inflection classes (Baerman, 2014), many (like Finnish) do not, but have complex systems of phonological alternations (Koskenniemi and Church, 1988). While the random-exemplar model can learn to compensate for these, the similar-exemplar model does not.

6 One-shot results

This section shows the results of one-shot learning. These experiments apply the multilingual and family models from the development languages to the surprise languages, without fine-tuning. For languages within development families, they use the appropriate family model; otherwise they use the

[8] A Swahili speaker confirms that some forms in the data appear artificially over-regularized (Martha Johnson p.c.).

multilingual model. Thus, the model's only access to information about the target language is via the provided exemplar.

Each experiment evaluates the results across five random exemplars per test instance (with replacement), but averages the results rather than applying majority selection. This computes the expected performance in the one-shot setting where only a single exemplar is available.

Results are shown in Table 2. One-shot learning is not competitive with the baseline fine-tuned system in any language family, but has some capacity to predict inflections in all families. Performance is generally better in families for which related languages were present in development.

The system trained with random exemplars achieves its best results on Tajik (Iranian: tgk, score 89%), Shona (Niger-Congo: sna, score 75%)[9], and Norwegian Nynorsk (Germanic: nno, score 42%). The system trained with similar exemplars achieves its best results on Shona (88%), Zarma (Songhay: dje, score 82%) and Tajik (79%). Notably, some of these high scores are achieved on languages that were difficult for the baseline systems; the score for Tajik beats the transformer baseline (56%), perhaps due to data sparsity, since baselines regularized using data hallucination perform better (93%).

Training with similar exemplars leads to clearly better results than random exemplars, a reversal of the trend observed with fine-tuning. This difference is particularly marked in Romance (53% average vs 5%). While the random-exemplar system is better at guessing what to do when the exemplar and target forms are divergent, this causes errors with unfamiliar languages. The system attempts to guess the correct inflection, rather than simply copying.

As an example, Table 3 shows an analysis of performance in Catalan (cat), selected because its results are fairly typical of the Romance family; the similar-exemplar system scores 53% while the random-exemplar system scores 12%. The table shows selected instances with different levels of exemplar match and mismatch. The first two, *arrissar* "curl" and *disputar* "discuss", match their exemplars well and are good cases for copying. The random-exemplar model gets these both wrong, segmenting incorrectly in the first and adding a spurious character in the second. The next two, *repetir* "repeat" and *engolir* "ingest", are mismatched with exemplars from a different inflection class; both systems make incorrect predictions, but the similar-exemplar system preserves the suffixes while the random-exemplar system does not. Finally, in the last example *llevar-se* "get up", the similar-exemplar model misinterprets the reflexive suffix *-se* as part of the verb stem, while the random-exemplar model fails to make any edit.

[9] As stated above, the Niger-Congo datasets are artificialized and probably does not represent the real difficulty of the inflection task.

Family	Random	Similarity	Base
Germanic (3)	29 (13)	38 (22)	80 (13)
Niger-Congo (1)	75 (0)	88 (0)	100 (0)
Uralic (5)	21 (9)	28 (12)	76 (26)
Afro-Asiatic (3)	7 (3)	26 (18)	96 (3)
Algic (1)	2 (0)	14 (0)	68 (0)
Dravidian (2)	7 (7)	13 (3)	85 (9)
Indic (4)	4 (5)	4 (2)	98 (3)
Iranian (3)	35 (39)	34 (32)	82 (19)
Romance (8)	6 (4)	53 (19)	99 (1)
Sino-Tibetan (1)	21 (0)	9 (0)	84 (0)
Siouan (1)	13 (0)	13 (0)	96 (0)
Songhay (1)	21 (0)	82 (0)	88 (0)
Southern Daly	4 (0)	6 (0)	90 (0)
Tungusic (1)	28 (0)	27 (0)	57 (0)
Turkic (9)	7 (8)	19 (11)	96 (7)
Uto-Aztecan (1)	33 (0)	30 (0)	81 (0)
Overall	14 (18)	30 (25)	90 (15)

Table 2: One-shot accuracy scores for models trained with random and similarity-based selection, compared to the baseline. Num. languages in family and score standard deviation across languages in parentheses. Families represented in development above the line, surprise families below.

