Interpreting Sequence-to-Sequence Models for Russian Inflectional Morphology

David L. King **Andrea D. Sims** **Micha Elsner**
The Ohio State University
{king.2138, sims.120, elsner.14}@osu.edu

Abstract

Morphological inflection, as an engineering task in NLP, has seen a rise in the use of neural sequence-to-sequence models (Kann and Schütze, 2016; Cotterell et al., 2018; Aharoni and Goldberg, 2017). While these outperform traditional systems based on edit rule induction, it is hard to interpret what they are learning in linguistic terms. We propose a new method of analyzing morphological sequence-to-sequence models which groups errors into linguistically meaningful classes, making what the model learns more transparent. As a case study, we analyze a seq2seq model on Russian, finding that semantic and lexically conditioned allomorphy (e.g. inanimate nouns like ZAVOD 'factory' and animates like OTEC 'father' have different, animacy-conditioned accusative forms) are responsible for its relatively low accuracy. Augmenting the model with word embeddings as a proxy for lexical semantics leads to significant improvements in predicted wordform accuracy.

1 Introduction

Neural sequence-to-sequence models excel at learning inflectional paradigms from incomplete input (Table 1 shows an example inflection problem.) These models, originally borrowed from neural machine translation (Bahdanau et al., 2014), read in a series of input tokens (e.g. characters, words) and output, or translate, them as another series. Although these models have become adept at mapping input to output sequences, like all neural models, they are relatively uninterpretable. We present a novel error analysis technique, based on previous systems for learning to inflect which relied on edit rule induction (Durrett and DeNero, 2013). By using this to interpret the output of a neural model, we can group errors into linguistically salient classes such as producing the wrong case form or incorrect inflection class.

Our broader linguistic contribution is to reconnect the inflection task to the descriptive literature on morphological systems. Neural models for inflection are now being applied as cognitive models of human learning in a variety of settings (Malouf, 2017; Silfverberg and Hulden, 2018; Kirov and Cotterell, 2018, and others). They are appealing cognitive models partly because of their high performance on benchmark tasks (Cotterell et al., 2016, and subsq.), and also because they make few assumptions about the morphological system they are trying to model, dispensing with overly restrictive notions of segmentable morphemes and discrete inflection classes. But while these constructs are theoretically troublesome, they are still important for *describing* many commonly-studied languages; without them, it is relatively difficult to discover what a particular model has and has not learned about a morphological system. This is often the key question which prevents us from using a general-purpose neural network system as a cognitive model (Gulordava et al., 2018). Our error analysis allows us to understand more clearly how the sequence-to-sequence model diverges from human behavior, giving us new information about its suitability as a cognitive model of the language learner.

As a case study, we apply our error analysis technique to Russian, one of the lowest-performing languages in SIGMORPHON 2016. We find a large class of errors in which the model incorrectly selects among lexically- or semantically-conditioned allomorphs. Russian has semantically-conditioned allomorphy in nouns and adjectives, and lexically-conditioned allomorphy (inflection classes) in nouns and verbs (Timberlake, 2004); Section 3 gives a brief introduction to the relevant phenomena. While these facts are commonly known to linguists, their importance to modeling the inflection task has not previously

Source	Features	Target

ABAŠ	pos=N,
	case=NOM,
	num=SG
		ABAŠ

JATAGAN	pos=N,
	case=INS,
	num=PL
		JATAGANAMI

Table 1: An example inflection problem: the task is to map the Source and Features to the correct, fully inflected Target.

been pointed out. Section 4 shows that these phenomena account for most of Russian's increased difficulty relative to the other languages. In Section 6, we provide lexical-semantic information to the model, decreasing errors due to semantic conditioning of nouns by 64% and of verbs by 88%.

