UNIVERSITY OF ICELAND

Communicating with computers with small languages

Hafsteinn Einarsson

SCHOOL OF ENGINEERING AND
NATURAL SCIENCES

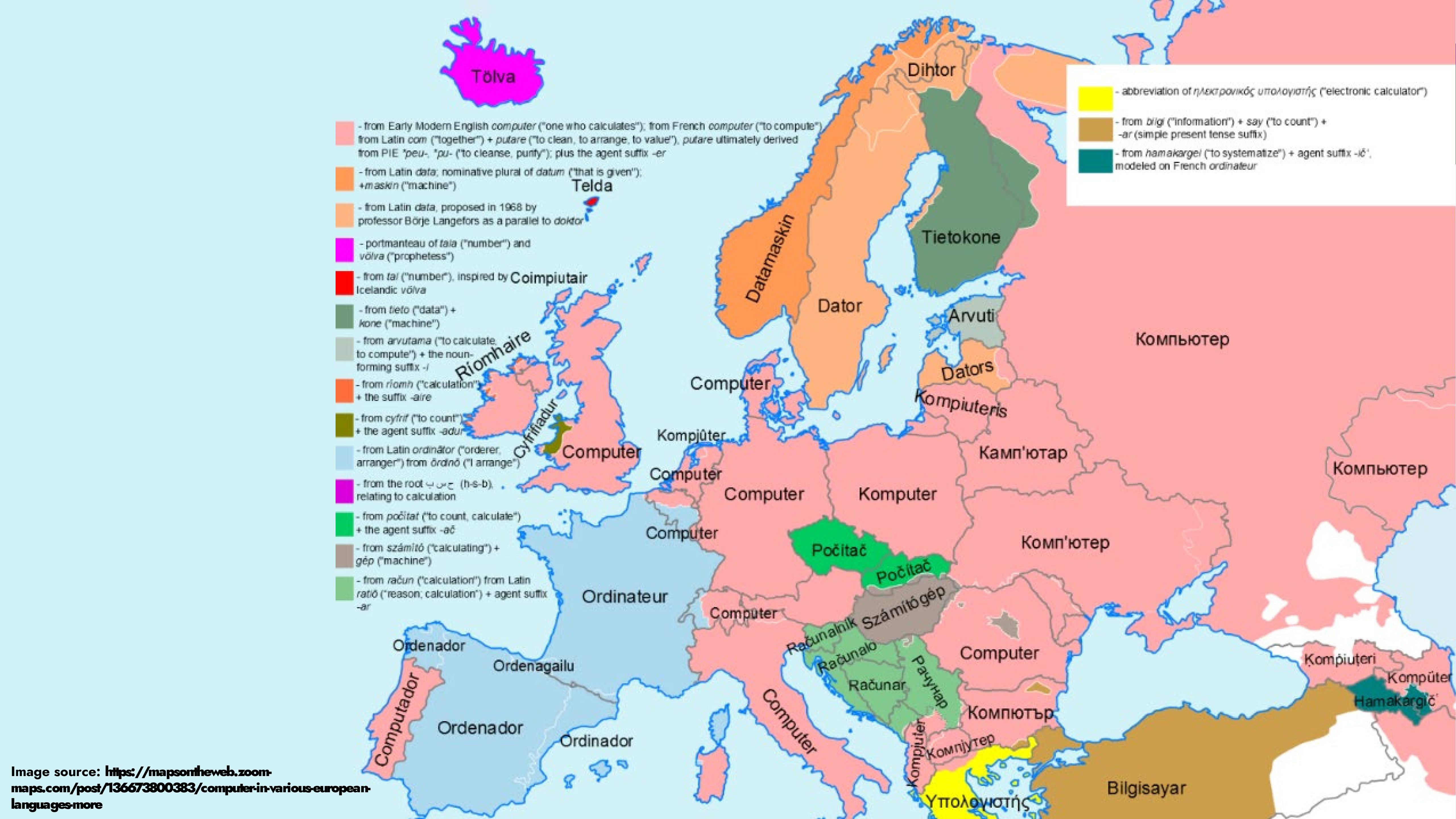
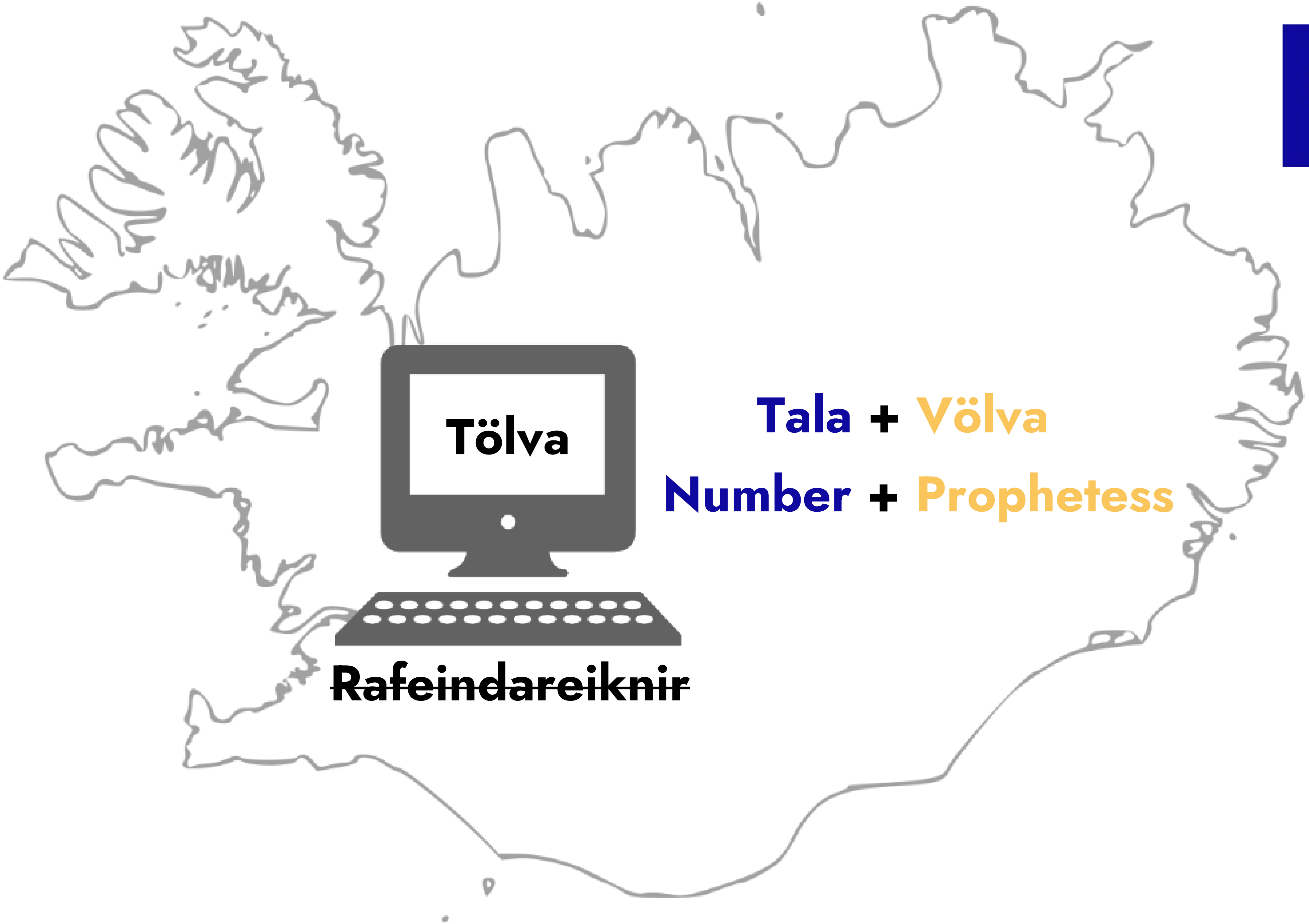

