Paulo Frazão '20
Advisor: Professor Li
Written Project Proposal (Thesis)

Project Title: Idioma – A Document-Driven Language Learning Application

Motivation and Goal:

Given both the current political climate and the air of distance and distrust that its effects have fostered among nations and cultures alike, one might argue that it has never been more important for people to try and understand one another. And there is no better tool to do so than language. Since the start of the 16th century (and perhaps even earlier, if we refrain from limiting our views to Western history), second-language learning has remained an integral part of teaching strategy throughout the world.[1]

Despite its centrality in developing cross-cultural awareness, empathy, and respect, the field of second-language acquisition has been remarkably stagnant in the past several years. Of course, it is important to recognize the utility of applications such as Duolingo, Memrise, and (for those trying to pick up a language in the early 2000s) Rosetta Stone. Ultimately, however, all such tools have failed to respond to current, cutting-edge revolutions in information processing techniques, machine learning and AI capabilities, and the petabyte-scale information store readily accessible through the internet. A truly modern second-language acquisition tool should be able to take advantage of these developments, while also providing the user with a seamless, entertaining experience.

Furthermore, many of these existing tools also fail to address a core aspect of second-language learning: reading comprehension. Applications like Duolingo simply attempt to reinforce grammar structures and vocabulary through repetition and memorization, which certainly works in a number of cases. This does leave the learner with no practical experience in reading, however, depriving them of the benefits that those exercises might provide.

The goal of my project, therefore, is to leverage the technologies mentioned above to deliver an application that targets the field of reading comprehension within the second-language acquisition process.

[1] Richards, Jack C., and Theodore S. Rodgers. *Approaches and Methods in Language Teaching*. Cambridge University Press, 2016.

Problem Background:

Unsurprisingly, there exists quite a bit of research in the field of second-language acquisition, particularly in relation to how it can be enhanced through technology and gamification. In my research process, I considered a number of popular applications, giving special attention to their strengths and to their weaknesses.

Two of the most popular such tools are Duolingo and Memrise. These applications aim to help the learner to develop a stronger vocabulary in their language of choice through gamified flashcards. As a former consumer of both, I can confirm that they indeed provide an excellent service; it is captivating to learn vocabulary in this way, and the scoring and awards can turn a five-minute review session into several hours of learning. Where they struggle, however, is in their provision of resources for other, more fundamental aspects of a language. To my knowledge, neither Memrise nor Duolingo provides much support for an individual to learn about grammar structures, for instance, which I would argue to be one of the most important aspects of any language. They also do not enable the student to view these words in context, another clear limitation.

There are other services, however, that are fueled by context. LinguaLift and HelloTalk are two good examples. The former provides more of a comprehensive curriculum for the learner, with support available from human tutors. The latter chooses instead to immerse the learner in the language through real conversations with other students, which are self-driven and can therefore accommodate any sort of pace or comfort level. These kinds of tools are very different from the standard language-learning models, and they bring a number of different benefits as a result. But they can also be very daunting, necessitate the active participation of multiple individuals, and ultimately prove to be a much larger time-sink than many students might be willing to undertake.

There also exists a fair bit of prior research in the field of ML-driven document scoring, which will be one of the key aspects of my project. This subject will be discussed in more detail below, but my approach will hinge on my ability to analyze foreign-language documents, score them based on their overall difficulty, and extract features relevant to the calculation of that difficulty. As such, I will be relying on a number of key existing methods and papers to develop my own algorithm, which I will be researching in greater depth in the coming weeks.

Approach:

My primary goal in this piece of independent work is twofold: to leverage new developments in the fields of machine learning and information accessibility to develop an application that provides users with a comprehensive introduction to their language of choice.

After various discussions with my advisor, I have decided to work towards this objective through the use of natural language processing techniques. I will be developing a web application that takes in a user's interests as well as some measure of their experience with a target language and then outputs a series of documents geared at helping them to learn that language in context. This application will keep up with a user, tracking their progression as they clear documents and preserving that progression through various gamification strategies such as levels and achievement awards. By making use of this tool, then, a student will be able to learn a second language on their own terms, practicing reading comprehension in context without sacrificing their own enjoyment in the process.

