Data Augmentation for Transformer-based G2P

University of Colorado **Boulder**

Zach Ryan, Mans Hulden
{zachary.j.ryan,mans.hulden}@colorado.edu
https://github.com/LonelyRider-cs/sig_shared_tasks

Overview

- An experiment on data augmentation for low-resource G2P (< 1000 examples)
- Use SIGMORPHON task 1 data
  - SIGMORPHON had no low-resource track so we sample uniformly from the original data to create data sets of 100 & 500 examples
- We use the *Transformer* throughout with the shared task baseline settings
- Use augmentation strategy based on 3 components (see right)
- Significant improvement with 100 examples, 500 examples, tapering off with full data set of 3,600 examples

(1) Align example data with MCMC

- Use an MCMC aligner, similar to EM aligners for 1-1 alignment, but faster
- Something like minimum-edit distance doesn't apply here since input-output alphabets are different
- Once alignments are learned, we extract all substring-pairs (input/output) from the beginning and end of the input-output pair
- We estimate which pairs are "reliable" g2p slices

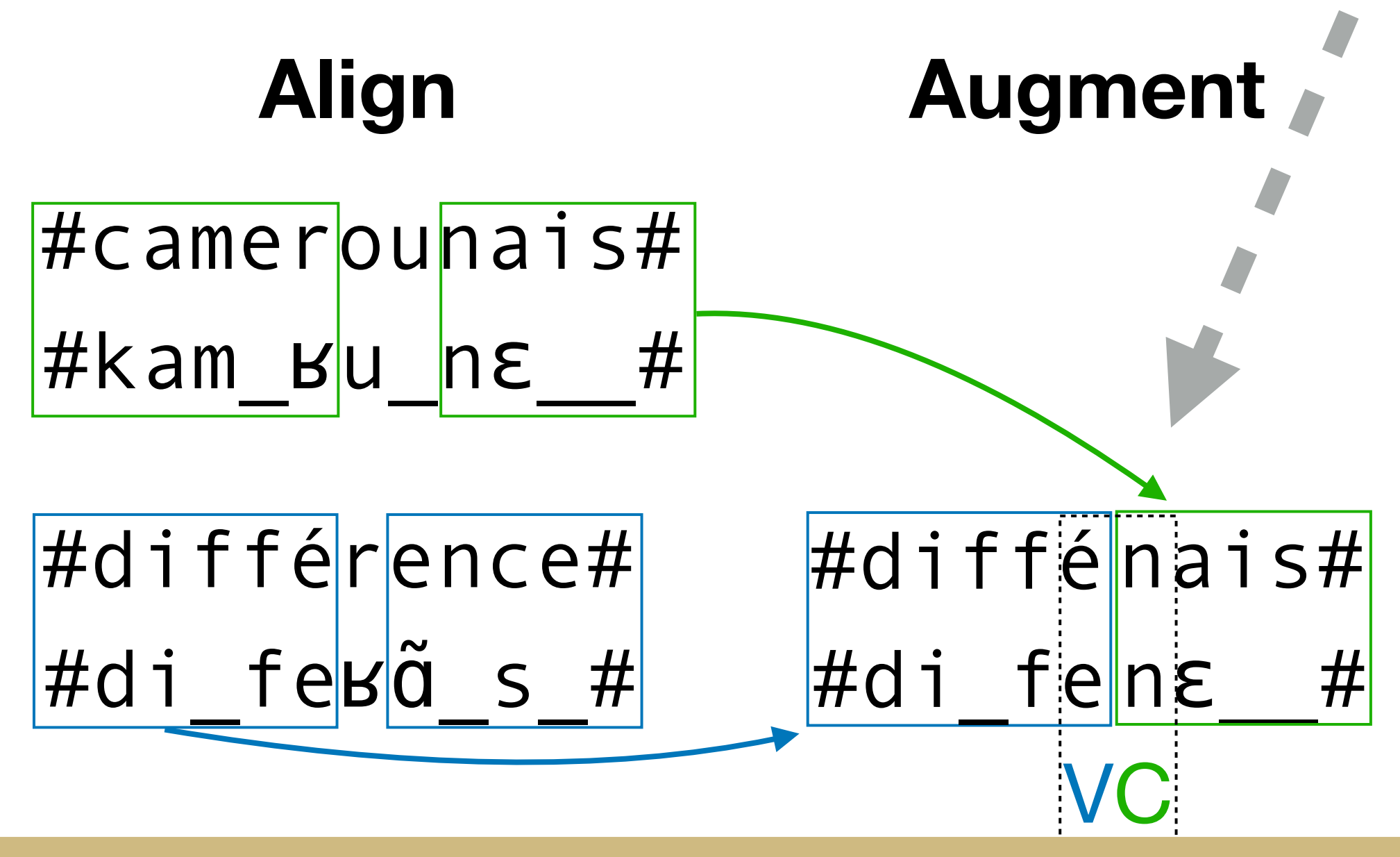
```
ta_xation
taksasjɔ̃_

déclaration
deklaʁasjɔ̃_

ambroisie
_ɑ̃bʁwazi_

commun
kɔmœ̃__

traite
tʁɛ_t_
```

(2) Extract consistent pieces

- Once we have all beginning and ending slices, we estimate the reliability of an i-o slice being consistent by

$$p(o|i) = \frac{\text{count}(i:o) + \alpha}{\sum_{\text{ANY}} \text{count}(i:\text{ANY}) + \alpha|\text{ANY}|}$$