Tölva
- from Early Modern English computer ("one who calculates"); from French computer ("to compute") from Latin com ("together") + putare ("to clean, to arrange, to value"), putare ultimately derived from PIE *peu-, *pu- ("to cleanse, purify"); plus the agent suffix -er
- from Latin data; nominative plural of datum ("that is given"); +maskin ("machine")
- from Latin data, proposed in 1968 by professor Börje Langefors as a parallel to doktor
- portmanteau of tala ("number") and völva ("prophetess")
- from tal ("number"), inspired by Icelandic völva
- from tieto ("data") + kone ("machine")
- from arvutama ("to calculate, to compute") + the noun-forming suffix -i
- from ríomh ("calculation") + the suffix -aire
- from cyfrif ("to count") + the agent suffix -adur
- from Latin ordinātor ("orderer, arranger") from ōrdinō ("I arrange")
- from the root ح س ب (h-s-b), relating to calculation
- from počítat ("to count, calculate") + the agent suffix -ač
- from számító ("calculating") + gép ("machine")
- from račun ("calculation") from Latin ratiō ("reason; calculation") + agent suffix -ar
- abbreviation of ηλεκτρονικός υπολογιστής ("electronic calculator")
- from bilgi ("information") + say ("to count") + -ar (simple present tense suffix)
- from hamakargel ("to systematize") + agent suffix -ič', modeled on French ordinateur
Telda
Dihtor
Datamaskin
Dator
Tietokone
Coimpiutair
Ríomhaire
Arvuti
Computer
Dators
Компьютер
Kompiuteris
Cyfrifiadur
Computer
Kompjûter
Камп'ютар
Computer
Компьютер
Computer
Komputer
Computer
Počítač
Počítač
Комп'ютер
Ordinateur
Computer
Számítógép
Računalnik
Računalo
Računar
Рачунар
Computer
Ķompiuteri
Kompüter
Hamakargič
Ordenador
Ordenagailu
Computer
Компютър
Computador
Ordenador
Ordinador
Kompjuter
Компјутер
Kompjuter
Υπολογιστής
Bilgisayar

Image source: https://mapsontheweb.zoom-maps.com/post/136673800383/computer-in-various-european-languages-more



Tölva

~~Rafeindareiknir~~

Tala + Völva

Number + Prophetess

Language Technology for Icelandic

UNIVERSITY OF ICELAND

- Initiative started back in 2018 with a €2.7M grant from the government
- The aim was to make Icelandic usable when communicating with or through computers.
- Projects
  - Speech-to-text
  - Text-to-speech
  - Machine translation
  - Grammatic error analysis
  - Dataset building
  - Summarization
  - Information retrieval
  - Sentiment analysis
  - Question answering systems
  - Chatbot systems

