Bilkent University

Department of Computer Engineering

Senior Design Project

Etymøn: A Data Visualization & Deep Learning Application for Etymological Clustering of Words

Final Report

Nashiha Ahmed, Mert İnan, Cholpon Mambetova, Utku Uçkun

Supervisor: Prof. Mehmet Koyutürk
Jury Members: Prof. Uğur Doğrusöz and Prof. Çiğdem Gündüz Demir

Final Report
May 3, 2018

This report is submitted to the Department of Computer Engineering of Bilkent University in partial fulfillment of the requirements of the Senior Design Project course CS491/2.

Table of Contents

Final Design Report

Etymøn: A Data Visualization & Deep-Learning Application for Etymological Clustering of Words

1. Introduction

Etymøn is an analysis and tracing tool for word origins in all languages. It is also a data visualization application that demonstrates network graphs of links between words in different languages. Etymøn can be considered as an interactive art experiment as well, with its aesthetic visuals. It also contains a "hallucination" section, where a Long Short-Term Memory (LSTM) neural network generates new connections for given words.

2. Final Architecture and Design

As mentioned in the previous reports, Etymøn uses client-server architecture. However, instead of the object-oriented class structure, modules based on the functionalities of the html structures are created. Figure 1 shows the UML diagram of the server side and Figure 2 shows the package structure of Etymøn client side.

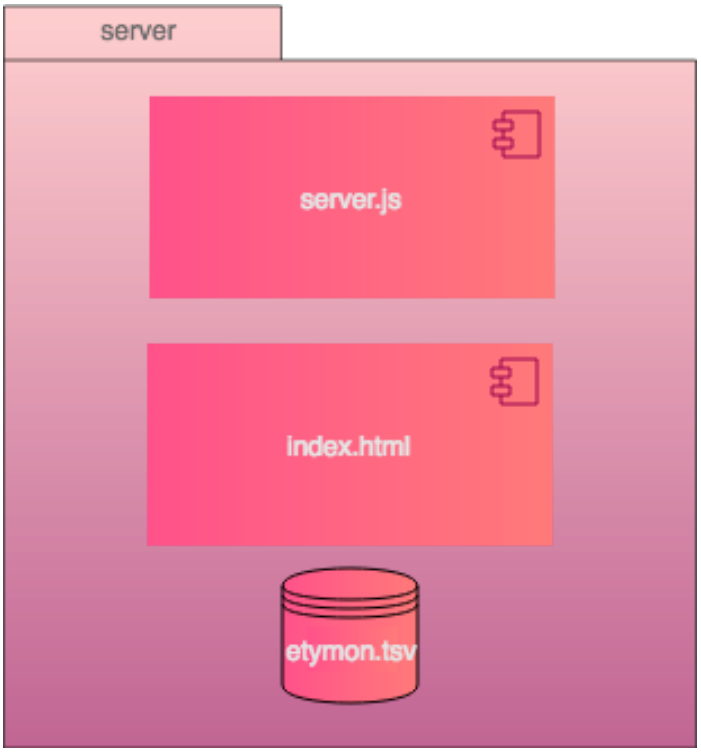

Figure 1 This figure shows the components of the server side of Etymøn.

Figure 2 This is the UML diagram for the components and packages of the client side of Etymøn.

These UML diagrams show main javascript files of the packages. Contents of these files are available in the GitHub repository of Etymøn.

3. Final Status

Etymøn functionally accomplishes all the requirements that were mentioned in the project specifications for search and hallucination parts. It does not contain the object recognition module yet, as it was deemed to be a low-priority sub-functionality of the project. Search functionality searches for an enquiry and demonstrates the word cloud (graph of etymologically-connected words) using cytoscape.js. Hallucination functionality uses LSTM to generate a theoretical word cloud for the inquiry. Screens of Etymøn can be seen in the following figures. Etymøn also strives for a good aesthetic experience for the user, hence it accomplished that in addition to the previous project specifications and it is still possible to improve the aesthetics of it.

