Utrecht University

The Effect of Space-Language Bias on Toponym Co-occurence Derived Networks

MSc Applied Data Science

Brecht Nijman
6683479

Project Supervisor	Dr. Evert Meijers
Second Examiner	Dr. Carolina Castaldi
Daily Supervisor	Tongjing Wang

July 8, 2022

Contents

Abstract

Toponym co-occurrence analysis on unstructured sources has been suggested as a possible method for obtaining data for the study of urban networks. This method is particularly beneficial for modelling international relations and the relations between smaller places. However, it also suffers from potentially introducing space-language bias into the networks. This paper creates a network of 151 European cities derived from English and French versions of Wikipedia using toponym co-occurrence analysis. City-pairs in the English and French language sphere tend to be over-represented in the respective data sources compared to the patterns expected from gravity modelling. Nevertheless both of the resulting networks fit expected patterns, showing the applicability of toponym co-occurrence analysis.

Keywords — urban networks, space-language bias, wikidata, toponym co-occurrence

1 Introduction

One of the tenets of urban network science is that a city can only ever be as well understood as its relation to those cities it is connected to. Frustratingly this understanding is hampered by a lack of data about these relationships. While there is of course data about individual cities, this is often not directly comparable as it is collected by different national or local bodies. Furthermore, their relationship to each other remains under-documented. Over 25 years ago Short, *et al.* (1996) called this lack of network level data the 'dirty little secret of world cities research', and this lack continues to be felt. A number of potential solution that has been suggested for this is that of toponym co-occurrence analysis.

Toponym co-occurrence analysis aims to use techniques developed from Natural Language Processing (NLP) in order to make use of the vast amounts of unstructured data that exists for the understanding of city networks. It relies on two important concepts. The first is that of semantic relatedness, the idea that words which co-occur together more often than expected can tell us something about the meaning of both words (Baker, 2016). The second is Tobler's often-cited first law of geography 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). The combination of these ideas is used with the idea that place names co-occurring more often reflects a stronger relationship between these two places in real life. The use of this method greatly expands the potentially available data for urban network analysis.

This method is particularly promising for its ability to identify relations between smaller places, even internationally (Meijers and Peris, 2019). Since it does not require the places or the relationships to be explicitly monitored. However, there are also important limitations to consider. Most important for this thesis is that semantic collocation, which toponym co-occurrence is a type of, will generally reproduce the biases or point of view of the corpus (the textual input) it is based on (Baker et al., 2008; Garg et al., 2018). Furthermore, reporting these relationships outside of their original context can obscure these biases. When taking this phenomenon into account, it can be used to identify such biases or cultural perspectives. In the context of toponym co-occurrence specifically Cooper, *et al* (2015) point to the potential for creating 'spatial narratives' which focus on the representation of place. Such methods often place the perception of place central, rather than trying to extract physical relationships between places from the text. Hu, *et al* (2017) refer to the fact that both these types of information, the physical, but also those which exist 'only in the perception of people' (p. 2429) are extracted as a benefit of the method. Their approach to differentiate these two types of relationships is mainly focused on labelling co-occurrences as one or the other. This approach offers much insight into how cultural relationships differentiate from physical ones in the text but is limited in its ability to take the effect of cultural perspective on the depiction of physical urban relationships into account. The differentiation of these relationships is particularly difficult.

Ignoring this phenomenon in toponym co-occurrence could be harmful to the quality of the resulting networks and the conclusions built thereon. Furthermore, it accidentally compound another issue in the field. Much ink has been spilled on the impact of *Anglocentric* scholarship and how this can create a hegemonic point of view (Hassink et al., 2019; Van Meeteren 2019). This issue is not limited to the actual scholarship but also extends to the data that this scholarship is built on. If toponym

co-occurrence derived networks are exclusively built on English sources, they might unwittingly reproduce an Anglocentric worldview. Hecht and Gergle (2009) have developed a method for identifying what they have named 'self-focus bias' in community maintained repositories such as Wikipedia. They identify self-focus as the overrepresentation of the 'home' region, that is the region where the language of the corpus is primary or dominant, in the corpus. They do not measure occurrences rather they measure how often a georeferenced article was linked to. They find that out of the 15 language editions of Wikipedia they cover only three do not have a home region as the one with the most links. English and French were found to be among the languages with the highest self-focus ratio (first and fourth respectively). There are some limitations to their measure. First of all, there are some regions for which it is more or less accurate to their physical position in the world to be the most referred to. Second, only a relatively small number of pages on Wikipedia is georeferenced. As a result, this bias does not necessarily translate directly in toponym co-occurrences.

Alternatively, Salvini and Fabriki (2016) find almost opposite results in the discussion of their world city network based on English Wikipedia. They define a co-occurrence of two places as the appearance of a link to the Wikipedia article about each place on the same Wikipedia article. This has some limitations considering Wikipedia Manual of Style's discouragement of linking to pages about well-known places and to linking to the same page more than once in one article ("MOS:OVERLINK", 2022; "MOS:REPEATLINK", 2022). Salvini and Fabriki use the correlation of the data extracted from English Wikipedia and data from French, German and Italian Wikipedias. They find the strong correlation between these to be an indication that a space-language bias might not exist. Considering the geographic scope of their research and the prominence of both European and American cities in their network, it might be worthwhile to see if this correlation holds when compared with language editions with their home regions outside of Europe or America.

Thus the aim of this paper is to build on this research regarding the effect of language on toponym co-occurrence analysis, while also making use of the benefits of this method to create a usable network for the analysis of urban networks in Europe. If the strength of the connection between two cities is considered to be an objective fact, it should not matter what language one speaks as to how related two places are. However, I hypothesize that this will not be reflected in the co-occurrences, and that co-occurrences involving cities where the input language of the analysis is spoken will appear more often than those that do not involve such cities. This is in line with the self-focus bias found by Hecht and Gergle 2009 regarding page links, as well as the information asymmetry in Wikipedia which continues to be present (Roy et al., 2021). The analysis will be done in the form of a case study using the French and English Wikipedia. Both language editions will form a large enough corpus for analysis, and are local to the geographic area covered by the network. The use of Wikipedia, ensures that both language corpora are created under similar circumstances so as not to accidentally measure differences in genre or medium. The answering of this issue will be done as follows. First, toponym co-occurrences will be extracted for both languages. Next, those co-occurrences must be transformed in such a way to allow for comparison between the French and English co-occurrences. Then the general pattern of the co-occurrences in both models will be described, followed by an attempt at the quantification of the effect of the choice of input language on the resulting co-occurrences.

