Linear Discriminative Learning: a competitive non-neural baseline for morphological inflection

Cheonkam Jeong[†]
Department of Linguistics
University of Arizona
Tucson, AZ, 85721, United States

Dominic Schmitz[†] and **Akhilesh Kakolu Ramarao**[†]
Department of English Language and Linguistics
Institute of English and American Studies
Heinrich-Heine-University
Düsseldorf, 40225, Germany

Anna Sophia Stein
Institute of Linguistics
Heinrich-Heine-University
Düsseldorf, 40225, Germany

Kevin Tang[✍]
Department of English Language and Linguistics
Institute of English and American Studies
Heinrich-Heine-University
Düsseldorf, 40225, Germany

Abstract

This paper presents our submission to the SIGMORPHON 2023 task 2 of Cognitively Plausible Morphophonological Generalization in Korean. We implemented both Linear Discriminative Learning and Transformer models and found that the Linear Discriminative Learning model trained on a combination of corpus and experimental data showed the best performance with the overall accuracy of around 83%. We found that the best model must be trained on both corpus data and the experimental data of one particular participant. Our examination of speaker-variability and speaker-specific information did not explain why a particular participant combined well with the corpus data. We recommend Linear Discriminative Learning models as a future non-neural baseline system, owning to its training speed, accuracy, model interpretability and cognitive plausibility. In order to improve the model performance, we suggest using bigger data and/or performing data augmentation and incorporating speaker- and item-specifics considerably.

1 Introduction

There has been a heated debate on whether human language users generate language in a generative manner (e.g., Chomsky and Halle (1968)) or an output-oriented manner (e.g., Prince and Smolensky (2004)). In accordance with the theoretical stance, computational models have been proposed. The generative approach is essentially rule-based at an abstract level. The Minimal Generalisation Learner (MGL) by Albright and Hayes (2003) is a traditional, symbolic rule learner. More recent rule-based computational approaches include Allen and Becker (2015) and Belth et al. (2021).

With the availability of large corpus data, output-oriented models have become widely popular. Output-oriented models can be rule-based or end-to-end. The former includes Prince and Smolensky (2004) and Lignos et al. (2009); in the former, a search procedure is implemented on a set of candidates, or outputs, in order to find the surface form that is most compatible with the underlying representation. On the other hand, output-oriented models can be rule-free. For instance, Malouf (2017) showcased a recurrent deep learning model to predict paradigm forms. Similarly, Kirov and Cotterell (2018) proposed a encoder-decoder network architecture to model linguistic phenomena.

Both approaches have their own advantages and disadvantages. However, in terms of model performance, output-oriented models surpass generative ones presumably due to difficulty of incorporating many other variations in conversation. Nevertheless, output-oriented models are not panacea. They are not as cognitively motivated, thus making them less appealing for cognitive research. Deep learning-based models, in particular, show great performance, but they lack interpretability. These shortcomings necessitate a hybrid model, which is i) cognitively motivated, ii) more agnostic than generative models, and iii) more transparent than deep learning-based models.

Recently, Baayen et al. (2019) proposed Linear Discriminative Learning (LDL), part of the discriminative lexicon (Chuang and Baayen, 2021). As the model follows the Rescorla-Wagner Rule and Widrow-Hoff Rules, some insight into human cognition can be obtained through model imple-

[†] These authors contributed equally to this work.

[✍] Senior and corresponding author: Kevin Tang (kevin.tang@hhu.de)

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 138–150
July 14, 2023 ©2023 Association for Computational Linguistics

mentation. Moreover, as it implements a linear mapping between form and meaning in simple two layers without any hidden layer, LDL features higher interpretability and embraces linguistic engineering. Considering these advantages, an LDL model was chosen as the main model for the SIGMORPHON 2023 shared task 2, which aims to generalize morphological inflections in Korean. A transformer model, which has been state-of-the-art models for various NLP tasks, was also implemented for comparison. The code and data are available here: `https://github.com/hhuslamlab/sigmorphon2023`

2 Related Work

While neural-based systems typically dominate SIGMORPHON challenges, perhaps because they generally perform well in morphological inflection tasks, their limitations can be better examined using wug testing. For instance, McCurdy et al. (2020) examined the ability of modern Encoder-Decoder (ED) architectures to inflect German plurals and concluded that ED does not show human-like variability as shown in wug data. In fact, recent SIGMORPHON challenges involve learning from corpora that better represent the actual linguistic input of children (such as child-directed speech) and evaluating on phonetically-transcribed spoken production by children or adults in corpora or in experiments. For instance, the SIGMORPHON 2021 shared Task 0 Part 2 was to predict the judgement ratings of wug words (Calderone et al., 2021) as opposed to using real words held-out from the training data as test data. Similarly, the SIGMORPHON 2022 challenge involved computational modeling of the data drawn from corpora of child-directed speech and evaluation on children's learning trajectories and erroneous productions (Kodner and Khalifa, 2022; Kakolu Ramarao et al., 2022).