Now, there are a number of clear benefits to this approach, especially in comparison to existing solutions. Firstly, it achieves the stated goal of promoting reading comprehension in a fun, intuitive way; after all, the best kind of reading is that which allows one to learn about their interests, and Idioma will be constructed to do just that. Furthermore, it will leverage a number of fairly novel strategies to ensure that it provides as appropriate a curriculum as is possible. I will be pulling Spanish-language documents (since this will be the test language around which I build this project) from all over the internet, scoring their difficulty and extracting the features that make them difficult, and then constructing an algorithm to dynamically offer these articles to users in the way stated above. This strategy will allow me to employ cutting-edge techniques in web-scraping, difficulty classification, and related problems in natural language processing.

This approach will also have the added benefit of being adaptable to a number of different language-learning scenarios. While I would like for anyone to be able to make an account and start using Idioma with no prior knowledge of Spanish (and so will be developing it with that goal in mind), I also envision it as an excellent classroom tool. I recall being assigned a series of articles in my AP Spanish class, most of which were useful for learning purposes but painfully dull as well. I imagine that replacing these readings with documents from Idioma, carefully chosen to ensure that they match a user's skillset and interests, would be a welcome change in many classroom settings.

As such, I believe that my approach will allow me to capitalize on the lack of state-of-the-art reading comprehension tools with a system that is both flexible and enjoyable for any kind of student.

Plan:

This section will cover my initial timetable for completing this project. I have included a Gantt chart below to showcase the schedule in its entirety; I will simply discuss the important sections and tasks listed on the chart. Please keep in mind that this is very tentative and so I expect this chart to change drastically over time.

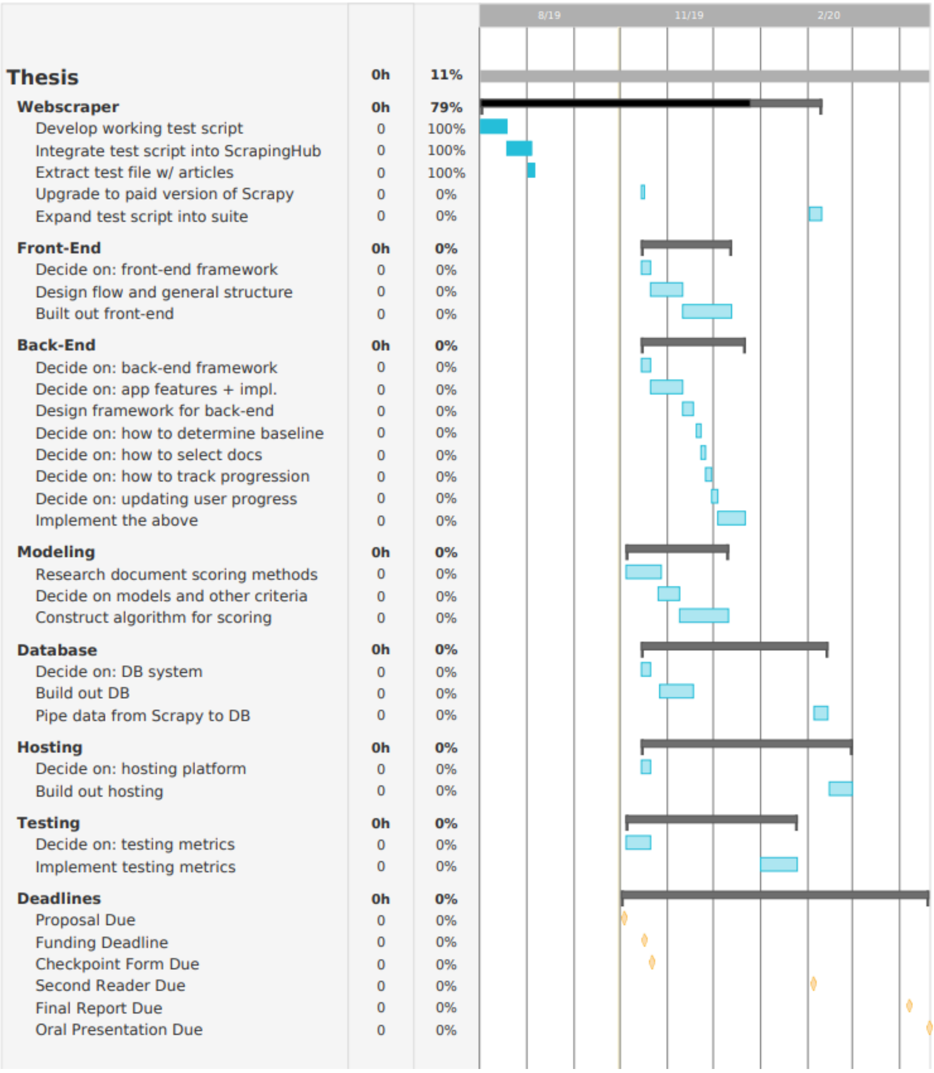
Given the full-stack nature of the project, I have broken up my thesis into seven modules (I have not included relevant deadlines for writing the paper at this point). These modules are: Web Scraper (to build my store of data for use in the models), Front-End, Back-End, Modeling, DB Design, Hosting, and Testing.

