Do RNN States Encode Abstract Phonological Alternations?

Miikka Silfverberg[ℨ]   **Francis Tyers**[ʕ]   **Garrett Nicolai**[ℨ]   **Mans Hulden**[χ]

[ℨ]University of British Columbia   [ʕ]Indiana University
[χ]University of Colorado

`first.last@ubc.ca`   `ftyers@iu.edu`   `first.last@colorado.edu`

Abstract

Sequence-to-sequence models have delivered impressive results in word formation tasks such as morphological inflection, often learning to model subtle morphophonological details with limited training data. Despite the performance, the opacity of neural models makes it difficult to determine whether complex generalizations are learned, or whether a kind of separate rote memorization of each morphophonological process takes place. To investigate whether complex alternations are simply memorized or whether there is some level of generalization across related sound changes in a sequence-to-sequence model, we perform several experiments on Finnish consonant gradation—a complex set of sound changes triggered in some words by certain suffixes. We find that our models often—though not always—encode 17 different consonant gradation processes in a handful of dimensions in the RNN. We also show that by scaling the activations in these dimensions we can control whether consonant gradation occurs and the direction of the gradation.

1 Introduction

Recent work on computational morphology demonstrates that neural networks can very effectively learn to inflect words, given adequate amounts of training data (Cotterell et al., 2016, 2017). However, in computational morphology and in NLP at large, the interpretability of neural models remains a serious concern (Doshi-Velez and Kim, 2017)—it is unclear how networks trained to inflect words actually accomplish their task. It is also unclear to which extent networks are able to learn linguistic generalizations from their input data instead of simply memorizing training examples and exhibiting a kind of nearest-neighbor behavior.

In this paper, we shed light on what kind of linguistic generalizations neural networks are capable of learning from data. We report on an in-
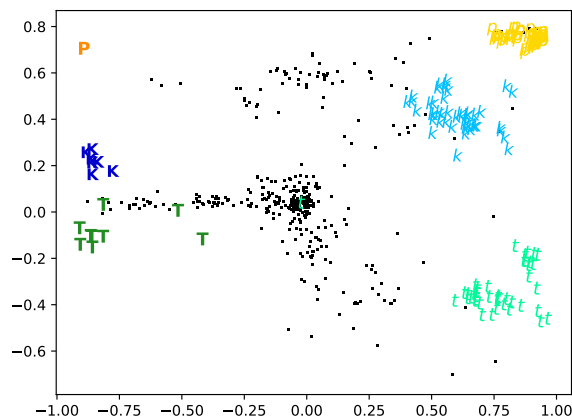

Figure 1: Scatter plot of the activation for two encoder hidden state dimensions which activate strongly during gradation. The letters **p**, **t** and **k** refer to examples which undergo gradation for the respective consonants. Lower-case characters mark direct (weakening) gradation and upper-case characters inverse (strengthening) gradation (see Section 3 for further details). Black dots · mark examples which do not undergo gradation.

vestigation into how *consonant gradation*, a particular morphophonological alternation which is common in Finnish and other Uralic languages, is encoded in the hidden states of an LSTM encoder-decoder model trained to perform word inflection. Specifically, we train character-based sequence-to-sequence models for inflection of Finnish nouns into the genitive case, an inflection type which commonly triggers consonant gradation.

Consonant gradation is a morphophonological alternation where voiceless stops **p**, **t** and **k** are lenited in certain positions (see Section 3 for further details). We first demonstrate that inflection networks tend to learn an abstract representation for consonant gradation, where the alternation is triggered by the same dimensions in encoder hidden states regardless of which stop **p**, **t** or **k** undergoes gradation. This echoes the treatment of gradation in linguistic literature (Hakulinen et al., 2004, §41) Nevertheless, we also find evidence that

5501

this behavior is not universal and that networks can sometimes fail to generalize gradation and instead learn to represent gradation using distinct dimensions for each stop **p**, **t** and **k**.

Our second contribution is to show that networks can learn a general representation encompassing both so-called *quantitative gradation* and *qualitative gradation* (these are further described in Section 3). This presents further evidence that the phonological representations learned by encoder-decoder models can learn to group linguistic generalizations that target different sounds.

As our third contribution, we show evidence of a remarkable property whereby directionality of gradation is encoded as positive or negative hidden state activations: Consonant gradation is called direct when the base form of a noun displays the strong grade (such as **kk**) and the genitive form displays the weak grade of a stop (such as **k**). In inverse (or 'strengthening') gradation, the opposite alternation occurs. We find hidden state dimensions which encode for the direction of gradation by a positive or negative activation. This behavior is demonstrated in Figure 1 where a negative activation of dimension 487 in the encoder hidden state marks inverse gradation of a stop, and positive activation instead marks direct gradation (see Section 6 for further discussion of this phenomenon).

2 Related Work

Interpretation of neural representations in recurrent neural models has been an active area of research over a long period of time starting with Elman (1990). However, representations in models of phonology have received less attention than many other subfields of NLP. Rodd (1997) investigates learning of Turkish vowel harmony by a character-based RNN language model trained on word forms. The paper investigates hidden state activations of RNN models while varying the hidden state dimensionality between 1 and 4. It presents evidence that RNN hidden states can capture Turkish vowel harmony patterns when a sufficient number of hidden dimensions are available. In a similar vein, Silfverberg et al. (2018) investigate phoneme representations for Finnish, Spanish and Turkish finding correlations between embedding representations and phonological distinctive features. Kolachina and Magyar (2019) present an investigation of phone embeddings learned using word2vec (Mikolov et al., 2013) for simulated data showing

that phone embeddings capture phonemic and allophonic relationships. They also show that phone embeddings capture co-occurrence restrictions for vowels well, while largely failing to do this for consonants. Our encoder representations, in contrast, are able to capture these co-occurrence restrictions.