Turning to studies from the field of laboratory phonology, there is a long history of training models on corpora to learn specific aspects of morpho-phonological grammar and evaluating their productivity with experimental data (wug-test and acceptability judgement). For instance, Jun (2010)'s study on stem-final obstruent variation in Korean trained a model with multiple stocastic rules (Albright and Hayes (2002)'s Paradigm Learning Model) in which the acquisition of morphology is based on the distributional pattern of the learning data, using the Sejong corpus, and evaluated on acceptability judgement data. Related to the linguistic phenomenon in this study, Albright and Kang (2009) conducted a computational modeling of inflected forms of Korean verbs using the Minimal Generalization Learner algorithm (Albright and Hayes, 2002) and evaluated the model's performance with attested child errors and historical changes.

Finally, there is a growing number of morphological inflection studies that use the Linear Discriminative Learning model (which will be introduced later) to train on corpus data and evaluate on experimental data (Nieder et al., 2021; Heitmeier et al., 2021; Chuang et al., 2020; Heitmeier and Baayen, 2020; Baayen et al., 2018) and they typical yielded relatively high performance (compared to traditional, symbolic rule learner) while being easy to interpret and cognitively motivated.

3 Task and Evaluation Details

We challenge the shared task 2 Cognitively Plausible Morphophonological Generalization in Korean. The aim of this task is to predict human responses to a generalization task (wug-test), considering high-frequency, low-frequency, and pseudoword items. This implies that human responses may vary depending on the word frequency and familiarity.

The phonological phenomenon to be tested through this task is Korean Post-Obstruent Tensification (henceforth, POT). In Korean, when a lenis consonant in the coda is followed by another obstruent, it can be tensified. However, when a consonant cluster occurs in the coda position, it undergoes Consonant Simplification (henceforth, CS) before POT if the following segment is either an obstruent or a sonorant. On the other hand, neither CS nor POT does occur when the following segment is a vowel.

Depending on the type of the deleted consonant, variation can occur, which can be affected by such speaker- and item-specific features as language familiarity and frequency. In this regard, the main task can be rephrased as predicting features as completely as possible in accordance with those in the answers. For this task, both corpus and experimental data are provided, which include variation patterns. Models are to be evaluated on the accuracy of the prediction of the feature vectors given the corresponding features in the answers from unseen participants.

4 Data

Both corpus and experimental data are provided for this task. The National institute of Korean Language (NIKL) Korean Dialogue corpus (NIKL, 2021) is provided as the main corpus data. All the word tokens affixed with `-lC verbs`, except for `-lh` final stems, are provided after manual annotation by the organizers. They further are categorized as `lC+Obstruent`, `lC+Sonorant`, and `lC+Vowel`, depending on the type of the consequent segment. `lC+Obstruent` and `lC+Sonorant` data each include target words with morphological boundaries and produced words with syllable breaks both in a Romanized form and Korean orthography. Whether target words undergo POT is also provided, as well as such features as obstruent deletion and lateral deletion pertinent to CS and POT.

On the other hand, `lC+Vowel` data only provide target words with morphological boundaries both in a Romanized form and Korean orthography. They do not include produced word information as this condition is not subject to POT and CS; thus, all the feature values of obstruent deletion and lateral deletion are labeled as 0 with the POT value being labeled as 0. Whether the target is lateralizated or nasalized is marked only in the `lC+Sonorant` data while labeled as `NA` in the others. The number of tokens in `lC+Obstruent` data is 876, that of tokens in `lC+Sonorant` data is 95, and that of tokens in `lC+Vowel` data is 2,525 with 514 types–1,485 if frequency information in `lC+Vowel` data is ignored. Thus, the total number of tokens in the NIKL data is 3,496.

In addition to the adult corpus, some part of a child spontaneous speech corpus, the Ko corpus (Ko et al., 2020), is provided. As with the NIKL data, the Ko corpus provides target words and their production but solely in a Romanized form. The POT value with the lateralization and nasalization feature values are labeled. As it does not include any lC verbs followed by a sonorant, both nasalization and lateralization are not applicable. A total of 336 tokens are provided.

In the case of experimental data, the experimental responses of 12 participants, in addition to 4 participants for the development data, are provided. They include what are included in the corpus data with the experimental specifics: the trial number, the trial ID, the subject ID, option, language familiarity, and word frequency. A total of 2,843 tokens are provided.

5 System Description

The main task is to accommodate as many variation patterns as possible. We navigated through all the corpus and experimental data, in which process we found inconsistencies in transcription in the data. In particular, the Ko corpus includes items transcribed in a very detailed manner, including phonological processes, such as deletion and insertion, other than those pertinent to the task. We also found that the target item does not always feature one-to-one mapping, but more-than-one mapping.

Based on the observations above, data preprocessing was of primary importance and was the most time-consuming component. We manually corrected the Ko corpus and automatically unified the transcription style. We then selected models adequate to this task. Considering the nature of the shared task that investigates morphological variations with both corpus and experimental data and the time constraint, LDL was chosen as the main model, along with a Transformer model that has demonstrated great performance in NLP. For each of the two modelling approaches we conducted two studies: Study 1 experimented with systems that train only on the corpus data and/or only on the experimental data; Study 2 used the insights from Study 1, and experimented with systems that train on both the corpus and the experimental data.

