Hugbúnaður til úrvinnslu á málsýnum íslenskra barna

Jón Benediktsson

Iðnaðarverkfræði-, vélaverkfræði- og tölvunarfræðideild
Háskóli Íslands
2014

Hugbúnaður til úrvinnslu á málsýnum íslenskra barna

Jón Benediktsson

10 eininga ritgerð sem er hluti af
Baccalaureus Scientiarum gráðu í hugbúnaðaverkfræði

Leiðbeinandi/endur
Ebba Þóra Hvannberg
Jóhanna Einarsdóttir

Iðnaðarverkfræði-, vélaverkfræði- og tölvunarfræðideild
Verkfræði- og náttúruvísindasvið
Háskóli Íslands
Reykjavík, mars 2014

Hugbúnaður til úrvinnslu á málsýnum íslenskra barna
Hugbúnaður til úrvinnslu á málsýnum barna
10 eininga ritgerð sem er hluti af *Baccalaureus Scientiarum* gráðu í Hugbúnaðaverkfræði

Iðnaðarverkfræði-, vélaverkfræði- og tölvunarfræðideild
Verkfræði- og náttúruvísindasvið
Háskóli Íslands
Hjarðarhaga 2-6
107 Reykjavík

Sími: 525 4000

Skráningarupplýsingar:
Jón Benediktsson, 2014, *Hugbúnaður til úrvinnslu á málsýnum íslenskra barna*, BS ritgerð, Iðnaðarverkfræði-, vélaverkfræði- og tölvunarfræðideild, Háskóli Íslands, 13 bls.

Prentun: 01
Reykjavík, mars 2014

Útdráttur

Skýrsla þessi lýsir hugbúnaðarlausn sem Jón Benediktsson vann að sem lokaverkefni fyrir Jóhönnu Einarsdóttur undir handleiðslu Ebbu Þóru Hvannberg og Jóhönnu Einarsdóttur á haustmisseri 2013.

Markmið verkefnisins gekk út á þróa hugbúnað við úrvinnslu á málsýnum íslenskra barna. Unnið var að því skapa hugbúnaðarlausn sem auðveldar úrvinnslu á málfræðiatriðum í máltjáningu barna fengna með málsýnum.

Abstract

This essay describes a software solution that Jón Benediktsson did as a final project for Jóhanna Einarsdóttir under guidance from Ebba Þóra Hvannberg and Jóhanna Einarsdóttir in the fall of 2013.

The goal of the project was to develop software to process language samples from Icelandic children. This software solution facilitates the processing of grammar items in the language expression of children from language samples.

Efnisyfirlit

Þakkir

Ég vil þakka leiðbeinendum mínum, þeim Ebbu Þóru Hvannberg og Jóhönnu Einarsdóttur.

Inngangur

Skýrsla þessi lýsir hugbúnaðarlausn sem Jón Benediktsson vann að sem lokaverkefni fyrir Jóhönnu Einarsdóttur undir handleiðslu Ebbu Þóru Hvannberg og Jóhönnu Einarsdóttur á haustmisseri 2013.

Markmið verkefnisins gekk út á þróa hugbúnað við úrvinnslu á málsýnum íslenskra barna. Unnið var að því skapa hugbúnaðarlausn sem auðveldar úrvinnslu á málfræðiatriðum í máltjáningu barna fengna með málsýnum.

Hugmynd verkefnisins og framtíðarsýn

Verkefnið byggði á 170 málsýnum sem Jóhanna Einarsdóttir hafði látið taka. Málsýnin eru af sjálfsprottnu tali barna á aldrinum 2 til 6 ára og sýna börnin við leik og að spjalla við fullorðin einstakling. Málsýnin voru afrituð á samræmdan hátt samkvæmt handbók og áreiðanleiki afritunar var tryggður.

Málsýnin gefa lýsandi mynd af málþekkingu barnsins bæði máltjáningu þess og málnotkun. Með þessu verkefni var unnið að því að hanna búnað sem auðveldar úrvinnslu á málfræðiþekkingu barna tekna með málsýnum.

Skýrsluhöfundur sér fyrir sér að með þessu verkefni hafi verið lagður grunnur að alhliða hugbúnaðarlausn þegar kemur að vinnu talmeinafræðinga með börn með frávik í málþroska. Í núverandi mynd mun hugbúnaðurinn nýtast talmeinafræðingum sem vinna með börn með málþroskaraskanir. Hægt er að kanna á nákvæman hátt hvaða málfræðiþekkingu barnið notar í sjálfsprottnu tali miðað við meðalgetu jafnaldra. Slíkar lausnir hafa verið til fyrir önnur tungumál en vegna einstakra eiginleika íslenskunnar og skorts á gögnum var ekki hægt að nýta þær hér á landi.

