Mapping International Culture: Perception of Identities in a Hundred Years of Google Books

Shilin Jia

September 22, 2023

Abstract

In this chapter, we propose a methodological framework for measuring how countries historically perceived other identities when they engaged in war and trade with others. The method relies on the use of n-grams in millions of digitized books that Google has scanned since 2009. Based on word co-occurrences in the Google Ngrams, we are able to train yearly neural-probabilistic word embeddings from the corpora. The word-embeddings allow us to extract meaningful dimensions from the space and map different identities to the dimensions. This enables us to take a dive into history and understand how different language communities perceived other identities. We demonstrate reasonable success in validating our measures with external measures on international war and trade. However, despite our success in obtaining promising results in American English, British English, and French, further effort is still needed for constructing good measurements in German, Italian, and Russian.

Introduction

With the development of social surveys, social scientists have been increasingly able to gauge the minds of the public and understand how the public thinks about and perceives various

social and political issues. One topic of interest that frequently makes newspaper headlines is how the general public in a nation perceives foreign countries in the world. Feeling thermometers have been widely used in social surveys for answering such questions. The feeling thermometer asks respondents how favorable their views are on a certain country, ranging from "cold" to "hot." In the United States, Gallup and the Chicago Council on Global Affairs have been including feeling thermometers in their national public opinion surveys since the 1970s (Schneider 1985). Since the early 2000s, Pew Research Center has been conducting a multi-national survey asking respondents in dozens of countries about their perceptions of foreign countries (Li 2021). Such longitudinal records allow researchers to develop a historical understanding of how the general public in a country changes their perceptions in response to external events. For instance, Pew's surveys show that Americans' perception of China has reached a historical low following the U.S. and China's recent clash in trade wars and on the spread of coronavirus (Li 2021; Pew Research Center 2020). Scholars found that public perceptions are correlated with foreign policies (Lee and Hong 2012). In democratic countries, such as the United States, change in public perception oftentimes even precedes change in national foreign policies (Page and Shapiro 1983). Studies on public perceptions could yield valuable insight into understanding international relations and provide explanations such as why democratic couturiers don't fight each other (Gries et al. 2020). However, because those surveys have been only been systematically conducted in recent times, how people in different nations historically perceived other nations remains a relatively unknown topic. It is impossible to travel back in time and ask people in the past about their perceptions. Also, perceptions of other identities could be inherently multi-dimensional, but a single-measure feeling thermometer ranging from "cold" to "hot" could be misleading (Li 2021). Scholars' understandings of this topic have so far still been relatively limited by the availability of data.

One way to measure human biases is through language (Dodds et al. 2015; Kozlowski, Taddy, and Evans 2019). Instead of asking people about their views on certain subjects,

researchers can observe what people say or write in their languages, and more specifically, in what contexts subjects of interest are mentioned. In this chapter, we propose a new research framework for understanding how literate elites in different language communities historically perceived other foreign identities. We use the Google Ngram Corpus to build yearly word-embedding spaces in 6 languages that are currently available. Then by projecting foreign national identities onto two in-group vs. out-group dimensions, we are able to obtain historical measures of foreign identities in these two dimensions. We show that in American English, British English, and French, there is consistent change in perceptions of other identities when their countries are at war with other countries.

Measuring Cultural Dimensions in Word Embeddings

Recent development in neural probabilistic language models has provided social scientists a way to map millions of words and documents from a gigantic corpus into a high-dimensional vector space (Bengio et al. 2003). Theoretically, this modeling approach resonates with structuralist theories and the distributional hypothesis in linguistics (Saussure 2011; Harris 1954; Firth 1957). The theory holds that the meaning of a word in a language is defined by its surrounding words, and words that appear in similar contexts have similar meanings. Neural probabilistic language models represent each word in a corpus as a vector and use the vectors to optimize the task of predicting words given their contexts. The vector representations, as by-products of the models, turn out to have high interpretative values. The models are able to map words that occur in similar contexts into proximate locations in a vector space.

The Word2Vec model and its skip-gram variant developed by Mikolov et al. has become a state-of-the-art technique for building vector representations of words from a large corpus (Mikolov, Sutskever, et al. 2013; Mikolov, Chen, et al. 2013). Based on a simplified two-layer neural network design, it is able to process huge volumes of texts fastly and efficiently. Computational linguists have utilized the model to study change of words' meaning over

time (Hamilton, Leskovec, and Jurafsky 2016).

Interestingly, computational linguists have also found that word-embedding models are not only able to map local clustering of words with similar meanings but are also capable of performing more complicated language tasks such as solving analogy tests. The models are able to solve standardized test questions in the form of "a is to b as c is to __" (Rumelhart and Abrahamson 1973; Mikolov, Sutskever, et al. 2013). A classic example is the question "man is to woman as king is to __." Word-embedding models trained on a reasonably large English corpus are usually able to yield $\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} \approx \overrightarrow{queen}$.The result suggests that $\overrightarrow{man} - \overrightarrow{woman}$ and $\overrightarrow{king} - \overrightarrow{queen}$ are parallel directions in the vector space, and there might exist a global gender dimension in the space. The reason why such a dimension exists is that "man" and "woman" are similar words in the sense that they are both nouns used to designate genders. They are oftentimes used interchangeably in texts except for the crucial difference that "man" is used more often in masculine contexts, and "woman" is used more often in feminine contexts. Subtracting the vector representations cancels out their similarity and keeps their difference. And a similar difference is also kept in taking the subtraction of "king - queen." The models are able to solve many different kinds of semantic and syntactic analogy tests such as "France is to Paris as Italy is to Rome" and "bad is to worse as big is to bigger." They are also able to learn explicit and implicit biases and stereotypes in languages. For instance, Bolukbasi et al. (2016) find that word embedding models could yield parallelograms such as "man" is to "computer programmer" as "woman" is to "homemaker" and "father" is to "doctor" as "mother" is to "nurse." Caliskan, Bryson, and Narayanan (2017) find that in their word-embedding models, African-American names are more associated with unpleasant words than European American names.

Kozlowski, Taddy, and Evans (2019) propose a method to project cultural items onto cultural dimensions. The idea is that antagonist pairs such as $\overrightarrow{man} - \overrightarrow{woman}$ and $\overrightarrow{rich} - \overrightarrow{poor}$ represent some global directions in the vector space. And by projecting words such as "scientist", "nurse", "hamburger" and "wine" to those directions, researchers are able to tell

how feminine/masculine and rich/poor each of these items is in a given linguistic community. Presumably, this measure would yield a higher loading of "king" and a lower loading of "queen" in a feminine/masculine dimension. The measure allows researchers to map as many items as there are in a corpus to the same linguistic dimension and yield a single measure of how these items score in that dimension. In our study, we apply this method to study international cultural perceptions.

Research Design

Building yearly word embedding models from the Google Ngram corpus

We trained our word-embedding models from 6 languages of the Google Ngram corpus. The Google Ngram corpus is one of the largest publicly available corpora that contain a huge amount of texts produced in human history. It is produced out of Google's massive effort in digitizing historical texts into Google Books. Although the original texts are copyrighted, Google releases all of the ngrams that appear in the texts. An ngram is n-words co-occurring together in a text. The maximum ngram length that Google supports is 5. The corpus gives a count of the number of times each ngram appears in the Google Books published in a given year. All n-grams that appear at least 40 times across all years are included. Researchers have used the ngram counts to tell the relative importance of different cultural items, identities, and political figures in human history (Michel et al. 2011; Junyan Jiang and Xie 2020). The following list is a snippet of 5-grams in UK English.

've a bloody good mind

've a decision to make

've a feeling that he

've a good mind ...

've a job for you

've a kind and loyal

've a legal right to

've a life to live

've a mind to show

've a mortal aversion to

've a notion in my

've a notion that if

've a plenty for them

've a pretty shrewd idea

The ngrams have also been used for learning language models (Hamilton, Leskovec, and Jurafsky 2016). In this study, we use the same corpus that has been previously used by many researchers to build yearly word-embedding models from 1900 to 2019. We also extend our research scope to 6 languages including the US and UK English, French, German, Italian, and Russian. Google uses the location of publishers to distinguish between the US and UK English.

When processing a normal text, Word2Vec uses a sliding window to predict and learn what other words would appear in a word's context window. Training a Word2Vec model on Google 5-grams is as if we are processing a normal text through a sliding window of 5. Google does not provide information about what specific books each n-gram appears in, and each year's corpus is released as a whole. Scholars using Google Ngrams also noted that the composition of Google Books could have changed significantly during the last century (Kozlowski, Taddy, and Evans 2019). Around 2009, the primary source of Google Book also changed from books scanned from libraries to digitized materials released directly by publishers. Although the Google Books corpus is large, it certainly does not represent all speakers in a language community. It is also probably not a representative sample of all

books published in a language. However, given its gigantic size, word-embedding models are able to learn common linguistic dimensions shared by authors in a language. The dimensions can also tell common biases and stereotypes in a language community.

We sampled from all 5-grams to construct our yearly corpus and trained our yearly word-embedding models. Figure 1 shows the sizes of the 5-gram corpora across time. Because the corpus sizes are dramatically different across years, to make our yearly models comparable, we randomly sampled the same number of 5-grams from each year's corpus. The yearly sample size we used is the smallest yearly corpus size in the full corpus. For instance, in British English, there were least publications in 1944 during the Second World War period, and the yearly sample size we used for other years is equivalent to the size of the 1944 corpus. To account for randomness in the texts, we bootstrapped 20 samples for each year using sampling with replacement. The bootstrapped samples would allow us to construct confidence intervals for our time series. Sampling with replacement has the effect of dropping rare 5-grams from our yearly corpus. It helps us to focus on central tendencies in the corpus. The Russian language had a major language reform in 1918 following the Bolshevik Revolution. Many modern words did not exist in the Russian corpus prior to 1918. So we chose 1918 as the beginning year of our Russian models. Unfortunately, for US-English, the corpus size is so large that training yearly models according to our framework would cost a huge amount of time. We were not able to train all of our models in time. With a compromise, we trained one model for each year using 20% of the least corpus size. The 20% of the least corpus is still larger than the yearly sample sizes of all other languages. Presumably, measures in US English would have shorter confidence bands than the ones observed in other languages. Table 1 summarizes information about our samples.