5.1 Linear Discriminative Model

LDL generates a system of form-meaning relations by discriminating between different forms and meanings, with forms and meanings being represented by numerical vectors. Form vectors are based either on segmental representations of various lengths, or on representations of acoustic transitions gleaned directly from the speech signal (Arnold et al., 2017; Shafaei-Bajestan et al., 2021). Meaning itself is taken to be a dynamic concept, being emergent from the context in which words are being used, and is represented by semantic vectors, similar to approaches in distributional semantics (Boleda, 2020). The idea is that if both forms and meanings can be expressed numerically, we can mathematically connect form and meaning, i.e. map meaning onto form and vice versa. In this system of mappings, the two sets of vectors are combined into matrices – a form matrix and a semantic matrix. The form vectors are mapped onto semantic vectors to model comprehension, and semantic vectors are mapped onto form vectors to model production. The mapping between them at the theoretical end-state of learning is predicted using multivariate multiple linear regression (hence the term 'Linear Discriminative Learning'). The network is simple and interpretable, because, in contrast to deep learning networks, it features just two layers (i.e. the form and meaning matrices), both of which are linguistically transparent.

5.1.1 Data Preprocessing

Due to the time limitation, we decided to winnow out the data that are only pertinent to POT and CS from the Ko corpus. To be specific, the Ko corpus includes tokens involving other phonological processes, like insertion, as well. For instance, there are 6 instances of `ipko` and the produced forms are `ikgo`, `lipgo`, `tipgo`, `ikgu`, `linkgo`, and `nipgo`. Considering POS and CS rules, the ideal outputs are `ipgo` and `ikgo`. Moreover, based on the tendency of negative vowelization of the mid rounded vowel in conversation, `ikgu` is another candidate. The others are also producible, especially considering that the Ko corpus consists of children speech, but they are definitely not canonical outputs from the pertinent rules. Thus, if the input and the output are hugely different from each other because of other phonological processes, the tokens were discarded. All the duplicated tokens were also removed. Thus, 286 tokens were left from 336 tokens. Lastly, morphological boundary and feature representation were manually incorporated following the style of the other data.

We also observed that there are inconsistencies in transcription style between the two corpus data and the experimental data. The following data transformations were performed on the corpus data. In the Production_R, the tense forms of the plosives `p*,t*,k*` are replaced with `b,d,g`, and those of the alveolar fricative and the alveolo-palatal affricate `S,c*` are replaced with `s*, J`, respectively. Moreover, there are several inconsistencies in transcription style between the input (Morphology_R) and the output (Production_R). First, the middle yin diphthong `yv` or `jv` in Morphology_R is replaced with `jv` or `yv` in Production_R. Second, the alveolar fricative `S` in Morphology_R is replaced with `s*`, but the reverse transformation is conducted in Production_R. Lastly, the tense stops in Korean orthography `P, T, K` in Morphology_R are replaced with `p, t, k`, except when they occur in the word-initial position.

As a result, the pre-processed data contained `s*` as phone representation. That is, a single phone is represented by two symbols. For triphones, this would lead to unwanted consequences: Triphones which contain information only on two phones, i.e. `s*` and any other phone, and triphones which contain only one of the two parts of `s*`. Therefore, `s*` was replaced with `S` for the implementations of LDL presented in the subsequent sections.

5.1.2 General Model Architecture

The form matrices C used for the present implementations of LDL consisted of triphones, i.e. sequences of three phones within a word form. Triphones overlap and can be understood as proxies for phonological transitions. In each word's individual form vector c, the presence of a triphone is marked with 1, while the absence is marked with 0. The form vectors of all words of a set of words constitute its C matrix and each row in such a C matrix represents a word form, while the columns of the C matrix represent all triphones of its underlying word set. Triphones were used as previous studies found overall good performance for triphones (e.g. Chuang et al. 2021; Schmitz et al. 2021).

The semantic matrices S used for the present implementations of LDL deviate from those usually found in studies using LDL. Commonly, semantics are introduced via semantic vectors obtained by methods of distributional semantics, e.g. via fastText (Bojanowski et al., 2016) or naive discriminative learning (Baayen et al., 2011). However, with the small amount of language data provided, the computation of such semantic vectors is barely feasible. While creating semantic vectors based on a larger corpus of Korean may be one option to solve this issue, we decided against this solution as it would mean using data that is not part of the current challenge. Instead, semantic vectors were created based on morphemes and in a binary fashion. That is, similar to the form vectors, in each word's individual semantic vector s, the presence of a morpheme is marked with 1, while the absence is marked with 0. The semantic vectors of all words of a set of words constitute its S matrix.