2 Data

2.1 Geographical Data

This project aims to create a city network for Europe using toponym co-occurrence analysis. This geographical scope allows for the exploration of the relationships of cities internationally, as well as domestically, which is often hampered by a lack of supranational data. The list of cities included are all those identified in the the ESPON *Study on Urban Functions* (2007) report as having a population of 300,000 or more. The list of 151 included cities can be viewed in Appendix B. 'Cities' in this context refers more precisely to 'Morphological Urban Areas' (or MUAs). With some exceptions, an MUA describes any municipality or group of municipalities with a population density of 650 people/km^2 and at least 20,000 inhabitants. Contiguous municipalities meeting this threshold form one MUA. Some

municipalities which do not meet the threshold may be included in an MUA if they are enclosed by municipalities that do (IGEAT, 2007). Generally, the MUAs in the ESPON data set are named after a central municipality or group of municipalities with a commonly shared name (e.g. 'London'). However, in some cases the name is a combination of multiple place names (e.g. 'Essen-Oberhausen'). This has implications for toponym co-occurrence analysis which will be further addressed in Section 3.1.

The data from the ESPON *Study on Urban Functions* (2007) was combined with the Euro Global Map (EGM) in order to obtain coordinates for each city.[1] The EGM dataset combines the data from various national mapping organisations within Europe. The two datasets were joined based on city name and country code. This allowed for the correct joining of 135 cities out of the dataset, and the incorrect inclusion of Leeds, Kent and Bremen, Geisa two villages which matched the country and toponym pattern of other cities in the list. These two were manually excluded. The majority of unmatched cities fell into the three following categories:

1. Name variants and translations ('Brussels' – 'Brussel')
2. Special characters ('Plovdiv' – 'Pl?vdiv')
3. Double names ('Essen-Oberhausen')

The first two issues affected so few cities that these were matched manually. The third issue was solved by matching to the first city in the pair. Finally two cities remained unmatched. Belfast, which had a different country code in the EGM data than the ESPON report, and Wuppertal which was not included at all in the EGM data. The coordinate for Wuppertal was added from OpenStreetMap data through QGIS.

The toponyms used in these datasets are either the local name or a transliteration of it, when the local language is not written in the Latin alphabet. These are often not the same names as those used in French or English. A French and English list of toponyms was compiled based on those names conforming to Wikipedia naming guidelines.[2] For finding co-occurrences, the assumption has been made that all Wikipedia articles follow the naming guidelines. The according toponyms for each language were added to the data set. Finally in order to allow for the measuring of a language effect on co-occurrence dummies were added for whether a city has English or French as one of its majority spoken languages.

2.2 Wikidata

Table 1: Wikidata Exploration

	English Wikipedia	**French Wikipedia**
n articles	6,488,754[a]	2,229,050[b]
size of dump (GB)	20.74 GB	5.65 GB
size extracted articles (GB)	17.17 GB	5.66 GB
n selected articles	509,894	274,639
mean article length (words)	1,142	1,042
size selected articles (GB)	3.61 GB	1.79 GB

[a] As of 23 April 2022 ("WP:SIZEWP", 2022)

[b] As of 22 April 2022 ("WP:STATS", 2022)

The contents of both language editions of Wikipedia used for this thesis were obtained from the Wikimedia dumps of April 20, 2022.[3] These come in a compressed XML format. In order to clean the

[1] Terms of license available at: https://www.mapsforeurope.org/licence

[2] Available at: https://en.wikipedia.org/wiki/Wikipedia:Naming_conventions_%28geographic_names%29#Alternative_names

[3] Available at https://dumps.wikimedia.org/enwiki/20220420/ and https://dumps.wikimedia.org/frwiki/20220420/

dumps WikiExtractor was used to extract the main body of each article in a clean document format (Attardi, 2015). Minor cleaning was done in the form of replacing special characters such as 'ł', 'ñ', and ''é with 'l', 'n' and 'e'. Other common preprocessing steps such as tokenisation, lemmatisation, or even removing upper case were not performed since these are likely to interfere with the identification of toponyms. For instance, there are multiple multi-token toponyms (e.g. 'The Hague') in the city data which could never match a single token or lemma. Capitalisation can serve to differentiate toponyms from common nouns or adjectives, 'Grenade' (Granada) and *grenade* (grenade) come to mind for French. Furthermore, in order to reduce the size of the data, a selection was made to only use articles with at least two toponyms. In this case the list of toponyms was split along individual city names, meaning that places such as 'Essen' and 'Oberhausen' were matched separately. No attention was paid to window-size at this time. The intention at this time was not to identify actual co-occurrences but merely to discard articles which do not contain any. Finally, some articles remained which consisted of a title only, these were discarded. The resulting data is about twenty to thirty percent the size of the original files (Table 1), and consists of 509,894 English Wikipedia articles with an average length of 1,142 words, and 274,639 French Wikipedia articles with an average length of 1,042 words.

3 Methods

3.1 Toponym Co-occurrences

Table 2: Example of an adjacency matrix with k toponyms.

	toponym 1	toponym 2	toponym 3	…	toponym k
toponym 1	0	0	0	…	0
toponym 2	0	0	0	…	0
toponym 3	0	0	0	…	0
…	…	…	…	…	…
toponym k	0	0	0	…	0

Co-occurrences were initially recorded in an adjacency matrix. In an adjacency matrix each toponym (node) is represented by a row and a column. A co-occurrence of two toponyms is recorded at the intersection of their row and column. So in the case of Table 2, if toponym 1 and toponym 3 co-occur the value at the intersection of row(toponym 1), column(toponym 3) and at the intersection of row(toponym 3), column(toponym 1) should be increased by 1. An empty adjacency matrix was created for both language instances of Wikipedia. Co-occurrences in articles were matched within a window-size of one paragraph. Previous research has found that paragraphs on Wikipedia are uniquely independent (Hecht and Raubal, 2008). They are often written in such a way that they can be read independently from the rest of the article they appear in. These characteristics makes the paragraph suitable as its own units of analysis. A paragraph makes for a rather large window for more general analysis of semantic relatedness. However, toponyms are relatively rare words in a corpus and thus require on average a greater distance to co-occur. Furthermore, distance decay has been observed to be less strong between toponyms in text than it is between physical places (Liu et al., 2014, Hu et al., 2017). In combination with the paragraph as a coherent unit, this makes the paragraph a useful unit of analysis.

