

Málþing Íslenskrar málnefndar
Enska í íslensku samfélagi
4. maí 2023

Nýjasta tækni og málvísindi

Einar Freyr Sigurðsson og Iris Edda Nowenstein



STOFNUN ÁRNA MAGNÚSSONAR
í íslenskum fræðum



**HÁSKÓLI
ÍSLANDS**

Hið endalausa kapphlaup við ensku

Getum við haldið í við ensku þegar kemur (mál)tækniframförum og stafrænu málumhverfi?

- Eða verður alltaf of lítið af íslensku í framtíðinni?
- Bæði í tækjunum okkar og málumhverfi barna og annarsmálshafa?

Alls kyns hindranir þegar kemur að því **að fá að velja íslensku**, stundum er:

- (Mál)tæknin ekki til staðar
 - Máltækniáætlun íslenskra stjórnvalda
- Framboðið ekki til staðar
 - Sjá t.d. umræðuna um Disney+
- Viljinn ekki til staðar
 - Vegna ofangreindra hindrana?

→ eigum við þá að gefast upp?

Til umræðu í dag: Nýjasta (mál)tæknin sem náð hefur flugi

- Stór mállíkön (e. *Large Language Models, LMM*) á borð við GPT-4
→ tæknin að baki ChatGPT frá OpenAI

Þrjár ástæður fyrir sérstakri umfjöllun:

- Mikill og hraður árangur sem þó er umdeildur
 - Gríðarlega öflugt tæki en veruleg hættu á misnotkun og skortur á reglugerðum
- **Nú þegar áberandi umræða um hvað stór mállíkön geta sagt okkur um málvísindi**
 - Læra börn mál eins og tölvur? Eða kannski ekki síður: Læra tölvur mál eins og börn?
- Áhersla á að hægt verði að nota gervigreind á íslensku
 - „Framtíðin svarar á íslensku“ - er það hægt?

Mikill en umdeildur árangur

Umfangsmikil umræða: Stór mállíkön og gervigreind nýttast í magnaða hluti en:

- Líka sköpun falsfrétta og upplýsingaóreiðu
 - Hlutverk markaðssetningar
- Það er skortur á gagnsæi, eftirliti og löggjöf
- Áhrif á umhverfið (Bender o.fl. 2021)

Önnur umræða: Hvað segja stór mállíkön okkur um mannlegt mál og málvísindi?

Hætti hjá Google til að vara við hættu af gervigreind

Kjartan Kjartansson skrifar · 2. maí 2023 11:55

Why Are We Letting the AI Crisis Just Happen?

Bad actors could seize on large language models to engineer falsehoods at unprecedented scale.

AI makes non-invasive mind-reading possible by turning thoughts into text

Advance raises prospect of new ways to restore speech in those struggling to communicate due to stroke or motor neurone disease

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

ARTIFICIAL INTELLIGENCE | MAR. 1, 2023

You Are Not a Parrot
And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this.

By Elizabeth Weil, a features writer at New York

Stór mállíkön og málvísindi

GPT-4 þykir hafa svo gott vald á ensku (þ.e. mannlegu máli) að innan málvísinda er talað um dauða kenninga Chomskys (sbr. Piantadosi 2023 vs. Katzir 2023)

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi^{a,b}

^aUC Berkeley, Psychology ^bHelen Wills Neuroscience Institute

Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023).

Roni Katzir

April 2023

What, if Anything, Can Large Language Models Teach Us About Human Language Acquisition?

Tal Linzen

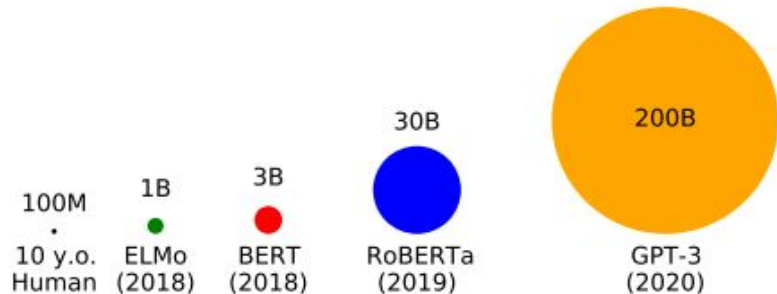
Læra tölvur málið eins og börn?

