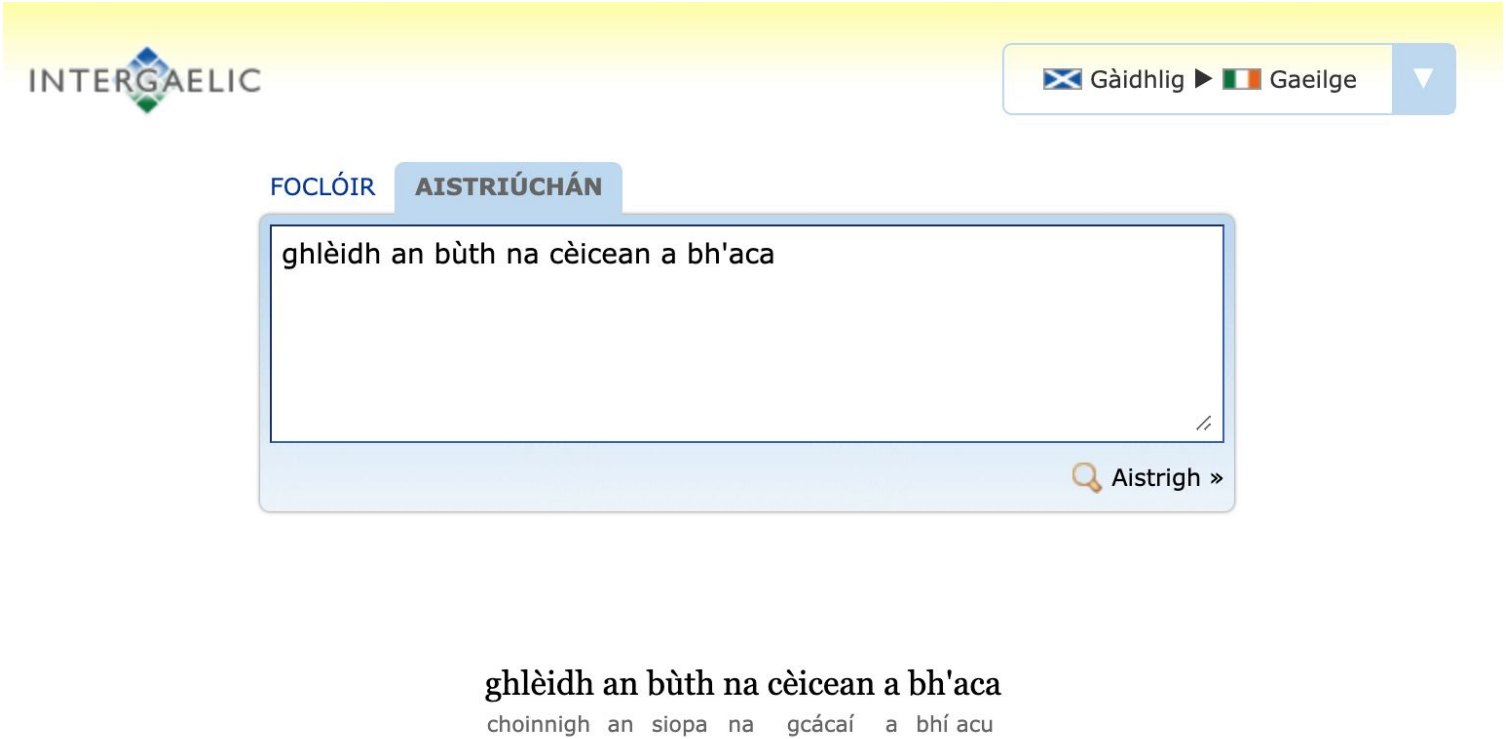
# Initial mutations as low-entropy features in neural language modeling

Kevin Scannell
Ollscoil Saint Louis / Acadamh na hOllscolaíochta Gaeilge
6 Feabhra 2020

# “Praistriúchán”

Irish portmanteau word:
“praiseach” = “a mess, a botch job”, “aistriúchán” = “translation”

# intergaelic.com



# Language modeling

- A language model (LM) is a probability distribution over sequences of words
- If S = “colorless green ideas…”, a language model assigns this a prob P(S):
- P(S) = P(colorless|^) P(green|colorless) P(ideas|colorless green) …
- Usually formulated and computed this way (word prob given history)
- LMs capture a lot! Pragmatics, syntax, real-world knowledge, …
- P(Carolina|We spent spring break in South) >
  P(Dakota|We spent spring break in South)
- P(is|The dog that chased the cat that chased the mice) >
  P(are|The dog that chased the cat that chased the mice)