2 Background

The inflection task described above is an instance of the *paradigm cell filling problem* (Ackerman et al., 2009), and models a situation which both computational and human learners face. For humans, the PCFP is closely related to the "wug test" (Berko, 1958): given some previously unseen word, how does a speaker produce a different inflected form? As Lignos and Yang (2016) and Blevins et al. (2017) point out, the same Zipfian distribution that makes other NLP tasks (e.g. MT) difficult is also at play in morphology, namely that no corpus will ever exist that has every wordform from every lexeme. For theoretical morphologists, the difficulty of the PCFP on average is a measure of the learnability of a morphological system, with implications for language typology (Ackerman et al., 2009; Ackerman and Malouf, 2013; Albright, 2002; Bonami and Beniamine, 2016; Sims and Parker, 2016).

Ackerman et al.'s (2009) formulation of the PCFP relies on a simple concatenative model in which words are divided into stems and affixes, and in which each affix is treated as a discrete value. Cotterell et al. (2018) points out that this model is ill-suited to dealing with phenomena like phonological alterations or stem suppletion. Newer models (Silfverberg and Hulden, 2018; Malouf, 2017; Cotterell et al., 2018) use sequence-to-sequence inflection models to avoid these shortcomings.

Faruqui et al. (2016) introduced the use of attention-based neural sequence-to-sequence learning for the inflection task, building on models from machine translation (Bahdanau et al., 2014). Their model treats input as a linear series where grammatical features and characters are encoded as one-hot embeddings and passed to a bidirectional encoder LSTM; output for each paradigm cell is produced by a separate decoder. Kann and Schütze (2016) extended Faruqui et al.'s architecture by using ensembling and by using a single decoder, shared across all output paradigm cells, to account for data sparsity. Later systems (Aharoni and Goldberg, 2017; Kann and Schütze, 2017) have made changes to the input representation and the architecture, for instance incorporating variants of hard attention and autoencoding. From a theoretical standpoint, all these models are "a-morphous" (Anderson, 1992) or "inferential-realizational" (Stump, 2001)— rather than assume a concatenative process which stitches discrete morphemes together into surface word forms, they learn a flexible, generalizable transduction, either between a stem and surface form (Anderson, 1992; Stump, 2001), or between pairs of surface forms (Albright, 2002; Blevins, 2006).

Some older learning-based inflection systems, such as Durrett and DeNero (2013), exploit sequence alignment across strings. Alignment-based systems essentially treat morphology as concatenation. While they do not perform full-scale morphological analysis (since they do not account for phonological alternations), in languages which are mostly concatenative, they do tend to isolate affix-like units as sequences of adjacent insertions or deletions. This property has been criticized in the neural literature (Faruqui et al., 2016) since it represents processes like vowel harmony by enumerating large sets of surface allomorphs, making the learning problem harder. We agree with these criticisms from the modeling standpoint, but we exploit the interpretability of the technique in our analysis of model results.

Our study of Russian concludes that semantically- and lexically-conditioned allomorphy constitutes a problem for current neural reinflection models. This is because such models are trained to map input to output character sequences; they do not typically have access to information about what the words they are inflecting *mean*. We show that, by providing word embeddings as meaning representations, we can reduce this source of error and bring Russian closer to the other languages studied in SIGMORPHON 2016.

Recently the NLP community has also pushed for greater transparency with neural models (xci, 2017; ana, 2019). Wilcox et al. (2018) showed that RNNs learn hierarchical structure in sentences like island constraints. Faruqui et al. demonstrated that RNNs can automatically learn which vowel pairs participate in vowel harmony alternation. Our error analysis allows us to interpret what neural models are learning, reconnecting inflection tasks to linguistic intuitions by generalizing over error classes.

3 Russian Inflectional Morphology

We select Russian as our language of analysis because it was among the three worst-performing languages in the SIGMORPHON 2016 shared task, falling 4+ percentage points behind the other languages. Problems with the design of the Navajo and Maltese datasets may have been the source of the problems with those languages,[1] but this cannot explain the Russian results. The discrepancy hints at some linguistic property which distinguishes Russian from the other languages. Below, we give an overview of the Russian morphological system, concentrating on nouns, verbs, and adjectives, the parts of speech targeted by the SIGMORPHON 2016 shared task.

Russian is an East Slavic language which, in line with other Slavic languages, makes heavy use of inflectional morphology. Russian nouns and verbs belong to inflectional *classes*: groups of words which share a common set of inflectional affixes.