With C and S, one can straightforwardly map forms onto meanings and meanings onto forms:

$$CF = S$$

$$SG = C$$

If one wants to predict the forms or semantics for words that are not yet part of the implementation, additional steps are required. Predicting semantics for newly introduced forms, one computes with C' denoting the Moore-Penrose generalised inverse. Using the transformation matrix F and a combined form matrix for previously and newly introduced forms $C_{combined}$, then

$$S = C_{combined}F$$

Using this method, previous studies have analysed the semantics of pseudowords (Cassani et al., 2020; Chuang et al., 2021; Schmitz et al., 2021). Adapting this method for the prediction of forms, as for the present study, one computes

$$G = S'C$$

Then, using the transformation matrix G and a combined semantic matrix for previously and newly introduced words, the following is solved:

$$C = S_{combined}G$$

Note that this method comes with an important caveat: Newly introduced words must not contain any triphones that are not part of the original C matrix when predicting their meaning, and, in the present case, they must not contain any morphemes that are not part of the original S matrix.

5.1.3 Study 1

For a first implementation of LDL, the following rationale was adopted. First, the combined data of the NIKL and Ko corpora were taken to represent the mental lexicon of a speaker of Korean. That is, we assumed that this knowledge is shared by all participants. Second, based on this shared prior knowledge, participants individually produced word forms during the experiment. Predicting these forms, and in turn the phonological processes underlying them, via prior knowledge was the aim of this implementation.

The combined NIKL and Ko corpus data were used as initial word set ($n = 632$ after duplicate removal). Based on the corpus data, C and S matrices were created following the specifications in Section 5.1.2. After obtaining the required transformation matrix G via $G = S'C$, G was based on the triphone to morpheme relations found in the corpus data it was trained on. In a next step, one would then use G to compute $C = S_{combined}G$. However, the experimental data contained 111 triphones (out of 247) that were not part of the corpus data. As G was not trained to predict these triphones, any further computations were rendered meaningless.

5.1.4 Study 2

Instead, a second LDL network was implemented. The rationale of this implementation was to first create individual networks for all sixteen train and dev participants. Each participant's network was trained on the combined corpus data and on their experimental data. In a second step, each of the sixteen participants and their networks were then used to predict all other participants' produced word forms. This provided insight in how far pertinent participants were able to predict other participant's productions, allowing the selection of a 'best' participant to then predict the test participants' produced word forms.

First, for each of the sixteen train and dev participants a data set containing the combined NIKL and Ko corpus data ($n = 632$ after duplicate removal) as well as their experimental data was created ($n = 175$ to $n = 180$). Based on this data set, C and S matrices were created and comprehension as well as production were modeled following the specifications outlined in Section 5.1.2.

Second, each of the sixteen participants' G matrices was used to predict the forms produced by all other train and dev participants in the experiment. In contrast to Section 5.1.3, this computation did not pose a problem as experiment items and their triphones were already introduced during the first step. As a result, we obtained prediction accuracies for all sixteen participants by all sixteen participants. Accuracy here refers to whether a word form was predicted correctly. The overall and individual accuracies for low-, high-frequency, and pseudoword items are available on GitHub.

Across all sixteen train and dev participants, it was found that participant 597515 clearly outperformed the other fifteen participants in terms of prediction accuracy across all experimental items. Their mean prediction accuracy across all experimental items was 73%, with 76% for low frequency, 71% for high frequency, and 73% for pseudoword items. Their overall Precision, Recall, and F1 scores for the training data are given in Table 1.

In an attempt to understand why this particular participant showed the best prediction results for the other fifteen train and dev participants and to find out whether we could determine differ-

	Precision	Recall	F1
simplify_delete_obstruent	0.48	0.57	0.43
simplify_delete_lateral	0.60	0.69	0.60
nasalization	0.60	0.69	0.60
lateralization	0.72	0.58	0.46
tensification	0.64	0.77	0.67

Table 1: Precision, Recall, and F1 of participant 597515 for the five phonological processes in the training data

ent 'best' participants for different participants to be predicted, we implemented three multiple regression models for each of the sixteen train and dev participants, i.e. one for high frequency, one for low frequency, and one for pseudoword items. For a given participant's multiple regression models, the dependent variable was the set of prediction accuracies reached by the other participants for that participant. As predictors, the biographical background information, LANGUAGEPREFERENCE and AGESTARTEDSPEAKING, were introduced. Across the sixteen low frequency item models, we found that one participant with a LANGUAGEPREFERENCE of 3 showed an effect for LANGUAGEPREFERENCE ($p = 0.02$). This presumably indicated that the other participant with a LANGUAGEPREFERENCE of 3 was the 'best' prediction candidate for this participant. Across the sixteen high frequency item models, we found that both participants with a LANGUAGEPREFERENCE of 3 showed an effect for LANGUAGEPREFERENCE ($p = 0.02$; $p = 0.0002$), indicating that they were each other's best prediction candidates. Another participant showed a barely significant effect of LANGUAGEPREFERENCE ($p = 0.046$), and yet another participant showed a significant effect of AGESTARTEDSPEAKING ($p = 0.03$) . Across the pseudoword item models, no effects were found. As these results were inconclusive, we decided to drop this attempt and to use participant 597515's G matrix to predict the forms, and hence the underlying phonological processes, for the seven test participants.

The predicted forms and their underlying representations were used to derive information on which of the five phonological processes of interest were predicted for a pertinent word form.

5.2 Neural Network

5.2.1 Data Preprocessing

See the data preprocessing steps in Section 5.1.1.