Language Technology has taken a leap forward

- See YouTube link: https://youtu.be/nm7gpwpQjG0

UNIVERSITY OF ICELAND

Building Icelandic language models

Vésteinn Snæbjarnarson

Haukur Barri Símonarson

Pétur Orri Ragnarsson

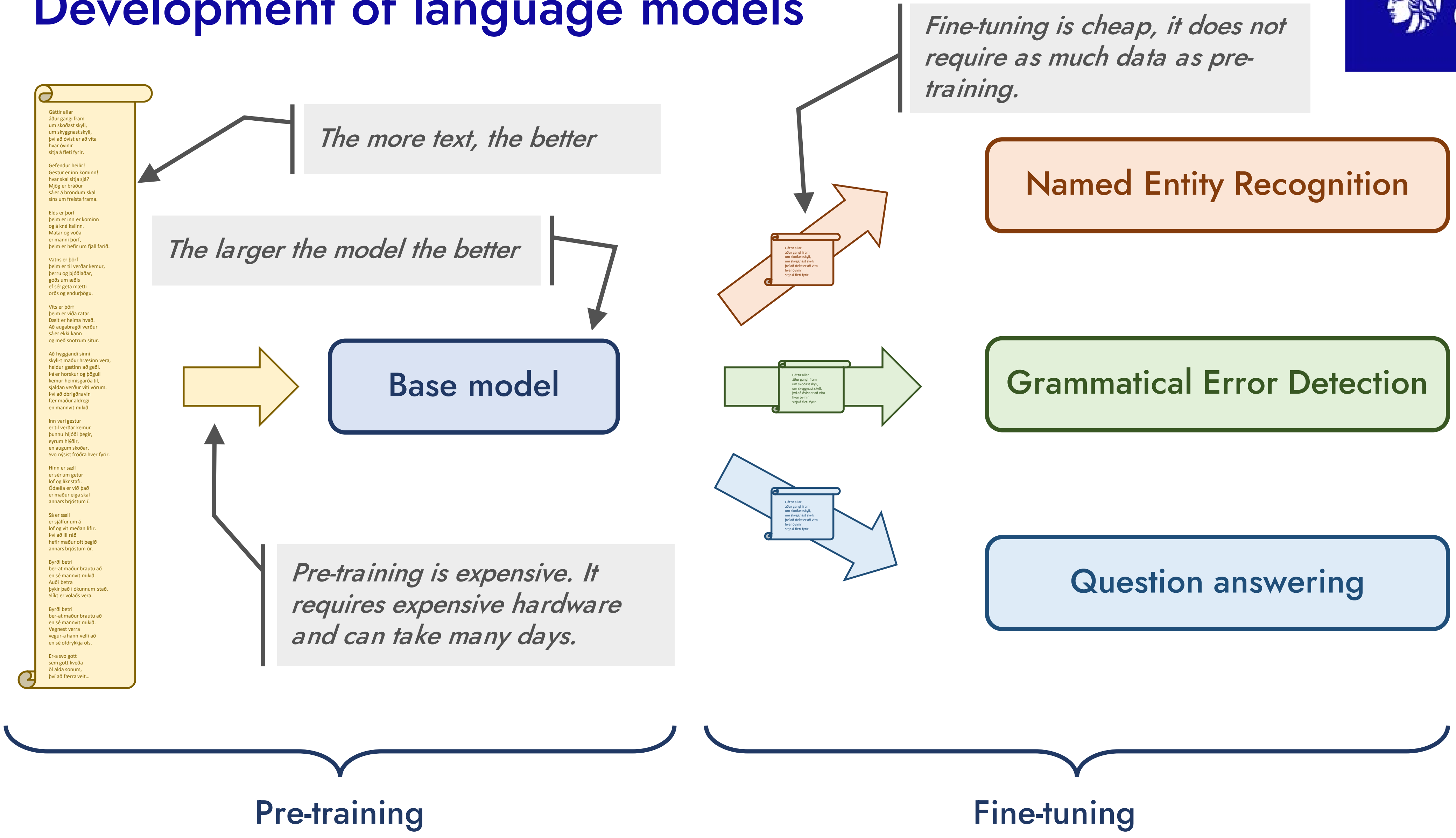
Svanhvít Lilja Ingólfsdóttir

Haukur Páll Jónsson

Vilhjálmur Þorsteinsson

Paper: A Warm Start and a Clean Crawled Corpus – A Recipe for Good Language Models (LREC 2022)

Development of language models



Gáttir allar
áður gangi fram
um skoðast skyli,
um skyggnast skyli,
því að óvíst er að vita
hvar óvinir
sitja á fleti fyrir.

Gefendur heilir!
Gestur er inn kominn!
hvar skal sitja sjá?
Mjög er bráður
sá er á bröndum skal
síns um freista frama.

Elds er þörf
þeim er inn er kominn
og á kné kalinn.
Matar og voða
er manni þörf,
þeim er hefir um fjall farið.

Vatns er þörf
þeim er til verðar kemur,
þerru og þjóðlaðar,
góðs um æðis
ef sér geta mætti
orðs og endurþögu.

Vits er þörf
þeim er víða ratar.
Dælt er heima hvað.
Að augabragði verður
sá er ekki kann
og með snotrum situr.

Að hyggjandi sinni
skyli-t maður hræsinn vera,
heldur gætinn að geði.
Þá er horskur og þögull
kemur heimisgarða til,
sjaldan verður víti vörum.
Því að óbrigðra vin
fær maður aldregi
en mannvit mikið.

Inn vari gestur
er til verðar kemur
þunnu hljóði þegir,
eyrum hlýðir,
en augum skoðar.
Svo nýsist fróðra hver fyrir.

Hinn er sæll
er sér um getur
lof og líknstafi.
Ódælla er við það
er maður eiga skal
annars brjóstum í.

Sá er sæll
er sjálfur um á
lof og vit meðan lifir.
Því að ill ráð
hefir maður oft þegið
annars brjóstum úr.

Byrði betri
ber-at maður brautu að
en sé mannvit mikið.
Auði betra
þykir það í ókunnum stað.
Slíkt er volaðs vera.

Byrði betri
ber-at maður brautu að
en sé mannvit mikið.
Vegnest verra
vegur-a hann velli að
en sé ofdrykkja öls.

Er-a svo gott
sem gott kveða
öl alda sonum,
því að færra veit…

The more text, the better

The larger the model the better

Base model

Pre-training is expensive. It requires expensive hardware and can take many days.

Fine-tuning is cheap, it does not require as much data as pre-training.

Named Entity Recognition

Grammatical Error Detection

Question answering

Pre-training

Fine-tuning

Pre-training data for Icelandic

UNIVERSITY OF ICELAND

Icelandic giga corpus (IGC)

8.2 GBs of curated Icelandic text from various sources (1,300M words)

The Icelandic Common Crawl Corpus (IC3)

4.9 GBs of cleaned Icelandic text from all websites using the .is top level domain

Icelandic language models

IceBERT

A model trained on curated text and text from the web
Training time: 96 days

IceBERT-IGC

A model only trained on curated text
Training time: 96 days

IceBERT-IC3

A model only trained on text from the web
Training time: 96 days

XLMR-IC3-7d

An existing multilingual model trained on text from the web
Training time: 7 days

IGC

IC3

Fine-tuning performance

	PoS-tagging	Named Entity Recognition	Grammatical error detection
IceBERT	98.33 🏆	91.43	70.11
IceBERT-IGC	98.27	91.10	71.33
IceBERT-IC3	98.30	91.29	70.95
XLMR-IC3-7d	98.20	92.52 🏆	75.87 🏆