- if p(oli) > our cutoff (0.98) we use the slice to create hallucinated words



Align

Augment

```
#camerounais#
#kam_ʁu_nɛ__#

#différence#
#di_feʁɑ̃_s_#

#diffénais#
#di_fenɛ__#
```

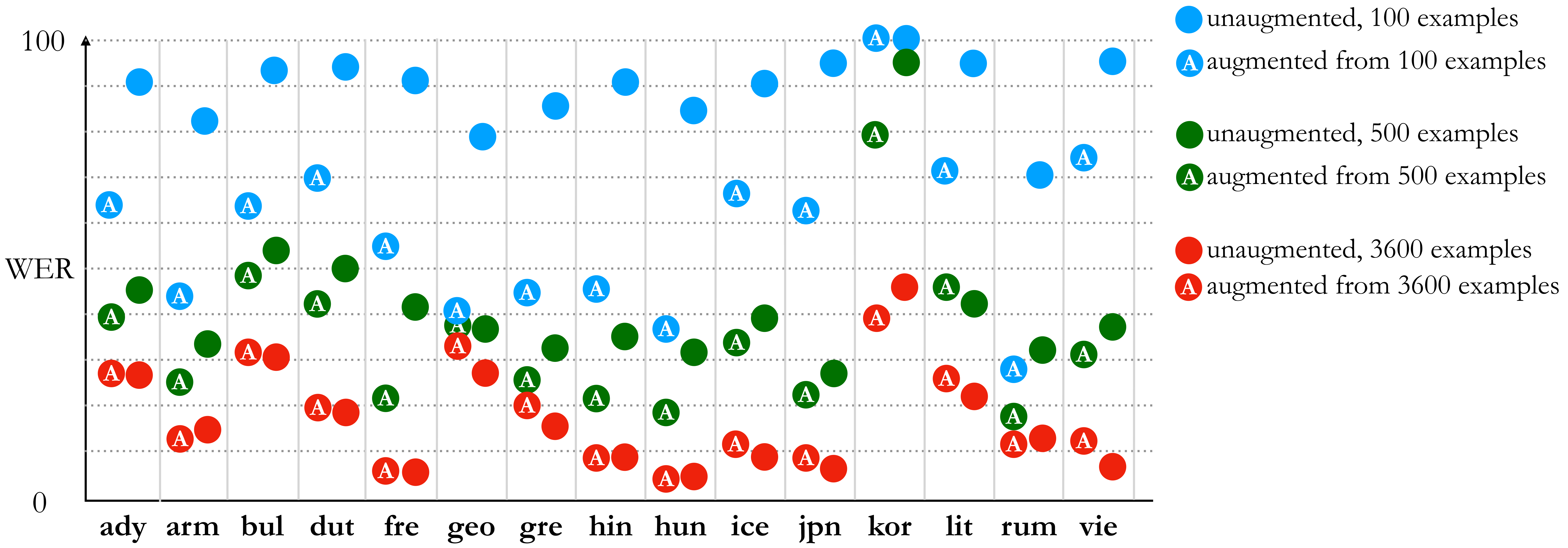
(3) Generate new data

- We also use an unsupervised algorithm to learn which phonemes are consonants and vowels
- We only splice together pieces where we get CV or VC at the juncture

procurions	p ʁ ɔ k y ʁ j ɔ̃
reconnaituer	ʁ ə k ɔ n ɛ t ɥ e
brancétude	b ʁ ɑ̃ ʃ e t y d
davasonnage	d a v ɑ̃ s ɔ n a ʒ
magazoulevard	m a g a z u l v a ʁ
oucoutume	w ɛ k u t y m
socendredi	s ɔ s ɑ̃ d ʁ ə d i
thapu	t a p y
sedi	s ə d i
sagementsier	s a ʒ m ɑ̃ z j e

Table 1: Example augmented French data from the original **min** data set that contains 100 examples. In total, 50,000 examples such as the ones shown here are created from each data set.

Results



Lang	100	100^{aug}	500	500^{aug}	full	full^{aug}
ady	90.22	**64.67**	45.33	**39.78**	**27.33**	27.78
arm	82.89	**45.33**	33.11	**24.89**	14.89	**13.33**
bul	93.56	**64.89**	53.78	**48.44**	**30.22**	32.22
dut	95.33	**69.11**	50.67	**42.00**	**18.22**	19.11
fre	91.56	**56.22**	41.78	**22.00**	**6.00**	6.22
geo	79.78	**40.89**	**37.33**	38.89	**27.78**	33.33
gre	86.00	**44.89**	32.00	**26.67**	**16.67**	20.67
hin	90.44	**46.22**	34.44	**21.33**	9.56	**9.11**
hun	84.89	**37.33**	31.78	**17.11**	4.67	**4.44**
ice	91.11	**66.89**	39.33	**33.78**	**9.56**	10.67
jpn	95.56	**62.22**	28.22	**22.00**	**6.67**	8.67
kor	**100.0**	**100.0**	95.78	**79.78**	46.22	**39.78**
lit	94.89	**70.89**	**42.89**	46.44	**21.78**	26.22
rum	70.67	**28.67**	31.56	**17.11**	12.22	**11.33**
vie	96.44	**74.89**	37.33	**30.89**	**7.11**	11.78

Table 2: Word error rate (WER) results on the test set when trained with 100 examples, 500 examples, and the full data set, compared to augmentation $(^{\text{aug}})$ for (100,500,3600) → 50,000 synthetic examples.