IceAlign - Search Tool for the Icelandic Parallel Corpus

MLT201F Language resources for software development and research

University of Iceland - School of Humanitie , Sæmundargata 2, IS-101 Reykjavik, Iceland

Egill Anton Hlöðversson
egilla14@ru.is

29. March 2019

Abstract

IceAlign allows language specialists and other users to quickly look up terms within any of the subcorpora in the Icelandic Parallel Corpus. With a highly interactive layout, that displays two aligned text columns, allows the user to easily compare the translation from the source text to the target text. IceAlign is a web application developed with the use of the Django web framework, and hosted at http://icealign.herokuapp.com. The source code is available at https://github.com/egillanton/icealign

1 Introduction

This paper describes the system design, implementation, and development process of the IceAlign project. This project is the first out of two final projects for the course *Language resources for software development and research* at the University of Iceland in cooperation with Reykjavik University (RU). This course is part of the MSc. Language Technology program at RU.

1.1 About the Project

This project aims to create a search tool, thought of as a working tool for linguists, students and other language specialists, capable of displaying two parallel texts in two different languages, allowing the users to compare the translation between the two texts. This web application has the name IceAlign, with reference to the words *Icelandic* and *Alignment*.

The application will improve upon Stofnun Árna Magnússonar í Íslenskum Fræðum (SÁM) current web interface at https://málheildir.árnastofnun.is/?mode=parallel, currently used as a corpus search tool for the Icelandic Parallel Corpus.