Stories, conceptual scenarios

Talmeinafræðingur fær til sín barn sem hann grunar að gæti verið með málþroskaröskun. Hann tekur málsýni af barninu, skrifar það upp og hleður inn á heimasíðu. Þá birtist honum greining á sýninu miðað við jafnaldra barnsins og hann getur þá metið hvort barnið sé með frávik í málþroska.

Rannsakandi vill skoða hvort að hann hafi rétt fyrir sér varðandi breytingar á tilteknum orðflokk með aldri. Hann skráir sig inn á heimasíðuna með rannsakandaleyfi og getur flett upp í gagnagrunninum til að sjá hvort að hann hefur rétt fyrir sér eða ekki.

Notendasögur

Ekki náðist að klára allar notendasögur verkefnisins. Ef einhver nemandi ætlaði sér að halda áfram með verkefnið síðar, leggur skýrsluhöfundur til að hann skoði vel óloknu notendasögurnar.

Heildarvörulisti

Í eftirfarandi töflu má sjá allar notendasögum vörunnar eins og þær eru í lok sprettsins.

Titill sögu eða stutt lýsing	Lokið?
Málsýni greint	✓
Greining borin saman við gagnagrunn	✓
Notendakerfi	✓
Uppfletting í gagnagrunn	✓
Leiðrétting ágiskanna	✓
Myndræn framsetning	✓
Yfirlit málsýna undirnotanda	ólokið
Word skjölum hlaðið inn	ólokið
Sjálfvirk afskráning notenda	ólokið

Nánari útfærsla notendasaga

Málsýni greint
Sem talmeinafræðingur get ég fengið málsýni greint með því að hlaða því inn á síðuna til þess að ég þurfi ekki að greina það sjálfur.

Verkefni 1
Tenging við IceNLP. Málsýnið er hlaðið inn í IceNLP til greiningar. Sjá seinni kafla fyrir umfjöllun um IceNLP.

Verkefni 2
Tölfræðigreining á sýni. Ýmis tölfræði tengd sýninu greind.

Greining borin saman við gagnagrunn
Sem talmeinafræðingur get ég séð hvernig að málsýnið kemur út miðað við meðaltal til þess að ég geti metið hvort barn sé með málþroskaröskvun eða ekki.

Verkefni 1
Gagnagrunnur smíðaður. Öll málsýni greind og tölfræði þeirra hlaðið inn í gagnagrunn.

Notendakerfi
Sem umsjónarmaður heimasíðunnar get ég stjórnað því hverjir hafa aðgang að henni til þess að ég eigi auðveldara með að rukka fólk fyrir aðgang.

Uppfletting í gagnagrunn
Sem rannsakandi get ég flett upp í gagnagrunninum til þess að rannsaka eiginleika málsýna.

Leiðrétting ágiskanna
Sem talmeinafræðingur get ég séð hvaða orð IceNLP giskaði á og dregið orð í rétta orðflokka til þess að ég geti verið viss um að greining málsýnisins sé sem réttust.

Myndræn framsetning
Sem talmeinafræðingur get ég séð niðurstöðurgreiningar myndrænt svo að ég sé fljótari að greina málsýni.

Yfirlit málsýna undirnotenda
Sem talmeinafræðingur get ég búið til svæði á síðunni fyrir hvern að skjólstæðingum mínum til þess að ég geti fylgst með þróun hans á heimasíðunni.

Word skjölum hlaðið inn
Sem talmeinafræðingur get ég hlaðið málsýni inn í Word skjali til þess að spara mér tíma.

Sjálfvirk afskráning notenda
Sem umsjónarmaður heimasíðunnar get ég stillt hvenær réttindi notanda renna út til þess að

auðvelda mér áskriftir.

IceNLP

IceNLP er opinn og frjáls hugbúnaður til að greina íslenskan texta. Hugbúnaðurinn samanstendur af eftirfarandi einingum: tilreiðara (e. *tokeniser*), giskara fyrir óþekkt orð (e. *unknown word guesser*), markara (e. *part-of-speech tagger*), lemmaldi (e. *lemmatiser*), þáttara (e. parser) og nafnaþekkjara (e. *named-entity recogniser*).

Hugbúnaðurinn var upphaflega þróaður í doktorsverkefni Hrafns Loftssonar á árunum 2004-2007 en síðan þá hafa m.a. nemendur í HR og HÍ komið að þróun einstakra eininga.