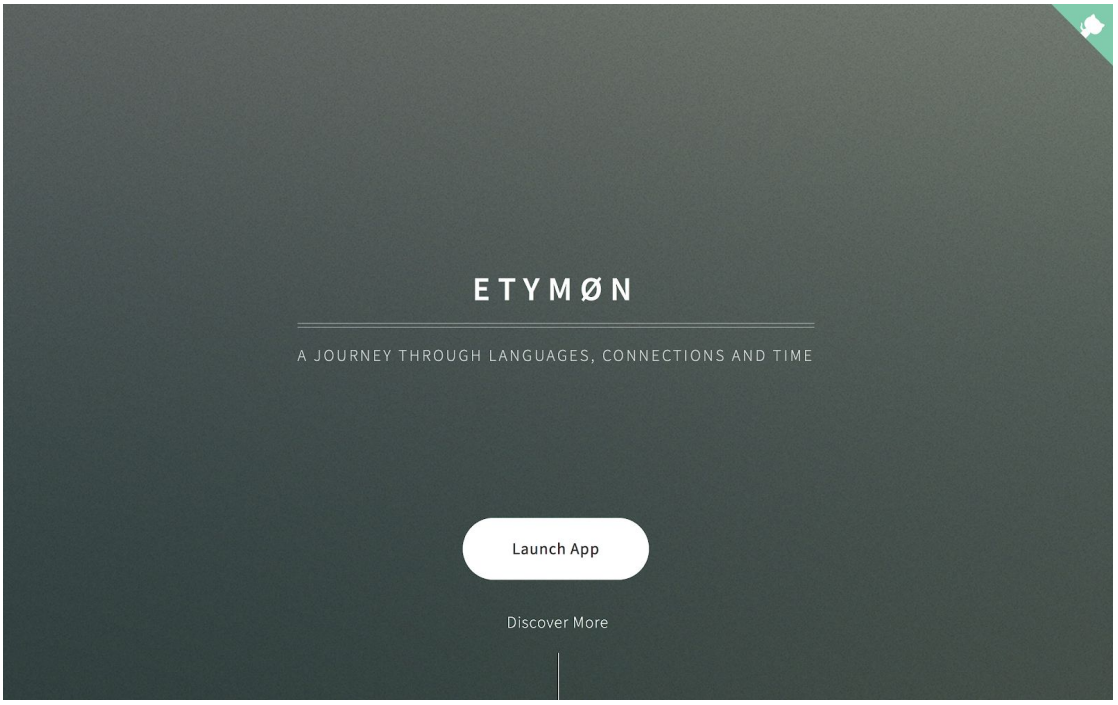

Figure 3 This figure is the landing screen of etymon.org.

Figure 4 This is the main screen of the Etymøn app, where the user is greeted by the moving language sea in the background.

