Gögn: lífæð máltækni og gervigreindar

Allar máltæknilausnir, hvort sem það eru spjallmenni, þýðingarvélar, talgervlar, talgreinar eða málrýnihugbúnaður, byggjast á miklu magni fjölbreyttra málgagna. Og máltæknin er að breyta lífi okkar og samfélögum um þessar mundir. Spurningin er löngu hætt að snúast um það hvort við eigum að þróa og nota máltækni. Í stað þess snýr hún að því hvernig landslag er að teiknast upp í máltækni og gervigreind. Er það tækni sem talar íslensku? Er það tækni sem hentar alls konar fólki við mismunandi aðstæður og kemur til móts við þarfir manneskjunnar? Eða er það einsleit tækni sem þvingar notendur til þess að aðlaga mál sitt og jafnvel suma þætti daglegs lífs að tækninni? Er það tækni sem er útilokandi eða inngildandi? Svörin við þessum spurningum velta að miklu leyti á því hvaða gögn eru til reiðu til þess að þróa tæknilausnir.

Máltækni fyrir íslensku þarf að þjálfa á íslenskum málgögnum. Þó að við njótum góðs af þróun algríma og þjálfunaraðferða, og einnig af grunnlíkönum sem þjálfuð eru á öðrum tungumálum, þá talar og skilur tæknin ekki íslensku nema hafa séð íslensk málgögn.

Eftir því sem tæknin að baki máltæknilausnum þróast og notkun þeirra verður útbreiddari, því meiri kröfur verðum við að gera til þess að lausnir henti notendum og aðstæðum hverju sinni. Við viljum að tæknin skilji unga sem aldna, konur, kvár og karla, Vestfirðinga, Norðlendinga, þau sem tala íslensku sem annað mál o.s.frv. Hún þarf að skilja formlegt mál og óformlegt, skapandi skrif og skýrslur, skilja stök orð og stuttar yrðingar í samtali og samhengi í löngum textum. Það þarf að skilja meira en orðanna hljóðan – það getur til dæmis verið verkefni að meta hvort texti eða tal lýsa ánægju eða óánægju, í hvaða samhengi eitthvað stendur, til dæmis varðandi söguna, fræðasvið, stjórnmálaskoðanir eða annað. Sömuleiðis viljum við að tæknin getið „tjáð sig“ á fjölbreyttan hátt en skili ekki sama málsniði í öllum aðstæðum.

Þekkjum gögnin

Við þurfum að vanda okkur sérstaklega vel þegar við látum tæknina taka ákvarðanir fyrir okkur á grundvelli greiningar á gögnum sem hún hefur yfir að ráða. Gögnin geta innihaldið bjaga sem endurspegla til dæmis ráðandi umræðu eða hlutdrægar ákvarðanir sem teknar hafa verið í gegnum tíðina t.d. varðandi ráðningar í störf. Við þurfum að þekkja gögnin sem við notum og vera meðvituð um hvaða gildrur gætu leynst í þeim.

Til þróunar á máltækni sem getur fengist við talað og ritað mál þarf bæði tal- og textagögn, sem og myndbandsupptökur þar sem hægt er að greina samhengi milli tals og varahreyfinga auk andlits- og líkamstjáningar.

Gagnasöfn eru útbúin á mismunandi hátt eftir tilgangi þeirra. Stóru mállíkönin og nýjustu taltæknilíkön vinna með gríðarlegt magn gagna. Hér erum við að tala um þúsundir eða tugþúsundir milljarða tóka af textum og hundruð þúsunda klukkustunda af tali. Til samanburðar inniheldur flaggskip íslenskra gagnasafna, Risamálheildin, um 2,5 milljarða orða, líklega um einn tvö þúsundasta af þjálfunargögnum stóru mállíkananna.

Með samstilltu átaki allra, það er opinbera geirans og einkafyrirtækja, almennings og sjálfstæðra tal- og textaframleiðenda, mætti þó lyfta grettistaki í því að færa íslenska máltækni upp á næsta stig, stig þar sem skortur á gögnum er ekki helsti akkillesarhællinn. Tæknin er til en hún er gagnslítil án gagna.