After obtaining our yearly samples, we trained yearly word-embedding models using the skip-gram model (Mikolov, Chen, et al. 2013). With the implementation of a negative sampling framework, what the skip-gram model does, in a nutshell, is to use a logistic regression to distinguish word pairs that actually occur in the corpus and randomly generated

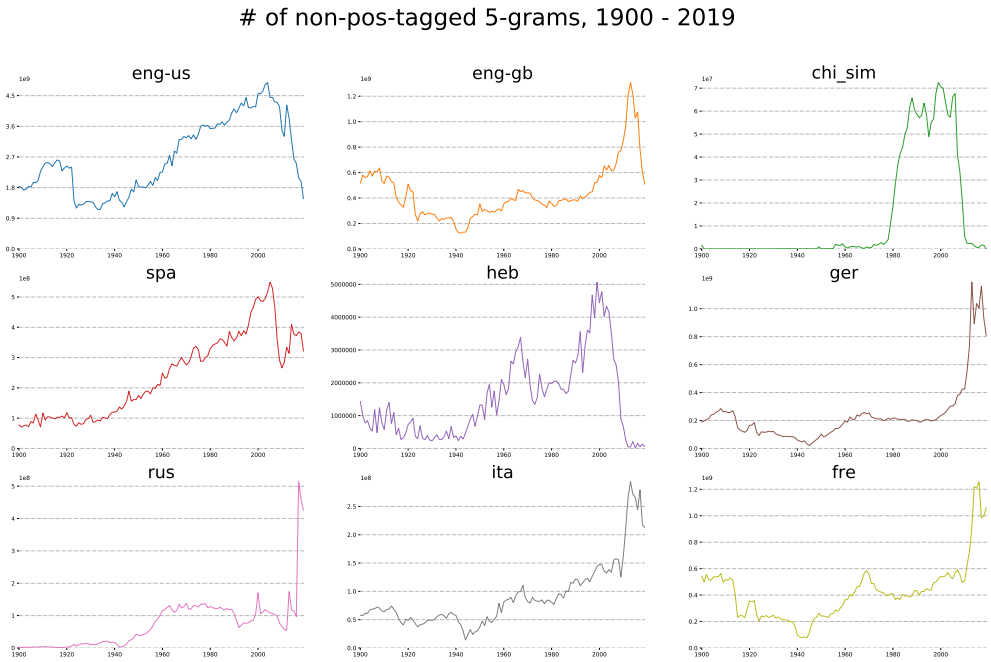

Figure 1: Total number of Google 5-grams in all available languages

Table 1: Sample sizes of our study

language	yearly sample size (total # of 5-grams)	bootstrapped samples	years
US English	2.3e8	1	1900-2019
UK English	1.2e8	20	1900- 2019
French	7.6e7	20	1900- 2019
German	2.1e7	20	1900- 2019
Italian	1.4e7	20	1900- 2019
Russian	1.4e6	20	1918- 2019

"negative" word pairs that don't appear in the corpus. Through stochastic gradient descent, it finds the best vector representations of words to maximize the log-likelihood of the logistic regressions. More details about the inner working of the model are explained in the appendix of this chapter.

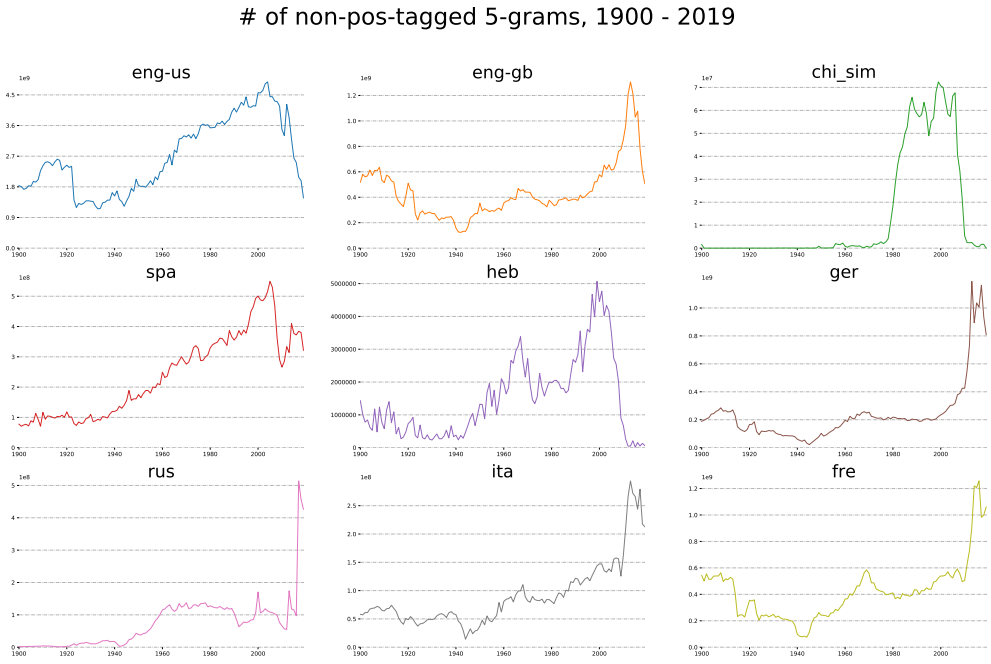

Figure 2: Total number of Google 5-grams in all available languages

External data on international war and trade

We used data on international war and trade to externally validate the soundness of our cultural measures[1]. The international war data comes from the Correlates of War Project [2]. The project has records of all international wars between nations from 1816 to 2007. It records whether a country is an ally or enemy of another country in a specific year. The international trade data comes from the Historical Bilateral Trade and Gravity Data set (TRADHIST) constructed by Fouquin, Hugot, et al. (2016). The dataset contains import

1. Special thanks to Tamara van der Does for compiling and formatting these data.
2. https://correlatesofwar.org/

and export trade volumes between nations from 1827 to 2014. It also has each country's historical GDP. We are interested in exploring whether our cultural measures correlate with international trade and war.

In-group vs. out-group dimensions

To construct our cultural measures, we focus on two analytical dimensions in in-group vs. out-group distinction. The first dimension we focus on is the "friend-enemy" dimension. As theorized by German political philosopher Carl Schmitt (1976), the friend-enemy distinction is a fundamental distinction in politics, which specifies with whom a political entity should form an alliance and make enemies in political struggle. We expect the dimension to be salient in a language community during wartime.

To measure more implicit cultural distinctions, we constructed a separate "we-they" dimension. Based on the social identity theory proposed by Tajfel et al. (1979), a basic process in people's everyday socialization is distinguishing who are part of "them" and who are part of "us." In this process, people use stereotypes to create group images for "others" and "us." This process is an origin of inter-group conflict in social life. It is more cultural and implicit. Word Embedding models provide us a tool to measure which identities are part of "them" and part of "us" in a language community. Because countries are not in direct conflicts most of the time, we expect that this dimension is more salient in telling subtle country-to-country relations during peacetime. We also expect that the we-they dimension is associated with how often members in a community interact with members in other communities. Therefore, it might have a stronger relationship with international trade.

Following Kozlowski, Taddy, and Evans (2019)'s method, for each yearly model, the dimension $\boldsymbol{d}$ is constructed as

$$\boldsymbol{d} = \sum_{w_i \in \text{positive-words}} \text{norm}(\boldsymbol{v}_{w_i}) - \sum_{w_j \in \text{negative-words}} \text{norm}(\boldsymbol{v}_{w_j}) \tag{1}$$

, and the loading of each selected identity-word in the dimension is the cosine similarity between its vector representation and $\boldsymbol{d}$.

Table gives the lists of positive and negative words we used in constructing the friend-enemy and we-they dimensions in 6 languages. Figure 3 shows that the chosen pairs in UK English are indeed parallel throughout history, at least in 2-dimensional visualizations. The visualization suggests that the dimensions exist.

Table 2: Postive and negative words used for constructing dimensions

language	dimension	positive words	negative words
US/UK English	friend-enemy	friend, ally	enemy, foe
US/UK English	we-they	we, us, our, ours, ourselves	they, them, their, theirs, themselves
French	friend-enemy	ami, amis, amie, amies	ennemi, ennemis, ennemie, ennemis
French	we-they	nous, on, notre, nos	ils, eux, leur, leurs
German	friend-enemy	freundin, freund	feind
German	we-they	wir	sie
Italian	friend-enemy	amica, amico	nemica, nemico
Italian	we-they	noi	esse, essi
Russian	friend-enemy	друзья, друзей, друзьям, друзей	враги, врагов, врагам, врагов
Russian	we-they	Мы, Нас, Нам, Нами, Нас	Они, Их, Им, Ими, Них

Results

Time series plots

After constructing the yearly in-group vs. out-group dimensions, we were able to project selected national identities to each dimension and make time series plots.