Russian nouns and adjectives have six primary cases—nominative, accusative, genitive, dative, locative, and instrumental—and two numbers, singular and plural. We follow the classification system of Timberlake (2004), which groups nouns into three primary inflection classes (I, II, and III) with subclasses (IA, IB, IIIA, IIIB, and IIIC).

Within these classes, however, the formation of the accusative is further subdivided based on semantics. Specifically, in class IA accusative singular and plural and in classes IB, II, and III accusative plurals, the accusative exhibits syncretism with either the genitive (for animates) or the nominative (for inanimates). In the case of the animate noun STUDENT (‘student’), for example, the nominative singular form is *student* and the accusative singular and genitive singular forms are both *studenta*. Conversely, for MESTO (‘place’), the accusative singular and nominative singular both have the form *mesto*, but the genitive singular is *mesta*. An example of how this phenomenon looks at the paradigm level for class IA can be seen in Figure 2.

Case	Singular	Plural
Nominative	∅, -’, -J, -IJ	-”, -I, -II
Accusative	N or G	
Genitive	-A, -JA, -IJA	-OV, -EJ, -EV, -IEV
Dative	-U, -JU, -IJU	-AM, -JAM, -IJAM
Instrumental	-OM, -EM, -IEM	-AMI, -JAMI, -IJAMI
Locative	-E, -II	-AX, -JAX, -IJAX

Table 2: An example of class IA, showing the effect of animacy in the orthography[2]across the singular and plural accusative forms, where *N* or *G* indicate where syncretism occurs in the accusative form based on animacy.

Adjectives in Russian must agree with case, gender, and number of the nouns they modify. They also exhibit the same syncretism in the plural and masculine singular forms, based on the animacy of the noun that the adjective modifies.[3]

Russian also has two verb classes based on what Timberlake calls a verb’s *thematic ligature* (i.e. a thematic vowel). A verb is either an *i-conjugation* verb or an *e-conjugation* verb, depending on the vowel used to create the present tense stem. For example, MOLČAT’ (‘to be silent’) forms the present tense stem with *-i* (namely *molč-i-*), making its second person singular form *molčiš’*. Likewise, for a verb like BROSAT’ (‘to toss’), its present tense stem is *brosae-*, formed with the theme vowel *-e*, making its second person singular form *brosaeš’*. For verbs with monomor-

[1] As announced by the SIGMORPHON shared task organizers.

[2] Examples in this paper are presented in scientific transliteration instead of Cyrillic for accessibility; our system processes Cyrillic characters.

[3] Predicative adjectives have an additional short form which only agrees with gender and number since they only use nominative suffixes. Active participles are inflected as adjectives.

t → č	d → ž	s → š	st → šč
k → č	z → ž	x → š	sk → šč
	g → ž		
p → pl	f → fl	m → ml	
b → bl	v → vl		

Table 3: Russian makes use of phonological alternation, which it encodes orthographically for some characters.

BUMAŽKA → BUMAŽEK (‘paper.DIM’)
NOM.SG → GEN.PL
Gold:

	b	u	m	a	ž		k	a
✓	b	u	m	a	ž	e	k	
						+e		-a

Predicted:

	b	u	m	a	ž		k	a
✗	b	u	m	a	ž	o	k	
						+o		-a

Table 4: Sample induced edit rules can be used to compare gold vs. predicted differences in the MED’s output for error mining. These automatic annotations we subsequently analyzed as missing insertions/deletion and erroneous insertions/deletions.

phemic bases, the class to which the verb belongs (and thus what theme vowel it combines with to form the present tense stem) is not normally thought to be predictable from its syntactic frame or its semantics. It is an idiosyncratic (i.e. lexically-conditioned) property which learners have to memorize for each verb they learn. For verbs with derived bases the situation is more complicated, since derivational suffixes systematically determine the inflection class of a verb. For example, verbs formed with the highly productive *-ova* suffix (*beseda*, ‘conversation’; *besed-ova-t’*, ‘converse’) always belong to the e-conjugation. Transitivity and inflection class are also sometimes related in derived verbs, although not perfectly predictably so. For instance, derived verbs formed with *-i* (e.g. *čist-yj*, ‘clean (adj)’; *čist-i-t’*, ‘clean (verb)’) tend to be transitive (Townsend, 1975).