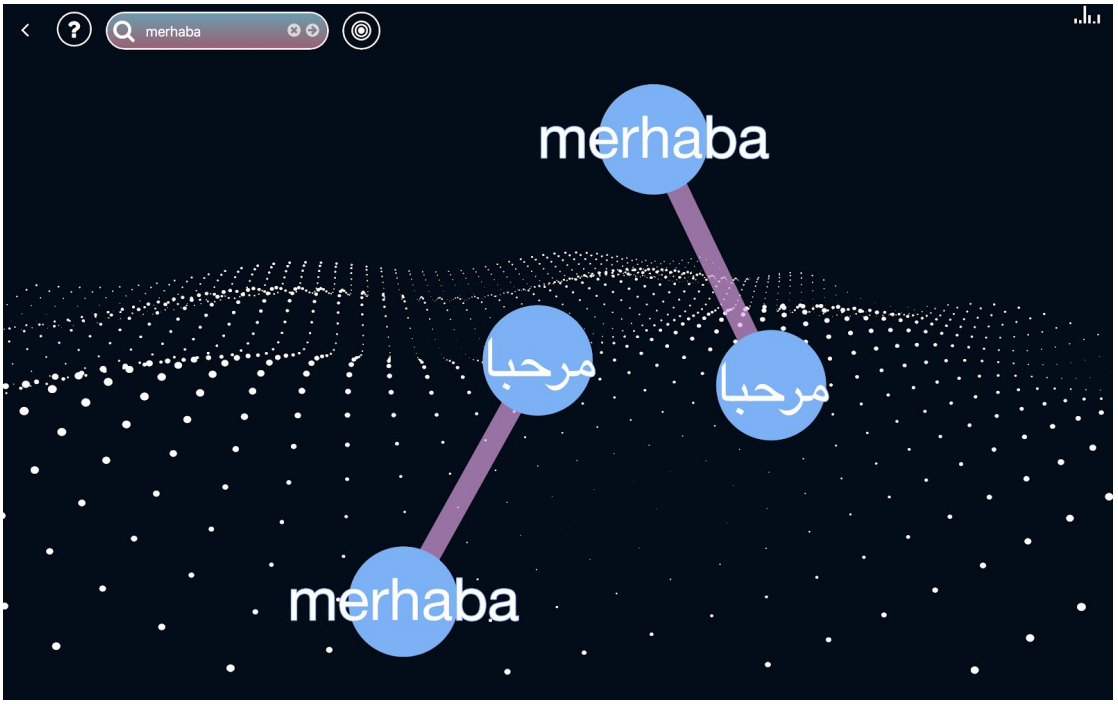

Figure 5 This figure shows a word cloud of the word "Merhaba" in Turkish when a user searches the word merhaba.

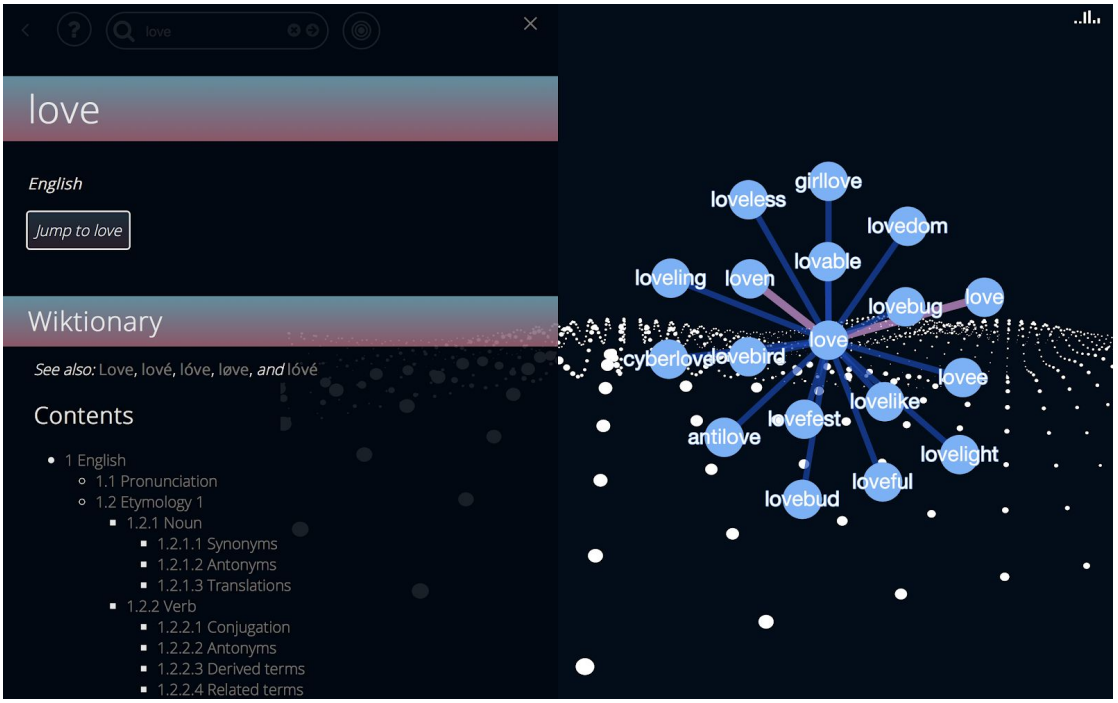

Figure 6 This figure shows the Wiktionary information of a word when it is clicked. It is located on the left side of the page in form of a panel. The user can jump to the word "love" by clicking the button on the top of the panel "Jump to love."

4. Impact of Engineering Solutions

4.1. Dataset

One of the functionalities Etymøn provides to the user is collecting and presenting already existing etymological information from web. We did not choose to do this lookup operation dynamic because we were not sure whether these websites would be available in the future. Thus, we downloaded and stored all the etymology related information from these websites. After parsing and extracting the useful informations from these files we decided to store this information in text format. We knew we had to store lots of words and relations between them and storing in text format was the most space efficient solution.

We also kept the dataset and backend logic as independent as possible. This way altering dataset become very easy. Etymology domain is very dynamic and we wanted to make Etymøn flexible.

4.2. Three.js and Cytoscape

Another functionalities Etymøn provides is allowing user to choose word relations in either cytoscape or three.js. We decided to have these options, since three.js despite looking beautiful is demanding. For those who are using Etymøn for research purposes, it may be more convenient to use Cytoscape, which is less demanding on the machine running Etymøn.

