30. janúar 2024

Skýrsla um tilraunir á sjálfvirkri flokkun á móðurmálstextum samkvæmt evrópska tungumálarammanum

Steinunn Rut Friðriksdóttir
srf2@hi.is

Verkefnið í hnotskurn

Í þessari skýrslu fjalla ég um þær tilraunir sem ég hef gert undanfarið í sambandi við sjálfvirka flokkun móðurmálstexta eftir hæfnistigum evrópska tungumálarammans (CEFR). Verkefnið hlaut styrk úr styrktarsjóði Áslaugar Hafliðadóttur að upphæð kr. 1.400.000 og var sú upphæð nýtt að fullu í þriggja mánaða launagreiðslur höfundar. Allur afrakstur verkefnisins sem fjallað er um hér má finna á Github. Eftirfarandi er stutt lýsing á hverjum verkþætti fyrir sig.

Python-skriftan **data_process.py** er einföld leið til þess að fara í gegnum öll skjöl í möppum sem merktar eru hverju hæfnistigi fyrir sig og útbúa úr þeim eitt, heildrænt skjal á sniðinu .csv (þ.e.a.s. *comma separated values*, eins konar Excel-snið sem tölvur eiga einfaldara með að vinna úr).

Python-skriftan **level_analyser.py** tekur .csv skjal sem inntak. Skjalið er með að lágmarki tveimur dálkum þar sem annar ber titilinn *text* og hinn *cefrlevel* og vísa þeir annars vegar til textana sem á að skoða og hins vegar til þess hæfnistigs sem viðkomandi texti fellur undir. Skriftan tekur til margra setningafræðilegra og beygingarfræðilegra þátta sem verður nánar fjallað um í kaflanum um þætti hér fyrir neðan. Skriftan veitir upplýsingar um það hversu hátt hlutfall einstakra þátta er að meðaltali í hverjum texta innan hvers hæfnistigs. Þessi greining er óháð einstökum líkönum sem kunna að vera valin við sjálfvirka flokkun textana og ætti því að veita innsýn inn í málfræðileg einkenni hvers hæfnistigs og þar með einfalda þá þáttagreiningu (e. *feature engineering*) sem þarf til þess að þjálfa sjálfvirkan flokkara.

Python-skriftan **gradientboosting.py** nýtir þá þætti sem voru skoðaðir í fyrrnefndri skriftu til þess að smíða þáttavigra sem tákna hvern texta innan þeirrar málheildar sem liggur til grundvallar þegar sjálfvirkur flokkari er þjálfaður. Í þessari skriftu er gert ráð fyrir því að inntaksgögnin séu á fyrrnefndu .csv sniði og innihaldi að lágmarki dálkana *text* og *cefrlevel*. Flokkarinn sem er þjálfaður með þessari skriftu er gradient boosting líkan (sjá umfjöllun í kaflanum um flokkarann) og frammistaða hans er metin með því að reikna meðalnámkvæmni (e. *accuracy*) eftir tífalda krossprófun (e. *10-fold cross validation*).

Textarnir sem liggja til grundvallar í þessum tilraunum eru tvenns konar. Annars vegar þeir textar sem Kolfinna Jónatansdóttir var búin að greina samkvæmt hæfnistigunum A1–C1 á haustmánuðum 2023 auk þess sem ég bætti inn tólf textum sem ég mat sem svo að féllu undir stig C2. Þessir textar einkennast annað hvort af sérhæfðum orðaforða tengdum ákveðnu fræðasviði eða talsverðri notkun orðtaka og/eða málshátta. Þessari flokkun ber að taka með nokkrum fyrirvara þar sem henni er bætt aftan við flokkun

Kolfinnu sem hefur töluvert meiri sérhæfingu á sviði íslensku sem annars máls heldur en ég. Súlurit sem gefa til kynna greiningu þessara texta er að finna í viðauka A. Hins vegar er um að ræða texta sem ég hef safnað sjálf úr Risamálheildinni og flokkað eftir hæfnistigum samkvæmt tilfinningu minni fyrir einkennum þeirra og ber því líka að taka með talsverðum fyrirvara, enda hafa þessir textar ekki verið yfirfarnir af sérfræðingum á sviði íslensku sem annars máls. Einnig er vert að nefna að fjöldi texta eftir hæfnistigum er mjög ójafn enda er erfitt að nálgast texta á lægri stigunum í Risamálheildinni sem er að langmestu leyti skrifuð af móðurmálshöfum fyrir móðurmálshafa. Þeir textar sem þó finnast á lægri stigum eru fyrst og fremst textar úr barnabókum og því e.t.v. ekki vel til þess fallnir að nýta í kennslu fullorðinna.

Textarnir sem hafa verið flokkaðir eru úr bókaundirmálheild Risamálheildarinnar, Wikipedia, bleikt.is og fréttablaðið.is. Á hæfnistigi A1 fannst enginn texti og því er það stig ekki haft með í súluritum í viðauka B. Á hæfnistigi A2 eru átta textar, á B1 eru 23 textar, á B2 eru 76 textar, á C1 eru 166 textar og á C2 eru 82 textar. Þessari flokkun til grundvallar leit ég fyrst og fremst á málfræðilega þætti, sambærilega við þá sem eru ræddir í kaflanum um þætti hér fyrir neðan. Einnig leit ég til umræðuefnis textana, þ.e.a.s. hvort þeir væru sérhæfðir eða almennir, hvort þeir lýstu fyrst og fremst manneskju (og þá sérstaklega ef um er að ræða fyrstu persónu frásögn) og annað slíkt. Munurinn á C1 og C2 er ekki mjög skýr en þó reyndi ég að horfa til þess hvort textinn innihéldi mjög sérhæfðan orðaforða og/eða mikið væri um orðtök. Ef nýta á þessi gögn til grundvallar frekari þjálfunar á sjálfvirkum flokkara þarf nauðsynlega að fara yfir þau og endurmeta hvort flokkunin mín er rétt eða ekki. Þessir textar nýtast þó vel til samanburðar við gögnin frá Kolfinnu og því er þáttagreining þessarra tveggja textasafna borin saman í kaflanum um samanburð þáttagreiningar hér fyrir neðan.

Þættirnir

Við val á þáttum til að greina horfði ég helst til meistararitgerðar Isidoru Glišić[1], meistararitgerðar Gísla Hvanndals Ólafssonar[2] og upplýsingum frá Stefanie Bade varðandi einkenni sem eiga við texta á hæfnistigunum B1 og B2 en auk þess kannaði ég ýmsa þætti sem mér fannst áhugaverðir. Við þáttagreininguna nýtti ég eftirfarandi tól: POS-pakkann frá CADIA sem byggir á ABLTagger, markara sem Steinþór Steingrímsson, Örvar Kárason og Hrafn Loftsson gerðu og hefur m.a. verið nýttur til þess að marka Risamálheildina; setningaþáttara Greynis frá Miðeind til að greina ýmsar gerðir aukasetninga í textunum; mállíkanið IceBERT frá Miðeind til þess að greypa orðin

[1] IsidoraGlišić. (2023). *Towards Automated Icelandic Skill Level Evaluation: A Deep Dive into L2 Error Corpus patterns and Classification* [meistararitgerð við Háskóla Íslands]. Skemman. http://hdl.handle.net/1946/45776

[2] Gísli Hvanndal Ólafsson. (2016). *Grammar and Linguistic Structures at Level A1 of Icelandic* [meistararitgerð við Vrije Universiteit Brussel, óútgefin].

í textunum; og pakkann Scikit learn til þess að fá tf-idf-vigra fyrir textana. Þáttavigrunum er með öðrum orðum skeytt við orðgreypingar og tf-idf-vigra í tilraun til þess sem mest aðgreinandi upplýsingar um einkenni textana á hverju hæfnistigi.

Alls voru þrjátíu þættir greindir og má skipta þeim upp í fjóra flokka. Fyrsti flokkurinn varðar setningafræðilega greiningu, þ.e.a.s. athugun á meðaltalshlutfalli tiltekinna setningagerða í textum hvers hæfnisstigs. Auk þess að kanna hlutfall málsgreina sem innihalda aukasetningar af öllum eftirfarandi gerðum (þ.e.a.s. óháð gerð aukasetningarinnar) er hlutfall eftirfarandi setningagerða kannað: skýringarsetningar, spurnaraukasetningar, tilvísunarsetningar, tíðarsetningar, tilgangssetningar, viðurkenningarsetningar, afleiðingarsetningar, orsakarsetningar, skilyrðissetningar og samanburðarsetningar. Í öllum tilfellum er heildarfjöldi málsgreina sem fá greiningu með Greyni (það er rétt að árétta að ekki allar málsgreinar fá þáttun og eru þær þá ekki taldar með í heildarfjölda) borinn saman við málsgreinar sem innihalda þá setningagerð sem er til skoðunar. Ekki er tekið tillit til þess hvort málsgreinarnar innihaldi fleira en eitt tilvik af setningagerðinni.

Annar flokkurinn athugar meðalhlutfall tiltekinna orðflokka af heildarfjölda allra orða í textum hvers hæfnisstigs, þ.e.a.s. hlutfall lýsingarorða, nafnorða, sagnorða, persónufornafna og smáorða (athugið að smáorð eru hér orð sem falla undir mörkin c og a í markaskrá MÍM-GULL en undir c falla samtengingar og nafnháttarmerki og undir a falla atviksorð, forsetningar og upphrópanir. Þar sem atviksorð og forsetningar eru ekki aðgreind í markaskránni falla atviksorð hér líka undir smáorð).

Þriðji flokkurinn er málfræðilegs eðlis og miðar almennt við meðahlutfall viðkomandi þáttar af heildarfjölda orða af tilteknum orðflokki. Undir þennan flokk fellur eftirfarandi: hlutfall sagna í viðtengingarhætti af öllum sögnum; hlutfall sagna í miðmynd af öllum sögnum; hlutfall sagna í lýsingarhætti þátíðar af öllum sögnum (hér er hugmyndin að fiska eftir þolmynd); hlutfall lýsingarorða í efsta stigi og aukafalli af öllum lýsingarorðum; hlutfall nafnorða með greini og í aukafalli af öllum nafnorðum; hlutfall spurnarfornafna í aukafalli af öllum spurnarfornöfnum; hlutfall lýsingarorða í efsta stigi af öllum lýsingarorðum; hlutfall persónufornafna í fyrstu persónu af öllum persónufornöfnum; og hlutfall sagna í þátíð (ekki lýsingarhætti þátíðar) af öllum sögnum.

Fjórði og síðasti flokkurinn tekur á atriðum sem tengjast fremur orðaforða og/eða beygingarlegri færni lesenda en beinum málfræðilegum atriðum. Undir hann falla eftirfarandi þættir: meðalhlutfall orða sem eru lengri en 6 bókstafir af öllum orðum; meðalhlutfall einstakra orðmynda af öllum orðum (þ.e.a.s. í hversu mörgum tilfellum er um endurtekinn orðaforða að ræða); meðalhlutfall einstakra sagnbeyginga af öllum sögnum (þ.e.a.s. hversu mikil fjölbreytni er í sagnbeygingum innan textans);

meðalhlutfall einstakra nafnorðabeyginga af öllum nafnorðum; og meðalhlutfall einstakra lýsingarorðabeyginga af öllum lýsingarorðum.

Samanburður þáttagreiningar

Í viðauka A og B má sjá upplýsingar um þáttagreiningu textasafns Kolfinnu annars vegar og þeirra texta sem ég safnaði úr Risamálheildinni hins vegar. Ætla má að viðauki A gefi réttari mynd af hæfnistigunum vegna þess að þar liggja til grundvallar textar yfirfarnir af sérfræðingi í íslensku sem öðru máli. Þó verður að taka til greina að textasafnið er of lítið til þess að hægt sé að alhæfa um einkenni hæfnistigana út frá þeim auk þess sem stigi C2 var bætt við og því allt eins líklegt að einhver munur sé á flokkun þeirra texta og hinna. Viðauki B er ætlaður til samanburðar. Þar er um talsvert fleiri texta að ræða sem ætti að gera það að verkum að einkenni textana komi betur fram og því auðveldara að draga ályktanir í samræmi við niðurstöðurnar. Á hinn bóginn verður að hafa það í huga að flokkun þessara texta hefur ekki verið yfirfarin af sérfræðingi í íslensku sem öðru máli og auk þess er dreifing texta innan flokkanna eins og áður segir mjög skökk, frá engum texta á stigi A1 yfir í 166 texta á stigi C1. Það hefur án efa áhrif á niðurstöðurnar. Eins eru textarnir mjög mislangir sem gæti að sama skapi haft áhrif.

Ef viðauki A er skoðaður má sjá að ekki allir þættirnir sem voru skoðaðir eru aðgreinandi fyrir hæfnistigin. Sem dæmi má nefna að meðalhlutfall nafn- og sagnorða helst mjög svipað í öllum tilfellum, nafnorðin telja 24-28% af heildarfjölda orða á öllum hæfnistigum og sagnorðin 14-17%. Þó eru ýmsir þættir sem eru sannarlega aðgreinandi fyrir hæfnistigin. Meðalhlutfall málsgreina sem innihalda aukasetningar vex í takt við hæfnistigin og má sjá þær niðurstöður endurspeglast í samsvarandi riti í viðauka B. Meðalhlutfall skilyrðissetninga er að sama skapi aðgreinandi þó svo að lítið sé um þær almennt í textunum og sama er upp á teningnum í viðauka B (þó svo að hlutfallið sé aðeins lægra í C2 en í C1 í því tilfelli). Samanburðarsetningar virðast aðgreinandi í viðauka A en óeðlilega hátt hlutfall þeirra á stigi B2 skekkir myndina í viðauka B. Þá má nefna að meðalhlutfall skýringarsetninga helst í hendur við hæfnistigin í viðauka A, nema í tilviki C2 þar sem hlutfallið er lægra en á C1. Sama gildir í viðauka B nema að þar er það B1 sem sker sig úr með óeðlilega lágt hlutfall.

Eins og áður segir er meðalhlutfall tiltekinna orðflokka af heildarorðafjölda textana almennt ekki aðgreinandi fyrir hæfnistigin. Þó má nefna að notkun persónufornafna er heldur hærri á A-stigunum sem gæti tengst því að í þeim er gjarnan fyrstu persónu frásögn. Ef litið er til málfræðilegu þáttanna er hins vegar ljóst að meðalhlutfall sagna í viðtengingarhætti eru aðgreinandi fyrir hæfnistigin en það sést bæði í viðauka A og B. Meðalhlutfall lýsingarorða í efsta stigi og aukafalli virðist aðgreinandi í viðauka A (þrátt

fyrir óeðlilega lágt hlutfall þeirra á stigi A2) en það endurspeglast ekki greinilega í viðauka B. Notkun efsta stigs lýsingarorða virðist fara lækkandi eftir því sem hæfni eykst og sést það í báðum viðaukum þó svo að A2 skeri sig úr í viðauka B og viðauki A sýni enga notkun þeirra á stigi A1. Notkun þátíðar sagna (sem eru ekki í lýsingarhætti þátíðar) er aðgreinandi í viðauka B en mun síður í viðauka A þó svo að megi greina stigvaxandi notkun hennar frá stigi A1 að B2. Þá má nefna að meðalhlutfall sagna í lýsingarhætti þátíðar greinir á milli texta af A-stigunum og öðrum stigum og það sama gildir um meðalhlutfall sagna í miðmynd. Nafnorð með greini í aukafalli birtast í mjög takmörkuðu upplagi á stigi A1 og gætu því greint að þann flokk sérstaklega. Að sama skapi er hlutfall persónufornafna í fyrstu persónu næstum því fjórfalt meira á stigi A1 en á hinum stigunum.

Þá er hlutfall orða sem eru lengri en 6 stafir svipað á öllum stigum nema A-stigunum þar sem það er umtalsvert lægra, einkum á A1. Meðalhlutfall einstakra orðmynda er ekki aðgreinandi fyrir hæfnistigin og helst mjög sambærileg á milli stiga í viðauka A en er af einhverjum ástæðum mun lægra á stigi C2 en hinum stigunum í viðauka B. Meðalhlutfall einstakra sagnbeyginga virðist færa lækkandi eftir því sem hæfnistigin hækka en sá þáttur getur þó varla talist aðgreinandi miðað við gögnin sem hér eru til skoðunar. Svipaða sögu er að segja af meðalhlutfalli nafnorða- og lýsingarorðabeyginga en þó er vert að nefna að hlutfall allra þessara orðflokka virðist óeðlilega lágt á C2 í viðauka B.

Flokkarinn

Gradient boosting (ísl. *stigulaflaukning*) er tölfræðiaðferð sem gengur út á að auka afl tiltölulega einfaldra flokkara með því að setja saman marga einfalda flokkara þar sem hver flokkari er látinn læra af og leiðrétta mistök síðasta flokkara. Flokkarinn byrjar á því að spá fyrir um gildi allra flokkunardæma með því að reikna logrann af líkunum á því að dæmin falli undir tiltekinn flokk miðað við meðaltalslíkur þeirra í þjálfunargögnunum. Mismunurinn á spágildunum og raungildunum er reiknaður og út frá þeim upplýsingum ásamt þáttunum sem liggja til grundvallar er fyrsta ákvörðunartréð (e. *decision tree*) smíðað. Mismunurinn á útkomu trésins og raungildum er margfaldað með námsafköstum (e. *learning rate*) líkansins og síðan lagt við útkomuna úr síðustu spám. Síðan er nýtt tré smíðað út frá mismuninum á raungildunum og spágildunum, spágildi þess reiknuð, mismunur þeirra og raungildana margfaldað með námsafköstunum og lögð við útkomuna úr öllum fyrri spám og svo koll af kolli þangað til hámarksfjölda trjáa er náð eða mismunurinn á spágildunum og raungildunum er orðin mjög lítill.

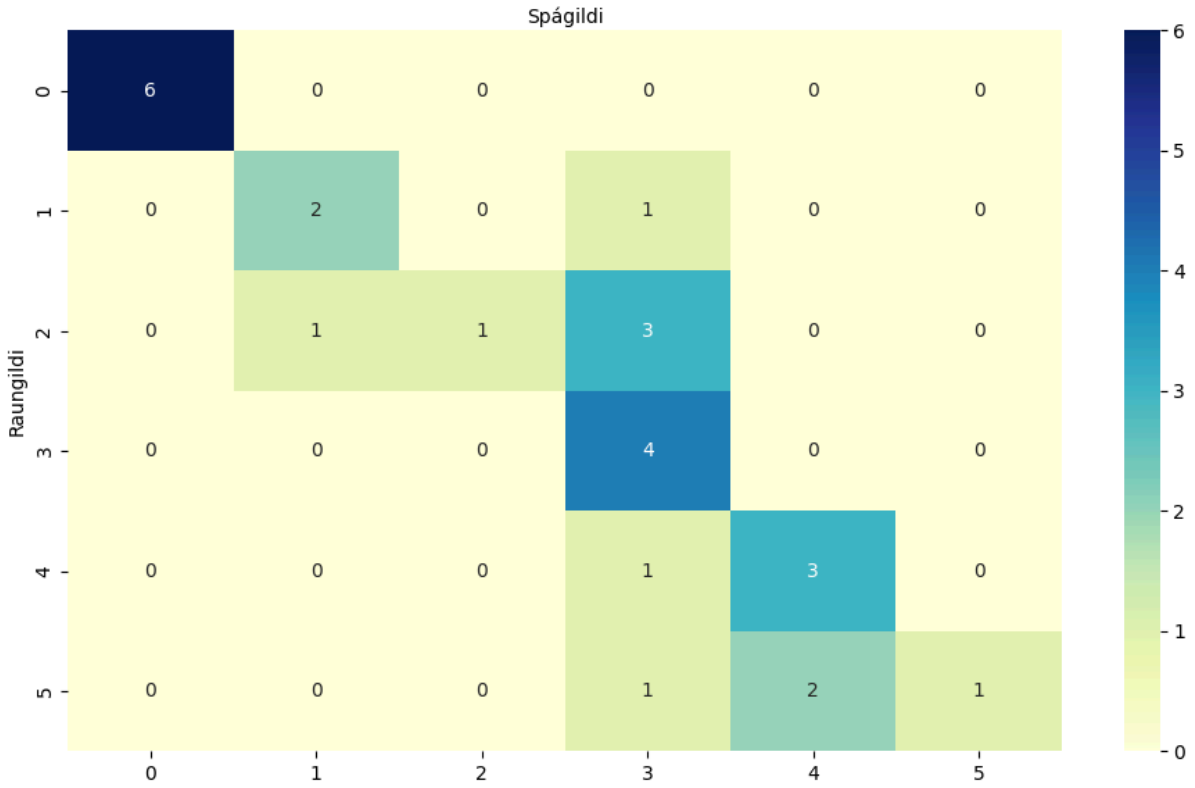
StandardScaler frá Scikit learn er beitt til þess að staðla þáttavigrana en það er gert með því að fjarlægja meðalgildið og kvarða vigrana miðað við einingardreifni (e. *unit variance*) þeirra.

Niðurstöður

Bestu niðurstöður flokkarans fengust með eftirfarandi hætti: Sérnöfn, ártöl, óbeygjanlegar tölur og prósentur eru hunsaðar alfarið, þ.e. orðmyndir sem eru markaðar sem eitthvað af ofantöldu eru útilokuð frá niðurstöðunum. Sautján þættir eru notaðir til grundvallar: Hlutfall lýsingarorða af öllum orðum; hlutfall sagnorða af öllum orðum; hlutfall forsetninga af öllum orðum; hlutfall málsgreina sem innihalda aukasetningar; hlutfall málsgreina sem innihalda spurnaraukasetningar; hlutfall málsgreina sem innihalda tilvísunarsetningar; hlutfall málsgreina sem innihalda tíðarsetningar; hlutfall sagna í viðtengingarhætti; hlutfall sagna í miðmynd; hlutfall sagna í þátíð; hlutfall lýsingarorða í efstastigi og aukafalli; hlutfall lýsingarorða í efsta stigi; hlutfall einstakra lýsingarorðabeyginga af heildarfjölda lýsingarorða; hlutfall nafnorða með greini í aukafalli; hlutfall einstakra nafnorðabeyginga af heildarfjölda nafnorða; hlutfall spurnarfornafna í aukafalli; og hlutfall persónufornafna í fyrstu persónu. Auk þessara þátta eru textarnir greyptir með IceBERT og tf-idf vigrum þeirra safnað.

Tíföld krossprófun með þáttavigrum eingöngu: 42,3% meðalnákvæmni (hæst: 62,5%, lægst 37,5%).
Tíföld krossprófun með þáttavigrum og tf-idf: 41,4% meðalnákvæmni (hæst 62,5%, lægst 25%).
Tíföld krossprófun með þáttavigrum og orðgreypingum: 34,5% meðalnákvæmni (hæst 62,5%, lægst 0%).
Tíföld krossprófun með þáttavigrum, tf-idf og orðgreypingum: 41,4% meðalnákvæmni (hæst 62,5%, lægst 25%).

Hér fyrir neðan er vafafylki (e. *confusion matrix*) sem fékkst við eina þjálfun líkansins þar sem nákvæmnishlutfallið var 65,4% (athugið að hér var ekki farið í krossprófun og því er líklegt að ofmátun útskýri tiltölulega góða frammistöðu líkansins). Á grafinu táknar 0 hæfnistig A1, 1 táknar A2, 2 táknar B1 og svo koll af kolli. Þar sem samræmi er á milli raungildis og spágildis hefur líkanið spáð rétt fyrir um flokkun viðkomandi texta. Eins og sést á líkanið ekki í neinum vandræðum með að finna texta á stigi A1 enda eru þeir textar mjög einkennandi fyrir stigið og flestir í samtalsformi. Eins virðist líkanið alltaf geta aðgreint texta á B2 en mig grunar að það sé afleiðing ofmátunar frekar en að auðvelt sé að aðgreina þá texta. Önnur stig valda nokkrum vafa. Það er áhugavert að sjá hvaða stig ruglast helst. Stig B1 er í flestum tilfellum ranglega flokkað sem stig B2 og því líklegt að það sé lítill munur þeirra á milli. Eins virðist líkanið ekki gera mikinn greinarmun á milli C1 og C2 þó svo að textar á C1 séu oftar flokkaðir rétt. Ástæða þess að C2 eru frekar greindir sem C1 gæti þó verið fólgin í því að þeim textum var bætt aftan við og hafa því ekki verið yfirfarnir af sérfræðingi í íslensku sem öðru máli.



Að endingu gerði ég tilraun með að nota öll gögn sem ég hafði til umráða til þess að endurþjálfa líkanið. Með öðrum orðum eru hér öll upprunalegu gögnin og öll gögnin sem ég safnaði úr Risamálheildinni með bráðabrigðaflokkun minni. Vegna þess hversu langir sumir aukatextarnir eru, einkum þeir sem komu úr bókaundirmálheildinni, tekur þjálfun á öllum gögnunum talsvert langan tíma og því ákvað ég að stytta textana þannig að aðeins væri unnið með fyrstu 1000 orðin úr hverjum texta. Úr tífaldri krossprófun þar sem fyrrnefnd þáttagreining er notuð eingöngu fæst 34,8% nákvæmni. Gögnin sem ég bætti við eru óyfirfarin og þar af leiðandi líkleg til þess að vera að einhverju leyti rangt flokkuð og auk þess er dreifingin á milli flokka mjög ójöfn eins og áður hefur verið nefnt. Það er mjög mikilvægt að fá meira magn af gögnum og jafna hlutföllin ef þjálfa á flokkara sem ræður við verkefnið. Ég hvet því eindregið til þess að gagnasöfnuninni verði haldið áfram.

Áframhald

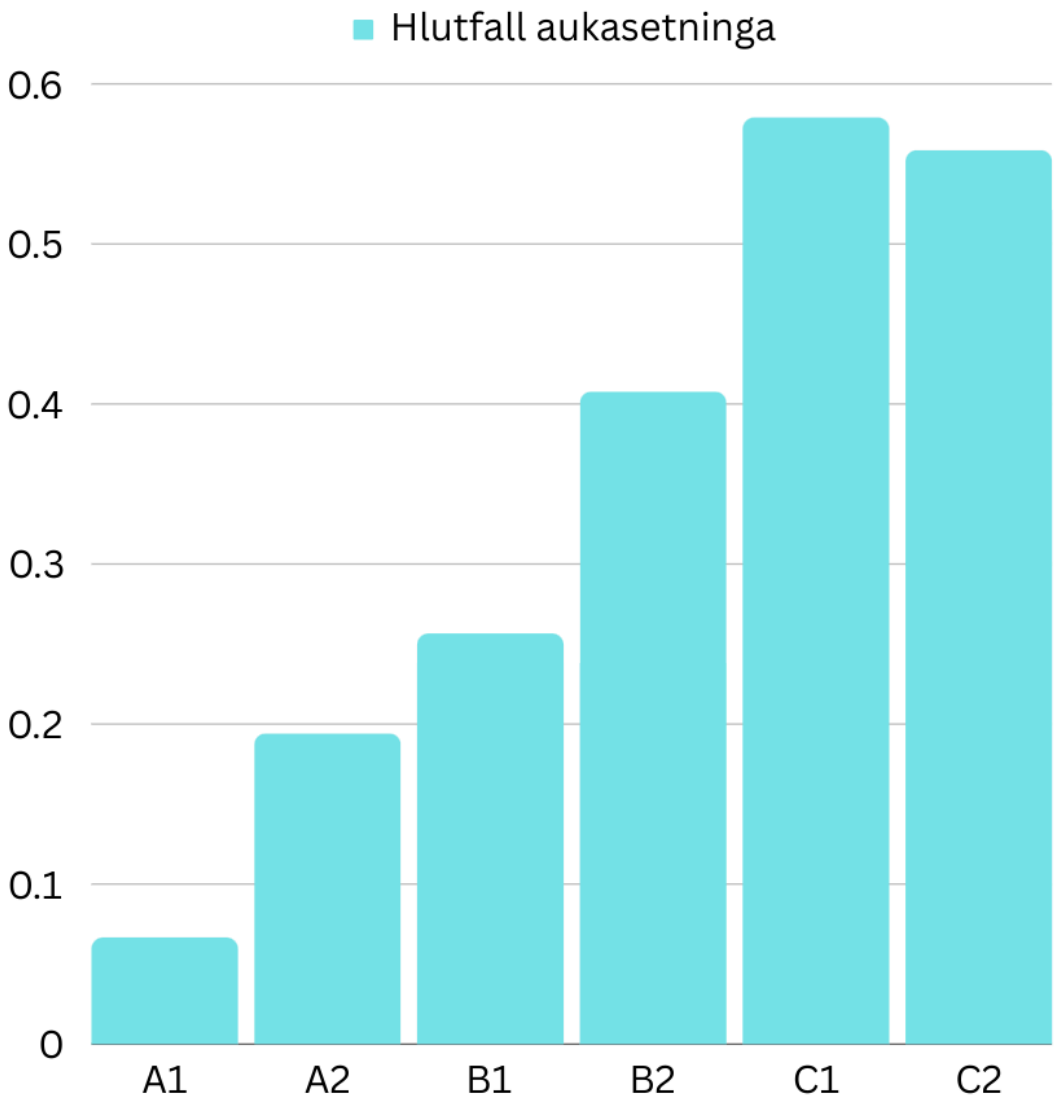
Eins og sést af niðurstöðunum er þessi flokkari langt frá því að vera fullkominn en hann má nýta til samanburðar við önnur líkön sem kunna að vera smíðuð í þessum tilgangi en eins mætti betrumbæta hann með því að skoða aðra og fleiri þætti. Það væri áhugavert að skoða fjölda aukasetninga innan hverrar málsgreinar en í þessari yfirferð

var eingöngu miðað við hvort málsgreinin innihaldi aukasetningu á annað borð eða ekki. Ég tel að best væri að skoða aðferðir sem taka merkingarleg atriði textana með í reikninginn auk málfræðilegu þáttanna enda er talsverður munur á umræðuefnum texta á stigi A1 og þeirra á stigi C1 og ég tel að það að greypa textana með IceBERT dugi ekki til eitt og sér til þess að ná fram þeim merkingarmun. Hugsanlega mætti beita sérþjálfuðum líkönum í þessum tilgangi, þ.e.a.s. einhverju á borð við afstöðugreiningarlíkönum (e. *sentiment analysis*) sem myndu þá senda tilsvarandi upplýsingar inn í þáttavigrana sem liggja til grundvallar í aðalflokkunarlíkaninu. Aðrar skölunaraðferðir gætu orðið til bóta sem og eftirvinnsluaðferðir. Ég tel jafnframt að það væri þess virði að bæta flokkarann með því að beita virku námi (e. *active learning*) á útkomur flokkarans á stærra safni texta úr Risamálheildinni eða öðrum textasöfnum. Þannig mætti að sama skapi safna mun stærra textasafni en liggur hér til grundvallar sem er nauðsynlegt til þess að bæta niðurstöður hvaða flokkara sem er. Eins væri gott að bera saman frammistöðu tölfræðilíkana á borð við það sem hefur verið rætt hér og skapandi líkana á borð við GPT-4. Það er von mín að þessar fyrstu niðurstöður komi að gagni í áframhaldandi vinnu með sjálfvirka flokkun á móðurmálstextum samkvæmt evrópska tungumálarammanum.