# Applications

- Almost all important language technologies use LMs at some level!
- Can be used generatively
- MT, ASR, etc. fundamentally generate text, conditioned on input
- Conversational agents (Turing test)
- Strong LM alone can do question answering, summarization, ...
- Better language models give better end-to-end performance, generally

# Neural language models

- A flood of recent papers on neural language modeling, big leaps forward
- Originally, feed-forward neural networks (Bengio et al, 2003)
- Various refinements + regularization of recurrent networks (LSTMs, etc.)
- Most recently the Transformer architecture (Vaswani et al, 2017)
- My current research involves applying these developments to Irish
- Want to discuss one small linguistically interesting piece of this today…

# Research on English != Research on Language

- Sites tracking SOTA for language modeling show English datasets only
- Research almost 100% (and implicitly!) focused on English
- The word “English” isn’t used even once in these groundbreaking papers:
  - Google Brain’s landmark 2016 paper “Exploring the limits of language modeling”
  - Melis et al’s “On the state of the art of evaluation in neural language models” (2017)
  - Dai et al’s “Transformer-XL” paper (2019)
  - New SOTA “Megatron-LM” paper (2019)
- New architectures are likely to only benefit languages with massive corpora
- Also are unlikely to work well for morphologically complex languages

# Celtic initial mutations

- Celtic languages have initial mutations usually triggered by context
- *bád seoil* “sailboat”, *mo b**h**ád seoil* “my sailboat”, *ár **m**bád seoil* “our sailboat”
- Gender: *fear* “man”, *an fear bocht* “the poor man”, but:
- *bean* “woman”, *an b**h**ean b**h**ocht* “the poor woman”
- Dative case: *ar an **m**bád seoil* “on the sailboat” (or, *ar an b**h**ád seoil*)
- Genitive plural:

| | | |
|---|---|---|
| *leithreas na* | | ***bh**fear* |
| toilet | DET.GEN.PL | men.GEN.PL |
| “the men’s toilet” | | |

- We consider five mutations: **none, lenition, eclipsis, t-prothesis, h-prothesis**

# Motivating examples

- This was (one of) Google's mistakes in the earlier image:
  **tríd an bóthar → tríd an **m**bóthar*
  through the road
- And Intergaelic too, tricked by VSO:
  **choinnigh an siopa na **g**cácaí a bhí acu*
  kept the shop the cakes that were at-them
  "the shop kept their cakes"
  (cf. *siopa na gcácaí* "the shop of the cakes", "the cake shop")

# Mutations as low-entropy features

- Celtic mutations carry very little information
- Usually determined by the previous two words and initial letter of target word
- Could remove them and one can almost always replace them unambiguously:

Deirtear go iompraíodh sí gunnaí ina carr, iad faoi ceilt i mála plúir.

Ní raibh Gaoth Dobhair ann mar ainm dúiche ná paróiste ar tús, ach mar ainm ar an gaoth / abhainn ónar baisteadh an ceantar, an cainéal nó an inbhear farraige idir an paróiste agus na Rosa, ar a tugtar an Gaoth go dtí an lá inniú, agus an abhainn.

# Mutations as low-entropy features

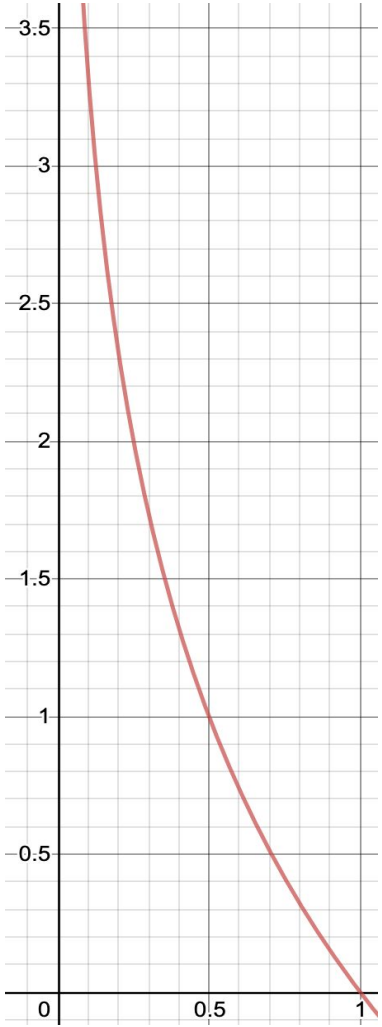
- Celtic mutations carry very little information
- Usually determined by the previous two words and initial letter of target word
- Could remove them and one can almost always replace them unambiguously:

Deirtear go n-iompraíodh sí gunnaí ina carr, iad faoi cheilt i mála plúir.

