

Towards explainable AI for Irish grammatical error correction

24 April 2023

Grammatical error correction

- Task takes a sentence as input, and outputs the sentence with errors corrected
- Easier problem: grammatical error *detection*
- Harder problem: Grammatical error detection/correction plus *explanations*
- Research is mostly on English; small datasets for maybe 20 languages total
- Definition of what constitutes an error is fuzzy; defined by the datasets!
- Exceptionally difficult task: SOTA F-scores in the 60's for English (low recall)

Celtic initial mutations

- Celtic languages have initial mutations usually triggered by context
- *bád seoil* “sailboat”, *mo bhád seoil* “my sailboat”, *ár mbád seoil* “our sailboat”
- Gender: *fear* “man”, *an fear bocht* “the poor man”, but:
- *bean* “woman”, *an bhean bhocht* “the poor woman”
- Dative case: *ar an mbád seoil* “on the sailboat” (or, *ar an bhád seoil*)
- Genitive plural: *leithreas na bhfear*
toilet DET.GEN.PL men.GEN.PL
“the men’s toilet”

Official Standard(s)

- First version published in 1958, revised in 2012 and then 2017
- Among other things, rules for when the various mutations should be used
- But...
 - Not fully aligned with existing grammars and dictionaries (including CB, FGB, NEID)
 - Standard form is “artificial” — doesn’t agree with the way native speakers use the language
 - Many divergences according to dialect
 - Even accepting all of this, rules leave quite a bit up to interpretation
 - So we see variance even among those committed to the standard
 - Upshot for us: no completely reliable training data exists!

One problematic example (of dozens)

- “saoirse cainte” or “saoirse chainte”?
- 177M word corpus: “saoirse cainte” 687 times and “saoirse chainte” 28 times
- “saoirse creidimh” or “saoirse chreidimh”?
- Same corpus: 85 times vs. 26 times

creideamh, *m.* (*gs. & npl. -dimh, gpl. ~*). Belief, faith; religion, creed. ~ **i nDia**, belief in (the existence of) God. ~ **sna sacraimintí**, belief in (the validity of) the sacraments. ~ **i luibheanna**, belief in (the efficacy of) herbs **An ~ fíor**, the true faith. **Ár g~ a admháil**, to profess our faith. **Níl ~ ná coinsias aige**, he has neither faith nor conscience. **A lucht an chreidimh bhig**, you of little faith. **Duine a thabhairt chun creidimh**, to bring s.o. to the faith; to bring s.o. round to one's point of view. **Tá siad ar aon chreideamh amháin**, they are of one faith, persuasion. **Is é mo chreideamh go**, it is my belief, conviction, that. **An ~ Caitliceach, Protastúnach**, the Catholic, Protestant, faith, religion. **Saoirse chreidimh**, religious freedom. ~ **polaitíochta**, political creed. (*Var: pl. creidiocha*)

saoirse², *f.* (*gs. ~*). Freedom. **1. Lit.** Status of freeman; nobility. ~ **cineáil**, nobility of race. **2.** Liberty, independence. ~ **na tíre**, the freedom of the country. ~ **cainte**, freedom of speech. ~ **coinsiasa**, liberty of conscience. ~ **é duine**, human freedom; personal liberty. **3.** Immunity, exemption. ~ **ó chánacha**, exemption from taxes. ~ **ar dhualgas**, freedom from an obligation. ~ **fónaimh**, exemption from service. **4. Lit.** Privilege. **Cinseal ag ~**, power and privilege. **5.** Honorary privilege. ~ **na cathrach**, the freedom of the city. **6.** Cheapness, inexpensiveness.

freedom

1 NOUN power, latitude

saoirse *fem4* 4) C M U

she has the freedom to make her own decisions **tá an tsaoir tá sí saor chun a cinntí féin a dhéanamh**
to live in freedom from fear **maireachtáil saor ó eagla**
freedom of movement **saoirse ghluaiseachta**
freedom of speech **saoirse labhartha, saoirse chainte**
freedom of thought **saoirse smaointeoireachta**
political freedom **saoirse pholaitiúil**
religious freedom **saoirse chreidimh, saoirse reiligiúin**

Phrases and Examples in other entries

| | |
|----------------------|--|
| curtailment » | it's a curtailment of freedom of speech is ciorrú ar shaoirse cainte é |
| enjoyment » | the government must facilitate the enjoyment of the freedom of speech caithfidh an rialtas saoirse cainte an phobail a éascú |
| freedom » | freedom of speech saoirse labhartha, saoirse chainte |
| limit » | they can't limit freedom of speech ní féidir leo srian a chur le saoirse cainte |
| profess » | they profess to support freedom of speech maíonn siad go bhfuil siad i bhfách le saoirse cainte |