I was able to begin the first of these modules over the summer. At this point, I've created a prototype of my web-scraper that successfully pulled Spanish-language documents from one news repository on the web. I aim to build this out more fully in February after the other, more pressing parts of my thesis have been fleshed out.

I've scheduled the front-end and back-end portions of my project to take place around the same time as I would like to be making design decisions within the same time frame for both. I expect the actual implementations of both to stretch way beyond the deadlines I've currently set, but I believe for these to be reasonable in the interim period.

I believe that the modeling will be one of the more intense modules, especially given my (current) lack of knowledge around NLP. Since I am currently taking a course on the subject, I've tasked myself with researching document scoring concurrently so that I can maximize the amount of information I glean about the subject. I expect to be refining my models well into the new year, and I will adjust my schedule when that time comes.

The DB design should be a comparatively simple part of the project, and so I may push that back farther than I've listed. I also determined that the hosting is of lower priority than the other modules, hence it's comparative lateness in my schedule. Lastly, I've also budgeted some time to reflect on how I will evaluate the results of my project. My current thoughts are shared below.



Thesis	**0h**	**11%**
Webscraper	**0h**	**79%**
Develop working test script	0	100%
Integrate test script into ScrapingHub	0	100%
Extract test file w/ articles	0	100%
Upgrade to paid version of Scrapy	0	0%
Expand test script into suite	0	0%
Front-End	**0h**	**0%**
Decide on: front-end framework	0	0%
Design flow and general structure	0	0%
Built out front-end	0	0%
Back-End	**0h**	**0%**
Decide on: back-end framework	0	0%
Decide on: app features + impl.	0	0%
Design framework for back-end	0	0%
Decide on: how to determine baseline	0	0%
Decide on: how to select docs	0	0%
Decide on: how to track progression	0	0%
Decide on: updating user progress	0	0%
Implement the above	0	0%
Modeling	**0h**	**0%**
Research document scoring methods	0	0%
Decide on models and other criteria	0	0%
Construct algorithm for scoring	0	0%
Database	**0h**	**0%**
Decide on: DB system	0	0%
Build out DB	0	0%
Pipe data from Scrapy to DB	0	0%
Hosting	**0h**	**0%**
Decide on: hosting platform	0	0%
Build out hosting	0	0%
Testing	**0h**	**0%**
Decide on: testing metrics	0	0%
Implement testing metrics	0	0%
Deadlines	**0h**	**0%**
Proposal Due	0	0%
Funding Deadline	0	0%
Checkpoint Form Due	0	0%
Second Reader Due	0	0%
Final Report Due	0	0%
Oral Presentation Due	0	0%

Evaluation:

There are a number of aspects of my project that I will have to evaluate thoroughly. The first is the suite of document scoring models. While I am not presently sure how one might evaluate the results of this type of model, I believe that I will be covering them in depth in my natural language processing class and so I hope to have a better understanding of that process soon. The second component that will need to be tested is the application itself. On the one hand, I'll need to make sure it's fairly bug-free and usable, which can be done through the typical means of student testing. On the other, I will need to find some way to gauge whether the application actually helps you to learn Spanish. At this point, it might be most efficient to sample from among the student body, taking people of various levels of comfort with Spanish and having them work with Idioma for a set period of time. Afterwards, I could ask them to fill out a qualitative survey to assess its effectiveness. I hope to find other, more quantitative methods to determine the value of the system that I will create.