Ní raibh Gaoth Dobhair ann mar ainm dúiche ná paróiste ar dtús, ach mar ainm ar an ghaoth / abhainn ónar baisteadh an ceantar, an cainéal nó an t-inbhear farraige idir an pharóiste agus na Rosa, ar a dtugtar an Gaoth go dtí an lá inniu, agus an abhainn.

# What is entropy?

- Repeat the above experiment, but now imagine that you have €1.00 to wager on each word
- For “...iad faoi ceilt”, you might bet €0.99 on lenition, and €0.0025 each on the other four possibilities
- For “idir an paróiste”, you might bet €0.75 on no mutation, €0.24 on lenition, and €0.003333 on the other three
- Whatever the correct mutation is, you **lose** an amount equal to $-\log_2$ of your bet (see graph)
- The entropy is your average loss per bet; it measures how hard it is to predict mutations. Our claim is that we can make near-optimal bets to make this loss very small!



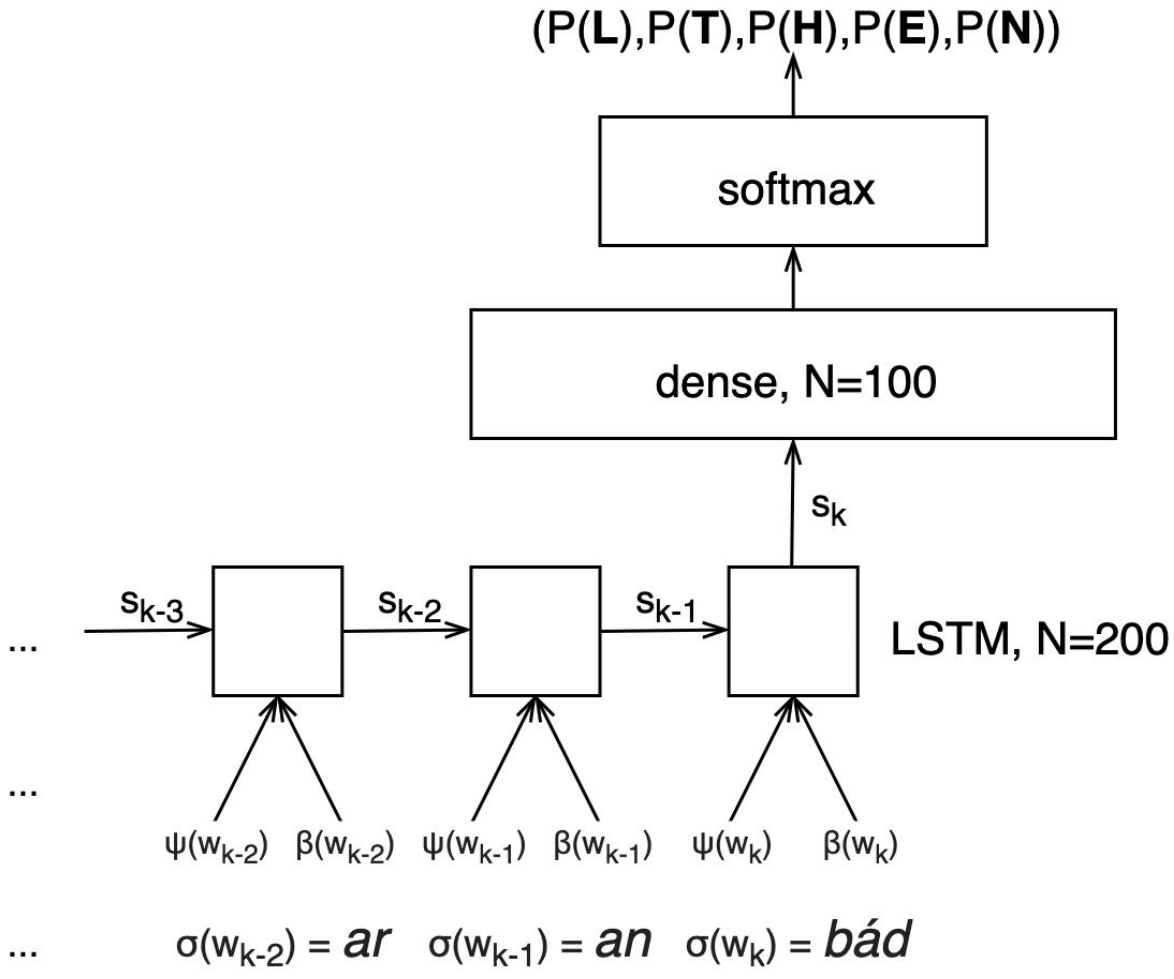
# A formula for entropy of mutations

- “Average number of bits per word carried by mutations”
- Let μ(w) be the mutation of w, and let σ(w) be w with its mutation removed
- Build a neural network model that predicts P(mutation | word history)
- Compute the $\log_2$ loss of this model on a test set

$$\Lambda = -\frac{1}{N}\sum_{i=1}^{N} \log_2 P(\mu(w_i)|\sigma(w_1)\ldots\sigma(w_i))$$