Mjög miklar rannsóknir á barnamáli

- Ílagið gegnir þar lykilhlutverki

Þess vegna er mjög mikilvægt að við getum líka rannsakað ílagið sem tölvurnar fá

- Það verður væntanlega hægt í náninni framtíð
- Nú er unnið að margvíslegum stórum mállíkönum og gervigreindarlausnum; sumt með opnum leyfum (Llama, Alpaca, Vicuna, Dolly, HuggingChat, GPT4All ...)
- En enn sem komið er er GPT-4 langbest
- ... og við höfum ekki aðgang að gögnunum sem líkanið er þjálfað á, þau eru óþekkt



Samanburður á stærð ílags (fjöldi mældur í tókum). Mynd frá Warstadt og Bowman (2022).

Stór mállíkön og málvísindi

Sleggjudómar um málvísindakenningar væru ótímabærir ef við miðuðum aðeins við íslenska virkni GPT-4 enda ílagið svo miklu minna

- En þó væntanlega nú þegar meira en t.a.m. börn hafa venjulega fengið við 10 ára aldur

Við getum t.d. borið saman „villur“ hjá GPT-4 og börnum

GPT-4 og íslenskan: Ýmis dæmi

Skoðum á næstu glærum ýmis íslensk málfræðidæmi þar sem við athuguðum m.a. „máltilfinningu“ GPT-4 (sjá einnig t.d. Þórhall Eypórsson 2023, Miðeind 2023 / <https://mideind.is/gpt.html>, Elmar Unnsteinsson 2023 / <https://twitter.com/unnsteinsson/status/1638142376991420417>)

GPT-4 og íslenskan



Er setningin "Það var borðað matinn" eðlilegt mál?



Já, setningin "Það var borðað matinn" er eðlilegt íslenskt mál. Setningin notar þolmynd (passive voice) og gefur til kynna að matnum hafi verið borðaður án þess að nefna hver gerði það. Þetta er algengt í íslenskri málfarsskrá og er notað til að leggja áherslu á atvikið (að borða matinn) frekar en þann sem gerir það (einhver borðandi).

GPT-4 hittir naglann á höfuðið



Er setningin "Það er alltaf borðað morgunmat nakinn" eðlilegt mál?



Setningin "Það er alltaf borðað morgunmat nakinn" er ekki alveg eðlilegt mál, því hún skortir einhverja tengingu við einstaklinginn sem er að borða morgunmatinn. Til að gera setninguna skýrari og eðlilegri, gætir þú orðað hana þannig:

"Ég borða alltaf morgunmatinn nakinn" eða "Hann/hún borðar alltaf morgunmatinn nakinn/nakin", eftir því kven er um að ræða.

GPT-4 og íslenskan: Villur sem börn gera líklega ekki



Hefurðu lesið Ofsa?



"Ofsa" er skáldsaga eftir íslenska höfundinn Arnaldur Indriðason, sem kom út árið 2004. Bókin er hluti af röð skáldsagna sem fjalla um lögreglumanninn Erlendur Sveinsson. "Ofsa" skartar spennuþræði, þar sem Erlendur, Elínborg og Sigurður Óli reyna að leysa mál um morð sem gerðu ráð fyrir óvenjulegum og flóknum aðferðum.

GPT-4 og íslenskan: Villur sem börn gera líklega ekki

7 ára barn spurt í morgun (4. maí 2023):

A: B, hefur þú lesið Ofsa?

B: Hvað er það?

A: Hvað er hvað?