A more systematic analysis computes an alignment-based edit rule for each system prediction (King et al., 2020) and counts the unique rules used to form one-shot predictions in the Catalan development set. Over 37105 instances, the random-exemplar model applies 626 unique edit rules, 20 of which appear in correct predictions. The similar-exemplar model applies 3137 unique rules, 154 of them correctly. The greater variety of both correct and incorrect outputs from the similar-exemplar model demonstrates its preference for faithfulness to the exemplar rather than remodeling the output to fit language-specific constraints.

7 Synthetic transfer experiments

When transfer learning fails, it can be difficult to tell whether the system has failed to represent a general morphological process, or whether it misapplies what it has learned due to mismatched lexical/phonological triggers. Experiments on artificial data can probe what abstract processes the model

Lemma	Exemplar	Rand. Sel.	Sim. Sel	Target
arrissar	posar : posarien	arrissaren	arrissarien	arrissarien
disputar	descriure : descriuria	disputarta	disputaria	disputaria
repetir	cremar : cremo	repetirer	repetio	repeteixo
engolir	forjar : forjava	engolire	engoliva	engolia
llevar-se	terminar : termino	llevar-se	llevor-se	llevo

Table 3: Development data from Catalan (Romance: cat) showing the outputs of two one-shot systems.

has learned to apply, the links between these processes and language families, and the environments in which they can operate.

A probing dataset is synthesized to model several morphological operations (Figure 2), including prefix/suffix affixation, reduplication and gemination. Affixation is typologically widespread (Bickel and Nichols, 2013) and appears in every development language on which the model was trained. Suffixation is more common in Germanic and Uralic; Oto-Manguean tonal morphology is also often represented via word-final diacritics.[10] Prefixing is more common in the Niger-Congo family.

Reduplication appears in three of the four Austronesian development languages, Tagalog, Hiligaynon and Cebuano (WAL, 2013), but not in the Maori dataset provided. The probe language has partial reduplication of the first syllable, as found in Tagalog and Hiligaynon. Previous work with artificial data demonstrates that sequence-to-sequence learners can learn fully abstract representations of reduplication (Prickett et al., 2018; Nelson et al., 2020; Haley and Wilson, 2021), but it has not been previously shown that networks trained on real data do this in a transferable way. In one-shot language transfer, reduplication instances are actually ambiguous. Given an instance *modi : _ :: gobu : gogobu*, there are two plausible interpretations, reduplicative *momodi* and affixal *gomodi*. Thus, analysis of reduplicative instances can be informative about the model's learned linkage between language family and typology.

Gemination is a inflectional process whereby a segment is lengthened to mark some morphological feature (Samek-Lodovici, 1992). The probe language geminates the last non-final consonant. None of the development languages have morphological gemination.

The probe languages use two alphabets: the first is a common subset of characters which appear in at least half the languages of every development family.[11] The second is a subset of Cyrillic characters intended to test transfer to a less-familiar orthography; a few Uralic development languages are written in Cyrillic. Each language has 90 random lemmas, sampled with the frames *CVCV*, *CVCVC*, *CVCVCVC*; affixal languages have 30 affixes of types *VCV*, *CV*, *CVCV*, plus 7 single-letter affixes. No probe lemma coincides with any real lemma, and no probe affix has frequency $> 5\%$ as a string prefix or suffix in any real language. Affixal languages contain an instance for every lemma/affix pair. Reduplication and gemination languages have one instance per lemma.