Verb stems can also undergo phonological alternation, in which the final consonant of a stem changes to another when being inflected for certain parts of the paradigm (e.g. EZDIT’ (‘to ride’) becomes *ezžu* in the first person present singular cell). Further common alternations can be seen in Table 3.

Finally, both nouns and verbs sometimes have morphological stress alternations within the paradigm. These tend to affect high token frequency lexemes, and are thus salient to speakers and learners, but do not affect the majority of words. Counted by type frequency, more than 97% of nouns have fixed stress throughout the paradigm (Brown et al., 2007). Stress alternations are not encoded orthographically.

4 Error Analysis

As mentioned in Section 2, some pre-neural systems for predicting a novel inflected wordform from a source wordform focused on inducing edit operations from one string to another using sequence alignment (Durrett and DeNero, 2013). These approaches model the differences between two strings as a series of *insert* and *delete* operations. While the alignment approach has been superseded by neural models with better performance, we re-apply it here in order to automatically compare and group predicted edit operations vs. gold edit operations. Rather than aligning source to target forms, we align the *gold* target form to the *proposed* target form from the system. For example, if a model learning English plurals incorrectly learned that the ending *-en* was productive, we would see a surplus of *-s* → *-en* errors.

Errors viewed in this way often have natural linguistic interpretations, especially when correlated with the paradigm cells in which they occur. As seen in Table 4, the model correctly predicted the zero genitive plural ending for the noun BUMAŽKA (‘paper.DIM’), but erroneously inserted an *o* (*bumažok*) instead of an *e* (*bumažek*). This is an example of stem alternation in nouns that occurs when there is a zero ending (i.e. nominative singular or genitive plural, depending on the class). The vowel inserted is always an *e* or an *o*, but in this case the wrong vowel was selected.

We used the 2016 SIGMORPHON dataset. Although ideally we would like to have had access to a dataset which more accurately encoded Russian phonology and stress, to our knowledge no such corpus exists. Using the SIGMORPHON dataset, we trained the original MED setup Kann and Schütze made publicly available[4] using the hyperparameters they specified. Other input forms, such

[4] http://cistern.cis.lmu.de/med/

as thos used by Cotterell et al. (2018), are possibly more realistic, but we wished to see why in a controlled setting (i.e. using citation forms) Russian underperformed as compared to languages like Spanish and German. We then extracted errors from the MED system's performance on the validation set, which had 1,591 wordform predictions in total. In using Durrett and DeNero's sequence alignment approach to isolate the differences in edit operations, we were able to annotate each error as a missing deletion (`-d`), an erroneous deletion (`+d`), a missing insertion (`-i`), or an erroneous insertion (`+i`). From here we were able to group erroneous outputs which contained the same edit operations. An example of how we compared and annotated each gold/prediction pair can be seen in Table 5. We can compare these to cases where the same edit operations occur in *correct* answers. This indicates whether an erroneous edit is entirely unattested (i.e. noise), or whether it represents a mis-application of a transform which would have been legitimate for a different source word or target paradigm cell.

We find that the system often produced nouns with the wrong case suffix. In 14% of the total errors, accounting for 29% of all errors affecting nouns, the MED system produced a form of the noun that exists, but corresponds to a different case than the target one. MED also produced verbs with inflections corresponding to the wrong inflection class. These cases account for 10% of the overall errors and 23% of the verb-specific errors. Other errors involved incorrect edits to the stem (in all parts of speech). These accounted for 72% of the overall error rate. These cases were often only a single edit away from the gold wordform, but were more drastic in other cases. We investigated how many of these edits represented mis-applied rules which had been observed elsewhere in training. Surprisingly, *every* erroneous edit rule discovered in the system output had been seen in the training data. We include examples of these error types in Table 7 and summarize the error rates in Table 6.