IGC

IC3

Takeaway messages

Use existing multilingual model as a starting point

Use cleaned text from the internet to train your models

Vésteinn Snæbjarnarson

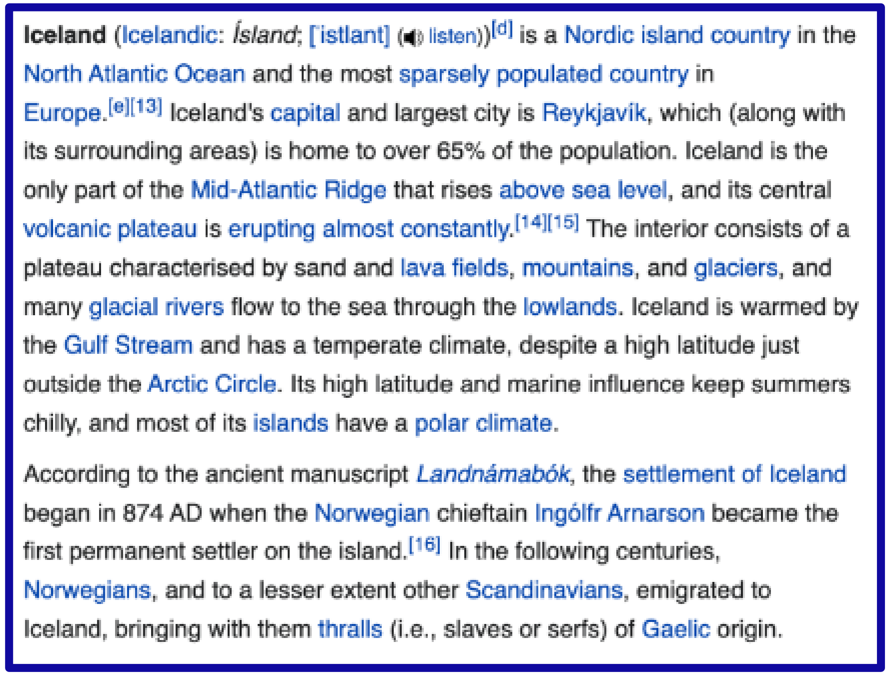
QA models

Paper: Cross-Lingual QA as a Stepping Stone for Monolingual Open QA in Icelandic (MIA @ NAACL 2022)

Reading comprehension

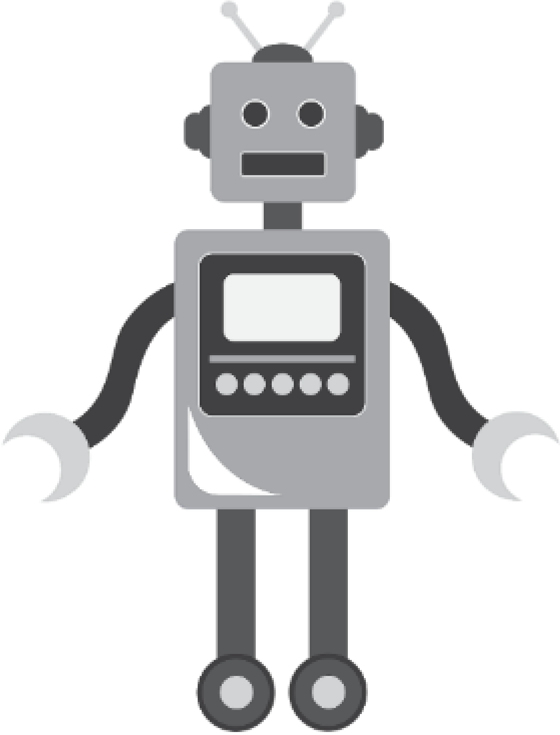
UNIVERSITY OF ICELAND

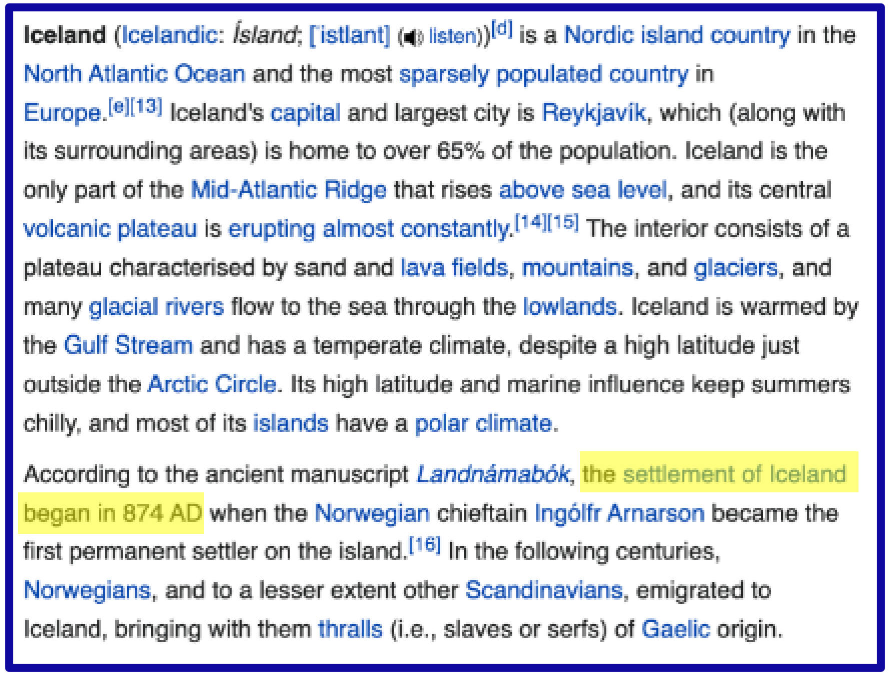
When did the settlement of Iceland begin?



