	Bengali	Hindi	Latin	Spanish
BLSTM-BLSTM	71.06/72.10	87.80/88.18	60.85/61.63	88.06/88.79
BGRU-BGRU	70.44/71.22	88.34/88.40	60.65/61.52	91.48/92.25
Lemming	74.10	90.35	57.19	58.89
Morfette	70.27	88.59	47.41	57.61

Table 5: Lemmatization accuracy (in %) on unseen words.

	Bengali	Hindi	Latin	Spanish
BLSTM-BLSTM	56.16/66.26	87.42/88.41	49.80/56.06	86.22/87.97
BGRU-BGRU	59.45/66.84	87.19/88.26	50.24/55.35	86.74/88.49

Table 6: Lemmatization accuracy (in %) on unseen words without using applicable edit trees in training.

plicable edit trees already learns to dispense the output probability to the legitimate classes over which, output restriction cannot yield much enhancement.

Results for Unseen Word Forms: Next, we discuss about the lemmatization performance on those words which were absent in the training set. Table 4 shows the proportion of unseen forms averaged over 4 folds on the datasets. In Table 5, we present the accuracy obtained by our models and the baselines. For Bengali and Hindi, Lemming produces the best results (74.10% and 90.35%). For Latin and Spanish, BLSTM-BLSTM and BGRU-BGRU obtain the highest accuracy (61.63% and 92.25%) respectively. In Spanish, our model gets the maximum improvement over the baselines. BGRU-BGRU beats Lemming with 33.36% margin (on average, out of 9,011 unseen forms, 3,005 more tokens are correctly lemmatized). Similar to the results in Table 2, the results in Table 5 evidences that restricting the output in applicable classes enhances the lemmatization performance. The maximum accuracy improvements due to the output restriction are: 1.04% in Bengali (from 71.06% to 72.10%), 0.38% in Hindi (from 87.80% to 88.18%) using BLSTM-BLSTM and 0.87% in Latin (from 60.65% to 61.52%), 0.77% in Spanish (from 91.48% to 92.25%) using BGRU-BGRU.

Further, we investigate the performance of our models trained without the applicable edit trees information, on the unseen word forms. The results are given in Table 6. As expected, for every model, the accuracy drops compared to the results shown in Table 5. The only exception that we find out is in the entry for Hindi with BLSTM-BLSTM. Though without restricting the output, the accuracy in Table 5 (87.80%) is higher than the corresponding value in Table 6 (87.42%), but after output restriction, the performance changes (88.18% in Table 5, 88.41% in Table 6) which reveals that only selecting the maximum probable class over the applicable ones would be a better option for the unseen word forms in Hindi.

	Sem. Embedding	Syn. Embedding
Bengali	90.76/91.02	86.61/86.82
Hindi	94.86/94.86	91.24/91.25
Latin	88.90/89.09	85.31/85.49
Spanish	97.95/98	96.07/96.10

Table 7: Results (in %) obtained using semantic and syntactic embeddings separately.

	# Sentences	# Word Tokens
Catalan	14,832	474,069
Dutch	13,050	197,925
Hungarian	1,351	31,584
Italian	13,402	282,611
Romanian	8,795	202,187

Table 8: Dataset statistics of the 5 additional languages.

Effects of Semantic and Syntactic Embeddings in Isolation: To understand the impact of the combined word vectors on the model's performance, we measure the accuracy experimenting with each one of them separately. While using the semantic embedding, only distributional word vectors are used for edit tree classification. On the other hand, to test the effect of the syntactic embedding exclusively, output from the character level recurrent network is fed to the second level BGRNN. We present the results in Table 7. For Bengali and Hindi, experiments are carried out with the BLSTM-BLSTM model as it gives better results for these languages compared to BGRU-BGRU (given in Table 2). Similarly for Latin and Spanish, the results obtained from BGRU-BGRU are reported. From the outcome of these experiments, use of semantic vec-