US-English

Figure 4 contains plots of selected national identities in the friend-enemy dimension in US-English. Without external validation, the plots make intuitive sense. Unsurprisingly, USA

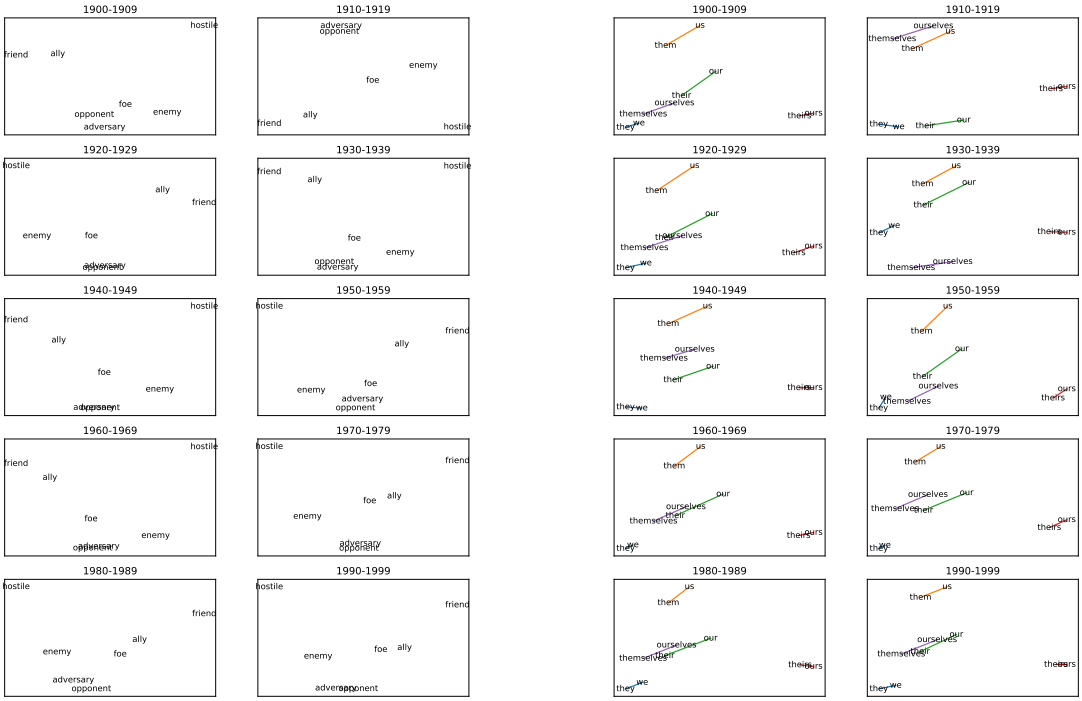

Figure 3: Friend-enemy and we-they pairs projected to 2D PCA dimensions

and Canada are the most friendly nations in American English. The two world wars had significant impacts on the friendliness of Germany, and Germany remained the least friendly western European country most of the time. Italy also experienced a similar drop during WWII but not WWI. Among Asian nations, Japan experienced a significant drop during WWII while China experienced a considerable boost. China and Japan's (as well as India's) friendliness scores were almost in parallel up till the outbreak of WWII. They also began to become indistinguishable from each other since the mid-1960s. Korea experienced a significant drop during the Korean War. Philippines as a former U.S. colony experienced a significant drop during WWII probably because of its independence movement. Vietnam experienced a significant drop around the time of the Vietnam War, but it bounced back after the 1970s and became indistinguishable from the Philippines. Thailand, on the other hand, has historically been a neutral country and is the most friendly Asian country most of the time. It nevertheless experienced a huge dip in WWII probably because it joined Japan's alliance during the middle of the war. Among Latin American countries, Mexico has always been the most friendly nation. Cuba experienced a huge drop around the time of the Cuban Revolution and remained low after that. Among countries in the Middle East, many countries including Iran, Afghanistan, and Libya experienced a significant drop around the time of the Islamic Revolution. Iraq continued its drop till the First Gulf War. It slightly bounded back after the war but dropped again around the time of the Second Gulf War. Lastly among Eastern European countries, the word "Soviet" has always been a highly hostile term. Its score started to decline even before the end of WWII. The image of Russia, however, was separate from that of the "Soviet." After the end of the Cold War, all former communist countries (except for Yugoslavia, which ceased its existence) seem to have enjoyed an increase. Overall, most of the time series conform to conventional understandings of the United States' foreign relations. Wars seem to have played a major role in influencing other nations' perceived friendliness in US English. It is also worth noticing that countries in the same regions seem to oftentimes co-vary a lot. However, there does not seem to be

any universal co-variation among all countries.

Figure 5 are the corresponding plots in the we-they dimension. War does not seem to play as much influence in those time series. And the graphs seem to be less interpretable in terms of major international events. For instance, although Germany was an enemy during the two world wars, its we-ness is not distinguishable from many other Western European nations including even America itself during wartime. Interestingly, although "UK" was a relatively novel abbreviation used in the history of American English, it became one of the most "us" nations in the latter half of the 21st century. However, "Britain," as an old word, remained the lowest "us" nation. The difference could tell some path dependency in the use of language. Among Asian nations, Japan also didn't experience a significant drop during WWII but had been in steady decrease till the end of the 1970s. After that, its we-ness started to bounce back along with China and Korea. Thailand, although is one of the most friendly Asian countries in US English, isn't as "us" as China, Japan, and Korea. More gradual and long-term international interactions in trade and immigration perhaps could explain the difference. Among Middle Eastern countries, Iraq also didn't experience a dramatic drop during the Gulf Wars. Lastly among Eastern European identities, "Russia" and "Soviet" started their great diverge since the middle of the 1980s probably because many Russians immigrated to the United States, and Russia itself also became a non-communist country. Overall, only a few well-known geopolitical crises, such as the Vietnam War, the Cuban missile crisis, and the Iranian Revolution seemed to have played a role in influencing the "us-ness" of other national identities. The we-they dimension is more cultural than political and is less reflective of high-level politics. Interestingly, although Mexico is the most friendly Latin American nation in US English, it is the least "us" identity in the we-they dimension. This suggests that frequent interactions do not necessarily lead to more incorporation. On the political level, Mexico has always been a close ally of the United States. But on the cultural level, its identity has always been more treated as "others" than "us."

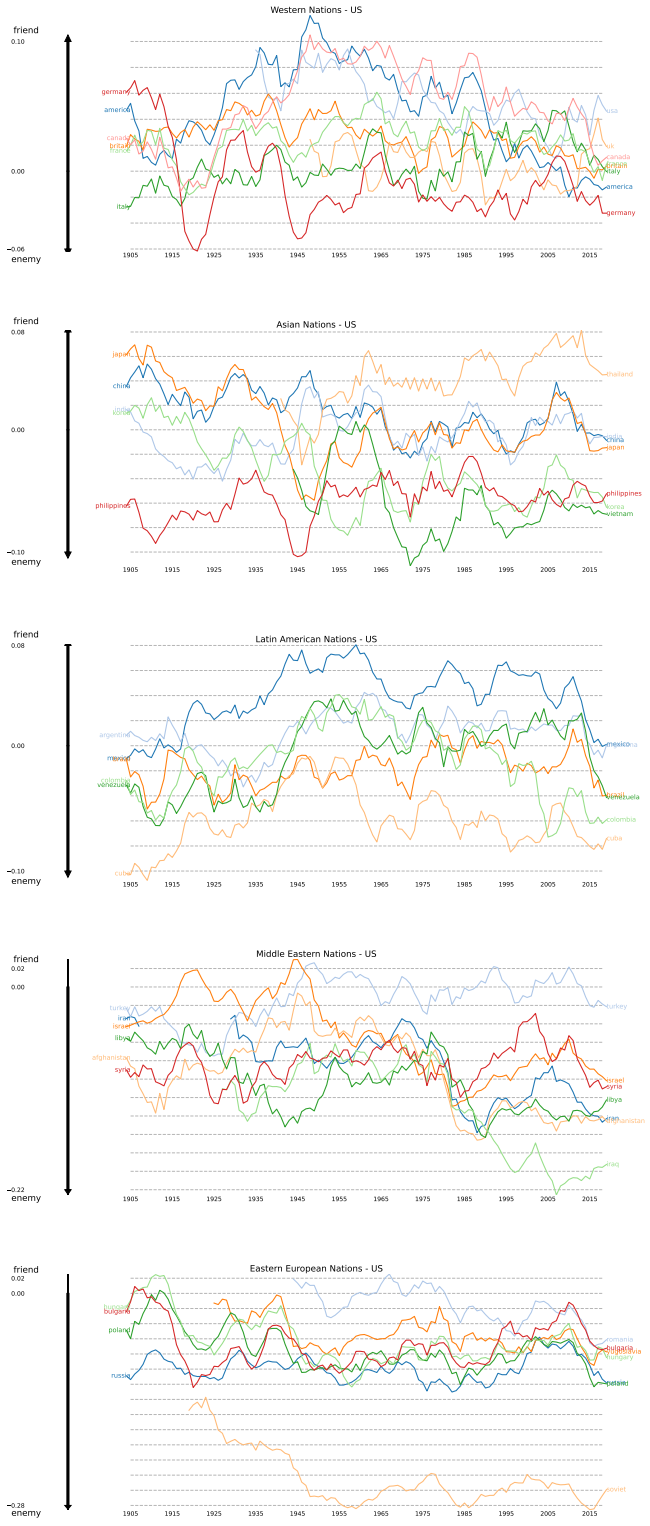

Figure 4: Loadings of selected countries in the friend-enemy dimension in US English