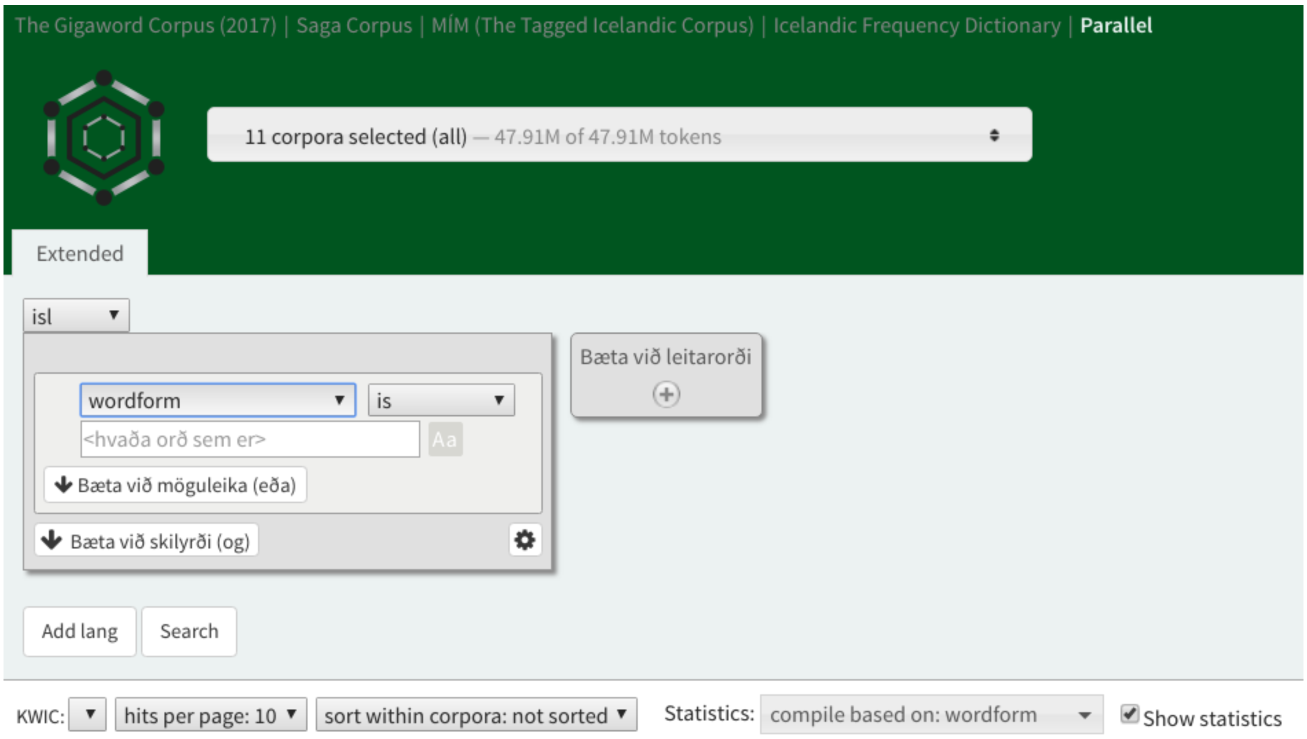
1.2 About the Icelandic Parallel Corpus

Icelandic Parallel Corpus (IPC) that was composed from 2017 to 2018 by SÁM is a collection of subcorpora. Each subcorpus consists of texts in Icelandic and English, where the lines of the paragraphs of the two languages are paired together. In total, the IPC contains 3,561,789 paired lines. The reading words of the Icelandic texts are a total of 38,927,999 but their English 42,684,702. As shown in table 1, the largest two subcorpora are the EEA (European Economic Area) documents and OpenSubtitles which is a collection of transcriptions for movies and television shows in English.

Subcorpus	**Lines**
The Bible	65.241
European Southern Hemisphere Observations (ESO) Releases	12.633
Statistic Office of Iceland - from website	2.288
Icelandic anthology	17.597
KDE 4	49.912
Classical literature	12.416
Prescriptions (European Medicines Agency)	404.333
OpenSubtitles	1.261.398
EEA documents	1.701.172
Tatoeba	8.263
Ubuntu	10.572
Total	3.561.789

Table 1: List of SÁM subcorpora's with corresponding number of lines for each subcorpus.

The current search tool for the IPC that is available is a web application based on Korp, a general corpus tool developed at Språkbanken. It is a general tool that allows for complex filtering and searches within a corpus. As seen in figure 1, the user can search by language and wordform.


The Gigaword Corpus (2017) | Saga Corpus | MÍM (The Tagged Icelandic Corpus) | Icelandic Frequency Dictionary | Parallel
Extended
Bæta við leitarorði
<hvaða orð sem er>
Aa
Bæta við möguleika (eða)
Bæta við skilyrði (og)
Add lang
Search

11 corpora selected (all) — 47.91M of 47.91M tokens

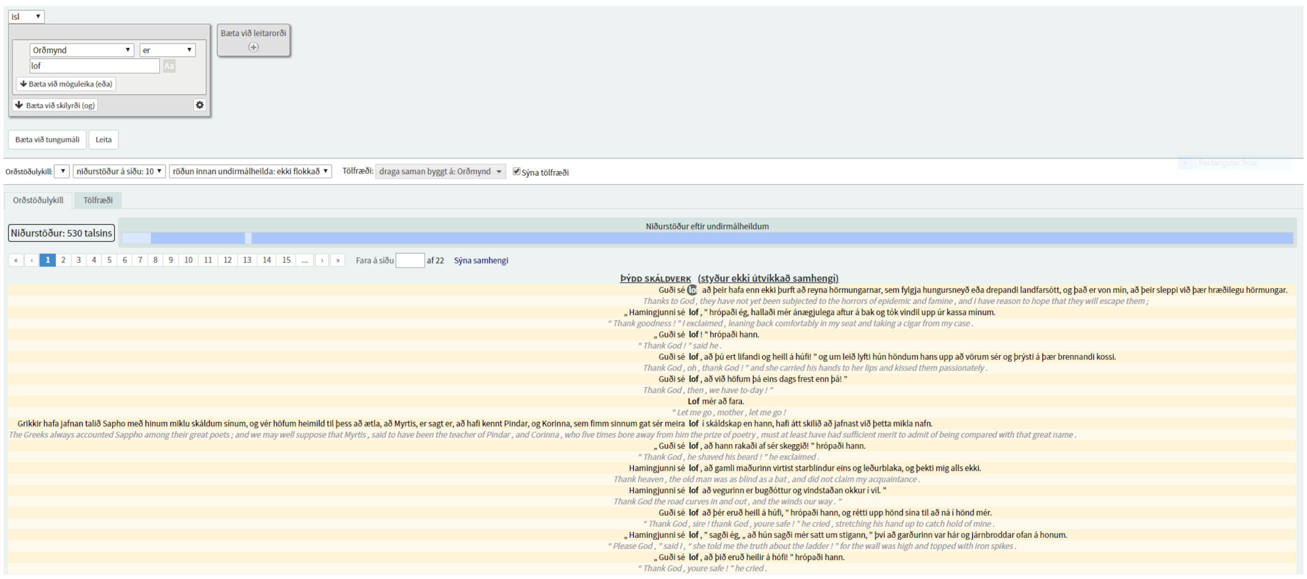
isl

wordform is

KWIC: | hits per page: 10 | sort within corpora: not sorted | Statistics: compile based on: wordform | ✓ Show statistics

Figure 1: Screen-capture from SÁM's Korp corpus tool on their website.