Viðauki A - Úr upphaflegu gögnunum (frá Kolfinnu)

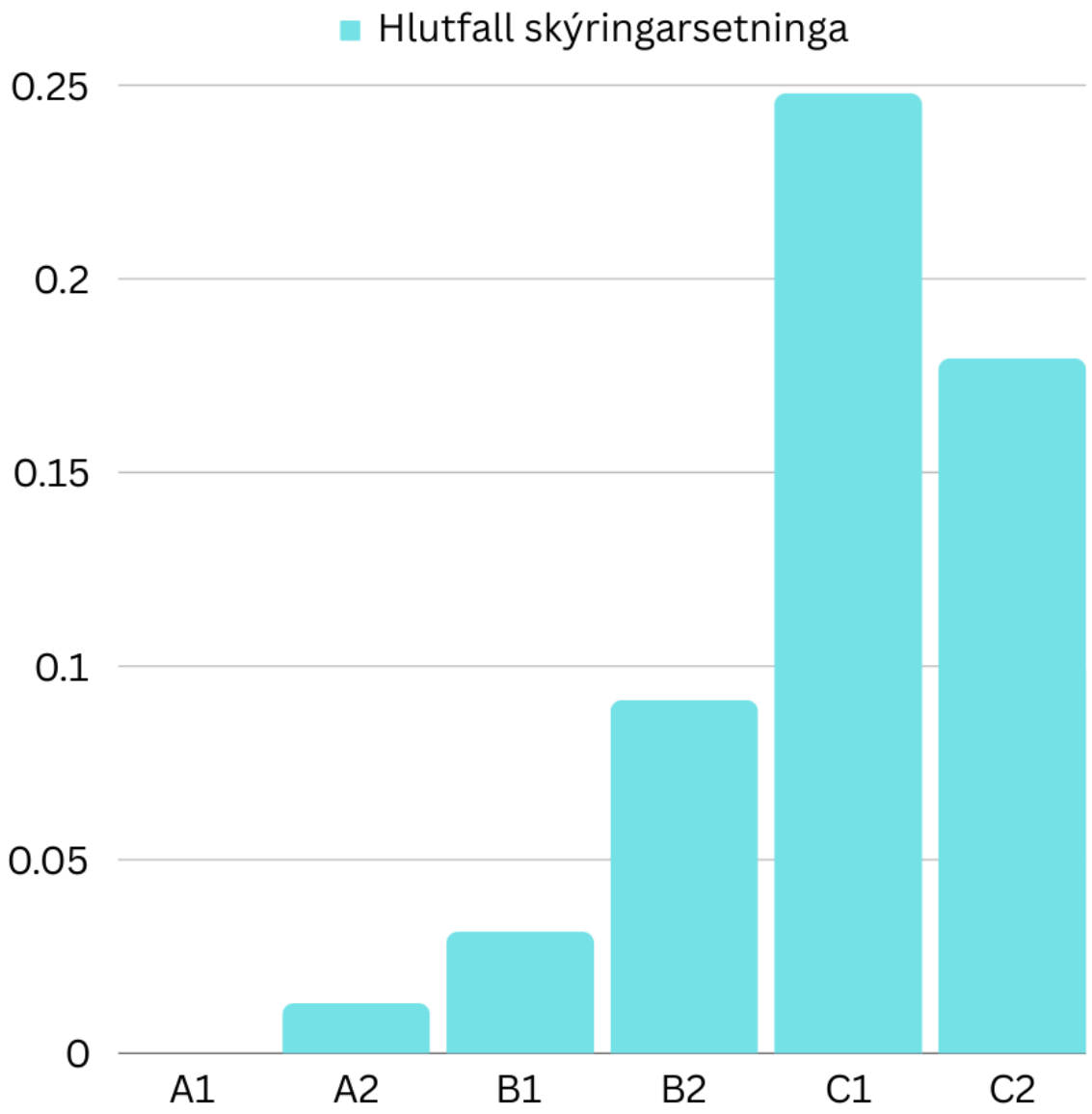
Meðaltalshlutfall málsgreina með aukasetningum í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda aukasetningu, óháð fjölda aukasetninga innan hverrar málsgreinar)

A1 0.06666666666666667
A2 0.19417989417989415
B1 0.2566724026249198
B2 0.4077692671271406
C1 0.5790058232365924
C2 0.5586528512400605



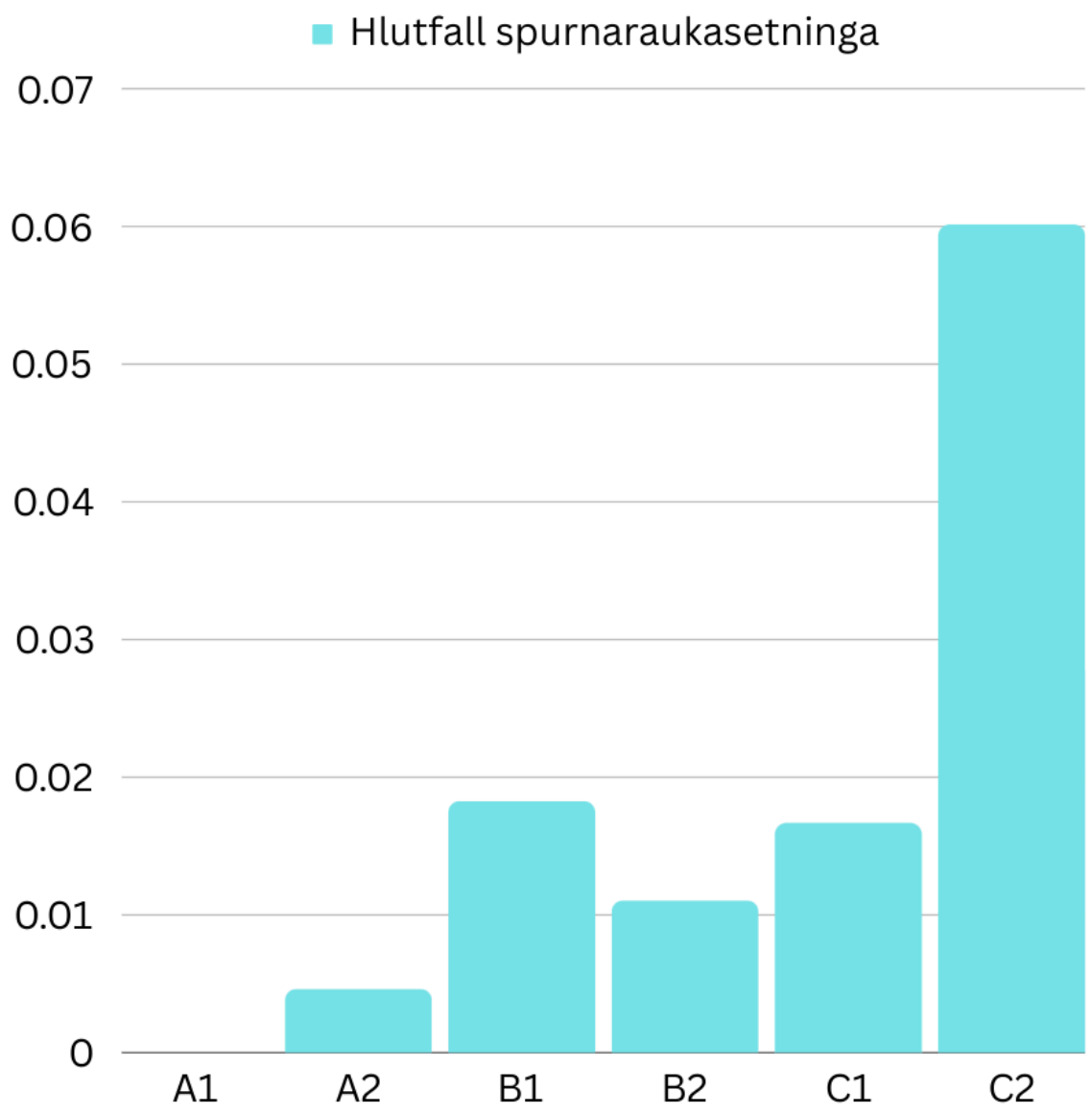
Meðaltalshlutfall málsgreina með skýringarsetningum í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda skýringarsetningu, óháð fjölda skýringarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.012962962962962963
B1 0.031432748538011694
B2 0.09122721552389071
C1 0.24791490560721335
C2 0.17947674902616761



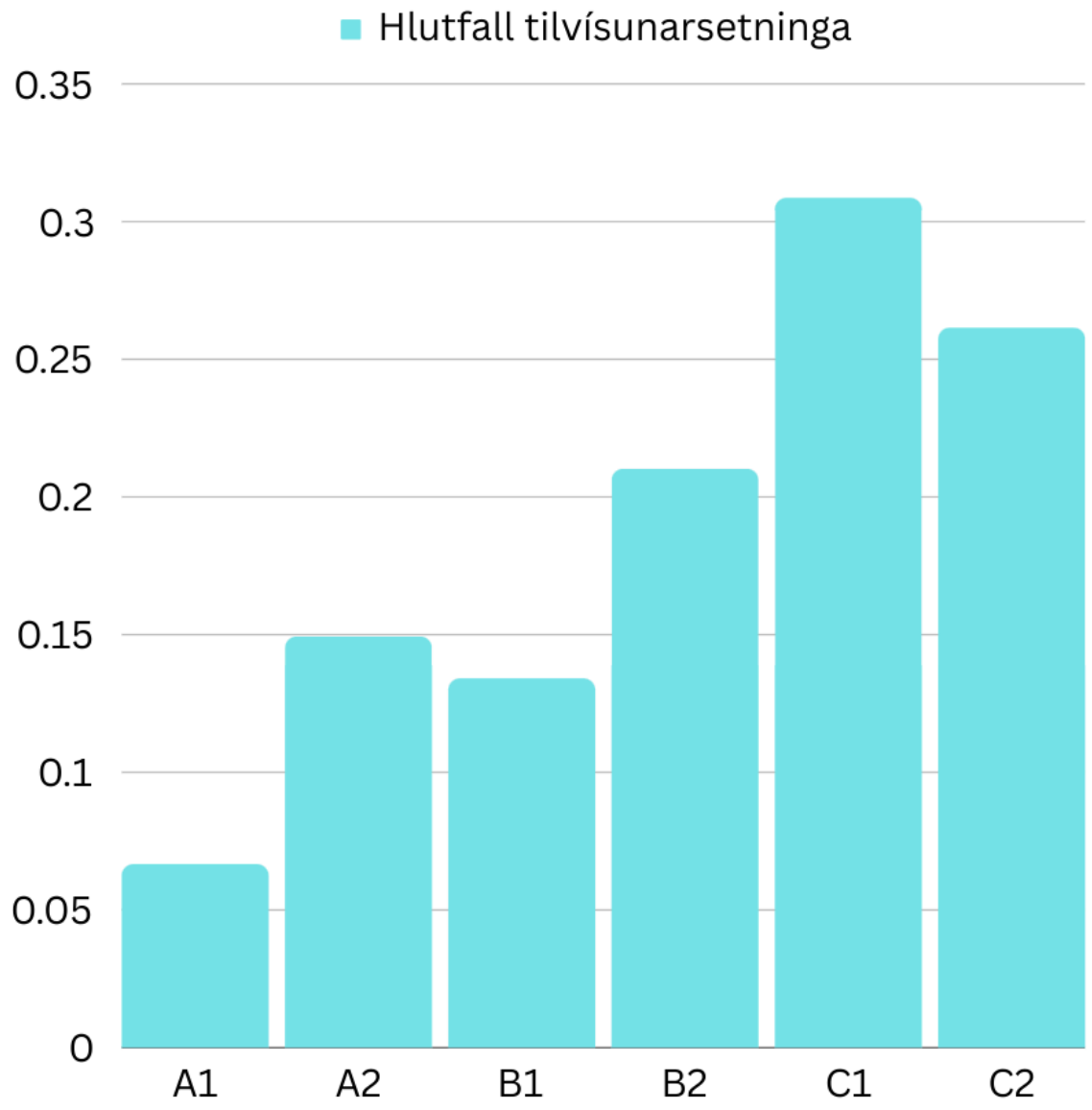
Meðaltalshlutfall málsgreina með spurnaraukasetningar í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda spurnaraukasetningu, óháð fjölda spurnaraukasetninga innan hverrar málsgreinar)

A1 0.0
A2 0.004629629629629629
B1 0.01827485380116959
B2 0.011054421768707483
C1 0.016687016687016686
C2 0.060143552076691614



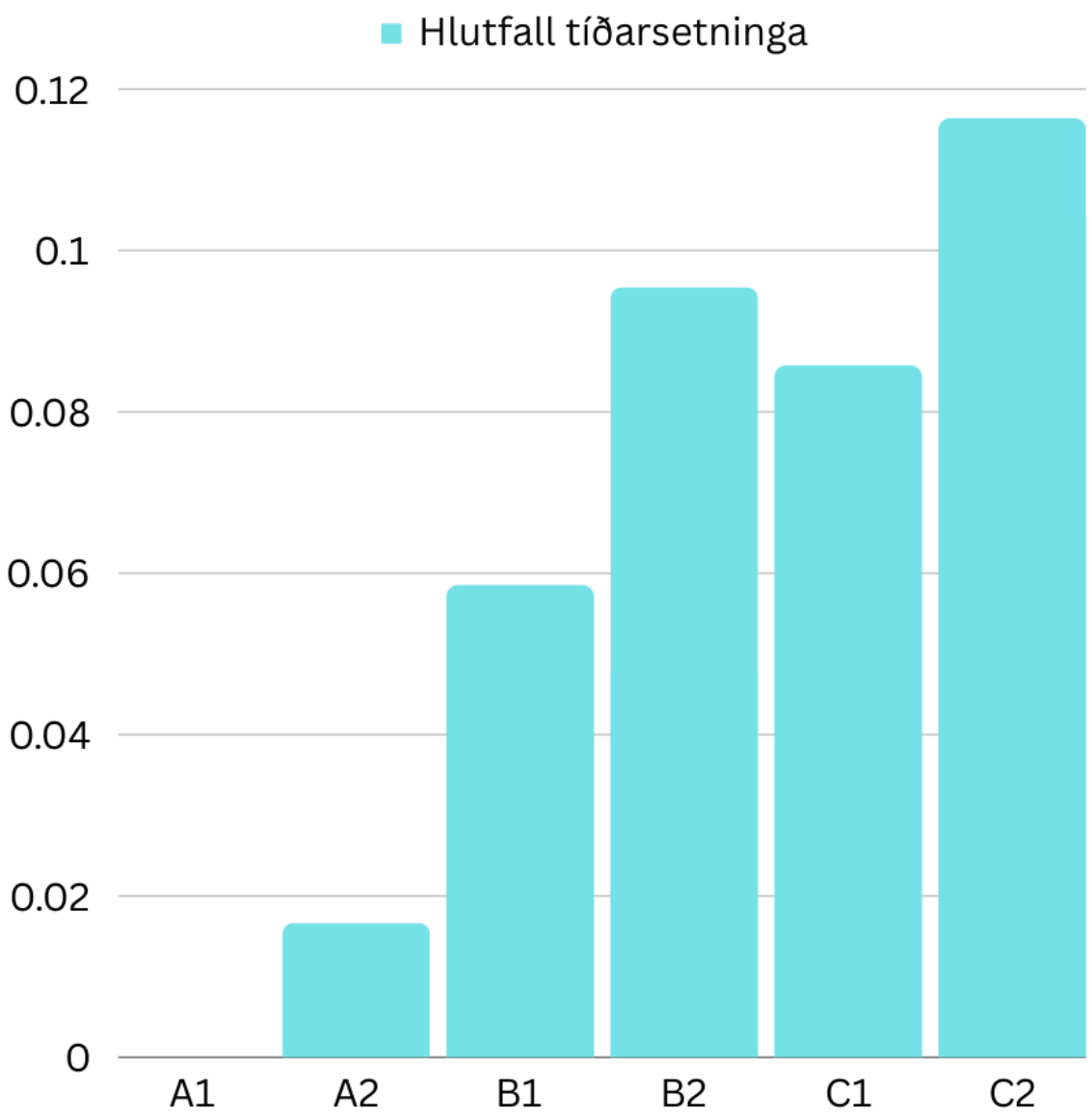
Meðaltalshlutfall málsgreina með tilvísunarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tilvísunarsetningu, óháð fjölda tilvísunarsetninga innan hverrar málsgreinar)

A1 0.06666666666666667
A2 0.14933862433862433
B1 0.13412210326540988
B2 0.21023148389790697
C1 0.30862449516295676
C2 0.2615068232800791



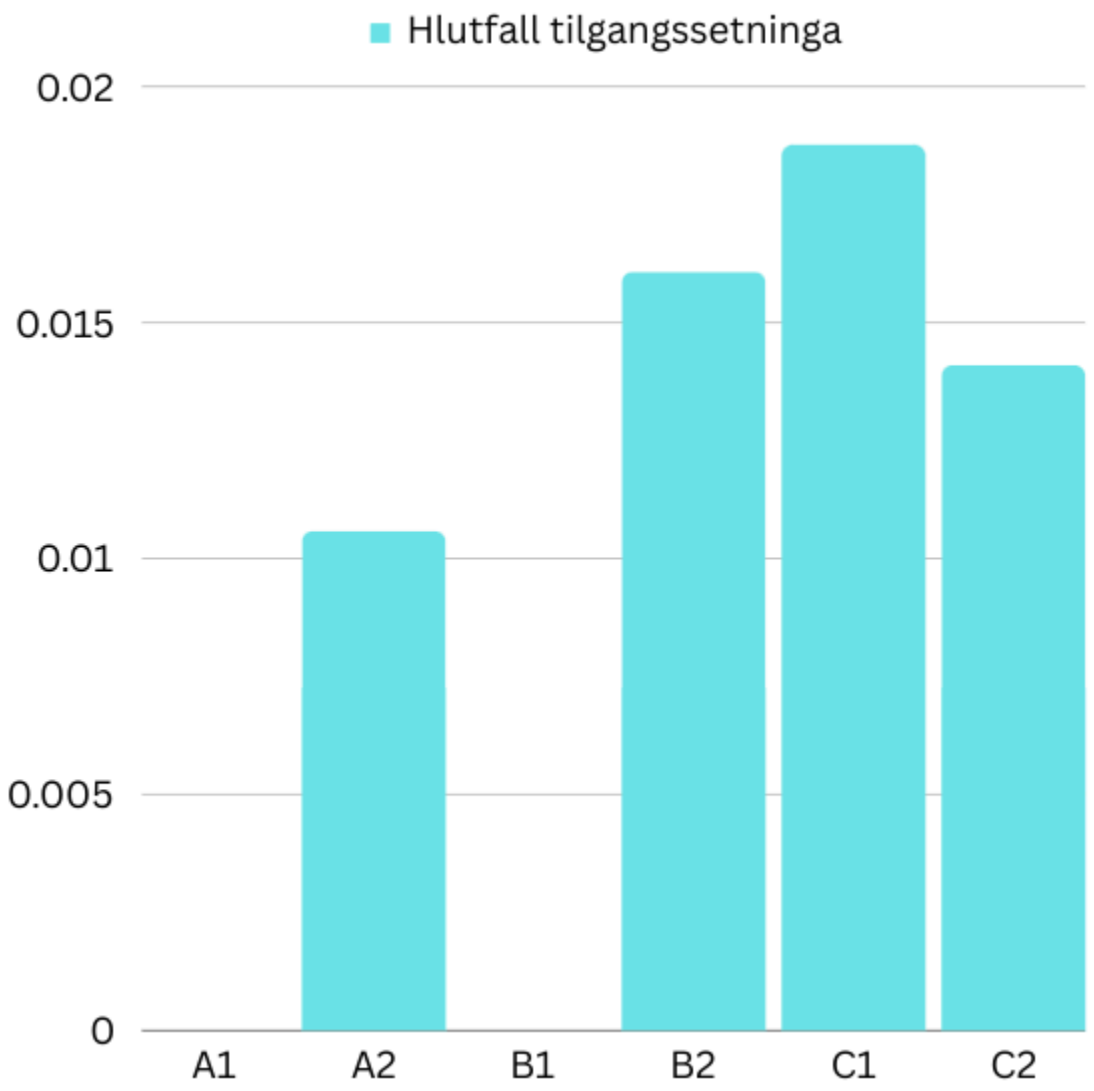
Meðaltalshlutfall málsgreina með tíðarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tíðarsetningu, óháð fjölda tíðarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.016666666666666666
B1 0.05853609718741298
B2 0.09543288329824683
C1 0.08576046460661844
C2 0.11639041406483269



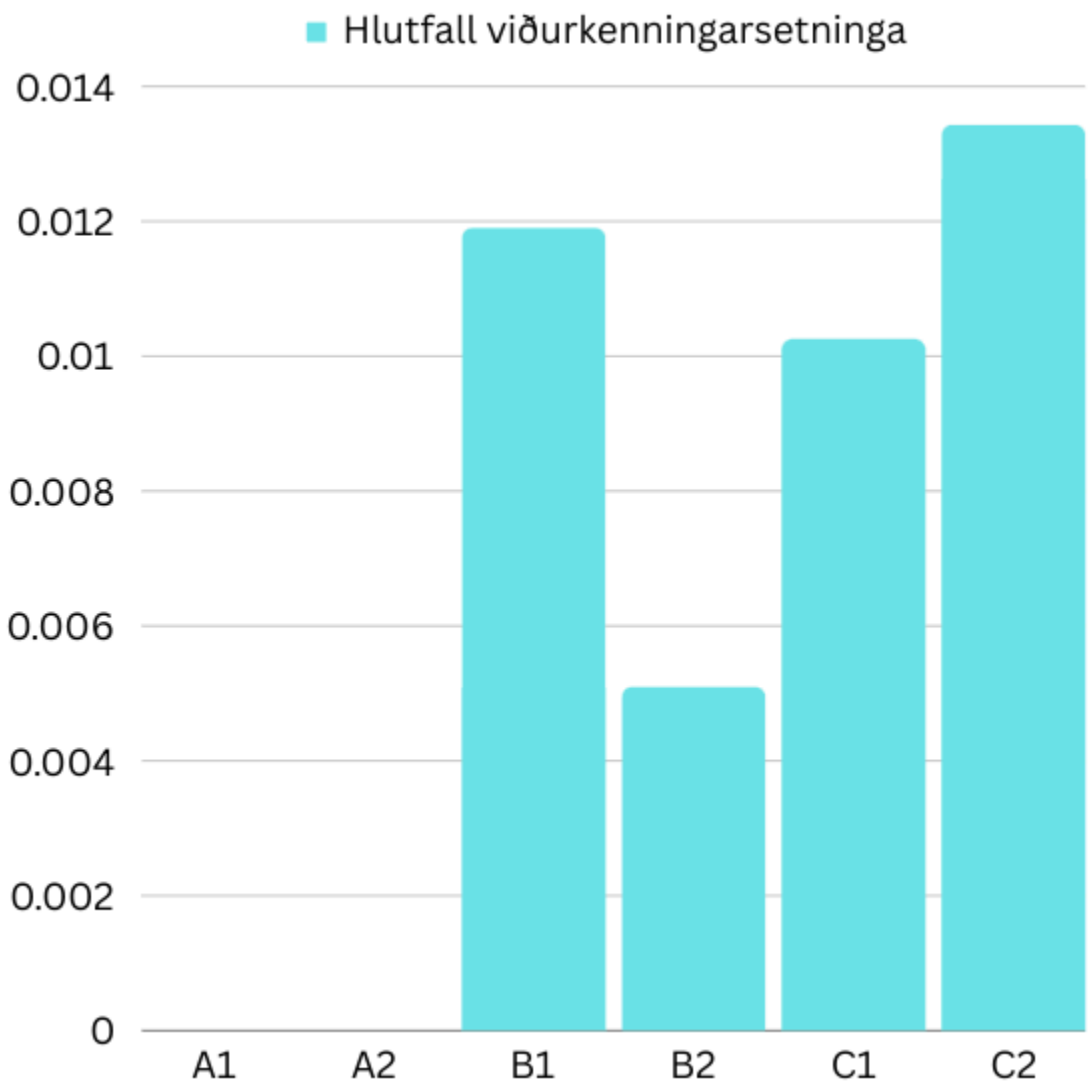
Meðaltalshlutfall málsgreina með tilgangssetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tilgangssetningu, óháð fjölda tilgangssetninga innan hverrar málsgreinar)

A1 0.0
A2 0.010582010582010581
B1 0.0
B2 0.016071428571428573
C1 0.018762718762718764
C2 0.01409527972027972



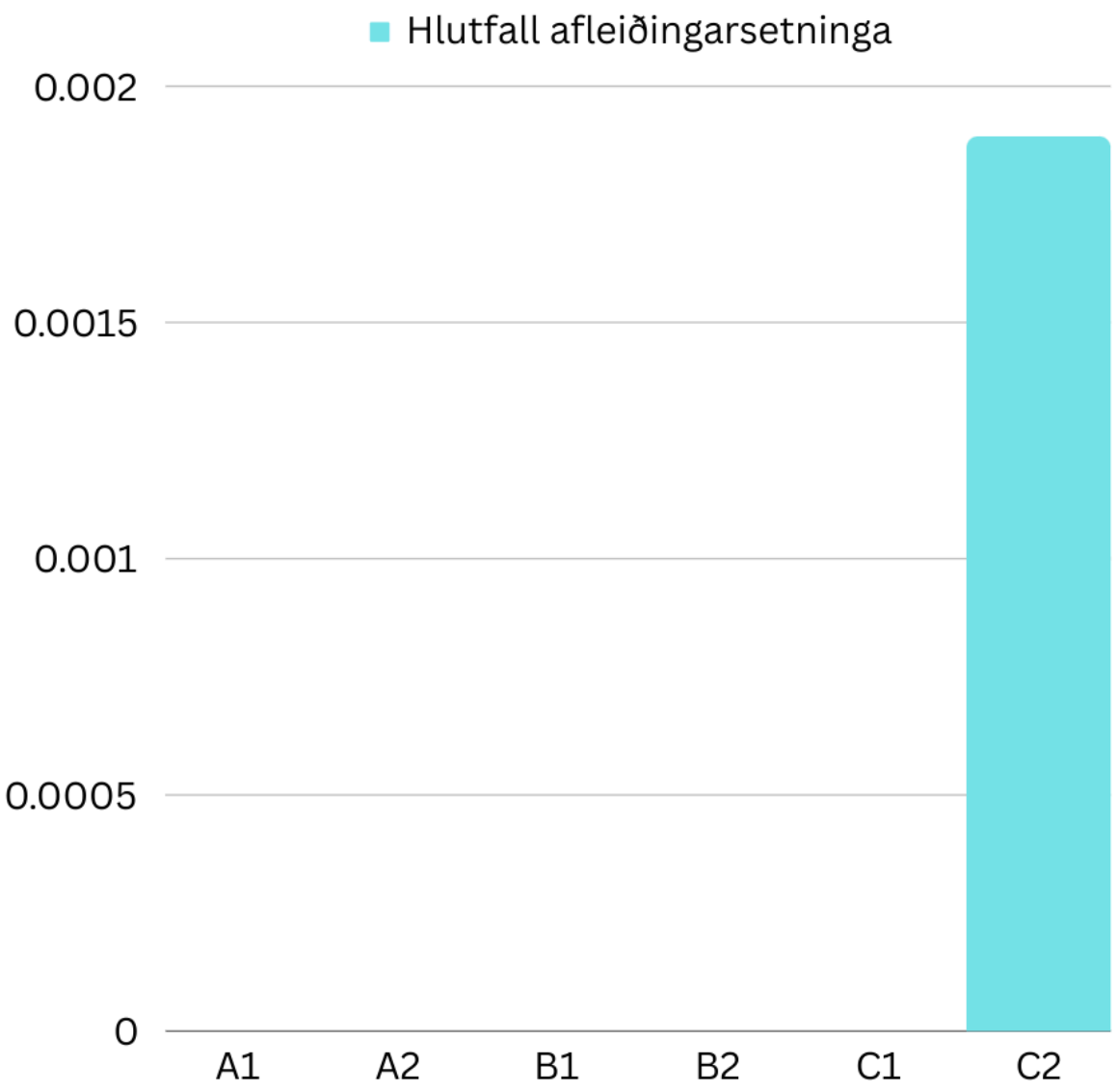
Meðaltalshlutfall málsgreina með viðurkenningarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda viðurkenningarsetningu, óháð fjölda viðurkenningarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.0
B1 0.011904761904761904
B2 0.00510204081632653
C1 0.010256410256410256
C2 0.013432400932400933



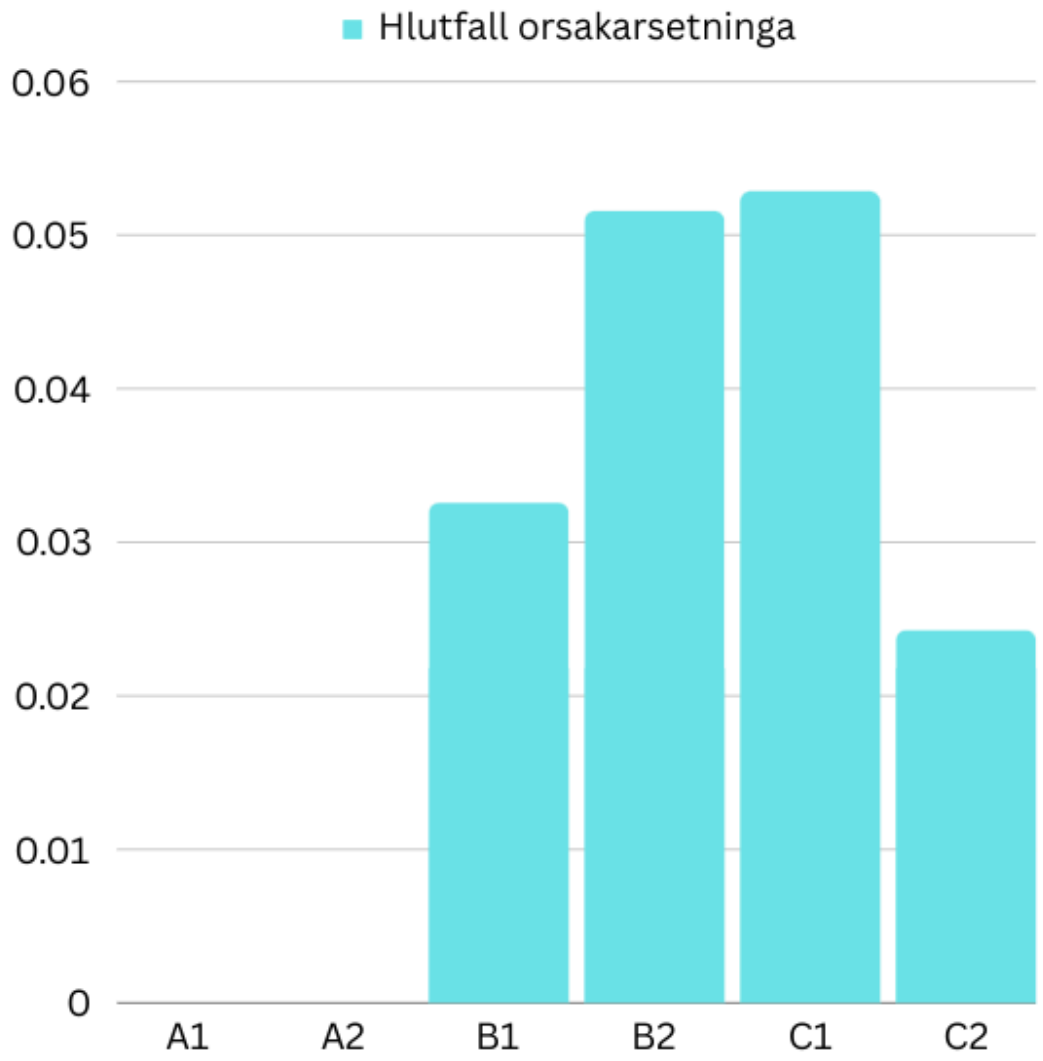
Meðaltalshlutfall málsgreina með afleiðingarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda afleiðingarsetningu, óháð fjölda afleiðingarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.0
B1 0.0
B2 0.0
C1 0.0
C2 0.0018939393939394