In order to identify toponyms for co-occurrence a couple of steps must first be taken. A number of the MUAs in the data set consist of more than one city, for example 'Bochum-Herne'. Matching the entire name as listed will return very few, perhaps even no, results since it is not actually the name of a single place. Furthermore, the collective entity 'Bochum-Herne' has no guideline name on either Wikipedia. Thus instances of 'Bochum' and 'Herne' must be identified separately and then matched to the same occurrence. A dictionary was created for this purpose by splitting any toponym with a '-' in it, and subsequently having each name as well as the paired name (should it occur) refer to the same MUA (See Appendix C). The list of keys[4] in this dictionary is then used as toponyms to be found in each article, these will subsequently be referred to as the variant toponyms. The variant toponyms

[4]The keys are the left-hand values in the dictionary.

are matched by regular expression,[5] making sure that they do not form the subsection of another word (e.g. 'Bari' in 'Baritone').

Each variant toponym found within the window is added to a list and then matched back to its official form with the dictionary. This list of toponyms is then matched into city-pairs with the adjacency matrix being updated by 1 for each city-pair which occurred in the window. This means that only one co-occurrence is counted per paragraph, even if a co-occurrence appears more than once. This whole process is done for both language versions, until each article has been iterated over. The resulting values reflect the number of paragraphs that have been found to have a co-occurrence of a given city-pair. Considering the different size of the English and French Wikipedia, direct comparison of these values is not possible. To facilitate this, adjusted co-occurrences were calculated according to Equation 1:

$$NCo(CP_L) = \frac{Co(CP_L)}{Par_L} \cdot 1000 \tag{1}$$

where NCo denotes the normalized co-occurrence, CP_L denotes a city-pair in language L, Co denotes the observed co-occurrence, and Par the total number of paragraphs in the corpus of that language (i.e. the maximum number of potential co-occurrences).

3.2 Modelling the networks

From the matrices two data sets are created, one of pairs and one of individual cities. The dataset of city-pairs includes their distance, co-occurrences and adjusted co-occurrences in both languages, as well as the following added variables:

- **Border**: The border variable is included in order to allow the gravity model to take the effect of national borders on co-occurrence into account.
- **Language spheres**: A French and English language sphere dummy have been included in the data set. A city-pair is identified as being in the French or English language sphere when one of the cities in the pair has either of these languages as (one of) its majority spoken language(s). Thus, the 'Paris–Rome' city-pair is in the French language sphere, the 'London–Madrid' city-pair in the English language sphere and the 'Paris–London' city-pair is in both. Language spheres are included to model the hypothesised effect of language on co-occurrences of places where the language is spoken.
- **Shared language sphere**: A stricter language sphere dummy which only marks a pair as being in either language sphere when both cities in the pair are within the same language sphere. So for instance 'Paris–Brussels' is in the French language sphere but 'Paris–Antwerp' is not.
- **Region**: There are four regional dummies North, West, South and Central East. A city-pair is marked as in this region when at least one city is in one of these regions as defined by EuroVoc.[6]

For the city-level data, the degree of each city is calculated as the sum of the co-occurrences a city appears in.[7] This is combined with the city data regarding population and location (See Section 2.1).

4 Results

4.1 Comparing the networks

Out of a possible 11,325 city-pairs 8,054 (71.1 %) occur in both networks. Out of the two, the network derived from English Wikipedia has a higher density[8] with 9,772 (86.3 %) city-pair co-occurrences

[5] Matching to the following pattern: `r'\b'+ toponym + r'\b'`

[6] EuroVoc is a body of the Publications Office of the European Union. Its definition of North, West, South and Central Eastern Europe is available at: https://eur-lex.europa.eu/browse/eurovoc.html?params=72#arrow_7206

[7] An introduction on degree centrality can be found in Chapter 7 of Newman (2018).

[8] Density is a network level measure dividing the number of actual ties in a network by the total number of potential ties (Newman, 2018).

compared to the French Wikipedia network in which 8,367 (73.9 %) city-pairs co-occur. While more co-occurrences appear in English it is worth to note that there are 313 co-occurrences which have been found only in the French Wikipedia. Mapping these networks can give us some additional insight. Figures 1a and 1b show adjusted co-occurrences (see Section 3.1 Eq. 1). The higher the adjusted co-occurrence the lighter the colour. As can be seen there are more high co-occurrence connections in France in the French Wikipedia derived model (Figure 1b) than the English Wikipedia one (Figure 1b). In fact out of the five highest co-occurrences in the French Wikipedia network all five involve at least one French city, and three of them involve only French cities (Table 3). The most common co-occurrences in the English Wikipedia network also mostly involve British cities. Four out of five city-pairs contain at least one British city, and two of those contain only British cities (Table 4). In fact only one city-pair is shared amongst the strongest five city-pair co-occurrences in both networks, namely 'Paris–London'. However, there are a number of pairs which are relatively close in size (no more than half or twice as large), such as 'Paris–Berlin', 'London–Berlin' and 'Paris–Rome'. These also happen to be all those pairs which do not fall entirely within a French or English speaking area. There is much less consensus between the models on the domestic relationships. For instance, the strongest relationship in the French Wikipedia network, 'Paris–Lyon', is more than 6 times stronger in this network than it is in the English one (Table 3). Among the strongest English Wikipedia ties 'London–Edinburgh' stands out which is almost 12 times stronger in the English Wikipedia network than it is in the French one (Table 3).

Table 3: Top 5 co-occurrences in the French Wikipedia network

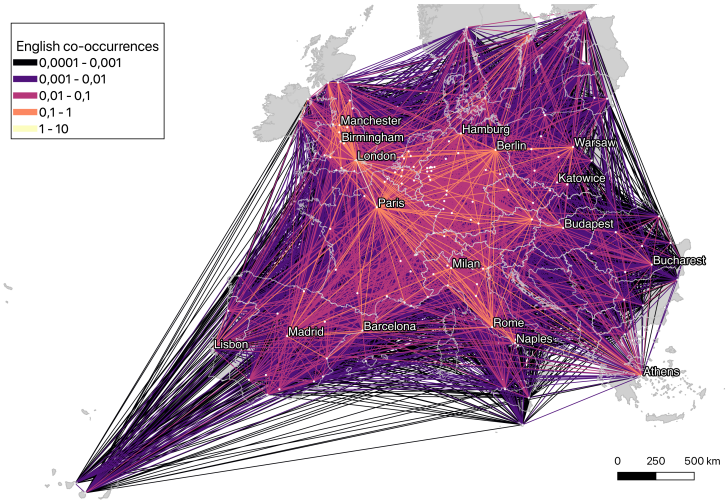
city A	city B	co-occurrence english	co-occurrence french	× larger in French
Paris	Lyon	0.343202	2.160905	6.296302
Paris	London	2.095814	1.787238	0.852766
Paris	Marseille	0.209351	1.245888	5.951180
Paris	Bordeaux	0.212513	1.132656	5.329812
Paris	Rome	0.653541	1.018424	1.558319

