Translation Flow Analysis of Wikidata Properties

Thibaut Chamard

CPE Lyon, France
thibaut.chamard@cpe.fr

Abstract. Wikidata is a collaboratively edited knowledge base. More then five thousand properties and fifty millions items are translated over two hundreds and fifty languages. The translation process of these properties shows relevant correlations between languages. This report will take a look on the translation flow of Wikidata label properties and offer a short analysis and some interesting visualizations.

Keywords: Information flow · Ontology development · Wikidata.

1 Introduction

Wikidata is a multilingual, collaborative and structured knowledge base. Humans actively edit, update and remove the elements of this database over more than two hundreds languages. Some tools exist to visualize those revisions (WDProp [3] [2]). Nevertheless, it could be interesting to analyze the information flow hidden in the translation process of the Wikidata label properties. Using frequent pattern mining methods and visualizations, this report aims to show how to detect correlations or translation patterns between different languages.

2 Data Gathering

In order to understand the properties translation flow, we have to look at the history of each property. Every time that a property is edited, Wikidata saves a message and the new state of the property in a log. This information is available on a Wikidata SPARQL endpoint[1] and a REST API[2].

As WDProd[3] do, it is possible to parse the message to detect when and with what language a property is created, updated or removed. However, the message is sometime confusing, especially on the property creation. To lower the number of mistakes, it is better to work directly on the historical states of a property. The bad point of this method is that it needs more storage and more computing, especially for the properties which were edited a large number of times.

[1] https://query.wikidata.org/
[2] https://www.wikidata.org/w/api.php
[3] https://tools.wmflabs.org/wdprop/index.html

3 Data Transformation and Filtering

3.1 Parsing

By parsing the revisions of each property, a database[4] given the property creation time of each property was created. This database has three fields, the property name, the timestamps, and the language.

In order to limit the computing time, the properties, with more than five thousands modifications, were discarded from the working scope. For example some properties have been modified by a bot (e.g. Taxon Name[5]) and the number of revision is just too important to be workable.

3.2 Volumetry

To date, 5193 properties exist on Wikidata. Our scope has 5126 properties and 234 different languages. **Table 1** shows the classical statistical tool and **Fig. 1** shows the distribution of this dataset. It is interesting to see that more than the half of the properties are translated into at least 15 languages.

Table 1. Statistical information of the number of translation per property

Statistical tool	Value
count	5126.000000
mean	21.513071
std	20.140921
min	2.000000
25%	8.000000
50%	15.000000
75%	27.000000
max	129

4 Data Analysis and Visualization

4.1 Frequent pattern mining

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with an important frequency The study of frequent patterns is important in order to discover interesting, unexpected and useful patterns in databases. The python package, orange3-associate[6], has been very useful for this work.

[4] https://github.com/chamard/wikidata_translation/blob/master/data/all_label_creations.csv
[5] https://www.wikidata.org/wiki/Property:P225
[6] https://orange3-associate.readthedocs.io/en/latest/scripting.html

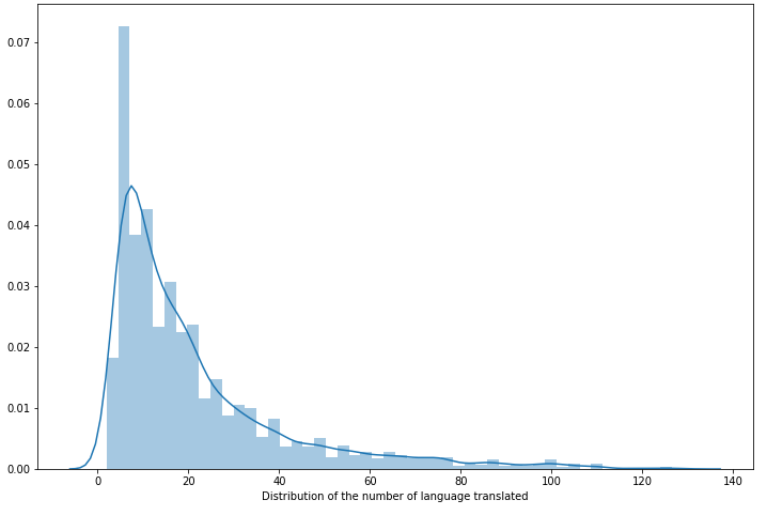

Fig. 1. Distribution of the number of translation per Wikidata property

Frequent itemsets: Using the Apriori algorithm, we detected the most frequent itemsets summarized in **Table 3**. It is interesting to see that all the properties are available in English. Although, the **Table 2** shows that the large majority of the properties were created in English.