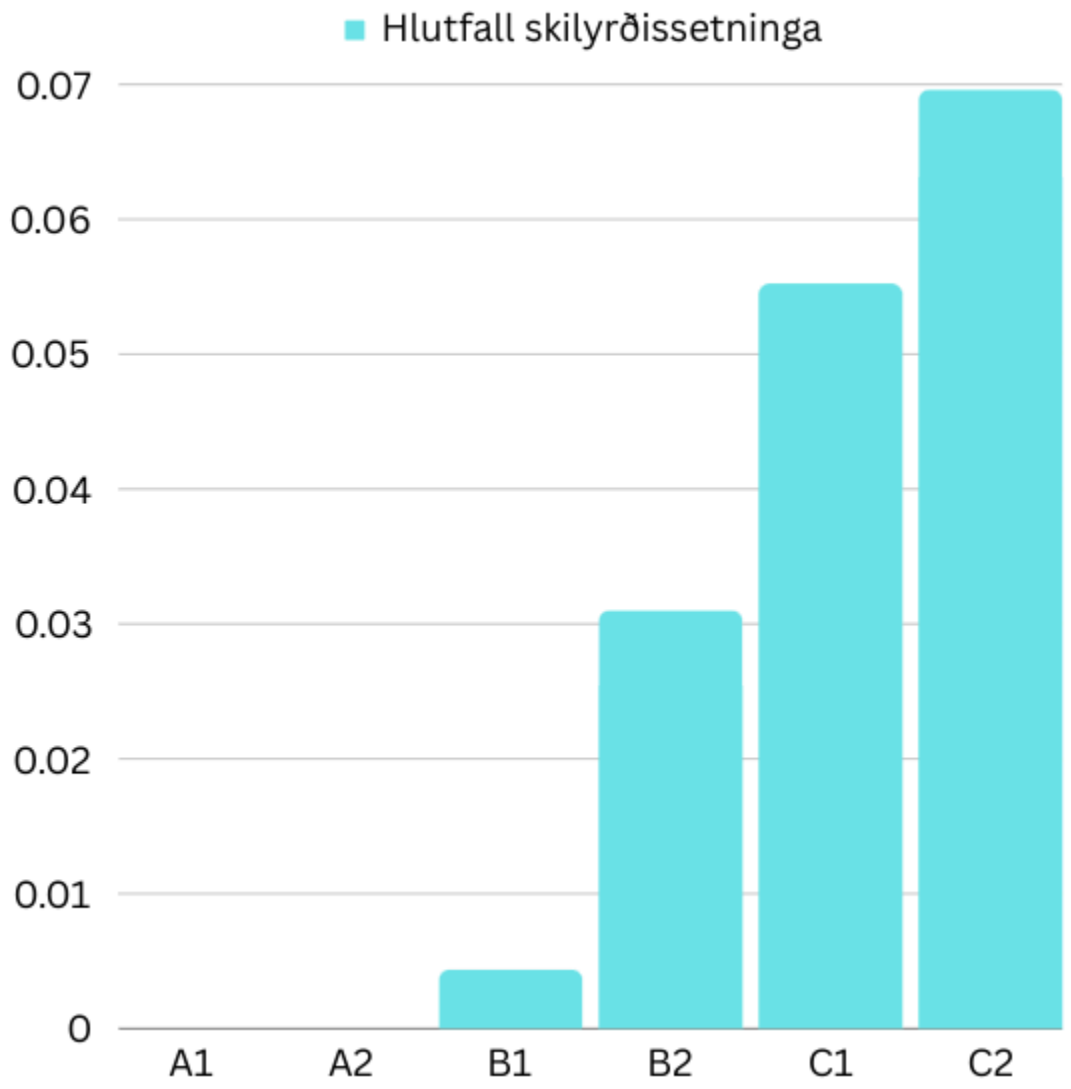
Meðaltalshlutfall málsgreina með orsakarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda orsakarsetningu, óháð fjölda orsakarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.0
B1 0.03258145363408521
B2 0.051587301587301584
C1 0.05286465671081056
C2 0.024288749579447252



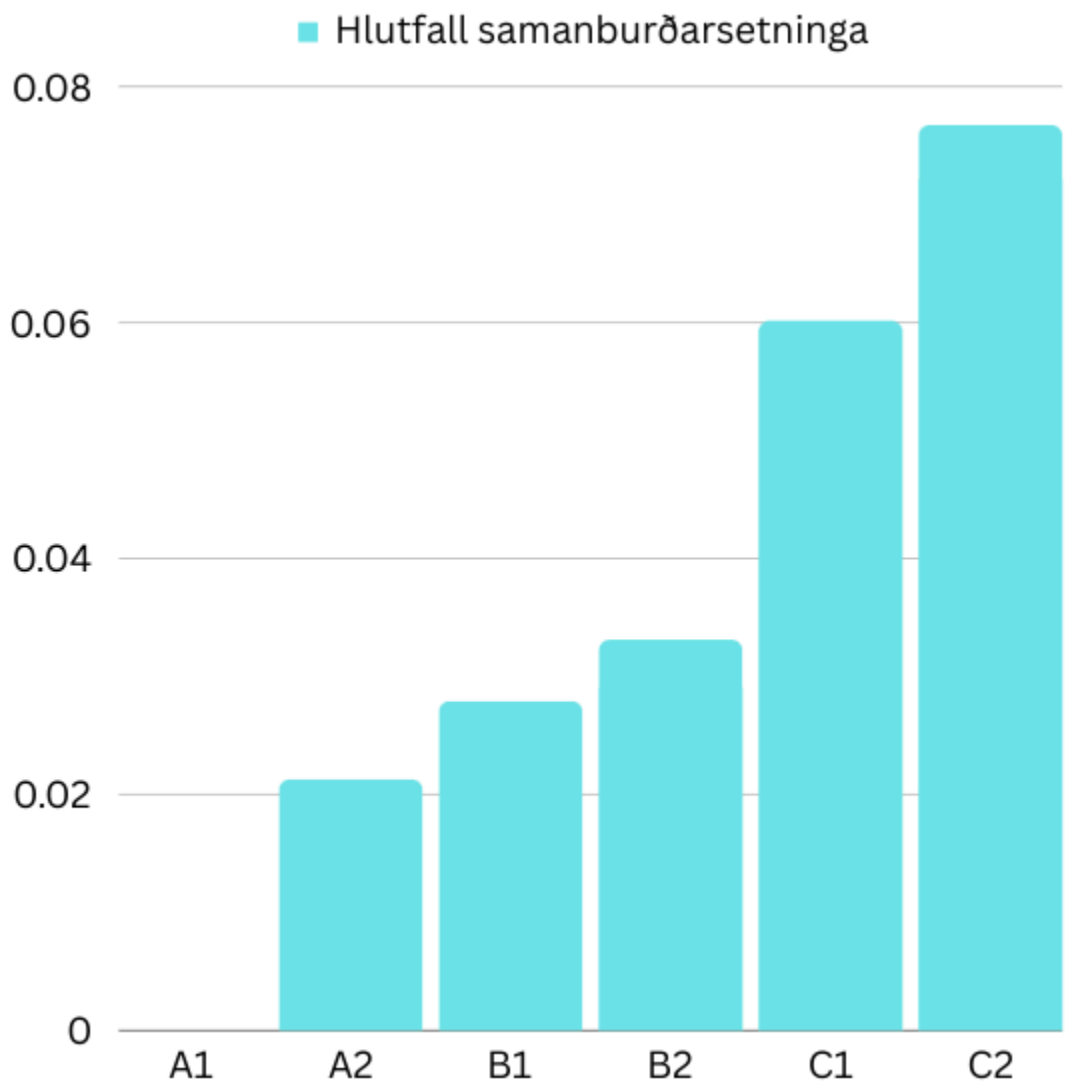
Meðaltalshlutfall málsgreina með skilyrðissetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda skilyrðissetningu, óháð fjölda skilyrðissetninga innan hverrar málsgreinar)

A1 0.0
A2 0.0
B1 0.0043859649122807015
B2 0.031012767425810904
C1 0.05524013024013024
C2 0.06961371961371961



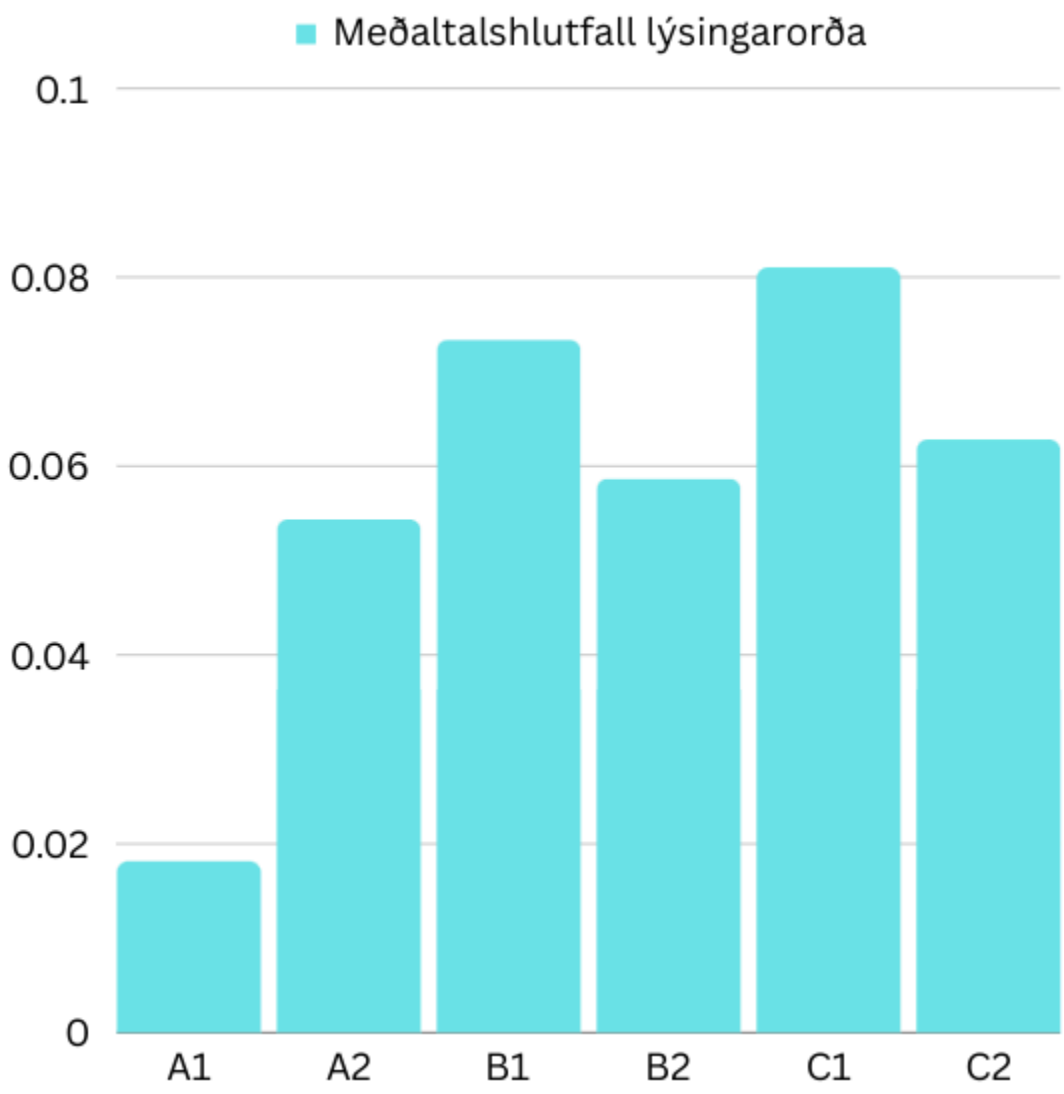
Meðaltalshlutfall málsgreina með samanburðarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda samanburðarsetningu, óháð fjölda samanburðarsetninga innan hverrar málsgreinar)

A1 0.0
A2 0.021296296296296296
B1 0.02792306854091751
B2 0.033126293995859216
C1 0.060144485144485144
C2 0.07672335201404969



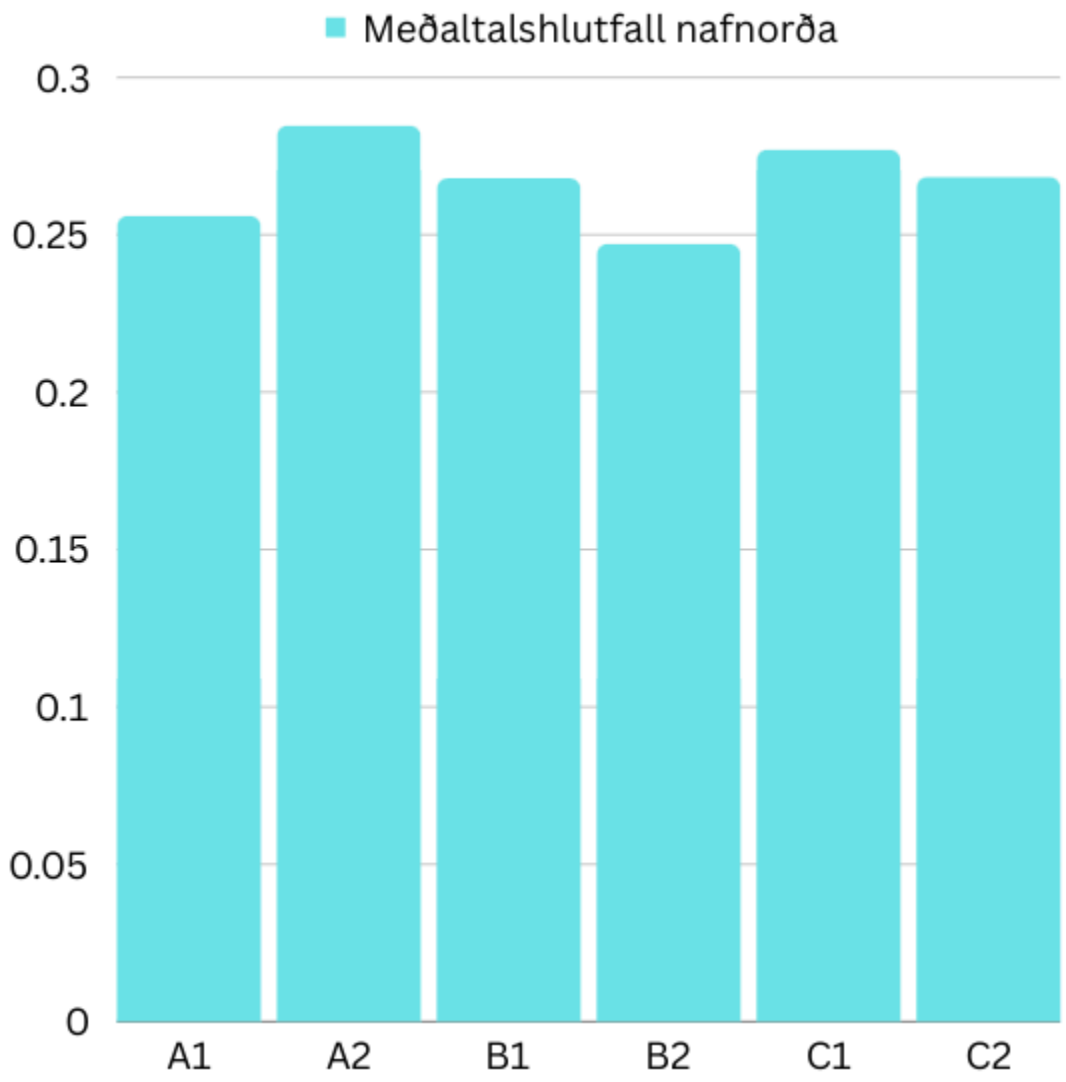
Meðaltalshlutfall lýsingarorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 0.018154251408968387
A2 0.05435285862366566
B1 0.07337618655849497
B2 0.058668705845064194
C1 0.0810388653368984
C2 0.06282053452525592



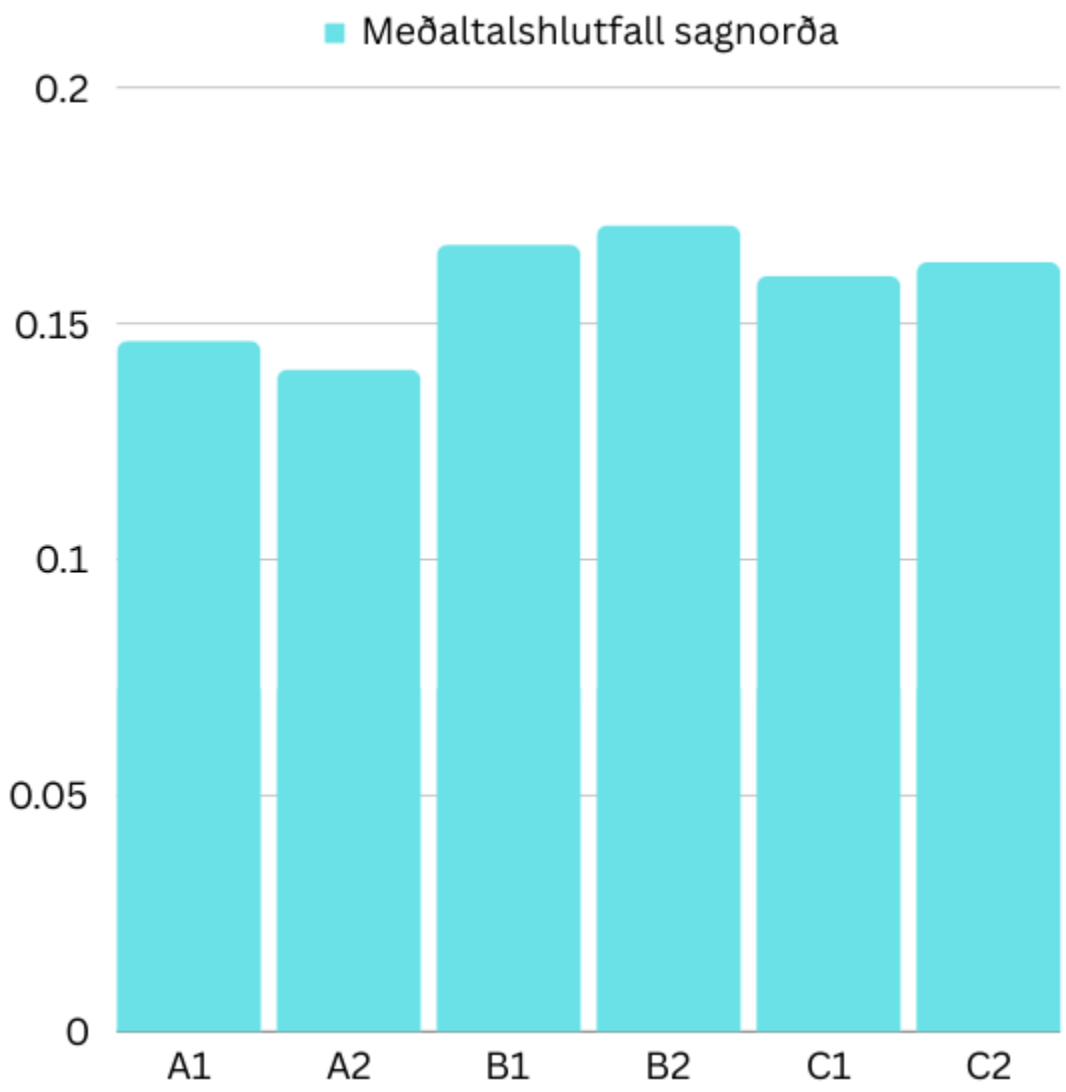
Meðaltalshlutfall nafnorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 0.2560154517445791
A2 0.2847183527825939
B1 0.26796284276918797
B2 0.24711585322297827
C1 0.27707375024642944
C2 0.268389671166604



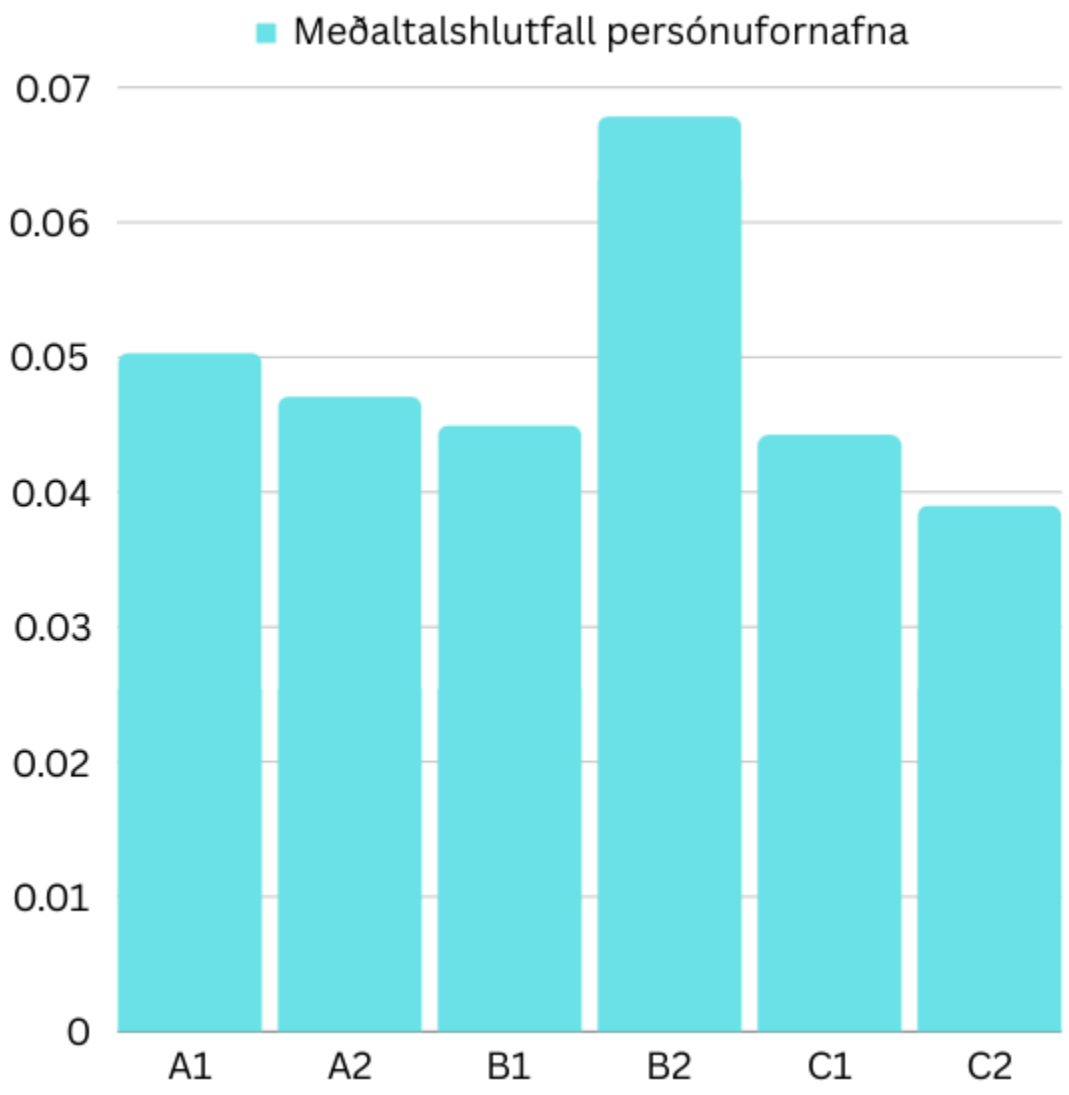
Meðaltalshlutfall sagnorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 0.14631590247404216
A2 0.14026857223992992
B1 0.16668250339354726
B2 0.170791087527943
C1 0.16006752544214267
C2 0.1629820658625848



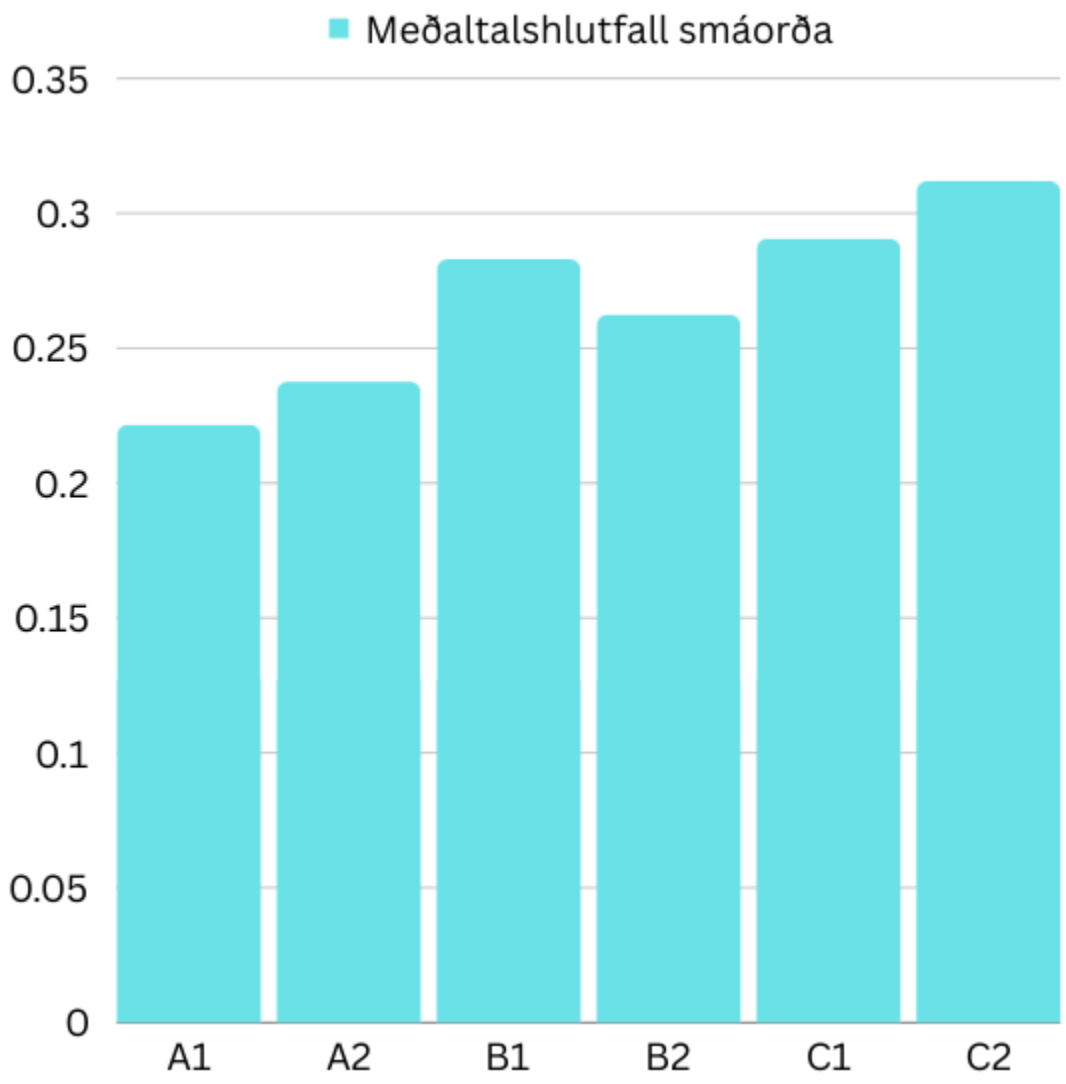
Meðaltalshlutfall persónufornafna af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 0.05031885635035015
A2 0.047078220091560874
B1 0.044951965704661756
B2 0.06785230199191793
C1 0.04424439680297247
C2 0.03900981083053389



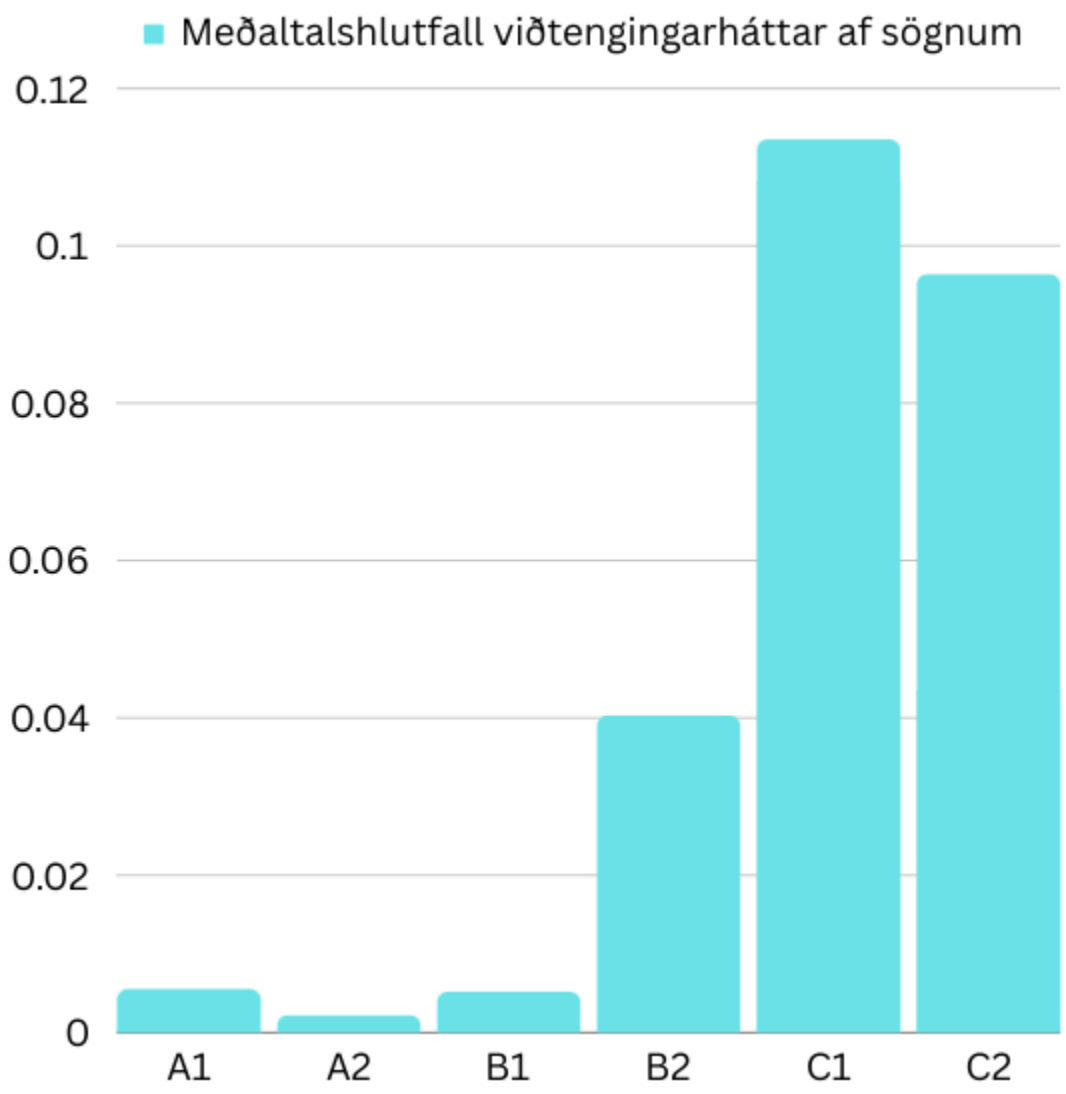
Meðaltalshlutfall smáorða (mörk c og a í markaskránni, ath inniheldur líka atviksorð) af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 0.22162160703429798
A2 0.23761320945043277
B1 0.2830010472939573
B2 0.2624985481922721
C1 0.29059999700537753
C2 0.31195502179602846