The model is prompted to inflect the probes as if they are members of each language family, and as members of a comparatively well-resourced language selected from those families, specifically Tagalog (tgl), German (deu), Mezquital Otomi (ote), Swahili (swa) and Finnish (fin), as well as the synthetic suffixing language used in pretraining (suff). In addition to checking whether the output matches, the table shows whether reduplicated instances have been correctly reduplicated (using a regular expression).

Table 4 shows the results. A comparison between the random-exemplar and similar-exemplar models confirms the hypothesis from above that random-exemplar models have less generalizable representations of morphological processes, especially prefixation and suffixation. While both models are capable of attaching affixes in the synthetic language, the random-exemplar model learns very language- and suffix-specific rules for applying these operations, leading to very low accuracy for copying generic affixes. Both models show less language-specific remodeling of affixes in the family-only setting than when the probes are labeled as part of a particular language; this effect is again more pronounced for the random-exemplar model.

Both models learn to reduplicate arbitrary CV syllables, but this process is mostly restricted to

[10] No Unicode normalization was performed; Oto-Manguean tone diacritics are treated as characters (as are parts of the complex characters of the Indic scripts). The placement of these diacritics within the word varies from language to language.

[11] Consonants *mpbntdrlskgh*, vowels *aeiou*.

Lemma *semet*		
Probe type	Exemplar	Target
Prefixing	kigu : igokigu	igosemet
Suffixing	kigu : kiguigo	semetigo
Reduplication	modi : momodi	sesemet
Gemination	bogu : boggu	semmet

Figure 2: Probe tasks illustrated for a single lemma.

Tagalog,[12], with some generalization to Austronesian. Most other languages interpret reduplication instances as affixes.

Only the similar-exemplar model gets any gemination instances correct, and these primarily in Uralic.[13] This is unsurprising, since the model was never trained with morphological gemination. It demonstrates that the model's representations of morphological processes represent the input typology and are not simply artifacts of the transformer architecture. While Uralic does not have gemination as an independent morphological process, alternations involving geminates do occur in some paradigms; the NOM.PL of *tikka* "dart" is *tikat*.[14] The model seems to have learned a little about gemination from this morphophonological process, but not a fully generalized representation.

Affixation remains relatively successful when using Cyrillic characters (suffixes more than prefixes), but for the most part, less so than with Latin characters, although in the random-exemplar model, Cyrillic suffixes are somewhat *more* accurate, probably due to less interference from language-specific knowledge. This substantiates the general finding (Murikinati et al., 2020) that transfer across scripts is more difficult than within-script. Cyrillic reduplication sees a much larger drop in accuracy. The difference is probably that simple affixation is phonologically uncomplicated, while reduplication requires phonological information about vowels and consonants.

8 Discussion

These experiments with real and synthetic transfer suggest some useful insights into the problematic findings of earlier transfer experiments. Why is Hungarian so successful as a source language for unrelated targets? Kann (2020) suggests that it is its agglutinative nature. The results shown here offer some speculative support for this view—perhaps the relative segmentability of prototypically agglutinative languages (Plank, 1999) acts like the similar-exemplar setting in the memory-based model, giving the source model a general bias for concatenative affixation, unpolluted by too many lexical and phonological alternations. As reported here, such a model is a promising starting point for inflection in many non-agglutinative systems, such as Romance verbs, which nevertheless are strongly concatenative.

Where transfer between related languages fails, it is conjecturally possible that the source model representations of edit operations are too closely linked to particular phonological and lexical properties of the source. This is clearly shown in the synthetic transfer experiments, where generic suffixation fails in Germanic and Uralic despite these families being strongly suffixing, because the system has learned to remodel its outputs to conform too closely to source-language templates.

More broadly, the synthetic experiments show a link between language typology and learning of morphological processes, suggesting that language structure, not only language relatedness, is key to successful transfer— transfer of structural principles can lead to improvements even without cognate words or affixes. For instance, successful reduplication appears only in Austronesian and successful gemination only in Uralic. A promising direction for future work would be to replace the language family feature with a set of typological feature indicators such as WALs properties (WAL, 2013), which might help the model to learn faster in low-resource target languages.

