Etymap Process Book

Jump to the story, directly to the demo (screencast) or to Github.

Words & etymologies interactive visualization

Proposal

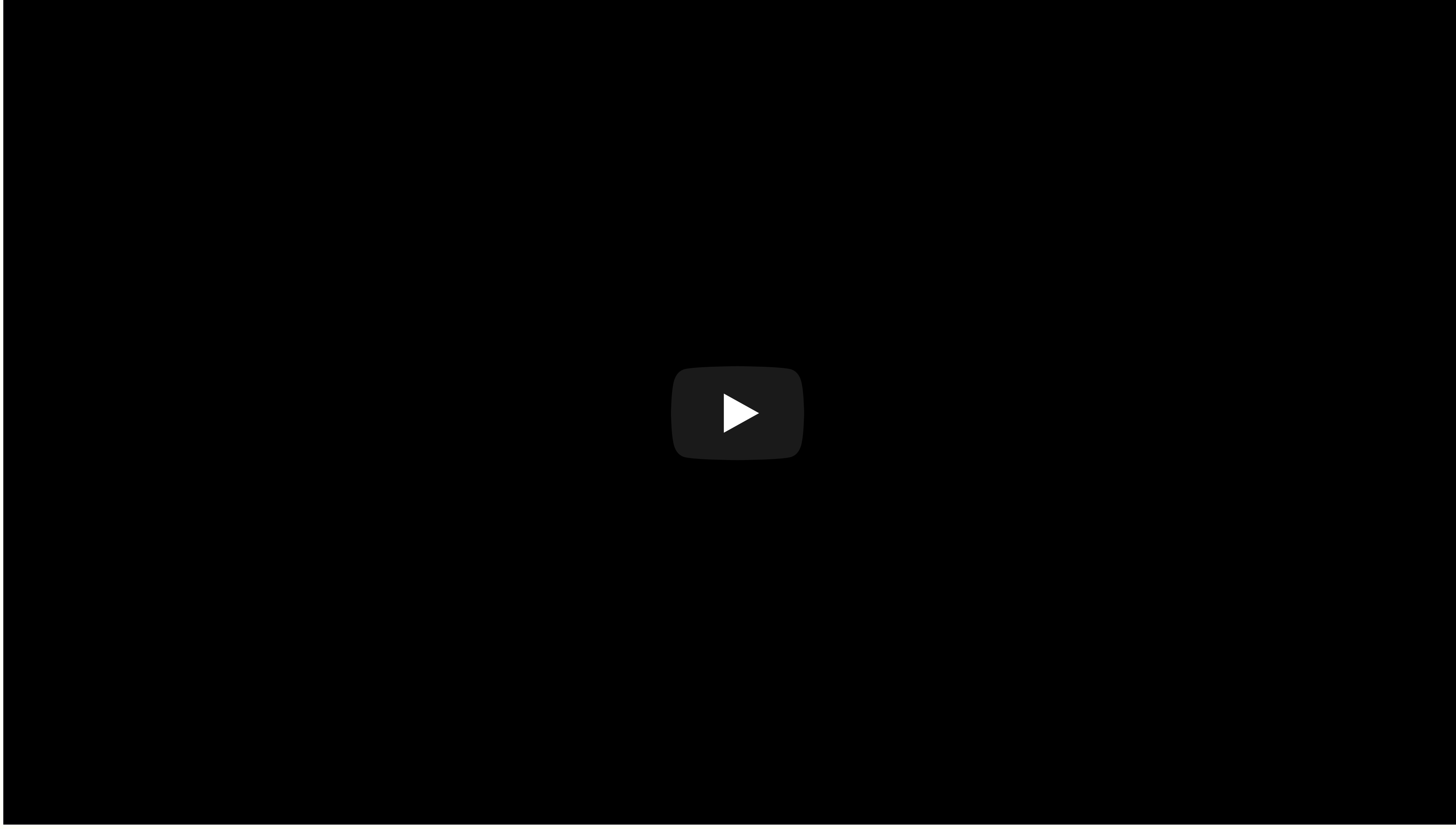
We are planning to build an interactive map which illustrates the etymology of words. Etymology is a subject which spans across time and space, and therefore creates a challenging and interesting problem to visualize.

In short, the main feature will allow the user to pick a word (or a group of words) and see its etymology as an insight (paths, directions, ancestors, periods, etc.) on the map. There are other interesting opportunities, for example, we could show words with different meanings which share a part of their etymology or other relations with them (synonyms, antonyms, homonyms).

The target audience would be linguists, historians and more generally anyone curious about the words we use everyday.

Some similar work can be found on the "etymology maps" subreddit. However most visualizations are not interactive and only project one aspect of the data.

Screencast



Initial wireframe

We will now describe and discuss our initial wireframe and the motivations behind each decision that went into making it. Of course, this only serves as a starting point. The closer we will get to a functional prototype, the more we will learn and understand about our dataset and how our visualization shows its features. The future insights we will gain, will therefore most likely change the visualization from the wireframe we describe here.

There are two main aspects of the data that we want to show. The first one being how languages relate to each other etymologically, and the second one is at a finer grain, namely the etymological tree of specific words. This already plays nicely into the "context and detail" principle, since it allows the user to see these two different levels of the data.

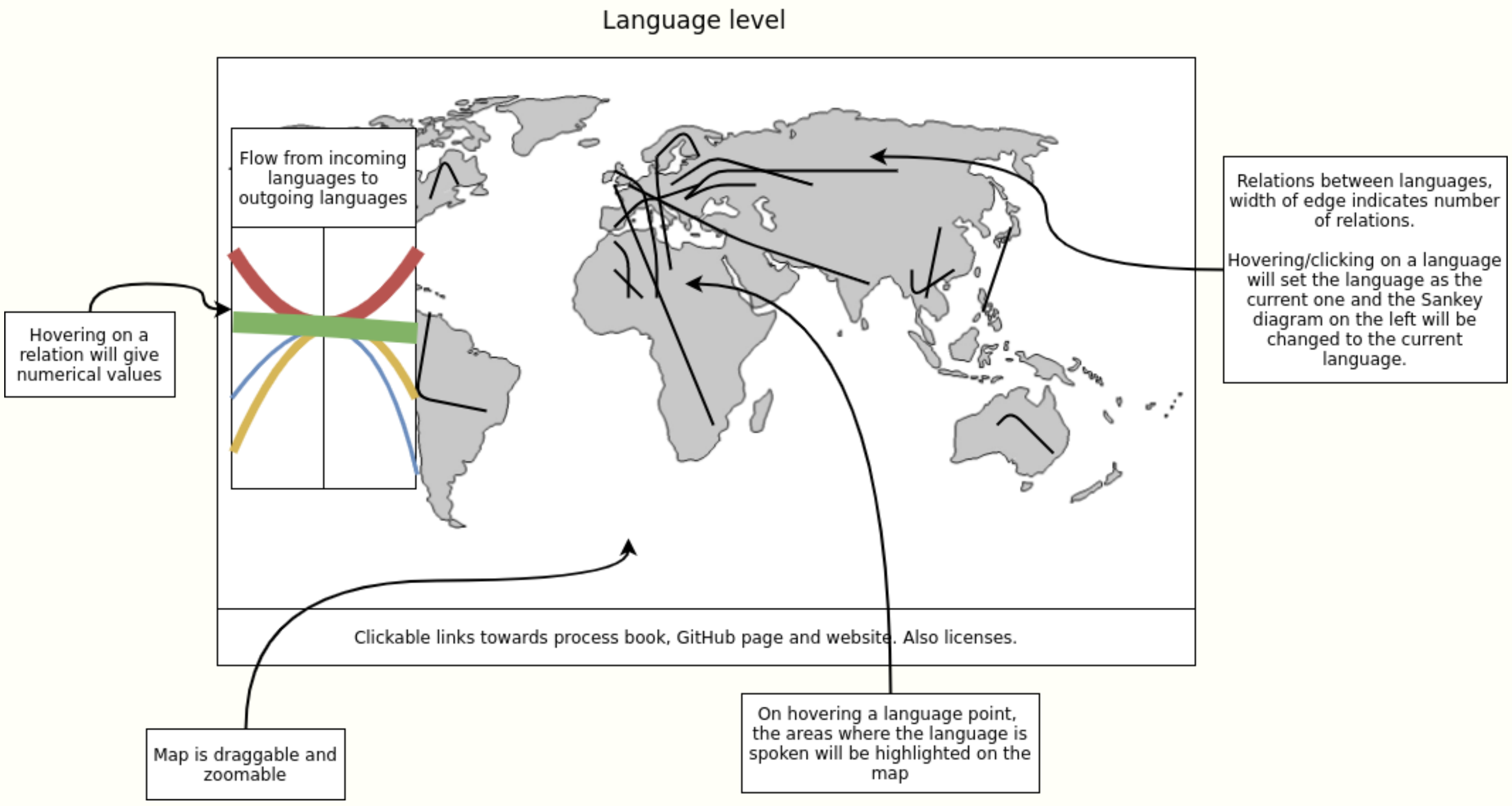
The core component of our visualization is a map. This has the added value of giving geographical information for these relationships. It could therefore potentially show how languages moved with time. The other obvious approach would be a graph, however it would make it much harder for the end-user to search for a specific language/word. There are a few downfalls of the map, namely the fact that certain areas have a much higher density of languages than others, hence making the visualization possibly cluttered. Additionally, certain languages may be spoken across several disjoint areas making the geographical information either hard to read or ambiguous.