Iceland (Icelandic: *Ísland*; [ˈistlant] (🔈 listen))[d] is a Nordic island country in the North Atlantic Ocean and the most sparsely populated country in Europe.[e][13] Iceland's capital and largest city is Reykjavík, which (along with its surrounding areas) is home to over 65% of the population. Iceland is the only part of the Mid-Atlantic Ridge that rises above sea level, and its central volcanic plateau is erupting almost constantly.[14][15] The interior consists of a plateau characterised by sand and lava fields, mountains, and glaciers, and many glacial rivers flow to the sea through the lowlands. Iceland is warmed by the Gulf Stream and has a temperate climate, despite a high latitude just outside the Arctic Circle. Its high latitude and marine influence keep summers chilly, and most of its islands have a polar climate.

According to the ancient manuscript *Landnámabók*, the settlement of Iceland began in 874 AD when the Norwegian chieftain Ingólfr Arnarson became the first permanent settler on the island.[16] In the following centuries, Norwegians, and to a lesser extent other Scandinavians, emigrated to Iceland, bringing with them thralls (i.e., slaves or serfs) of Gaelic origin.



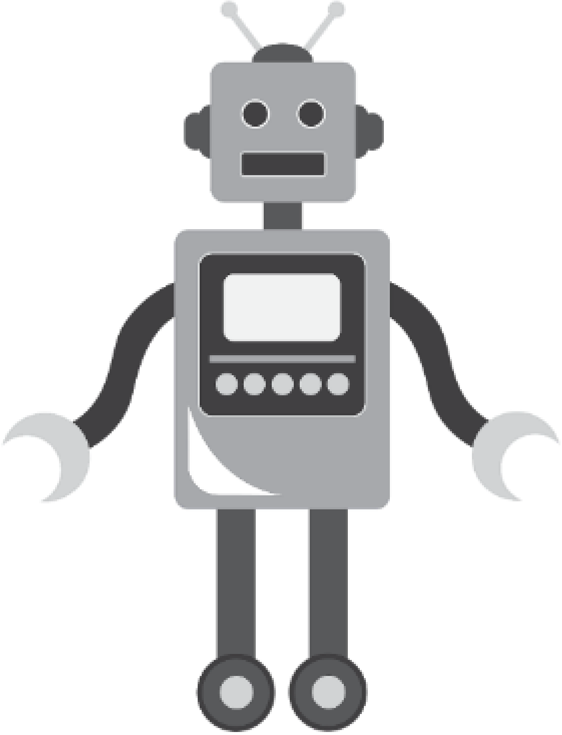


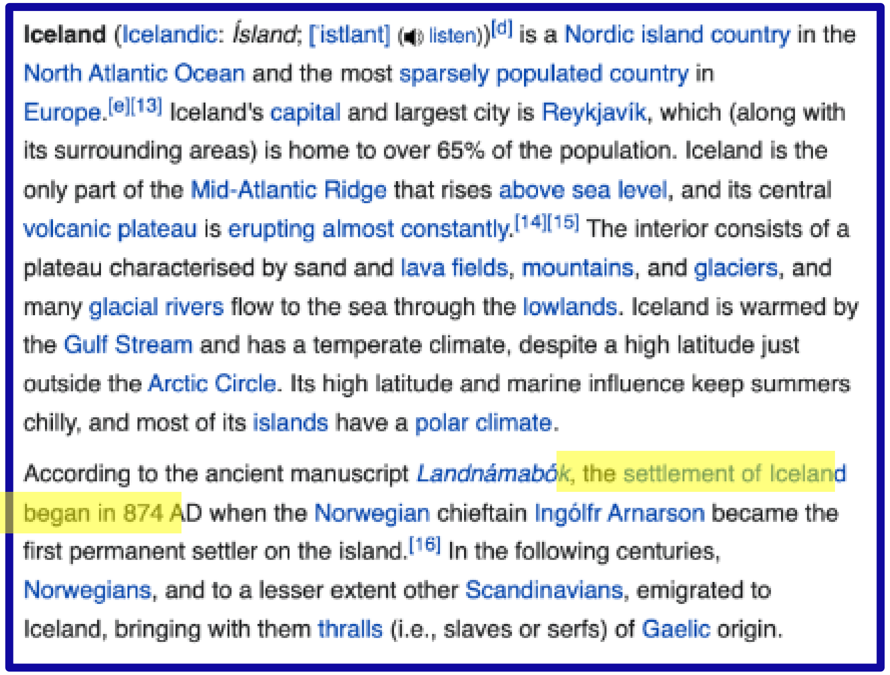
Iceland (Icelandic: *Ísland*; [ˈistlant] (🔈 listen))[d] is a Nordic island country in the North Atlantic Ocean and the most sparsely populated country in Europe.[e][13] Iceland's capital and largest city is Reykjavík, which (along with its surrounding areas) is home to over 65% of the population. Iceland is the only part of the Mid-Atlantic Ridge that rises above sea level, and its central volcanic plateau is erupting almost constantly.[14][15] The interior consists of a plateau characterised by sand and lava fields, mountains, and glaciers, and many glacial rivers flow to the sea through the lowlands. Iceland is warmed by the Gulf Stream and has a temperate climate, despite a high latitude just outside the Arctic Circle. Its high latitude and marine influence keep summers chilly, and most of its islands have a polar climate.

According to the ancient manuscript *Landnámabók*, the settlement of Iceland began in 874 AD when the Norwegian chieftain Ingólfr Arnarson became the first permanent settler on the island.[16] In the following centuries, Norwegians, and to a lesser extent other Scandinavians, emigrated to Iceland, bringing with them thralls (i.e., slaves or serfs) of Gaelic origin.

Open QA

When did the settlement of Iceland begin?





Iceland (Icelandic: *Ísland*; [ˈistlant] (listen))[d] is a Nordic island country in the North Atlantic Ocean and the most sparsely populated country in Europe.[e][13] Iceland's capital and largest city is Reykjavík, which (along with its surrounding areas) is home to over 65% of the population. Iceland is the only part of the Mid-Atlantic Ridge that rises above sea level, and its central volcanic plateau is erupting almost constantly.[14][15] The interior consists of a plateau characterised by sand and lava fields, mountains, and glaciers, and many glacial rivers flow to the sea through the lowlands. Iceland is warmed by the Gulf Stream and has a temperate climate, despite a high latitude just outside the Arctic Circle. Its high latitude and marine influence keep summers chilly, and most of its islands have a polar climate.