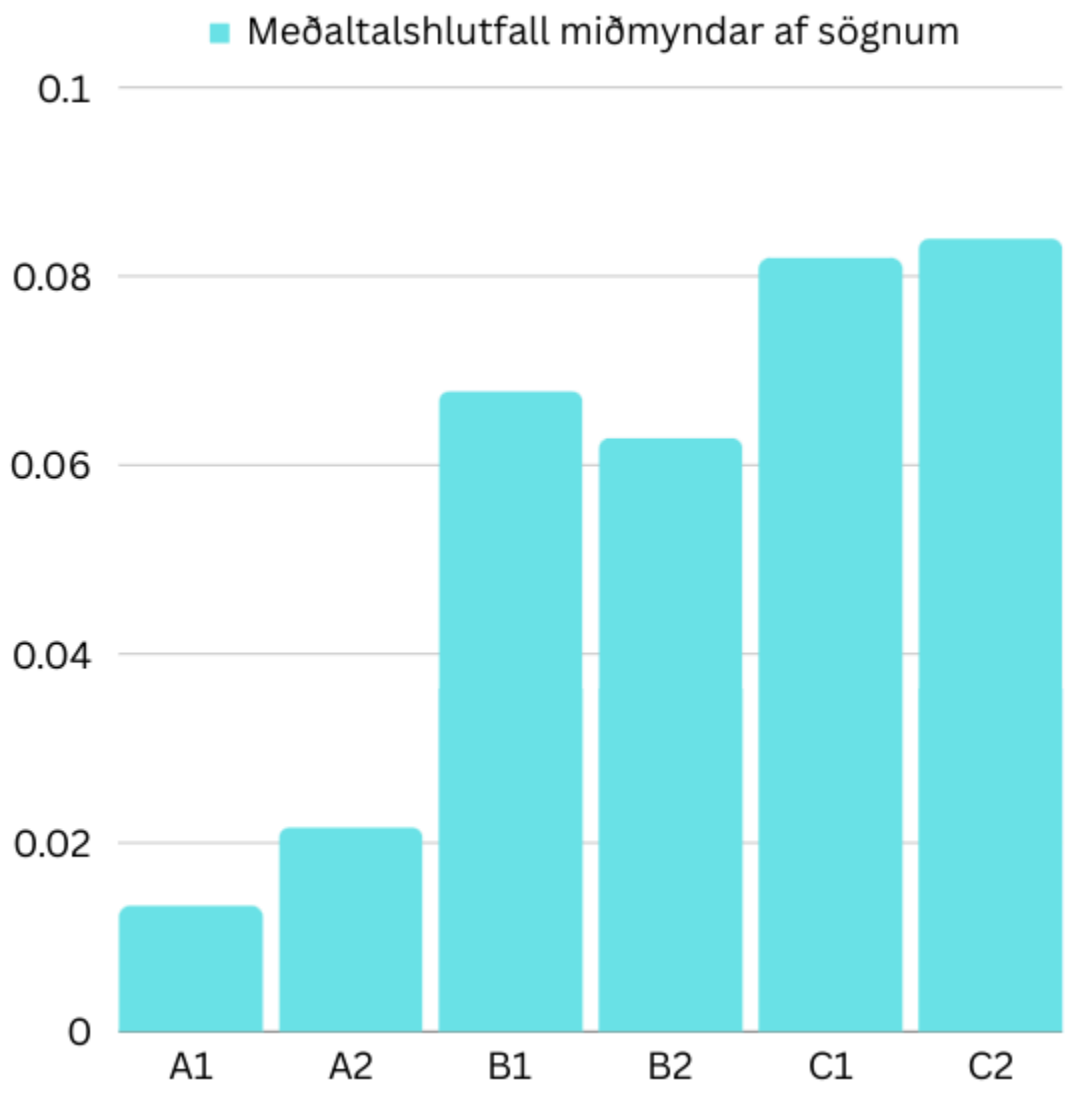
Meðaltalshlutfall viðtengingarháttar af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 0.005555555555555555
A2 0.0022522522522522522
B1 0.005208333333333333
B2 0.04032892729683187
C1 0.11353387759498526
C2 0.0964144211714603



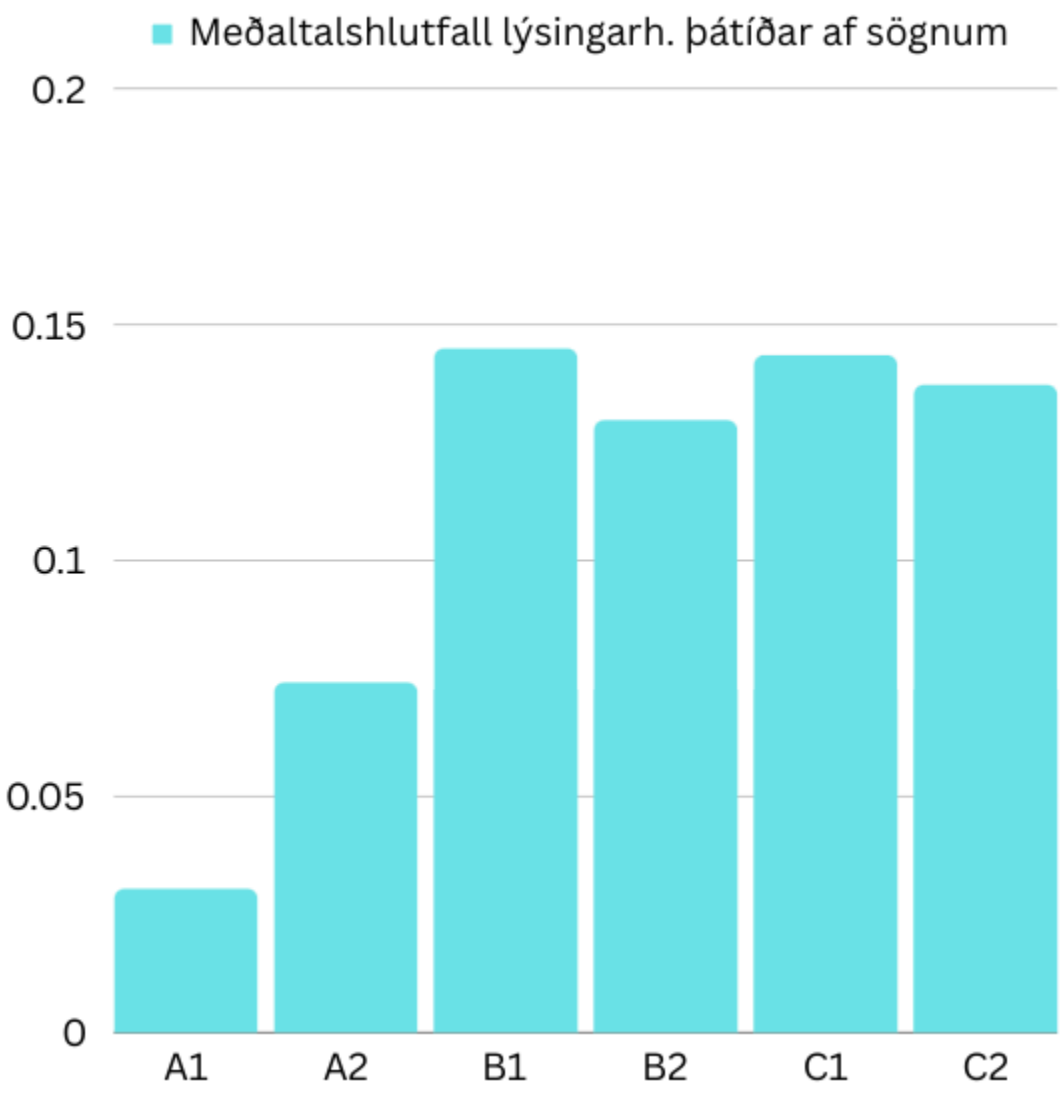
Meðaltalshlutfall miðmyndar af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 0.013333333333333334
A2 0.021648101737089503
B1 0.06783604299928926
B2 0.06287443031875568
C1 0.0819370526597026
C2 0.08399307152948414



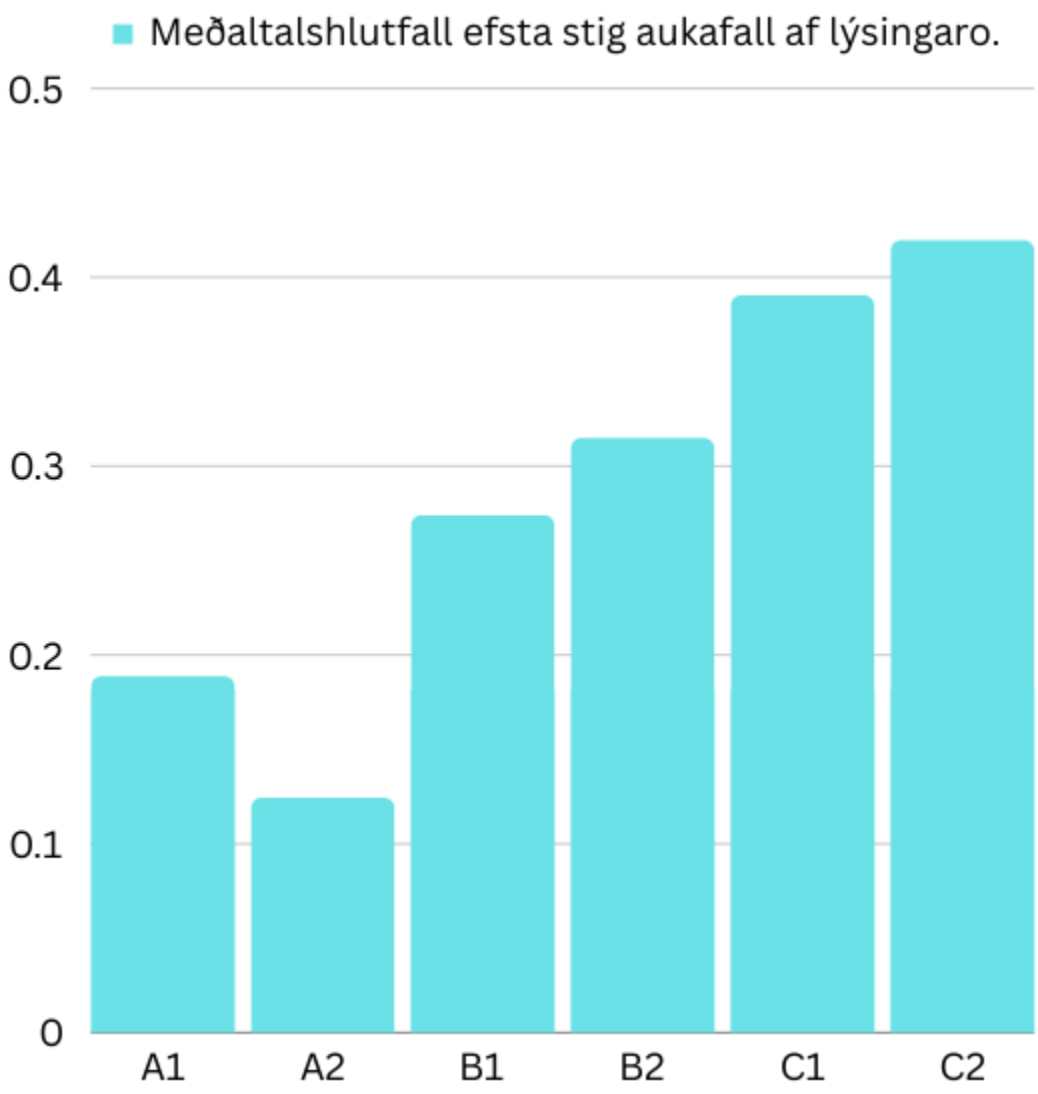
Meðaltalshlutfall lýsingarháttar þátíðar (oftast þolmynd) af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 0.03055555555555555
A2 0.07424385336100107
B1 0.14499378109452735
B2 0.12980072574248858
C1 0.1435538465745244
C2 0.13731490031372703



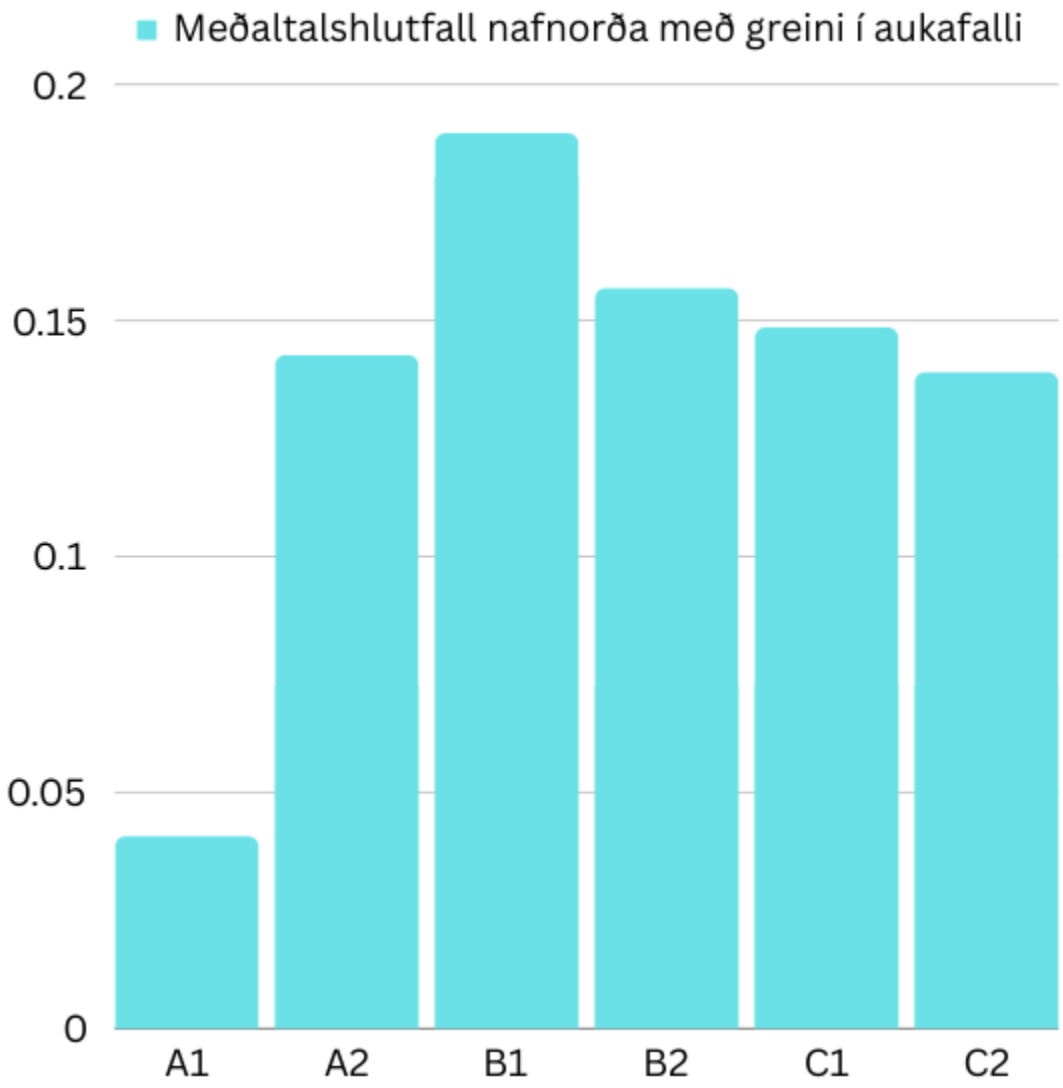
Meðaltalshlutfall lýsingarorða í efsta stigi í öðru en nefnifalli af öllum lýsingarorðum í hverjum texta á hverju hæfnistigi

A1 0.1888888888888889
A2 0.12448646125116712
B1 0.2741140668700737
B2 0.31500273953281477
C1 0.3905421674894437
C2 0.41975402814165624



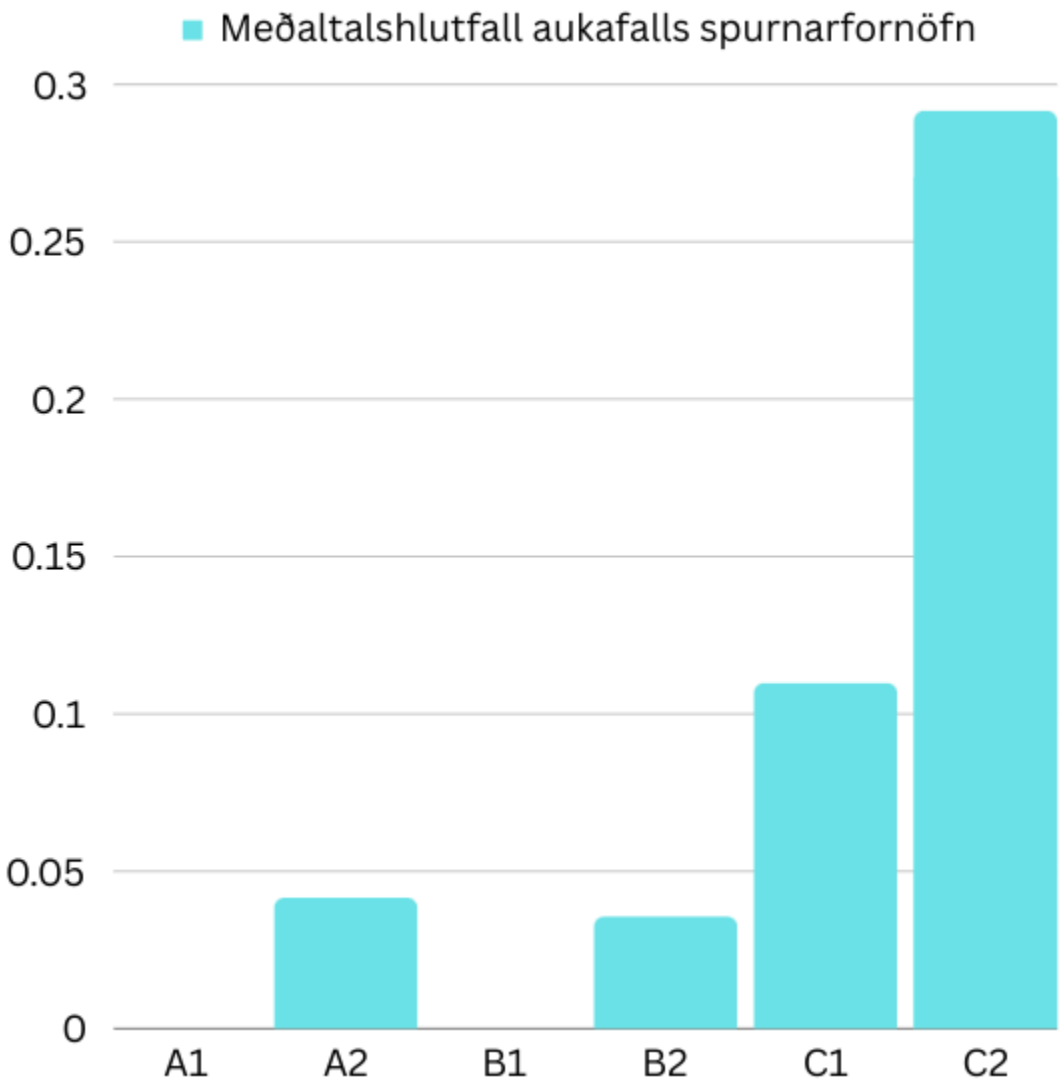
Meðaltalshlutfall nafnorða með greini í öðru en nefnifalli af öllum nafnorðum (ATH sérnöfn hafa áhrif, eru tekin með hér) í hverjum texta á hverju hæfnistigi

A1 0.0407936507936508
A2 0.1427068859062403
B1 0.1896942007052614
B2 0.1568849291834212
C1 0.1485376411355324
C2 0.13908667794899812



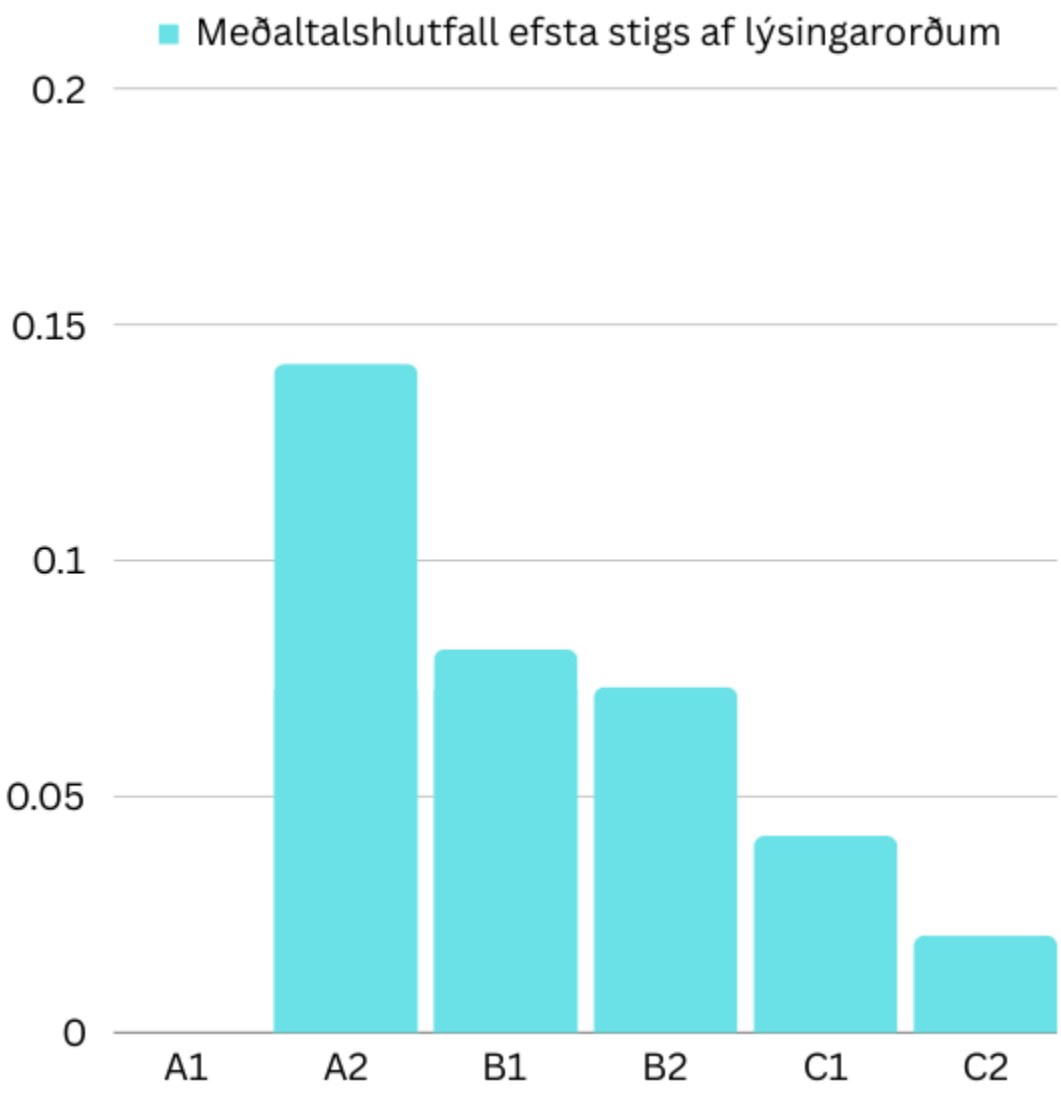
Meðaltalshlutfall spurnarfornafna í öðru en nefnifalli af öllum spurnarfornöfnum í hverjum texta á hverju hæfnistigi

A1 0.0
A2 0.041666666666666664
B1 0.0
B2 0.03571428571428571
C1 0.10989010989010989
C2 0.2916666666666667



Meðaltalshlutfall lýsingarorða í efsta stigi af öllum lýsingarorðum í hverjum texta á hverju hæfnistigi

A1 0.0
A2 0.14166666666666666
B1 0.08120474798106377
B2 0.0732163200208313
C1 0.041697952229031054
C2 0.020589147999107552

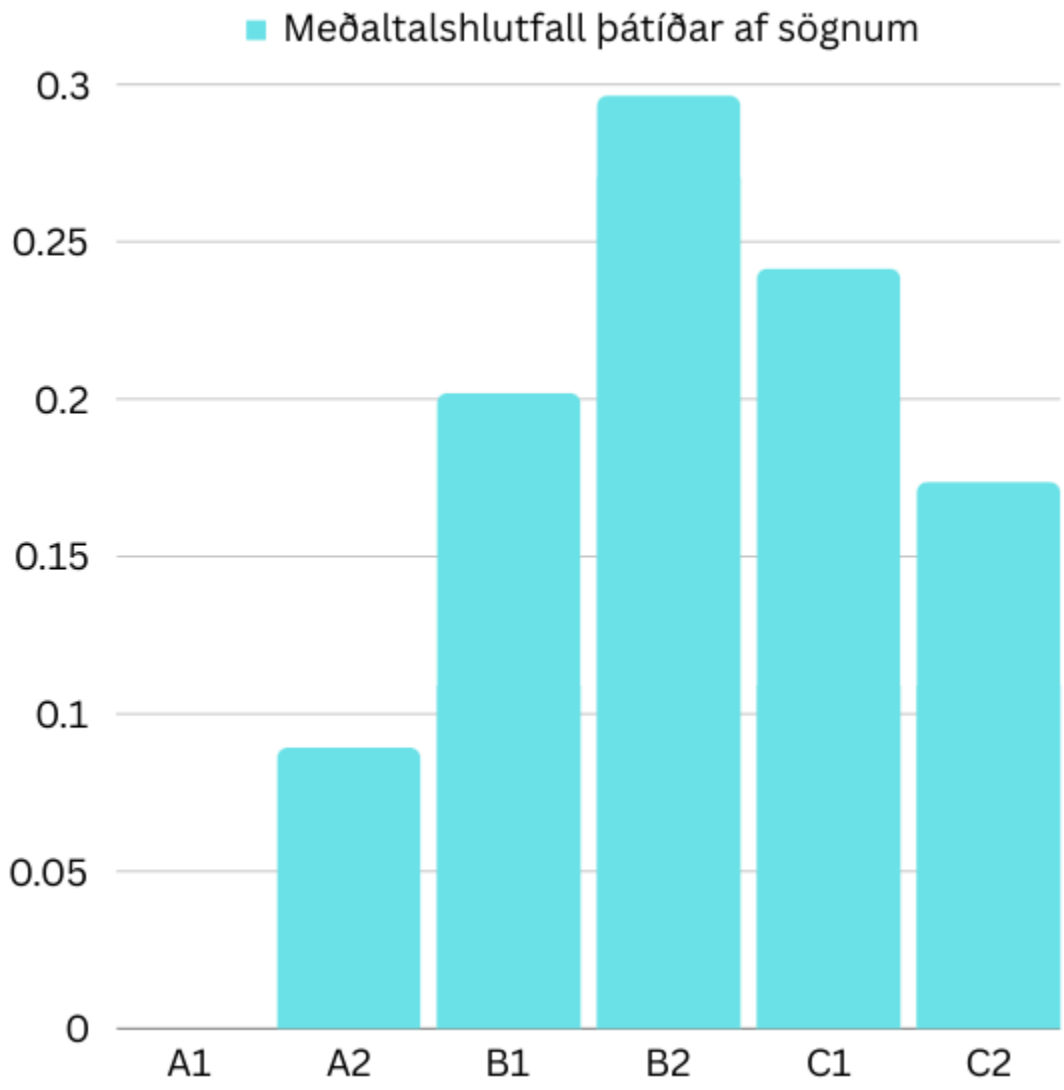


Meðaltalshlutfall persónufornafna í fyrstu persónu af öllum persónufornöfnum í hverjum texta á hverju hæfnistigi

A1 0.4534920634920635
A2 0.10750481402655315
B1 0.16737204237204237
B2 0.10208719851576994
C1 0.08707987918514234
C2 0.11246367767537123

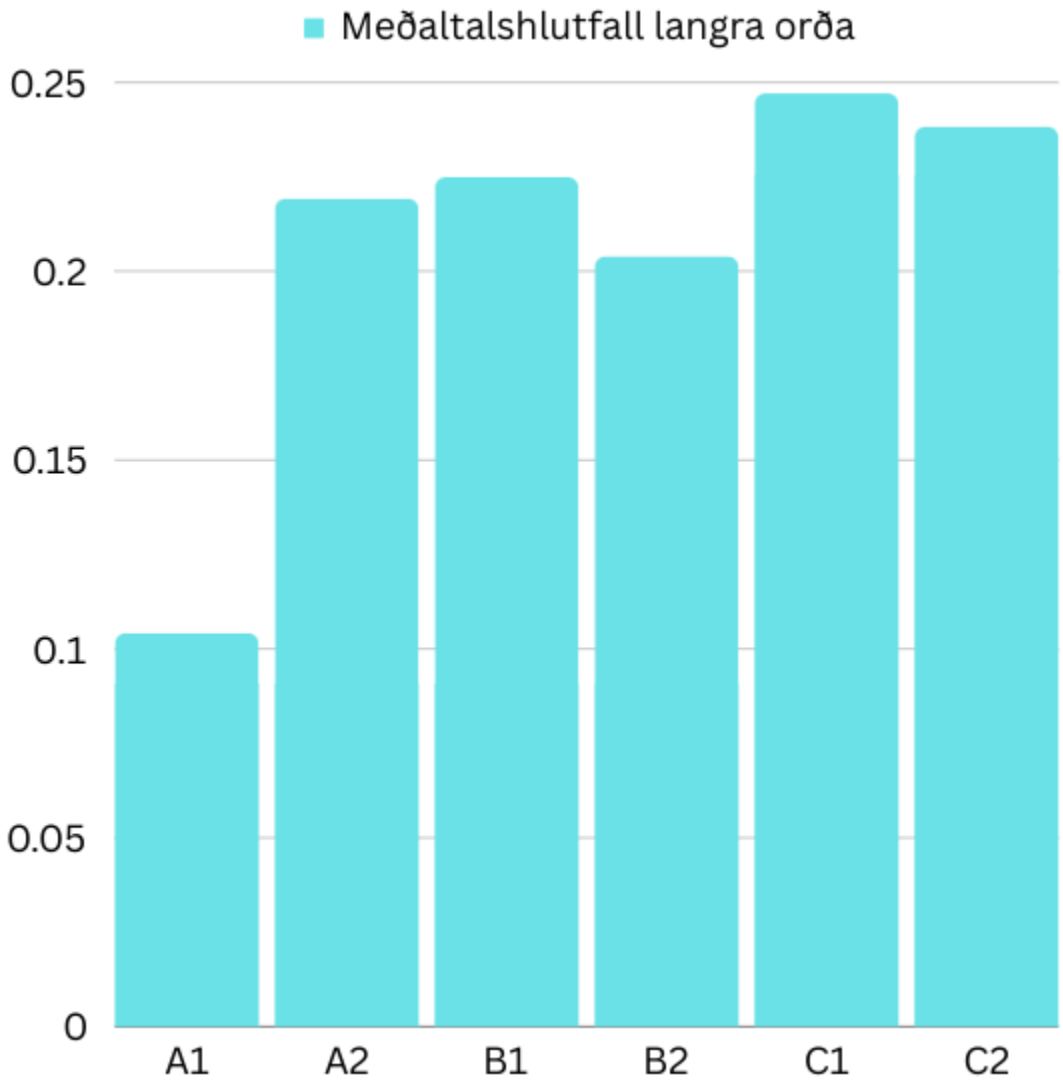
Meðaltalshlutfall sagna í þátíð (ekki lýsingarháttur þátíðar) af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 0.0
A2 0.0892998261419314
B1 0.2019754279199242
B2 0.2964736183142704
C1 0.24152408104033898
C2 0.1737096944276736



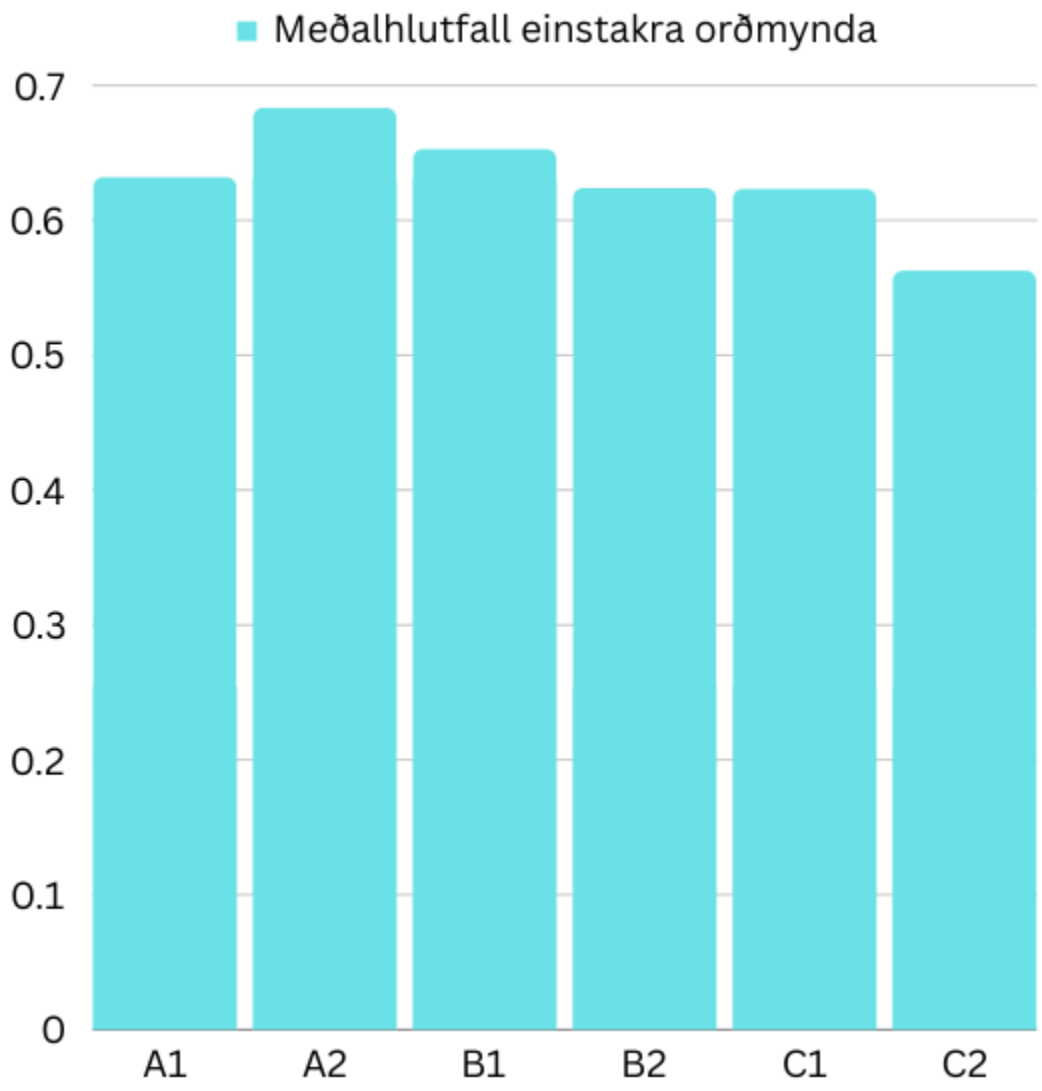
Meðaltalshlutfall langra orða (orð lengri en 6 stafir) (ATH sérnöfn hafa áhrif, þau eru meðtalin hér) í hverjum texta á hverju hæfnistigi