Arkitektúr og tæknilegt platform

Í upphafi verkefnisins var ákveðið að láta það vera hugbúnað sem þjónustu (e. SaaS) í stað þess að láta gera forrit sem stæði eitt. Nokkrar ástæður lágu að baki þessarar ákvörðunar:

- Notendur nota forritið á netinu og þurfa ekki að setja neitt upp á tölvunum sínum. Tillit þurfti að taka til þess að tæknigeta notenda er mismikil og með þessum hætti er forritið *ósýnilegt* notendanum - hann heldur í raun að hann sé bara að nota venjulega heimasíðu.
- Auðvelt er að uppfæra forritið fyrir alla notendur á sama tíma.
- Umsjónarmaður heimasíðunnar getur stýrt aðgangi allra notanda. Þannig er hægt að láta notendur gerast áskrifendur að þjónustunni, því hægt er að loka aðgangi þeirra kjósi þeir að segja upp áskriftinni.

Nokkrir valkostir voru skoðaðir þegar kom að vali á forritunarmáli hugbúnaðarins. Þau mál sem skýrsluhöfundur skoðaði voru Python, Ruby og PHP. Öll hafa þessi mál sína kosti og galla að mati skýrsluhöfundar, PHP er rótgróið og vinsælt mál til að skrifa vefsíður í en ekki sérstaklega kennt í HÍ, Ruby með Rails er hins vegar nokkurs konar *tískumál* og mikið notað í SaaS geiranum en ekki heldur kennt sérstaklega í HÍ. Python fellur þarna mitt á milli og hefur þann kost að vera notað í kúrsum í HÍ svo það varð fyrir valinu. Hugmyndin var þá að ef einhver nemandi vildi bæta við verkefnið, væri auðveldast að hann þekkti forritunarmálið.

Þegar Python var valið kom næst að því að velja vefgrind fyrir forritið. Í upphafi var valinn vefgrindin web.py þar sem skýrsluhöfundur hafði áður notast við hana í kúrsum HÍ. Þegar kom að því að skrifa notendalíkan forritsins kom í ljós að það var of flókið að gera á öruggan hátt fyrir web.py og því var skipt yfir í vefgrindina Django með sqlite3 gagnagrunn.

Þegar kom að því að finna hýsingu fyrir forritið var fyrsti valkostur að hýsa það á neti háskólans. Ekki gafst færi á að setja upp Apache vefþjón fyrir heimasíðuna í tæka tíð, svo skoða þurfti erlenda hýsingaraðila. Þeir sem skoðaðir voru voru Google App Engine og Heruko. Google App

Engine kallaði á meiri breytingar á forritinu en Heroku svo á endanum varð Heroku fyrir valinu. Þó þýddi það að breyta þurfti um gagnagrunnskerfi og notast forritið núna við postgresql.

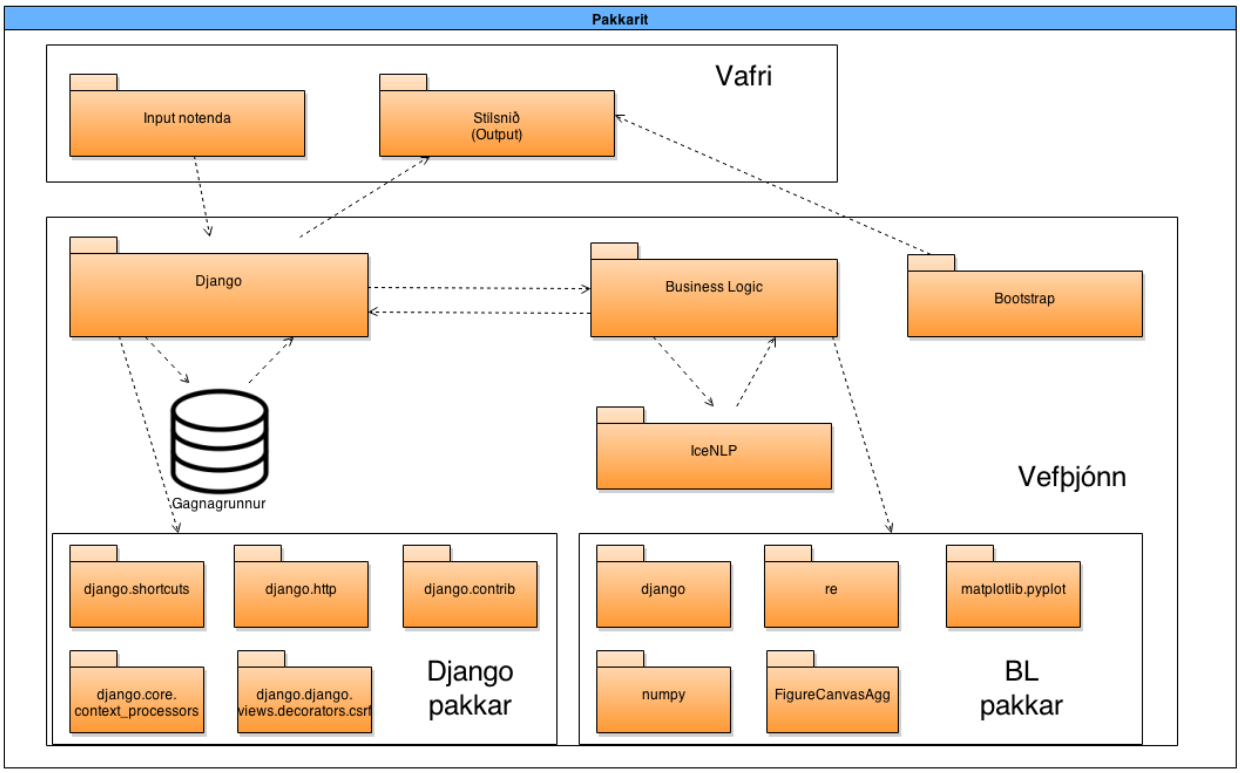
Fyrir framendagrind varð Bootstrap fyrir valinu. Ekki mikill tími fór í að rannsaka framendagrindur, en skýrsluhöfundur mat það svo að útlitið kæmi betur út með Bootstrap en án þess og því varð það fyrir valinu.