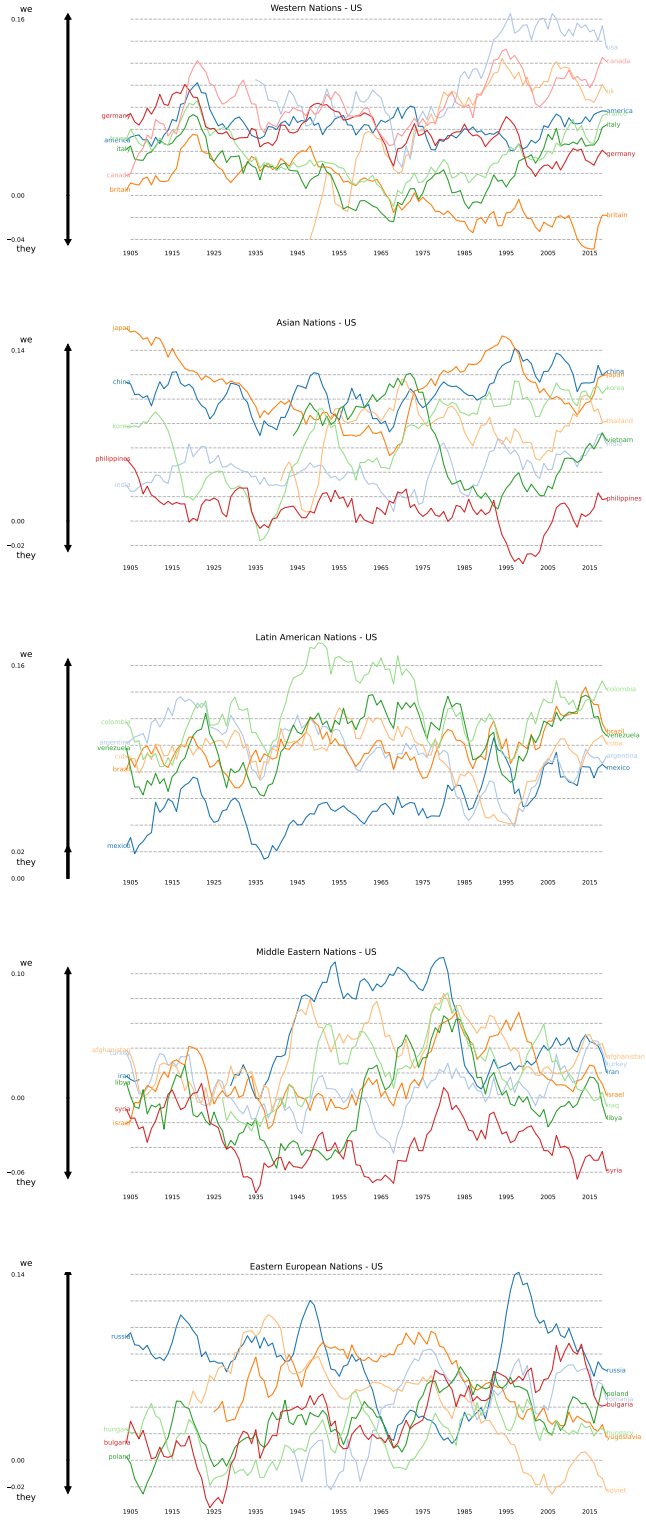

Figure 5: Loadings of selected countries in the we-they dimension in US English

UK-English

Figure 6 and 7 are our results in UK English. Because we had 20 bootstrapped samples, we were able to create 90% confidence intervals for our measures. The confidence intervals help to tell that a lot of times, observed differences in the means are not statistically significant. There is a considerable degree of randomness in countries' loadings, which could depend on which 5-grams were drawn into our samples. However, the graphs are still able to pronounce some major historical changes. "America" and "USA" have always been the most friendly Western European identities in British English. They also seemed to enjoy some special favor during WWII and the beginning of the Cold War. Germany and Italy experienced significant drops during WWII. France also seemed to have experienced a drop during WWII probably due to its surrender. However, during WWI, as allies of the UK, France and Italy both experienced a boost contrary to Germany. Japan experienced a huge drop amid WWII. Thailand is again relatively always a more friendly nation. Burma, as a former British colony, enjoyed the same rank as India before WWII. However, probably because it severed its tie with the British Empire after gaining independence, its friendliness continued to drop in the second half of the 20th century. Among Latin American countries, Cuba clearly stands out as an outlier after the Cuban Revolution. Among Middle Eastern countries, Iran, Libya, Afghanistan, and Iraq followed similar paths around the middle of the 20th century with Iraq dropping to the lowest during the time of the Gulf Wars. Loadings in the we-they dimension also in general tell a different story than the friend-enemy dimension. Germany has always been more "us" than "they." Japan and China are the most "us" Asian countries. Latin American and East European countries are mostly indistinguishable from each other. "Russia" and "Soviet" experienced the same divide since the end of the 1980s. Again, countries in the same regions seem to co-vary. But there does not seem to be any global co-variation. We also projected the constituent nations of the United Kingdom to the we-they dimension. Interestingly, the use of "UK" seems to generally complement the use of "Britain," "England," "Scotland," and "Wales." When "we" become more "UK," "we" become

less “Britain,” “England,” “Scotland,” and “Wales.” All the sub-national identities seemed to be in decline following the world wars. However, the trend has been reversed since the 1980s.

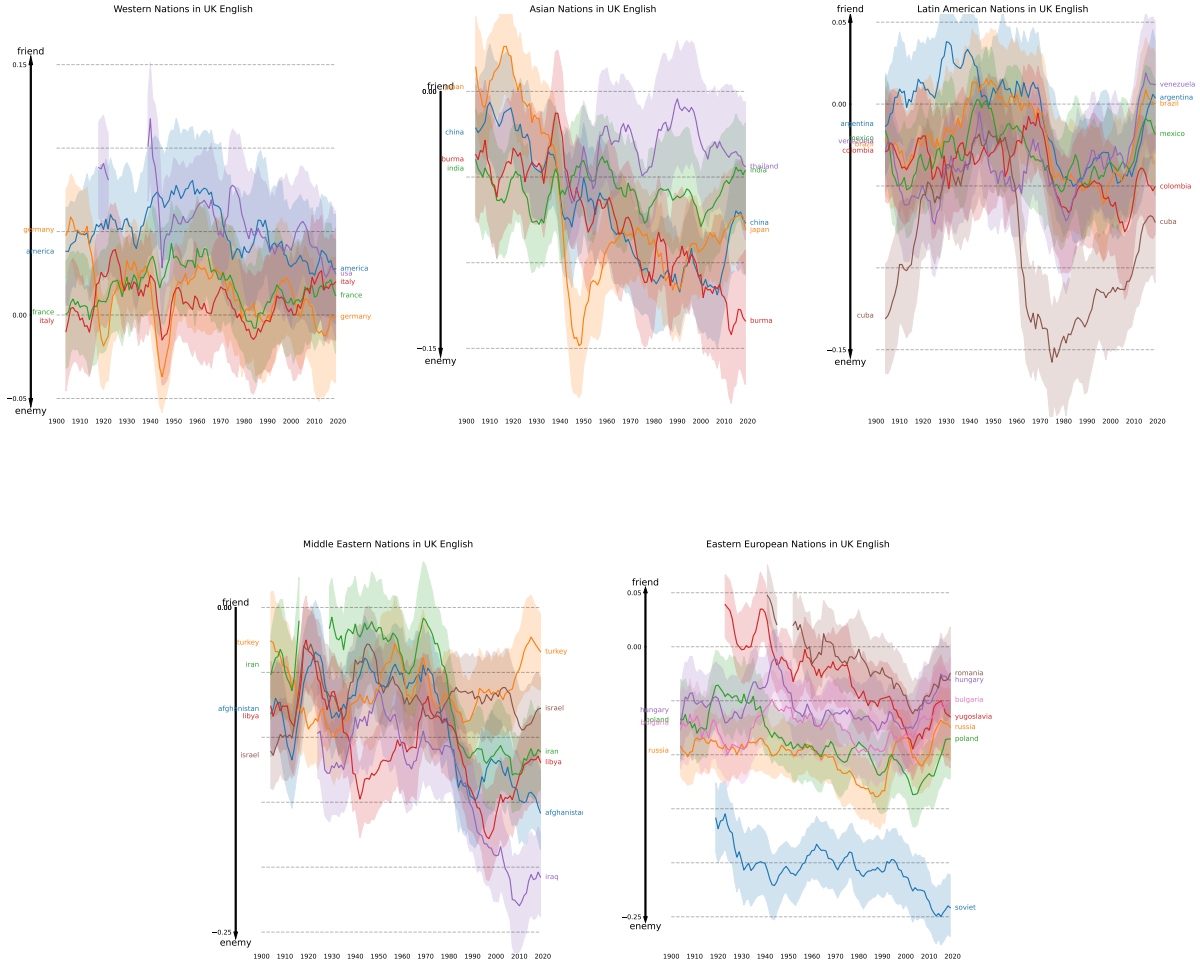

Figure 6: Loadings of selected countries in the friend-enemy dimension in UK English

French

The French corpus is the third largest corpus that we have. Figure 8 and 9 are our results in French. The stories that the graphs tell are very similar to the US and UK stories with some French ingredients. The two world wars are clearly pronounced in the friend-enemy plots. Although France was perhaps seen as a German collaborator in the eyes of the British, it