5.2.2 Model Architecture

Our model closely follows the formulation of the encoder-decoder Transformer for character-level transduction model proposed by Wu et al. (2021). We use multi-headed Transformers with self-attention and implement them with Fairseq (Ott et al., 2019) tool, a PyTorch-based sequence modeling toolkit. Both Encoder and Decoder have four layers with four attention heads, an embedding size of 256 and hidden layer size of 1,024. We use Adam Optimizer (Kingma and Ba, 2015), with an initial learning rate of 0.001, a batch size of 400, 0.1 label smoothing and 1.0 gradient clip threshold. Models are trained for a maximum of 3,000 optimizer updates. Checkpoints are saved every 10 epochs. Beam search is used at the decoding time with a beam width of 5.

The checkpoint with the smallest loss on the development data is chosen as the best model.

For the evaluation, we consider the models' *sequence accuracy* (henceforth, *accuracy*), where only instances for which the entire output sequence equals the target are considered correct.

5.2.3 Study 1

The inputs to each model are the individual characters of the romanized. For example, for the model trained against the raw NIKL dataset, the input is `J a l p - k o` and the output is `J a l . g o`

Model Training Three models were trained on i) the raw NIKL dataset (with a total of 1485 tokens), ii) the raw Ko corpus (with a total of, and iii) the combined datasets (NIKL and Ko). The data in each model were split into train (70%), dev (10%), test (20%) sets. The sequence accuracies of the three models are 71.7% (raw NIKL)[1], 35.3% (raw Ko) and 65.5% (the combined dataset). Furthermore, we trained a model on all experimental items following the same train-test split as stated above and the accuracy was found to be 69.4%.

While these models were evaluated on a different set of test data, their accuracies can nonetheless suggest how the different datasets should be used in Study 2 (Section 5.2.4). Training on the Ko corpus alone is unlikely to be sufficient as it yielded the lowest accuracy. While combining NIKL with Ko yielded a poorer model compared to just using

[1] We experimented with a model using the NIKL dataset but without syllable boundaries, and it yielded an accuracy of 71.6% – a negligible difference compared to the model with syllable boundaries 71.7%

NIKL alone, the Ko corpus should not be excluded given that it is arguably more ecologically valid than NIKL and the amount of training data is already small in this challenge. Finally, training only on experimental items resulted in a comparable performance as the combined dataset. This model was not used as it was explicitly discouraged by the challenge.

To determine how well a model trained only on corpus data would perform on the experimental data, we evaluated the best model (trained on raw NIKL) against the experimental data (and removed the syllable boundaries in the predictions to match the transcription style of the experimental data) and it yielded a much lower accuracy of 29.9%, suggesting that we should incorporate the experimental data as part of training.

5.2.4 Study 2

In this study, we primarily used the pre-processed dataset (using methods described in section 5.2.1) that consists of both NIKL and child spontaneous speech dataset. We then incorporate parts of experimental data along with the combined dataset during both training and development phase. The test data provided by the organizers is used during the testing phase.

Model Training We first incorporated models with training the combined dataset using i) productions from best participant and ii) productions from worst participant, as development data , that yielded accuracy scores of 43.8% and 39.4%.

Next, we trained a model on the combined dataset (NIKL, and the Ko corpus) with the productions from 4 best participants and using responses from a random participant as development data which produced an accuracy score of 68.1%.

Finally, a model was trained on all participants' except the best participant's responses with the combined dataset (NIKL, and the Ko corpus) and using the productions from best participant as development data that yielded an accuracy of 69.2%. The accuracies of this model for: i) low-frequency ii) high-frequency and iii) pseudoword items are 64.4%, 77.7% and 65.38% respectively. The overall Precision, Recall and F1 scores for the five phonological processes in the test data are given in Appendix A.

	Precision	Recall	F1
simplify_delete_obstruent	0.69	0.70	0.67
simplify_delete_lateral	0.79	0.75	0.75
nasalization	0.79	0.75	0.75
lateralization	0.44	0.53	0.41
tensification	0.98	0.98	0.98

Table 2: Precision, Recall, and F1 of participant 597515 for the five phonological processes in the test data

5.3 Results

Predicting the seven test participants' productions using the 'best' participant's LDL network as detailed in Section 5.1.4, an overall accuracy of 83.32% was reached. The accuracies of this model for: i) low-frequency ii) high-frequency and iii) pseudoword items are 83.56%, 83.58% and 82.84%. The overall Precision, Recall, and F1 scores for the test data are given in Table 2. The model performed best on tensification (F1: 0.98), and worst on lateralisation (F1: 0.41). The mean perplexity scores for the train and dev as well as for the test data are 2.11 and 1.97 respectively. The performance of the model on the test data is similar to that on the training data (Table 1) with the exception of simplify delete obstruent being better predicted than lateralization in the test data. Comparing to the best Transformer model, LDL performed better in terms of the overall accuracies of the model; however, the relative performances of the five phonological processes (Precision, Recall and F1 scores) are largely the same (Appendix A).