Rule-based system (2000–2004)

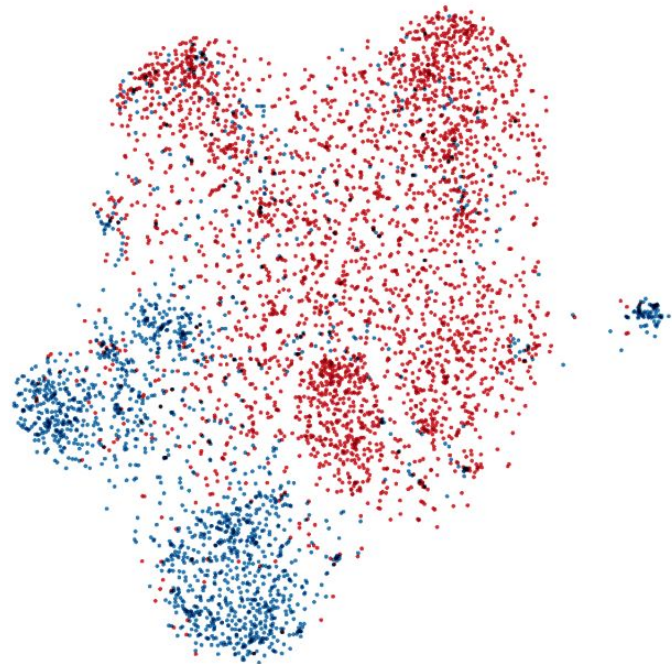
- This initial attempt was based on explicit rules
- Perform part-of-speech tagging, and then apply pattern-matching rules
- Exceptions, and exceptions to the exceptions, etc. (2814 rules in all)
 - Positive: possible to provide explanations/diagnoses when errors are flagged
 - Negative: usual brittleness of rule-based systems
 - Negative: tedious to implement all the rules; exceptionally hard to maintain them
 - Negative: some rules depend on semantics or syntax and can't be encoded in this formalism

Unsupervised approach

- If you sacrifice the explanations, can predict mutations in an unsupervised way
- Formulate this as a tagging problem with five tags: N, S, U, T, H
- Can (almost) remove mutations algorithmically => unlimited training data
- I knew this in 2004, but it was hard to get right using n-gram language models
- Some rules depend on subtle character-level info (slender endings, etc)
- Generalized parallel backoff (Bilmes and Kirchhoff)

Neural system: 17–18 October 2019

- Neural model: eliminates the hard parts of the statistical approach
- LSTM layer(s), BiLSTM at character level
- No need to hand-select features; no complicated backoff schemes
- Achieves much higher accuracy than previous approaches
- Character-based component learns gender other relevant features (“snideog”)
- Word-based component learns sometimes subtle contextual clues (“Ó Baoill”)



Digression: entropy of mutations

- Network has a five-fold softmax at the top, trained to minimize cross-entropy
- Best model achieves entropy *very* close to zero; under 0.07 bits per word
- Informally, “Irish initial mutations carry virtually no information”, empirically
- Even stronger; more than half of the loss caused by errors in the test set!
- Remainder are things like 3rd person possessives (well-known)
- Some with zero impact: direct vs. indirect relatives

Critique

- Learns mutations as used by the language community, “errors” included
- Makes some inexplicable mistakes; difficult to debug
- Easy things are easy for it; less effective on the cases humans find difficult
- **No explanations given**

How do we get the explanations back?

- Want to keep it as a tagging problem but with an enhanced set of tags
- Augment each tag with a section of the official standard that “explains” it
- *Tá an bhean/S+10.2.1 ag canadh*
- *Tá an doineann/N+10.2.1.e1 ag maolú*
- Just need to produce millions of training examples, somehow :/

10.2.1 I nDIAIDH AN AILT

Cuirtear séimhiú ar an ainmfhocal i ndiaidh an ailt (mura *d*, *t* nó *s* an túschonsan)–

(a) san ainmneach uatha baininscneach, e.g., *an chathair*; *an ghloine*; *an fhuascailt*:

Tá **an bhean** ag canadh.

Ar dhún sé **an fhuinneog**?

