Óravíddir: Network Visualization of the Icelandic Vocabulary

Eva María Jónsdóttir, Jón Hilmar Jónsson, and Trausti Dagsson[0000−0001−8537−8771]

The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland
arnastofnun@arnastofnun.is

Abstract. This paper describes an interactive exhibition about the vocabulary of the Icelandic language. The exhibition is called *Óravíddir – Orðaforðinn í nýju ljósi* (e. Vastness – The Vocabulary in a New Light) and was opened at the Culture House in Reykjavík, a part of The National Museum of Iceland in May 2019. The exhibition used data from the word database *Íslenskt orðanet* (The Icelandic Word Web) and illustrates semantic relations between words in a three-dimensional visualization. The paper introduces *Íslenskt orðanet* followed by a description on how the data was used to create the network graph visualization. Then we discuss the setup of the exhibition and finally we conclude by reflecting on future possibilities and further development.

Keywords: Network, Lexicography, Exhibition, Word net, Visualization.

1 About Íslenskt orðanet

The online word database *Íslenskt orðanet* (The Icelandic Word Web) was launched in 2016. The database can be used much like a thesaurus and contains a description of Icelandic words and multiword expressions. The focus is on giving examples of usage, where lexical and syntactic relations mirror semantic association and relationships. Parallel structures, i.e. words or expressions linked by 'og' (e. 'and') play a major role, as their analysis shows how closely related different items are, giving the description of synonyms a new and more dynamic aspect. The single-word and multi-word entries are linked to related entries in the database, where a thorough analysis of parallel structures gives a graded overview of semantic relatedness and proximity. The main work on *Íslenskt orðanet* is carried out by research professor emeritus Jón Hilmar Jónsson at The Árni Magnússon Institute for Icelandic Studies[1] [2].

The most basic usage is search for synonyms and antonyms but the search results also consist of closely related words connected through common partners in parallel structures. The search for the icelandic word for sea (icelandic *haf*) results for example in the synonyms *mar*, *sjór* and *sær* but it includes also related words like *ey* (en. *island*), *fjall* (en. *mountain*), *strönd* (en. *coast*) and *land* (en. *country*, *landmass*). More closely related words or "near-synonyms" are compounds like *úthaf* (en. *ocean*), *hafsvæði* (en. *marine areas*), *innhaf* (en. *marginal sea*) and *reginhaf* (en. *vast ocean*).

Íslenskt orðanet currently (october 2019) includes 102097 single headwords and 93876 multiword expressions which also are searchable.

2 From data to visualization

Network graphs are a popular and convenient way of displaying data entities that have many-to-many connections to each other. This is especially the case in collections where each entity can have a connection to any other entity in the collection. In network graphs, entities are presented as nodes and the connections as edges.

Network graphs can be a suitable way of displaying the types of interconnected entities as described above and can be of great help to find clusters in the network, groups of entities that have unusually many relationships to other entities. They should however be looked at as only one of the tools to approach a dataset and to help researchers get an overview of their data but not one that gives an absolute answer.

Since the data in *Íslenskt orðanet* consists of entities which are interconnected it was quite straightforward to attempt to visualize it as a network graph. Similar visualization projects have emerged outside of Iceland, for example visualization tools for the data from The Princeton University WordNet[3], a visualization tool for the Slovene wordnet[4] and a tool to view semantic relations between words in Warlpiri, an Australian Aboriginal language[5]. Our project is the first attempt to visualize an Icelandic lexical database in such a way as far as we know. The database technology we choose was Neo4j. Neo4j has support for the Cypher query language which enables developers to write queries to traverse the network and analyze relationships between specific nodes or nodes with specific attributes. Neo4j can be either set up as desktop service or on a web server where a middle layer, often an application programming interface (API) is used to send queries to and deliver results from the database and a user interface[6].

For the visualization we choose a subset of *Íslenskt orðanet* which contains information on lexical items which are related through common partners in parallel structures, consisting of words or expressions linked by 'og' (e. 'and') in a textual context. Since the subset includes relations like *horse – dog*, *horse – sheep*, *sheep – human*, *human – nation* and so on, the chosen subset is richly interconnected. First attempts at importing it into the graph database showed us that almost every node in the dataset had connections to other nodes, thus the entire dataset could be used to produce a very large network graph. The data was imported from a database table via script that converts each pair of related words to a Cypher query which are run via the Neo4j API interface. The database had been configured beforehand to ensure that each pair could only exist once to avoid misleading duplicates in the end result. To further ensure accuracy, the word entity ID's from the original database was used instead of using the headword itself as identifier. At the end of this process, we had a graph database consisting of 24828 nodes and 264974 edges.