6 Variability: Items and Participants

To examine the variability of the phenomenon, Shannon entropy (base 2) (Shannon, 1948) was used to quantify how variable are the items in the experimental data and how variable are the participants. In this study, we considered sixteen possible combinations of the five phonological processes (therefore sixteen events in entropy's term) (See Appendix C). With sixteen combinations, the highest possible entropy value is 4 which means each combination has a probability of 1/16 indicating a high level of variability, and the lowest possible entropy is 0 which means there is only one attested combination indicating no variability. For detailed analyses, see Appendix B.

First, we computed the by-item entropy values by computing the proportion of the sixteen response combinations using the sixteen participants (training and development). The 180 experimental items have a mean entropy of 0.584. Pseudoword items have the highest mean entropy (0.612), followed by high-frequency items (0.596) and low-frequency items (0.544). These entropy values suggest that the experimental items in general have low variability and unsurprisingly the pseudoword items were particularly variable compared to the real words. However, these differences in entropy values were not statistically significant ($ps > 0.3842$).

Second, we computed by-participant entropy values by computing the proportion of the sixteen response combinations. Across all the experimental items, the sixteen participants have a mean entropy of 2.143. Participants show the lowest mean entropy with the high-frequency items (2.049) , followed by low-frequency items (2.111) and pseudoword items (2.112). However, these differences in entropy values were not statistically significant ($ps > 0.2542$). Our 'best' participant 597515 has an entropy of 2.192 across all items, 2.176 for high-frequency, 2.154 for low-frequency and 2.140 for pseudoword items, with all values similar to their means. Therefore, the participant's superiority is not purely due to their responses being more variable.

7 Discussion and Conclusion

We demonstrated that LDL is capable of modelling morphological inflection trained on limited corpus and experimental data. Its performance is competitive to that of the Transformer model that we experimented with. Past SIGMORPHON shared tasks (2017–2022) with a focus on morphological inflection have generally received more neural-based systems than non-neural ones and found that neural-based ones tend to be superior (Kodner and Khalifa, 2022; Kodner et al., 2022; Pimentel et al., 2021; Vylomova et al., 2020; McCarthy et al., 2019; Cotterell et al., 2018, 2017, 2016). Amongst the submitted non-neural systems, LDL has never been utilized. Our study cannot conclude that LDL is superior to the transformer architecture as the latter was not fully optimized. However, it has great potential to serve as a non-neural baseline system for future shared tasks as well as allowing researchers to conduct rapid experiments, because of its architecture simplicity, performance (with accuracies from 59% to 99% in a range of languages, e.g. Heitmeier et al. 2021; Schmitz et al. 2021; Stein and Plag 2021; Chuang et al. 2020; Baayen et al. 2019) and speed (in our study one model required on average 35 seconds of CPU processing on an i7-9750H 2.60 GHz system with 32 GB memory).

Our study found that training on the corpus data alone was insufficient and that our models require at least one participant's experimental data in order to inflect the experimental items well. However, from an ecological perspective, a model should only be trained on the corpus data (NIKL and Ko), excluding the experimental data, as the corpus data serve to represent the participants' actual linguistic input. The corpus data we have are likely unrepresentative of the actual linguistic input of our participants. Firstly, the verbs were not embedded in an utterance, and even if the full utterances were used the overall amount of data would still be small with only 53,000 words from the Ko corpus, and 900,000 phrases from the NIKL corpus. Based on spoken speech input alone, Brysbaert et al. (2016) estimated that for American English, the total input from social interactions (in a dialogue) would be equal to 11.688 million word tokens per year and a 20-year-old would have been exposed to about 234 million word tokens. Using a much larger speech-like or transcribed corpus such as SUBTLEX-KR (Tang and de Chene, 2014) (90 million eojeols) is a promising approach for examining morphological inflection patterns (see de Chene (2014) on regularisation in Korean noun inflection).

Our item variability analyses suggest that the three item types (high-, low-frequency and pseudowords) are not particularly different in their variability. This might be reflecting how the LDL model reported in Section 5.3 performed similarly with them (high: 83.58%, low: 83.56%, and pseudowords: 82.84%). However, the best Transfomer model was sensitive to item types yielding a higher accuracy for the high-frequency items (77.7%), than the low-frequency (64.4%) and pseudoword (65.38%) items.

Our attempts in understanding why the 'best' participant was the best in predicting individual participants' productions were not successful. Response variability was unable to explain why our 'best' participant was the best, as it has neither high nor low in variability compared to the other 15 participants. Our regression analyses predicting individual participant accuracies using the participants' demographics was inconclusive. While one may assume that the LDL prediction results should improve when one predicts speakers of similar backgrounds, the nonetheless satisfying LDL prediction results suggest that demographics-matching was not needed. Overall, our results suggest that LDL is suitable for tasks such as the one at hand.

Limitations

The small amount of training data provided in this shared task poses a challenge for models that need large amounts of data to reliably learn linguistic patterns. While we did not employ any data augmentation techniques, we suggest future work to train the models on all the possible feature combinations (weighted with the probabilities as the experimental data) for the stems in the two corpora.