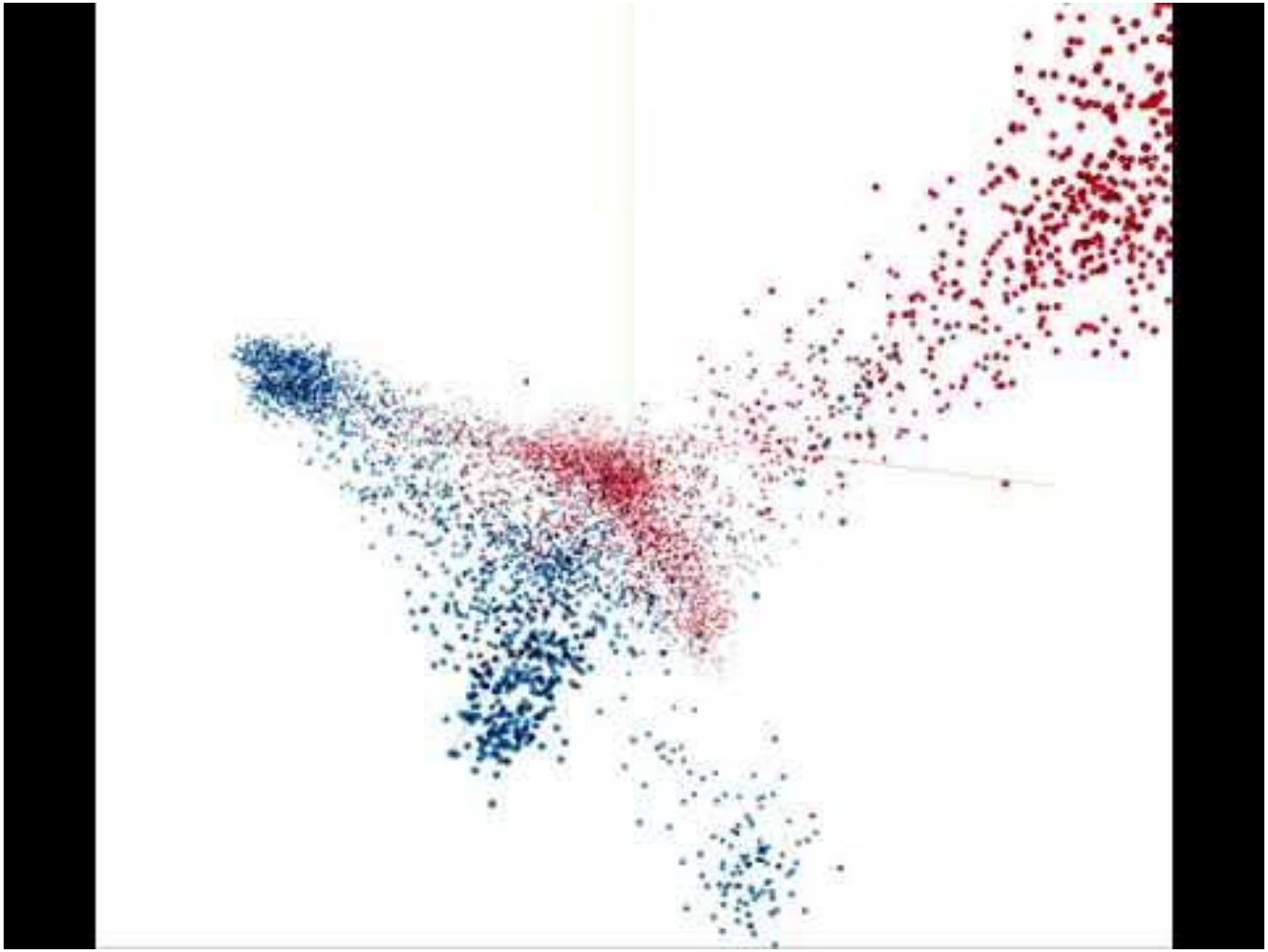
# Factored language models

- Word-based LMs don’t see that *bád, bhád, mbád* are really the same word
- Since “bád” is most common, harder to predict collocations like “bhád seoil”
- Standard solution: factored language models (Bilmes and Kirchhoff, 2003)
- View each word w as a bundle of features
- Factor P(w) as a product of feature probabilities conditioned on earlier features
- In our case this is simple! Features are the demutated word and the mutation
  e.g., P(*bhád* | … *mo*) = P(*bád* | … *mo*) P( **lenition** | … mo bád)


(P(L),P(T),P(H),P(E),P(N))
softmax
dense, N=100
$s_k$
$s_{k-3}$
$s_{k-2}$
$s_{k-1}$
LSTM, N=200
...
...
$\psi(w_{k-2})$ $\beta(w_{k-2})$ $\psi(w_{k-1})$ $\beta(w_{k-1})$ $\psi(w_k)$ $\beta(w_k)$
... $\sigma(w_{k-2}) = ar$ $\sigma(w_{k-1}) = an$ $\sigma(w_k) = bád$

# Results

- 2.32193 ($\log_2 5$) bits/word for random labels
- 0.75917 bits/word using label prior probabilities
- 0.40571 bits/word using unigram model (label distribution per word)
- 0.10710 bits/word using trigram model
- **0.06949** bits/word: NN trained on 50M words, 100k vocabulary, 15 epochs
- More than ⅓ of the loss comes from human errors in test corpus!



# Applications

- Improved LM for Irish when used in a factored model on demutated words
- Data-driven grammar checking which robustly handles variant spellings, etc.
- Data-driven estimate of information-theoretic content of mutation system
- Large (quantifiable) divergence between official standard(s) and actual usage