Figure 2 is an example of how the results are shown by searching for a wordform, the two texts are aligned horizontally, where the source text in English is in light gray color, and the target text in Icelandic is in default style except for the word that matches the word form, that word is bold.



isl

Orðmynd er lof

Orðstöðulykill: | niðurstöður á síðu: 10 | röðun innan undirmálheilda: ekki flokkað | Tölfræði: draga saman byggt á: Orðmynd | ✓ Sýna tölfræði

Niðurstöður: 530 talsins

af 22

ÞÝDD SKÁLDVERK (styður ekki útvíkkað samhengi)

Guði sé **lof** að þeir hafa enn ekki þurft að reyna hörmungarnar, sem fylgja hungursneyð eða drepandi landfarsótt, og það er von mín, að þeir sleppi við þær hræðilegu hörmungar.
Thanks to God , they have not yet been subjected to the horrors of epidemic and famine , and I have reason to hope that they will escape them ;
„ Hamingjunni sé **lof** , " hrópaði ég, hallaði mér ánægjulega aftur á bak og tók vindil upp úr kassa mínum.
" Thank goodness ! " I exclaimed , leaning back comfortably in my seat and taking a cigar from my case .
„ Guði sé **lof** ! " hrópaði hann.
" Thank God ! " said he .
Guði sé **lof** , að þú ert lifandi og heill á húfi! " og um leið lyfti hún höndum hans upp að vörum sér og þrýsti á þær brennandi kossi.
Thank God , oh , thank God ! " and she carried his hands to her lips and kissed them passionately .
Guði sé **lof** , að við höfum þá eins dags frest enn þá! "
Thank God , then , we have to-day ! "
Lof mér að fara.
" Let me go , mother , let me go !
Grikkir hafa jafnan talið Sapho með hinum miklu skáldum sínum, og vér höfum heimild til þess að ætla, að Myrtis, er sagt er, að hafi kennt Pindar, og Korinna, sem fimm sinnum gat sér meira **lof** í skáldskap en hann, hafi átt skilið að jafnast við þetta mikla nafn.
The Greeks always accounted Sappho among their great poets ; and we may well suppose that Myrtis , said to have been the teacher of Pindar , and Corinna , who five times bore away from him the prize of poetry , must at least have had sufficient merit to admit of being compared with that great name .
„ Guði sé **lof** , að hann rakaði af sér skeggið! " hrópaði hann.
" Thank God , he shaved his beard ! " he exclaimed .
Hamingjunni sé **lof** , að gamli maðurinn virtist starblindur eins og leðurblaka, og þekti mig alls ekki.
Thank heaven , the old man was as blind as a bat , and did not claim my acquaintance .
Hamingjunni sé **lof** að vegurinn er bugðóttur og vindstaðan okkur í vil. "
Thank God the road curves in and out , and the winds our way . "
Guði sé **lof** að þér eruð heill á húfi, " hrópaði hann, og rétti upp hönd sína til að ná í hönd mér.
" Thank God , sire ! thank God , youre safe ! " he cried , stretching his hand up to catch hold of mine .
„ Hamingjunni sé **lof** , " sagði ég, „ að hún sagði mér satt um stigann, " því að garðurinn var hár og járnbroddar ofan á honum.
" Please God , " said I , " she told me the truth about the ladder ! " for the wall was high and topped with iron spikes .
„ Guði sé **lof** , að þið eruð heilir á húfi! " hrópaði hann.
" Thank God , youre safe ! " he cried .

Figure 2: Screen-capture from SÁM's Korp corpus tool for the wordform *lof* in Icelandic.

With the currently available corpus tool, there are several points that we need to address the interface regarding the usability:

1. The search capability is quite complex for a user that is using the tool for the first time.
2. The font size for the source and the target text is small

3. The source text is hard to read due to contrast with the yellow background.

The goal of this project is to implement a web application that will address these points directly. In section 2 we will talk about the methodology, were we look at the capabilities of the Django web framework and how we can use it to create our web application. In section 3 we will talk about the initial design and see how they can better address the points mentioned here above. In section 4 we talk about our result and what particular challenges we had to face, and how we solved them. Finally, in section 5 we give a brief conclusion of our work and point out some future improvements.

2 Methodology

In this section, we briefly go over the software and tool used for this implementing this project, and some of the main libraries used. For the full list of libraries, take a look at the *requirements.txt* file in the root of the project. For further information about the setup procedure, refer to the *READEME.md* markdown file in the root of the project.