Beguš (2020b) investigates representations learned by a generative adversarial network or GAN (Goodfellow et al., 2014) trained on audio recordings of speech, showing that some of the latent variables of the GAN correspond to phonological features of the speech signal: specifically the presence or absence of the fricative [s] in the output of the network and the amplitude of frication. They show that manipulation of the variables changes these features in a predictable manner. Similarly to our work, Beguš (2020b) also scales state activations and observes the effect on the output of the network. In a related investigation of reduplication, Beguš (2020a) train GAN models on speech and identify variables which trigger reduplication in the speech signal.

Extensive work exists on linguistic probing experiments for neural representations (Conneau et al., 2018a,b; Clark et al., 2019). A recent probing paper by Torroba Hennigen et al. (2020) is more directly related to our work. They present a decomposable probe for finding small sets of hidden states which encode for linguistically relevant information, particularly morphosyntactic information. Our work shares the aim of not only identifying if information is present in a neural system, but also examining how it is represented. However, we additionally perform experiments on manipulating network activations and examine how such manipulations influence the outputs of the network.

Our approach was inspired by the now-classic paper on visualization and interpretation of recurrent networks by Karpathy et al. (2015) in that we also seek individual interpretable dimensions. The work by Dalvi et al. (2019) on analyzing individual neurons in networks trained for linguistic tasks (POS tagging as well as semantic and morphological tagging) is more closely related to the present work. They present a general methodology for uncovering neurons which encode linguistic information by training a classifier to predict linguistic features of the input based on the representations generated by the network. They also show that it is possible to manipulate specific neurons to force the

	Nominative	Genitive	Gloss
Quantitative	pappi	papin	'priest'
	kenttä	kentän	'field'
	kiukku	kiukun	'anger'
	ripe	rippeen	'remain(s)'
	laite	laitteen	'device'
	liike	liikkeen	'motion'
Qualitative	sopu	sovun	'agreement'
	johto	johdon	'lead'
	aika	ajan	'time'
	kyky	kyvyn	'skill'
	olento	olennon	'creature'
	kenkä	kengän	'shoe'
	silta	sillan	'bridge'
	rumpu	rummun	'drum'
	ranne	ranteen	'wrist'
	salko	salon	'pole'

Table 1: Examples of various kinds of consonant gradation in Finnish. Adding the genitive suffix **-n** closes the final syllable triggering gradation.

network to generate a particular linguistic feature. Meyes et al. (2019) investigate the effect of scaling network activations, something they call *ablation* studies. They train classifiers and look at how classification performance varies when zeroing out the activations for particular states or groups of states.

3 Consonant Gradation

Consonant Gradation (CG), common in many Uralic languages, is a set of assimilation and lenition processes, usually targeting the final syllable in a word stem. Historically the trigger for the alternation has been purely phonological, but in Finnish, the alternation is no longer entirely predictable from the phonological structure (Karlsson, 2017).[1] The trigger for gradation is usually an affix that closes the final syllable, such as the genitive **-n**, e.g. **katto** ~ **katon** ('roof' sg. nom. ~ sg.[2] gen.). The overall process is divided into *quantitative gradation* where, for example, geminate **pp**, **tt**, **kk** alternate with their non-geminate counterparts, **p**, **t**, **k**, and *qualitative gradation* where a large variety of lenition and assimilation processes are found. For example, strong grade **k** can alternate with the weakened **j**, **v**, **g**, etc. See Table 1 for a summary of these types of gradation processes found in our data set. The lenited or elided forms are commonly called the *weak grade* (e.g. **katon**) and the alter-

nant the *strong grade* (e.g. **katto**). Sometimes the weak and strong grades appear in the inverse position, i.e. the weak grade appears with open syllables as in **rike** ~ **rikkeen** ('offense' sg. nom. ~ sg. gen.).

While quantitative gradation remains productive in the language, many stems from more recent loanwords in particular, do not tend to alternate qualitatively; for example **auto** ~ **auton**, *audon ('car' sg. nom. ~ sg. gen.). Speakers must therefore know the lexical status of each stem to inflect it correctly. Our data set includes both gradating and non-gradating lexemes.

The advantages of studying Finnish consonant gradation in this context is that the set of sound changes is very diverse, but that the trigger for all of them is the same. Also, the Finnish writing system is very phonemic and surface-oriented and therefore no conversion to an IPA representation is necessary to reveal the sound changes that occur as a result of gradation.

Of particular interest to us is that there are many similar-looking alternations in Finnish that are not a result of consonant gradation, but paradigmatic variation. For example, **varis** ('crow' sg. nom.) is inflected **variksen** in the sg. gen. form. Note the similarity of this alternation to the actual CG case of **liike** ('motion' sg. nom.) ~ **liikkeen** (sg. gen.) which also involves a $\emptyset$ ~ k alternation. It is therefore of some interest to observe whether neural inflection models encode the two cases differently in some respect.

In total we count 17 different types of lenition or fortition falling under the rubric of consonant gradation in our data set; an example of each type is shown in Table 1.

4 Methods

This section presents our nominative → genitive inflection models and our approach to finding encoder hidden state dimensions which are associated with consonant gradation.

4.1 Inflection Models

As our inflection model, we use the well-known attentional BiLSTM encoder-decoder model which was presented by Bahdanau et al. (2014) and first applied to inflection by Kann and Schütze (2016). This neural model transduces a nominative input form which is represented as a sequence of characters $x_{[1:T]}$ of length T into a genitive output form

---

[1]Nevertheless, the alternation is fully determined by the phonological context in the nominative and genitive cases which are the focus of this paper. Therefore, we treat gradation as a phonological alternation.