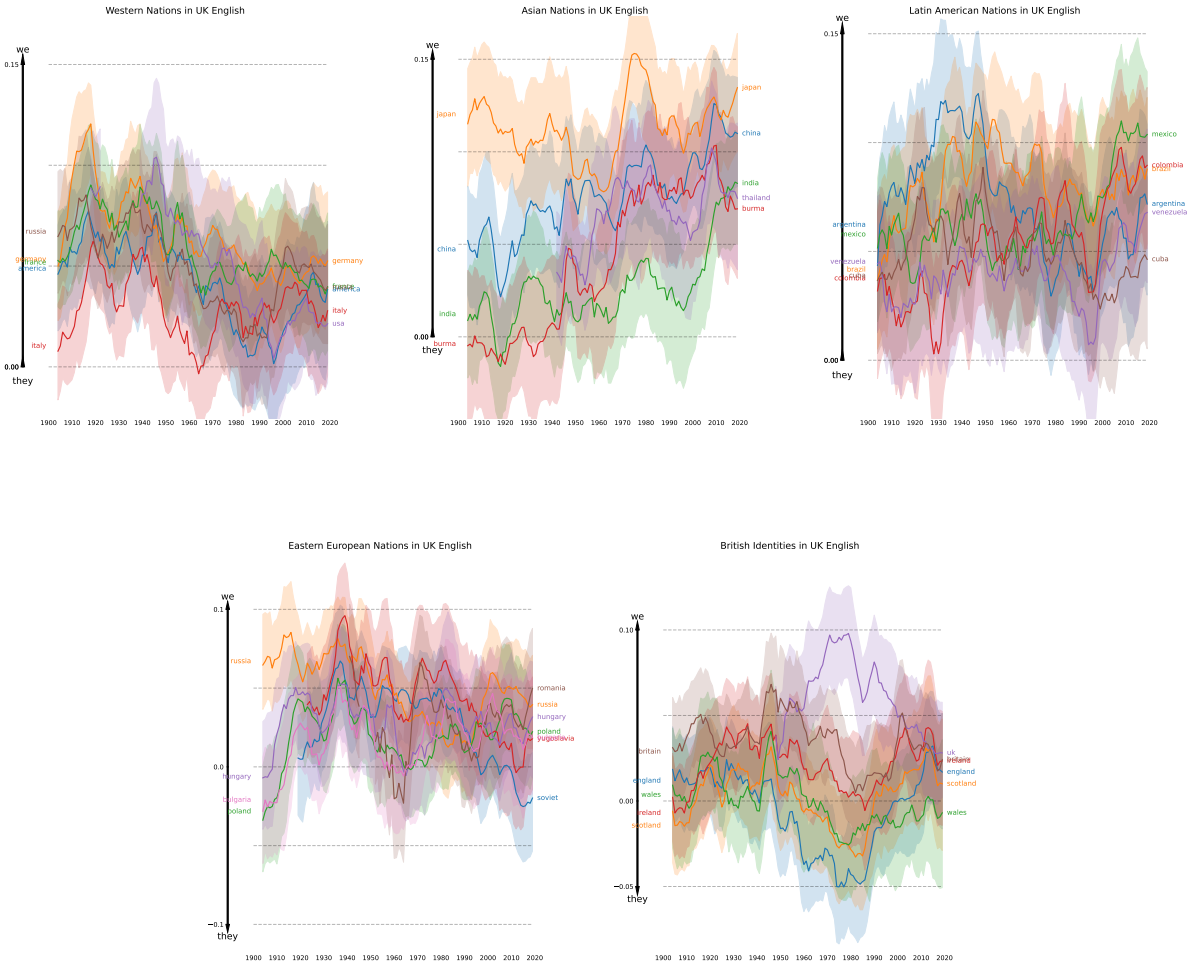

Figure 7: Loadings of selected countries in the we-they dimension in UK English

sees Germany majorly as an existential threat in its own language. Japan and Italy were both seen as enemies during WWII. Canada is unsurprisingly the most friendly Western European nation probably due to its close linguistic tie with France. Vietnam's friendliness dropped significantly after it claimed its independence. However, the image of "Indochine" remained relatively unchanged. Vietnam reverted back to the level of Indochina in the late 1980s, and these two identities became indistinguishable again as Vietnam gradually re-opened and rebuilt its tie with its former ruler. To French authors, although Vietnam was lost, the image of Indochina remains in their imaginary. The dropping of the image of Cuba and the deterioration of the relationships with the Muslim world seem to be universal across American English, British English, and French. Interestingly, the friendliness of Algeria started to deteriorate before the general decline of the friendliness of the Muslim world. The decline should correspond to the Algerian Independence War. "Soviet" is also the least friendly identity that is only comparable to that of Germany during world wars with Russia enjoying a different path. In the cultural dimension, we also observe some similar general trends and some special French flavors. For instance, although Britain was politically an ally during world wars, it has always been more "them" than "us" in French publications. The friendliness of Indochina although remained largely unaffected by the independence of Vietnam, there is a clear sign that Indochina, which was one of the most "us-ish" identities at the beginning of the century, is no longer part of "us." The same happened to Algeria.

Less desirable results in German, Italian, and Russian

Unfortunately, we do not achieve the same level of interpretability in modeling German, Italian and Russian. Our results are shown in Figure 10-12. Some trends seem to be still interpretable. For instance, in German, there seems to be the same Clash of Civilization trend between the Western World and Muslim Countries. However, we will not go into detail in trying to interpret the graphs. Two basic problems observed in these graphs make us doubt the value of interpreting these graphs. First. there seems to be wide global co-