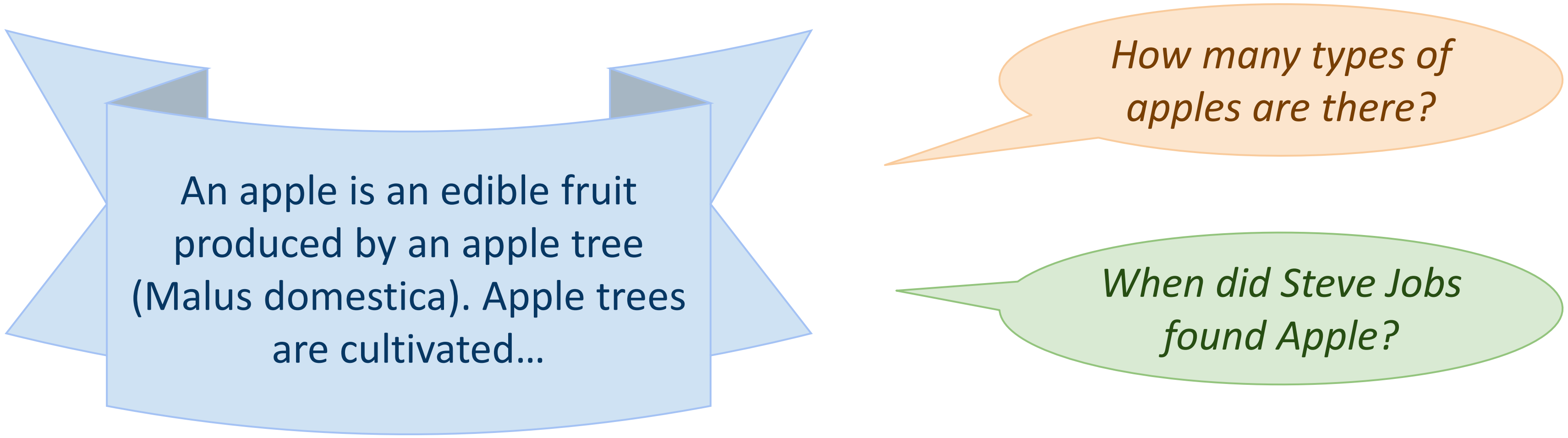
According to the ancient manuscript *Landnámabók*, the settlement of Iceland began in 874 AD when the Norwegian chieftain Ingólfr Arnarson became the first permanent settler on the island.[16] In the following centuries, Norwegians, and to a lesser extent other Scandinavians, emigrated to Iceland, bringing with them thralls (i.e., slaves or serfs) of Gaelic origin.

Creating a dataset for QA in Icelandic

We applied an existing method to create QA datasets such that the task would not be too easy (Clark et al., 2020).

1

Question elicitation



An apple is an edible fruit produced by an apple tree (Malus domestica). Apple trees are cultivated…

How many types of apples are there?

When did Steve Jobs found Apple?

13,740 questions were created

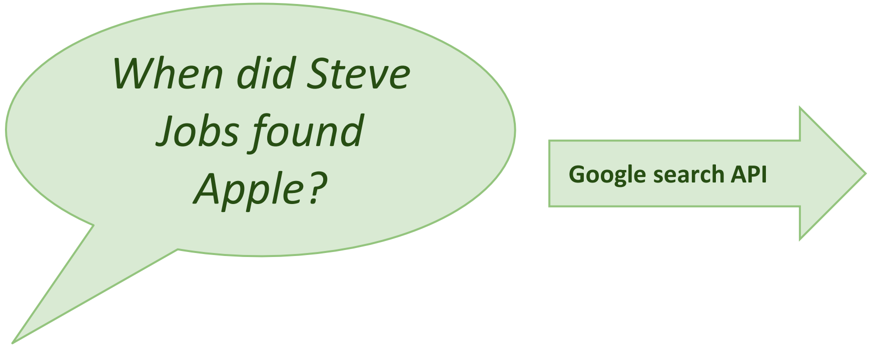
Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, *8*, 454-470.

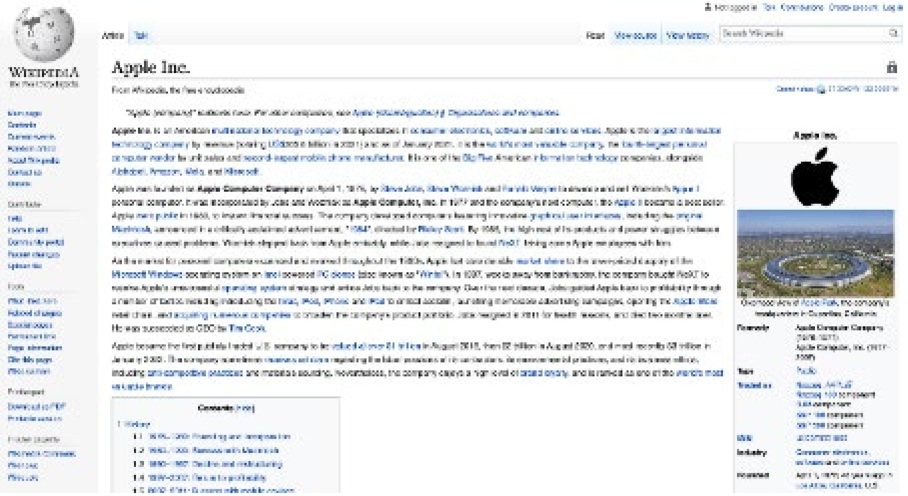
Creating a dataset for QA in Icelandic

We applied an existing method to create QA datasets such that the task would not be too easy (Clark et al., 2020).

2

Article retrieval





Apple Inc.