B: Hvað er þessi Ofsi?

GPT-4 og íslenskan: Villur sem börn gera líklega ekki

GPT-4 bjó til liðinn *vegna kjarasamningsbundinna launahækkanir*

- Jafnvel þótt börn geti notað þolfall á andlagi forsetningarinnar *vegna* er óeðlilegt að það kæmi misræmi innan nafnliðarins þannig að lýsingarorðið væri í eignarfalli en nafnorðið í nefnifalli/þolfalli

GPT-4 og íslenskan: Villur sem börn gera líklega ekki



Er setningin "Launavísitalan hækkaði um 0,9% í mars, aðallega vegna kjarasamningsbundinna launahækkanir Eflingarfélaga." eðlilegt mál? Er liðurinn "vegna kjarasamningsbundinna launahækkanir Eflingarfélaga" eðlilegt mál?



Já, setningin "Launavísitalan hækkaði um 0,9% í mars, aðallega vegna kjarasamningsbundinna launahækkanir Eflingarfélaga." er eðlilegt mál og gagnlegt til að lýsa því hvernig launavísitala hefur hækkað yfir ákveðinn tíma.

Liðurinn "vegna kjarasamningsbundinna launahækkanir Eflingarfélaga" er einnig eðlilegt mál. Hann lýsir því að aðalástæðan fyrir hækkun launavísitölu er vegna launahækkanir sem eru skilgreindar í kjarasamningum milli Eflingarfélaga og atvinnurekendur.

GPT-4 og íslenskan: Villur sem börn gera ekki

Tilvísunarsetningar vefjast ekki fyrir börnum

Í þeim er eyða sem samsvarar orði eða lið í móðursetningu:

- Manneskjan sem ég hélt að þú hefðir séð Guðmund tala við ____ í gær er farin heim til sín.

Við getum ekki fyllt upp í eyðuna (hér *Sigurð*):

- Manneskjan sem ég hélt að þú hefðir séð Guðmund tala við **Sigurð** í gær er farin heim til sín.

En þetta vefst fyrir GPT-4! (líka í ensku, sbr. Katzir 2023)

GPT-4 og íslenskan: Villur sem börn gera ekki



Geturðu sagt mér hvort eftirfarandi setning sé eðlilegt mál?

Manneskjan sem ég hélt að þú hefðir séð Guðmund tala við Sigurð í gær er farin heim til sín.



Já, setningin er eðlilegt mál. Hér er hennar merking:

Manneskjan (sá eða sú sem þú hefur áhuga á eða vilt vita af hverju hefur farið heim) sem ég hélt að þú hefðir séð (ég hélt að þú hefðir séð atvikið eða samræðuna) Guðmund tala við Sigurð í gær (atvik eða samræða sem átti sér stað í gær) er farin heim til sín (manneskjan er komin heim, ferli hennar er lokið eða henni langar heim).

Máltökuskeið barna og GPT-4

Ekki eins góður árangur og í ensku ennþá (bæði innihald og form)
en mjög mikill þó!

Verður alltaf of lítið af íslensku ílagi?

Hversu mikið íslenskt ílag þarf í raun og veru samanborið við börn?

Málskilningur og málmyndun GPT-4 bendir til að það sé talsverður munur á
málfræði þess og málfræði barna á máltökuskeiði

Ómögulegt að sigra? Skiptir samt höfuðmáli að vera með!

Hægt að ná verulegum árangri í íslenskri máltækni þótt enskan muni sennilega alltaf hafa forskot

- Nýjasta tæknin nýtist nú þegar almenningi (t.d. <https://greynir.is/news>) og getur sagt okkur ýmislegt um íslensk málvísindi
 - Þurfum samt að hafa varann á!