2.1 Django Web Framework

There are many frameworks out there; for this project, was chose the Django web framework for three main reasons:

- Uses the Python programming language server side.
- Comes with a built-in admin interface for accessing the database during runtime.
- Has an integrated database out of the box.

Additionally, to these main points, Django follows the MVT (model, view, template) architecture, and comes with a large set of libraries of pre-built functionality. That way it should take the minimum code to get a full framework capability.

2.2 Bash and Python

The first step to be able to populate our Postgres database we had to obtain our dataset from SÁM, we did that by requesting a copy at http://malfong.is/?pg=samhlida. The IPC dataset is a directory of 11 *.tmx* files, so we need to preprocess the data to be able to run our population script on the files. For this, we created a Bash script that extracted the value pair for the source and target text and outputted it into new tab separated *.txt* files, where the first column is the source text in English while the other column is in Icelandic.

Later we had to change the script, so it slit down the larger .txt files into smaller files composed of 50.000 lines each. This extra step was done to the memory restriction on the Heroku hosting service that we selected.
Haven created these .txt files, it allowed us with simple for loop in Python to walk through each line of the file and create an entry into our database.

The source file for our Bash script for preprocessing is accessible in the scripts folder of our project with the name *tmx2txt.sh.* As well, our Python code can as well be found in the scripts folder inside of our project with the name *populate_ samhlida.py.* For utility purpose, we included another Python file with the name *clean_ database.py* that deletes all entries within our database.

Django (Current Version 2.1.7) is available at bluehttps://www.djangoproject.com/

2.3 UIKit Front-End Framework

When it comes to styling our front-end HTML code, there is an option creating custom CSS style files, or import a third-party library and add specific class attributes to the HTML elements. For minimalist and free to use front-end library, we decided to go with UIKit that offers excellent range off class attributes and as a well good collection of icons that we decided to use for example in our navigation bar.

UIKit (Current Version 3.0.3) is available at bluehttps://getuikit.com/.

3 Design

In this section, we will address how we imaged at the beginning how we wanted to address the main issues of the current corpus search tool, mentioned in section 1.

3.1 Landing Page

To address a part of the first issue, about the experience for the first time users, we wanted to create a simple landing page that would direct the user to his destination, in our case, the desired corpus. As we can see in the lof-fi prototype in figure 3, we list all the available subcorpora in a menu in the navigation bar that we have placed on the right-hand side.