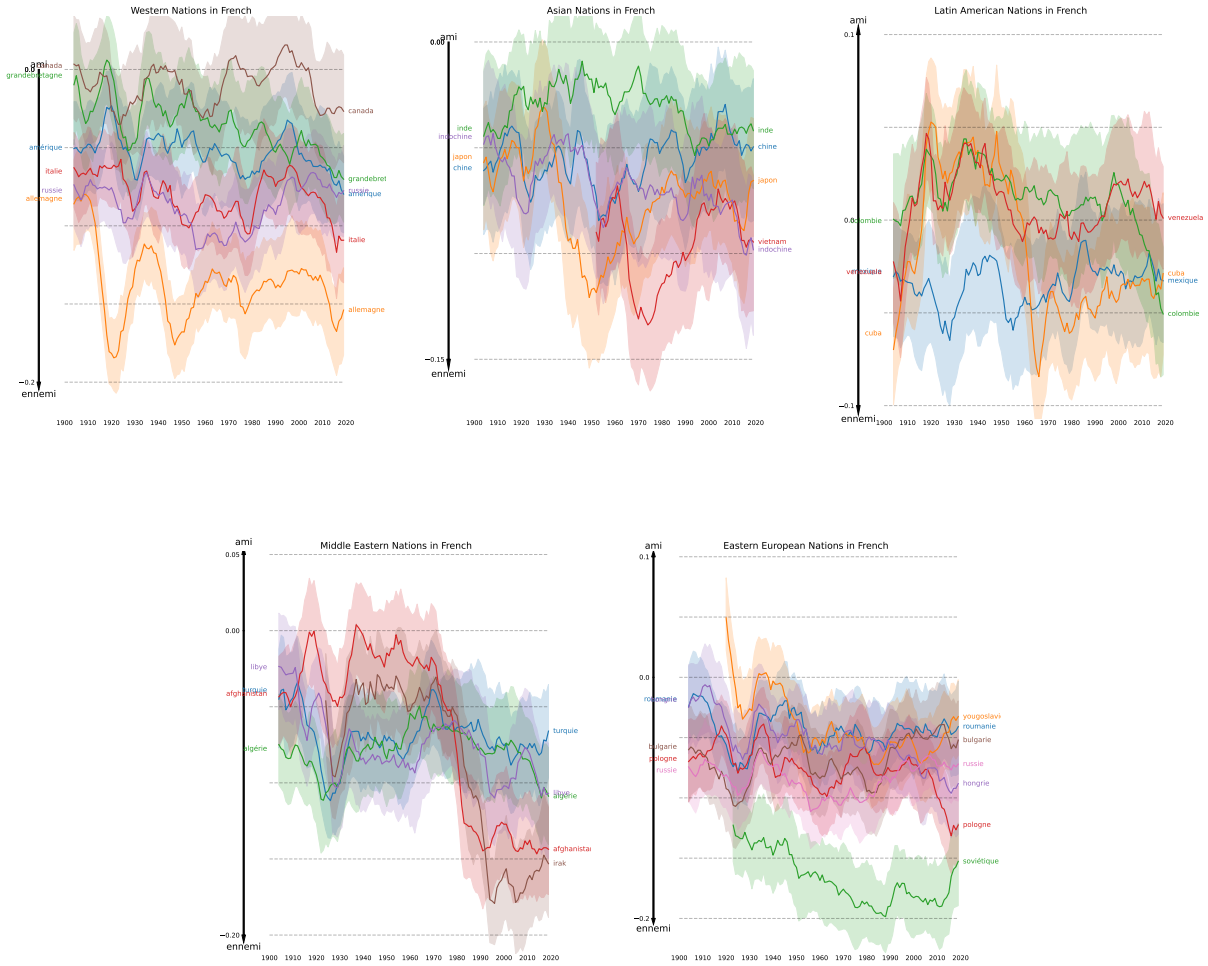

Figure 8: Loadings of selected countries in the friend-enemy dimension in French

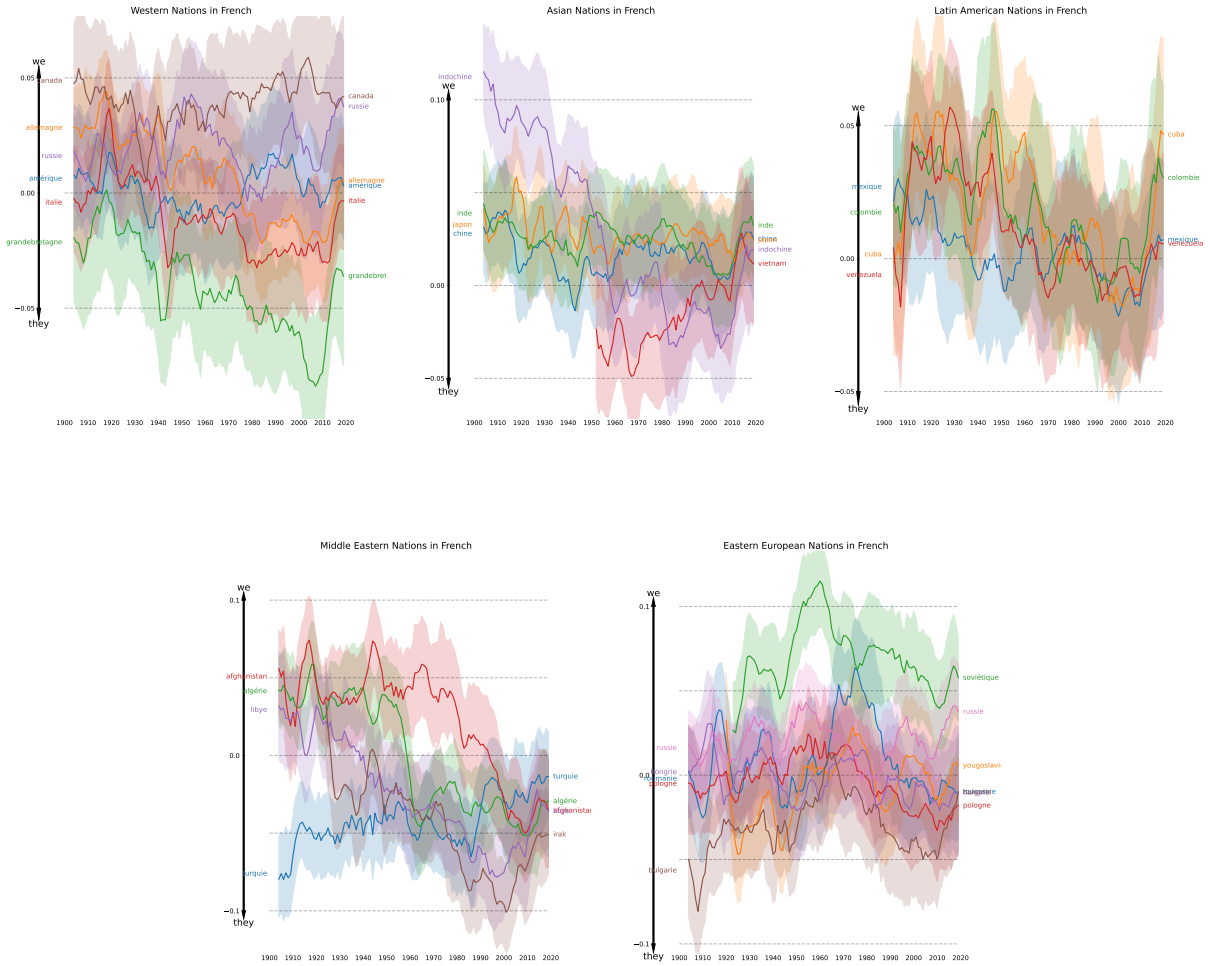

Figure 9: Loadings of selected countries in the we-they dimension in French

variation among not just countries in the same region but all countries. This makes us wonder whether there is some spatial distortion in the word-embedding spaces. We observed similar spatial distortions when we first trained our English and French models. We were able to correct the problem by adopting a more methodologically consistent sampling scheme. (The details of our correction are documented in the Appendix.) However, even after applying the same sampling scheme to all corpora, we still observe spatial distortion in the word-embedding spaces of German, Italian and Russian. Because we are not able to exactly identify the cause of spatial distortion, we cannot tell whether the observed trends are due to some real historical processes or some noises introduced by our models. Second, some historical trends reflected in those graphs do not make intuitive sense. For instance, Japan and Italy were seen more as enemies rather than friends in German during WWII. The same is also true for Austria during WWI. We will discuss potential causes of the unfavorable results and try to continue to improve the quality of our models after this study.

Stretch of Dimensions during War Time

Lastly, one thing that we do observe across corpora is the stretch of dimensions during the world wars. Figure 13 plots the cosine similarities between positive words and negative words for all languages and all years. The higher the score is, the more the positive and negative words are used in dissimilar contexts. Except in US English, there is some significant stretch of the friend-enemy dimension at some point in history in all other languages. In British English, French, German, and Italian, the stretch all happened during the two world wars. During war time, the word "friend," and "enemy" become more dissimilar to each other in terms of their contextual usage. In Russian, a stretch also occurred, but it seems to correspond to the Great Purge rather than WWII. A stretch in the we-they dimension seemed to occur later in WWII. In other languages, during the same time that the stretch happened, there did not seem to be any significant stretch in the we-they dimensions. US English seems to experience less stretch of dimensions in all time. The distinction of friend

Figure 10: Loadings of selected countries in the friend-enemy and we-they dimensions in Italian

Figure 11: Loadings of selected countries in the friend-enemy and we-they dimensions in German

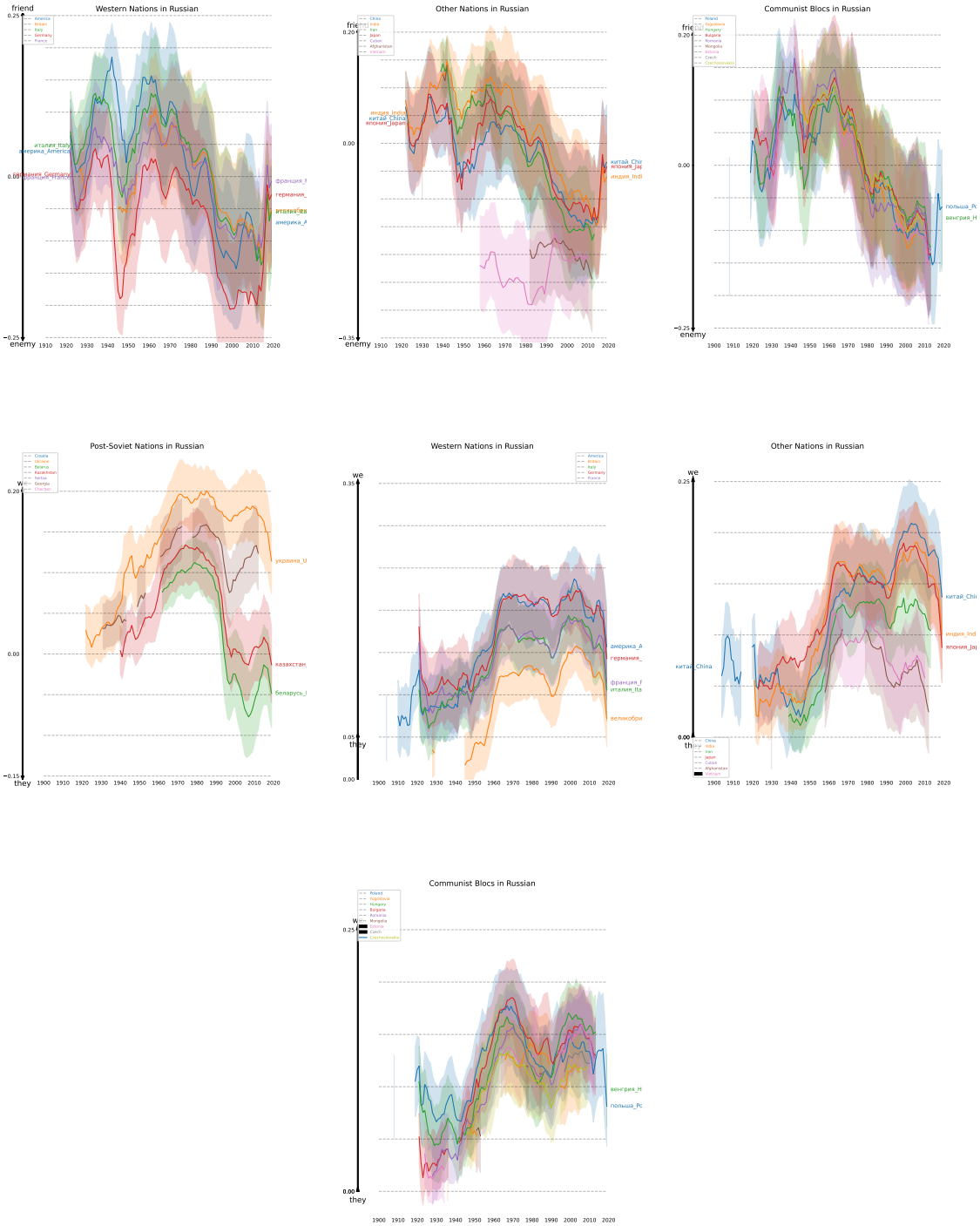

Figure 12: Loadings of selected countries in the friend-enemy and we-they dimensions in Russian

and enemy were at their historical low during war time. However, the drop was not as dramatic as it is observed in other languages. It could be due to the fact that the two world wars did not happen on America's home continent and were not existential wars to the United States. Because some spatial distortion seems to also occur in the we-they dimensions of the German, Italian, and Russian spaces, we cannot tell whether the stretch of the friend-enemy dimensions could have anything to do with the spatial distortion observed in German, Italian and Russian plots presented in the preceding section. However, it could still nevertheless have an impact on the dimensional loadings we observed. This could mean that even though we applied the same measure to all years, the measuring ruler itself may not be consistent across years especially during war time. It is like when the special theory of relativity applies to the physical space, not only do things in the space change, but the space itself is changing during those special moments. The same amount of distance in two spaces may not be easily comparable.

Confirmatory analysis

All of our analyses thus far have been based on subjective interpretation. We cannot rule out the possibility that the interpretations only lie in our eyes while in fact they only correspond to some random noises. In the last part of this chapter, we will try to externally validate our measures. In our confirmatory analysis, we run two mixed-effect models on all of the 6 languages to confirm the relationship between perceptions and war and trade. To make our analyses comparable, for each host language, we selected 30 other countries as guest countries for evaluating their friendliness and we-ness. Then, we want to test whether their friendliness or we-ness is correlated with external measures on international trade and war. Although our external dataset almost covers the entirety of the time period we are interested in, many yearly observations are missing due to either lack of statistical records or simply the fact that some nations did not exist in some years. So for each host language, we select 30 guest countries with least missing values. Table 3 gives the guest countries being selected