Töflulýsing gagnagrunns

Eftirfarandi er töflulýsingin úr postgresql:

```
Attribute |  Type   |      Modifier
-----------+---------+--------------------
 malsyninr    | integer |
 aldur    | real | not null default 0
 heildsegdir    | integer | not null default 0
 heildmls    | real | not null default 0
 heildordafj    | integer | not null default 0
 barnmls    | real | not null default 0
 barnordafj    | integer | not null default 0
 barnmisordfj    | integer | not null default 0
 fjoskiljanlegra    | integer | not null default 0
 fjjanei    | integer | not null default 0
 malfraedivillur    | integer | not null default 0
 so    | integer | not null default 0
 sonh    | integer | not null default 0
 sopb    | integer | not null default 0
 sotht    | integer | not null default 0
 fn    | integer | not null default 0
 fnpfn    | integer | not null default 0
 fnabfn    | integer | not null default 0
 no    | integer | not null default 0
 nonf    | integer | not null default 0
 nothf    | integer | not null default 0
 nothgf    | integer | not null default 0
 noef    | integer | not null default 0
 nokk    | integer | not null default 0
 nokvk    | integer | not null default 0
 nohk    | integer | not null default 0
 noet    | integer | not null default 0
 noft    | integer | not null default 0
 lo    | integer | not null default 0
 lokk    | integer | not null default 0
 lokvk    | integer | not null default 0
 lohk    | integer | not null default 0
 lovb    | integer | not null default 0
 losb    | integer | not null default 0
 fs    | integer | not null default 0
 st    | integer | not null default 0
```