Many of the noun case errors involve the accusative case, and in particular, an incorrect choice between semantically-conditioned alternatives. As discussed in Section 3, the accusative is syncretic with the genitive or the nominative, conditioned on animacy. In these errors, the system proposes an accusative which matches a correctly inflected form of the word, but not the right one. For instance, the first row of Table 7 shows the proposed accusative of OZNOB 'the chills, shaking'. This matches the genitive form rather than the nominative, which we can easily diagnose by looking for cells in the gold paradigm where the +A edit rule appears.

Verb errors tend to involve alternations characteristic of confusion between i- and e-conjugation verbs. Stem edits often introduce or delete sounds which participate in phonologically motivated alternations, but are not restricted to the contexts in which those alternations legitimately appear.

Error type		Form
Case	✗	OZNOB-A
	✓	OZNOB-∅
	✗	MEXANIZM-OV
	✓	MEXANIZM-Y
Verb class	✗	DOŽD-I-Š'SJA
	✓	DOŽD-E-Š'SJA
Stem edits	✗	REZG-G-OVORČIVY
	✓	RAZ-∅-GOVORČIVY
	✗	ZA-P-O-ŠČ-ENNYJ
	✓	ZA-K-A-Č-ENNYJ
	✗	SANKTPETE-TE-R-B-BUR-B-ŽCAM
	✓	SANKTPETE-∅-R-∅-BUR-∅-ŽCAM

Table 7: Examples of the three main error groups we found produced by the MED system on the 2016 SIGMORPHON dataset. An ✗ is an incorrect prediction and the ✓ below is the gold wordform. Empty set symbols (∅) indicate an erroneous insertion.

5 Model Improvements

In this section, we incorporate a proxy for lexical semantics into the model input representations, leading to improved results. This is useful from a practical standpoint, but also as a clear demonstration that semantic conditioning was responsible for many of the errors which we discussed in the previous section.

As our source for semantic information, we use word embeddings (Mikolov et al., 2013; Socher et al., 2013; Xu et al., 2015). We concatenate the output from the bidirectional encoder with the citation form's embedding. Equipped with this information, the model should be able to learn phenomena like the animacy-dependent syncretism

Gold	Predicted	Rule	Annotation	Category
ABSOLJUTISTA	ABSOLJUŠČISTA	-T+Š+Č	+d+i+i	phonological alternation
DERŽIŠ'SJA	DERŽAEŠ'SJA	-I+A+E	+d+i+i	verb class
ABDOMEN	ABDOMENA	+A	+i	animacy

Table 5: An example of the annotation we performed, where '-' indicates 'missing' and '+' indicates 'erroneous'. Additionally, 'i' indicates 'insertion' and 'd' deletion, so '-i' and '+d' is a missing insertion and erroneous deletion respectively. Collating the grammatical information in the dataset with these annotation allowed us generalize over the errors.

Error type	Percentage	Error Number
Noun class	14%	20
Verb class	10%	15
Stem edits	72%	128

Table 6: A summary of the results from our errors analysis. Results do not sum to 100% since these are only the most frequent errors and can co-occur.

discussed above. We have no *a priori* reason to expect the model to improve its performance on verb class errors, since class membership is a lexical property of the verb stem and not semantically conditioned. However, verbal derivational morphology can affect a verb's meaning and also determines its inflection class, so an indirect effect of semantics is possible. We show below that embeddings are also helpful for verbs, an issue we return to in Section 7.

We modified the original MED code, built in Blocks,[5] so that the output from the encoder could be concatenated with the 300-dimensional word embedding from Kutuzov and Andreev (2015). Since using these embeddings more than doubles the parameter space of the MED system, the model takes longer to converge. We therefore allowed the system to train up to 50 epochs, instead of the 20 Kann and Schütze needed for their models to converge. Both the original MED system and our modified version use early stopping. Once the model has converged, we evaluate system performance by measuring accuracy at the word level.

6 Results

The overall accuracy rates of a single trained MED system and our system are shown in Table 8. Following Kann and Schütze (2016), we also train and evaluate ensembles of five models (Table 9). In each case, our model performs about one percentage point better (significant using McNemar's test). The jump in significance scores between the validation and test is due to the relative sizes of these datasets (1,591 and 22,334, respectively).