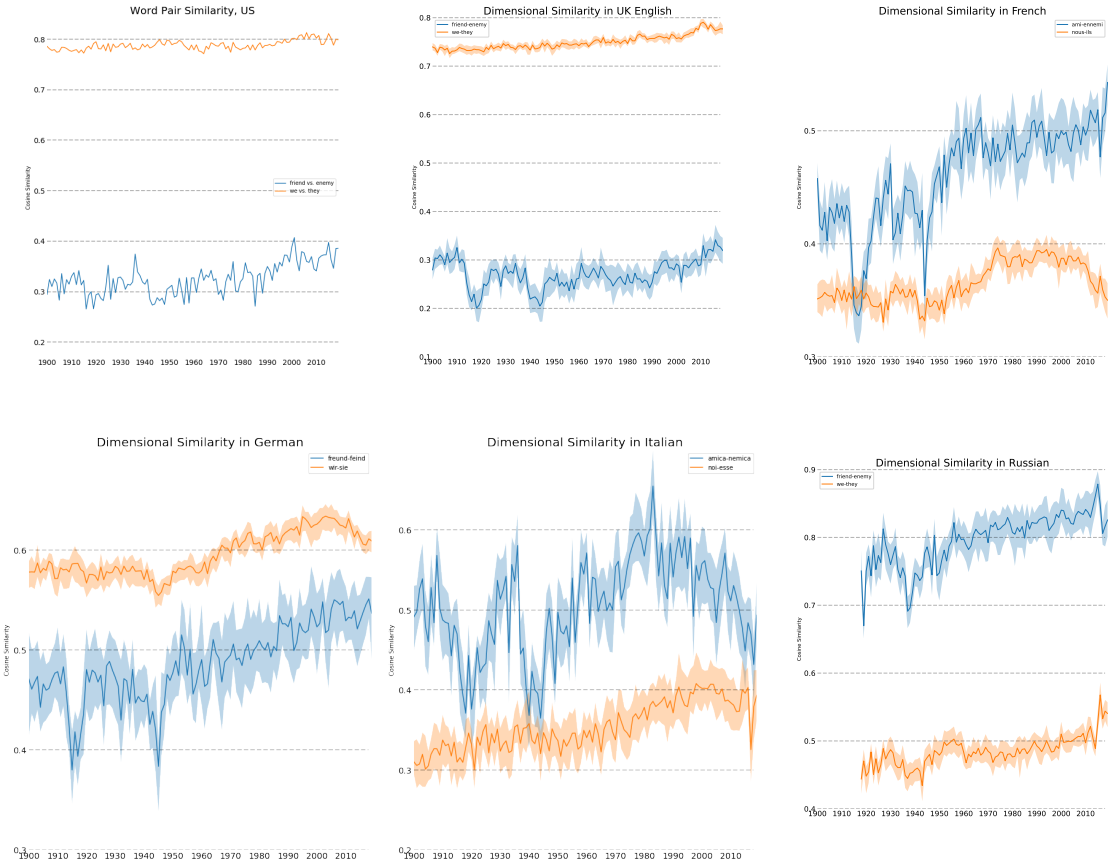

Figure 13: Dimensional similarities in 6 languages

into our analysis. For UK and USA, because the abbreviations either do not exist in some languages or were frequently used only in recent time, we used the words "America" and "Britain" to obtain their country loadings. For all other countries, we translated their names using Google Translate.

Table 3: Selected guest countries in our confirmatory analysis

host language	guest countries
US-English	Peru, France, UK, Portugal, Brazil, Indonesia, Chile, Canada, Cuba, Denmark, Italy, China, Argentina, Spain, Japan, Sweden, Colombia, Netherlands, Norway, Egypt, Belgium, Switzerland, Guyana, Finland, Greece, Algeria, Australia, Germany, Philippines
UK-English	USA, Spain, France, Russia, Canada, Indonesia, Sweden, Netherlands, Belgium, Argentina, Chile, Greece, Guyana, China, Finland, Italy, Norway, Ghana, Colombia, Portugal, Brazil, Morocco, Denmark, Algeria, Thailand, Japan, Philippines, Belize, Malta, Egypt
French	Spain, Brazil, Britain, Indonesia, China, Portugal, Canada, Haiti, Italy, Japan, Denmark, Argentina, Cuba, Colombia, America, Egypt, Russia, Chile, Philippines, Peru, Switzerland, Mexico, Belize, Algeria, Netherlands, Norway, Sweden, Greece, Guyana, Belgium
German	USA, Belgium, Netherlands, France, Sweden, Spain, Finland, UK, Denmark, Chile, Portugal, Canada, Argentina, Japan, Norway, Italy, Nigeria, Ghana, Russia, Switzerland, Mexico, Uruguay, Romania, Greece, Algeria, Egypt, Bulgaria, Yugoslavia, Peru, Colombia
Italian	France, Spain, USA, Sweden, Belgium, Netherlands, Chile, UK, Greece, Switzerland, Russia, Portugal, Norway, Argentina, Brazil, Canada, Japan, Germany, Egypt, Finland, Uruguay, Australia, Yugoslavia, Romania, Denmark, Bulgaria, Tunisia, China, Morocco, Indonesia
Russian	Spain, France, UK, Sweden, Belgium, USA, Netherlands, China, Finland, Italy, Greece, Romania, Norway, Portugal, Germany, Egypt, Denmark, Japan, Switzerland, Yugoslavia, Australia, Iran, Brazil, Canada, Bulgaria, Chile, Algeria, Bermuda, Morocco, Tunisia

In our mixed-effect regressions, the units of observations are country-pairs. Because we run separate regressions for each host language, the observations are essentially the guest countries. Based on the graphs presented in the previous section, it looks like in some of the languages, there is some systematic spatial distortion around 2009, which is about the time point Google changed its source of books. To rule out the effect, we only included yearly observations of country-pairs from 1900 to 2009. Missing observations are dropped from the

dataset. For host languages with bootstrapped samples, we measured the friendliness and we-ness of each guest country in each sample and pooled all of the observations into a single dataset. For each regression, we included the random effects of all the 30 guest countries, all the observation-years, and all the yearly samples. Even though we controlled for the random effects of every sample, pooling all the samples into a single dataset may still not be the best use of the bootstrapped samples because every guest country is still observed 20 times every year in the pooled dataset. Their errors might still be correlated. In our future analysis, we might consider treating each sample as an imputed dataset in a multiple imputation, running a regression on every sample, and then pooling all the results together with corrected standard errors.

In our regressions, we would like to predict the friendliness and we-ness of guest countries based on war and trade variables. To account for unobserved heterogeneity of guest-countries, we used Δfriendliness$_t$ = friendliness$_t$ - friendliness$_{t-1}$ as well as Δwe-ness$_t$ as our dependent variables. The fixed-effect variables include war_ally (whether a guest country is an ally of the host country in a war), war_enemy (whether it is an enemy of the host country in a war), Δlog_gdp_ratio = $\Delta(\log(\text{gdp}_{\text{host}}) - \log(\text{gdp}_{\text{guest}}))$, Δlog_trade_12, which is the log of the yearly export volume from the host country to the guest country, and Δlog_trade_21, which is the log of the yearly import volume. For both the friend-enemy and we-they dimensions, We included their lagged terms and the other dimensional measure in the fixed effect as well.

Results from US English, UK English and French are reported in Table 4 and 5. War_enemy consistently has a negative effect on the perceived friendliness of guest countries. The effects are also consistent with our interpretation of the time-series graphs. Being in war with another country significantly reduces the perceived friendliness of that country. War_ally has a positive effect on friendliness in US and UK English. However, the effect sizes are much smaller than that of war_enemy. In French, the ally effect is not present probably due to France's complicated involvement in WWII. We-ness consistently has a positive effect

on friendliness, which suggests the two dimensions are somewhat correlated. The effect of trade on friendliness is less clear. In the UK, exporting more leads to less friendliness, and importing more leads to more friendliness. However, in French, the effect of export is in the opposite direction, and the effect of import is null. In the US, neither import nor export has an effect. it is worth noticing that the standard errors are smaller for UK and French simply because there are more observations. Given that the effect sizes are quite weak, it is safer to conclude that trade has no major impact on friendliness. The same is true for GDP ratio. The significant effect of the lag term suggests that the model may not adequately control for temporal auto-correlation. Including more lag terms might also help us reveal more patterns. The results do not reveal any causal direction although we implicitly assumed that it is more likely that the perceptions are influenced by external events rather than vice versa. On the other hand, Table 5 suggests that the we-ness of guest countries is correlated with neither war nor trade.

Regression results from German, Italian and Russian are reported in Table 6 and 7. Although some effects are significant in some regressions, the signs are not consistent across languages. it is also quite counter-intuitive to interpret some effects. For instance, in Russian, war_enemy leads to more friendliness. As we observed some strange properties in the time series plots, we would rather not further interpret the regression results in those languages.

Conclusion and Discussion

Our investment in constructing perception measures is rewarded with some return. The time series constructed from the US-English, UK-English, and French word-embedding models seem to be highly interpretable. Our interpretations are also consistent with the results of our confirmatory analysis. Some patterns are also highly consistent across these three languages. The friend-enemy measures are highly responsive to wars and perhaps some other major geopolitical events. Some countries, such as Canada and Thailand, are consistently

Table 4: Mixed-effect regression of friendliness in US English, UK English, and French

	Dependent variable:		
	Δfriendliness		
	(US)	(UK)	(French)
war_ally	0.006**	0.003***	0.001
	(0.003)	(0.001)	(0.001)
war_enemy	-0.013**	-0.010***	-0.005***
	(0.006)	(0.001)	(0.001)
$\Delta\text{friendliness}_{t-1}$	-0.456***	-0.482***	-0.479***
	(0.016)	(0.004)	(0.004)
Δwe-ness	0.039**	0.045***	0.045***
	(0.017)	(0.004)	(0.004)
Δlog_gdp_ratio	-0.003	-0.001	-0.002*
	(0.004)	(0.001)	(0.001)
Δlog_trade_12	0.0004	-0.001**	0.001***
	(0.0005)	(0.0003)	(0.0002)
Δlog_trade_21	0.0004	0.0005**	-0.0002
	(0.0005)	(0.0002)	(0.0002)
Constant	-0.00001	0.001	-0.00004
	(0.002)	(0.001)	(0.001)
Observations	2,878	53,563	53,469
Log Likelihood	6,078.754	106,051.200	110,892.300
Akaike Inf. Crit.	-12,133.510	-212,078.500	-221,760.600
Bayesian Inf. Crit.	-12,061.930	-211,971.800	-221,654.000
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 5: Mixed-effect regression of we-ness in US English, UK English, and French