Table 4: Top 5 co-occurrences in the English Wikipedia network

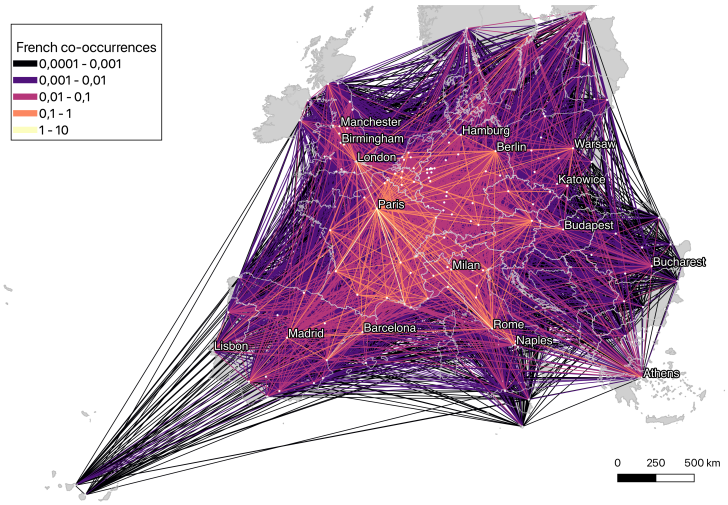
city A	city B	co-occurrence english	co-occurrence french	× larger in English
Paris	London	2.095814	1.787238	1.172655
London	Manchester	0.993389	0.177175	5.606819
London	Edinburgh	0.909553	0.076099	11.952271
Paris	Berlin	0.756827	0.786465	0.962315
London	Berlin	0.734694	0.413631	1.776207

This pattern of city-pairs from within the English or French language zone being much more strongly connected in the Wikipedia of their respective language, seems to hold relatively true throughout. Figure 2a shows the adjusted co-occurrences in both models plotted against each other. If a point appears above the diagonal it means that the corresponding city-pair has a higher co-occurrence in French Wikipedia than in English Wikipedia. Below the line it is the other way around. It can be seen that English city-pairs (city-pairs where both cities are in the English language sphere) trend towards the bottom, while French city-pairs tend towards the top. There is one point in neither language sphere that has a much stronger co-occurrence in French Wikipedia than in English. This is 'Valencia-Grenoble', and this difference is potentially due to a disambiguation issue which only occurs in French.[9] This separation is even more clearly observed at a city level. Figure 2b shows the adjusted degrees of the cities in the French and English Wikipedia derived models. Cities in the French language sphere are shown to have a higher adjusted degree in French while those in the English language sphere have

[9] 'Valencia' in French is called 'Valence'. However, there is also a French city named 'Valence' which is close to 'Grenoble' and thus potentially likely to co-occur.



(a) English Wikipedia Co-occurrences.

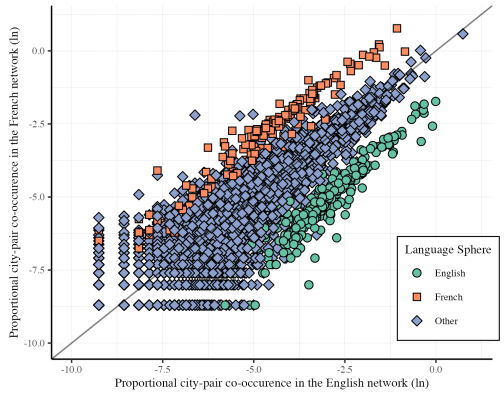


(b) French Wikipedia Co-occurrences.

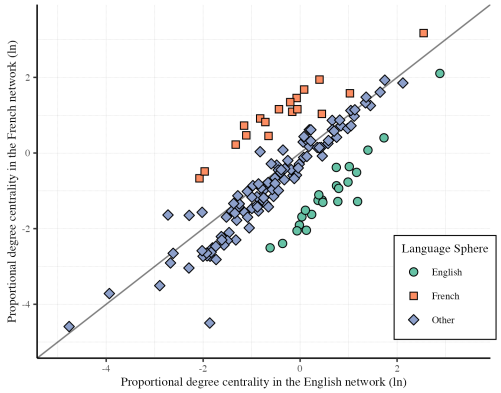
Figure 1: Observed adjusted co-occurrences in the European city network derived from English and French Wikipedia.

Table 5: Correlations between the properties of the French and English Wikipedia derived networks.

Property	Pearson's r
all shared city-pairs (co-occurrence) (N = 8054)	0.6414027
city-pairs (co-occurrence) (excl English and French pairs) (N = 7648)	0.9465477
cities (degree) (N = 151)	0.7363526
cities (degree) (excl. English and French pairs) (N=110)	0.9676624



(a) Scatter plot of the adjusted city-pair co-occurrences of the French and English city networks.



(b) Scatter plot of the adjusted degree centrality of each city in the French and English city networks.

Figure 2: Comparing French and English Wikipedia derived networks.

a higher degree in English. Cities which do not belong to either of these categories generally fall closer to the line of consensus, with the notable exception of Las Palmas.[10] Furthermore, it is interesting to note that while the correlation of either of the French and English properties (co-occurrence and degree) is not that strong ($r < .8$), this value is much increased when only considering the cities and city-pairs which are not within the French or English language sphere. Indicating a much greater level of consensus regarding these places (Table 5).

4.2 Gravity Model

Comparing the two networks to each other has shown some notable differences. In order to assess how either of these co-occurrence networks stand up compared to the expected pattern gravity modelling was used. Equation 2 shows the formula for the baseline gravity model:

$$C_{ab} = b_0 + b_1 \cdot Pop_a + b_2 \cdot Pop_b + b_3 \cdot Dist_{ab} \tag{2}$$

Where C_{ab} refers to the co-occurrence of cities A and B, Pop_a to the population of city A, Pop_b the population of city B and $Dist_{ab}$ to the distance between these cities. The gravity model was fit only on pairs occurring in both networks. Table 6 shows the results of four different models for each of the networks. The base model (model 1 and 4) fits the English Wikipedia co-occurrences better than the