Table 2. Language of the properties at its creation

Language	count	percentage
English	4503	89.59%
French	288	5.73%
German	134	2.67%
Dutch	46	0.92%
Italian	31	0.62%
Chinese	29	0.58%

Association rules: Using the FP-growth frequent pattern mining algorithm [4] implemented by Orange3-associate, it is possible to detect the most frequent association rules. This algorithm works well on the associations with an important confidence (Defined as itemset_support / antecedent_support), but it is not adapted to analyze rare associations. Indeed the computing time is very long if

Table 3. Most frequent itemsets

Number of items	Most frequent itemsets	Support
1	English	100%
	Arabic	99.1%
	Ukrainian	97.3%
	French	97.3%
	Dutch	75.7%
	Spanish	69.2%
	Macedonian	68.6%
	German	61.2%
2 (English pairs removed)	Arabic - French	96.6%
	Arabic - Ukrainian	96.6%
	Ukrainian - French	95.0%
	Ukrainian - Dutch	75.3%
	French - Dutch	74.9%
3 (English triples removed)	Arabic - Ukrainian - French	94.3%
	Arabic - Ukrainian - Dutch	75.1%
	Ukrainian - French - Dutch	74.6%
	Arabic - French - Dutch	74.5%
	Arabic - Ukrainian - Spanish	68.1%

you want to analyze less frequent languages. However, let's compare the association of English, Spanish and Catalan in **Table 4** which can also be represented like on **Fig 2**. It is interesting to see that even if Spanish and Catalan have a strong correlation, it is still two independent languages. In fact, only 75.8% of Spanish properties are translated into Catalan and 89.3% of Catalan properties are translated into Spanish.

Table 4. English - Spanish - Catalan association rules for the Wikidata properties

Itemset languages	antecedent languages	Confidence
Spanish - Catalonian	English	100%
Catalonian	Spanish	89.3%
Spanish	Catalonian	75.8%
English	Spanish - Catalonian	52.5%

It is also interesting to take a look at the two first languages translated in **Table 5**. Indeed we can see that the French community is very active in the creation of the new properties. Indeed, 27.72% of the time, the French is the first language translated when a property is created in English. We can say the same thing about German. Only 61.2% of the properties are translated into German but 11.38% of the time, German was the language translated right after the property creation in English.

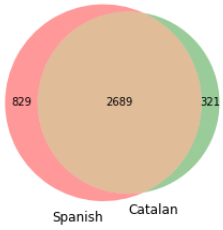

Fig. 2. Venn graph showing the intersection between the Spanish and Catalan translations of the Wikidata properties

Table 5. Association rules for the 2 first translations of the properties at its creation

First language	Second language	count	percentage
English	French	1393	27.72%
	German	572	11.38%
	Russian	423	8.42%
	Dutch	347	6.90%
French	English	270	5.37%
English	173	173	3.44%
	164	164	3.26%

4.2 Network Visualization

Sometimes, it is easier to understand the information by visualizing it, and it is particularly true for the study of graphs. On this part, we assume that the translation flow follows a temporal pattern. For every properties, a language is linked to the followed and the following language in the temporal order of creation. The visualization of this graph allows to understand the correlation between several languages without having any knowledge in linguistic.

For example, the **Fig. 3** shows a strong link between Tagalog and Kapampangan. Those two languages are actually two Philippians languages with 2 and 22 millions native speaker.