[illegible]

Article found for 9,060 questions (65.9%)

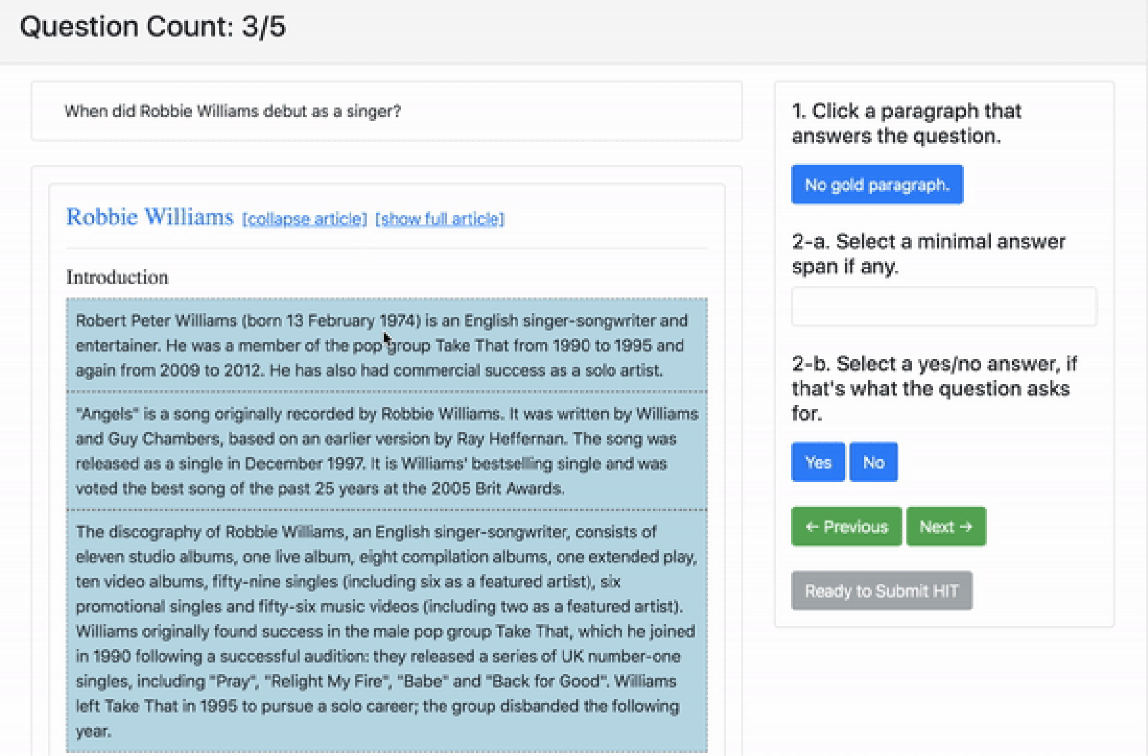
Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, *8*, 454-470.

Creating a dataset for QA in Icelandic

We applied an existing method to create QA datasets such that the task would not be too easy (Clark et al., 2020).

3

Answer labelling



Question Count: 3/5

When did Robbie Williams debut as a singer?

Robbie Williams [collapse article] [show full article]

Introduction

Robert Peter Williams (born 13 February 1974) is an English singer-songwriter and entertainer. He was a member of the pop group Take That from 1990 to 1995 and again from 2009 to 2012. He has also had commercial success as a solo artist.

"Angels" is a song originally recorded by Robbie Williams. It was written by Williams and Guy Chambers, based on an earlier version by Ray Heffernan. The song was released as a single in December 1997. It is Williams' bestselling single and was voted the best song of the past 25 years at the 2005 Brit Awards.

The discography of Robbie Williams, an English singer-songwriter, consists of eleven studio albums, one live album, eight compilation albums, one extended play, ten video albums, fifty-nine singles (including six as a featured artist), six promotional singles and fifty-six music videos (including two as a featured artist). Williams originally found success in the male pop group Take That, which he joined in 1990 following a successful audition: they released a series of UK number-one singles, including "Pray", "Relight My Fire", "Babe" and "Back for Good". Williams left Take That in 1995 to pursue a solo career; the group disbanded the following year.

1. Click a paragraph that answers the question.

No gold paragraph.

2-a. Select a minimal answer span if any.

2-b. Select a yes/no answer, if that's what the question asks for.