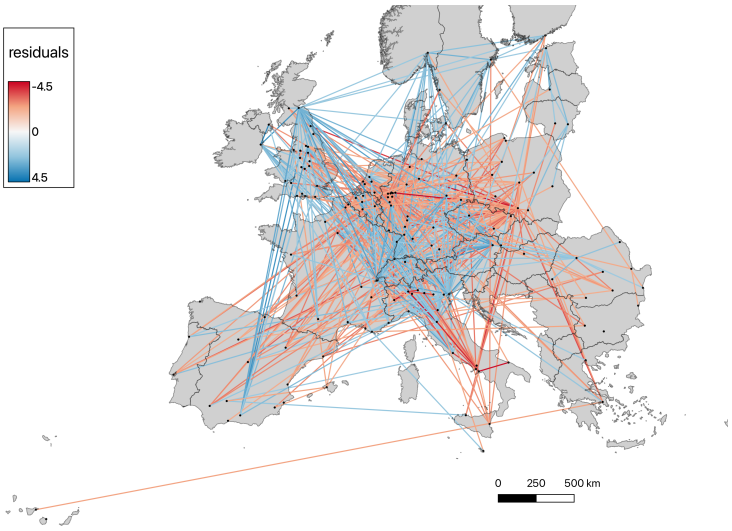
[10] On French Wikipedia the name guidelines dictates that Las Palmas should be written 'Las Palmas de Gran Canaria', while English Wikipedia simply uses 'Las Palmas'. The latter is more sensitive to false positives while the former might result in false negatives when the full name is not used.

French ones, with Adjusted R^2 of .54 and .48 respectively. The fit continues to be better for the English co-occurrences than the French ones with the extension of the model but the difference decreases. The first extension of the model (models 2 and 5) includes a dummy for whether a national border exists between the two cities. The presence of a national border is shown to have a negative effect on the co-occurrence of city-pairs on French and English Wikipedia. The last version of the model represented (models 3 and 6) here includes the dummies for the English and French language sphere. This is the best fitting model for both sets of co-occurrences. But the improvement of the Adjusted R^2 between the previous model and the current one is much greater for the French co-occurrence model. For both models the related language sphere is shown to have a significant positive relation on the co-occurrences.

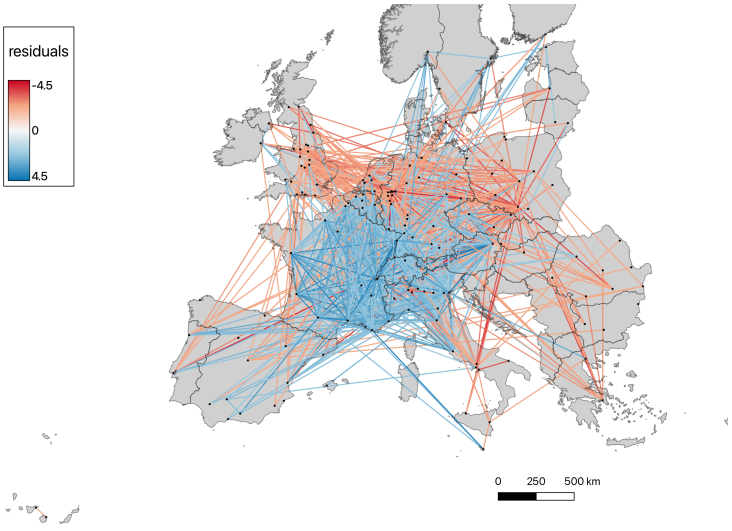
Mapping the residuals of the second model (model 2 and 4 in Table 6), the base gravity model with the border dummy included, reveals some distinct patterns. This reflects the patterns that would be revealed should space-language bias not be taken into account. Figures 3a and 3b show the residuals of the model fitted on co-occurrences in English and French Wikipedia respectively. Negative residuals (in red) indicate that the observed co-occurrence is smaller than would be expected from the gravity model. Positive residuals (in blue) indicate that the observed value is larger than expected. In both cases, though more so in the French case, the co-occurrences involving cities in the respective language sphere tend to be greater than expected. As a result the pattern in the very west of Europe is quite different. The pattern of residuals in Eastern Europe and Germany is much more similar. Co-occurrences in these regions, especially long ones tend to occur less than the models predict.

Table 6: Gravity model, language sphere influence and toponym co-occurrences.

	Dependent variable:					
	English Co-occurrence (ln)			French Co-occurrence (ln)		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	−3.374***	−4.842***	−4.922***	−3.978***	−5.016***	−6.079***
	(0.218)	(0.203)	(0.206)	(0.231)	(0.226)	(0.212)
Population A (ln)	0.899***	0.897***	0.893***	0.831***	0.829***	0.867***
	(0.019)	(0.018)	(0.018)	(0.021)	(0.020)	(0.018)
Population B (ln)	1.271***	1.296***	1.301***	1.230***	1.247***	1.303***
	(0.033)	(0.030)	(0.030)	(0.035)	(0.034)	(0.031)
Distance (ln)	−1.155***	−0.726***	−0.724***	−1.064***	−0.761***	−0.672***
	(0.018)	(0.020)	(0.020)	(0.019)	(0.022)	(0.020)
Border		−1.771***	−1.792***		−1.252***	−1.527***
		(0.045)	(0.046)		(0.050)	(0.048)
French language sphere			0.058**			0.890***
			(0.029)			(0.030)
English language sphere			0.249***			−0.538***
			(0.028)			(0.029)
N city-pairs	8,054	8,054	8,054	8,054	8,054	8,054
Adjusted R^2	0.534	0.608	0.612	0.473	0.510	0.583
F Statistic	3,080.889***	3,126.122***	2,117.323***	2,405.895***	2,095.932***	1,877.410***
Note:						*p<0.1; **p<0.05; ***p<0.01



(a) English Wikipedia (model 2).



(b) French Wikipedia (model 4).

Figure 3: Difference between predicted and observed co-occurrences in the European network derived from English and French Wikipedia (model 2 and 4).

5 Discussion

The results indicate that in this instance language choice has a significant effect on the outcome of toponym co-occurrence analysis (Table 6). They furthermore show that outside of the language sphere of either source there is a fairly strong consensus (Table 5). This is promising as it could indicate a potential method for correcting models for space-language bias. However the consensus might well be a shared cultural view between the French and English language communities. Further research is needed to confirm whether this. Particularly since the lower-than expected co-occurrences between Eastern European cities and Western Europe could be the real result of historical divides.

Another possible reason for the relatively low co-occurrences in Eastern Europe may be related to Named Entity Recognition (NER) and disambiguation issues. Contested histories result in places having multiple placenames, for example the many Polish cities which have alternate German placenames ('Sczcezin' – 'Stettin', 'Gdansk' – 'Danzig', 'Bydgoszcz' – 'Bromberg'). Places such as this are more likely to not to meet the toponym assumption of this paper. Namely, that all articles on Wikipedia follow the provided guidelines. Wikipedia aims to use place names by consensus, but if little consensus exists it is more likely to be subject to change which name should be used. With more frequent change it is more likely that some pages are out of date and not using the current name. Pages discussing historical events may refer to places by their historical name, even if guidelines suggest that the modern name should be used at least once this would not outweigh this effect.