System	Val	Test
MED base system	90.03	88.88
MED + word embeddings	91.95*	90.06****

Table 8: Overall results on the validation and test set, using only a single trained model (ensemble of 1). Significance is reported using McNemar's test where * indicates $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$.

System	Val	Test
MED base system	92.14	91.49
MED + word embeddings	93.33*	92.38****

Table 9: Overall results on the validation and test set, using an ensemble of 5 trained models (ensemble of 5). Significance is reported using McNemar's test where * indicates $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$.

We reapply our error analysis to determine error reduction rates by error category. Reductions were largest in noun cases and verb class, with a reduction of more than 50% for both. As seen in Table 10, stem edit errors were least improved. For a breakdown of errors by noun class, see Table 12.

We conduct two other experiments to rule out alternate accounts of the performance increase. First, our model with word embeddings has access to higher-dimensional input for decoding (600 dimensions vs. 300), and therefore to more parameters. We ran a model with 600-dimensional embeddings but no word embeddings, in order to test whether this could be responsible for the gain, but found no significant differences from our baseline system.

Second, we do not expect the word embedding system to encode inflectional information directly (since it operates at the word level with no access

[5] https://github.com/mila-udem/blocks

Error type	Decrease in error rate	Current error rate
Noun class	64.2%	5%
Verb class	88.1%	1%
Stem edits	44.1%	40%

Table 10: Overall error reduction rates in all three error types we considered.

to character information). However, we make absolutely sure that this is not the case by retraining the word embeddings on a stemmed version of our Russian corpus (processed with the NLTK stemmer (Bird, 2006)). Performance using these word embeddings is not significantly different from our results using regular word embeddings.

The error reduction rates by category which we report above are based on the relatively small SIGMORPHON 2016 validation set, and do not represent enough data to conduct statistical analyses by category or paradigm cell. To further break down the improvements quantitatively, we created secondary evaluation sets containing more items. For nouns, we created a secondary evaluation set with the Universal Dependency RusSynTag corpus[6] since it annotates both animacy and gender. We removed any nouns that did not have a 1-to-1 feature correspondence with the SIGMORPHON dataset.[7] This gave us a new evaluation set of 48,590 wordforms. Similarly, we also built a second evaluation set of 25,000 verb forms from Unimorph (Kirov et al., 2016). Although verb conjugation class is not directly annotated, we extracted that information from the second person singular present indicative form. In both cases, we removed any word form that also occurred in the training data.

As seen in Table 11, using word embeddings almost halved the error rate of e-conjugation verbs. It is important to note that the citation form supplied often requires less editing to make an i-conjugation verb than an e-conjugation verb since the citation form often has the *-i* theme vowel. Since the model has a strong preference for reproducing the input, our modification has minimal effect for i-conjugation verbs.

Verbs	Error count	Total words	Error rate
i-conj	163	516	0.3159
e-conj	430	3191	0.1348
With embeddings			
i-conj	161		0.3120
e-conj	273		0.0856

Table 11: Verb class-specific error reduction rates from 25,000 randomly sampled verb forms from the Unimorph Russian dataset.

Noun class	SG/PL	Error rate	Error rate+	Total count
IA	SG	0.2487	0.2132	2340
	PL	0.4244	0.3839	1555
IB	SG	0.0239	0.0427	1170
	PL	0.1818	0.1439	396
II	SG	0.0542	0.0274	1753
	PL	0.1826	0.1366	805
IIIA	SG	0.0736	0.0851	611
	PL	0.3016	0.1905	126

Table 12: Noun class-specific error reduction rates in the accusative case from 48,590 randomly sampled noun forms from the Universal Dependency RusSynTag dataset. "Error rate+" indicates the error rate after adding word embeddings to the MED system. IIIB and IIIC are not included since there are few nouns and no accusative errors were produced for them by the MED system.

Table 12 shows the general reduction in errors caused by adding word embeddings in various classes of the accusative. We note that errors in accusative forms increase only in class/number combinations that do not exhibit animacy-conditioned syncretism (i.e. singular of classes IB and IIIA).