We decided on using a map as it gives an initial structure to the underlying graph and can be removed if the previous mentioned pitfalls, or those to come, are clearly detrimental to the visualization.

As described earlier we want to have have two levels of visualization. We will informally call them the "language level" and the "word level" respectively from now on.

We will now describe them individually and the interactions which allow the user to switch between them.

⊕

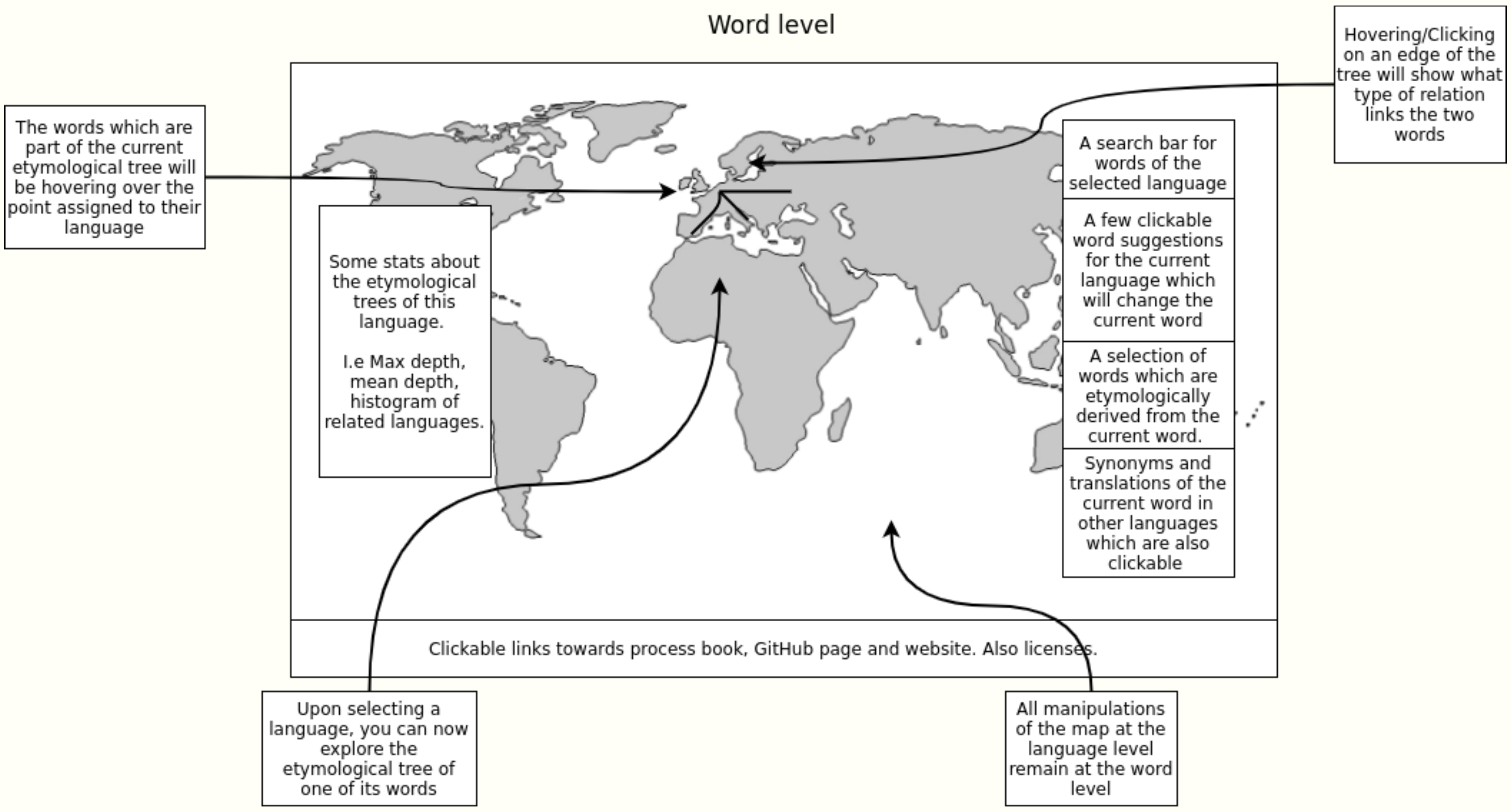

Language level
Flow from incoming languages to outgoing languages
Relations between languages, width of edge indicates number of relations.
Hovering/clicking on a language will set the language as the current one and the Sankey diagram on the left will be changed to the current language.
Hovering on a relation will give numerical values
Clickable links towards process book, GitHub page and website. Also licenses.
Map is draggable and zoomable
On hovering a language point, the areas where the language is spoken will be highlighted on the map

For the language level, there will be bundled edges between languages which are assigned to areas on the map. The edges will represent the relations between words of the linked languages. Above each language area there will be the name of the language assigned to it, it will also highlight the area when hovering on it. These names will be clickable and allow the user to select a language. By selecting a language, a panel to the left of the map will appear which holds a Sankey diagram of the incoming and outgoing relations (per language) to the selected language. Hovering on this diagram will give additional numerical information to the user. Of course, the underlying map will have all the usual interactions one can expect from a map, namely dragging and zooming.