To make it possible to view the visualization on different devices and to easily enable interaction we used web development technology to produce the graph. When it comes to visualization in a web browser there are many options available. The emergence of vast variety of JavaScript frameworks last years make it relatively easy to plug a visualization component into a web application project. The D3.js framework is a powerful and popular framework which can render various types of visualizations, including force-directed network graph which uses force simulation to position edges in a network in such a way that clusters of edges with more interconnections than others are positioned closer together[7]. The D3.js framework was used for prototyping but for the production visualization we wanted something that could provide more visually appealing experience for the viewers. The large number of data lead us to a visualization in a three dimensional space. Alternative would have been a two dimensional network graph but considering the number of nodes we choose a three dimensional graph which would enable us to view the graph from different angles. For the 3d visualization we used a JavaScript code library called 3d-force-graph which uses part of the D3.js framework as well as the WebGL technology to render the network graph using GPU (Graphics processing unit)[8]. The base rendering engine is built using the Three.js code library which is developed to make it easier to render 3d models in a web browser[9]. Running 3d content in a web browser can however be challenging when it comes to performance and speed. Our solution was therefore to render the words as two dimensional text sprites instead of 3d objects. The edges were rendered as triangular prisms meaning that each edge only required six vector points. Limiting the total points used in the graphic significantly increased performance speed.

3 The Exhibition

The National Museum of Iceland has its main exhibition in the main building which was opened in 1950. The older museum building, *Safnahúsið* (e. The Culture House) which also housed The National Library of Iceland and National Archives is still used today for exhibitions from various museums and cultural institutes[10]. The Árni Magnússon Institute for Icelandic Studies has an agreement with the Culture House which states that the institute will have a one year exhibition in one of the exhibition rooms every three years, switching places with The National Library and National Archives. In June 2019, The Árni Magnússon Institute for Icelandic Studies opened the exhibition which was built around the network visualization of *Íslenskt orðanet*.

For the exhibition we rendered a video of the network graph where the virtual camera is moving through the network space (figure 1). The video, which is approximately 15 minutes long demonstrates a flight between random words where it stops briefly in front of a word with the neighbor words around it. The speed of the flight is slow to make the viewer able to easily observe the words that travel past the camera. In the end the video stops at nearly the same position it started on to enable us to seamlessly loop it. The video does not display the whole graph produced from the database but rather a subset based on connections between a set of chosen words. We defined a list of words related to nature (f.ex. flora and landscape features), weather, seismic activity, culture (f.ex. arts, humanities, literature, education), society and production (f.ex. recycling, sustainability, production, oil, heavy industry) and travels. By this we wanted to have a subtle focus on environmental issues and to draw attention to the vast number of Icelandic words related to landscape and weather in particular. The graph database was queried for the prechoosen words and all related words at maximum two nodes distance. This way we produced a graph containing approximately 20000 words.

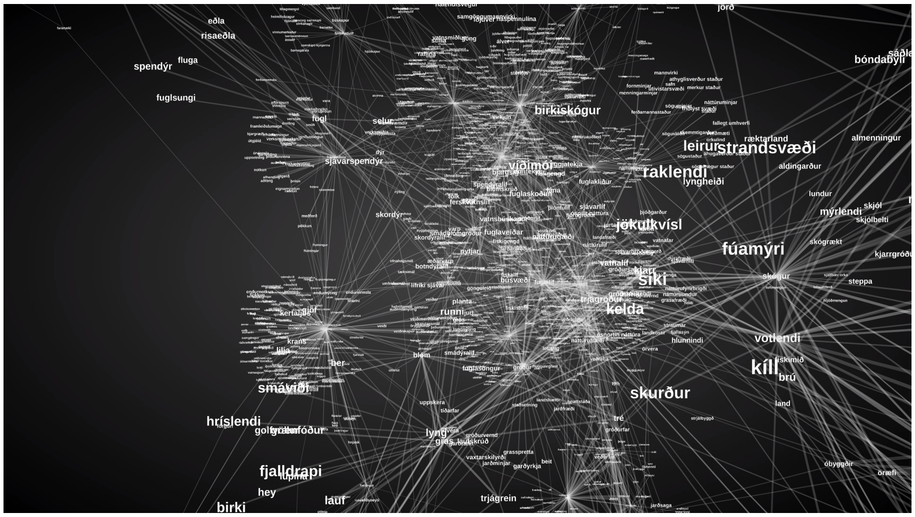

Fig. 1. Screenshot of the rendered video.