A1 0.1042074320478958
A2 0.2192312579106553
B1 0.22490559910009533
B2 0.2039544288507499
C1 0.24711762466874412
C2 0.23825870150926423



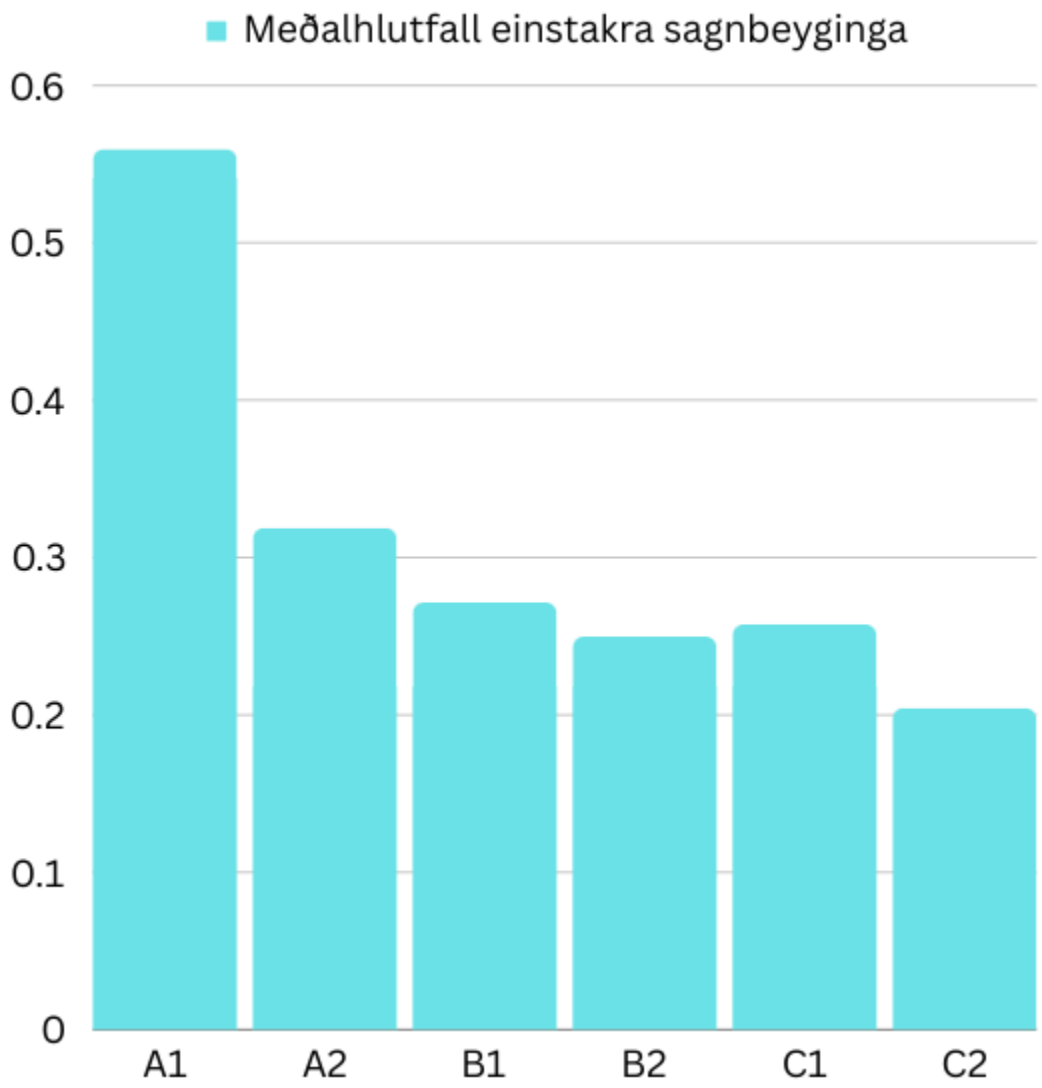
Meðalhlutfall einstakra orðmynda miðað við heildarfjölda orðmynda í hverjum texta á hverju hæfnistigi

A1 0.6321713613722922
A2 0.6835859500697276
B1 0.6529844828322037
B2 0.624007686100751
C1 0.6234258394783321
C2 0.5627908141818788



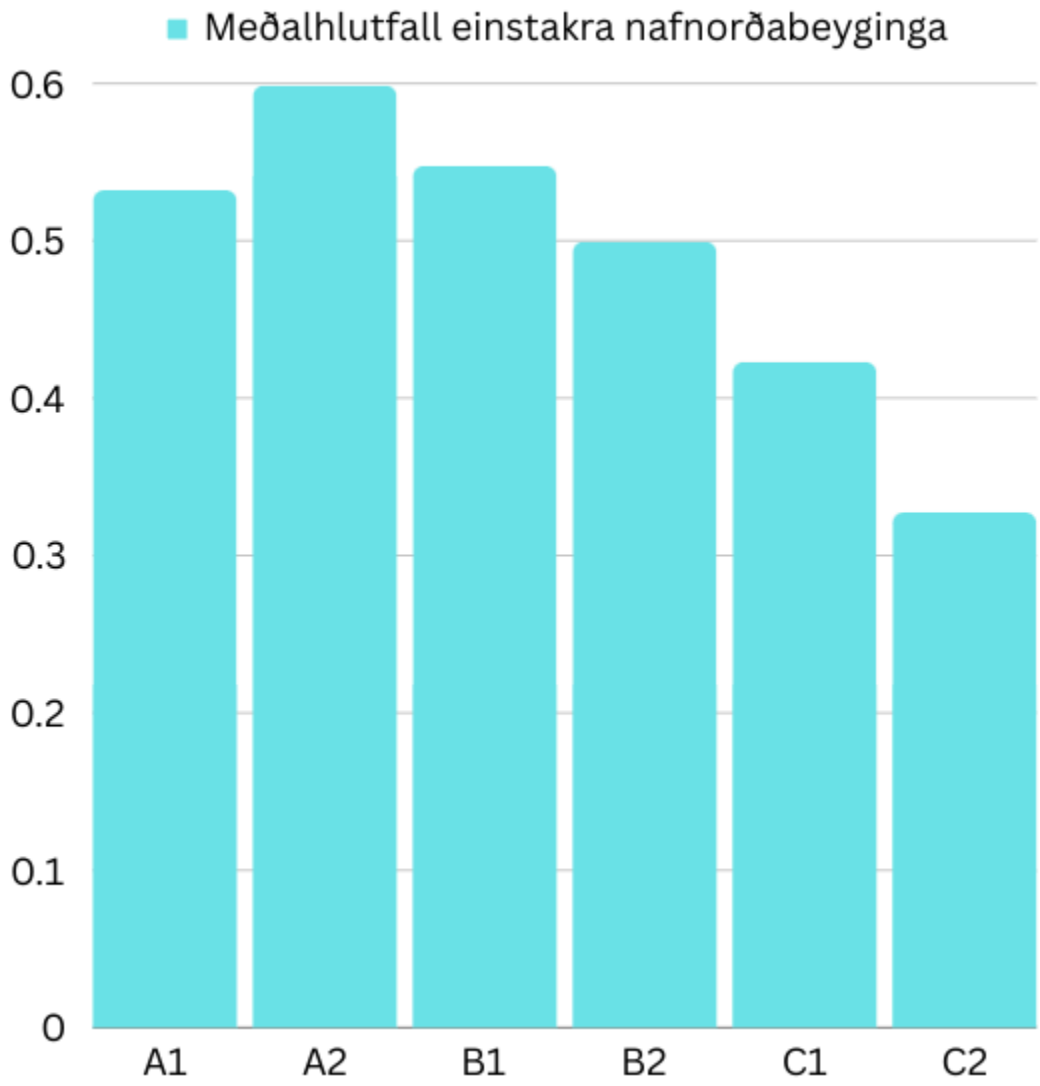
Meðalhlutfall einstakra sagnbeyginga (þ.e. einstakra marka sagna) miðað við heildarfjölda sagna í hverjum texta á hverju hæfnistigi

A1 0.5594708994708995
A2 0.3187235631376782
B1 0.27149908937455575
B2 0.2497513726434087
C1 0.25747262601002535
C2 0.2043064043136832



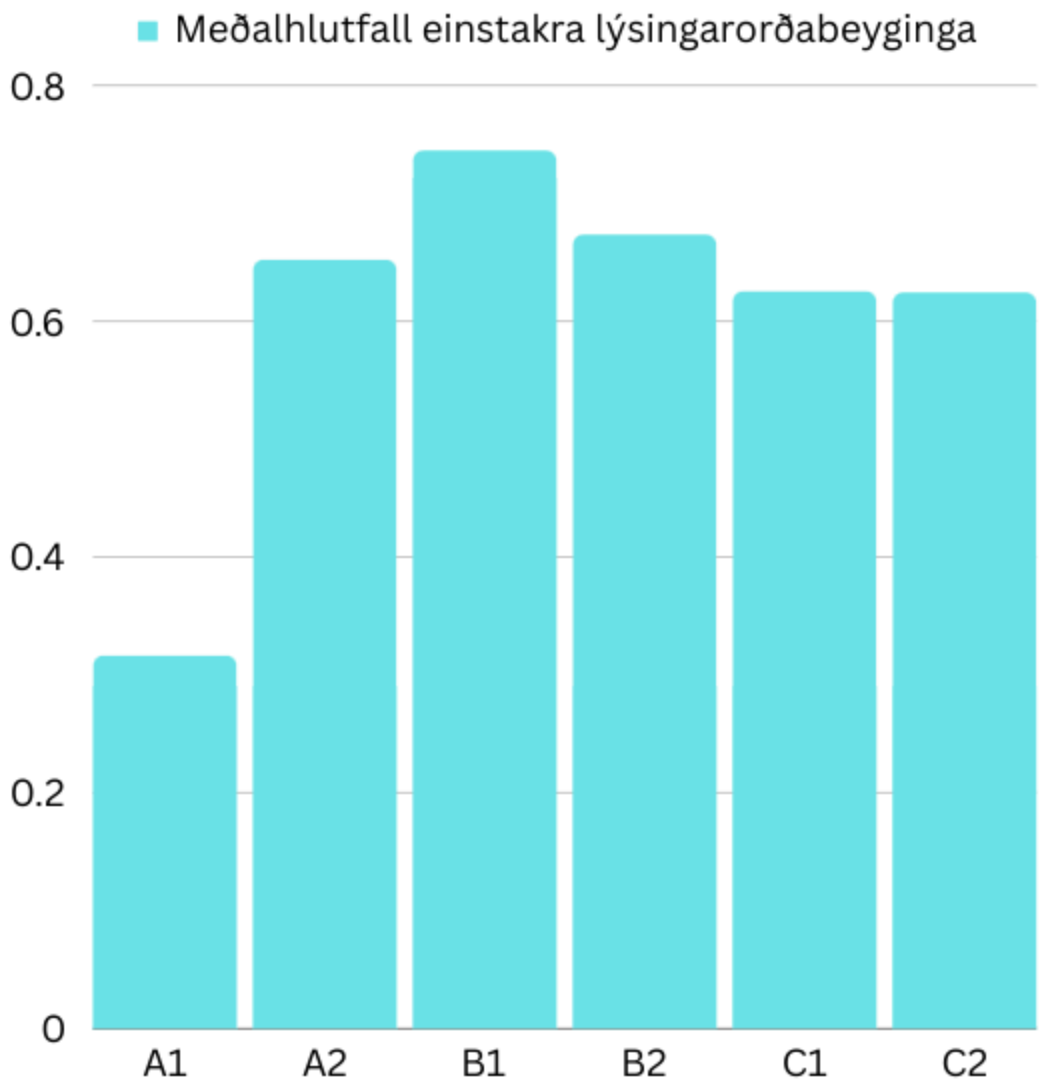
Meðalhlutfall einstakra nafnorðabeyginga (þ.e. einstakra marka nafnorða) miðað við heildarfjölda nafnorða í hverjum texta á hverju hæfnistigi

A1 0.532132741985683
A2 0.5986388717238553
B1 0.5475258849762537
B2 0.49941367008294923
C1 0.4229725955298073
C2 0.3274257858225133



Meðalhlutfall einstakra lýsingarorðabeyginga (þ.e. einstakra marka lýsingarorða) miðað við heildarfjölda lýsingarorða í hverjum texta á hverju hæfnistigi

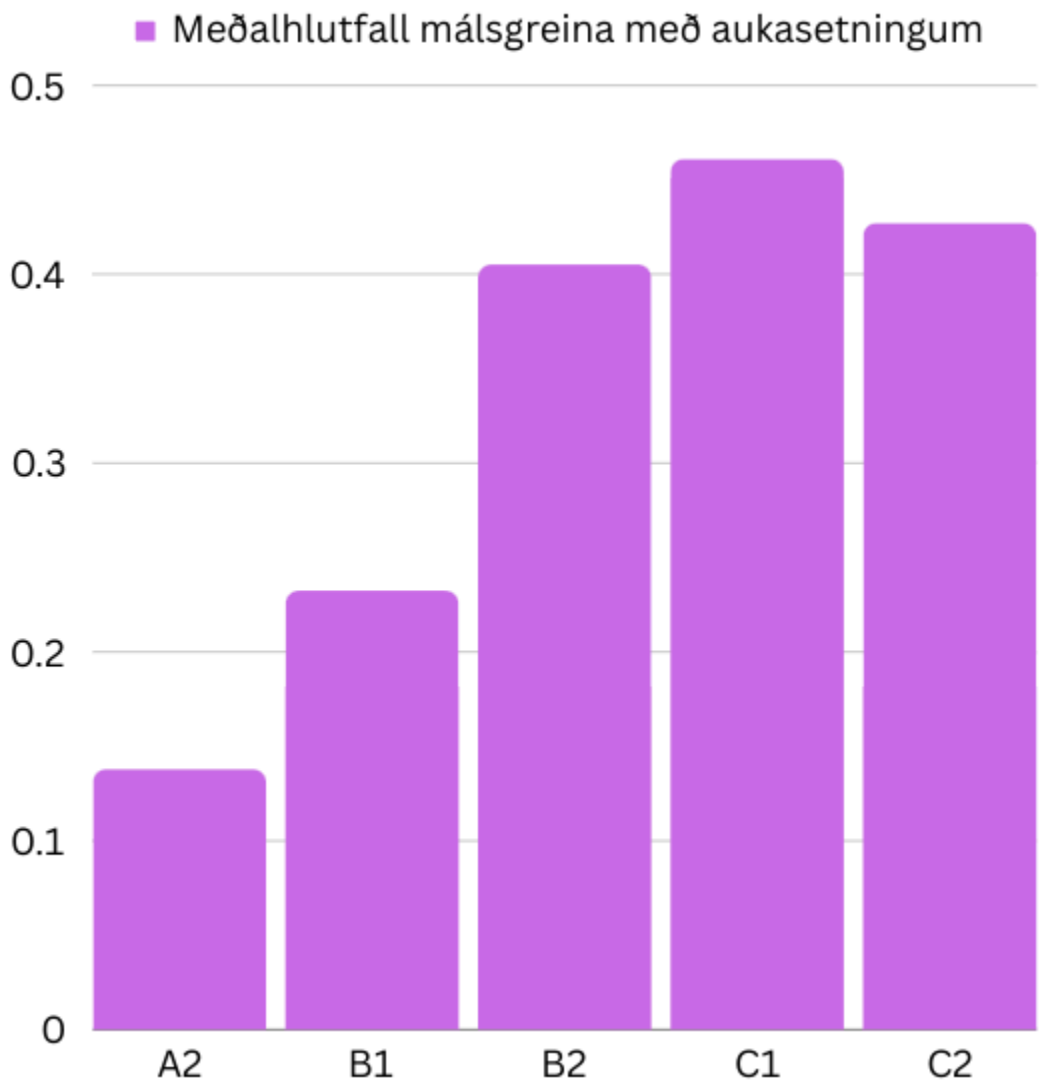
A1 0.31666666666666665
A2 0.6522175536881419
B1 0.7448841381836232
B2 0.6735923954908917
C1 0.6255459460118787
C2 0.6244943129626364


Meðalhlutfall einstakra lýsingarorðabeyginga
0.8
0.6
0.4
0.2
0
A1
A2
B1
B2
C1
C2

Viðauki B - Úr Risamálheildargögnunum

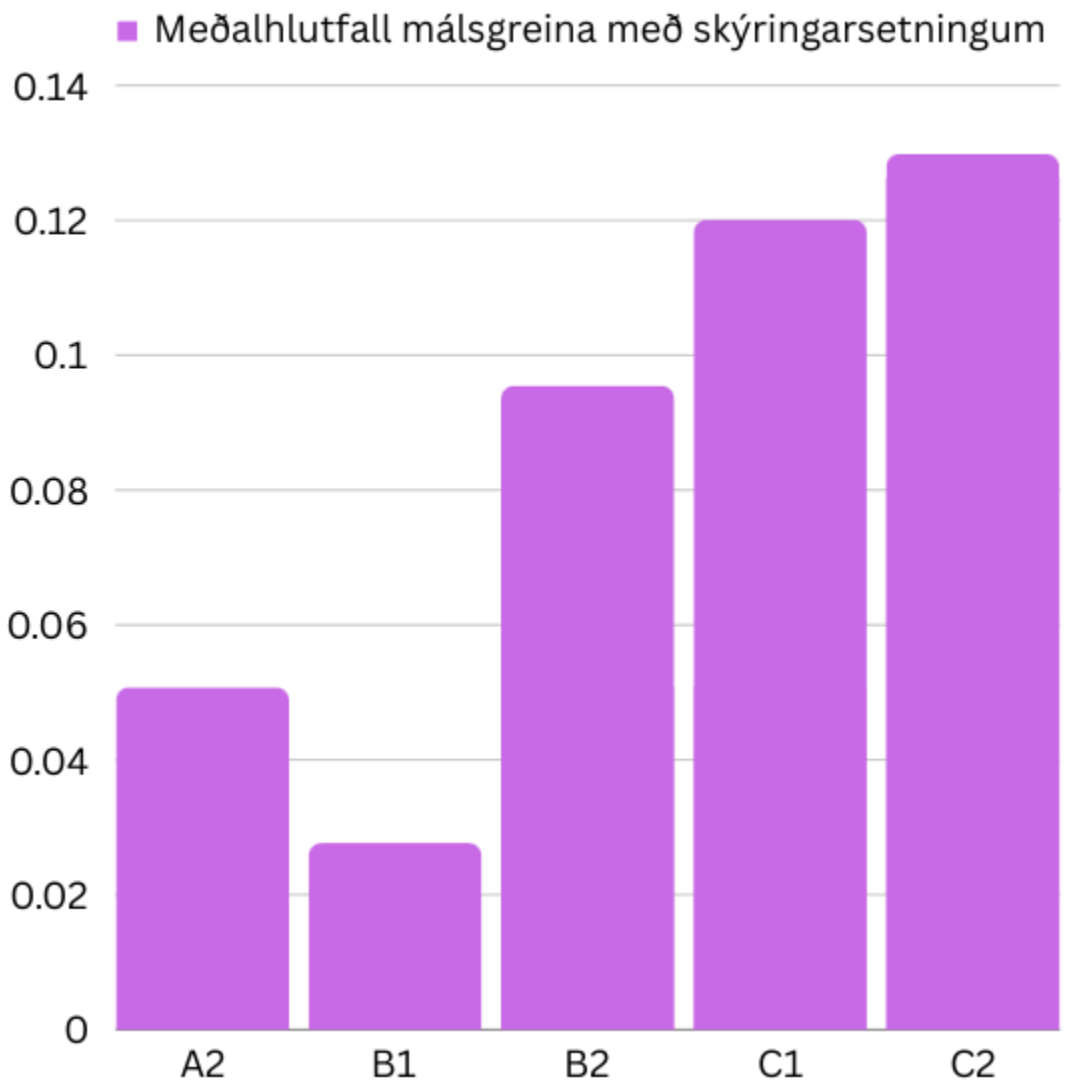
Meðaltalshlutfall málsgreina með aukasetningum í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda aukasetningu, óháð fjölda aukasetninga innan hverrar málsgreinar)

A1 n/a
A2 0.13796616976890375
B1 0.23247265138750653
B2 0.4051838338043445
C1 0.4610639042619768
C2 0.4271404568394142



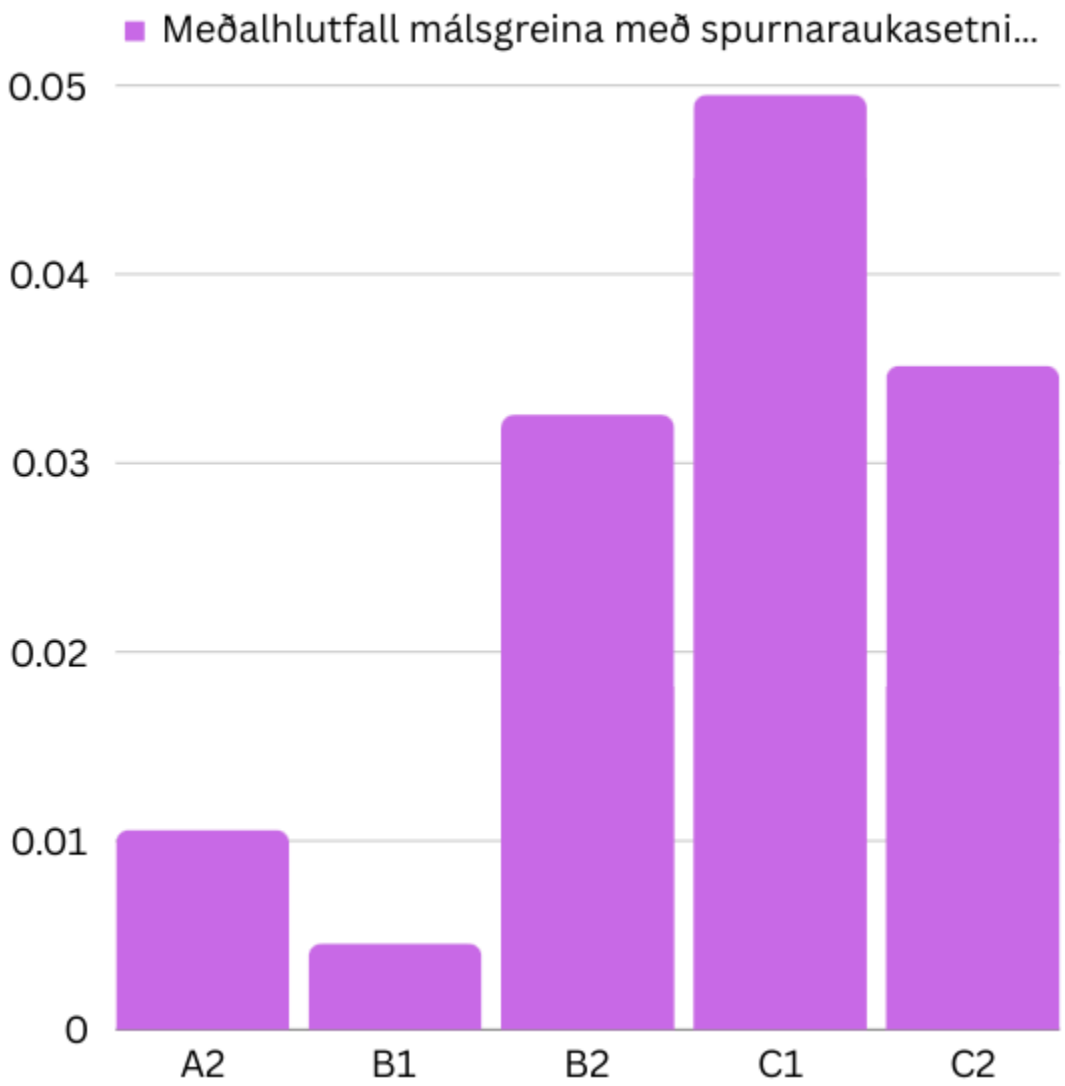
Meðaltalshlutfall málsgreina með skýringarsetningum í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda skýringarsetningu, óháð fjölda skýringarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.05076610594998266
B1 0.027688177563058306
B2 0.0954889208750336
C1 0.12009421359605166
C2 0.1298429915843455



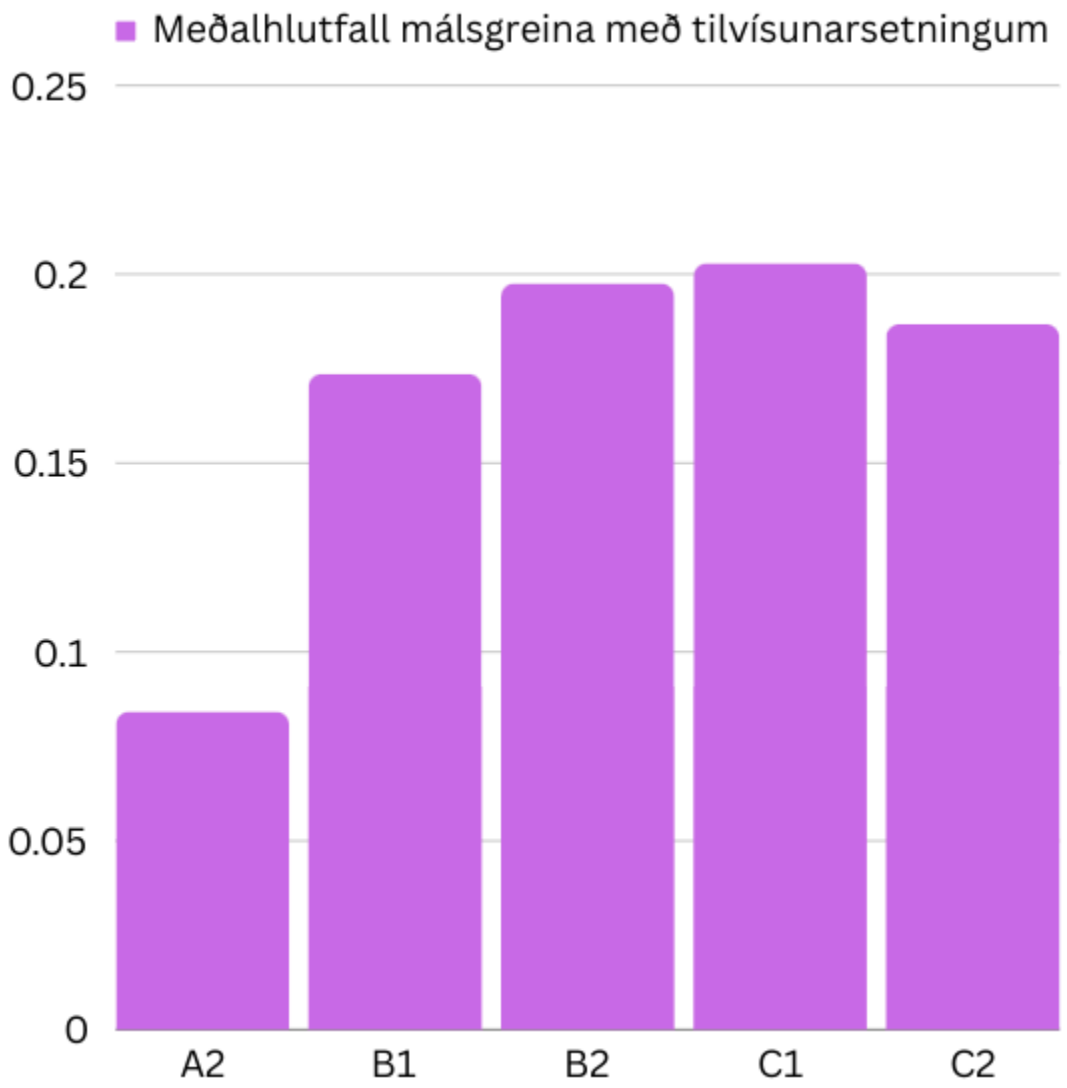
Meðaltalshlutfall málsgreina með spurnaraukasetningar í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda spurnaraukasetningu, óháð fjölda spurnaraukasetninga innan hverrar málsgreinar)

A1 n/a
A2 0.010584312019718718
B1 0.004565463398617829
B2 0.032565543092424605
C1 0.049505278573085
C2 0.03515747345091815



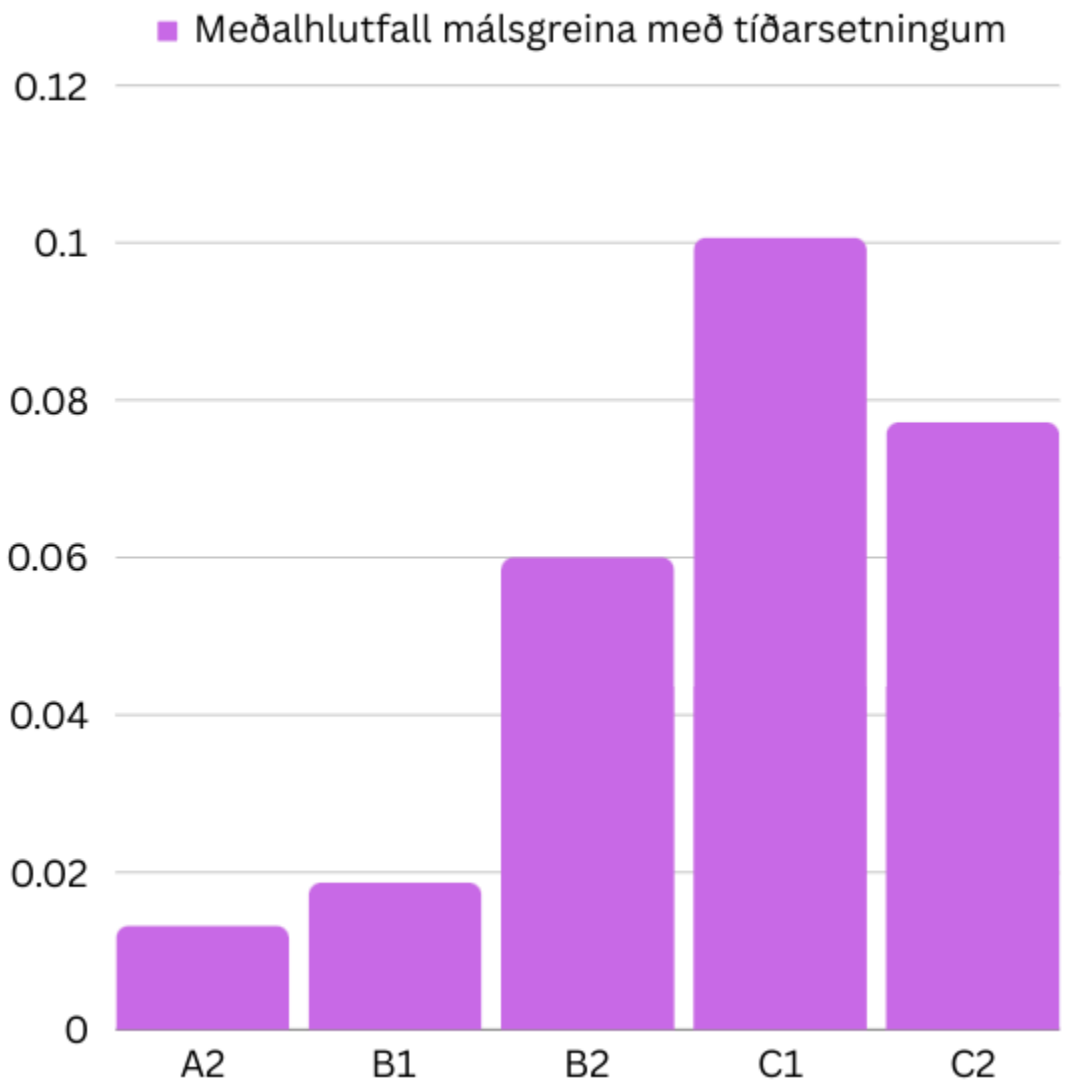
Meðaltalshlutfall málsgreina með tilvísunarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tilvísunarsetningu, óháð fjölda tilvísunarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.08420908840612301
B1 0.17362360917158942
B2 0.1975825648891701
C1 0.202915979437659
C2 0.1867860824152803