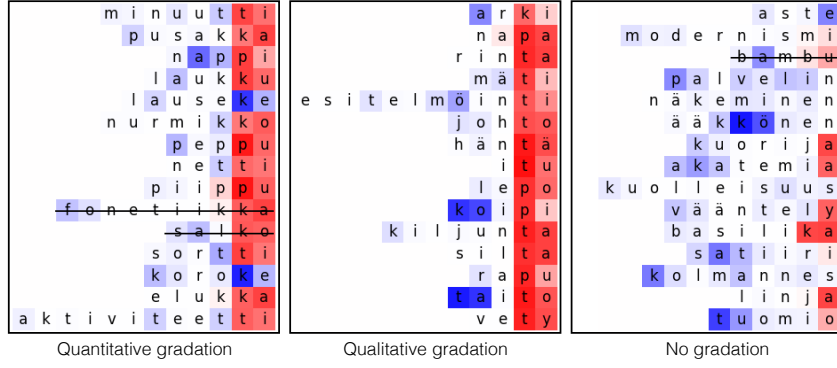[2]Underlining marks the position affected by gradation.

Figure 2: Activation of encoder hidden state dimension 487 in model 3 when (1) quantitative gradation, (2) qualitative gradation and (3) no gradation occurs in the input form. Red squares indicate positive and blue squares negative activation. Forms with strikethrough are inflected incorrectly by the decoder: ***fonetikan** (**fonetiikan**), ***salkon** (**salon**) and ***bammun** (**bambun**).

$y_{[1:S]}$ of length S.

The encoder network in an attentional model generates one hidden state vector $\mathbf{h}_t = \mathbf{f}_t \oplus \mathbf{b}_t \in \mathbb{R}^{2n}$ for every position in the input sequence. Due to the bidirectionality of the encoder, the hidden state vector is a concatenation of a forward state $\mathbf{f}_t \in \mathbb{R}^n$ and a backward state $\mathbf{b}_t \in \mathbb{R}^n$. We refer to the vectors $\mathbf{f}_t$ as hidden states and the elements in the vectors $\mathbf{f}_t[d]$ as *activations*. Here $d \in \{1, 2, 3, ..., 2n\}$ is called a *dimension*.

4.2 Finding Dimensions Associated with Gradation

Our aim is to investigate encoder hidden state dimensions d which are associated with gradation. To this end, we extract the encoder hidden state activations $\mathbf{h}_1[d], ..., \mathbf{h}_T[d]$ for each example $(x_{[1:T]}, y_{[1:S]})$ in our development set and dimension $d \in \{1, 2, 3, ..., 2n\}$. In order to find dimensions which activate strongly at positions where gradation occurs, we compare the mean activation of each dimension for forms which undergo gradation and forms which do not.

Let $a_X(d)$, as defined by Equation (1) below, be the mean activation for dimension d in a set of encoder hidden states X. For each dimension d, we extract the mean activation $a_G(d)$, where G is the set of encoder hidden states at positions where gradation occurs. As explained in Section 3, gradation applies to the final stop in word forms which undergo gradation. Usually, this would refer to position $T - 1$ in a string of length T as in **tupa** 'cottage sg. nom.', where **p** undergoes gradation, but can also happen at position $T - 2$ as in the form **ratas** 'wheel sg. nom.', where **t** undergoes gradation.

$$a_X(d) = \frac{\sum_{h \in X} \mathbf{h}[d]}{|X|} \qquad (1)$$

The mean activation $a_G(d)$ is compared to the activation $a_N(d)$ of dimension d at the penultimate position $T - 1$ in base forms of length T which do not undergo gradation. In order to specifically capture dimensions which encode for gradation as opposed to simply encoding for consonants, we limit this examination to base forms like **kana** 'chicken sg. nom.' and **auto** 'car sg. nom.', where the penultimate character is a consonant. We retrieve the top-N dimensions d where the difference in mean activation $|a_N(d) - a_G(d)|$ is maximized and consider these candidate dimensions for gradation.

5 Data

Our dataset was produced by taking the most frequent 5,000 lexemes tagged as singular nominative nouns from the Turku Dependency Treebank (Haverinen et al., 2014) and generating the singular genitive forms using the OmorFi finite-state morphological transducer (Pirinen, 2015). We excluded compound nouns (e.g. **ammattikorkeakoulututkinnoista** 'from the professional high-school examinations') and words marked as nouns which contained punctuation or numerals (e.g. **G8-neuvottelut** 'G8 negotiations', **2000-luvulla** 'in the 2000s', **°C:ssa** 'in °C' etc.). Loan words were included, both unadapted such as **workshop** and **bungalow** and partially or fully adapted such as **brosyyri** 'brochure' and **samppanja** 'champagne'. This gave a total of 4,797 nominative–genitive pairs. We randomly ordered them and then split these into disjoint sets: 90% for

	Ungrad.	Grad.		
		Qual.	Quant.	Total
p	35	14	40	54
t	159	24	30	54
k	128	4	50	54
Total:	322	42	120	162

Table 2: Composition of the manually annotated validation set. The **Ungrad.** column refers to forms that do not exhibit gradation in the genitive, **Grad.** those forms that exhibit gradation and **Qual.** is qualitative and **Quant.** is quantitative gradation.

training (4,317 pairs) and 10% for validation.

We then took the validation set (479 pairs) and annotated them for: gradation (yes, no), type of gradation (qualitative, quantitative), consonant (**p**, **t**, **k**) and direction (direct, inverse). This gave a total of 84 examples of nouns exhibiting consonant gradation.

This set was heavily skewed towards **t** gradation (54 out of 84 examples).[3] So we randomly sampled another 84 words from the frequency list, which were not found in the training data or in the existing validation set and which contained **p** and **k**, and annotated them and added them to the validation set. Statistics on the composition of the hand-annotated dataset can be found in Table 2 and the full data is freely available on GitHub.[4]

6 Experiments and Results

We investigate representation of consonant gradation in encoder hidden states in the following way: As explained in Section 4.2, we identify individual dimensions in encoder hidden states which activate strongly during gradation regardless of the identity of the consonant undergoing gradation. We then investigate the association of these states using two experiments: we (1) perform significance tests on a held-out dataset to determine if the states activate significantly more strongly when gradation occurs, and (2) scale the state activations and observe the effect on the output of the network.