5. Contemporary Issues

5.1. Machine Learning

Currently Etymøn uses an LSTM model to create the graph itself. However, it may be useful to generate the word cloud based on specific words instead of the graphs. As, machine learning is of contemporary importance, it would be appropriate for Etymøn to use a different model, and improve its usefulness in the etymological origin-finding process. This can be accomplished by showing the confidence levels for each generated word when another model is used.

5.2. Intellectual Property and Legal Issues

Each team member and the system itself was responsible with the data or code that they used in the project and gave credit in order to protect intellectual property rights of the users or experts in linguistics. The data we retrieved from websites were attributed and used for academic purposes. We also have not monetized our software, therefore there can be no legal disputes.

5.3. Natural Language Processing

Even though Etymøn uses already present etymological information, natural language processing is relevant in several aspects. First of all, while scraping data from websites, natural language processing can be employed to better extract information from the websites. Secondly, it would prove to be useful in generation of hallucinated origin words in the LSTM, as certain key features of words can be extracted using natural language processing and used in machine learning.

5.4. Linguistics

Relations between languages and language families are changing with new archaeological discoveries. As new literature pieces, alphabets and cultures are unearthed etymological origins of languages and words become open to reinterpretations. Therefore, there can be variations between etymological origin information of the same word among different sources. We experienced same issue in our application as well. Some of the sources we used had contradicting results and we are aware that further discoveries may nullify some of our work. On the other hand, Etymøn dataset is easy to change and these new discoveries can be integrated without much effort.

6. New Tools & Technologies Used

6.1. Cytoscape.js

Main data visualization of the network graphs is undertaken by cytoscape.js. As cytoscape.js is a client-side application that renders graphs using source and target nodes and edges, it is used to demonstrate the word clouds.

Different colors of edges are used for different languages. Edge and node information is received from the server side and processed inside the client-side javascript code.

Cytoscape has the flexibility of adapting to the inputs. It also contains algorithms to perfectly organize a given graph. CoSE Layout is used to create nodes which have the searched word in the center and the relations around the word arranged

in a circle. This enables the user to clearly see the relations based on their connections to the searched word.

6.2. Three.js

As Etymøn is also an art experience, quality animations are necessary to attract users' attention. This is accomplished by the use of client-side 3D rendering javascript library called three.js.

Three.js allows smooth movement of particles. This idea and several examples on the three.js documentation website enabled Etymøn to have a language sea. This sea contains white particles that are aesthetically moving in waves when Etymøn is first opened.

Full transition from cytoscape.js to three.js is envisioned as such a transition would improve the aesthetic quality and smoothness of Etymøn. Yet, due to heavy processing load of particles in three.js, this transition may not preferred, as certain users may use Etymøn just to search for etymological purposes rather than aesthetic experiences.

6.3. Node.js

Node.js is the javascript library for server-side setup. This library is used to deploy the application on Google Cloud. It is also employed to test Etymøn locally on the computers for its projected behavior on the server.

Certain libraries of node.js were necessary during the production of the server. These include python-virtualenv, express and body-parser. Python virtual environment package was used for machine learning code. Express and body-parser were used to set up http requests.

6.4. Express.js

Express.js is a framework that handles http requests such as POST, SEND and GET. Etymøn's server uses two different POST requests: one for regular word search and another for machine learning inquiries. It is used in company with node.js.

6.5. Keras

Keras is an abstraction library of neural networks on the tensorflow library. It is used in Etymøn to create a sequential model of a neural network. The model contains an LSTM layer. Ease of use of Keras makes it suitable for small applications of neural networks, hence it was the best choice to be used in the project.

Keras is used inside a python virtual environment in combination with tensorflow. This integration was necessary to be used in Google Cloud platform. LSTM model uses the tab-separated file as input to produce etymological connections.

6.6. Tensorflow

Tensorflow is an open-source machine learning library developed by Google. It is used in the machine learning part of Etymøn. Keras is dependent on tensorflow

library to work. Hence, it is used to create the LSTM code for hallucinating new connections.

H5Py is used in combination with tensorflow to output the weight vectors of the LSTM model.

6.7. Google Cloud Console

Google Cloud Console is the location of the server of Etymøn. It holds the code that searches the enquiry through the dataset using bash commands and node.js. Google Cloud Console provides resources to analyze the activity of the server and has tutorials on how to get started easily, hence, it was chosen as the main node.js server. Free credits were used for this application.