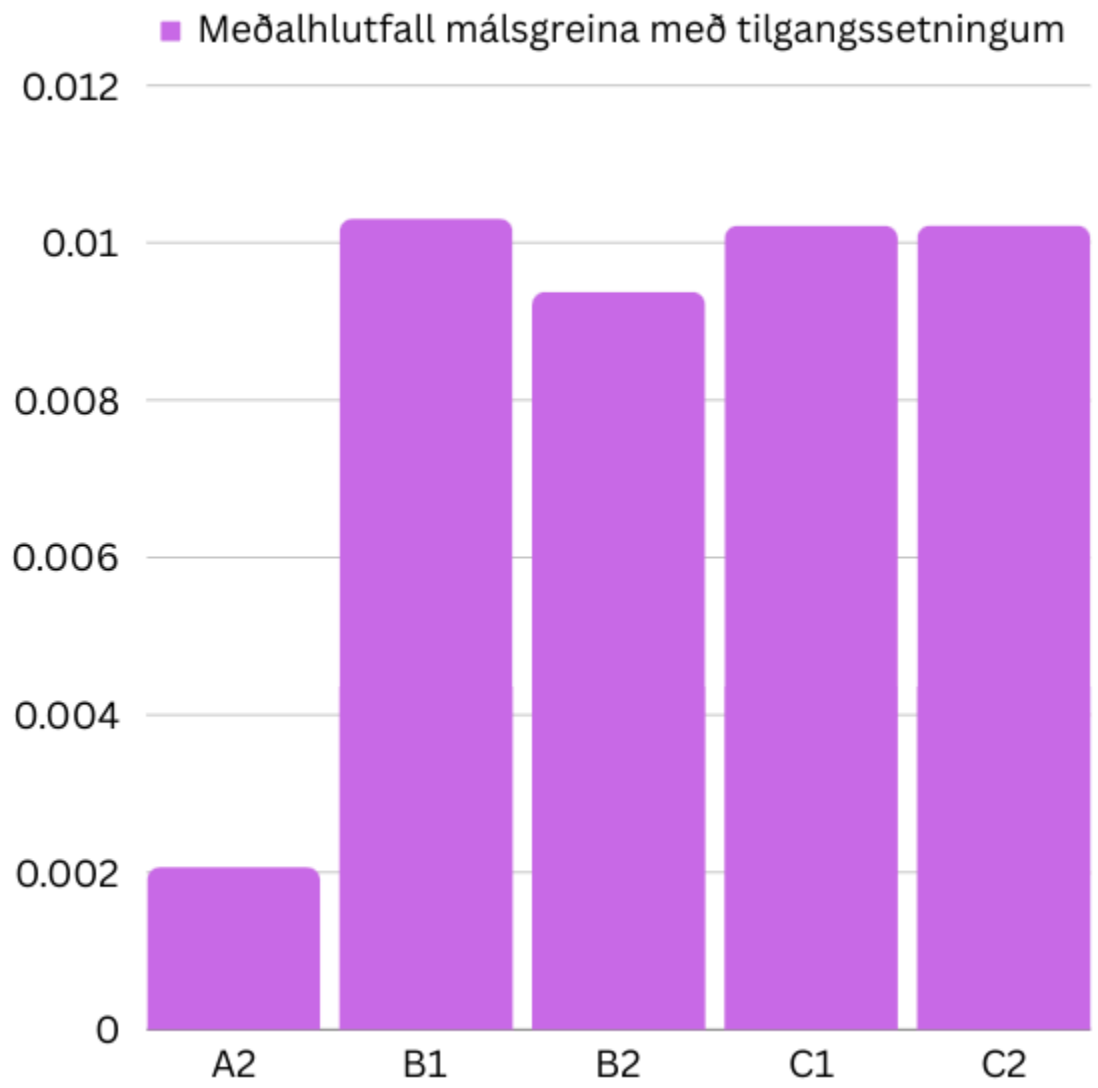
Meðaltalshlutfall málsgreina með tíðarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tíðarsetningu, óháð fjölda tíðarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.013237555256582741
B1 0.01872019096848347
B2 0.060060672909219064
C1 0.10068899144030755
C2 0.07721737496137368



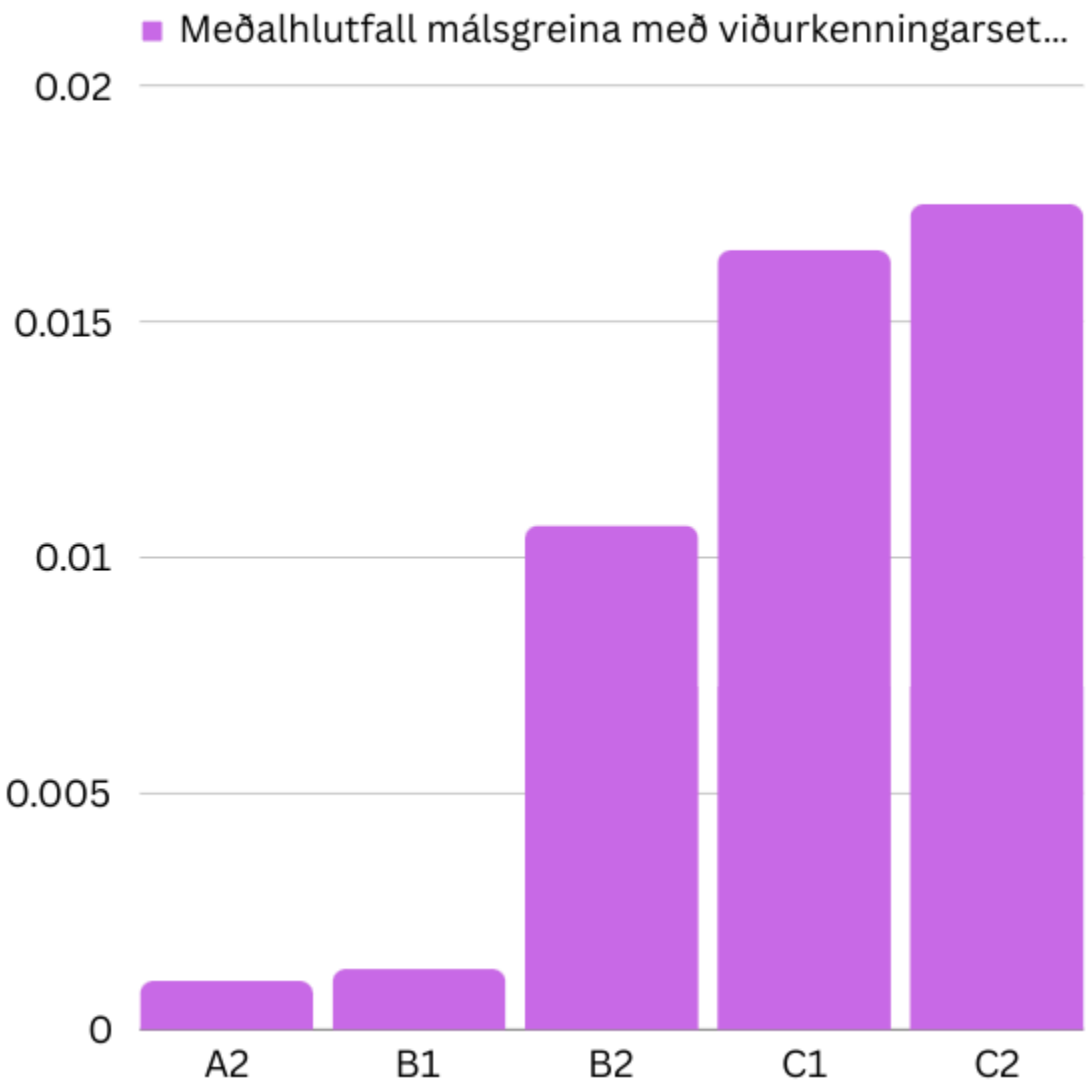
Meðaltalshlutfall málsgreina með tilgangssetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda tilgangssetningu, óháð fjölda tilgangssetninga innan hverrar málsgreinar)

A1 n/a
A2 0.002066115702479339
B1 0.010306698657360284
B2 0.009377237999729665
C1 0.010218872189711178
C2 0.01022093808704551



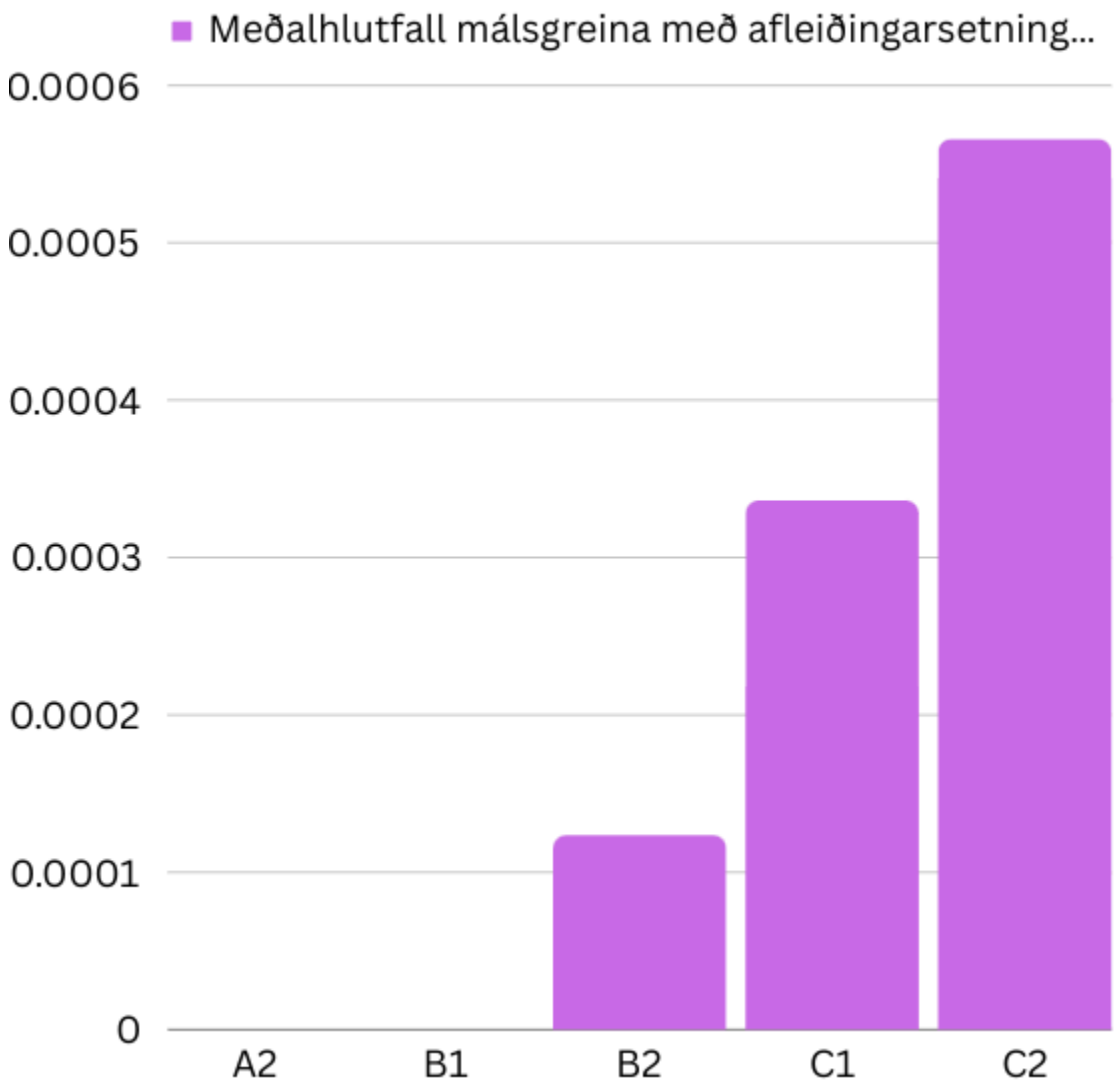
Meðaltalshlutfall málsgreina með viðurkenningarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda viðurkenningarsetningu, óháð fjölda viðurkenningarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.0010330578512396695
B1 0.0012860519455403139
B2 0.01068019541854092
C1 0.016509988936961848
C2 0.017484148827394363



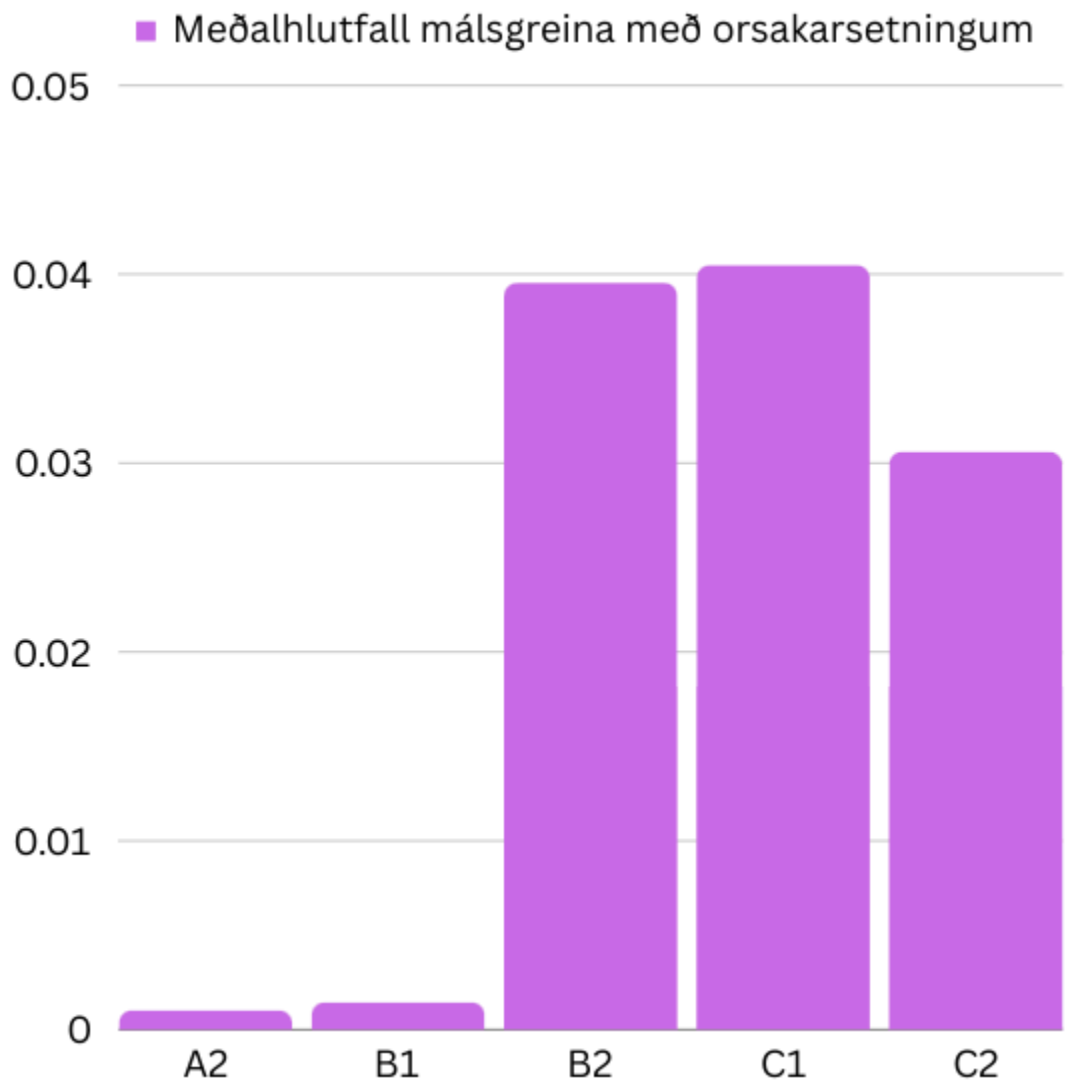
Meðaltalshlutfall málsgreina með afleiðingarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda afleiðingarsetningu, óháð fjölda afleiðingarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.0
B1 0.0
B2 0.00012366887502036214
C1 0.0003363004907292873
C2 0.000565979319287455



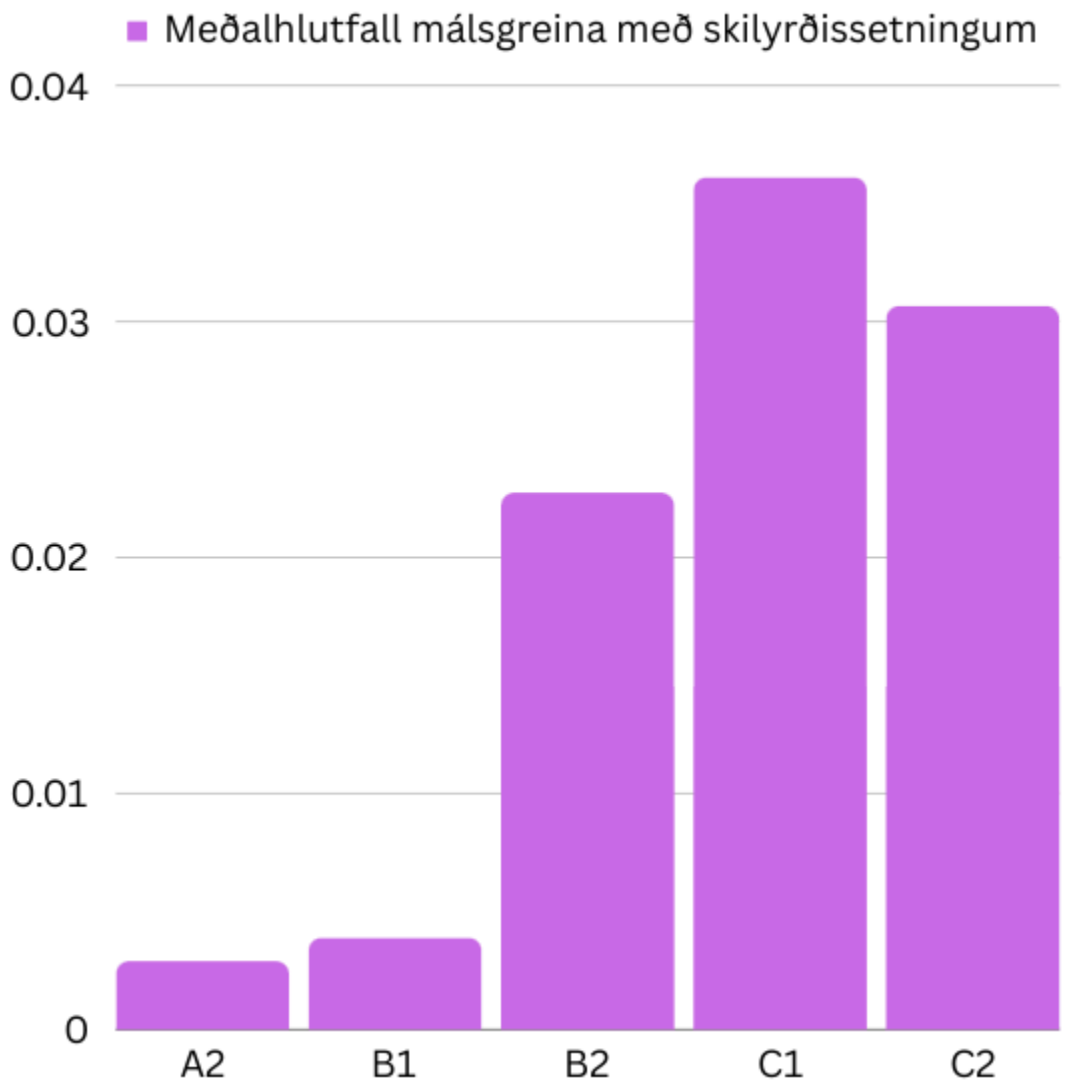
Meðaltalshlutfall málsgreina með orsakarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda orsakarsetningu, óháð fjölda orsakarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.0010330578512396695
B1 0.0014561420516051896
B2 0.03955081544036223
C1 0.04048651419309143
C2 0.030608289172021543



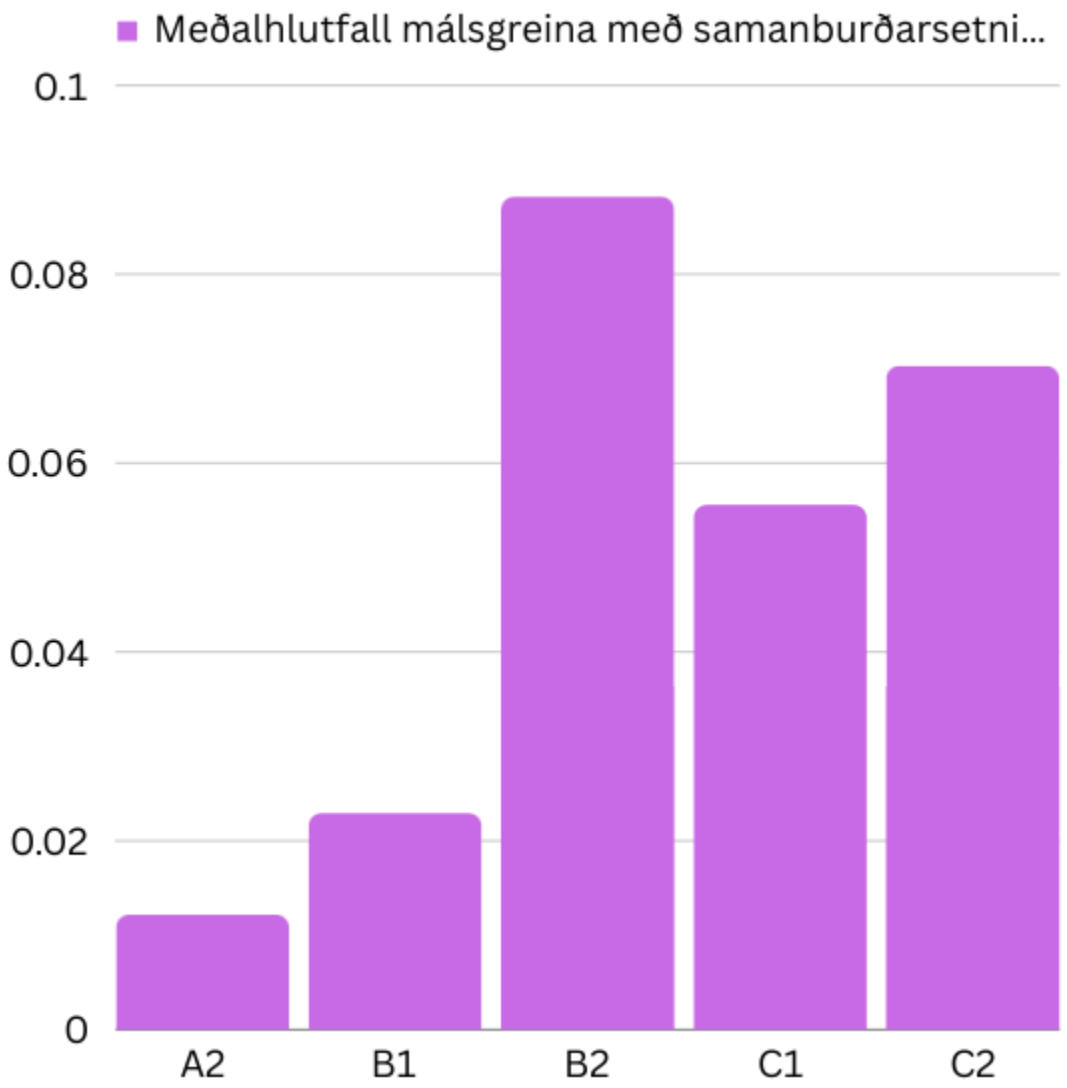
Meðaltalshlutfall málsgreina með skilyrðissetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda skilyrðissetningu, óháð fjölda skilyrðissetninga innan hverrar málsgreinar)

A1 n/a
A2 0.0029069767441860465
B1 0.003896863837228637
B2 0.022750622957121797
C1 0.036109190079785865
C2 0.03065997699547019



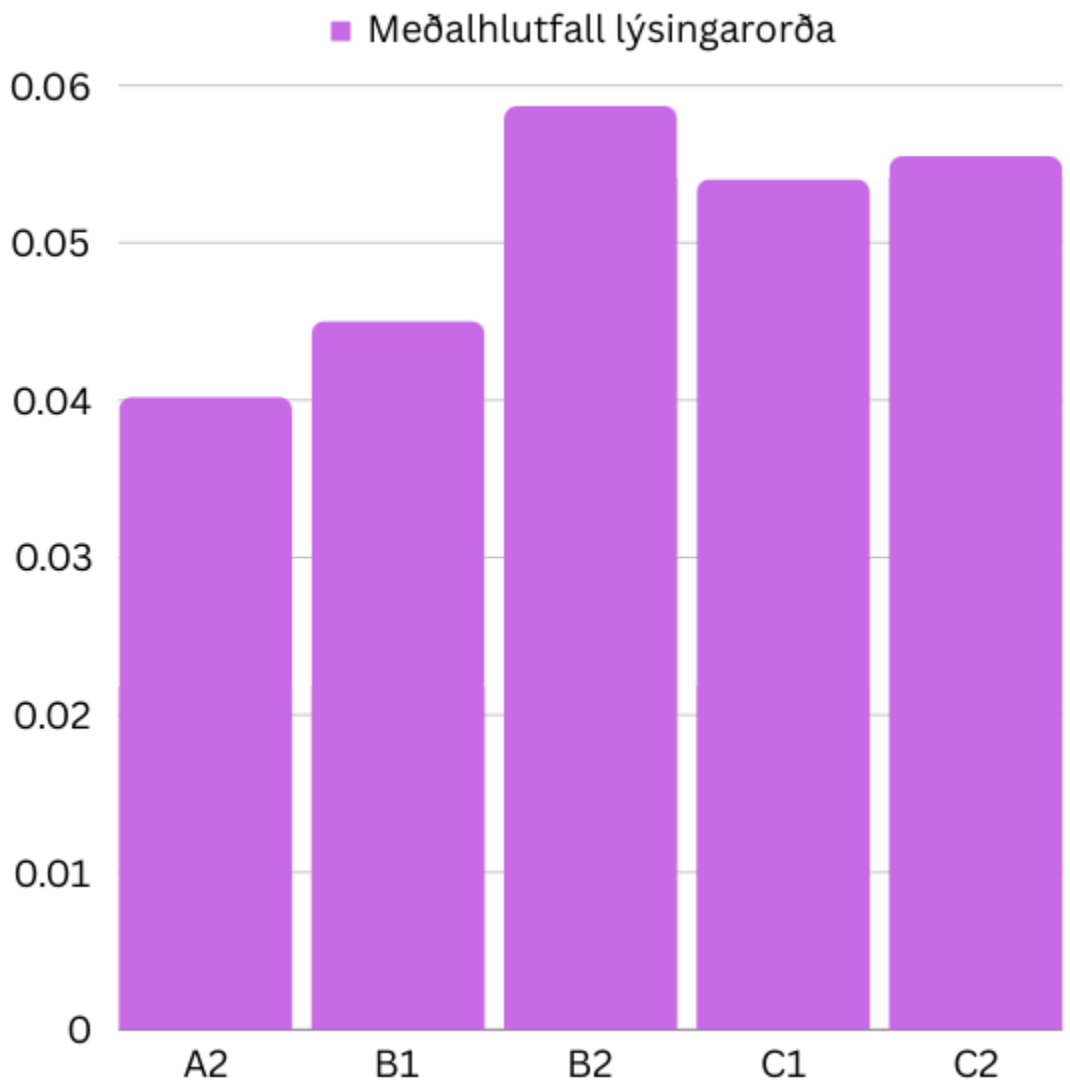
Meðaltalshlutfall málsgreina með samanburðarsetningu í hverjum texta á hverju hæfnistigi (hversu margar málsgreinar innihalda samanburðarsetningu, óháð fjölda samanburðarsetninga innan hverrar málsgreinar)

A1 n/a
A2 0.012191660071783607
B1 0.022942163859551227
B2 0.08822308107453554
C1 0.05559345224805598
C2 0.0703059649149253



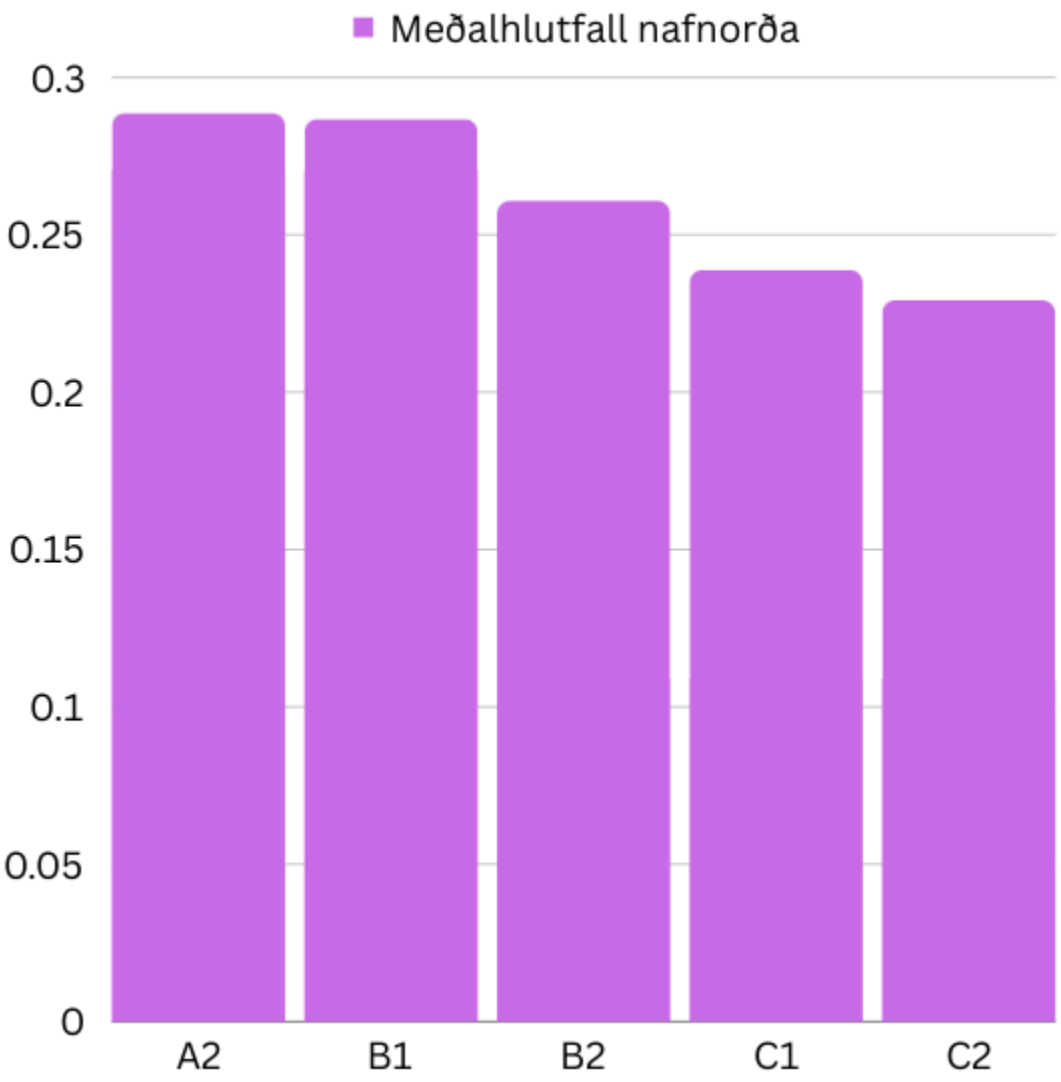
Meðaltalshlutfall lýsingarorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.04020809673608151
B1 0.04502414756314852
B2 0.05869779199249374
C1 0.05404396235284822
C2 0.0555046255728585