We have chosen to use bundled edges as we are only interested in the distribution of relations per language at this level. For the same reason, the Sankey diagram is an appropriate visualization to have. The reason we want both to be shown at the same time is because it will allow the user to isolate a language. It also allows for easier comparisons between languages, because two languages might have similar distributions but could be geographically quite different - which will make it hard for the user to compare them. This Sankey diagram proposes a sort of "standard" representation of these relationships.

In order to switch to the word level, we are not quite sure what is the best/most appropriate interaction. We are currently considering a double-click on the language or the bundled edges. We could also have an interaction on the Sankey diagram and not only the map.

⊕


Word level
The words which are part of the current etymological tree will be hovering over the point assigned to their language
Some stats about the etymological trees of this language.
I.e Max depth, mean depth, histogram of related languages.
Hovering/Clicking on an edge of the tree will show what type of relation links the two words
A search bar for words of the selected language
A few clickable word suggestions for the current language which will change the current word
A selection of words which are etymologically derived from the current word.
Synonyms and translations of the current word in other languages which are also clickable
Clickable links towards process book, GitHub page and website. Also licenses.
Upon selecting a language, you can now explore the etymological tree of one of its words
All manipulations of the map at the language level remain at the word level

Once the user entered the word level, he will have the exact same interactions with the map as previously. The languages will still have the same interactions too (highlighting areas, etc.). Upon first entering the word level, a random word will be selected. There will now be edges on the map which represent the etymological tree of this word. On the right side, there will be a panel allowing the user to search for a specific word, or pick among suggested words to examine. Whereas on the left there will be some statistics about the etymological trees of the current language: max depth of trees, mean depth, histogram of related languages.

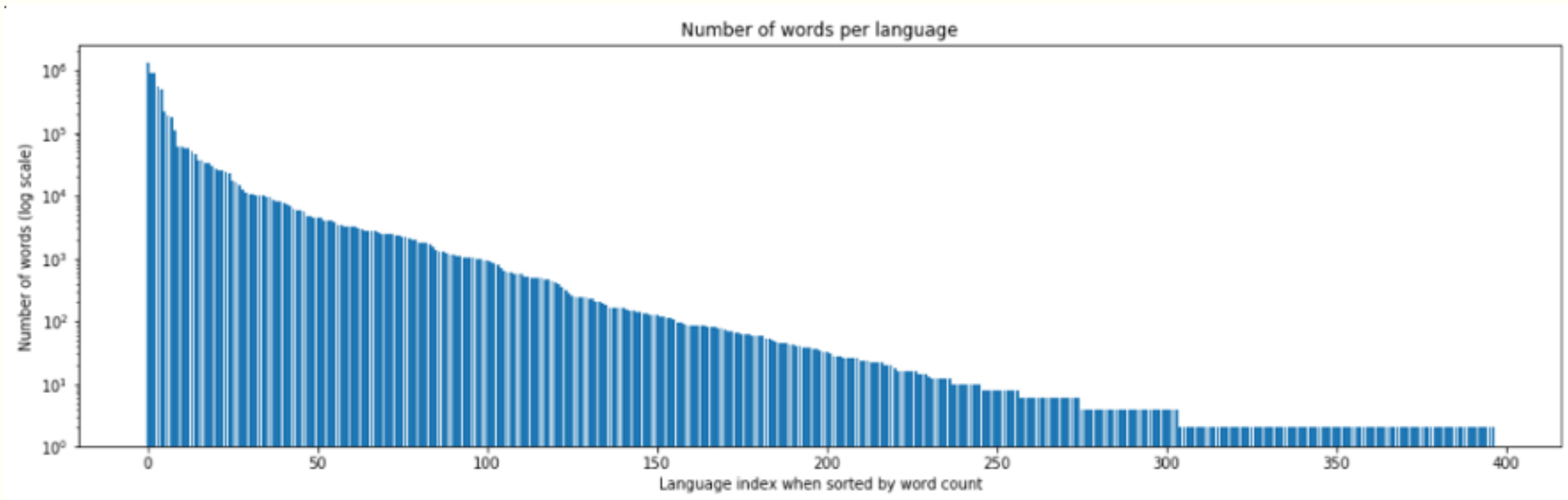
Etymwn dataset

The core dataset on which our data visualization will rely is the Etymological Wordnet abbreviated Etymwn from here on. This dataset is a great starting point as it describes the etymological relations between words which is the foundation of what we are trying to show.

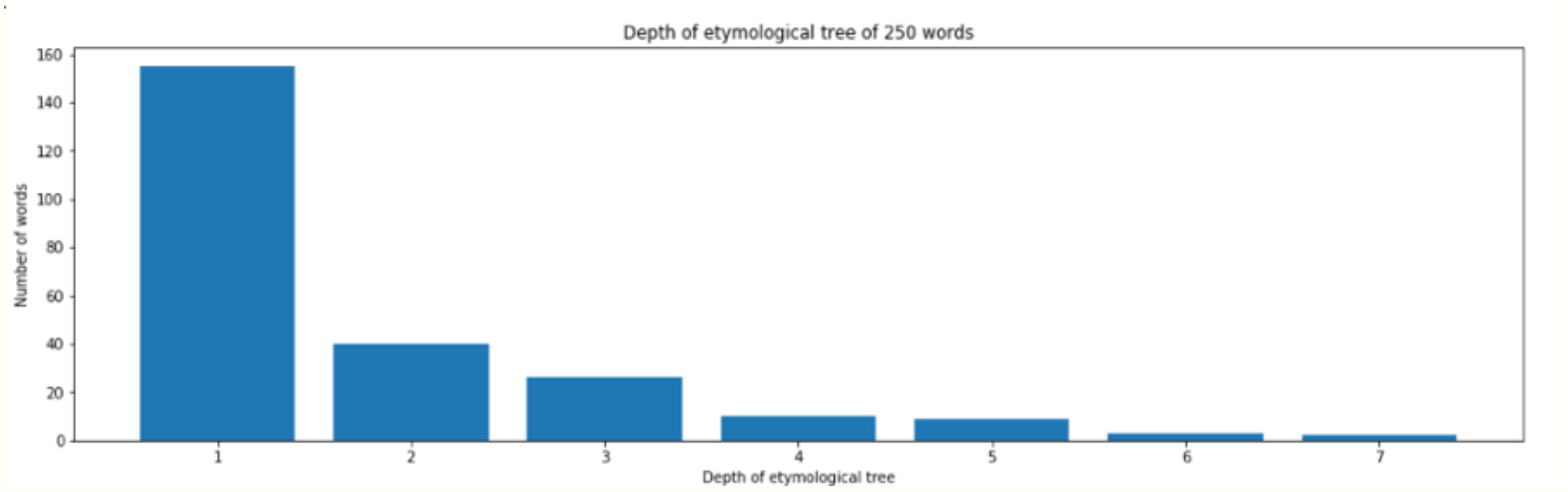
The Etymwn dataset is built on data available from Wikitionary. The entire dataset compromises about 2.5m unique words which are used in 6m different relations. Given the nature of the source on which it is built on, we expect some bias (in completeness) towards North American and European languages. This will also be something to show on the data visualization. How for example African languages are less documented and hence less represented on the visualization. In order to better understand the data at hand, we started by doing a simple exploration of it.

The goal is to see how large each one of our main features is and to already gain insights as to what would be particularly interesting to highlight in our visualization. By "main" features, we mean: the number of unique words per language, the number of languages, how many languages are related to a given language and the etymological tree of a given word.