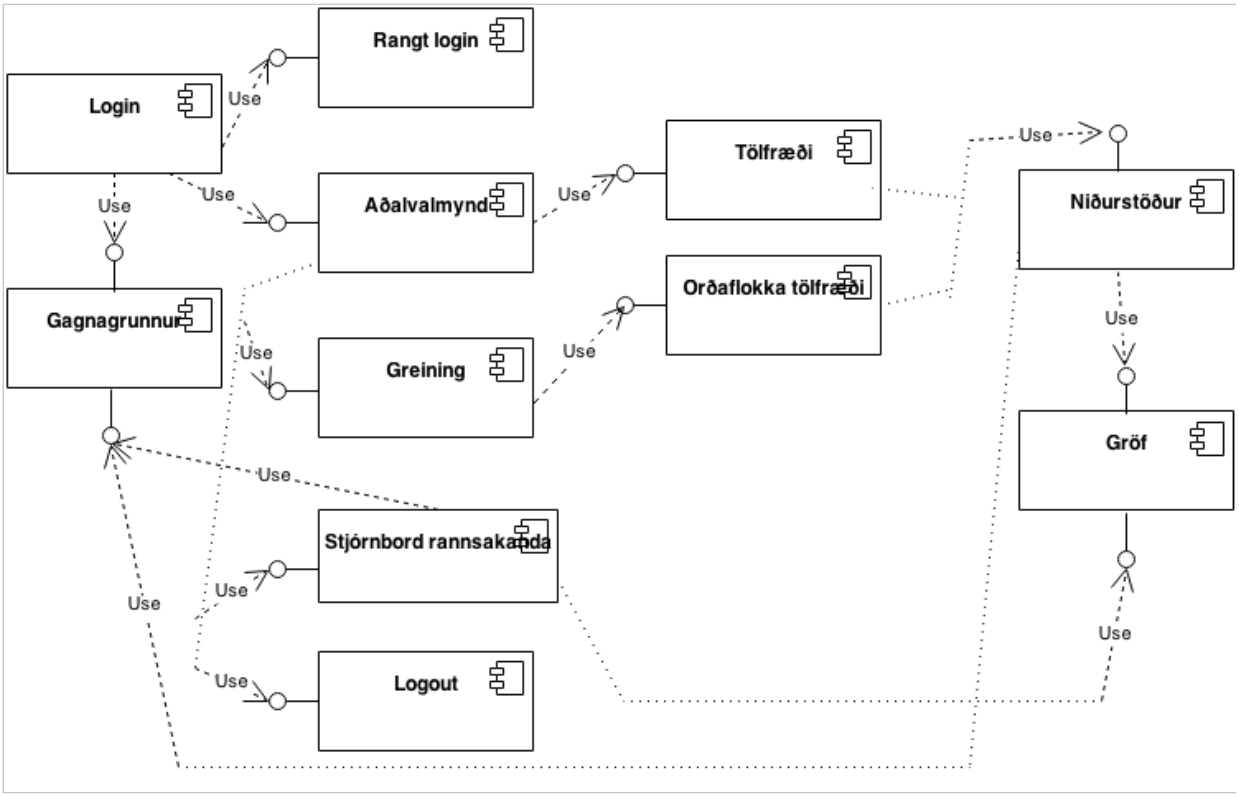
Pakkarit



Fjallað hefur verið um Bootstrap og IceNLP hér að ofan. Þeir pakkar sem ekki hefur verið fjallað um eru:

django.shortcuts - notaður til að rendera python kóða yfir í HTML
django.http - notaðir fyrir http redirect
django.contrib/django.core.context_processors - notaðir í notendakerfi
django.django.view.decorators.csrf - notaður fyrir CSRF token
django/FigureCanvasAgg - notaðir til að vista dínamískar myndir á serverinn
re - notaður til að geta skrifað regular expression í python kóða
numpy/matplotlib.pyplot - notaðir til að teikna gröf

Kompónentarit



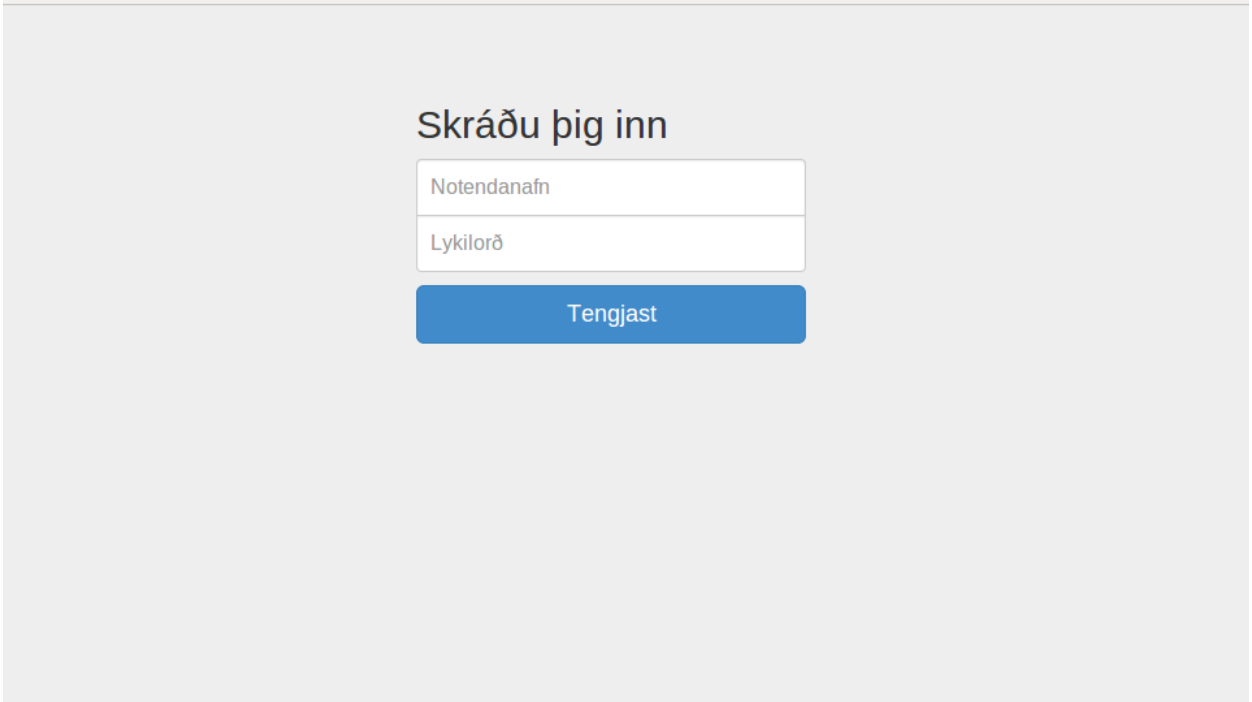
Hver kompónent er safn klasa/vef stílsniða sem hægt væri að skipta inn og út úr forritinu.