Yes No

← Previous Next →

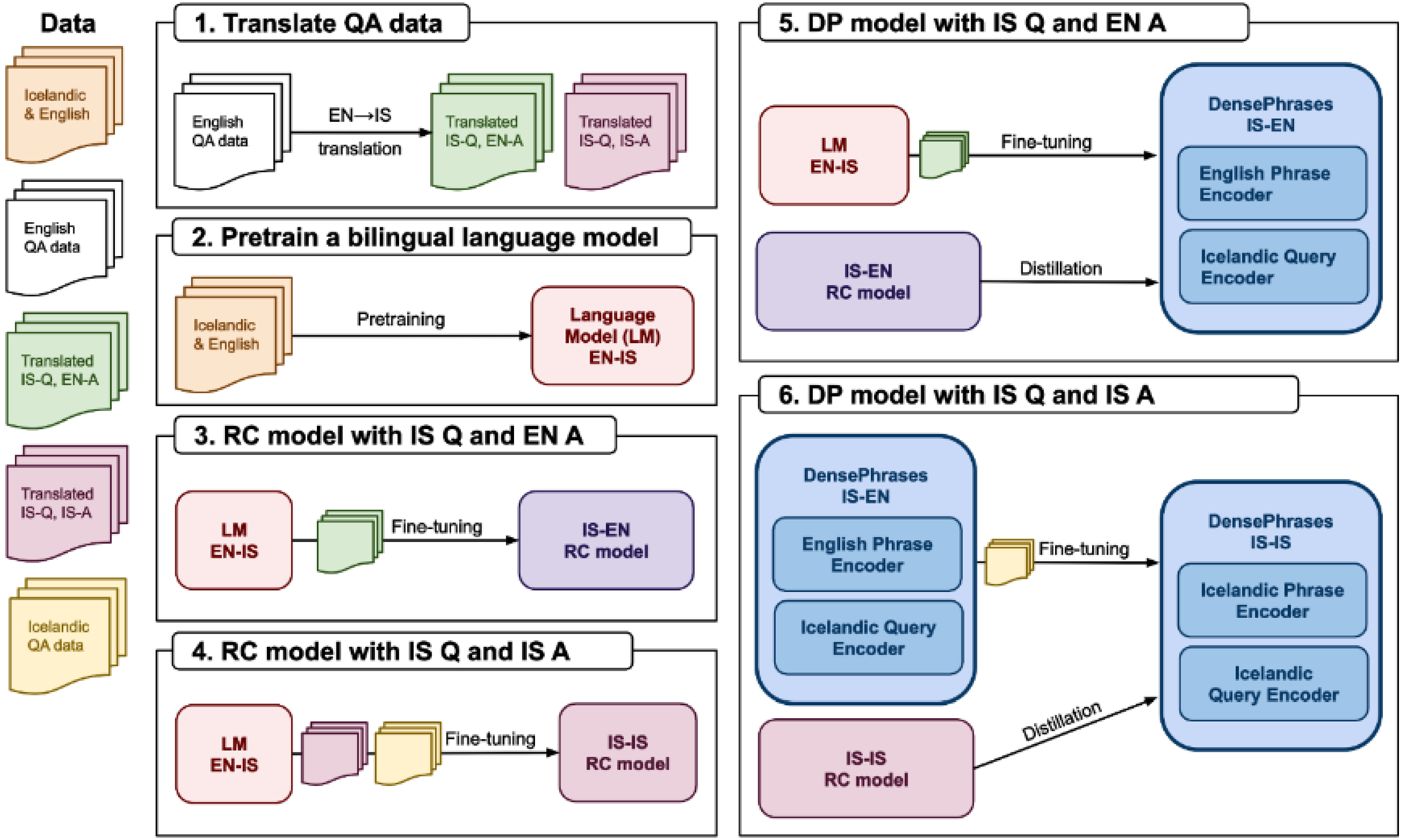
Ready to Submit HIT

18,378 labelled question-passage pairs from 1,400 unique Wikipedia articles

Answer found in 5,405 cases

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, *8*, 454-470.

Model development

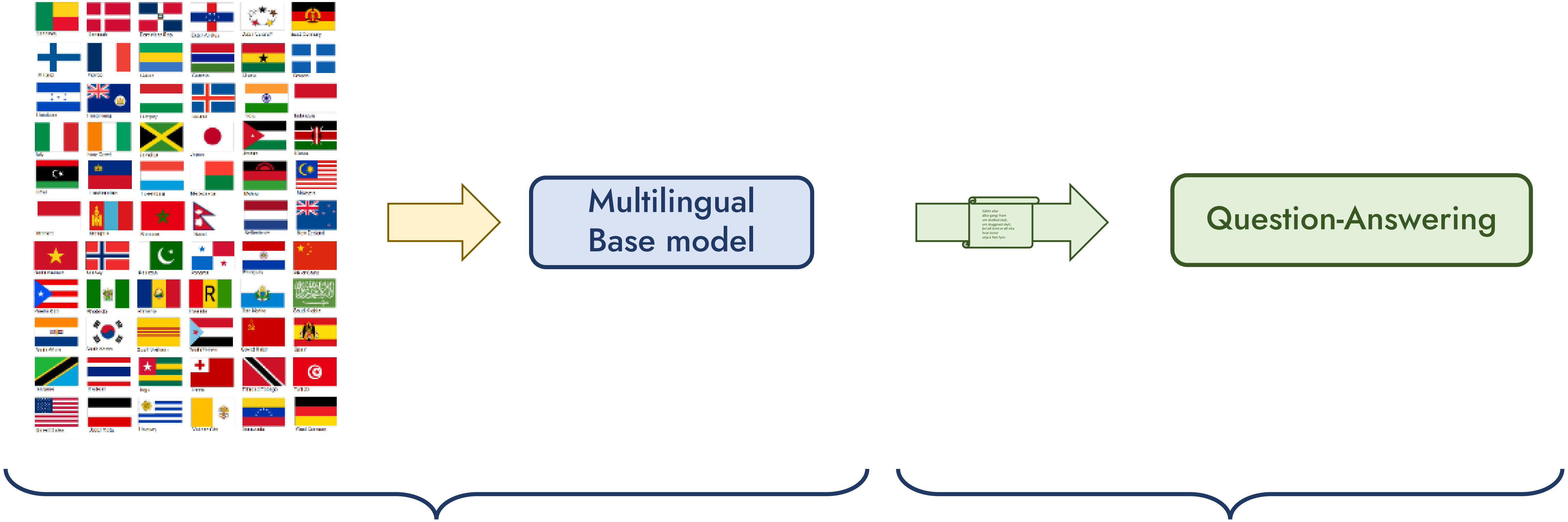


Evaluation

The method had an F1 score of 18.8 (out of 100) on the test dataset.

The model's answer matched the annotators answer exactly in 9.7% of cases.

Crosslingual transfer



Multilingual Base model

Gáttir allar
áður gangi fram
um skoðast skyli,
um skyggnast skyli,
því að óvíst er að vita
hvar óvinir
sitja á fleti fyrir.

Question-Answering

Pre-training

Fine-tuning

Evaluation of crosslingual transfer

Our method reached an F1 score of 18.8.

We also evaluated a new multilingual method (Asai et al., 2021)

- The base model was trained in a multilingual manner (with Icelandic included)
- The model was not fine-tuned on Icelandic
- That method reached an F1 score of 28.6 (out of 100)!
- Exact match with annotator in 15% of cases.

Asai, A., Yu, X., Kasai, J., & Hajishirzi, H. (2021). One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, *34*.

Thank you!

UNIVERSITY OF ICELAND

The Icelandic word for computer is tölva

Use existing multilingual model as a starting point

Use cleaned text from the internet to train your models

Growing Wikipedia in your language has many benefits

Evaluation datasets are necessary

Crosslingual transfer might save us, eventually