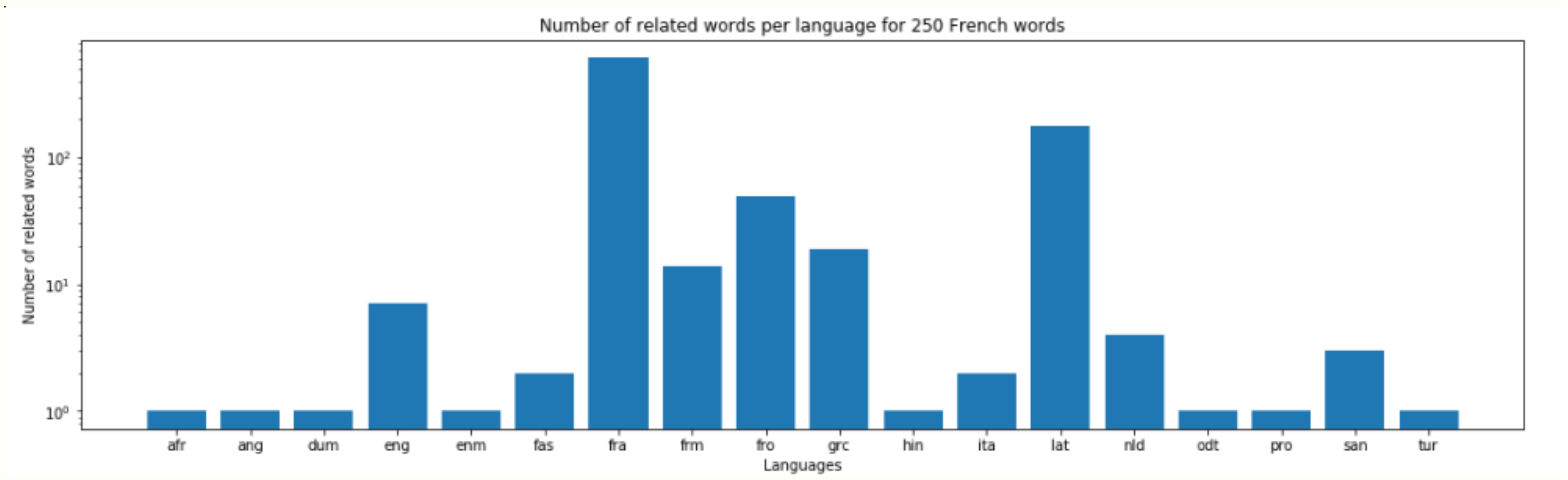
The dataset has eight different relations possible between two words. We are only interested in certain of them, as other might be ambiguous or not add any information to the etymological origins of words (i.e knowing that "nicely" is a derived form of "nice"). By performing this initial filter, we practically cut the dataset into half, having 2.7m relations remaining.

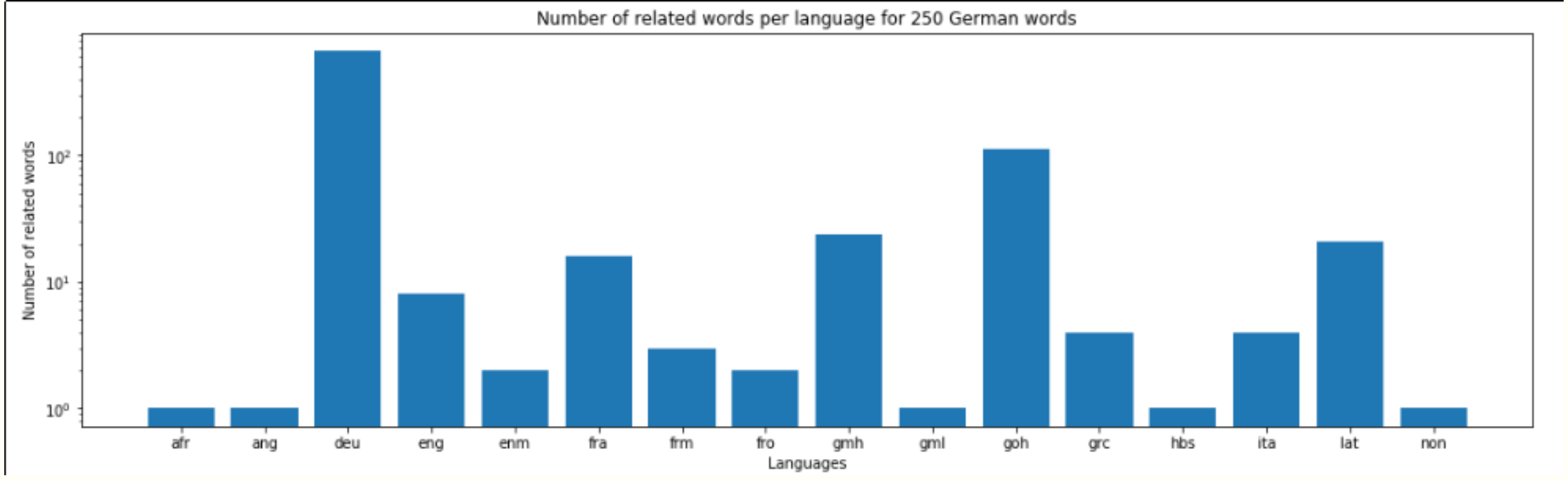


The first insight is about the number of words per language. The dataset only holds words which have an etymological relationship, hence all other words are discarded. On the figure above, we can see that due to this there are languages that have very few words. This could also be the result of a poorly documented language on Wikitionary. Either way, we will have to deal with the tail of the distribution in the visualization.



The depth of the etymological tree for a given word is a very important aspect in our visualization. Depending on this depth the tree might be very large and hence given a sense of clutter which we will have to deal with. The figure above illustrates the depth of the tree for 250 words sampled uniformly from the whole dataset. The main insight we gain from this is that these trees are relatively shallow, and will therefore not be computationally difficult to draw. This could potentially allow us to draw trees of multiple words at the same time to gain other insights about the data.





Finally, we explored the relationships for two languages in particular : German and French. This figure illustrates the languages to which these are related (directly or indirectly) by sampling 250 words from the language and constructing their etymological tree. As described in our wireframe, we would have only wanted to plot this feature for directly related languages and not using the full tree as it is currently done in this figure. Given that there are not that many languages related even indirectly to a given language, we could adapt our visualization to show this feature instead or add it in.

TL;DR:
We use the Etymological Wordnet dataset. From a simple exploration of it's data we understood that certain languages will have very few words, and that more generally the graph of related languages or the etymological trees of words are pretty shallow allowing us to potentially draw more of these simultaneously than we had planned.

Giving more thought to synonyms

We think that our visualization could greatly benefit from the added insights "synonyms" bring. The quotes are used here to actually describe all words share a same meaning, they can therefore span across different languages. You could consider them as traditional synonyms and translations.

These would particularly be interesting to visualize since it could show if there are languages which do not have a specific word for something. They could also be used to see if all the synonyms and translations of a word share parts of their respective etymological tree. For example we could see if all the "synonyms" of "flat" have "πλατύς" in Ancient Greek as their etymological origin.

The way this would be visualized is at the word level of our described wireframe. As previously discussed, the visualization would have a list which would allow the user to pick synonyms of the current selected word. The user would still have to go back and forth between synonyms to see if they share parts of their etymological tree. Given this inconvenience, we are currently prototyping a version where synonyms would have different colored edges. The overlapping parts of their trees would be highlighted and weighted accordingly. Most of these decisions will be made once we have a clear idea of how often these cases occur.