Two other extensions might bring the memory-based model closer to the state of the art in supervised inflection prediction. First, although the SigMorphon 2020 datasets are balanced by paradigm cell, real datasets are Zipfian, with sparse coverage of cells (Blevins et al., 2017; Lignos and Yang, 2018). For languages with large paradigms, the model thus requires the capacity to fill cells for which no exemplar can be retrieved, perhaps using a variant of adaptive source selection (Erdmann et al., 2020; Kann et al., 2017a). Second, the similar-exemplar model performs better in one-shot transfer experiments, but is hampered in the su-

[12]The random-exemplar model has low accuracy for reduplication in Tagalog because it appends spurious Tagalog prefixes to the output, another example of a language-specific rule. However, the regular expression check confirms that reduplication is performed correctly.

[13]Because of this poor performance, Cyrillic gemination was not tested.

[14]See Silfverberg et al. (2021) for a fuller investigation of generalizable representations of gradation processes in Finnish noun paradigms.

Model	Fam/Lg.	Pref	Pref (Cyrl)	Suff	Suff (Cyrl)	Redup.	Redup. (Cyrl)	Gem.
Rand.	austro	62	36	26	38	0 (10)	0	0
	austro/tgl	0	1	0	0	28 (90)	3 (7)	0
	ger	1	0	25	36	0 (3)	0	0
	ger/deu	0	0	8	10	0 (3)	0	0
	n-congo	92	55	40	41	0 (3)	0	0
	n-congo/swa	100	76	36	25	0 (3)	0	0
	oto	20	15	21	33	0 (3)	0	0
	oto/ote	35	30	1	9	0 (3)	0	0
	uralic	3	0	23	34	0 (3)	0	0
	uralic/fin	0	0	7	22	0 (3)	0	0
	synth	84	62	97	91	0 (3)	0	0
	synth/suff	28	1	100	97	0 (3)	0	0
Sim.	austro	86	75	94	85	30 (30)	0	0
	austro/tgl	30	35	75	63	88 (88)	8 (8)	0
	ger	85	55	99	96	3 (3)	0	8
	ger/deu	86	55	99	98	0	0	5
	n-congo	99	96	98	93	0 (3)	0	3
	n-congo/swa	99	98	88	57	0	0	0
	oto	88	76	95	87	18 (18)	0	0
	oto/ote	96	84	59	17	5 (5)	0	0
	uralic	59	10	97	95	0	0	17
	uralic/fin	52	4	98	98	0	0	12
	synth	94	84	99	95	8 (10)	0	2
	synth/suff	86	42	100	99	0	0	2

Table 4: Accuracy of synthetic probe tasks presented as different language and language family. (Cyrl) indicates Cyrillic alphabet. Parentheses in reduplication columns show frequency of correct CV reduplication.

pervised setting by train-test mismatch. Selecting training exemplars using a classifier which could also be used at inference time would reduce this mismatch. These experiments are left for future work.

Finally, since the memory-based architecture is cognitively inspired, it might be adapted as a cognitive model of language learning in contact situations. Work on this learning process suggests that speakers find it much easier to learn new exponents than to learn new morphological processes (Dorian, 1978; Mithun, 2020). Thus, the impact of source-language transfer may indeed be most significant in cases where the L1 and L2 (source and target) languages differ in the abstract mechanisms of inflection rather than the specifics. Historical contact-induced change provides evidence for this viewpoint in the form of systems which have changed to employ the same processes as a contact language. For example, Cappadocian Greek has become agglutinative through its extensive contact with Turkish (Janse, 2004). For other examples, see Green (1995); Thomason (2001).