Owning to the lack of time and computing resources, we did not fully optimize our transformer models and we did not fully utilize and explore i) speaker-specific information, especially for the transformer models, ii) token frequency information in the corpora, as we assumed extension of morphological patterns is based on type, not token, frequency (Bybee, 2001; Pierrehumbert, 2001). Furthermore, we did not experiment with training models with either high-frequency, low-frequency and pseudoword items. It is possible that some speakers' high/low/pseudoword items would be better served as part of the training set.

The LDL model in Section 5.1.3 was not able to evaluate the experimental items due to the unattested triphones. This shortcoming can be mitigated by using phonological features (Tang and Baer-Henney, 2023).

Ethics Statement

In the beginning of the challenge, we discovered the answers of the experimental test data were accidentally released early by the organisers. We immediately informed the organisers and as requested, we deleted the test data and committed to not use it until it was officially released. Our work was trained on speech corpora of adults which were recorded with ethics approval. The broader impact of the work includes i) improving how morphological inflection models can be trained with low-resource languages or phenomena, ii) developing speaker-specific morphological inflection models, iii) establishing a new baseline model architecture (LDL) that has a low carbon footprint.

CRediT authorship contribution statement

CJ, DS, and AKR contributed equally to this work. KT served as the senior and corresponding author on this paper.

We follow the CRediT taxonomy[2]. Conceptualization: KT; Data curation: CJ, AKR; Formal Analysis: AS, DS; Investigation: KT, DS, CJ, AKR; Methodology: KT; Supervision: KT; Visualization: AS; and Writing – original draft: KT and Writing – review & editing: KT, AS, CJ, DS, AKR.

Acknowledgements

We gratefully acknowledge the support of the central HPC system "HILBERT" at Heinrich-Heine-University, Düsseldorf.

References

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, page 58–69, USA. Association for Computational Linguistics.

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

Adam Albright and Yoonjung Kang. 2009. Predicting innovative alternations in Korean verb paradigms. *Current issues in unity and diversity of languages: Collection of the papers selected from the CIL 18, held at Korea University in Seoul*, pages 1–20.

Blake Allen and Michael Becker. 2015. Learning alternations from surface forms with sublexical phonology. *Unpublished manuscript. Available as lingbuzz/002503.*

Denis Arnold, Fabian Tomaschek, Konstantin Sering, Florence Lopez, and R. Harald Baayen. 2017. Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, 12:e0174623.

R. Harald Baayen, Yu-Ying Chuang, and James P. Blevins. 2018. Inflectional morphology with linear mappings. *The Mental Lexicon*, 13(2):230–268.

R. Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins. 2019. The discriminative lexicon: A unified computational model for

[2] https://www.ucl.ac.uk/library/research-support/open-access/credit-taxonomy

the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019:1–39.

R. Harald Baayen, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–481.

Caleb Belth, Sarah Payne, Deniz Beser, Jordan Kodner, and Charles Yang. 2021. The greedy and recursive search for morphological productivity.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, 7:1116.

Joan Bybee. 2001. *Phonology and language use*. Cambridge University Press, Cambridge.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 274–282, Online. ACL.

Giovanni Cassani, Yu-Ying Chuang, and R Harald Baayen. 2020. On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(4):621.

Noam Chomsky and Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Yu-Ying Chuang and R. Harald Baayen. 2021. Discriminative learning and the lexicon: NDL and LDL. *Oxford Research Encyclopedia of Linguistics*.

Yu-Ying Chuang, Kaidi Lõo, James P. Blevins, and R. Harald Baayen. 2020. Estonian case inflection made simple: A case study in word and paradigm morphology with linear discriminative learning. In Lívia Körtvélyessy and Pavol Štekauer, editors, *Complex Words: Advances in Morphology*, page 119–141. Cambridge University Press.

Yu-Ying Chuang, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix, and R Harald Baayen. 2021. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior research methods*, 53:945–976.

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

Brent de Chene. 2014. Probability matching versus probability maximization in morphophonology: The case of Korean noun inflection. *Theoretical and applied linguistics at Kobe Shoin*, 17:1–13.

Maria Heitmeier and R. Harald Baayen. 2020. Simulating phonological and semantic impairment of english tense inflection with linear discriminative learning. *The Mental Lexicon*, 15(3):385–421.

Maria Heitmeier, Yu-Ying Chuang, and R. Harald Baayen. 2021. Modeling morphology with linear discriminative learning: Considerations and design choices. *Frontiers in Psychology*, 12.

Jongho Jun. 2010. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics*, 19:137–179.

Akhilesh Kakolu Ramarao, Yulia Zinova, Kevin Tang, and Ruben van de Vijver. 2022. HeiMorph at SIGMORPHON 2022 shared task on morphological acquisition trajectories. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 236–239, Seattle, Washington. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Eon-Suk Ko, Jinyoung Jo, Kyung-Woon On, and Byoung-Tak Zhang. 2020. Introducing the Ko corpus of Korean mother–child interaction. *Frontiers in Psychology*, 11:602623.

Jordan Kodner and Salam Khalifa. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 157–175, Seattle, Washington. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Constantine Lignos, Erwin Chan, Mitchell P. Marcus, and Charles Yang. 2009. A rule-based unsupervised morphology learning framework. In *Working Notes for CLEF 2009 Workshop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009*, volume 1175 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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

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.