6.1 Training Details

We train ten encoder-decoder models with different random initializations for inflection using the OpenNMT toolkit (Klein et al., 2018). We use a 2-layer BiLSTM encoder with hidden dimension
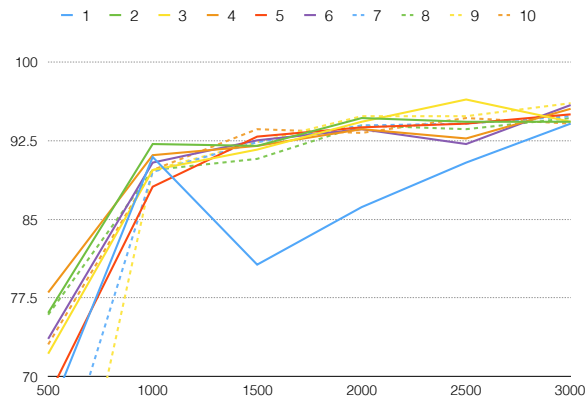
Figure 3: Percent inflection accuracy on the development set for steps 500-3,000 during the training of our ten inflection models. The figure demonstrates that changes in accuracy on the development set are modest after step 2,000.

250. Due to the bidirectionality of the encoder this results in 500-dimensional hidden states (consisting of a forward and backward hidden state). Our model uses 500-dimensional character embeddings both in the encoder and decoder and we use an attentional decoder with 250-dimensional hidden states. The model is trained for a total of 3,000 steps using stochastic gradient descent and a batch size of 64. See Figure 3 for a plot of the development accuracy during the training process. As can be seen, changes in development accuracy are modes after training step 2,000.

We report inflection accuracy for our ten inflection models measured on held-out data in Table 3. The accuracy is reported separately for forms undergoing gradation and forms not undergoing gradation. In addition, we report an overall accuracy for all forms. We can see that the mean performance is close 95% for all forms and performance tends to be higher on forms undergoing gradation than other forms.

6.2 Investigation of State Activations

We randomly split our development set into two disjoint parts of equal size. The first part of the development set we use to discover the top-5 encoder hidden state dimensions which are strongly associated with gradation (as described in Section 4.2). The rest of the development set is used for significance testing. We perform a two-sided t-test to check if the mean activations of our top-5 dimensions differ significantly (at the 99.5% significance level) between positions which undergo gradation

Model	# States	+Grad.	-Grad.	Overall
1	2	94.4	94.0	94.1
2	2	95.1	94.0	94.3
3	1	94.4	94.3	94.3
4	3	96.9	95.0	95.6
5	3	95.1	95.0	95.0
6	3	97.5	95.3	95.9
7	1	95.7	94.5	94.9
8	0	95.7	94.3	94.7
9	0	96.9	95.8	96.1
10	2	96.9	93.0	94.1

Table 3: Percent inflection accuracy for 10 NOM to GEN models trained using different random seeds. The column # States refers to the number of states found in Table 4 that have significant activations for all gradation types.

and positions which do not undergo gradation. As explained in Section 4.2, we limit this examination to nominative forms where the penultimate character is a consonant to better zone in on gradation.

Table 4 shows the results separately for **p**, **t** and **k** gradation. The table also shows results for qualitative and quantitative gradation. We can see that eight of the ten models contain at least one dimension where activation is significantly stronger for all stops **p**, **t** and **k** undergoing gradation than other stem-final consonants indicating that these states are associated with gradation in general rather than gradation of one of the individual consonants **p**, **t**, or **k**. We note that these dimensions also typically activate both for qualitative and quantitative gradation indicating that the network has learned an abstraction for both types of gradation.

6.3 Scaling State Activations

As a direct test of the effect of hidden state dimensions on gradation, we scale the activations of dimensions which are strongly associated with gradation. Our hypothesis is that negatively scaling these dimensions will prevent forms from undergoing gradation.

We experiment on a dataset consisting of all development examples which undergo gradation. For each nominative input form such as **luuk̲ku**, we identify the correct gold standard genitive form **luuk̲un** (where **kk** $\rightarrow$ **k** alternation has applied) and an alternate output form ***luuk̲kun** which is correct apart from the fact that the form has not undergone gradation. We then compute (1) the number of gold standard forms, (2) the number of alternate forms, and (3) the number of nonce forms

generated by our models. Nonce forms here refer to erroneous outputs like ***luukuukuukkun** which do not belong in category (2).

We scale the hidden state activations at positions where gradation occurs, that is at the final stop in the nominative form, before feeding the encoder hidden states into the decoder. For each input form, we scale the top-N encoder hidden states which are associated with gradation according to the mapping $a \mapsto x \cdot a$ where x varies between 1 and -25. The number of states which are scaled (that is N) is tuned for maximal effect on the number of alternate forms which are generated.

Figure 4 shows the results for the scaling experiment when tuning N.[5] The first graph shows that for most models the number of alternate forms first increases when the scaling factor x approaches -25, and then gradually decreases. As the number of alternate forms increases, the number of gold standard forms undergoing gradation naturally decreases as demonstrated by the second graph. We also see an increase in the number of nonce forms which do not belong to either category. This is to be expected as scaling represents a deviation from learned model weights which disturbs the network.

The effect of scaling varies between models: When scaling activations for Model 9, over half of the output forms do not undergo gradation. In contrast, for Model 7, the best scaling factor only produces around 7% of non-gradating output forms. Crucially, however, we do see an effect for nearly all models (apart from model 8). Contrast this with Figure 5 which shows results when scaling a set of five random states instead of states which are associated with gradation, showing that scaling of randomly sampled states has very small if any effect on the number of alternate forms produced by the models. Based on the graphs in Figure 4, scaling has very limited effect on Model 8. Even when scaling by $a = -25$, there is only a small decrease in the number of gold standard forms and a corresponding small increase in nonce forms. This might be evidence of a more redundant representation of information in Model 8, whereby scaling a few states will not strongly perturb the network.