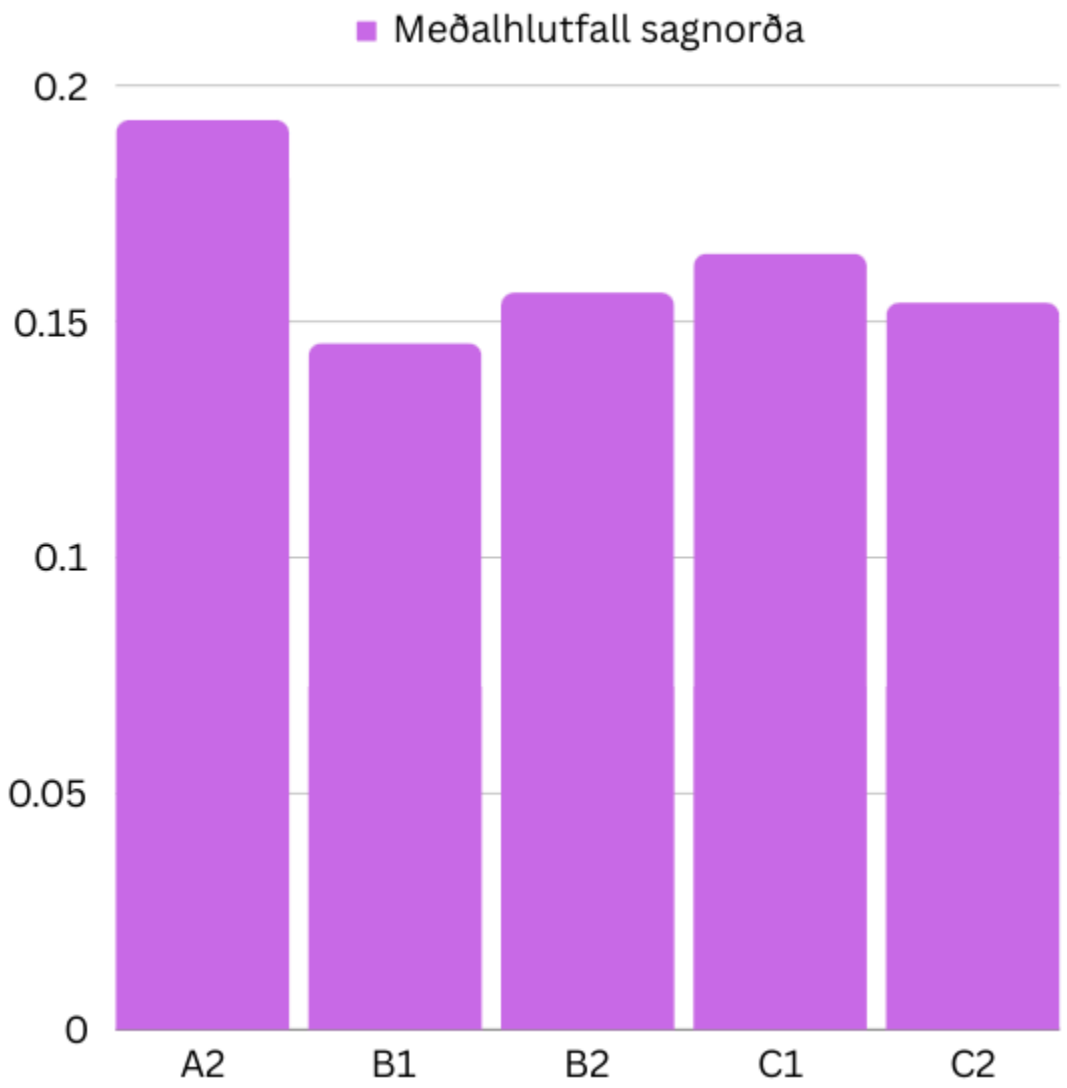
Meðaltalshlutfall nafnorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.2887020255143254
B1 0.28677644545481085
B2 0.2608325805098298
C1 0.23888357069263966
C2 0.2292317327461842



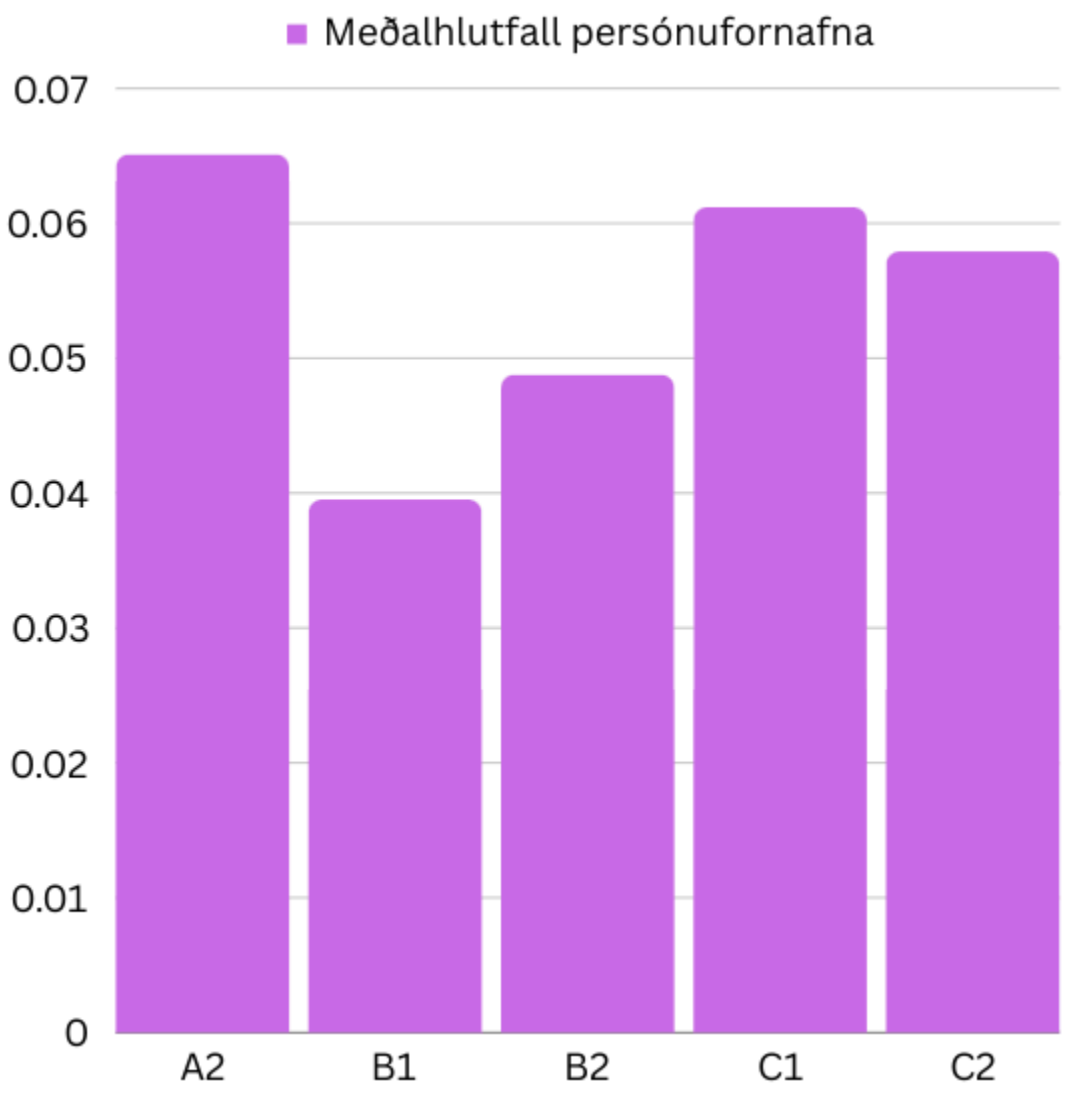
Meðaltalshlutfall sagnorða af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.19266002061208
B1 0.14541640468302394
B2 0.15612755545003107
C1 0.16440598713718438
C2 0.15400071082142697



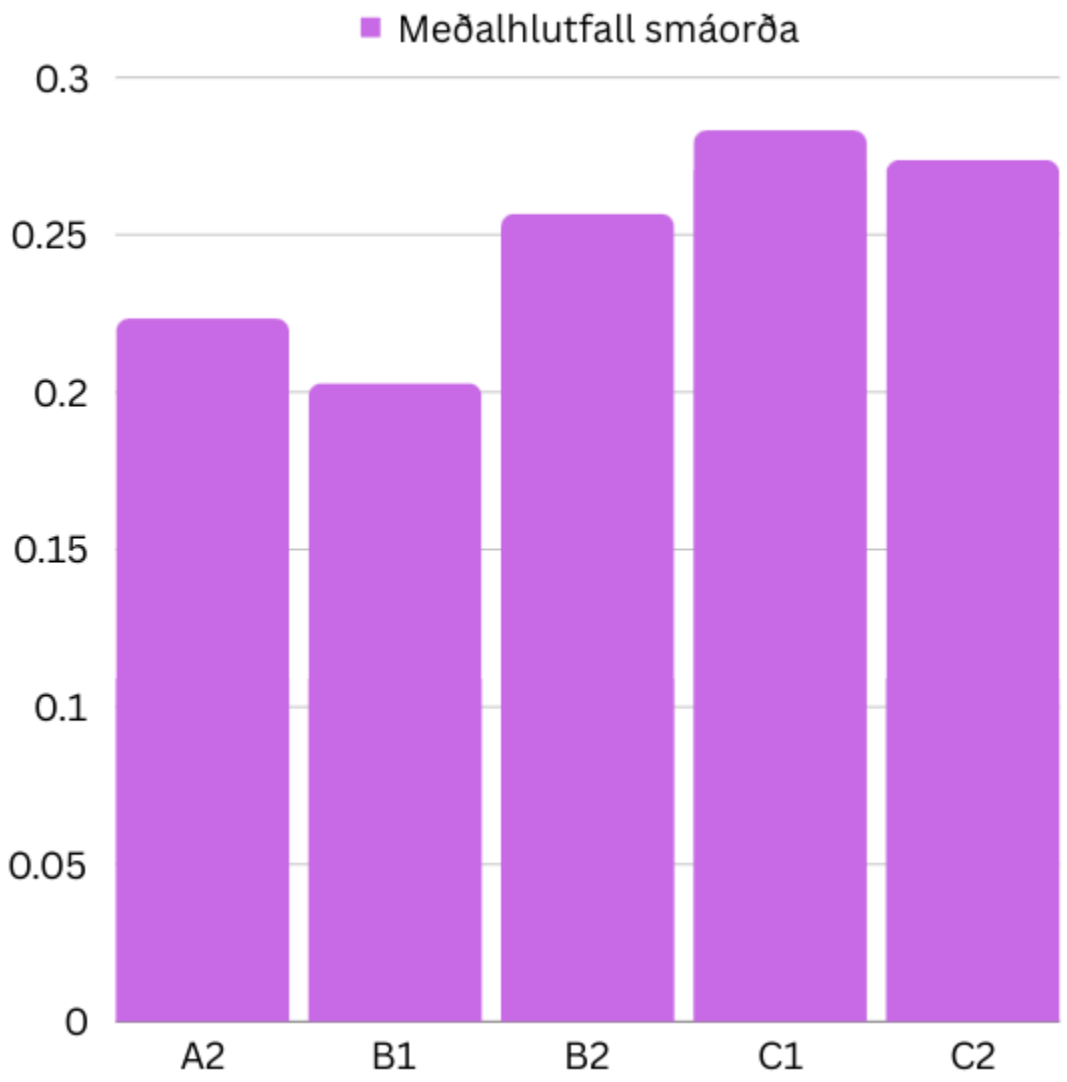
Meðaltalshlutfall persónufornafna af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.06511809205572618
B1 0.03952244282369139
B2 0.04878332266191778
C1 0.06118822388091197
C2 0.05791772843038041



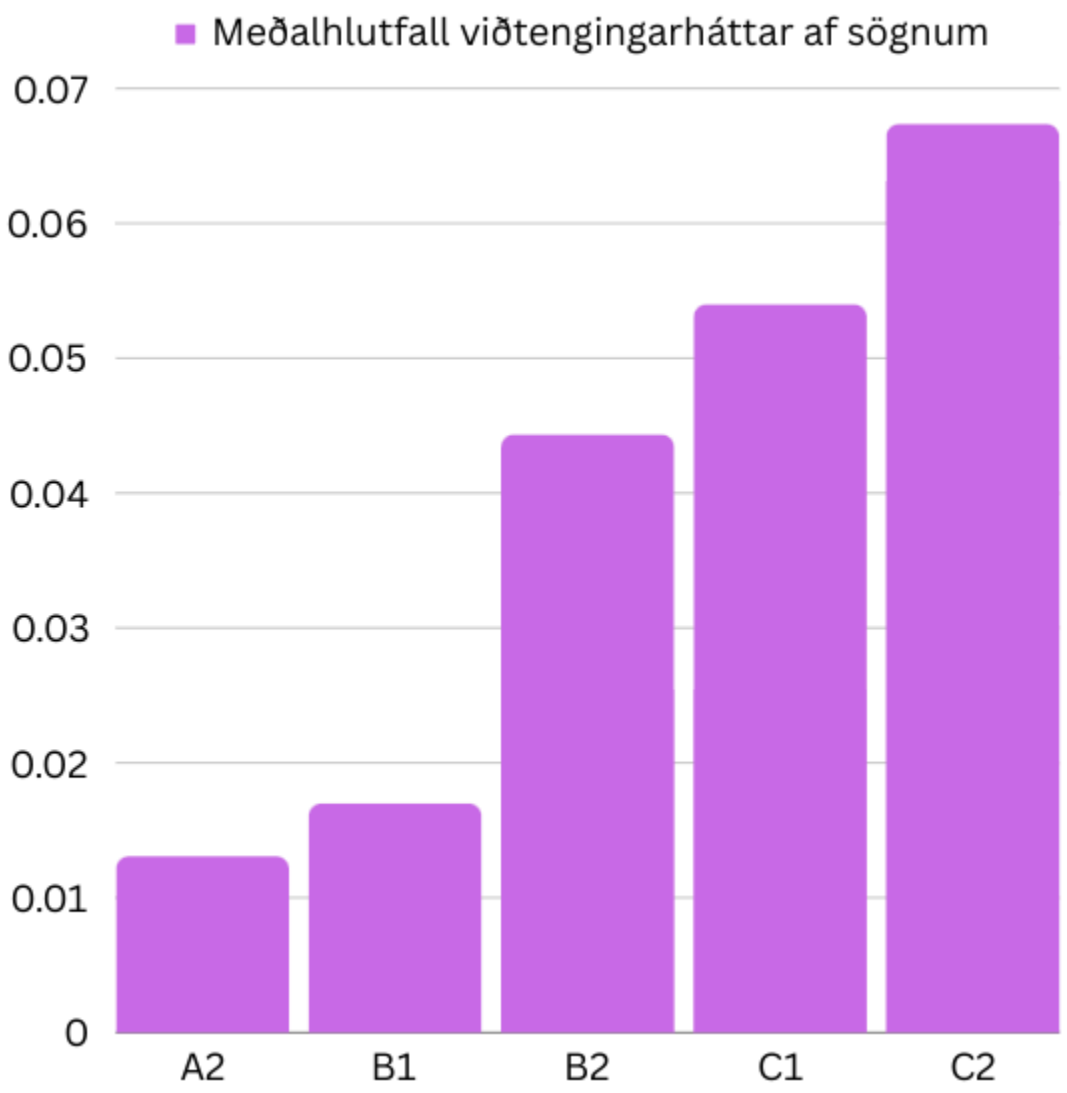
Meðaltalshlutfall smáorða (mörk c og a í markaskránni, ath inniheldur líka atviksorð) af heildarfjölda orða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.22340862000992376
B1 0.2028588731478079
B2 0.25669498640056465
C1 0.28319280625225945
C2 0.2737623206979762



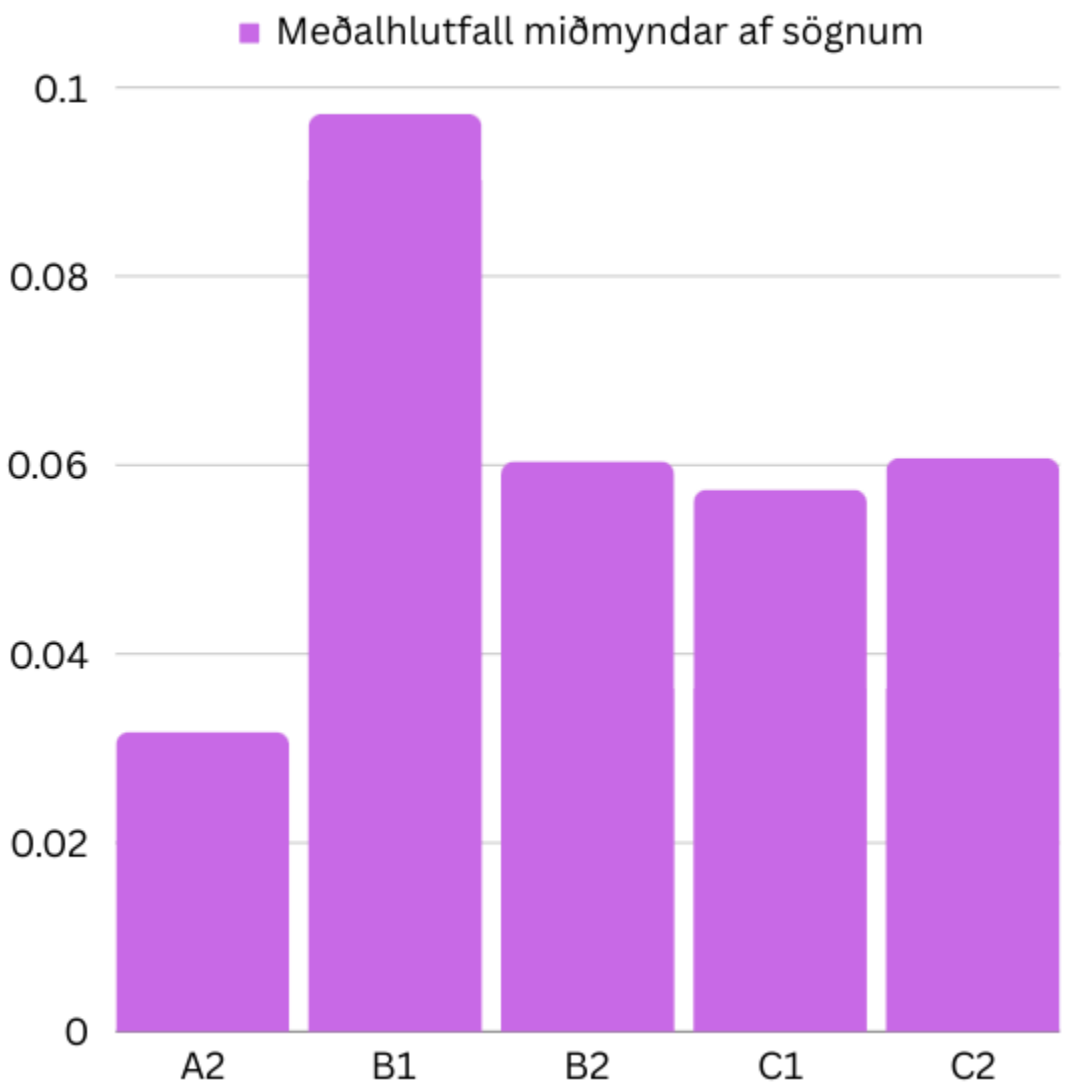
Meðaltalshlutfall viðtengingarháttar af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.013089976714281422
B1 0.017001859289667024
B2 0.044359950048865865
C1 0.054008723142991856
C2 0.06737803708400497



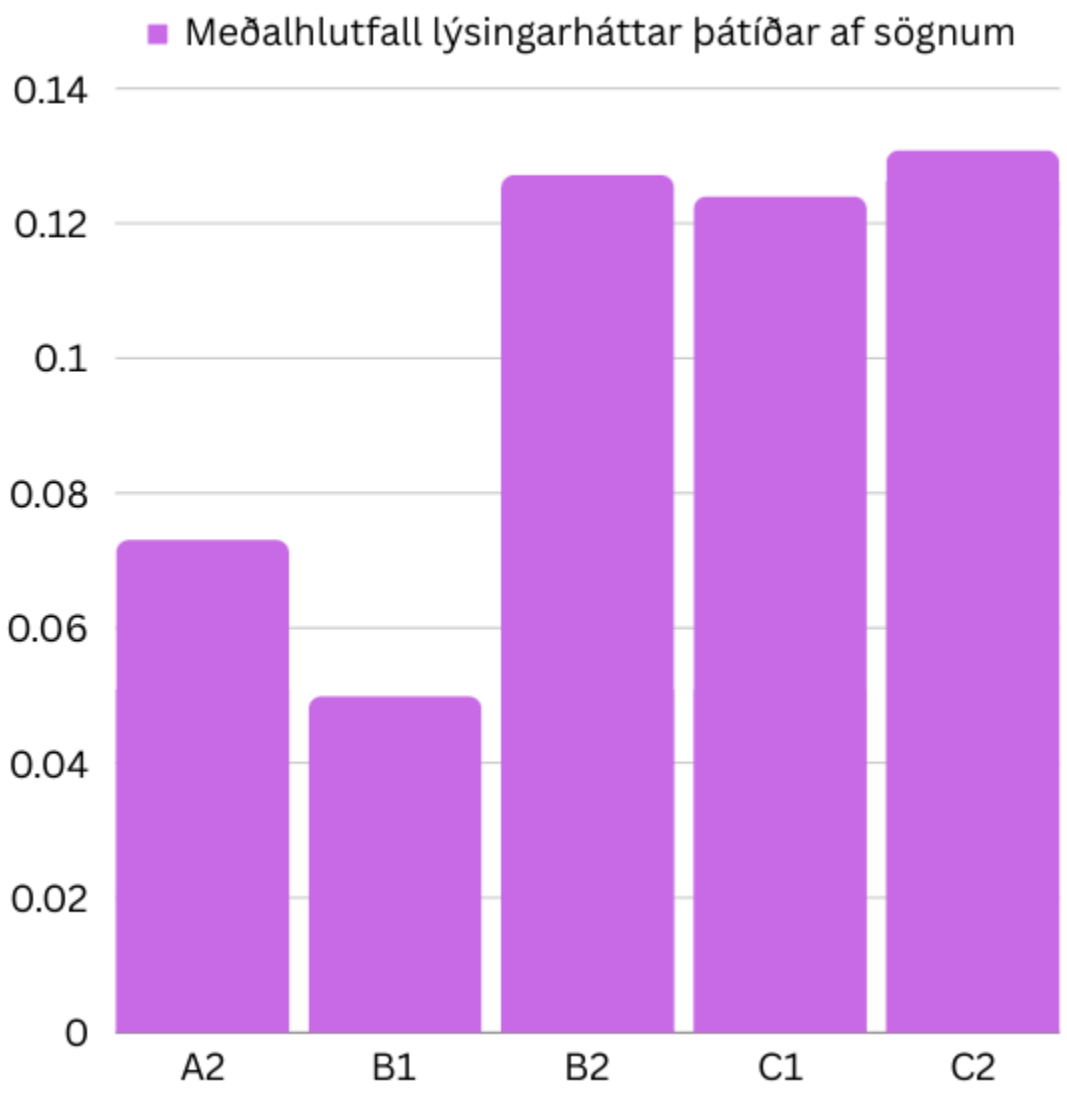
Meðaltalshlutfall miðmyndar af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.031755396235172755
B1 0.09719420227946637
B2 0.060398640972629904
C1 0.05741527923343043
C2 0.060754732478941684



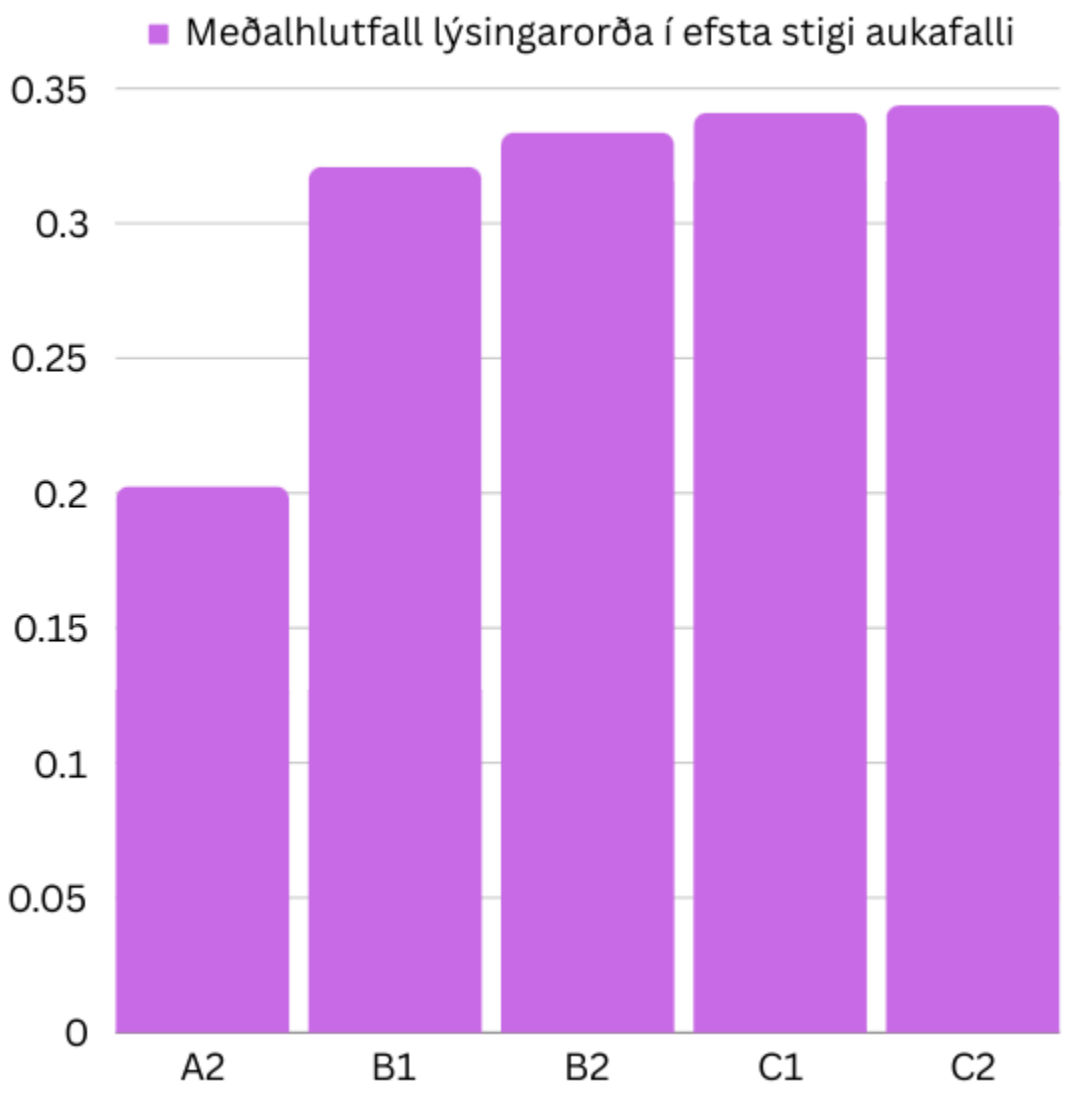
Meðaltalshlutfall lýsingarháttar þátíðar (oftast þolmynd) af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.07307850064073497
B1 0.04990063399165268
B2 0.12716332968421643
C1 0.12401229123783837
C2 0.1308533279106131



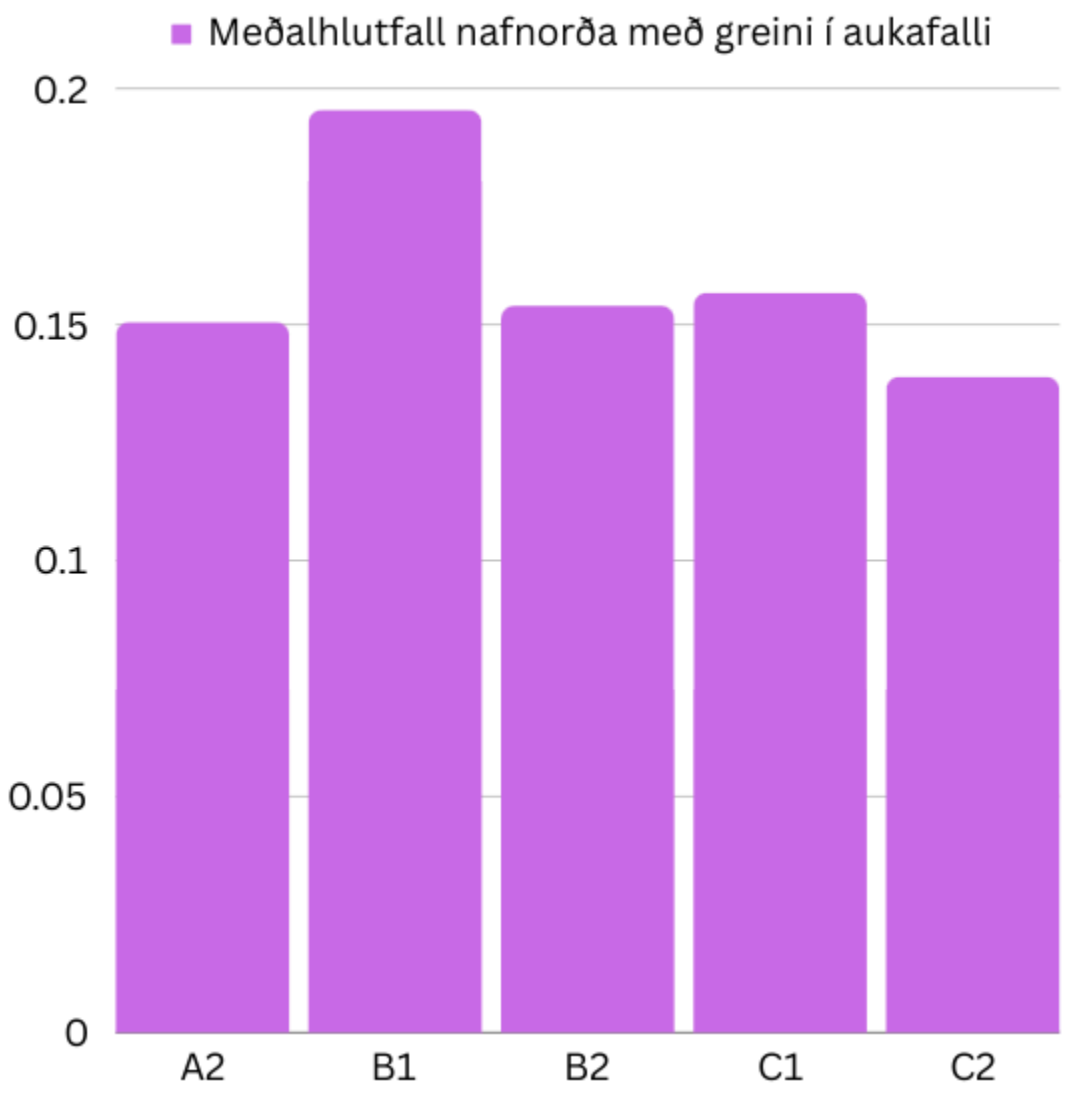
Meðaltalshlutfall lýsingarorða í efsta stigi í öðru en nefnifalli af öllum lýsingarorðum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.2025346477478964
B1 0.32091911893498115
B2 0.3337141622732768
C1 0.34099035558201046
C2 0.3437550113999318