Kate McCurdy, Adam Lopez, and Sharon Goldwater. 2020. Conditioning, but on which distribution? grammatical gender in German plural inflection. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 59–65, Online. Association for Computational Linguistics.

Jessica Nieder, Yu-Ying Chuang, Ruben van de Vijver, and R. H Baayen. 2021. A discriminative lexicon approach to word comprehension, production and processing: Maltese plurals.

NIKL. 2021. NIKL Korean dialogue corpus (audio) 2020(v.1.3).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Janet Pierrehumbert. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee and Paul Hopper, editors, *Frequency and the emergence of linguistic structure*, pages 137–157. John Benjamins, Amsterdam.

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

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Dominic Schmitz, Ingo Plag, Dinah Baer-Henney, and Simon David Stein. 2021. Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, 12.

Elnaz Shafaei-Bajestan, Masoumeh Moradipour-Tari, Peter Uhrig, and R. Harald Baayen. 2021. Ldl-auris: a computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience*, pages 1–28.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Simon David Stein and Ingo Plag. 2021. Morpho-phonetic effects in speech production: Modeling the acoustic duration of english derived words with linear discriminative learning. *Frontiers in Psychology*, 12.

Kevin Tang and Dinah Baer-Henney. 2023. Modelling L1 and the artificial language during artificial language learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1):1–54.

Kevin Tang and Brent de Chene. 2014. A new corpus of colloquial Korean and its applications. Presented at The 14th Laboratory Phonology Conference (LabPhon 14), Tachikawa, Tokyo, Japan.

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. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1901–1907. Association for Computational Linguistics.

A Appendix: Evaluation metrics for the Neural Network model

	Precision	Recall	F1
simplify_delete_obstruent	0.70	0.71	0.68
simplify_delete_lateral	0.79	0.72	0.72
nasalization	0.79	0.72	0.72
lateralization	0.44	0.53	0.41
tensification	0.98	0.98	0.98

Table 3: Precision, Recall, and F1 for the five phonological processes in the test data for the best performing neural network model

B Appendix: Variability analyses

	mean	sd	min.	max.
All	0.58	0.52	0.00	1.68
High-frequency	0.60	0.54	0.00	1.68
Low-frequency	0.54	0.52	0.00	1.47
Pseudoword	0.61	0.50	0.00	1.65

Table 4: Summary statistics of by-item entropy: Mean, standard deviation (sd), minimum (min.) and maximum (max.) entropy values of all items computed over all items, as well as subsets of items (high-frequency, low-frequency and pseudoword items)

	mean	sd	min.	max.
All	2.14	0.14	1.83	2.33
High-frequency	2.05	0.16	1.67	2.25
Low-frequency	2.11	0.10	1.88	2.25
Pseudoword	2.11	0.15	1.85	2.33

Table 5: Summary statistics of by-participant entropy: Mean, standard deviation (sd), minimum (min.) and maximum (max.) entropy values of all participants computed over all items, as well as subsets of items (high-frequency, low-frequency and pseudoword items)

participant	all	pseudoword	low	high
597515	2.19	2.14	2.15	2.18
592117	2.20	2.11	2.14	2.18
563118	2.19	2.13	2.15	2.10
556014	2.26	2.25	2.19	2.21
578085	2.16	2.04	2.23	2.03
559838	2.14	2.14	2.05	2.00
589028	2.03	2.01	2.04	1.99
594939	2.05	1.97	2.07	2.02
581952	2.22	2.25	2.19	2.02
565631	1.89	1.85	1.95	1.79
578698	2.24	2.25	2.19	2.18
556505	2.23	2.25	2.08	2.13
592166	2.26	2.22	2.25	2.18
556033	2.33	2.33	2.15	2.21
585660	1.84	1.87	1.88	1.67
575760	2.04	2.00	2.07	1.91

Table 6: Breakdown of by-participant entropy values: Entropy values for all participants in the experimental dataset (excluding the test set) computed over all items, as well as subsets of items (pseudoword, low-frequency (low) and high-frequency (high) items)

C Appendix: Feature combinations

Tens.	Nasal.	L del.	C del.	Lateral.
0	N/A	0	0	N/A
N/A	0	0	0	0
1	N/A	0	1	N/A
N/A	1	N/A	N/A	N/A
1	N/A	N/A	N/A	N/A
0	N/A	N/A	N/A	N/A
N/A	0	1	0	0
1	N/A	0	0	N/A
N/A	1	1	0	0
N/A	0	0	1	0
N/A	0	N/A	N/A	N/A
1	N/A	1	0	N/A
N/A	1	0	0	0
0	N/A	1	0	N/A
1	N/A	1	1	N/A
0	N/A	0	1	N/A
N/A	0	0	1	1

Table 7: Feature combinations used in the entropy calculation. The features are tensification (Tens.), nasalization (Nasal.), laterial deletion (L del.), obstruent deletion (C del.) and lateralization (Lateral.). The combination '1, N/A, 1, 1, N/A' was excluded as it had only one attestation across the dev and train set.