Lastly, further insight into language specific patterns may be gained by labelling the co-occurrences. This would allow to understand whether the same connections are understood similarly in both language communities. Further research on disambiguation and the labelling of connections has been undertaken in the other two theses in this project, Dieder van Rijen's *Classifying and labeling the relationships between cities with high levels of co-occurrence on the English Wikipedia* and Kevin O'Driscoll's *Analysis of Toponym Co-occurrences on Social Media.*

6 Conclusion

This thesis used toponym co-occurrence analysis to create two city-network data sets covering 151 cities in Europe, which can be used for the study of these relationships. It derived these networks from French and English Wikipedia and compared them in order to assess the potential effects of space-language bias. In this regard, it has found that in the case of French and English Wikipedia the ties involving cities where these languages are spoken tend to be overestimated compared to the others. This indicates the existence of a space-language bias affecting toponym co-occurrence derived networks. Nevertheless, both French and English Wikipedia co-occurrences fit the gravity model quite well, and have a high level of consensus outside of their respective language spheres. This could potentially be used as a way to mitigate the aforementioned bias. However this would require additional insight in order to confirm that this pattern holds true.

References

Attardi, G. (2015). *Wikiextractor*. GitHub. https://github.com/attardi/wikiextractor

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, *21*(2), 139–164. https://doi.org/10.1075/ijcl.21.2.01bak

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306. https://doi.org/10.1177/0957926508088962

Cooper, D., Gregory, I. N., Hardie, A., & Rayson, P. (2015). Spatializing and Analyzing Digital Texts: Corpora, GIS and Places. In D. J. Bodenhamer, J. Corrigan, & T. M. Harris (Eds.), *Deep Maps and Spatial Narratives* (pp. 150–178). Indiana University Press.

EuroGeographics. (2022). EuroGlobalMap (EGM).

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Hassink, R., Gong, H., & Marques, P. (2019). Moving beyond Anglo-American economic geography. *International Journal of Urban Sciences*, *23*(2), 149–169. https://doi.org/10.1080/12265934.2018.1469426
_eprint: https://doi.org/10.1080/12265934.2018.1469426

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. *Proceedings of the 2009 International Conference on Communities and Technologies*, 11–19. https://doi.org/10.1145/1556460.1556463

Hecht, B., & Raubal, M. (2008). GeoSR: Geographically Explore Semantic Relations in World Knowledge. In L. Bernard, A. Friis-Christensen, & H. Pundt (Eds.), *The European Information Society: Taking Geoinformation Science One Step Further* (pp. 95–113). Springer. https://doi.org/10.1007/978-3-540-78946-8_6

Hu, Y., Ye, X., & Shaw, S.-L. (2017). Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science*, *31*(12), 2427–2451. https://doi.org/10.1080/13658816.2017.1367797

IGEAT. (2007). ESPON project 1.4.3 Study on Urban Functions.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing Relatedness by Toponym Co-Occurrences on Web Pages: Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *Transactions in GIS*, *18*(1), 89–107. https://doi.org/10.1111/tgis.12023

Meijers, E., & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. *International Journal of Urban Sciences*, *23*(2), 246–268. https://doi.org/10.1080/12265934.2018.1497526

MOS:OVERLINK. (2022). Retrieved June 30, 2022, from https://en.wikipedia.org/w/index.php?title=MOS:OVERLINK

MOS:REPEATLINK. (2022). Retrieved June 30, 2022, from https://en.wikipedia.org/w/index.php?title=MOS:REPEATLINK

Newman, M. (2018). *Networks*. Oxford University Press.

Roy, D., Bhatia, S., & Jain, P. (2021). Information asymmetry in Wikipedia across different languages: A statistical analysis. *Journal of the Association for Information Science and Technology*, *73*(3), 347–361. https://doi.org/10.1002/asi.24553
_eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24553

Salvini, M. M., & Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multirelational world city network. *Environment and Planning B: Planning and Design*, *43*(1), 228–248. https://doi.org/10.1177/0265813515603868

Short, J. R., Kim, Y., Kuus, M., & Wells, H. (1996). The Dirty Little Secret of World Cities Research: Data Problems in Comparative Analysis. *International Journal of Urban and Regional Research*, *20*(4), 697–717. https://doi.org/10.1111/j.1468-2427.1996.tb00343.x
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2427.1996.tb00343.x

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240. https://doi.org/10.2307/143141

van Meeteren, M. (2019). On geography's skewed transnationalization, anglophone hegemony, and qualified optimism toward an engaged pluralist future; A reply to Hassink, Gong and Marques. *International Journal of Urban Sciences*, *23*(2), 181–190. https://doi.org/10.1080/12265934.2018.1467273
_eprint: https://doi.org/10.1080/12265934.2018.1467273

WP:SIZEWP. (2022). Retrieved June 30, 2022, from https://web.archive.org/web/20220423181542/https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

WP:STATS. (2022). Retrieved June 30, 2022, from https://web.archive.org/web/20220422145209/https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Statistiques

Appendix A. Software Packages

Appendix A.1. QGIS

Used QGIS version 3.16.12-Hannover

Appendix A.2. Python

Used Python 3.10.4

packages:

- pandas 1.4.2
- wikiextractor 3.0.6
- numpy 1.22.3
- shapely 1.8.2
- geopandas 0.10.2

Appendix A.3. R

Used R version 4.1.1 (2021-08-10)
packages:

- stargazer 5.2.3
- tidyverse 1.3.1
- ggplot2 3.3.5

Appendix B. List of cities

Table 7: Cities included in the network.