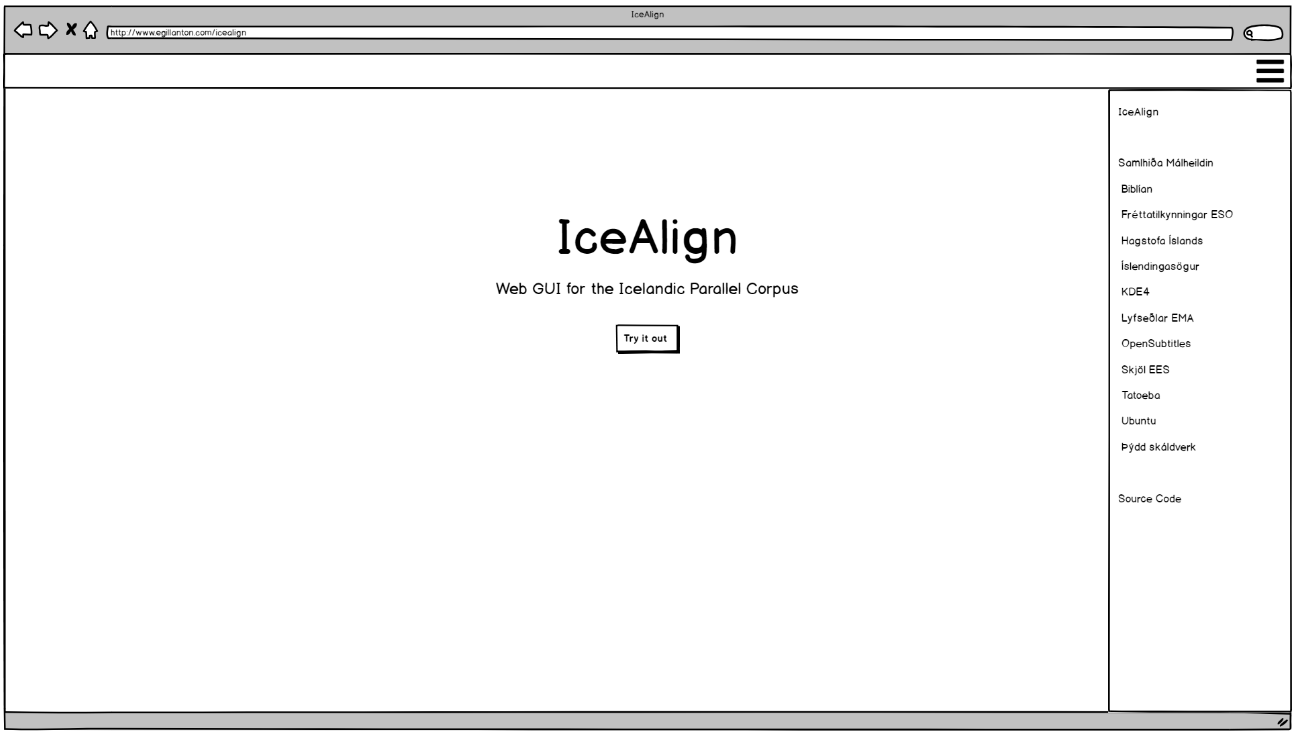
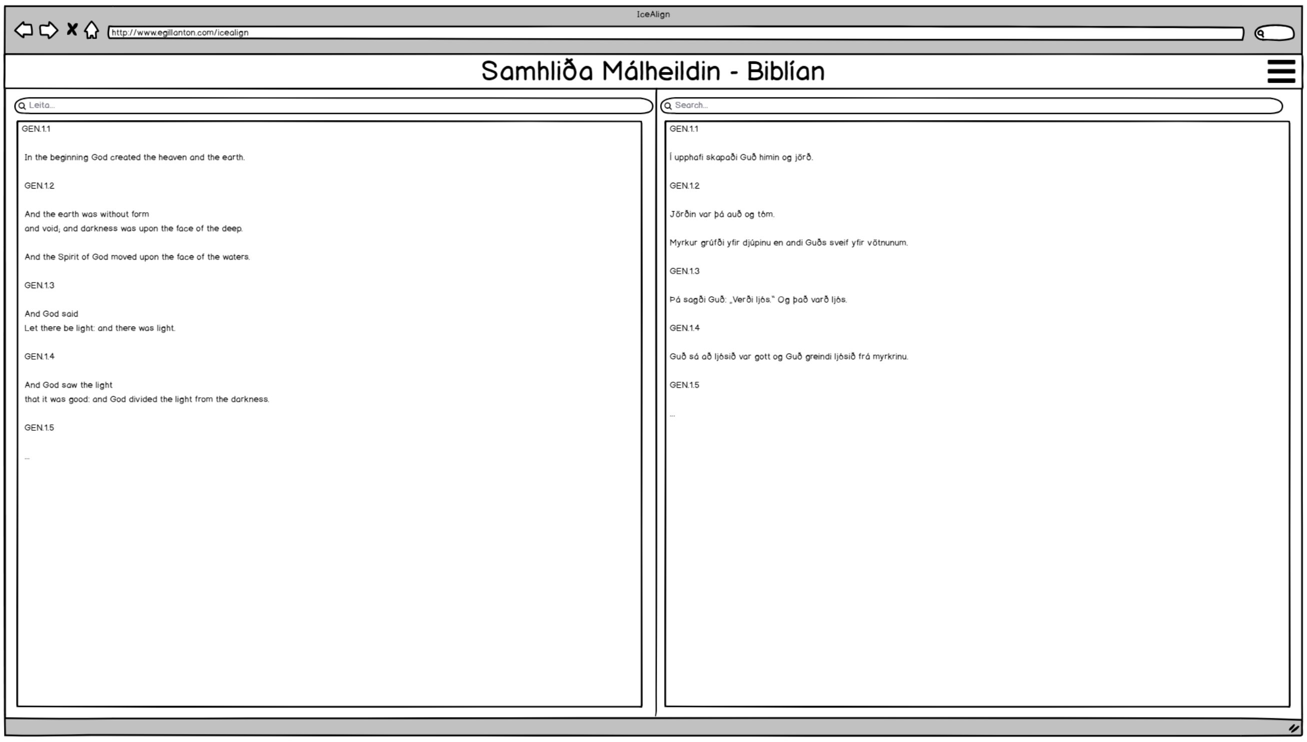


Figure 3: Lo-fi prototype design for the landing page.