The most relevant and complete dataset we found is the UWN / MENTA dataset. It is based on WordNet which groups words into cognitive synonyms. The UWN dataset does the same, but additionally incorporates other languages - not only english. There are a few particularities about this dataset, namely the fact that each pair of words which share a meaning have a weight assigned to their link. This weight defines how close the meanings are to each other. We are not truly interested in this aspect of the data, we therefore entirely dropped this aspect of the data and assumed that as long as there is a relationship it should be considered a synonym or translation.

We have not yet done a in-depth data exploration of this dataset. However, we have checked that the languages documented in the Etymwn overlap with this set. It is indeed the case. We do still want to check if they both suffer from the same skew - if some languages have significantly more data than others.

Partial first prototype

In the initial wireframe we had also discussed some interactions when hovering on language names on the map. In most cases, the amount of language points or edges depending on the user input is again too high to make this interaction appropriate. In short, it is practically impossible or even simply annoying to move the mouse away from any data on the map - the user would therefore always be highlighting data from something he did not actually select. This was by no means something important to the visualization, but rather a quick and easy way to get more information about languages without selecting them. We removed this entirely, and do not intend to add it again. However, by adding a router to the visualization, the user will be able to quickly push and pop his history and can therefore still rapidly get quick information about a language and still go back to his original selection.

The bundled edges for each relationship between two languages was not a bad idea. However, given that we have these two clear levels of data, there is no actual point in having individual edges per relationship. In fact, it is more consistent to just have a single link between the two languages. This level of the visualization is not intended for detail at all. Of course, we need a way to get down to the detail, but this is definitely not the best approach. Clicking on a single edge is difficult, and doesn't really allow to explore the interesting words. Given the minimal benefit of this, we will continuing with single edges between languages instead of a bundle of edges. We still need to find an accurate way of representing the weights of these edges, it could be done by color or width for example. This is a particularly hard problem given the logarithmic number of words per language there is (as we saw in the data exploration).

Now we can move on to the world level.

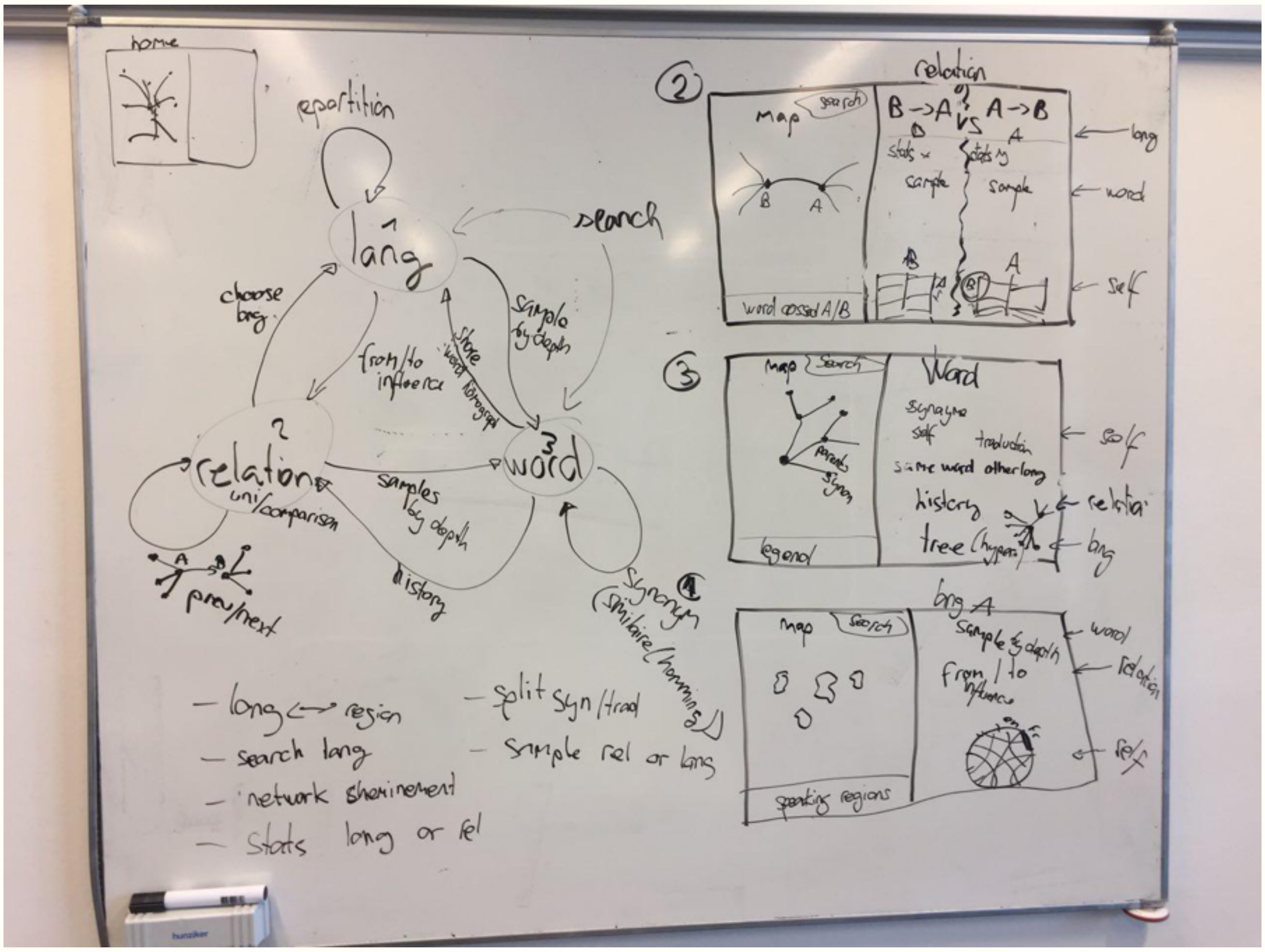
We are currently able to build and draw both the tree of all derived words for a given word and the tree of all words which constitute it's etymological origin. There are some obvious challenges which we had not thought about in the initial wireframe. In our data exploration we had only evaluated the depth of these trees, but not at all their number of words. This can lead to the problem where Latin words for example are not particularly deep, but are the origin of practically all European languages making the tree very wide. This is a problem we will have to spend additional time on and for which we do not have an immediate solution.

Refining our visualization's direction

From the previous prototype we already noticed and fixed some of the obvious flaws which are mostly related to usability. We will now take the time to also talk about what our visualization is doing right/wrong more generally. These opinions are based on our own personal views of it as well as feedback from users which had not heard about this project at all.

The map plays an essential role in locating well-known relations, but it's hard to accurately display data on it without having a bias due to the projection, or other spatial constraints. We initially only wanted the side panels as places to give additional information, but in reality these should be used as the main source for precise measures. They are a core part of the visualization and should therefore be more prominent, or at least more informative.

A second big change for our visualization, is the transition between words and languages. They initially seem quite closely related, but in reality for our visualization we need another layer in between them. The reason for this is that it's very interesting to inspect all words which are part of relationship between two given languages. For example, it can be interesting to examine the words which are related to English and Japanese to understand their overall nature. We therefore want to add what we call a "relation level" which allows the user to specifically target an etymological relationship between two languages.



The figure above illustrates the core new direction for our visualization.

On the left side of it, we have illustrated our visualization as a simple state machine between the three levels (word, language and relation). The edges between the states are the aspects of the data we can use to switch between them.

On the right side, there are extremely simple wireframes of each level.