We can see, another interesting example on **Fig. 4** shows the relationships between the different kind of Chinese. To get this visualization, I removed all the languages between two Chinese languages in order to get a relevant flow of information. We can see that Chinese have is a complex language with a traditional Chinese, two simplified Chinese and a Chinese for different region (China, Macao, Singapore, Taiwan and Malaysia). We can clearly imagine the following pattern:

- 1. Set in simplified Chinese
- 2. Translated into Traditional Chinese (Navy Blue strings)
- 3. Translated into one regional Chinese (Orange strings)

– 4. Translated into the other regional Chinese (Pink strings)

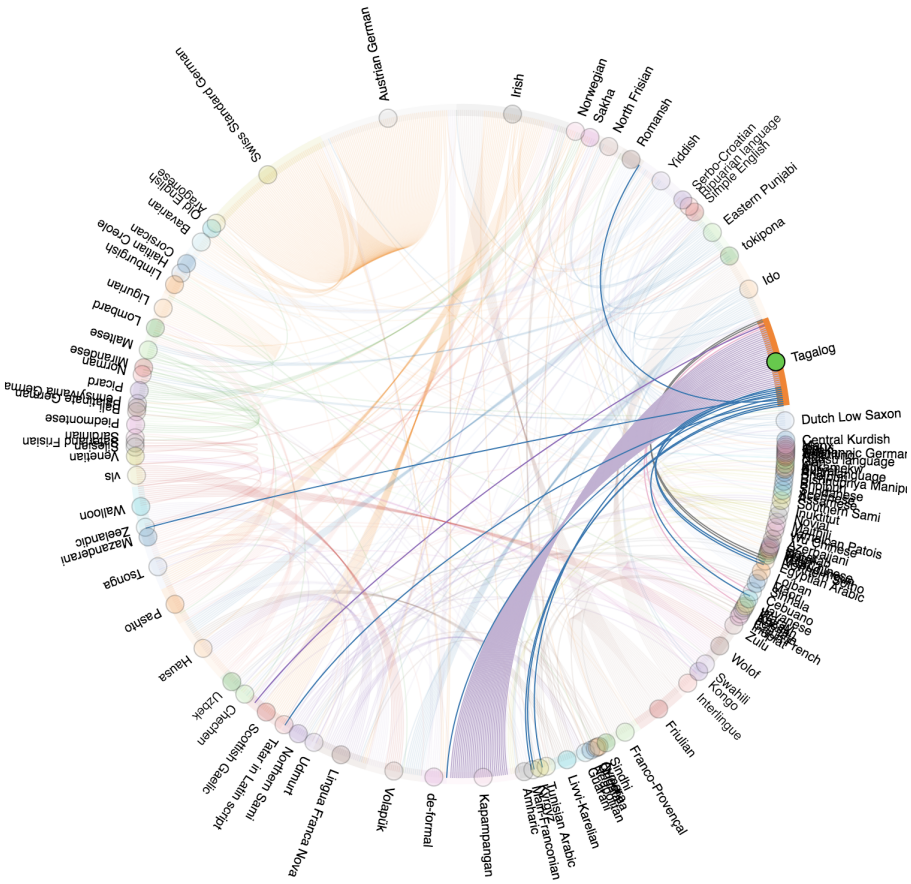

Fig. 3. Graph showing the links between the 100 least important languages used on the Wikidata properties

5 Conclusion and Discussion

The flow of information is a complex theory and the study of the data available on Wikidata offers a better understanding of the phenomenon of translation. In fact, throw different method of pattern mining or visualization, it is possible to show the relationship between several languages.