3.2 Corpus Search Page

For the main search page for each corpus, we imaged to have two columns of text split down in the center of the screen, having the source text and the target text in each column. Each pair of correlating paragraphs should be placed so that they would align with one another. There should be two search bars, one for searching in the source language and the other of the target text. We imagined this functionality would be a significant main functionality for the web application. To type in any word for given language, find all the related entries with in the corresponding language. For simplicity, we want to implement a simple search that compares every entry if the term appears in any form in that entry, with no restriction on word separation or case of the text. A lo-fi prototype can been in figure 4.


IceAlign
http://www.egillanton.com/icealign
Samhliða Málheildin - Biblían
Leita...
Search...

GEN.1.1

In the beginning God created the heaven and the earth.

GEN.1.2

And the earth was without form
and void; and darkness was upon the face of the deep.

And the Spirit of God moved upon the face of the waters.

GEN.1.3

And God said
Let there be light: and there was light.

GEN.1.4

And God saw the light
that it was good: and God divided the light from the darkness.

GEN.1.5

...

GEN.1.1

Í upphafi skapaði Guð himin og jörð.

GEN.1.2

Jörðin var þá auð og tóm.

Myrkur grúfði yfir djúpinu en andi Guðs sveif yfir vötnunum.

GEN.1.3

Þá sagði Guð: „Verði ljós." Og það varð ljós.

GEN.1.4

Guð sá að ljósið var gott og Guð greindi ljósið frá myrkrinu.

GEN.1.5

...

Figure 4: Lo-fi prototype design for the corpus search tool.

4 Results

In this section we will summarize the procedure and what were the main obstacles that stood in our way and how we were then able to by pass them.

4.1 Front-end

For our front-end results, everything went as planned besides one thing. Using a front end library such as UIKit restricts our capability of overriding the design with our custom style code. As we wanted for our text to be aligned entirely, we were not able to enforce that due to the restriction that I mentioned. We can see in figure 6 how there is a small difference between the height of two texts, and as the list grew longer it was harder to see the two correlating text. So the solution was to restrict how many results appeared on each page, which in return speed up the client process due to fewer data in his browser cache. Another feature that was added to make it even more apparent for the user what two texts belong together, a hover effect was then added so that when hovering over an entry in either one of the columns, the correlating text in the other column would also get highlighted. This functionality was implemented with small Jquery Javascript functionality.

For our landing page, as seen in figure 5, nothing was out of order in comparison to our lo-fi prototype in figure 3.