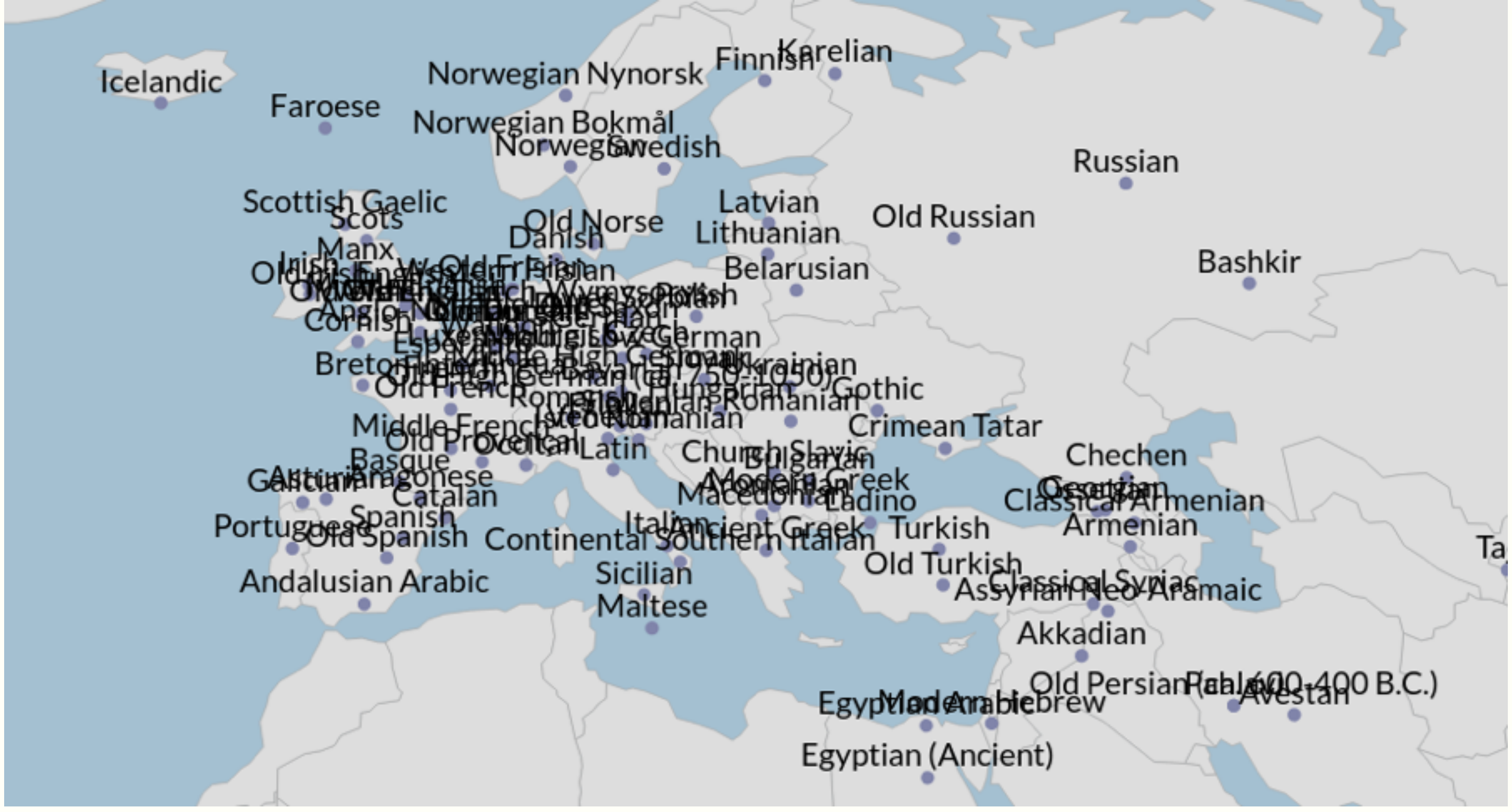
The language level will have a chord diagram of the largest influences the currently selected language has. Clicking on a specific relation will select it and bring the user to the relation level. The Sankey diagram will still remain here too, and will also have clickable relations. The reasoning behind adding a chord diagram is that it will allow the user to go to relations which do not directly involve the currently selected language. Otherwise he would have to go back to the map and search for the language for which he wants to know a relationship. This is clearly not ideal.

For the word level, we will simply add clickable edges on the word graph, so that the user can go to the language relation which holds that word relation.

The relation level will hold a Sankey diagram as did the language level, but for each language in the relation. It will easily allow the user to navigate to other similar relations as well as compare how important a language is to another. More importantly, it will also have sampled words from each language which are part of the relation. This will fulfill the initial motivations behind adding the relation level.

Improving how to display language names

We have already discussed how to display language names on the map quite a bit. We finally found a decent solution to this issue.



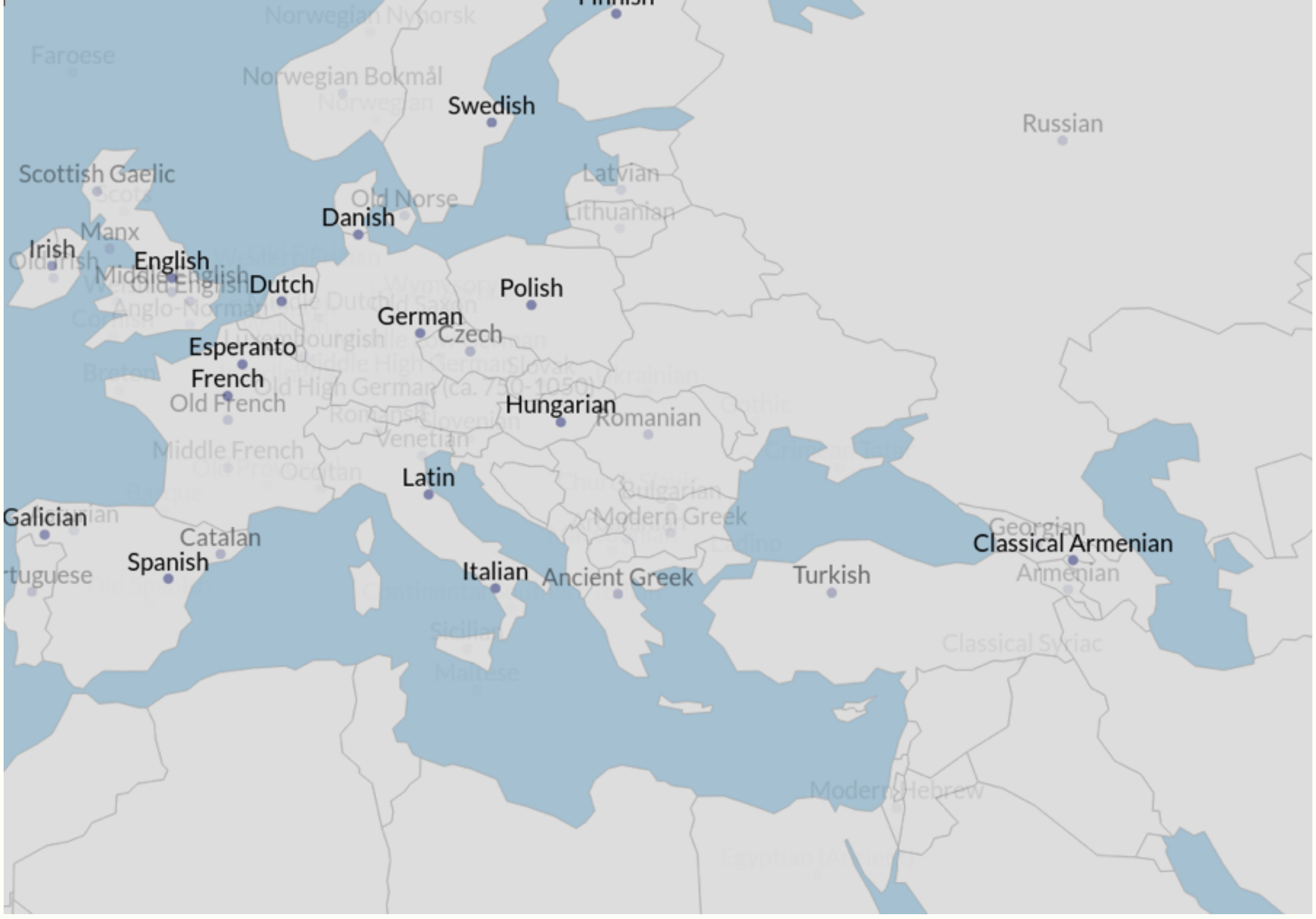
The first approach was to show them all, but the overlapping was absolutely unreadable as you can see on the screenshot above. There are workarounds, we could have spaced some of the language points better and fixed the minimum zoom level such that we ensure that there is enough space between the points. But the drawbacks of these modifications are huge. The user would need to drag the map around more to get from a point A to a point B (because he can't zoom out anymore). He would also spend more time searching for a language point if we move it away from where it is traditionally known to be. These are particularly bad downsides since, as we established in the "Refining our visualization's direction" article, the map is useful for quick switching and searching of languages. It would therefore be going directly against our goal.