9 Conclusion

The results of this paper demonstrate that the proposed cognitive mechanism of memory-based analogy provides a relatively strong basis for inflection prediction. Performance in a supervised setting is strongest in languages without large numbers of inflection classes, and requires training exemplars to be selected in the same way as test exemplars. Memory-based analogy also provides a foundation for one-shot transfer; in this case, training exemplars should closely match the elicited inflections, so that the model learns to copy rather than reconstruct the output form. One-shot transfer using this mechanism can achieve higher accuracy than previously thought, even when no genetically related languages are available in training. Scores vary widely, but can be over 80% for some languages.

Finally, this paper provides new evidence about what kinds of abstract information (beyond character correspondences) is transferred between languages when learning to inflect. The model learns general processes for prefixation and suffixation which apply (to some extent) across character sets, but its application of these can be disrupted by language-specific morpho-phonological rules. It also learns to reduplicate arbitrary CV sequences, but applies this process only when targeting a language with reduplication. Learning of morphological processes in general appears to be driven by the input typology. The discussion argues that the usefulness of general representations for prefixation and suffixation accounts for the puzzling effectiveness of agglutinative languages as transfer sources reported in previous research.

Acknowledgments

This research is deeply indebted to ideas contributed by Andrea Sims. I am also grateful to members of LING 5802 in autumn 2020 at Ohio State, and to the three anonymous reviewers for their comments and suggestions. Parts of this work were run on the Ohio Supercomputer (OSC, 1987).

References

2013. World atlas of language structures online. Available online at https://wals.info/, accessed 3 June 2020.

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the Sixth Meeting of the Association for Computational Linguistics Special Interest Group in Computational Phonology in Philadelphia, July 2002*, pages 58–69.

Maria Alegre and Peter Gordon. 1998. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40:41–61.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Matthew Baerman. 2014. Covert systematicity in a distributionally complex system. *Journal of Linguistics*, pages 1–47.

Balthasar Bickel and Johanna Nichols. 2013. Fusion of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

James P. Blevins, Petar Milin, and Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins, and Huba Bartos, editors, *Perspectives on morphological organization: Data and analyses*, pages 141–158. Brill.

Antal van den Bosch and Walter Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292, College Park, Maryland, USA. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Brian Butterworth. 1983. Lexical representation. In Brian Butterworth, editor, *Language production, vol. 2: Development, writing and other language processes*, pages 257–294. Academic Press.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Joan Bybee and Carol Moder. 1983. Morphological classes as natural categories. *Language*, 59(2):251–270.

Nancy C. Dorian. 1978. The fate of morphological complexity in language death: Evidence from East Sutherland Gaelic. *Language*, 54(3):590–609.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Alexander Erdmann, Tom Kenter, Markus Becker, and Christian Schallhart. 2020. Frugal paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8248–8273, Online. Association for Computational Linguistics.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Ian Green. 1995. The death of 'prefixing': contact induced typological change in northern australia. In *Annual Meeting of the Berkeley Linguistics Society*, volume 21, pages 414–425.

Coleman Haley and Colin Wilson. 2021. Deep neural networks easily learn unnatural infixation and reduplication patterns. *Proceedings of the Society for Computation in Linguistics*, 4(1):427–433.

Hans Henrich Hock and Brian D. Joseph. 1996. *Language history, language change and language relationship: An introduction to historical and comparative linguistics*. Mouton de Gruyter.

Ray Jackendoff and Jenny Audring. 2020. *The texture of the lexicon: Relational Morphology and the Parallel Architecture*. Oxford University Press.

Mark Janse. 2004. Animacy, definiteness, and case in Cappadocian and other Asia Minor Greek dialects. *Journal of Greek linguistics*, 5(1):3–26.

Huiming Jin and Katharina Kann. 2017. Exploring cross-lingual transfer of morphological knowledge in sequence-to-sequence models. In *Proceedings of*

the First Workshop on Subword and Character Level Models in NLP, pages 70–75, Copenhagen, Denmark. Association for Computational Linguistics.