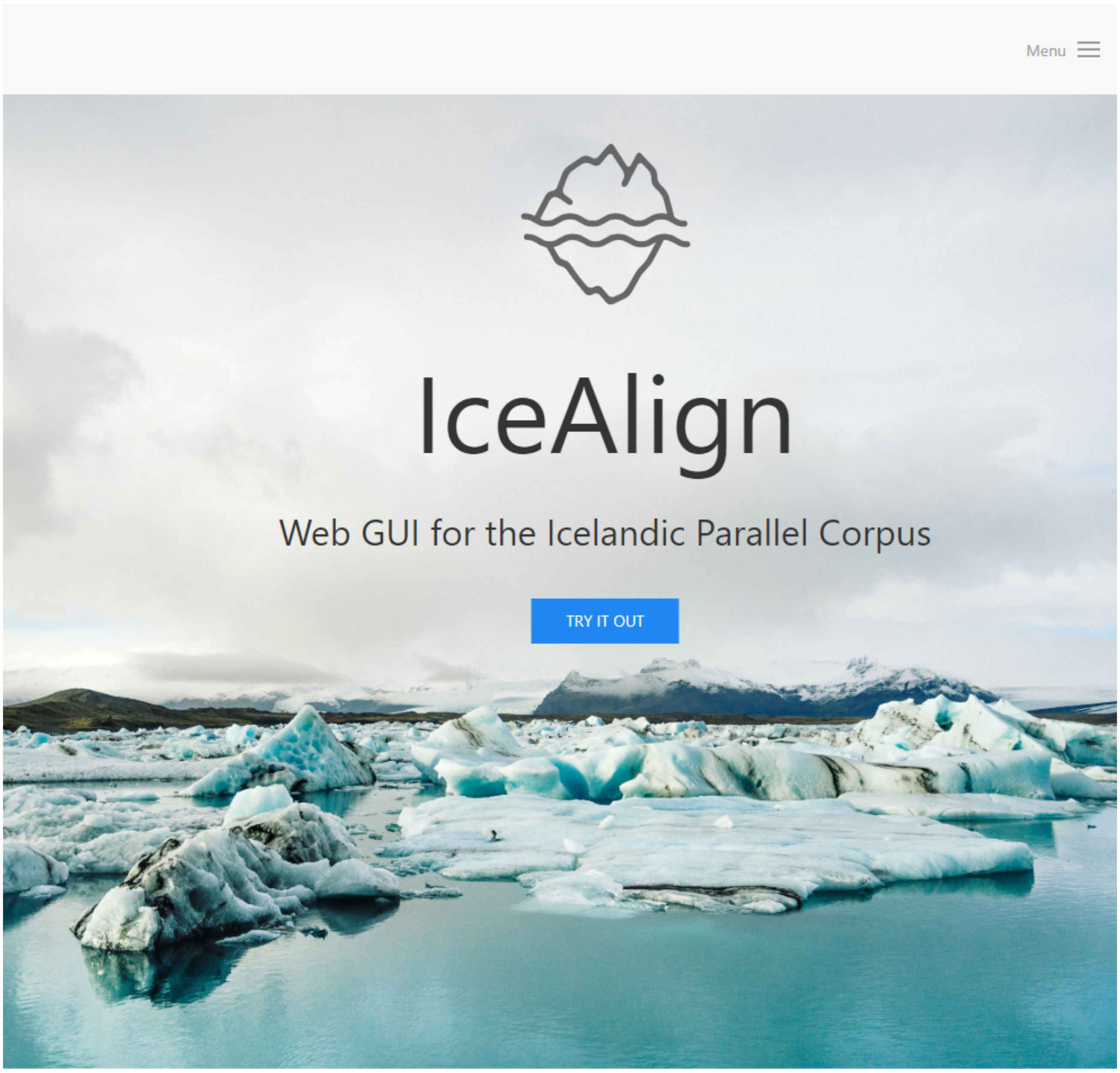
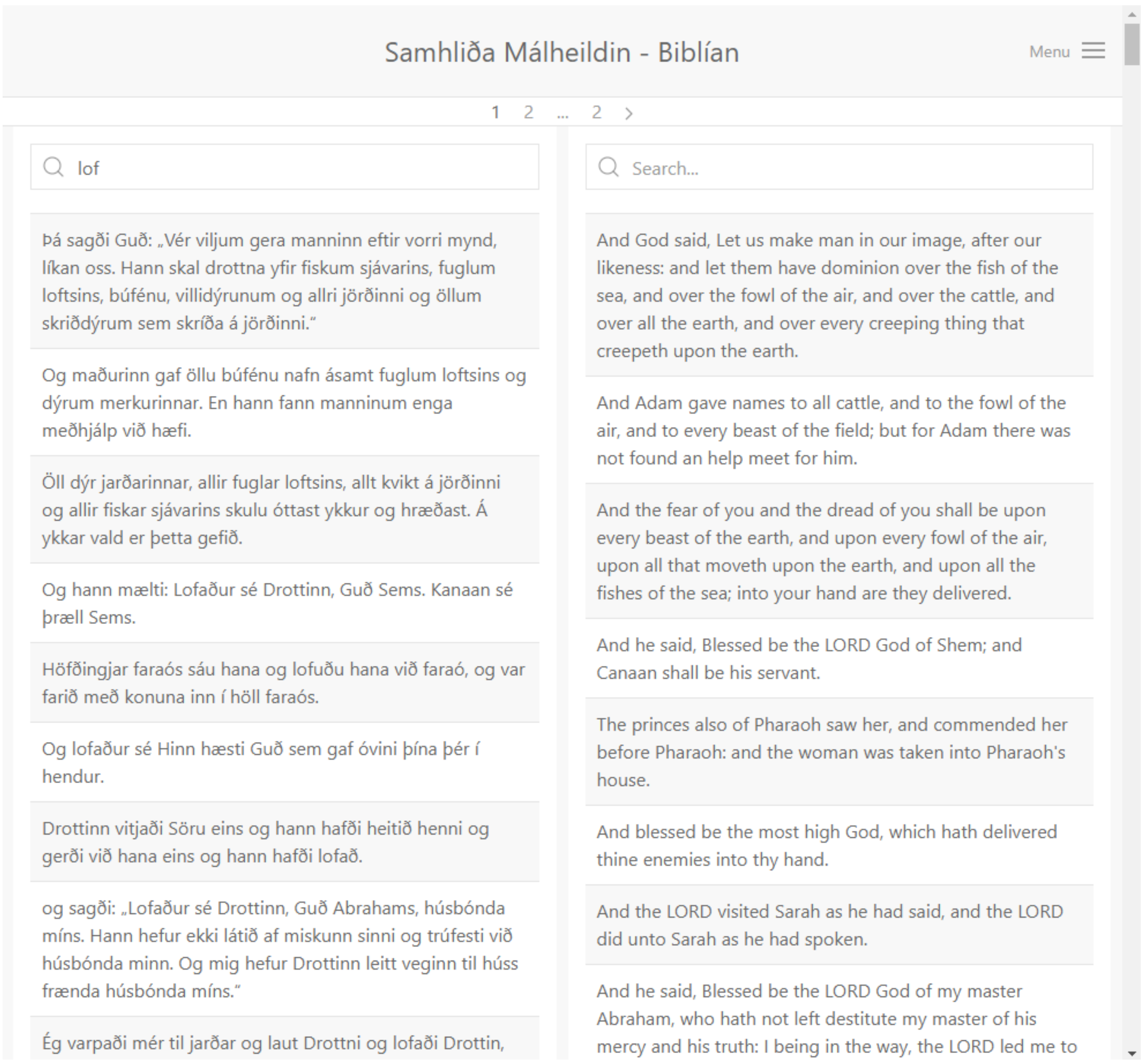