6.8. Google Machine Learning Engine API

Google Machine Learning Engine API was used to train the LSTM model on the cloud, using faster processors. Google provides multiple sets of processors of arbitrary capacity to be used on machine learning model training. As the original input to the Etymøn machine learning module contains more than 300,000,000 characters, it is computationally expensive to run the model. However, as Google does not provide high end processors for free-credit users, it was decided afterwards to rely on subsets of the input data to be trained on local computers.

7. Use of Resources

Mainly online resources were used to cover multitudes of areas, ranging from website development to machine learning model creation. These resources were vital in understanding and developing on different platforms.

7.1. Resources for the Website

Main resources used for Etymøn's website are documentations about HTML, CSS and javascript. W3Schools and CodePen were appropriate resources on this matter. Three.js official website was also used for documentation and examples. The resources are described as follows:

- GitHub corners code that links to the GitHub repository of the code and has an animation of a cat [1].
- HTML5Up template for the HTML responsive website design [2].
- Sample code for the audio icon with bar animation [3].
- Sample code for filling animation of search and hallucination buttons on hover [4].
- Sample code for search button and hallucination button animation of enlarging on click [5].
- Resource about creating navigation bars that pop up from the side of the window [6].
- Documentation and examples of Three.js [7].

7.2. Resources for the Server

Server side code required javascript knowledge that was new. Hence, the following resources were useful in order to understand server architectures and creating http requests.

- Resource on executing unix commands in node.js [8].

- Tutorial on node.js server setup [9].
- Server sample code using node.js and express.js [10].
- Express.js documentation [11].
- Google Cloud Console quickstart with node.js [12].

7.3. Resources for Etymological Information

Etymological information was present on a diverse set of websites online such as etymonline.com and nisanyansozluk.com. These information had to be scraped from HTML. Here are the resources that were used while scraping:

- Nisanyan Sozluk Turkish etymology resource [13].
- Etymønline English etymology resource [14].
- Etymological WordNet resource [15].
- Wiktionary Resource to display etymological and meaning information [16].

7.4. Resources for Animation and Data Visualization

Animation and data visualization is also done in javascript in order to increase compatibility with the client-side code. Hence, javascript solutions such as cytoscape.js and three.js had to be learnt. The following resources were used while implementing in those libraries:

- Waves animation resource [17].
- Cytoscape.js documentation [18].
- Cytoscape.js twitter tutorial [19].

7.5. Resources for Machine Learning

Neural network creation is done using Keras and tensorflow. As these libraries have their own specific function uses, documentation and tutorials of them were useful. Furthermore, tutorials on implementing a whole LSTM network were used to create the model.

- Training LSTM Networks [20].
- Keras documentation [21].
- Tensorflow documentation [22].
- Google Cloud Machine Learning Engine documentation [23].

7.6. Miscellaneous Resources

These resources were of general help while developing python code and certain other API requests.

- Sample code on looking up a word using Wiktionary API [24].
- Python BeautifulSoup Documentation [25].
- Python regular expressions documentation [26].

8. User's Manual

8.1. General Information

8.1.1. System Overview

Etymøn is an analysis and tracing tool for word origins in all languages. It is also a data visualization application that demonstrates network graphs of links between words in different languages. Etymøn can be considered as an interactive art experiment as well, with its aesthetic visuals. It also contains a "hallucination" section, where a Long Short-Term Memory (LSTM) neural network generates new connections for given words.

Etymøn is currently available on web-browsers, such as Mozilla Firefox, Internet Explorer, Safari, and Google Chrome. Etymøn is also compatible with mobile phones and can be accessed through the phone's web browser.

8.1.2. Points of Contact

The points of contact (POCs) that may be needed for any troubleshooting, information, or assistance purposes are the creators of the website and team members of the Etymøn team. They are as follows:

- Mert Inan: mert.inan@ug.bilkent.edu.tr
- Utku Uckun: utku.uckun@ug.bilkent.edu.tr
- Nashina Ahmed: nashiha.ahmed@ug.bilkent.edu.tr
- Cholpon Mambetova: cholpon.mambetova@ug.bilkent.edu.tr

8.1.3. Definitions, Acronyms and Abbreviations

Some definitions of the Etymøn jargon are provided.