Universal Dependencies

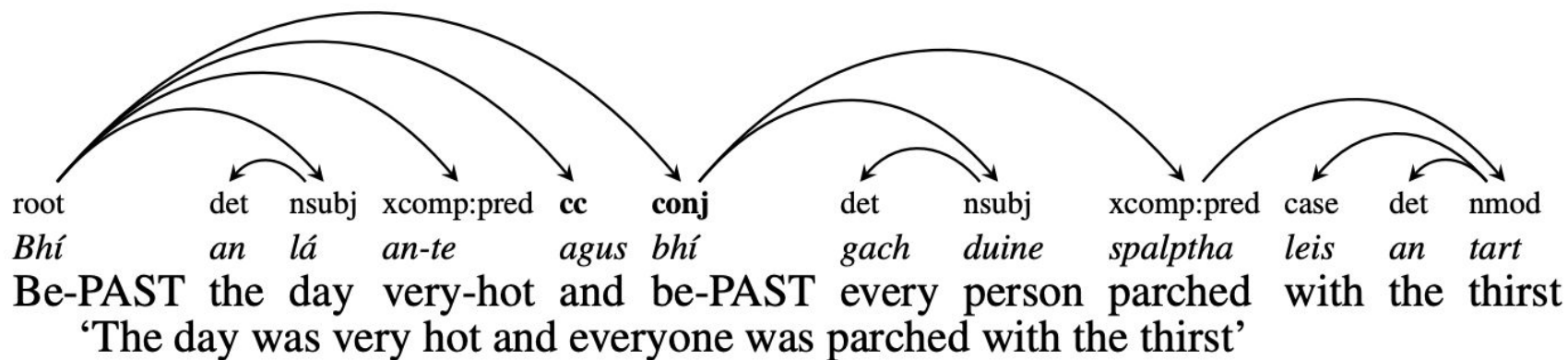


Figure 4 in Lynn, Teresa and Foster, Jennifer (2016) *Universal dependencies for Irish*. In: *Second Celtic Language Technology Workshop. (CLTW 2016)*, 4 July 2016, Paris, France.

Celtic UD treebanks

- Irish: Teresa Lynn's [Ph.D. thesis](#) (2016)
- Scottish Gaelic: [Colin Batchelor](#) (2019)
- Manx Gaelic: [Scannell](#) (2020)
- Welsh: [Heinecke and Tyers](#) (2019)
- Breton: [Tyers and Ravishankar](#) (2018)
- Cornish: ????
- Considerable effort has gone into harmonizing annotation schemes
- Allows cross-linguistic comparison and transfer learning

CoNLL-U format (with features)

sent_id = 465

text = Bhí sí naoi mbliana agus leathchéad.

| | | | | | | |
|---|------------|------------|-------|-------------------------------------|---|------------|
| 1 | Bhí | bí | VERB | Form=Len Mood=Ind Tense=Past | 0 | root |
| 2 | sí | sí | PRON | Gender=Fem Number=Sing Person=3 | 1 | nsubj |
| 3 | naoi | naoi | NUM | NumType=Card | 4 | nummod |
| 4 | mbliana | bliain | NOUN | Case=NomAcc Form=Ecl Gender=... | 1 | xcomp:pred |
| 5 | agus | agus | CCONJ | _ | 6 | cc |
| 6 | leathchéad | leathchéad | NOUN | Case=NomAcc Gender=Masc Number=Sing | 4 | conj |
| 7 | . | . | PUNCT | _ | 1 | punct |

Feature Prediction

- I developed a suite a QA tools for checking the Irish treebank(s) in 2021
- <https://github.com/kscanne/grammatach>
- Implements “constraints” on feature values based on dependency relations
- For our purposes, enough that the “Form” feature captures mutations
- Refactored this code to parallel the rules in the standard to the extent possible


```

if hd['lemma'] in gadata.feminineGroups:
    return [Constraint('!Len', '10.2.7.j: Do not lenite a genitive noun after various feminine noun for groups or organizations')]

if self['lemma'] in ['dlí', 'sí']:
    return [Constraint('!Len', '10.2.7.k: Do not lenite “dlí” or “sí” after a feminine noun')]

if self['lemma'] in ['bliain', 'coicis', 'mí', 'seachtain']:
    return [Constraint('!Len', '10.2.7.l: Do not lenite various units of time after a feminine noun')]

if self.isQualifiedNoun():
    if hd['lemma'] in ['beirt', 'dís']:
        return [Constraint('Len', '10.2.7.m.e1: Lenite nouns after “beirt” or “dís” even if the genitive noun is modified by an adjective or another noun')]
    return [Constraint('!Len', '10.2.7.m: Do not lenite after a feminine noun if the genitive noun is modified by an adjective or another noun')]

```

- (j) Ní shéimhítear an dara hainmfhocal i ndiaidh na bhfocal *aicme, comhairle, corparáid, cuideachta, earnáil, feidhmeannacht, foireann, gníomhaireacht, institiúid, oifig, rannóg, roinn, scéim, seirbhís*, e.g., *aicme ceannais; comhairle contae; corparáid baincéireachta; earnáil gnó; foireann bainistíochta; institiúid breisoidreachais; oifig preasa; rannóg pearsanra; roinn cosanta*:

Ní foláir dóibh an **scéim marcála** a choigeartú.
Cuireann an gnólacht **seirbhís comhairleoireachta** ar fáil.