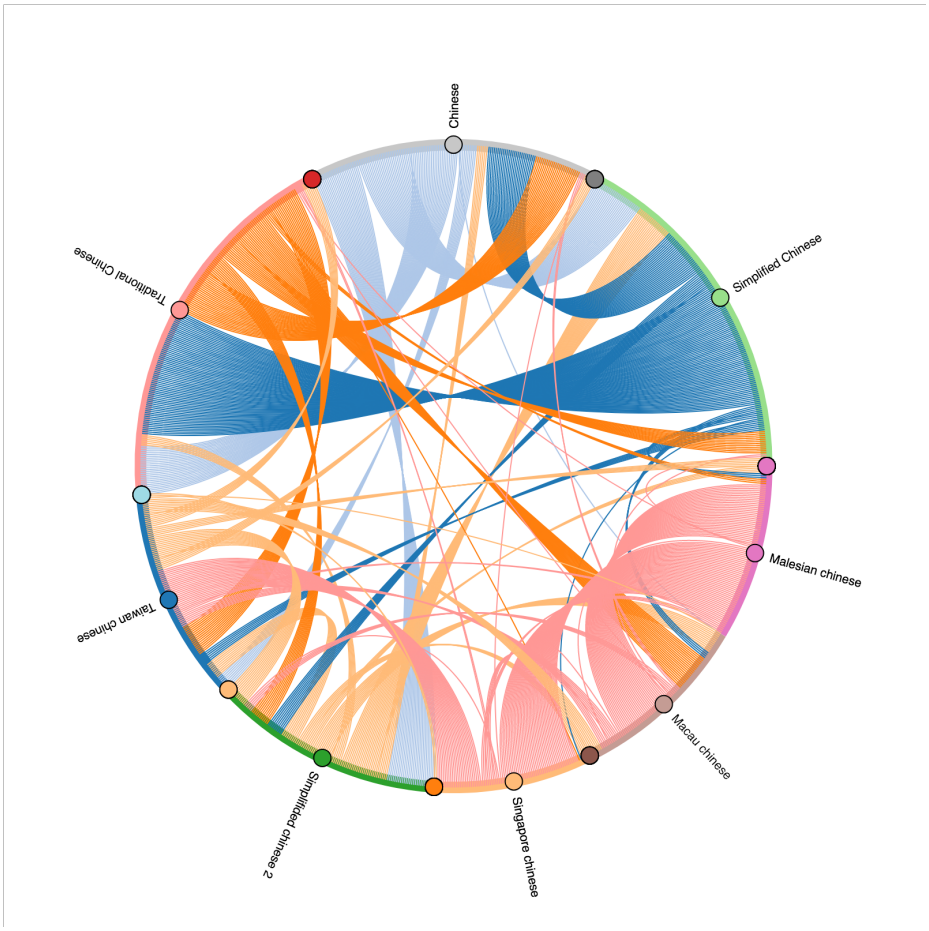

Fig. 4. Graph showing the links between all the different kind of Chinese used on the Wikidata properties

This report was focused on the analysis on the translation of the Wikidata label properties, but it would also be interesting to do the same work on the description or the aliases. Indeed, the translation of the description could follow a different pattern and the analysis of the aliases might be very complex because it is a list of elements. Besides, It could be possible to work on the items, rather than the properties. With almost 50 million items the analysis might offer interesting results. However, the issue will be to get the data. Indeed, the way I gather the information is not really scalable. Maybe, it is possible to find a way to parse the information related to the translations directly with SPARQL.

Moreover, I only worked on the creation process. It could be interesting to study other actions like the updates or the vandalisms. Nevertheless, all the code of my work is available on my Github [7]. The method and the code is widely reusable and improvable. It is a good sample of what can be done. Even if Orange3 is a good package, it could be interesting to develop a better way to get the association rules. Indeed the method is not good if you want to study the association rules of the rarest languages.

References

1. Erik Borra, David Laniado, Esther Weltevrede, Michele Mauri, Giovanni Magni, Tommaso Venturini, Paolo Ciuccarelli, Richard Rogers and Andreas Kaltenbrunner: A Platform for Visually Exploring the Development of Wikipedia Articles
2. John Samuel: Analyzing and Visualizing Translation Patterns of Wikidata Properties. 2018
3. John Samuel: Towards Understanding and Improving Multilingual Collaborative Ontology Development in Wikidata, 2018
4. J. Han, J. Pei, Y. Yin, R. Mao: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. 2004. https://www.cs.sfu.ca/ jpei/publications/dami03_fpgrowth.pdf

[7] https://github.com/chamard/wikidata_translation