	Toponym	CC	Pop (x 1000)		Toponym	CC	Pop (x 1000)
1	Paris	FR	9591	41	Lille	FR	953
2	London	UK	8256	42	Lodz	PL	919
3	Madrid	ES	4955	43	Marseille	FR	862
4	Berlin	DE	3776	44	Antwerp	BE	830
5	Milan	IT	3698	45	Bilbao	ES	822
6	Barcelona	ES	3659	46	Newcastle	UK	814
7	Athens	GR	3331	47	Krakow	PL	807
8	Rome	IT	2532	48	Bochum-Herne	DE	804
9	Birmingham	UK	2363	49	Thessaloniki	GR	777
10	Lisbon	PT	2315	50	Nuremberg	DE	769
11	Naples	IT	2308	51	Riga	LV	764
12	Katowice	PL	2279	52	Duisburg	DE	758
13	Manchester	UK	2207	53	Dortmund	DE	750
14	Hamburg	DE	2123	54	Hanover	DE	747
15	Budapest	HU	2123	55	Zürich	CH	718
16	Bucharest	RO	2064	56	Oslo	NO	712
17	Warsaw	PL	2004	57	Bremen	DE	709
18	Stuttgart	DE	1735	58	Dresden	DE	697
19	Vienna	AT	1674	59	Sheffield	UK	693
20	Munich	DE	1647	60	Palermo	IT	680
21	Brussels	BE	1498	61	Poznan	PL	679
22	Stockholm	SE	1479	62	Gelsenkirchen-Bottrop	DE	666
23	Frankfurt	DE	1462	63	Bordeaux	FR	652
24	Cologne	DE	1398	64	Wroclaw	PL	634
25	Copenhagen	DK	1360	65	Gothenburg	SE	627
26	Valencia	ES	1318	66	Zaragoza	ES	615
27	Turin	IT	1309	67	Genoa	IT	611
28	Glasgow	UK	1228	68	Catania	IT	602
29	Prague	CZ	1175	69	The Hague	NL	589
30	Lyon	FR	1175	70	Toulouse	FR	588
31	Sofia	BG	1174	71	Bristol	UK	568
32	Liverpool	UK	1170	72	Vilnius	LT	554
33	Porto	PT	1163	73	Saarbrücken	DE	552
34	Seville	ES	1082	74	Malaga	ES	543
35	Dublin	IE	1070	75	Nantes	FR	536
36	Helsinki	FI	1065	76	Leeds	UK	534
37	Amsterdam	NL	1052	77	Nottingham	UK	532
38	Rotterdam	NL	1025	78	Florence	IT	525
39	Düsseldorf	DE	1016	79	Gdansk	PL	519
40	Essen-Oberhausen	DE	986	80	Leipzig	DE	516

Table 8: Cities included in the network, cont'd.

	Toponym	CC	Pop (x 1000)
81	Mannheim	DE	508
82	Belfast	UK	501
83	Portsmouth	UK	500
84	Venice	IT	483
85	Edinburgh	UK	478
86	Murcia	ES	476
87	Nice	FR	472
88	Liège	BE	451
89	Bratislava	SK	444
90	Leicester	UK	442
91	Karlsruhe	DE	440
92	Bergamo	IT	438
93	Palma de Mallorca	ES	433
94	Bologna	IT	432
95	Bielefeld	DE	419
96	Rouen	FR	419
97	Strasbourg	FR	417
98	Tallinn	EE	416
99	Szczecin	PL	416
100	Grenoble	FR	415
101	Bari	IT	411
102	Toulon	FR	410
103	Brighton	UK	410
104	Darmstadt	DE	407
105	Wuppertal	DE	395
106	Utrecht	NL	390
107	Bournemouth	UK	390
108	Middlesbrough	UK	389
109	Geneva	CH	388
110	Bydgoszcz	PL	383
111	Basel	CH	381
112	Kaunas	LT	379
113	Brno	CZ	376
114	Southampton	UK	376
115	Lens	FR	374
116	Augsburg	DE	371
117	Padua	IT	370
118	Ostrava	CZ	365
119	Las Palmas	ES	365
120	Constanta	RO	364
121	Castellammare di Stabia-Torre Annunziata	IT	362
122	Stoke	UK	359
123	Santa Cruz de Tenerife	ES	357
124	Lublin	PL	354
125	Cardiff	UK	353
126	Iasi	RO	349
127	Plovdiv	BG	341
128	Bradford	UK	341
129	Alicante	ES	339
130	Cluj-Napoca	RO	332
131	Granada	ES	330
132	Timisoara	RO	328
133	Brescia	IT	327
134	Galati	RO	325
135	Montpellier	FR	323
136	Varna	BG	322
137	Verona	IT	320
138	Busto Arsizio	IT	320
139	Valladolid	ES	318
140	Eindhoven	NL	316
141	Charleroi	BE	314
142	Cordoba	ES	314
143	A Coruna	ES	311
144	Craiova	RO	311
145	Caserta	IT	308
146	Coventry	UK	308
147	Brasov	RO	307
148	Bonn	DE	306
149	Valletta	MT	301
150	Ghent	BE	300
151	Gdynia	PL	300