Refactoring

Klasarnir sem sáu um að teikna gröf voru allir sameinaðir í einn klasa sem tekur inn öll breytunöfnu dínamískt. Þetta kallar á langar slóðir á myndirnar en þar sem notendur verða aldrei vitni að þeim og var þessi ákvörðun tekin. Breyturnar eru margar og mismunandi fyrir hvert graf og því þurfti nýi klasinn að vera mjög sveiganlegur, en skýrsluhöfundur mat það svo að það kæmi betur út að vera með einn ofur klasa sem gerði öll gröfin í stað nokkra sem héldu utan um nokkrar mismunandi týpur af gröfum.

Viðmót

Í upphafi birtist notanda skjár þar sem hann þarf að skrá sig inn.



Þegar hann hefur skráð sig inn tekur við úrvinnslu skjárinn. Notandi þarf að afrita málsýnið í stóra gluggan að neðan.

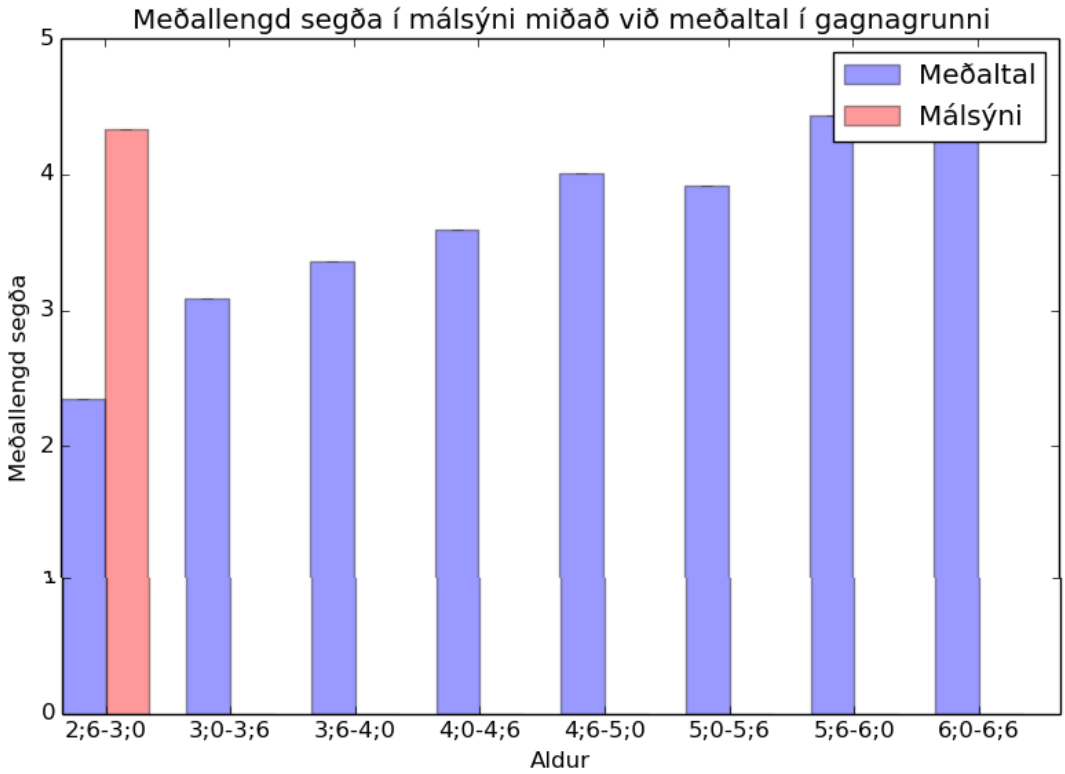
Þegar hann hefur ýtt á skoða málsýni birtast honum niðurstöðurnar. Notendasaga: Málsýni greint,Greining borin saman við gagnagrunn, Myndræn framsetning