Katharina Kann. 2020. Acquisition of inflectional morphology in artificial neural networks with prior knowledge. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 144–154, New York, New York. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017a. Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 514–524, Valencia, Spain. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017b. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2017. Unlabeled data for morphological generation with character-based sequence-to-sequence models. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 76–81, Copenhagen, Denmark. Association for Computational Linguistics.

David King, Andrea Sims, and Micha Elsner. 2020. Interpreting sequence-to-sequence models for Russian inflectional morphology. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 481–490, New York, New York. Association for Computational Linguistics.

Kimmo Koskenniemi and Kenneth Ward Church. 1988. Complexity, two-level morphology and Finnish. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Andrea Krott, R Harald Baayen, and Robert Schreuder. 2001. Analogy in morphology: modeling the choice of linking morphemes in dutch.

Constantine Lignos and Charles Yang. 2018. Morphology and language acquisition. In Andrew Hippisley and Gregory T. Stump, editors, *Cambridge handbook of morphology*, pages 765–791. Cambridge University Press.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2020. Analogy models for neural word inflection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2861–2878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Petar Milin, Laurie Beth Feldman, Michael Ramscar, Roberta A. Hendrick, and R. Harald Baayen. 2017. Discrimination in lexical decision. *PLoS ONE*, 12(2):e0171935.

Marianne Mithun. 2020. Where is morphological complexity? In Peter Arkadiev and Francesco Gardani, editors, *The complexities of morphology*, pages 306–327. Oxford University Press.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

Max Nelson, Hossep Dolatian, Jonathan Rawski, and Brandon Prickett. 2020. Probing RNN encoder-decoder generalization of subregular functions using reduplication. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 167–178, New York, New York. Association for Computational Linguistics.

Garrett Nicolai, Bradley Hauer, Adam St Arnaud, and Grzegorz Kondrak. 2016. Morphological reinflection via discriminative string transduction. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 31–35, Berlin, Germany. Association for Computational Linguistics.

OSC. 1987. Ohio supercomputer center.

Frans Plank. 1999. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340.

Brandon Prickett, Aaron Traylor, and Joe Pater. 2018. Seq2Seq models with dropout can learn generalizable reduplication. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

Vieri Samek-Lodovici. 1992. A unified analysis of crosslinguistic morphological gemination. In *Proceedings of CONSOLE*, volume 1, pages 265–283. Citeseer.

Miikka Silfverberg, Francis Tyers, Garrett Nicolai, and Mans Hulden. 2021. Do RNN states encode abstract phonological alternations? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5501–5513. Association for Computational Linguistics.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.

Royal Skousen. 1989. *Analogical modeling of language*. Springer Science & Business Media.

Sarah Grey Thomason. 2001. Contact-induced typological change. In *Language typology and language universals: An international handbook*, volume 2, pages 1640–1648.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint arXiv:2005.10213*.

A Full results

For replicability, this appendix provides full results for all languages, as 0-1 accuracy on the official test datasets. The reported baseline is TRM-SINGLE, copied from https://docs.google.com/spreadsheets/d/1ODFRnHuwN-mvGtzXA1sNdCi-jNqZjiE-i9jRxZCK0kg. Scores for supervised systems on the development languages are shown in Table 5 and scores for one-shot systems on surprise languages are shown in Table 6. See Vylomova et al. (2020) for language abbreviation definitions.