The second approach is the exact opposite, hide everything until it is selected or hovered. Obviously, the problem here is that it becomes a guessing and memory game. As a user, you will have to imagine what language is assigned to a point, and if you check it by hovering, you will have to remember until you do not need it anymore. Again, it goes without saying that this is completely defeating the purpose of our map.

As most things, the solution is neither black or white. We have already established that the map should help the user quickly search for information he knows. What this means in practice is that commonly-known languages should always be shown. Now, for more "obscure" languages, the goal would be to only show them if we know that the user is looking for their name. The way we could detect this, is if the user is centered and/or zoomed on it. We kind of get the best of both worlds by doing this. Finally, we would also want to put the spotlight on the "larger" (from an etymological relations point of view) languages.

Of course, there is no ideal solution, the previous paragraph simply establishes what we want/need. Now it's all about compromising.

The more we thought about it the more this problem looks like how Google Maps deals with location names. They also have to deal with well-known searched locations, as well as more obscure places, and their solution is similar to what we describe above.

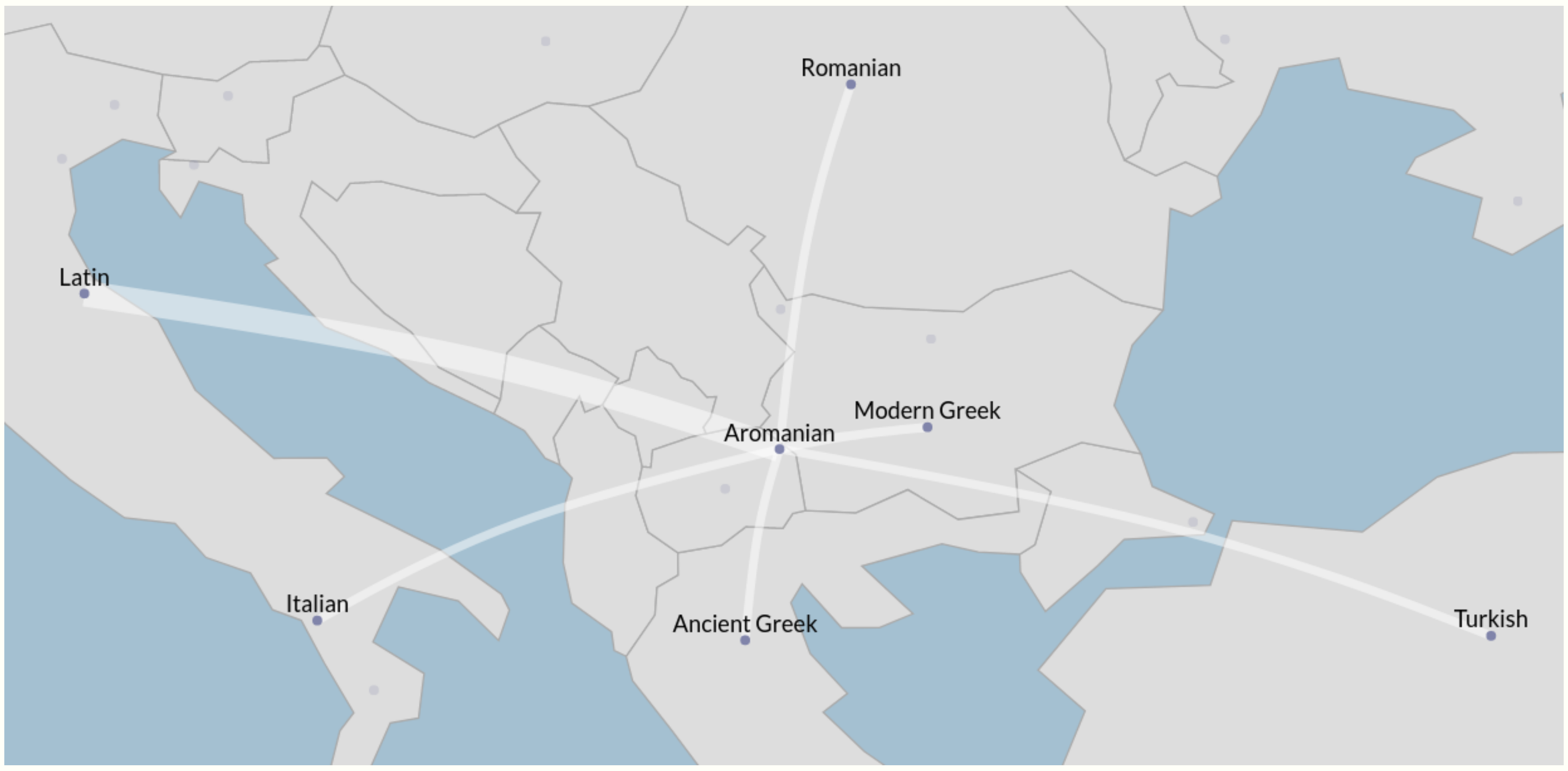


Unfortunately, we don't have a clear established hierarchy of well-known languages. However, we do have an established order of languages from their number of etymological relations. Our dataset is also quite biased towards well-known languages - this is simple the nature of Wikitionnary contributions. We can therefore use this as our metric to show language names. We have decided to use an opacity gradient instead of discrete steps to show the language names. This creates a nice smooth transition, and already gives the user the information if he's looking for it or doesn't stand out too much if he isn't looking for them.

As part of a feedback loop, we asked the initial users what they thought of this new version we propose. All agreed that it was much better than the two extreme solutions and only argued about how certain languages should be more/less prominent.