Source: ESPON 2007

Appendix C. Dictionary

```
{'Paris': 'Paris',
'London': 'London',
'Madrid': 'Madrid',
'Berlin': 'Berlin',
'Milan': 'Milan',
'Barcelona': 'Barcelona',
'Athens': 'Athens',
'Rome': 'Rome',
'Birmingham': 'Birmingham',
'Lisbon': 'Lisbon',
'Naples': 'Naples',
'Katowice': 'Katowice',
'Manchester': 'Manchester',
'Hamburg': 'Hamburg',
'Budapest': 'Budapest',
'Bucharest': 'Bucharest',
'Warsaw': 'Warsaw',
'Stuttgart': 'Stuttgart',
'Vienna': 'Vienna',
'Munich': 'Munich',
'Brussels': 'Brussels',
'Stockholm': 'Stockholm',
'Frankfurt': 'Frankfurt',
'Cologne': 'Cologne',
'Copenhagen': 'Copenhagen',
'Valencia': 'Valencia',
'Turin': 'Turin',
'Glasgow': 'Glasgow',
'Prague': 'Prague',
'Lyon': 'Lyon',
'Sofia': 'Sofia',
'Liverpool': 'Liverpool',
'Porto': 'Porto',
'Seville': 'Seville',
'Dublin': 'Dublin',
'Helsinki': 'Helsinki',
'Amsterdam': 'Amsterdam',
'Rotterdam': 'Rotterdam',
'Dusseldorf': 'Dusseldorf',
'Essen': 'Essen-Oberhausen',
'Oberhausen': 'Essen-Oberhausen',
'Essen-Oberhausen': 'Essen-Oberhausen',
'Lille': 'Lille',
'Lodz': 'Lodz',
'Marseille': 'Marseille',
'Antwerp': 'Antwerp',
'Bilbao': 'Bilbao',
'Newcastle': 'Newcastle',
'Krakow': 'Krakow',
'Bochum': 'Bochum-Herne',
'Herne': 'Bochum-Herne',
'Bochum-Herne': 'Bochum-Herne',
'Thessaloniki': 'Thessaloniki',
'Nuremberg': 'Nuremberg',
'Riga': 'Riga',
'Duisburg': 'Duisburg',
'Dortmund': 'Dortmund',
'Hanover': 'Hanover',
'Zurich': 'Zurich',
'Oslo': 'Oslo',
'Bremen': 'Bremen',
'Dresden': 'Dresden',
'Sheffield': 'Sheffield',
'Palermo': 'Palermo',
'Poznan': 'Poznan',
'Gelsenkirchen': 'Gelsenkirchen-Bottrop',
'Bottrop': 'Gelsenkirchen-Bottrop',
'Gelsenkirchen-Bottrop': 'Gelsenkirchen-Bottrop',
'Bordeaux': 'Bordeaux',
'Wroclaw': 'Wroclaw',
'Gothenburg': 'Gothenburg',
'Zaragoza': 'Zaragoza',
'Genoa': 'Genoa',
'Catania': 'Catania',
'The Hague': 'The Hague',
'Toulouse': 'Toulouse',
'Bristol': 'Bristol',
'Vilnius': 'Vilnius',
'Saarbrucken': 'Saarbrucken',
'Malaga': 'Malaga',
'Nantes': 'Nantes',
'Leeds': 'Leeds',
'Nottingham': 'Nottingham',
'Florence': 'Florence',
'Gdansk': 'Gdansk',
'Leipzig': 'Leipzig',
'Mannheim': 'Mannheim',
'Belfast': 'Belfast',
'Portsmouth': 'Portsmouth',
'Venice': 'Venice',
'Edinburgh': 'Edinburgh',
'Murcia': 'Murcia',
'Nice': 'Nice',
'Liege': 'Liege',
'Bratislava': 'Bratislava',
'Leicester': 'Leicester',
'Karlsruhe': 'Karlsruhe',
'Bergamo': 'Bergamo',
'Palma de Mallorca': 'Palma de Mallorca',
'Bologna': 'Bologna',
'Bielefeld': 'Bielefeld',
'Rouen': 'Rouen',
'Strasbourg': 'Strasbourg',
'Tallinn': 'Tallinn',
'Szczecin': 'Szczecin',
'Grenoble': 'Grenoble',
'Bari': 'Bari',
'Toulon': 'Toulon',
'Brighton': 'Brighton',
'Darmstadt': 'Darmstadt',
'Wuppertal': 'Wuppertal',
'Utrecht': 'Utrecht',
'Bournemouth': 'Bournemouth',
'Middlesbrough': 'Middlesbrough',
'Geneva': 'Geneva',
'Bydgoszcz': 'Bydgoszcz',
```

```
'Basel': 'Basel',
'Kaunas': 'Kaunas',
'Brno': 'Brno',
'Southampton': 'Southampton',
'Lens': 'Lens',
'Augsburg': 'Augsburg',
'Padua': 'Padua',
'Ostrava': 'Ostrava',
'Las Palmas': 'Las Palmas',
'Constanta': 'Constanta',
'Castellammare di Stabia': 'Castellammare di Stabia-Torre Annunziata',
'Torre Annunziata': 'Castellammare di Stabia-Torre Annunziata',
'Castellammare di Stabia-Torre Annunziata': 'Castellammare di Stabia-Torre Annunziata',
'Stoke': 'Stoke',
'Santa Cruz de Tenerife': 'Santa Cruz de Tenerife',
'Lublin': 'Lublin',
'Cardiff': 'Cardiff',
'Iasi': 'Iasi',
'Plovdiv': 'Plovdiv',
'Bradford': 'Bradford',
'Alicante': 'Alicante',
'Cluj': 'Cluj-Napoca',
'Napoca': 'Cluj-Napoca',
'Cluj-Napoca': 'Cluj-Napoca',
'Granada': 'Granada',
'Timisoara': 'Timisoara',
'Brescia': 'Brescia',
'Galati': 'Galati',
'Montpellier': 'Montpellier',
'Varna': 'Varna',
'Verona': 'Verona',
'Busto Arsizio': 'Busto Arsizio',
'Valladolid': 'Valladolid',
'Eindhoven': 'Eindhoven',
'Charleroi': 'Charleroi',
'Cordoba': 'Cordoba',
'A Coruna': 'A Coruna',
'Craiova': 'Craiova',
'Caserta': 'Caserta',
'Coventry': 'Coventry',
'Brasov': 'Brasov',
'Bonn': 'Bonn',
'Valletta': 'Valletta',
'Ghent': 'Ghent',
'Gdynia': 'Gdynia'}
```

Appendix D. Input

Inputs in general can be found at: https://github.com/seriousdeejay/citynet/tree/main/src/input

However some inputs are not available in the repository due to storage limitations or for licensing reasons. Notably the text input (Wikipedia dumps) are not included but instructions are given for how to obtain these.

- **Script for downloading wikidump**:
  By Diederik van Rijen.
  https://github.com/seriousdeejay/citynet/blob/main/src/analyses/1wikidump%20Downloading%20and%20Parsing.ipynb
- **wikiextractor command**:
  `python -m wikiextractor.WikiExtractor <Wikipedia dump file> -o <output_path>`
- **The list of cities** (IGEAT, 2007) **with added French and English toponyms**
  https://github.com/seriousdeejay/citynet/blob/main/src/input/List_of_cities_300k.csv
- **The joined ESPON** (IGEAT, 2007) **and EGM data** (EuroGeographics, 2022) **data**
  https://github.com/seriousdeejay/citynet/tree/main/src/input/maps

Appendix E. Code and Output

Appendix E.1. Code

All code for this thesis can be found at:
`https://github.com/seriousdeejay/citynet/tree/main/src/analyses/Brecht`

- **Preprocessing Functions**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/preprocessing_functions.py`
- **1 Article Selection and Matrix Construction**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/01_selection_and_matrix_construction.ipynb`
- **2 Wikidata Exploration**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/02_wiki_exploration.ipynb`
- **3 Match Coordinates and Create Dummies**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/03_match_coordinates.ipynb`
- **4 Edges**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/04_edges.ipynb`
- **5 Gravity Model**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/05_gravity_model.Rmd`
- **6 Nodes**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/analyses/Brecht/06_nodes.ipynb`

Appendix E.2. Output

- **French co-occurrence matrix**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/output/fr_matrix.csv`
- **English co-occurrence matrix**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/output/en_matrix.csv`
- **Edges with occurrences for all links (incl. 0 co-occurrence links)**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/output/edges.csv`
- **Edges (citylinks) with occurrences and predictions**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/output/edges_nz.csv`
- **Nodes with occurrences and predictions**:
  `https://github.com/seriousdeejay/citynet/blob/main/src/output/nodes.csv`
- **Outputs as .shp files (for mapping)**:[11]
  `https://github.com/seriousdeejay/citynet/tree/main/src/output/maps`

[11] .shp files must be kept in the same directory as their accompanying .crg, .dbg, .prj, and .shx files.