Figure 5: IceAlign landing page



Samhliða Málheilditn - Biblían

Menu

1 2 ... 2 >

lof

Search...

Þá sagði Guð: „Vér viljum gera manninn eftir vorri mynd, líkan oss. Hann skal drottna yfir fiskum sjávarins, fuglum loftsins, búfénu, villidýrunum og allri jörðinni og öllum skriðdýrum sem skríða á jörðinni.“	And God said, Let us make man in our image, after our likeness: and let them have dominion over the fish of the sea, and over the fowl of the air, and over the cattle, and over all the earth, and over every creeping thing that creepeth upon the earth.
Og maðurinn gaf öllu búfénu nafn ásamt fuglum loftsins og dýrum merkurinnar. En hann fann manninum enga meðhjálp við hæfi.	And Adam gave names to all cattle, and to the fowl of the air, and to every beast of the field; but for Adam there was not found an help meet for him.
Öll dýr jarðarinnar, allir fuglar loftsins, allt kvikt á jörðinni og allir fiskar sjávarins skulu óttast ykkur og hræðast. Á ykkar vald er þetta gefið.	And the fear of you and the dread of you shall be upon every beast of the earth, and upon every fowl of the air, upon all that moveth upon the earth, and upon all the fishes of the sea; into your hand are they delivered.
Og hann mælti: Lofaður sé Drottinn, Guð Sems. Kanaan sé þræll Sems.	And he said, Blessed be the LORD God of Shem; and Canaan shall be his servant.
Höfðingjar faraós sáu hana og lofuðu hana við faraó, og var farið með konuna inn í höll faraós.	The princes also of Pharaoh saw her, and commended her before Pharaoh: and the woman was taken into Pharaoh's house.
Og lofaður sé Hinn hæsti Guð sem gaf óvini þína þér í hendur.	And blessed be the most high God, which hath delivered thine enemies into thy hand.
Drottinn vitjaði Söru eins og hann hafði heitið henni og gerði við hana eins og hann hafði lofað.	And the LORD visited Sarah as he had said, and the LORD did unto Sarah as he had spoken.
og sagði: „Lofaður sé Drottinn, Guð Abrahams, húsbónda míns. Hann hefur ekki látið af miskunn sinni og trúfesti við húsbónda minn. Og mig hefur Drottinn leitt veginn til húss frænda húsbónda míns.“	And he said, Blessed be the LORD God of my master Abraham, who hath not left destitute my master of his mercy and his truth: I being in the way, the LORD led me to
Ég varpaði mér til jarðar og laut Drottni og lofaði Drottin,	

Figure 6: IceAlign corpus search tool

4.2 Back-end

Using Django is proper consideration when deciding a web framework where many features come out of the box. Such as built-in database, an SQLite database. However, when it was time to deploy our project to the hosting service, Heroku had a restriction that the Django projects would need to use their own Heroku Postgres database. This reconfiguration was defiantly the biggest obstacle in our way when deploying our project to their cloud service. It took an approximately one working day to tweak the settings for the project and setting up a new database that we had to populate on the server-side.

5 Conclusion

Even though this project fulfilled its purpose by improving search usability for the IPC, there are still several missing improvements. Here are some of the following features that can be done to enhance usability:

1. Display the total number of search entries found.
2. Add a tooltip functionality when hovering over an example and offer the user to copy the entry.
3. Allow regex search capability in the search bars.
4. Under the Samhlida tab, add a checkbox functionality that allows specifying which sub-corpora to search from.
5. add export capability for .json, .tvb, and .csv files.