7 Discussion

What inflectionally useful information is present in the word embeddings? As previously stated, we assume that word embeddings give good clues for noun animacy, but verbs form is not directly conditioned by semantic properties, so we have no *a priori* reason to assume they will indicate verb conjugation. To test whether these features can be derived from the embeddings, we construct maxent classifiers,[8] with only word embeddings as

[6] Freely available here: https://github.com/UniversalDependencies/UD_Russian-SynTagRus.git.

[7] These were generally cases where features were missing in the Universal Dependency corpus that were present in the SIGMORPHON corpus.

[8] We use Daumé III (2004)'s implementation available here: http://users.umiacs.umd.edu/~hal/

features, for two binary classification tasks: animate vs. inanimate for nouns and i-conjugation vs. e-conjugation for verbs. Using the same two datasets described in Section 6 for testing nouns and verb class error reduction, we extracted the verb class and animacy annotation along with the citation form's word embedding to create a classification task. With a baseline accuracy rate of 80% for both tasks (i.e. selecting the majority class), both classifiers were more than 98% correct.

We were unsurprised that animacy could be detected in this way, since word embeddings are already used in high-performance models for this kind of lexical feature (Moore et al., 2013; Rubinstein et al., 2015). The model's success for verbs is more surprising. One possible explanation is that Russian verb classes are indirectly related to lexical semantics (Aktionsart). As noted above, derivational suffixes determine the inflection class membership of verbs. Some derivational affixes also create verbs with predictable lexical aspectual properties (e.g. *-nu* creates semelfactives) (Isačenko, 1960; Janda, 2007; Dickey and Janda, 2009), and these semantic properties might be detectable from word embeddings alone. [9] Another possibility is that the predictability of verb class reflects the historical origins of some Russian verbs. Subclasses of verbs borrowed from Church Slavonic tend to have predictable assignments to classes, and also to be more bookish, abstract or metaphorical than native Russian terms (Townsend, 1975; Cubberley, 2002), which may render them recognizable to a distributional system. In any case, the classifier results validate our explanation of why our model improves by showing that the word embeddings do contain the information which the model needs to accurately predict semantically-conditioned allomorphs.

At a higher level, this highlights the issue of semantic conditioning as one which should be taken seriously in models of inflection and the PCFP. Current neural models, which take only word *forms* but not *meanings* as input, are insensitive to this kind of conditioning. They therefore yield overestimates of how difficult it is to acquire and use some morphological systems, such as Russian.

Although our error analysis methods and model extension focused on LMU's 2016 implementation of MED, more recent systems (Aharoni and Goldberg, 2017; Kann and Schütze, 2017) are subject to the same criticisms, since they use the same input representation. In this paper, we focus on Russian, as a language with lower-than-average performance in an inflection task and with a well-described system of inflection classes and alternations. However, we believe it is worth looking for similar effects in less well-studied languages as well, particularly given the wide range of languages now represented in Unimorph (Kirov et al., 2016).

8 Conclusion

Neural networks are a promising technology for cognitive models of a variety of language processing tasks. Their ability to learn flexible representations of complex, multidimensional data allows them to cover a wide range of linguistic phenomena which were difficult to model in more traditional frameworks. In morphology, this corresponds to adopting an "a-morphous" framework in which we do not need to commit to the existence of troublesome constructs like segmentable morphemes. But the adoption of neural nets as cognitive models has demanded a new focus on interpretation. It has become increasingly clear that networks are useful models only to the extent that we can compare what they are learning to what humans learn, and that this is a challenging area of research in its own right.

This work presents a new way to evaluate morphological inflection systems in a linguistically sensitive manner by repurposing previous work in edit rule induction to analyze and group error types. This allows us to attribute errors in inflection generation to specific, interpretable phenomena. We make our code and our expanded datasets publicly available for future use.[10]

We use this new method to discover that semantically- and lexically-conditioned allomorphy are responsible for a shortfall in inflection performance (and thus an overestimate of PCFP complexity) for Russian. Using word embeddings as

megam/version0_3/.

[9] Since the data are not tagged for derivational morphology or lexical aspect, it is difficult to assess whether this is a cause of the model's improvement. Given that certain lexical aspects align more naturally with one grammatical aspectual value (perfective or imperfective), we examined whether there is a relationship between verb class and grammatical aspect. We found no correlation in the training or validation data, but this does not rule out the possibility of a lexical semantic effect.