7 Discussion and Conclusions

Figure 2 shows the activation for a hidden state dimension which is strongly associated with grada-

---

[5]For completeness, all results for scaling the top-N states where N varies between 1 and 5 are shown in Appendix A.

Model 1

Gradation	Dimension				
	458	**252**	395	489	396
K	**0.247**	**0.162**	0.048	**0.156**	**0.649**
P	**0.291**	**0.350**	**0.390**	**0.283**	0.056
T	**0.175**	**0.062**	0.022	0.043	**0.193**
Qual.	**0.259**	**0.176**	**0.200**	**0.136**	0.069
Quant.	**0.252**	**0.218**	**0.199**	**0.194**	**0.245**

Model 2

Gradation	Dimension				
	255	127	453	**399**	**469**
K	0.080	**0.285**	**0.395**	**0.319**	**0.281**
P	**0.557**	**0.354**	**0.371**	**0.260**	**0.228**
T	**0.168**	0.158	**0.122**	**0.233**	**0.277**
Qual.	**0.330**	0.047	0.132	**0.242**	**0.224**
Quant.	**0.314**	**0.368**	**0.335**	**0.277**	**0.252**

Model 3

Gradation	Dimension				
	487	207	**203**	484	221
K	0.205	**0.412**	**0.393**	**0.263**	**0.276**
P	**0.717**	**0.471**	**0.248**	**0.622**	**0.366**
T	**0.567**	0.080	**0.324**	**0.475**	0.127
Qual.	**0.684**	**0.254**	**0.268**	0.118	**0.249**
Quant.	**0.440**	**0.423**	**0.323**	**0.358**	**0.289**

Model 4

Gradation	Dimension				
	349	**163**	124	**416**	254
K	**0.229**	**0.498**	**0.540**	**0.078**	**0.364**
P	**0.260**	**0.081**	**0.115**	**0.414**	0.099
T	**0.386**	**0.193**	**0.082**	**0.070**	0.139
Qual.	**0.222**	**0.122**	0.006	**0.294**	0.068
Quant.	**0.270**	**0.290**	**0.297**	**0.214**	**0.273**

Model 5

Gradation	Dimension				
	345	**109**	408	390	**191**
K	**0.272**	**0.206**	**0.775**	0.041	**0.225**
P	**0.466**	**0.381**	**0.105**	**0.498**	**0.183**
T	**0.498**	**0.186**	0.057	**0.161**	**0.257**
Qual.	**0.397**	**0.241**	0.068	**0.206**	**0.162**
Quant.	**0.395**	**0.296**	**0.298**	**0.224**	**0.222**

Model 6

Gradation	Dimension				
	480	**22**	418	**48**	320
K	**0.303**	**0.240**	**0.257**	**0.191**	**0.651**
P	**0.394**	**0.353**	**0.346**	**0.297**	0.045
T	**0.593**	**0.073**	0.012	**0.198**	0.055
Qual.	**0.320**	**0.140**	**0.180**	**0.162**	0.018
Quant.	**0.388**	**0.294**	**0.275**	**0.262**	**0.267**

Model 7

Gradation	Dimension				
	142	108	230	207	**405**
K	**0.212**	**0.223**	**0.138**	0.068	**0.309**
P	**0.319**	**0.452**	**0.421**	**0.438**	**0.169**
T	0.172	**0.227**	0.032	0.028	**0.122**
Qual.	0.180	0.099	0.148	0.170	**0.166**
Quant.	**0.268**	**0.282**	**0.278**	**0.232**	**0.226**

Model 8

Gradation	Dimension				
	441	234	238	283	125
K	0.033	**0.272**	**0.191**	0.075	0.013
P	**0.463**	**0.243**	**0.240**	**0.352**	**0.260**
T	0.038	0.085	0.025	0.087	**0.241**
Qual.	**0.238**	0.080	**0.154**	0.149	**0.192**
Quant.	**0.238**	**0.227**	**0.193**	**0.185**	**0.164**

Model 9

Gradation	Dimension				
	446	459	438	380	351
K	0.184	**0.375**	**0.487**	**0.157**	**0.109**
P	**0.416**	**0.293**	**0.142**	**0.283**	**0.379**
T	**0.208**	0.030	**0.075**	0.124	**0.135**
Qual.	**0.258**	**0.137**	0.044	**0.242**	0.117
Quant.	**0.293**	**0.314**	**0.282**	**0.196**	**0.220**

Model 10

Gradation	Dimension				
	56	270	284	**367**	306
K	**0.563**	0.172	0.108	**0.357**	**0.438**
P	**0.153**	**0.326**	**0.429**	**0.145**	0.132
T	**0.194**	**0.177**	0.064	**0.076**	0.075
Qual.	**0.173**	**0.240**	**0.215**	**0.120**	0.158
Quant.	**0.348**	**0.232**	**0.218**	**0.235**	**0.305**