- The *Language Sea* is the first view that the user is greeted with. It is a zoomed out map of the most abundant words graphed together to make a sea like shape.
- The *Word Cloud* is a local graph for words clustered close to one each other.
- *Word Hallucinations* are brand-new words created from another word using neural networks.
- *App* denotes Application
- *LSTM* denotres Long Short-Term Memory neural network

8.2. System Summary

As aforementioned, Etymøn allows users to search for the etymology of any word in all languages. It shows this etymology in a network diagram. In addition, Etymøn provides word definitions and explanations through Wiktionary. Etymøn also creates word hallucinations when a user enters a source word. The user can also view the different relationships between words in a network diagram explained in further sections.

8.2.1. System Configuration

Etymøn has platform support for all web browsers on desktop and mobile devices. It features one main screen for the app that can be accessed through the website etymon.org by launching the app. Users can search for the etymology of a word by clicking the search button. They can create word hallucinations by using the hallucinate button. Finally, users can view instructions by clicking the help button. Lastly, users can turn background music on or off by clicking the music icon on the top right hand corner of the screen. Users can zoom in and out of the animation and navigate the 3D animation using their cursor. Users can also move the nodes of the network diagram of an etymology word cloud using their cursor. They can jump to another word by clicking the 'Jump to Word" button in the Wiktionary information of a word.

8.2.2. Data Flows

Users can input a word using their keyboards. The word can be changed after it has been typed by clicking the search or hallucinate fields. The word cannot be changed after it has been submitted, but a new word can be entered into the search or hallucinate fields.

8.2.3. User Access Levels

All users can only view information but cannot directly add or edit new information to the system.

8.3. Getting Started

8.3.1. Accessing the Page

The app can be accessed through the website etymon.org by clicking the button "Launch App." Alternative, the app can be accessed via the following link: http://etymon.org/website/etymon/index.html. The app can be accessed from all web browsers on phone and desktop.

8.3.2. System Menu

8.3.3. Searching a Word

After a user has entered the app, the user can utilize the functionality of searching the etymology of a word by clicking on and entering a word into the search field by clicking the search button which expands on click.

The word can then be searched by using the enter button on the keyboard or clicking the right arrow button on the screen using a cursor. The word can be deleted or edited by using the backspace on keyboard or clicking on the cross button on the search bar.

- **The Word Cloud**

  When a word is searched and a network diagram, the word cloud, denoting the etymology of the word appears, some edges may be colored according to the relationships between words. The searched word is in the center of the cloud. Pink edges denote the source etymology of the word. Blue edges denote words that come from the searched word. Therefore the searched word (in the center of the word cloud) is the etymology of the word at the end of a blue edge. Green edges denote etymologically related words. Light blue edges denote similar words.

  When a word from a word cloud is selected the panel with information on the word appears on the left side of the screen.

- **Word Definitions**

  The panel shows the word definition that is real time scrapped from Wiktionary website. The definition is very detailed and shows all possible ways of pronunciation, definitions with example, etc. that is available on Wiktionary.

- **Jumping to another Word**

  After the Wiktionary information of a word on the left side of the screen appears via a panel, the user can jump to that word so that that word is searched directly on Etymøn and its word cloud will appear. The user can do this by clicking the "Jump to [selected word]" button in the top of the panel.

8.3.4. Creating a Word Hallucination

The creation of a word hallucination is a similar process to the searching of a word. The hallucinate button denoted with an icon that contains

concentric circles can be clicked. Upon clicking the button, the hallucinate bar expands so the user can enter a word to hallucinate. The word can be deleted by using the cross button in the hallucinate bar or by using backspace on the keyboard. The word can be entered to hallucinate by pressing the enter button on the keyboard or clicking the right arrow in the hallucinate bar.

8.3.5. Background Music

The background music can be turned off and on using the button on the top right hand corner of the screen. When the bars are animating, the music is on. When the bars are still, the music is off. The music is on by default.

8.3.6. Exiting the System

The system can be exited by simply closing the window or tab on which Etymøn is running.

9. References

[1] Tholman, "tholman/github-corners," GitHub, 26-Sep-2017. [Online]. Available: https://github.com/tholman/github-corners. [Accessed: 03-May-2018].

[2] "HTML5 UP," HTML5 UP. [Online]. Available: https://html5up.net/. [Accessed: 03-May-2018].

[3] "Audio Icon," CodePen. [Online]. Available: https://codepen.io/joannaong/pen/olkzh. [Accessed: 03-May-2018].