Efling íslensku er þó ekki háð því að enskan (eða önnur mál) sé sett upp sem ógn

- Enskan er ekki alltaf á kostnað íslenskunnar og jákvæð viðhorf til tungumála skipta máli (Iris Edda Nowenstein og Sigríður Sigurjónsdóttir 2021)
- Enskan getur verið mikilvægt verkfæri í íslensku (mál)samfélagi, t.d. þegar kemur að hagsmunum og mannréttindum hópa sem hafa ekki vald á íslensku
 - mikilvægt að átta sig á mállegum forréttindum og valdaójafnvægi (sbr. einnig umræðu um bjaga í mállíkönum, Bender o.fl. 2021, Agnes Sólmundsdóttir 2021)
 - Íslenska sem val og valdefling





Geturðu sagt mér hvort eftirfarandi setningar séu eðlilegt mál?

- 1) María trúir ekki þeim sögusögnum að lögreglan hafi handtekið manninn.
- 2) Hvaða mann trúir María ekki þeim sögusögnum að lögreglan hafi handtekið?
- 3) Hversu mörgum borgum á María bræður sem búa í?
- 4) María á bræður sem búa í hversu mörgum borgum?
- 5) Hverjum kyssti María hundinn sinn þegar hún beið eftir?



1. María trúir ekki þeim sögusögnum að lögreglan hafi handtekið manninn. (Eðlilegt mál)
2. Hvaða mann trúir María ekki þeim sögusögnum að lögreglan hafi handtekið? (Eðlilegt mál)
3. Hversu mörgum borgum á María bræður sem búa í? (Ekki eðlilegt mál, rétt mynd væri: "Í hversu mörgum borgum á María bræður sem búa?")
4. María á bræður sem búa í hversu mörgum borgum? (Eðlilegt mál)
5. Hverjum kyssti María hundinn sinn þegar hún beið eftir? (Ekki eðlilegt mál, rétt mynd væri: "Hverjum kyssti María hundinn sinn þegar hún beið eftir honum/henni?")

GPT-4 og íslenskan: Fleiri dæmi



Er setningin "Það má ekki borða morgunmatinn nakinn" eðlilegt mál?



Já, setningin "Það má ekki borða morgunmatinn nakinn" er eðlilegt íslenskt mál. Setningin gefur til kynna að það sé óheimilt eða óviðeigandi að borða morgunmatinn án fötum.

GPT-4 og íslenskan: Fleiri dæmi



Hvert er nefnifallið af "Ofsa"?



Nefnifall orðsins "ofsa" er "ofsi". Orðið er karlkyns og er með því regluleg beyging: nefnifall "ofsi", þolfall "ofsa", þágufall "ofsa" og eignarfall "ofsans".

GPT-4 og íslenskan: Fleiri dæmi



Hvernig er "ofsi" með og án greinis í öllum föllum og báðum tölum?



Hér er beyging orðsins "ofsi" með og án greinis í öllum föllum og báðum tölum:

Eintala:

- Nefnifall: ofsi, ofsið
- Polfall: ofsa, ofsann
- Þágufall: ofsa, ofsnum
- Eignarfall: ofsans, ofsansins

Fleirtala:

- Nefnifall: ofsar, ofsarnir
- Polfall: ofsa, ofsana
- Þágufall: ofsum, ofsunum
- Eignarfall: ofsa, ofsanna

GPT-4 og íslenskan: Fleiri dæmi



Hvert er nefnifallið af "launahækkunar"?



Nefnifall orðsins "launahækkunar" er "launahækkunar" í eintölu og "launahækkunar" í fleirtölu. Orðið er samsett úr "laun" (salary/wage) og "hækkunar" (increases), sem merkir að það vísar til nokkurs konar hækkaðra launa eða launahækkana.



Hvernig er "launahækkanir" með og án greinis í öllum föllum og báðum tölum?