[10] https://github.com/DavidLKing/SCiL-20.

a proxy for lexical semantics allows us to supplement the model's input and greatly reduce this source of error. In the future, we will investigate which other languages might show semantically-conditioned allomorphy, potentially even discovering semantic effects in languages where they were not previously known to exist. We will also apply our analysis technique to other models and languages, helping to close the gap between neural reinflection systems and full-scale cognitive models of the PCFP.

9 Acknowledgements

We especially thank Michael White, the Ohio State Computational Linguistics and Language Acquisition discussion groups, Ekaterina Kolbasova, and Katya Rouzina!

References

(2017). *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*, Dundee, United Kingdom. Association for Computational Linguistics.

(2019). *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy. Association for Computational Linguistics.

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in grammar: Form and acquisition*, pages 54–82.

Ackerman, F. and Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, pages 429–464.

Aharoni, R. and Goldberg, Y. (2017). Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2004–2015.

Albright, A. C. (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

Anderson, S. R. (1992). *A-morphous morphology*, volume 62. Cambridge University Press.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2-3):150–177.

Bird, S. (2006). Nltk: The natural language toolkit. In *COLING ACL 2006*, page 69.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42:531–573.

Blevins, J. P., Milin, P., and Ramscar, M. (2017). The Zipfian paradigm cell filling problem. *Perspectives on morphological organization: Data and analyses*, 10:141.

Bonami, O. and Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Brown, D., Corbett, G. G., Fraser, N., Hippisley, A., and Timberlake, A. (2007). Russian noun stress and Network Morphology. *Journal of Linguistics*, 34:53–107.

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2018). On the complexity and typology of inflectional morphological systems. *arXiv preprint arXiv:1807.02747*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Cubberley, P. (2002). *Russian: A linguistic introduction*. Cambridge University Press.

Daumé III, H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/.

Dickey, S. M. and Janda, L. (2009). *Xoxotnul, sxitril*: The relationship between semelfactives formed with *-nu-* and *s-* in Russian. *Russian Linguistics*, 33:229–248.

Durrett, G. and DeNero, J. (2013). Supervised learning of complete morphological paradigms. In *HLT-NAACL*, pages 1185–1195.

Faruqui, M., Tsvetkov, Y., Neubig, G., and Dyer, C. (2016). Morphological inflection generation using character sequence to sequence learning. In *Proc. of NAACL*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Isačenko, A. (1960). *Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim – Čast' vtoraja: morfologija*. Slovackoj Akademii Nauk.

Janda, L. (2007). Aspectual clusters of Russian verbs. *Studies in Language*, 31:607–648.

Kann, K. and Schütze, H. (2016). MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. *ACL 2016*, page 62.

Kann, K. and Schütze, H. (2017). The LMU system for the CoNLL-SIGMORPHON 2017 shared task on universal morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 40–48. Association for Computational Linguistics.

Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *arXiv preprint arXiv:1807.04783*.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Kutuzov, A. and Andreev, I. (2015). Texts in, meaning out: neural language models in semantic similarity task for russian. *Dialog 2015*.

Lignos, C. and Yang, C. (2016). *Morphology and Language Acquisition*, page 743764. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Moore, J., Burges, C. J., Renshaw, E., and Yih, W.-t. (2013). Animacy detection with voting models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 55–60.

Rubinstein, D., Levi, E., Schwartz, R., and Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 726–730.

Silfverberg, M. and Hulden, M. (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.

Sims, A. D. and Parker, J. (2016). How inflection class systems work: On the informativity of implicative structure. *Word Structure*, 9(2):215–239.

Socher, R., Bauer, J., Manning, C. D., et al. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 455–465.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press.

Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge University Press.

Townsend, C. E. (1975). *Russian word-formation*. Slavica Publishers.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? *EMNLP 2018*, page 211.

Xu, W., Auli, M., and Clark, S. (2015). CCG supertagging with a recurrent neural network. In *ACL (2)*, pages 250–255.