Á undanförnum árum hefur á Íslandi byggst upp góð þekking á því að útbúa vönduð gagnasöfn þar sem gæði eru metin fram yfir magn. Vinnan hefur ekki aðeins snúið að því að fínþjálfa líkön til ákveðinna nota, til dæmis til þess að tala íslensku eða til þess að standa sig sérlega vel í málrýni, heldur einnig að því að prófa og meta gæði mismunandi lausna. Líkan sem stendur sig best í ákveðinni tegund verkefna er ekki endilega besta líkanið fyrir öll verkefni og allar aðstæður. Hér þarf ekki eingöngu að taka tillit til gæða útkomu heldur þarf oft að líta til þátta eins og hraða og hagkvæmni. Það er því nauðsynlegt að hafa aðferðir til þess að meta gagn og gæði mismunandi lausna á mismunandi verkefnum.

Tæknilegt sjálfstæði

Í samhengi tækninnar erum við ekki eyja. Mikil áhersla hefur verið lögð á það undanfarin ár að stóru tæknifyrirtækin í Bandaríkjunum hugi að íslensku og öðrum smærri tungumálum í sinni tækni, en við þurfum einnig að líta mun meira til Evrópu. Það er í erfðamengi Evrópu að huga að fjölbreytni, þar sem til dæmis öll tungumál álfunnar eiga að fá sitt pláss. Á meðan hefur það verið aðalsmerki bandarísku fyrirtækjanna að huga fyrst og fremst að þeim tungumálum sem flestir tala eins og ensku, kínversku og spænsku. Íslenska hefur þó fengið mjög mikið pláss miðað við höfðatölu, þökk sé þeirri vinnu sem farið hefur fram hér heima bæði á tækni- og samskiptasviðinu.

Við þurfum að huga að því að vera ekki of háð stórfyrirtækjunum og vinna í sameiningu að því að efla íslenska og evrópska máltækni sem getur þjónað okkar málsvæðum á eigin forsendum. Það er strategísk ákvörðun að halda sjálf um taumana, ekki bara elta heldur að vera í fararbroddi í því að þróa tækni í hæsta gæðaflokki okkur til hagsbóta. Og ég veit að ég er farin að endurtaka mig: grunnurinn liggur í gögnum.

Við þurfum að byggja upp ferla og innviði til þess að sem flestir sjái hag sinn í því, eða a.m.k. að það sé ekki íþyngjandi, að leggja til gögn til þróunar íslenskrar máltækni. Grunnurinn hefur verið lagður gegnum íslensku máltækniáætlunina. Sú vinna skilaði mikilli þekkingu á þessu sviði og við höfum nú reynslu af tilreiðingu fjölbreyttra gagnasafna.

Síðan þarf að setja vélina í gang, ef svo má segja, og halda áfram að safna. Lífæðin þarf stöðugt að renna. Úrelt gögn er sama og úrelt tækni. Ég hvet þig, lesandi góður, til þess að skoða hvaða gögn á þínum vegum gætu komið að gagni í því verkefni að stuðla að lifandi, manneskjuvænni tækni til framtíðar.

Hvernig getur þú eða þitt fyrirtæki lagt til gögn?

Almannarómur hefur sett upp *Heimildagátt* til þess að taka við textagögnum frá fyrirtækjum: https://almannaromur.is/thin-islenska-er-malid

Árnastofnun og Grammatek vinna um þessar mundir að umfangsmikilli söfnun talgagna, til dæmis úr hlaðvörpum og öðru fjölmiðlaefni. Tengiliður hjá Grammateki er Anna Björk Nikulásdóttir (anna@grammatek.com) og hjá Árnastofnun Einar Freyr Sigurðsson

(einar.freyr.sigurdsson@arnastofnun.is), sem auk þess getur tekið við hvers kyns fyrirspurnum varðandi málgögn.

Til upplýsingar

Unnið upp úr fyrirlestri á málþingi European Language Data Space (LDS), Veröld, hús Vigdísar, 9. október 2025.

Mynd

Anna Björk Nikulásdóttir.