Hér er beyging orðsins "launahækkanir" með og án greinis í öllum föllum og báðum tölum:

Eintala:

- Nefnifall (Nominative): launahækkan (án greinis), launahækkanin (með greini)
- Polfall (Accusative): launahækkan (án greinis), launahækkunina (með greini)
- Þágufall (Dative): launahækkan (án greinis), launahækkaninni (með greini)
- Eignarfall (Genitive): launahækkunar (án greinis), launahækkunarinnar (með greini)

Fleirtala:

- Nefnifall (Nominative): launahækkanir (án greinis), launahækkanirnar (með greini)
- Polfall (Accusative): launahækkanir (án greinis), launahækkanirnar (með greini)
- Þágufall (Dative): launahækkunum (án greinis), launahækkunum (með greini)
- Eignarfall (Genitive): launahækkananna (án greinis), launahækkananna (með greini)

GPT-4 og íslenskan: Fleiri dæmi



Hvert er nefnifallið af "hækkanir"?



Nefnifall orðsins "hækkanir" er "hækkanir" í fleirtölu. Það er lýsingarháttur þátíðar af sögninni "að hækka" og merkir nokkrar hækkanir eða aukningar.



Hvernig er "hækkanir" með og án greinis í öllum föllum og báðum tölum?



Hér er beyging "hækkanir" með og án greinis í öllum föllum og báðum tölum:

Eintala:

- Nefnifall án greinis: hækkan
- Nefnifall með greini: hækkanin
- Þágufall án greinis: hækkan
- Þágufall með greini: hækkanina
- Polfall án greinis: hækkan
- Polfall með greini: hækkanina
- Eignarfall án greinis: hækkunar
- Eignarfall með greini: hækkunarinnar

Fleirtala:

- Nefnifall án greinis: hækkanir
- Nefnifall með greini: hækkanirnar
- Þágufall án greinis: hækkanir
- Þágufall með greini: hækkanirnar
- Polfall án greinis: hækkanir
- Polfall með greini: hækkanirnar
- Eignarfall án greinis: hækkana
- Eignarfall með greini: hækkanna

GPT-4 og íslenskan: Fleiri dæmi



Hvert er nefnifallið af "kjarasamningsbundinn"?



Nefnifallið af "kjarasamningsbundinn" er "kjarasamningsbundinn" því þetta er lýsingarháttur þátíðar af lýsingarháttur þátíðar lýsingarhætti þátíðar samsettu lýsingarorðsins. Hér er þó að athuga að "kjarasamningsbundinn" er einnig samsett orð, sem er myndað úr "kjarasamningur" og "bundinn".



Hvernig er "kjarasamningsbundinn" í kvenkyni í öllum föllum og báðum tölum?



Hér er beyging "kjarasamningsbundinn" í kvenkyni í öllum föllum og báðum tölum:

Eintala:

- Nefnifall: kjarasamningsbundin
- Polfall: kjarasamningsbundna
- Þágufall: kjarasamningsbundinni
- Eignarfall: kjarasamningsbundinnar

Fleirtala:

- Nefnifall: kjarasamningsbundnar
- Polfall: kjarasamningsbundnar
- Þágufall: kjarasamningsbundnum
- Eignarfall: kjarasamningsbundinna

Dæmi um setningar með orðið "kjarasamningsbundinn" í kvenkyni:

- Konan er kjarasamningsbundin.
- Hann sá kjarasamningsbundna konu.
- Hann tók þátt í málum kjarasamningsbundinnar konu.
- Hann gaf blómin kjarasamningsbundinni konu.

GPT-4 og íslenskan: Fleiri dæmi



Er liðurinn "vegna kjarasamningsbundinna launahækkana Eflingarfélaga" eðlilegt mál?



Já, liðurinn "vegna kjarasamningsbundinna launahækkana Eflingarfélaga" er eðlilegt mál. Hann lýsir því að aðalástæðan fyrir hækkun launavísitölu eða öðrum breytingum á launum er vegna launahækkana sem eru skilgreindar í kjarasamningum milli Eflingarfélaga og atvinnurekendur.