Lang	Fam	Rand	Sim	Base
ang	Indo-Eur: Germanic	72	19	78
azg	Oto-Manguean	94	22	95
ceb	Austronesian	79	69	84
cly	Oto-Manguean	82	19	91
cpa	Oto-Manguean	74	33	91
ctp	Oto-Manguean	43	15	60
czn	Oto-Manguean	83	32	80
dan	Indo-Eur: Germanic	75	42	75
deu	Indo-Eur: Germanic	93	62	98
eng	Indo-Eur: Germanic	97	67	97
est	Uralic	94	47	95
fin	Uralic	100	39	100
frr	Indo-Eur: Germanic	81	39	87
gaa	Niger-Congo	100	100	98
gmh	Indo-Eur: Germanic	94	75	91
hil	Austronesian	97	74	98
isl	Indo-Eur: Germanic	88	37	97
izh	Uralic	85	33	87
kon	Niger-Congo	99	99	98
krl	Uralic	99	36	99
lin	Niger-Congo	100	100	100
liv	Uralic	93	54	96
lug	Niger-Congo	90	74	91
mao	Austronesian	71	57	52
mdf	Uralic	92	67	94
mhr	Uralic	91	67	93
mlg	Austronesian	100	100	100
myv	Uralic	93	61	94
nld	Indo-Eur: Germanic	99	61	99
nob	Indo-Eur: Germanic	75	47	76
nya	Niger-Congo	100	100	100
ote	Oto-Manguean	99	80	99
otm	Oto-Manguean	98	46	98
pei	Oto-Manguean	65	17	72
sme	Uralic	99	31	100
sot	Niger-Congo	100	100	98
swa	Niger-Congo	100	100	100
swe	Indo-Eur: Germanic	97	59	99
tgl	Austronesian	69	35	72
vep	Uralic	83	28	84
vot	Uralic	81	41	86
xty	Oto-Manguean	90	79	91
zpv	Oto-Manguean	87	46	85
zul	Niger-Congo	92	83	92
Overall		89	57	90
Stdev		12	26	11

Table 5: Zero-one test-set accuracy scores by language for SigMorphon 2020 development languages (supervised).

Lang	Fam	Rand	Sim	Base
ast	Indo-Eur: Romance	2	64	100
aze	Turkic	9	17	81
bak	Turkic	15	14	100
ben	Indo-Aryan	1	4	99
bod	Sino-Tibetan	21	9	84
cat	Indo-Eur: Romance	12	53	100
cre	Algic	2	14	68
crh	Turkic	24	45	99
dak	Siouan	13	13	96
dje	Nilo-Saharan	21	82	88
evn	Tungusic	28	27	57
fas	Indo-Eur: Iranian	2	13	100
frm	Indo-Eur: Romance	7	73	100
fur	Indo-Eur: Romance	11	19	100
glg	Indo-Eur: Romance	9	59	100
gml	Indo-Eur: Germanic	11	11	62
gsw	Indo-Eur: Germanic	33	64	93
hin	Indo-Aryan	0	1	100
kan	Dravidian	13	16	76
kaz	Turkic	0	7	98
kir	Turkic	2	6	98
kjh	Turkic	11	11	100
kpv	Uralic	17	47	97
lld	Indo-Eur: Romance	3	68	99
lud	Uralic	22	14	32
mlt	Afro-Asiatic	10	13	97
mwf	Australian	4	6	90
nno	Indo-Eur: Germanic	42	40	86
olo	Uralic	37	33	94
ood	Uto-Aztecan	33	30	81
orm	Afro-Asiatic	2	52	99
pus	Indo-Eur: Iranian	13	9	90
san	Indo-Aryan	13	5	93
sna	Niger-Congo	75	88	100
syc	Afro-Asiatic	8	13	91
tel	Dravidian	0	10	95
tgk	Indo-Eur: Iranian	89	79	56
tuk	Turkic	0	21	86
udm	Uralic	11	30	98
uig	Turkic	0	26	99
urd	Indo-Aryan	2	7	99
uzb	Turkic	0	21	100
vec	Indo-Eur: Romance	2	62	100
vro	Uralic	17	17	61
xno	Indo-Eur: Romance	2	22	96
Overall		14	30	90
Stdev		18	25	15

Table 6: Zero-one test-set accuracy scores by language for SigMorphon 2020 surprise languages (one-shot).