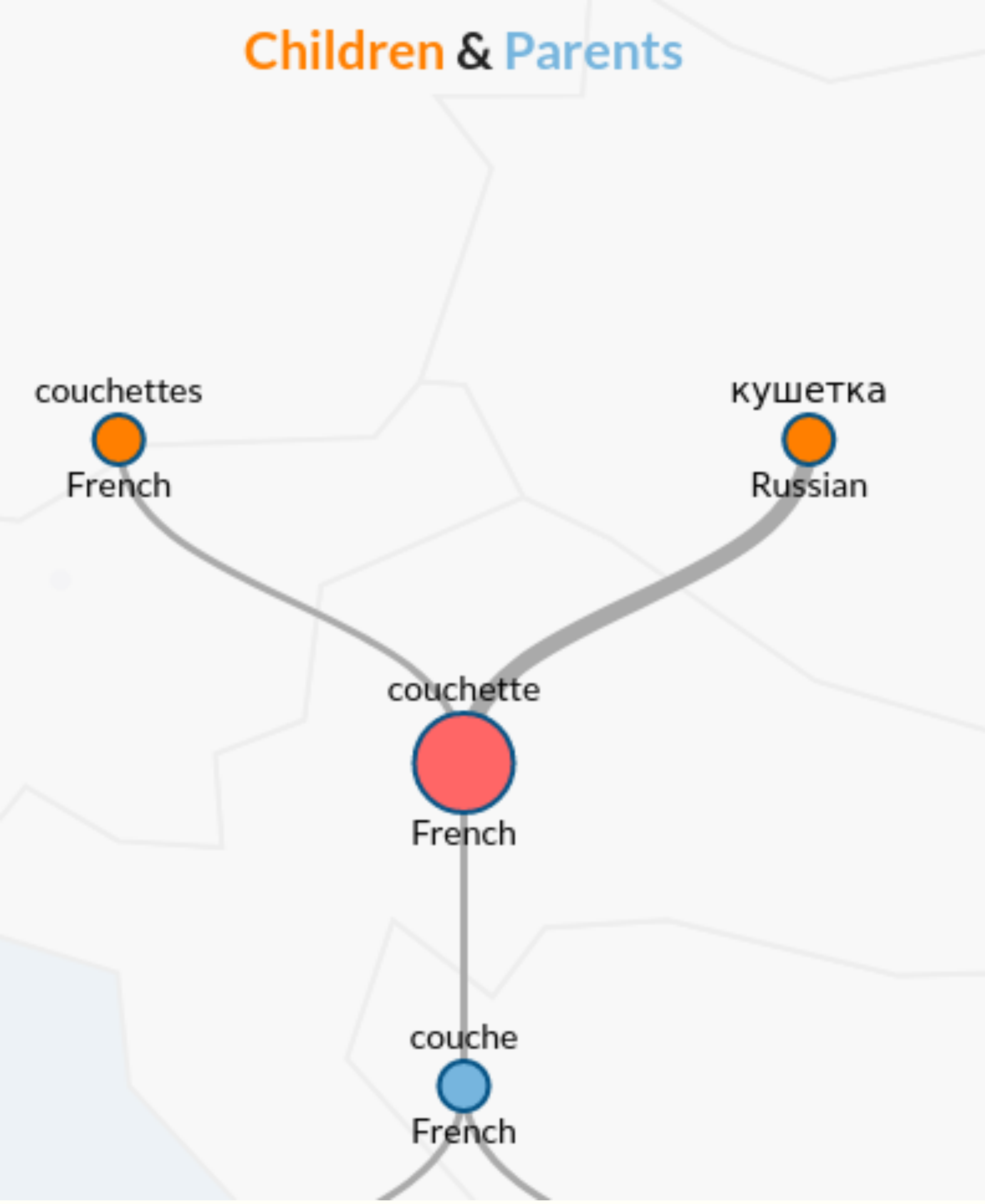
The beauty is in the details

In this process book, we mainly documented the large changes of our visualization. However, there were a lot of much smaller changes which lead to an overall better user experience which we did not mention at all. The user might not even notice these features, even though they improve his understanding and intimacy with our visualization. We are now going to quickly go over some of these.



The first one is at the language level. When selecting a particular language all other language names disappear and their points become slightly faded. This focuses the attention on what has just been selected.

On the side panel, every time the current selection appears, it will be in a red color. On the Sankey diagram, on the chord diagram, or even the word graph, the user's selection is always in red. This is the kind of consistency which makes the user rapidly familiar with the viz' without even noticing why.



The etymological word has it's legend in the title – the colors for children and parents. This avoids the additional clutter of having a small legend. Additionally, the edges in the tree which are larger are clickable because they are edges where the language switches.

Evaluation

As we unfortunately come to the end of this project, it is time for us to evaluate our work.

First of all, we would like to start with the fact that we are quite satisfied with the stories our visualization can show. From the beginning, we had always discussed what kind of words we would like to spotlight, and if it would actually be possible to see how wars, colonies, inventions, migration, etc. had influenced languages. In our story and the screencast we give two typical examples of words which clearly show how the Spanish colonization practically imported words to all of Europe. These are the aspects we wanted to highlight since day one, and we did it.

However, we do have to recognize that there are flaws. Most of which come from the fact that our dataset was far from perfect. As the previous examples can show, a single word relation can be the origin of an entire story. Therefore the slightest hole in the dataset might deny this possibility to our visualization entirely. The dataset skew towards Western languages is also unfortunate, there are plenty of interesting stories to tell through Asian etymology. We would also not have to use logarithmic scales if all languages were equally documented. For completeness, it would also have been nice to have the exact regions where languages were spoken.

During the final steps of this project, we also started to realise that if we had time information about these relationships we would be able to pinpoint certain historical influences. It would also serve as a confirmation of the stories we can currently tell.

There are two core things we learned through this project. The first is that your data will never be perfect, and that you should expect spending at least twice as much time on parsing/cleaning that what you initially think – never rely on your data being perfect for your visualization to work. The second one is that some of the best work goes unnoticed, everything that is intuitive, or that the end user does unconsciously is without a doubt the hardest results to achieve in an interactive visualization.

Peer assessment

Arnaud

Preparation My teammates were prepared at each meeting, with their task of the

week done.
Contribution We each did an equal part of the project, and when a dilemma occurred (for instance, a choice of color), we gave each other feedback on what we thought was best.
Respect for others' ideas There was some discussion during the topic selection for this project, both of my teammates offered their own ideas, while giving suggestions on what could be improved.
Flexibility There was a slight disagreement concerning the selection of the topic, Teo wanted to do a viz about neural nets, but Nicolas and I preferred the current topic. In the end Teo was convicted by our arguments, and generously agreed on our topic.

Nicolas
Preparation They were always prepared.
Contribution I do not feel like anyone was being counterproductive during the whole project.
Respect for others' ideas: Yes, every new feature was discussed as a team.
Flexibility I do not recall any major disagreements which were blocking our progress.

Teo
Preparation All members were ready and motivated.
Contribution Everyone one was focus moving forward and bringing the project alive equally.
Respect for others' ideas: Great ideas have been gathered during brainstorming session and then nicely discussed and selected.
Flexibility Partners were always flexible and open to new ideas.

Inspiration & references

1. Mledoze, Countries
2. Gerard de Melo, Etymological Wordnet: Tracing the History of Words
3. Glottolog, Languages coordinates
4. SIL ISO 639-3, Macrolanguage mappings
5. Yago-Naga, UWN / MENTA: Towards a Universal Multilingual Wordnet

6. Dave Liepmann, Tufte CSS