Table 4: Mean differences in activation strength for dimension d where we first find the top-5 states associated with gradation using 50% of the development data and then perform significance tests using the remaining 50% of the development data. We present results for 10 different random initializations of model parameters. We compare activation when **k**, **p** or **t** gradation occurs to activation at **-CV** word endings where gradation does not occur. We also report results for qualitative and quantitative gradation irrespective of the consonant undergoing gradation. Statistically significant differences in activation strength at the 99.5% significance level are shown in bold face. Dimensions with significantly stronger association for all stops as well as qualitative and quantitative gradation are marked using a **gray box**.
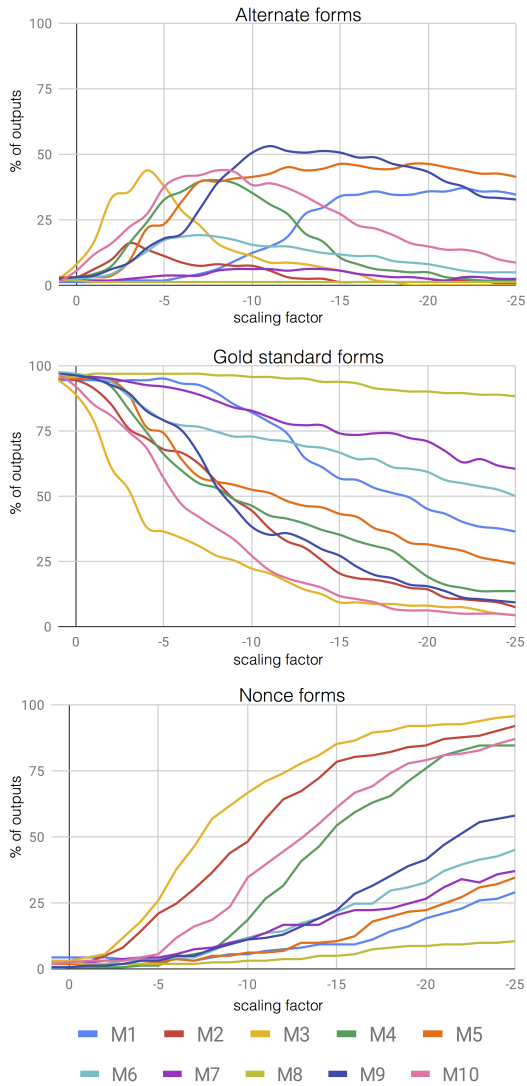
Figure 4: The amount (in %) of gold standard outputs, alternate outputs not displaying gradation and nonce forms which are generated when scaling top encoder hidden dimensions associated with gradation.
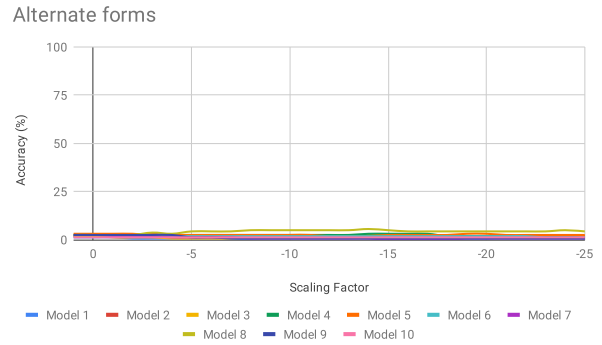

Figure 5: The amount (in %) of alternate outputs not displaying gradation when five randomly sampled encoder hidden dimensions are scaled.

prompted us to investigate hidden state activations more directly using the scaling experiments described in Section 6.3.

Figure 1 shows a scatter plot of two encoder hidden state dimensions (487 and 484 in model 3) which activate strongly during gradation. Each point in the plot corresponds to one example in our development dataset. Clearly, examples which do not undergo gradation cluster around $(0, 0)$.[7] In contrast, gradation for **k** and **p** lead to a positive activation for state 484, whereas **t**-gradation gives a negative activation. Moreover, direct gradation results in a positive activation for state 487 and inverse gradation gives a negative activation. Examples which do not undergo gradation can also have high values for 484 (> 0.4). Many of these examples end in **-jV**, **-vV** or **-mV** which could actually be examples where inverse gradation occurs but it happens not to be the case for these particular ones. Examples where the activation for 484 is low (< -0.5) span a small number of forms ending **-tV**, **-bV**, and **-gV**. There is also a substantial number of non-gradating forms where the activation for 484 is > 0.5. Most of these fall into the **linnoitus** 'fortress' / **linnoituksen** 'fortress sg. gen.' patterns where a **k** is inserted in the penultimate syllable. This alternation bears great resemblance to gradation as mentioned in Section 3. There are also a few examples of the type **tase** 'balance sheet' / **taseen** 'balance sheet sg. gen.' where the stem-final vowel is doubled displaying large activation for 484. This is perhaps somewhat harder to explain. However, note that this vowel doubling fre-

tion: dimension 487 in model 3. This dimension displays positive activation for consonants undergoing direct gradation as in **lau<u>kku</u>** 'bag sg. nom.' ∼ **lau<u>ku</u>n** 'bag sg. gen.'. Remarkably, the state displays negative activation for consonants undergoing inverse gradation as in the example **lause<u>ke</u>** 'phrase' where **k** is strengthened into a geminate **kk** resulting in the genitive form **lause<u>kke</u>en** 'phrase-GEN'. This effect can be seen both in forms where quantitative and qualitative gradation occurs. However, as the example **basilika** 'basil' in the third heat map demonstrates, dimension 487 can also activate strongly when no gradation occurs.[6] This

state 487, our model still correctly inflects **basilika** into **basilikan** instead of applying gradation, which would give a form like ***basilijan** or ***basilian**.

[7]The single **t** at $(0, 0)$ represents the pair **olu<u>t</u>** ∼ **oluen**, where **t** → ∅. This is an extremely infrequent gradation type.

---

[6]The form **basilika** is a loan word and would probably undergo gradation if it were a native Finnish word. It is noteworthy, however, that regardless of the strong activation of

quently co-occurs with gradation as in **tarvike** 'accessory', **tarvikkeen** 'accessory sg. gen.'.