# Which mutations carry information?

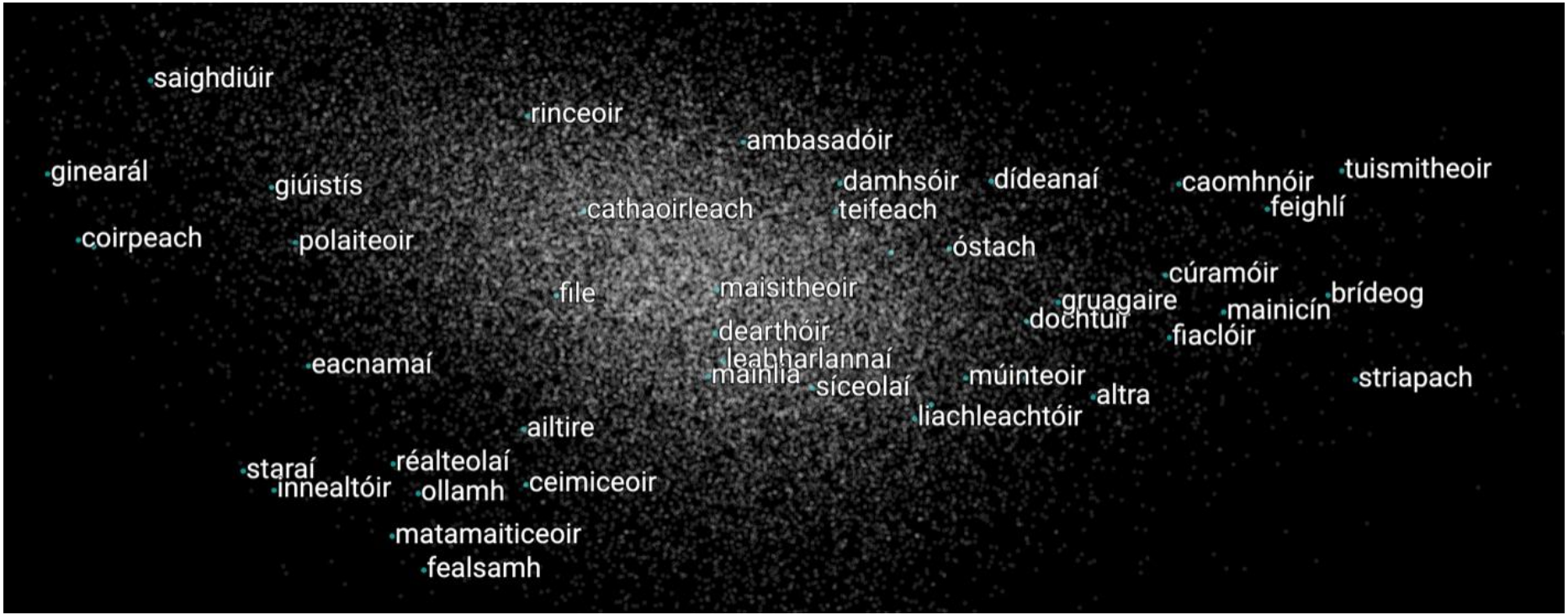
- Of 10000 examples, correct label was assigned P<0.5 167 times, 98.3% correct
- These 167 examples contribute 77% of the total loss!
- 61 of 167 are grammatical errors in the test file
- 30 were assigned low prob only because of lack of context to the right
- 23 were correct but non-standard forms (*dhom*, e.g.)
- 16 relate to some form of the third person possessive (*a*, *ina, faoina, …*)
- 9 are dialect differences: lenition vs. eclipsis in the dative
- Various assorted others
- (Note the many cases *not* here, e.g. indirect vs. direct relativizer, etc.)

# Digression: orthographic transparency

- This approach only works for Irish and Scottish Gaelic
- Four of the five mutations in Irish can be trivially and algorithmically removed
- h-prothesis cannot, in general: (***h****amhlaidh* vs. *hidrigin*)
- Even with a dictionary, some ambiguity: *aiste* “essay” vs. *haiste* “hatch”
- I strip all h’s and let the neural networks figure it out!
- Scottish Gaelic is transparent in all cases (they write h-)
- Welsh, Cornish, Breton, and Manx Gaelic are not at all transparent!

# Gender bias in training

- *tá sé/sí ina m**h**úinteoir/múinteoir*
  is he/she in-his/her teacher
  “he/she is a teacher”
- Discard words with gender baked-in: máistir/máistreás, siúr/bráthair, etc.