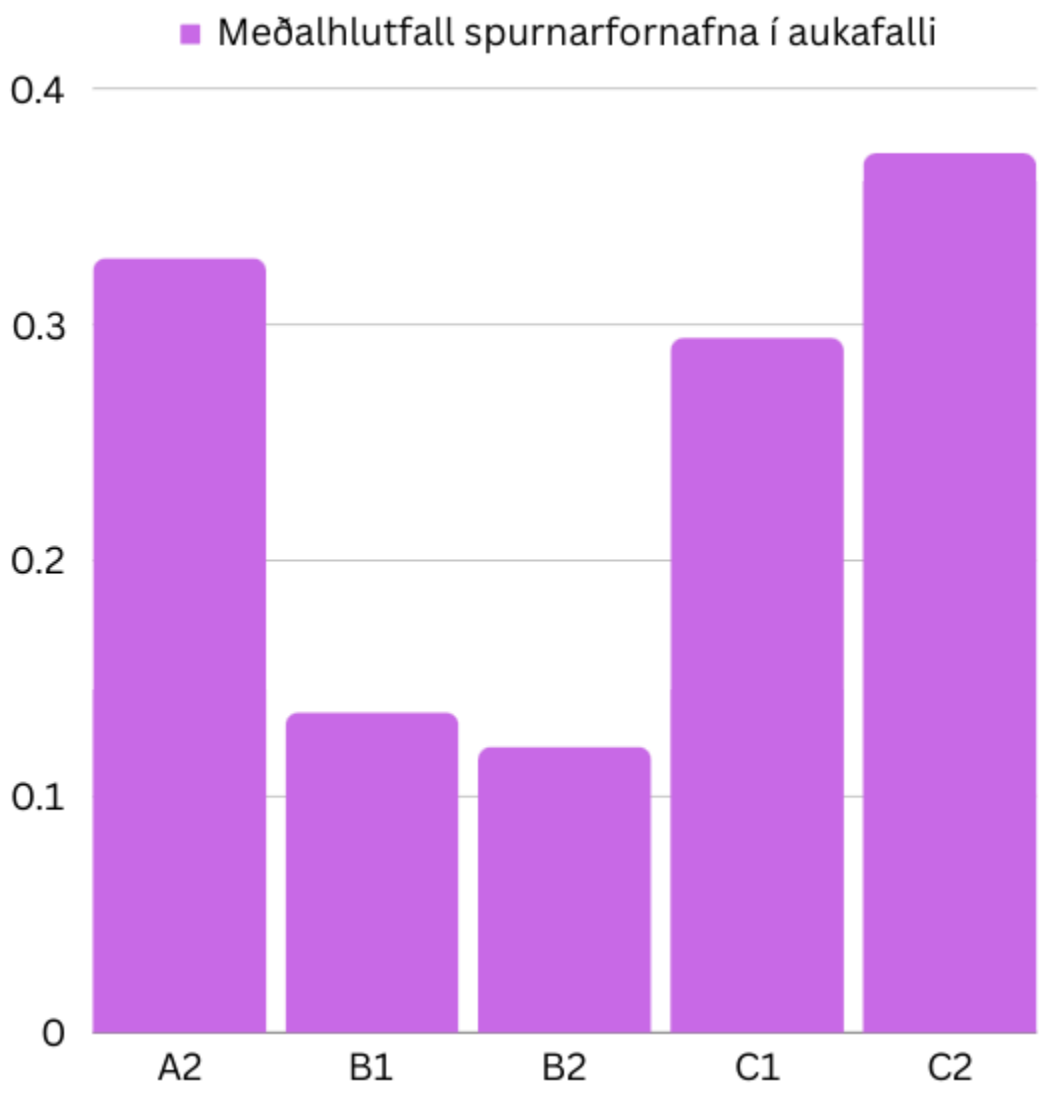
Meðaltalshlutfall nafnorða með greini í öðru en nefnifalli af öllum nafnorðum (ATH sérnöfn hafa áhrif, eru tekin með hér) í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.1505153791601293
B1 0.19544909361431537
B2 0.15398794421193218
C1 0.15672701698227703
C2 0.1389416697106636



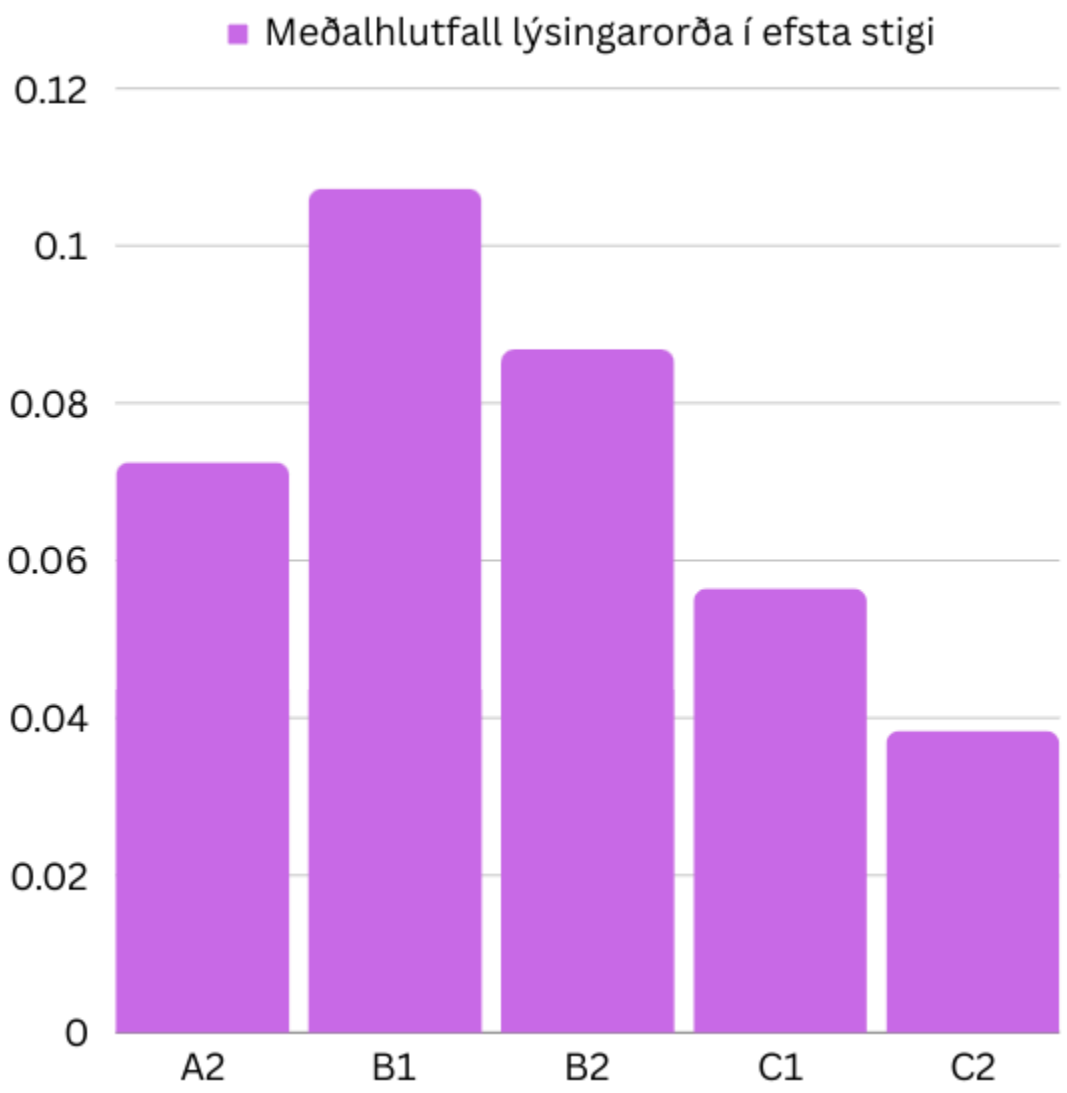
Meðaltalshlutfall spurnarfornafna í öðru en nefnifalli af öllum spurnarfornöfnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.328125
B1 0.1357135276755215
B2 0.12113341235738803
C1 0.29439003444632517
C2 0.372607861301508



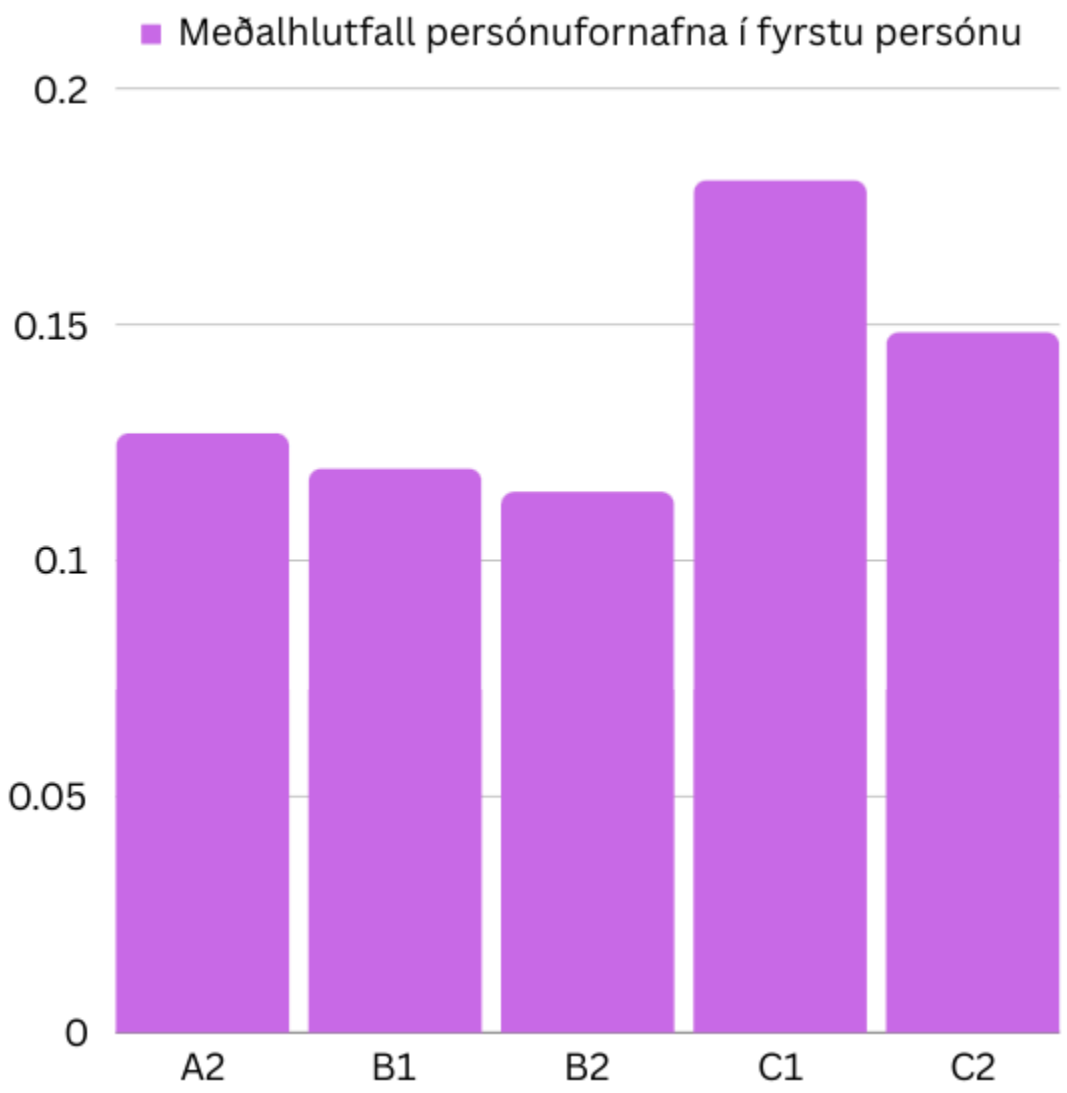
Meðaltalshlutfall lýsingarorða í efsta stigi af öllum lýsingarorðum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.07249216300940438
B1 0.10726350583372948
B2 0.086841058424151
C1 0.05646237244215576
C2 0.038371758511507724



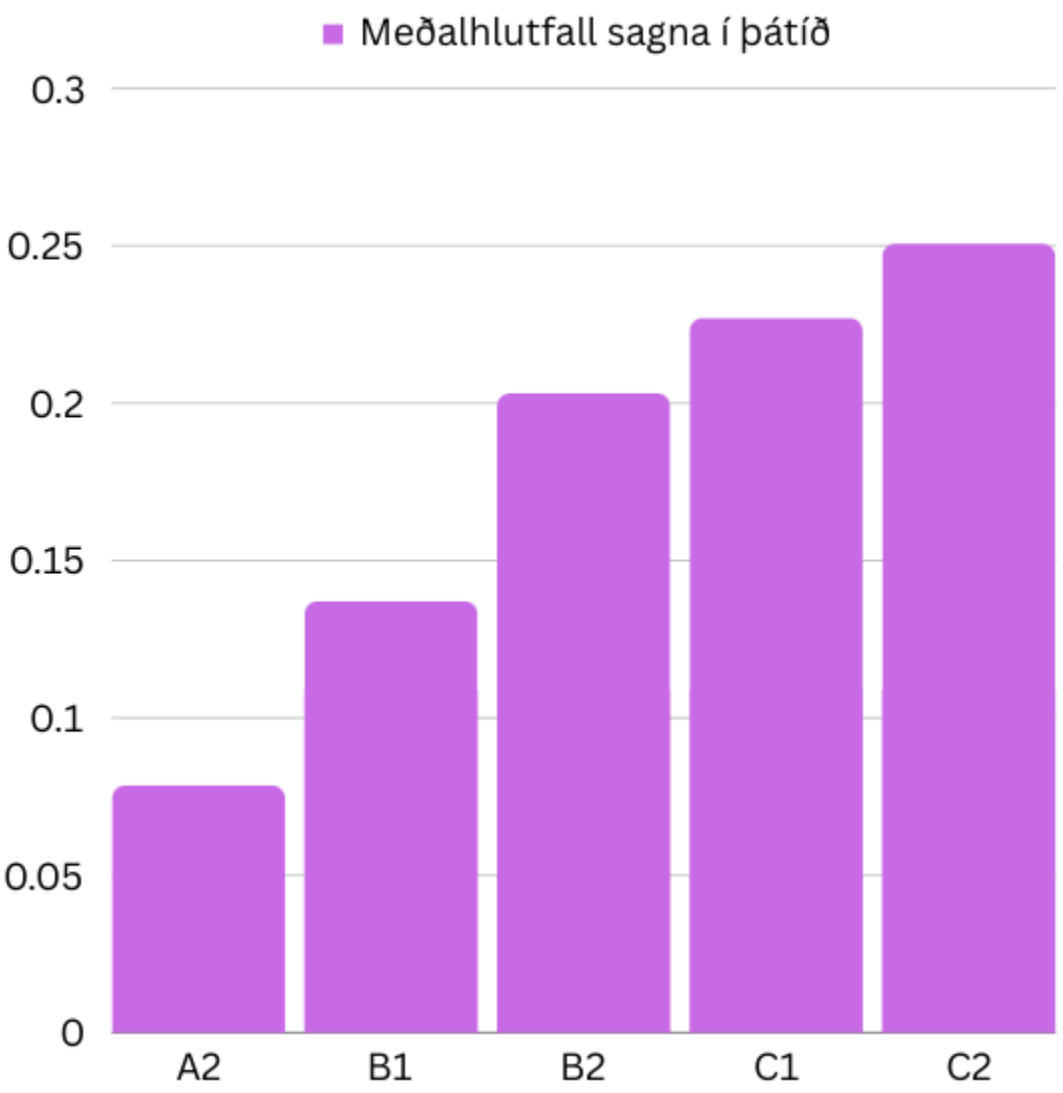
Meðaltalshlutfall persónufornafna í fyrstu persónu af öllum persónufornöfnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.12700022842912745
B1 0.1195343025235114
B2 0.11458888972597876
C1 0.18058830895026437
C2 0.14841938703868582



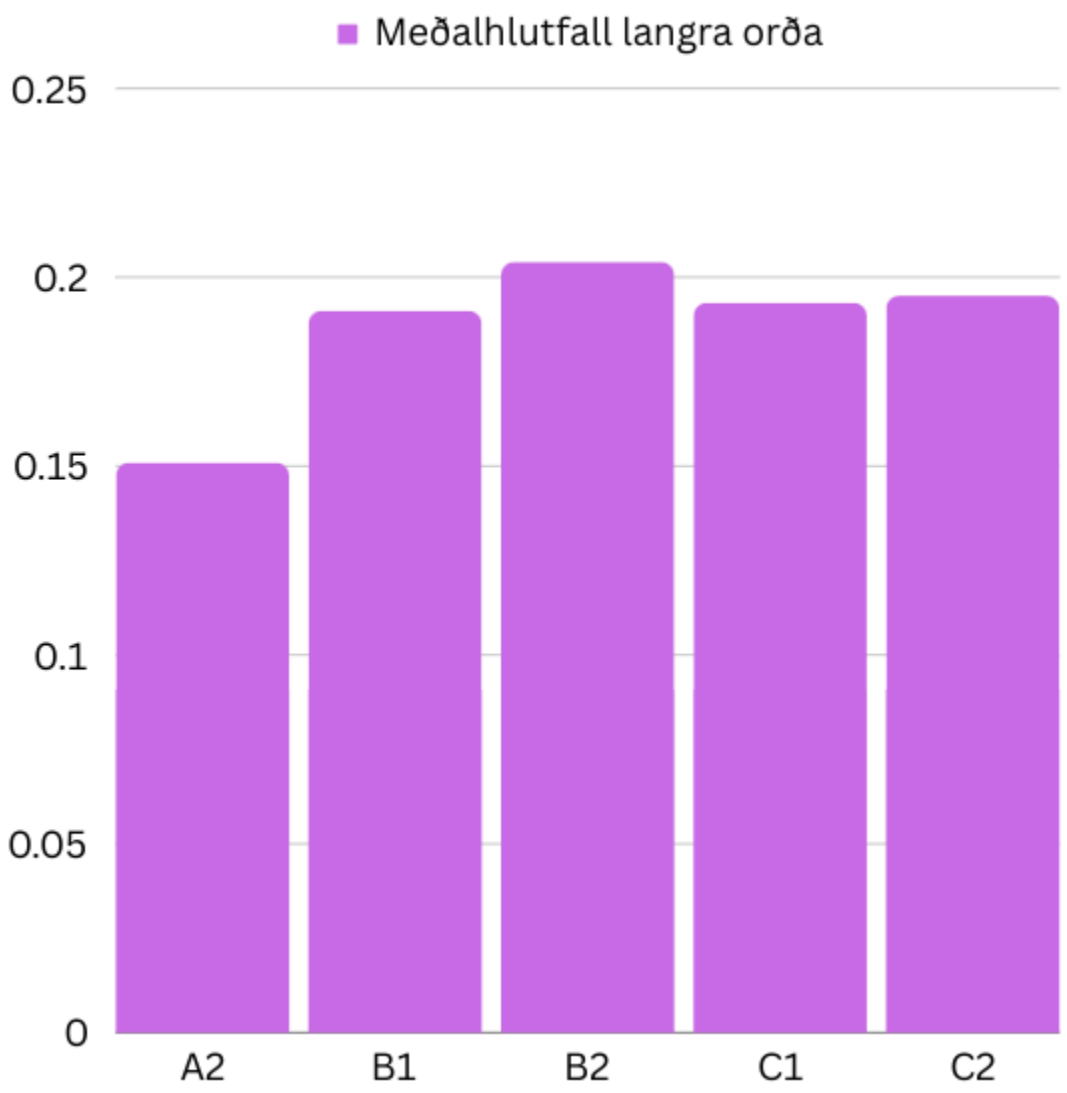
Meðaltalshlutfall sagna í þátíð (ekki lýsingarháttur þátíðar) af öllum sögnum í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.07858328403688934
B1 0.13708977143825862
B2 0.20316387391957674
C1 0.22708330369431726
C2 0.25069850383401193



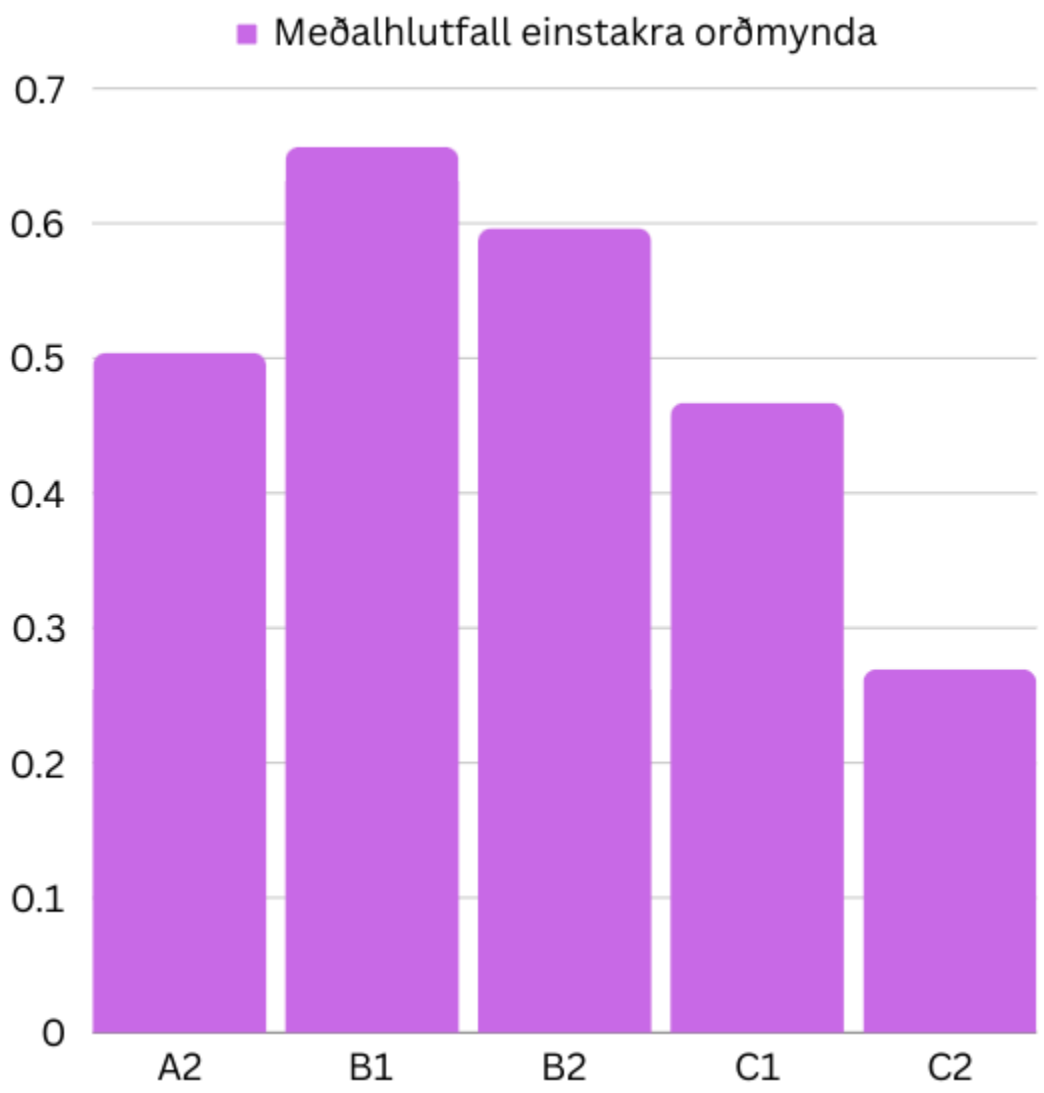
Meðaltalshlutfall langra orða (orð lengri en 6 stafir) (ATH sérnöfn hafa áhrif, þau eru meðtalin hér) í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.15093187448541062
B1 0.19113290331975996
B2 0.20405924639003653
C1 0.19325039036229236
C2 0.19513977402276136



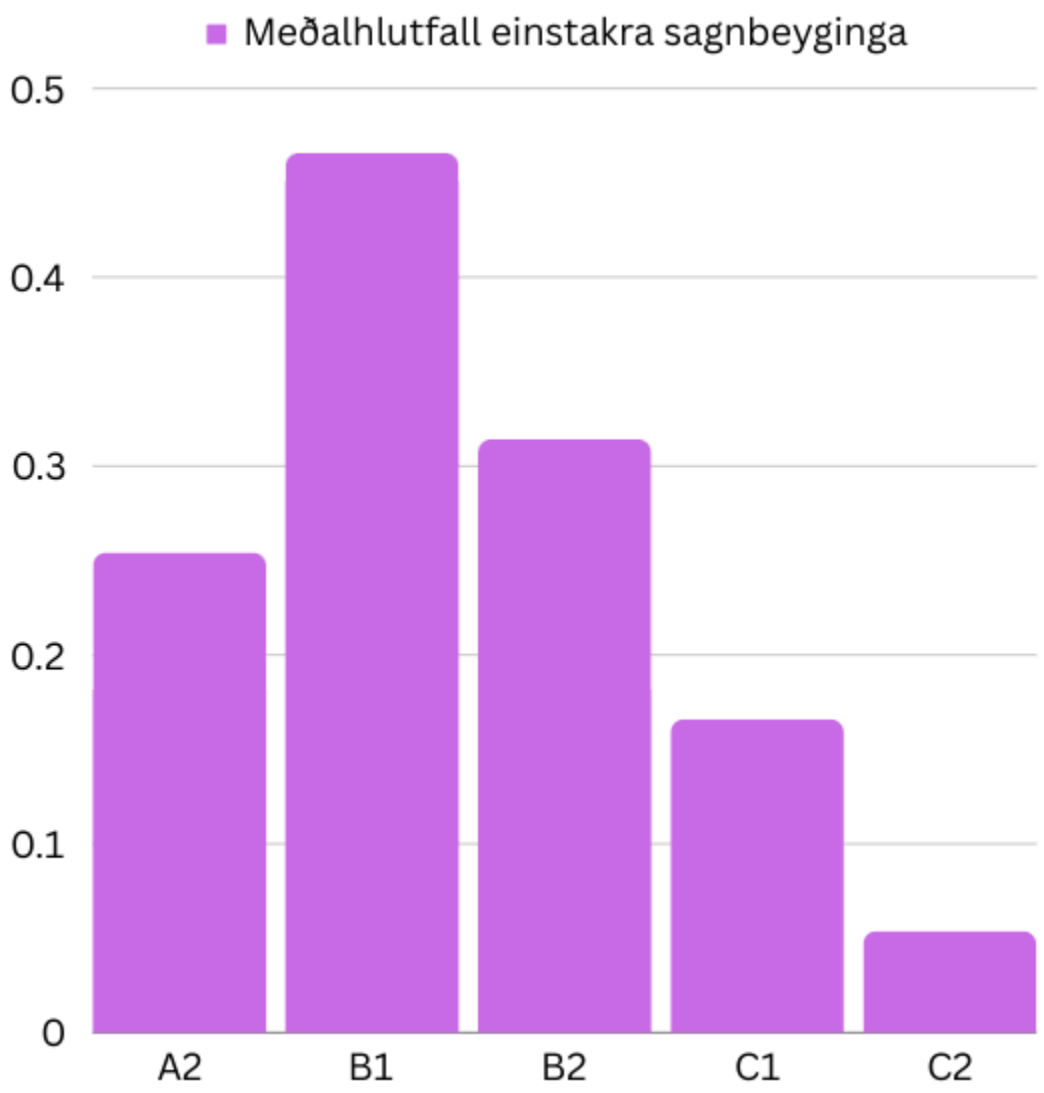
Meðalhlutfall einstakra orðmynda miðað við heildarfjölda orðmynda í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.5040877183654199
B1 0.656497216084069
B2 0.5963565878149738
C1 0.46701532603169305
C2 0.26938045865583415



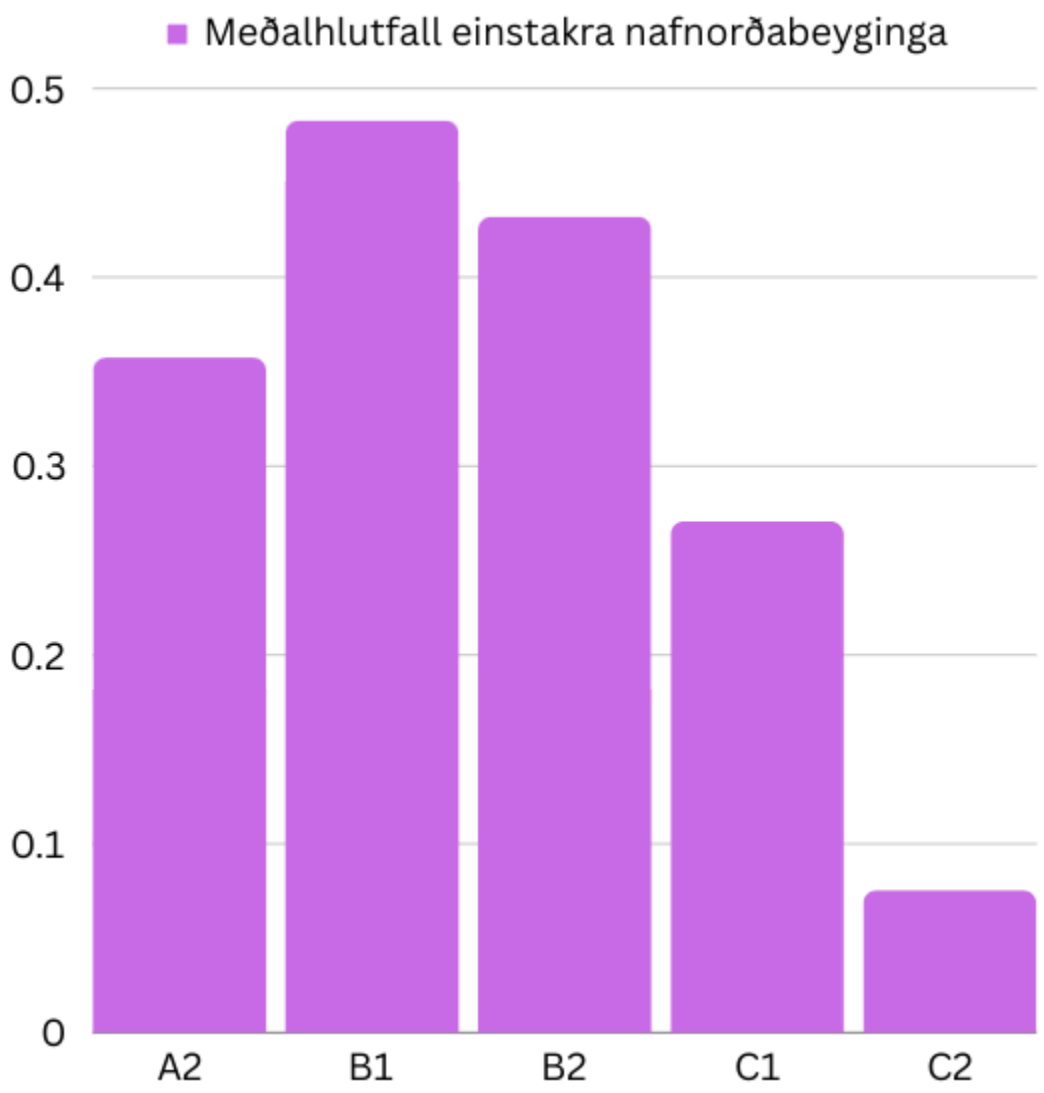
Meðalhlutfall einstakra sagnbeyginga (þ.e. einstakra marka sagna) miðað við heildarfjölda sagna í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.25426683089618624
B1 0.4657976781046413
B2 0.3142562659210544
C1 0.1659690568632481
C2 0.05381113407549933



Meðalhlutfall einstakra nafnorðabeyginga (þ.e. einstakra marka nafnorða) miðað við heildarfjölda nafnorða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.3574979713395616
B1 0.48301359236952324
B2 0.4320967621142061
C1 0.2707491368974242
C2 0.07543369291888684



Meðalhlutfall einstakra lýsingarorðabeyginga (þ.e. einstakra marka lýsingarorða) miðað við heildarfjölda lýsingarorða í hverjum texta á hverju hæfnistigi

A1 n/a
A2 0.6369173243322518
B1 0.6975898430318357
B2 0.6601523618368946
C1 0.4936388072147415
C2 0.16502581003468297