In our experiments we found that the system would sometimes output a gradated form even when the exact type of gradation was not present in the training data, for example **ba<u>mb</u>u** ∼ **ba<u>mm</u>un** ('bamboo' sg. nom. ∼ sg. gen.). Since Finnish natively lacks **b** and **g**, examples of gradation with these consonants are rare. However, it is indeed the case that loanwords that include such voiced stops do undergo gradation, e.g. **du<u>b</u>ata** ∼ **du<u>bb</u>aan** ('to dub' inf. ∼ 1p sg. pres. sg.) (Voutilainen, 2008). Since native Finnish speakers seem to extend gradation from voiceless stops to their voiced counterparts in loanwords, the question whether neural models can exhibit such generalizing behavior as well is an interesting one. Our initial investigations into whether the similarity of the learned embeddings for **p** and **b** could trigger such generalizations across similar sounds failed to identify a clear reason for the behavior, and we leave a detailed study of this to future work.

We have presented an investigation of encoder representations of phonological alternations, specifically consonant gradation in Finnish. We found evidence of a generalized representation of gradation covering all stops which undergo gradation and different types of gradation. We also found that scaling hidden states can "switch off" gradation, prompting the model to generate alternate forms which do not display gradation. Moreover, the direction of gradation can be encoded as positive vs. negative hidden dimension activation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gašper Beguš. 2020a. Identity-based patterns in deep convolutional networks: Generative adversarial phonology and reduplication. *arXiv preprint arXiv:2009.06110*.

Gašper Beguš. 2020b. Modeling unsupervised phonetic and phonological learning in generative adversarial phonology. *Proceedings of the Society for Computation in Linguistics*, 3(1):138–148.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. 2004. *Iso Suomen Kielioppi [Comprehensive grammar of Finnish]*. Suomalaisen Kirjallisuuden Seura [Finnish Literature Society].

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531.

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.

Fred Karlsson. 2017. *Finnish: A comprehensive grammar*. Routledge.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169.

R. Meyes, M. Lu, C. Waubert de Puiseau, and T. Meisen. 2019. Ablation studies to uncover structure of learned representations in artificial neural networks. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 185–191. The Steering Committee of The World Congress in Computer Science.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tommi A. Pirinen. 2015. Development and use of computational morphology of Finnish in the open source and open science era: Notes on experiences with Omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Jennifer Rodd. 1997. Recurrent neural-network learning of phonological regularities in Turkish. In *CoNLL97: Computational Natural Language Learning*.

Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Eero Voutilainen. 2008. Bloggaaminen laajentaa astevaihtelua [Blogging extends consonant gradation]. https://www.kotus.fi/nyt/kolumnit_artikkelit_ja_esitelmat/kieli-ikkuna_(1996_2010)/bloggaaminen_laajentaa_astevaihtelua.

A Appendix: Scaling experiments

This appendix contains all results for the scaling experiment presented in Section 6. Figure 6 presents the amount of alternate forms produced by each model when 1 - 5 top gradation encoding hidden state dimensions associated with gradation are scaled. Figure 7 presents results for the gold standard forms undergoing gradation. For each model, we also present results for scaling a set of five randomly selected encoder hidden state dimensions.