- Of remaining 446 occupations, male mutation is more common for *434* (sic!)
- Exceptions: altra, aoi-léachtóir, comhláithreoir, comhordaitheoir, comhstiúrthóir, cuiditheoir, damhsóir, fidléir, gnáthurlabhra, mainicín, striapach, tréidlia
- Strongest male bias: ardcheannasaí, coirnéal, gineárál, giúistís, iascaire, marcach, misinéir, óglach, peileadóir, píobaire, printíseach, seanchaí, tosaí


saighdiúir
rinceoir
ambasadóir
ginearál
giúistís
damhsóir
dídeanaí
caomhnóir
tuismitheoir
cathaoirleach
teifeach
feighlí
coirpeach
polaiteoir
óstach
cúramóir
file
maisitheoir
gruagaire
brídeog
dochtúir
mainicín
dearthóir
fiaclóir
eacnamaí
leabharlannaí
mainlia
múinteoir
striapach
síceolaí
altra
liachleachtóir
ailtire
staraí
réalteolaí
innealtóir
ollamh
ceimiceoir
matamaiticeoir
fealsamh

# Gender bias in mutation prediction

Predicted Mutation

| Actual Mutation | Masculine | Feminine | Plural | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| Masculine | 189 | 7 | 2 | 0.9594 | 0.9545 | 0.9570 |
| Feminine | 8 | 29 | 2 | 0.8056 | 0.7436 | 0.7733 |

# Thank you! / Go raibh maith agaibh!

- https://cs.slu.edu/~scannell/
- https://cadhan.com/
- http://crubadan.org/
- http://indigenoustweets.com/
- http://chuala.me/
- http://intergaelic.com/
- http://corpas.ria.ie/
- https://github.com/kscanne/