- (k) Ní shéimhítear na hainmfhocail *dlí* ná *sí*, i ndiaidh ainmfhocal baininsneach, e.g., *feidhm dlí; bó sí*:

Cuirfidh an rialtas cóip den **ionstraim dlí** ar fáil.
Deirtear gur comhartha an mhí-áidh an **long sí** a fheiceáil.

- (l) Ní shéimhítear an dara hainmfhocal más ceann de na haonaid tomhais bheachta *bliain, coicis, mí* nó *seachtain* an dara hainmfhocal, e.g., *saoire míosa; tréimhse bliana; moill bliana*:

Cuireadh tús leis an **bhféile seachtaine** le hócáid i dteach an Mhéara.
Bhí **scíth coicise** de dhíth uirthi tar éis na hoibre.

- (m) Ní shéimhítear an dara hainmfhocal má tá an dara hainmfhocal, atá sa ghinideach, cáilithe le hainmfhocal eile nó le haidiacht, e.g., *oíche gaoithe móire; bileog páipéir dúigh; gloine beorach fuaraithe*:

Ní raibh **deoir bainne úir** fágtha sa teach.
Cuireadh an **obair tionscadail aistriúcháin** i gcrích inné.

Eisceacht: nuair is é *beirt* nó *dís* an chéad ainmfhocal, e.g., *beirt bhan mhisniúla* nó nuair is é *beirt* an dara hainmfhocal, e.g., *páistí bheirt bhan*.

Corpus-level survey of errors

| Error | Sample error | Total | Errors | % |
|--------------|-------------------------------------|-------|--------|---------|
| 10.2.5.a.e1 | um bhainistíocht | 2 | 2 | 100.00% |
| 10.3.1.d.ii: | sé chúrsa páirtaimseartha | 1 | 1 | 100.00% |
| 10.3.1.d.i: | A dhá chúrsa páirtaimseartha | 6 | 3 | 50.00% |
| 10.2.6: | Shc in aghaidh Binse na Státseirbhí | 40 | 16 | 40.00% |
| 10.2.8.a: | D faoi chórais theilifíse | 23 | 5 | 21.74% |
| 10.11.2.a: | S an tríú áit | 21 | 4 | 19.05% |
| 10.2.7.j: | Dc eanáil chruach | 69 | 10 | 14.49% |
| 10.4.1.i.e1: | tángthas | 8 | 1 | 12.50% |
| 10.2.4.c.e1 | trí throigh | 34 | 4 | 11.76% |
| 10.2.5.c.e1 | gan chomhaoin luachmhar | 9 | 1 | 11.11% |
| 10.2.9: | Do le linn an tséasúir mháirseála | 243 | 15 | 6.17% |
| 10.2.1.c: | Al don bhfile | 645 | 39 | 6.05% |
| 10.2.8.b: | D ar chumais fhiontraíochta | 17 | 1 | 5.88% |
| 10.2.5.a: | Sl faoi clár | 146 | 7 | 4.79% |
| 10.2.3.b: | Al chuile duine | 24 | 1 | 4.17% |
| 10.6.3: | Shc ocht bliana | 49 | 2 | 4.08% |
| 10.11.3: | Sh le iascaire | 218 | 8 | 3.67% |
| 10.4.2.b: | Fc inar dhúirt sé | 141 | 5 | 3.55% |

Synthetic training data

- Produce dependency parses for sentences in a large corpus
- Ignore tokens with incorrect features according to QA scripts
- (Mix of incorrect parses and corpus examples not compliant with standard)
- Each remaining token is attached to one or more constraints
- Intentionally violate those constraints by changing the mutation and add tag:
 - Correct sentence: *Tá an bhean ag canadh*
 - Constraint on “bhean” requires the S tag with reference 10.2.1
 - Can violate this two ways: *Tá an bean/S+10.2.1 ag canadh*
Tá an mbean/S+10.2.1 ag canadh
- Training works much like the previous paper (test set in progress)

Benefits and future directions

- Neatly handles the issue of non-standard texts in training
- Can oversample rare mutation contexts that the previous model failed on
- No reason to limit to mutation features (errors in agreement verb tense, etc.)
- Approach should apply neatly to the 4 other Celtic languages with treebanks

Thank you! / Go raibh maith agaibh!

- <https://cadhan.com/>
- <https://github.com/kscanne/>