[4] "Button Fill Animation," CodePen. [Online]. Available: https://codepen.io/davekilljoy/pen/wHAvb. [Accessed: 03-May-2018].

[5] "Awesome Search Button with Input Animation," CodePen. [Online]. Available: https://codepen.io/ey_intuitive/pen/vlcgf. [Accessed: 03-May-2018].

[6] "How TO - Side Navigation," How To Create a Side Navigation Menu. [Online]. Available: https://www.w3schools.com/howto/howto_js_sidenav.asp. [Accessed: 03-May-2018].

[7] "three.js," Three.js official website. [Online]. Available: https://threejs.org. [Accessed: 03-May-2018].

[8] L. Pollard, "Execute A Unix Command With Node.js - DZone Web Dev," dzone.com, 15-Aug-2010. [Online]. Available: https://dzone.com/articles/execute-unix-command-nodejs. [Accessed: 03-May-2018].

[9] "Build a simple Weather App with Node.js in just 16 lines of code," codeburst, 20-Jun-2017. [Online]. Available: https://codeburst.io/build-a-simple-weather-app-with-node-js-in-just-16-lines-of-code-32261690901d. [Accessed: 03-May-2018].

[10] "Build a Weather Website in 30 minutes with Node.js Express OpenWeather," codeburst, 26-Jun-2017. [Online]. Available: https://codeburst.io/build-a-weather-website-in-30-minutes-with-node-js-express-openweather-a317f904897b. [Accessed: 03-May-2018].

[11] "Express - Node.js web application framework," Express - Node.js web application framework. [Online]. Available: https://expressjs.com/. [Accessed: 03-May-2018].

[12] "Quickstart for Node.js in the App Engine Flexible Environment | Node.js | Google Cloud," Google. [Online]. Available: https://cloud.google.com/nodejs/getting-started/hello-world. [Accessed: 03-May-2018].

[13] "Nişanyan - Türkçe Etimolojik Sözlük," Nişanyan - Türkçe Etimolojik Sözlük. [Online]. Available: http://www.nisanyansozluk.com/. [Accessed: 03-May-2018].

[14] "Online Etymology Dictionary," Online Etymology Dictionary. [Online]. Available: https://www.etymonline.com/. [Accessed: 03-May-2018].

[15] Etymological Wordnet. [Online]. Available: http://www1.icsi.berkeley.edu/~demelo/etymwn/. [Accessed: 03-May-2018].

[16] "Wiktionary," Wiktionary. [Online]. Available: http://wiktionary.org/. [Accessed: 03-May-2018].

[17] Mrdoob, "three.js," GitHub. [Online]. Available: https://github.com/mrdoob/three.js/blob/master/examples/canvas_particles_waves.html. [Accessed: 03-May-2018].

[18] M. Franz, "Cytoscape.js," Cytoscape.js. [Online]. Available: http://js.cytoscape.org/. [Accessed: 03-May-2018].

[19] "Graphing a social network," Graphing a social network · Cytoscape.js. [Online]. Available: http://blog.js.cytoscape.org/2016/07/04/social-network/. [Accessed: 03-May-2018].

[20] "Text Generation With LSTM Recurrent Neural Networks in Python with Keras," Machine Learning Mastery, 08-Jan-2018. [Online]. Available: https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/. [Accessed: 03-May-2018].

[21] "Keras: The Python Deep Learning library," Keras Documentation. [Online]. Available: https://keras.io/. [Accessed: 03-May-2018].

[22] "Installing TensorFlow | TensorFlow," TensorFlow. [Online]. Available: https://www.tensorflow.org/install/. [Accessed: 03-May-2018].
[23] "Predictive Analytics - Cloud Machine Learning Engine | Google Cloud," Google. [Online]. Available: https://cloud.google.com/ml-engine/. [Accessed: 03-May-2018].
[24] Nichtich, "Look up a word in Wiktionary via MediaWiki API and show the Wiktionary page," Gist. [Online]. Available: https://gist.github.com/nichtich/674522. [Accessed: 03-May-2018].
[25] "Beautiful Soup Documentation," Beautiful Soup Documentation - Beautiful Soup 4.4.0 documentation. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed: 03-May-2018].
[26] "7.2. re - Regular expression operations," 7.2. re - Regular expression operations - Python 2.7.15 documentation. [Online]. Available: https://docs.python.org/2/library/re.html. [Accessed: 03-May-2018].