The video was projected to a wall accompanied with two touch screens where users can interact with the data. On one of the screens guests can select a word resulting in a graph emerging on the screen showing related words at maximum two nodes distance. On the other screen users can select two words which results in a graph visualizing all shortest paths between the two words. These touch screens were powered by Raspberry Pi computers because we wanted the appearance of the screens to be minimalistic and because they are physically quite small, the Raspberries gave us a chance to install them on the back of the screens. Using Raspberry Pi introduced a challenge since their power is limited and the three dimensional graphic tended to be heavy when rendering larger numbers of nodes and edges. A reasonable solution was to run the operating system (Raspian) running only the web browser Google Chrome displaying an optimized version of the web application for visualizing network for a single word and the shortest path between two words (figure 2). The optimization of the visualization consists of applying parameters to the WebGL process of the browser to limit the rendering quality of the graphic as well as limiting the 3d models vertices by rendering the edges between the nodes as three-sided prisms instead of cylinders.

On the surrounding walls we put up embossed letters, forming words whose use is declining and rising today based on word lists from the databases *Nýyrðavefurinn* (e. The Neologism Database)[11] and *Risamálheildin* (en. The Gigaword Corpus)[12] both of which are administered by The Árni Magnússon Institute for Icelandic Studies. This was to emphasize how words come and go, similar to the visual experience from the video projected on the wall.

Along with the exhibition, we opened a website with information about the exhibition and where users can also visualize paths between two words, the website is accessible at the url http://oraviddir.arnastofnun.is.

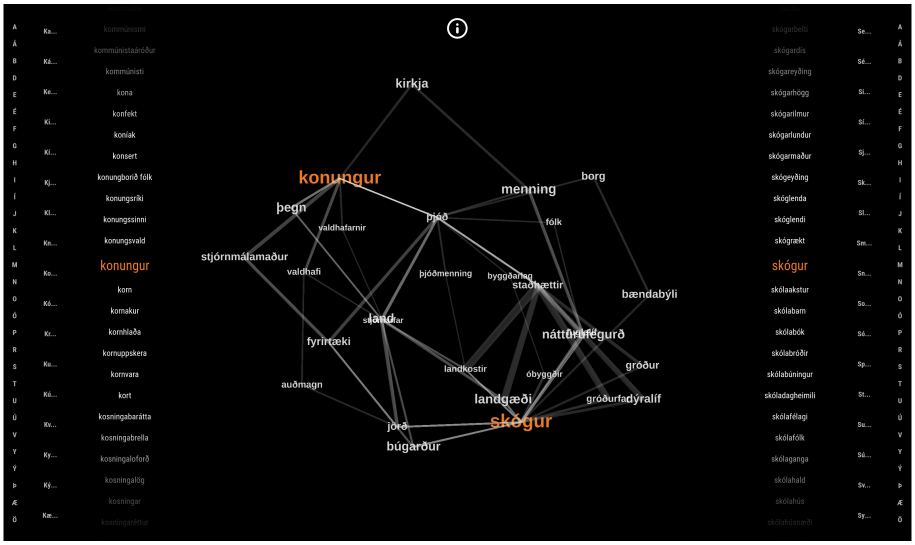

Fig. 2. The touch screen interface showing relations between the words *skógur* (e. *forest*) and *konungur* (e. *king*).

The name *Óravíddir – Orðaforðinn í nýju ljósi* (e. Vastness – The Vocabulary in a New Light) was chosen as a name for the exhibition. The word *óravíddir* roughly translates to "vastness". It is coined from two words, *óri* (e. fantasy, imagine) and *vídd* (e. dimension) and can also be translated to "vast dimensions". This name can therefore refer to the visual aspect of the large word graph where words are drawn white on a black background, strongly resembling stars, nebulas and constellations in space. The word is also often used together with the word *himingeimur* (e. space; *óravíddir himingeimsins*, "the vastness of space").

4 Conclusion and further work

The work on the exhibition was not intended to be an end product but rather one way to approach the data and an example of how such data can be mediated using highly visual components. The idea in the future is to make this exhibition mobile by moving it to other parts of the country from time to time. We want to encourage schools to bring their students to see it and use it as a starting point for conversation about the language. Further on we want to utilize the graph database technology to further develop the website for *Íslenskt orðanet* (http://ordanet.arnastofnun.is) in a way where users can browse the data in more graphical way, a solution similar to WordVis which is based on Princeton WordNet[3].