Niðurstöður

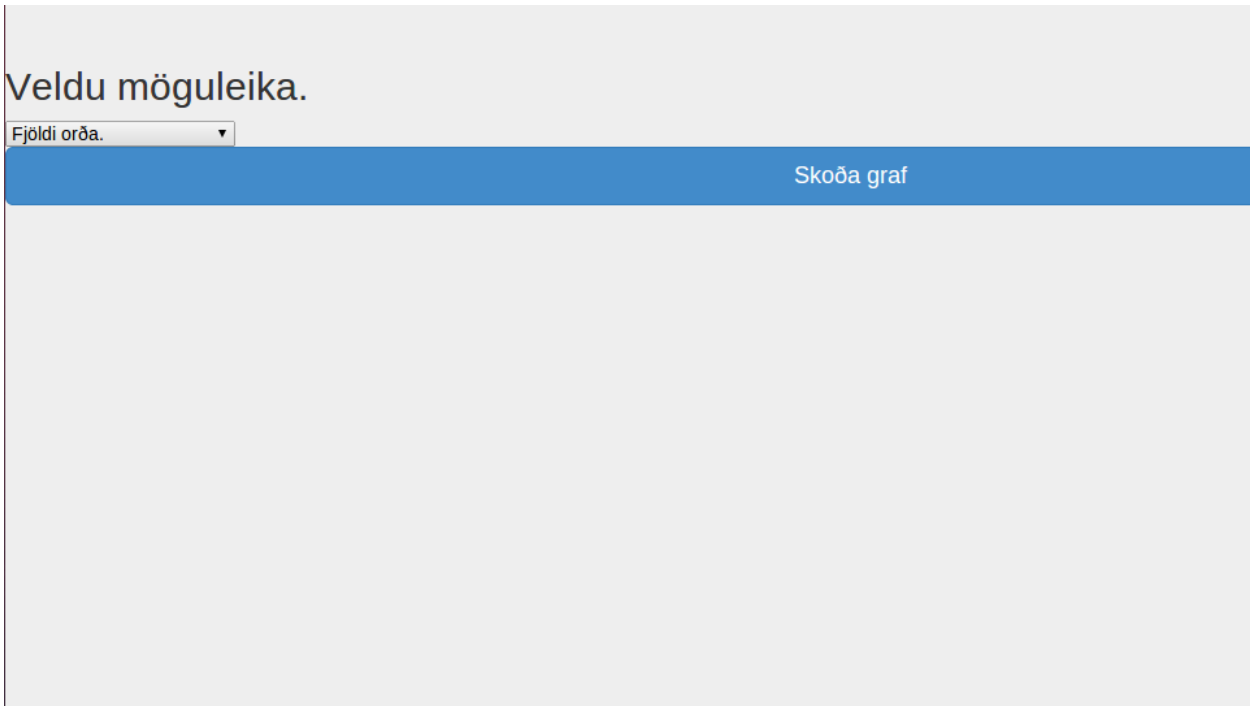
Kyn: drengur
Aldur: 2

Meðallengd segða

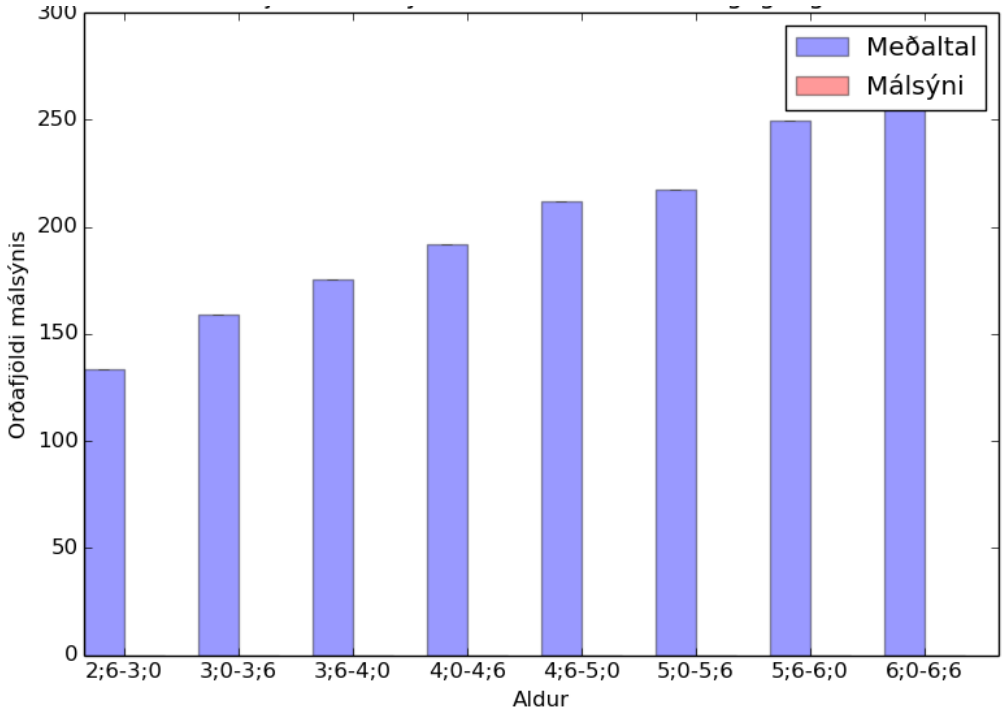
Meðallengd segða: 4.33333333333



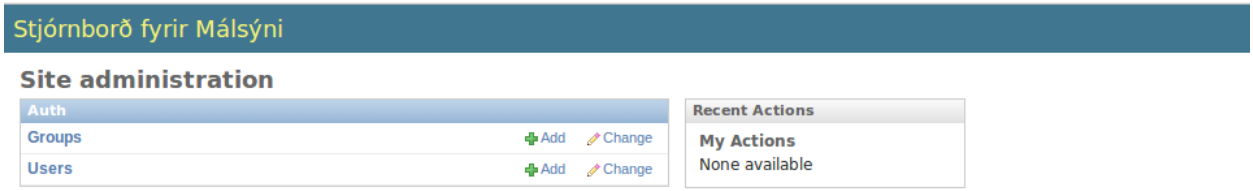
Sé notandi með aukin réttindi getur hann ýtt á stjórnborð þar sem honum birtist listi yfir hluti til að skoða úr gagnagrunni. Notendasaga: Uppfletting í gagnagrunn



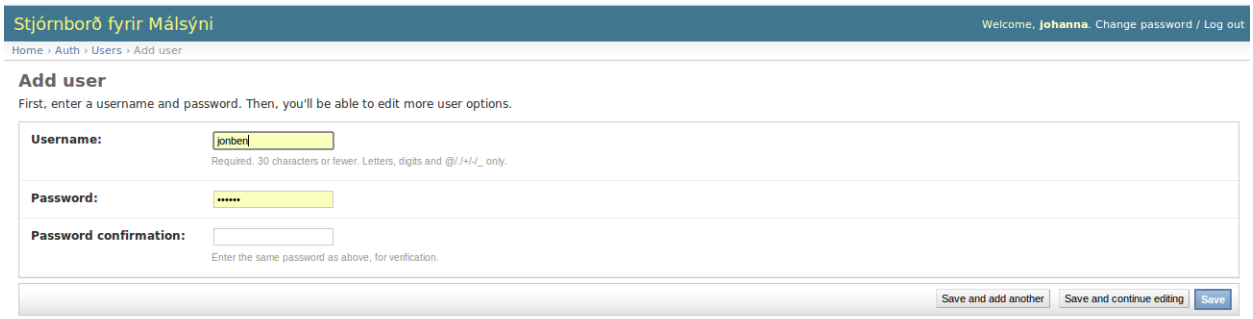
Þegar hann hefur ýtt á skoða graf, birtist honum grafið sem hann bað um. Notendasaga: Uppfletting í gagnagrunn, Myndræn framsetning



Til að stjórna notendum skráir stjórnandi síðunnar sig inn á admin síðuna og honum birtist stjórnborðið. Notendasaga: Notendakerfi



Þar getur hann bætt við notendum eða tekið út notendur. Notendasaga: Notendakerfi



Giski IceNLP á orð er notandanum boðið að fara yfir giskið og leiðrétta ef það var rangt.
Notendasaga: Leiðrétting ágiskanna

Niðurstöður

IceNLP þurfti að giska á eftirfarandi orð:

Nafnorð • Árbæjar
Lýsingarorð
Sögn
Atviksorð og forsetningar
Samtenging
Fornafn
Greinir
Tala

Vista

Viðauki 1 - Klasar

Hér eftir fylgir listi af klösum forritsins ásamt stuttri lýsingu.

Login

Sér um innskráningu notanda

Invalid

Sér um röng notendanöfn/lykilorð

Loggedin

Tekur á móti rétt innskráðum notenda

Greining

Notar IceNLP til að greina málsýnið

Rannsakandastjb

Sér um stjórnborð rannsakanda

Logout

Sér um útskráningu

Giskord

Sér um giskorð

Angiskorda

Sér um sýnið án giskorða

Ordflokkatolfraedi

Sér um tölfræði orðflokka

Villur

Sér um villur málsýnis

Tolfraedimalsynis

Sér um tölfræði sem tengist ekki orðflokkum

Giskordaflokkun

Sér um flokkun giskorða.

Nidurstodur

Sér um niðurstöður

graf

Sér um gröf.