As Figures 6 and 7 show, the effect of scaling dimensions associated with gradation has a clear positive effect on the number of output forms which do not undergo gradation. In contrast, scaling randomly selected encode hidden state dimensions has small effect overall on the number of these output forms although it does tend to reduce the number of gold standard outputs undergoing gradation. This means that the number of nonce output forms still increases when the scaling factor approaches -25 as might be expected because we are deviating from the learned models parameters.
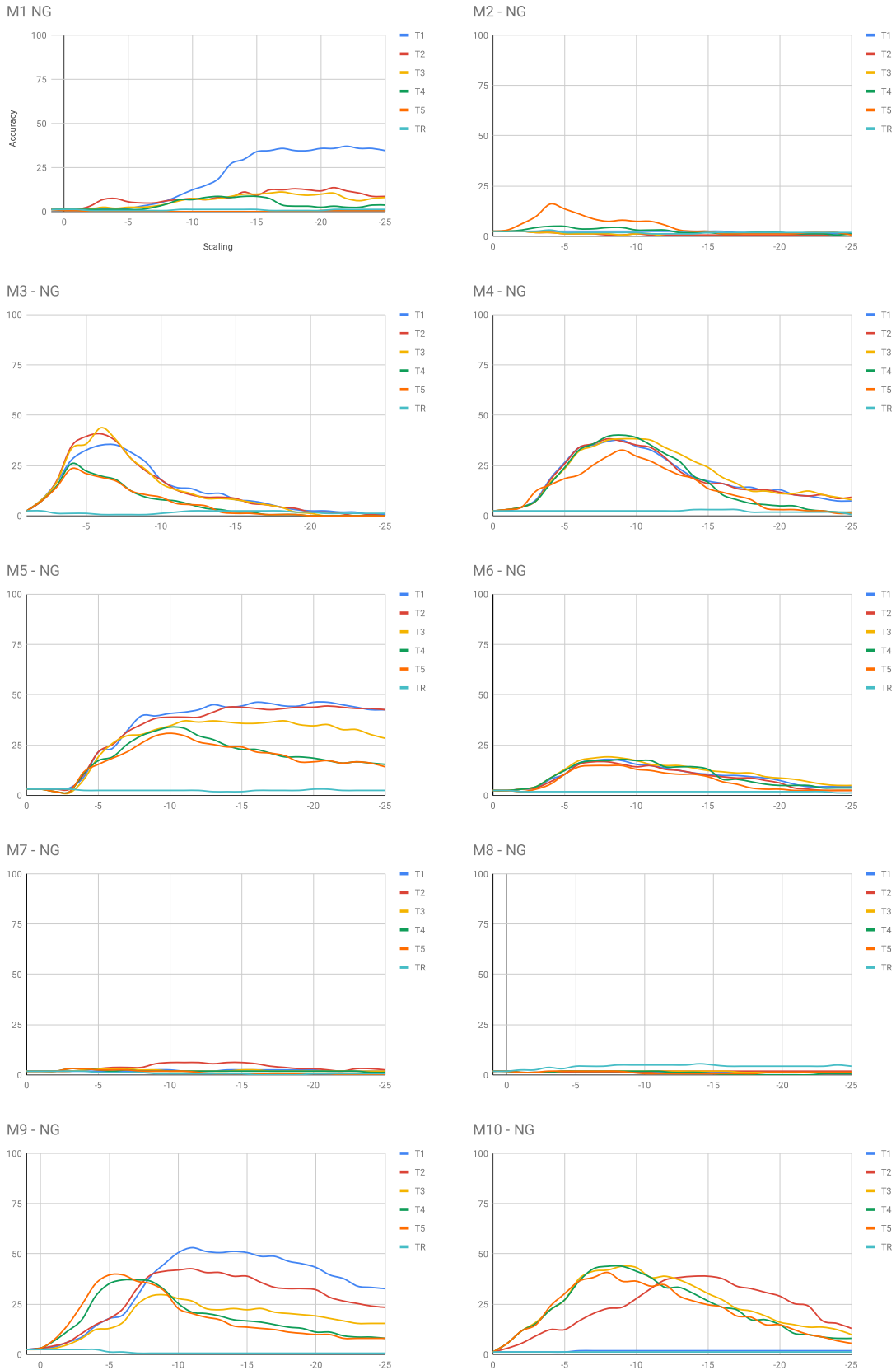
Figure 6: Results when scaling the activations for the top 1-5 (T1 - T5) encoder dimensions associated with gradation for each of out ten models M1 - M10. These graphs show the amount of outputs which do not undergo gradation which are produced when encoder dimensions are scaled. As comparison, the green TR graph shows the effect of scaling 5 randomly selected encoder dimensions.
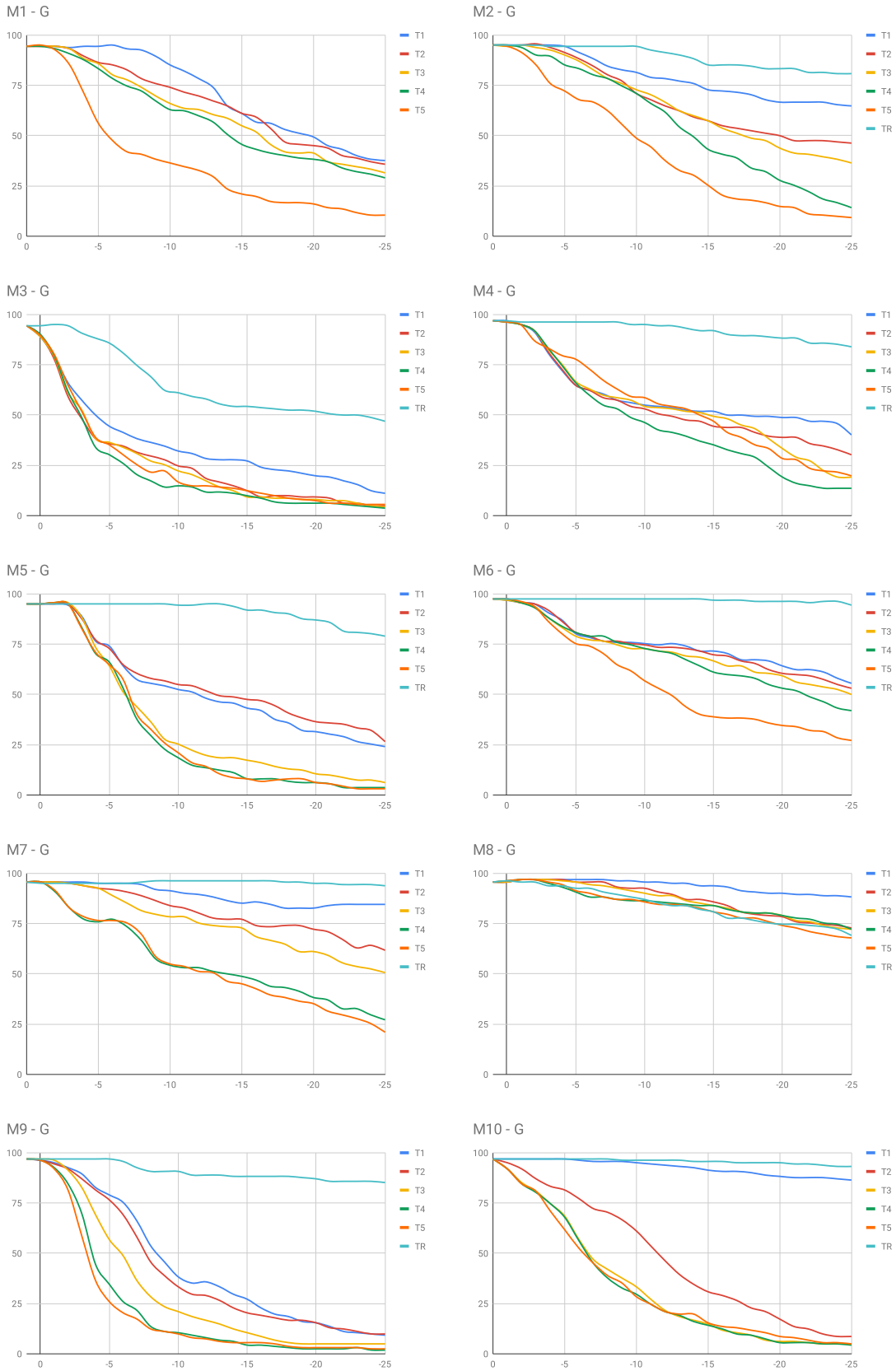
Figure 7: Results when scaling the activations for the top 1-5 (T1 - T5) encoder dimensions associated with gradation for each of out ten models M1 - M10. These graphs show the amount of gold standard outputs undergoing gradation which are produced when encoder dimensions are scaled. As comparison, the green TR graph shows the effect of scaling 5 randomly selected encoder dimensions.

5513