Our aim with this project is to illustrate that the Icelandic language is rich with words which might be rarely used today due to changes in society and language use. With the exhibition we want visitors to view the vocabulary as a large interconnected space which contains a vast number of words, some of which are commonly used but

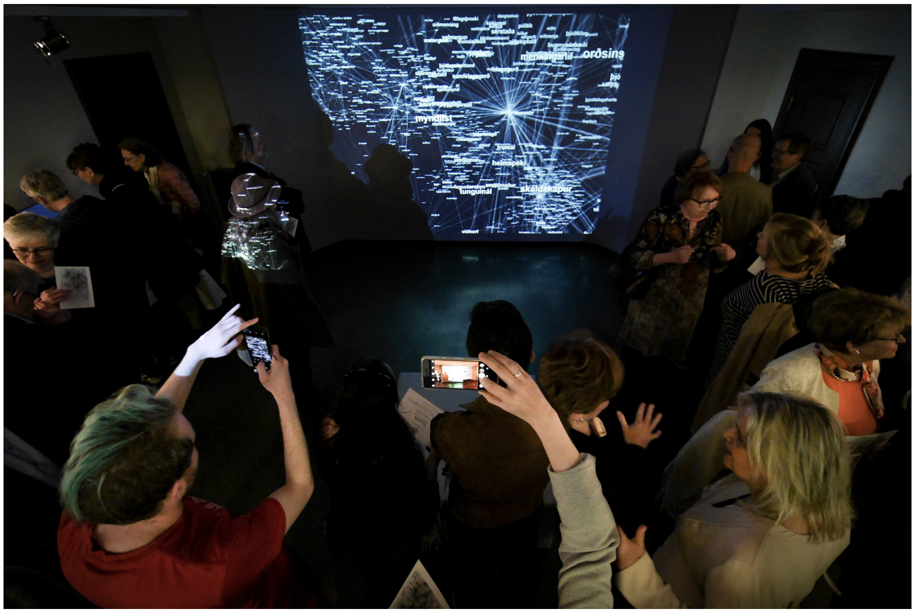

Fig. 3. Guests at the exhibition opening (photo: Sigurður Stefán Jónsson/The Árni Magnússon Institute for Icelandic Studies).

others that are less used but nevertheless interesting. By utilizing visualization and interaction we hope to raise interest in the language and creative language use among younger generations.

References

1. Jónsson, J. H.: Íslenskt orðanet: Tekstbasert kartlegging og presentasjon av leksikalske relasjoner. In: Nordiske Studier i Leksikografi 14. Rapport fra 14. Konference om Leksikografi i Norden. Reykjavík, pp. 1–17 (2017).
2. Rögnvaldsson, E.: Íslenskt orðanet: a treasure for writers and word lovers. In: LexicoNordica. Fackspråk i nordiska ordböcker, vol. 25, pp. 313–328 (2018).
3. Vercruysse, S., Kuiper M.: WordVis: JavaScript and Animation to Visualize the WordNet Relational Dictionary. In: Advances in Intelligent Systems and Computing. Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, pp. 137–145. https://doi.org/10.1007/978-3-642-31603-6 (2012).
4. Fišer, D., Novak, J.: Visualising sloWNet. In: Proceedings of eLex 2011, Bled, pp. 76–82 http://www.trojina.si/elex2011/elex2011_proceedings.pdf (2011).
5. Manning, C., Jansz, K., Indurkhya, N.: Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri. In: Literary and Linguistic Computing, vol. 16, no. 2, pp. 135–151. https://doi.org/10.1093/llc/16.2.135 (2001).
6. Neo4j Graph Platform – The Leader in Graph Databases. https://neo4j.com/
7. Bostock, M.: D3.Js – Data-Driven Documents. https://d3js.org/
8. Asturiano, V.: 3D Force-Directed Graph. https://vasturiano.github.io/3d-force-graph/

9. Three.js – JavaScript 3D library. https://threejs.org/
10. Safnahúsið – The Icelandic Culture House, http://www.culturehouse.is/information/
11. Dagsson, T., Þorbergsdóttir, Á., Steingrímsson, S.: Nýyrðavefurinn: A Website for Collection and Dissemination of Icelandic Neologisms. In: 4th Digital Humanities in the Nordic Countries, Copenhagen. (2019).
12. Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E.: An Icelandic Gigaword Corpus. Nordiske Studier. In: Nordiske Studier i Leksikografi 14. Rapport fra 14. Konference om Leksikografi i Norden. Reykjavík, pp. 246–254 (2017).