	Dependent variable:		
		Δwe-ness	
	(US)	(UK)	(French)
war_ally	0.002	-0.002*	-0.003***
	(0.003)	(0.001)	(0.001)
war_enemy	-0.003	0.001	-0.006***
	(0.006)	(0.001)	(0.001)
Δwe-ness$_{t-1}$	-0.519***	-0.483***	-0.481***
	(0.016)	(0.004)	(0.004)
Δfriendliness	0.005	0.042***	0.042***
	(0.016)	(0.004)	(0.003)
Δlog_gdp_ratio	-0.003	0.001	-0.001*
	(0.004)	(0.001)	(0.001)
Δlog_trade_12	-0.0002	-0.0003	-0.0004***
	(0.0005)	(0.0003)	(0.0002)
Δlog_trade_21	-0.00002	0.0003*	0.0004***
	(0.0004)	(0.0002)	(0.0001)
Constant	0.0002	0.001	-0.0001
	(0.002)	(0.001)	(0.001)
Observations	2,878	53,563	53,469
Log Likelihood	6,252.854	108,371.200	115,264.600
Akaike Inf. Crit.	-12,481.710	-216,718.400	-230,505.300
Bayesian Inf. Crit.	-12,410.130	-216,611.700	-230,398.600
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 6: Mixed effect regression of friendliness in German, Italian and Russian

	Dependent variable:		
	Δfriendliness		
	(German)	(Italian)	(Russian)
war_ally	-0.003**	0.001	-0.003
	(0.001)	(0.001)	(0.012)
war_enemy	-0.0004	-0.010***	0.030***
	(0.001)	(0.002)	(0.005)
$\Delta\text{friendliness}_{t-1}$	-0.496***	-0.493***	-0.475***
	(0.004)	(0.004)	(0.004)
Δwe-ness	0.001	0.023***	0.130***
	(0.004)	(0.004)	(0.008)
Δlog_gdp_ratio	0.001	0.003***	-0.007***
	(0.001)	(0.001)	(0.002)
Δlog_trade_12	-0.00002	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0001)
Δlog_trade_21	0.0002**	0.002***	-0.0003**
	(0.0001)	(0.0003)	(0.0001)
Constant	0.0003	0.0003	-0.001
	(0.001)	(0.002)	(0.004)
Observations	55,251	50,932	39,292
Log Likelihood	108,939.000	99,703.730	61,944.600
Akaike Inf. Crit.	-217,854.100	-199,383.500	-123,865.200
Bayesian Inf. Crit.	-217,747.100	-199,277.400	-123,762.300
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 7: Mixed effect regression of we-ness in German, Italian, and Russian

	Dependent variable:		
		Δwe-ness	
	(German)	(Italian)	(Russian)
war_ally	-0.002*	-0.003**	-0.016**
	(0.001)	(0.001)	(0.007)
war_enemy	0.002**	0.001	-0.006**
	(0.001)	(0.002)	(0.003)
Δwe-ness$_{t-1}$	-0.498***	-0.494***	-0.487***
	(0.004)	(0.004)	(0.004)
Δfriendliness	-0.004	0.018***	0.037***
	(0.003)	(0.003)	(0.002)
Δlog_gdp_ratio	-0.002***	-0.001	0.001
	(0.001)	(0.001)	(0.001)
Δlog_trade_12	-0.0001	0.0002	-0.0001
	(0.0001)	(0.0002)	(0.0001)
Δlog_trade_21	0.0003***	-0.001**	-0.00002
	(0.0001)	(0.0003)	(0.0001)
Constant	0.0003	0.0002	0.002
	(0.001)	(0.001)	(0.002)
Observations	55,251	50,932	39,292
Log Likelihood	119,841.800	106,389.000	84,735.500
Akaike Inf. Crit.	-239,659.700	-212,754.000	-169,447.000
Bayesian Inf. Crit.	-239,552.700	-212,647.900	-169,344.000
Note:		*p<0.1; **p<0.05; ***p<0.01	

more friendly in the perceptions of these three Western language communities. The host countries' relationships with their former colonies dropped significantly when the colonies successfully declared their independence. There also seems to be some co-variation among all countries in the same region. Because we do not observe any global co-variation, we suspect that these regional co-variations are real, and there could be some high-level patterns beyond the level of nations. For instance, the perceived friendliness of Muslim countries seemed to all drop at the same time during the 1960s and 1970s. The Iranian Revolution does not seem to be a single coincident that led to the decline. Although we do not have access to the other end of the story, the clash of civilizations between the Western World and the Muslim World seams to have begun way before the end of the Cold War and perhaps even before the First Oil Crisis.

However, results in German, Italian and Russian are much less interpretable. Potentially, many reasons could help explain our failure. First of all, the corpus sizes of these three languages are much smaller. The strange phenomena we observed in their word-embedding spaces could simply be due to randomness in the yearly samples. If the corpus sizes are the cause, fixing the problem wouldn't be too hard. We do have more 5-grams in later years' corpora. For earlier years, we used all we have, but we can pool several year's 5-grams together to build larger corpora. Second, there could be severe selection bias in how the German and Italian corpora were selected. Pro-Fascist books might be banned after the world wars and did not get selected into Google Books. There also existed two Germanys during a long period of time. Authors in East and West Germany might not share the same linguistic dimensions. Results from that period might be misleading. Third, none of the factors listed above could easily explain the spatial distortion observed in the word embedding spaces. They could correspond to some real historical development. However given that we observed similar patterns in English and French corpora before and were able to correct the problem, the problems observed in the German, Italian and Russian spaces are more likely introduced by the word-embedding models. We will still need to understand

the inner working of the models better in order to fix the problems.

References

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A neural probabilistic language model." *Journal of machine learning research* 3 (Feb): 1137–1155.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in Neural Information Processing Systems* 29.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356 (6334): 183–186.

Dodds, Peter Sheridan, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, et al. 2015. "Human language reveals a universal positivity bias." *Proceedings of the National Academy of Sciences* 112 (8): 2389–2394.

Firth, J. 1957. "A Synopsis of Linguistic Theory 1930-1955." In *Studies in Linguistic Analysis.* Reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow. Philological Society, Oxford.

Fouquin, Michel, Jules Hugot, et al. 2016. *Two centuries of bilateral trade and gravity data: 1827-2014.* Technical report. Universidad Javeriana-BogotÃ!‘

Gries, Peter, Andrew Fox, Yiming Jing, Matthias Mader, Thomas J. Scotto, and Jason Reifler. 2020. "A new measure of the 'democratic peace': what country feeling thermometer data can teach us about the drivers of American and Western European foreign policy." *Political Research Exchange* 2 (1): 1716630.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky. 2016. "Cultural shift or linguistic drift? comparing two computational measures of semantic change." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing,* 2016:2116. NIH Public Access.

Harris, Zellig S. 1954. "Distributional structure." *Word* 10 (2-3): 146–162.

Junyan Jiang, Tianyang Xi, and Haojun Xie. 2020. "In the Shadows of Great Men: Leadership Turnovers and Power Dynamics in Autocracies." *Working Paper,* https://doi.org/https://dx.doi.org/10.2139/ssrn.3586255.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The geometry of culture: Analyzing the meanings of class through word Embeddings." *American Sociological Review* 84 (5): 905–949.

Lee, Suman, and Hyehyun Hong. 2012. "International public relations' influence on media coverage and public perceptions of foreign countries." Public Relations History, *Public Relations Review* 38 (3): 491–493.

Li, Xiaojun. 2021. "More than meets the eye: Understanding perceptions of China beyond the favorable–unfavorable dichotomy." *Studies in Comparative International Development* 56 (1): 68–86.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, null null, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781.*

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems,* 3111–3119.

Page, Benjamin I., and Robert Y. Shapiro. 1983. "Effects of public opinion on policy." *American Political Science Review* 77 (1): 175–190. https://doi.org/10.2307/1956018.

Pew Research Center. 2020. "Unfavorable Views of China Reach Historic Highs in Many Countries." https://www.pewresearch.org/global/2020/10/06/unfavorable-views-of-china-reach-historic-highs-in-many-countries/.

Rumelhart, David E, and Adele A Abrahamson. 1973. "A model for analogical reasoning." *Cognitive Psychology* 5 (1): 1–28.

Saussure, Ferdinand de. 2011. *Course in General Linguistics.* New York: Columbia University Press.

Schmitt, Carl. 1976. *The Concept of the Political.* New Brunswick, N.J.: Rutgers University Press.

Schneider, William. 1985. "Peace and strength: American public opinion on national security." In *The Public and Atlantic Defense,* 321–64. Rowman & Littlefield.

Tajfel, Henri, John C Turner, William G Austin, and Stephen Worchel. 1979. "An Integrative Theory of Intergroup Conflict." *Organizational identity: A reader* 56 (65): 9780203505984–16.

Methodological Appendix

Effect of Corpus Size

As shown in Figure 1, the corpus sizes of Google Books are dramatically different over years. In all the Western European languages except for Spanish, years during the Second World War are associated with the least corpus sizes.

It turns out that corpus size also has a huge impact on word-embedding spaces. In my preliminary stage, I trained three rounds of models:

- In the first round, I sampled 1% of the ngrams from every year and trained yearly models.
- In the second round, I used the same corpus as used in round 1 but restricted the vocab size to the same number of most frequent words (n = 11,000) every year. In other words, the model ignored words outside the yearly frequency ranges during training so that the learned vector spaces would have the same number of rows and columns.
- In the third round, I drew a 5% weighted sample from all year's 5-grams. I calculated n_y, the total number of 5-grams in each year y and assigned $1/n_y$ as the weight for all 5-grams in that year. Basically, I over-sampled from smaller years and down-sampled from bigger years so that every year has roughly the same amount of 5-grams drawn into the sample. Then, I trained word-embedding models from the yearly samples.

After each round, I projected some pre-chosen identities to the "friend-enemy" and "we-they" dimensions in each year's word-embedding space using Kozlowski, Taddy, and Evans (2019)'s method. The results are in Figure 14 and 15. Initially, by look at the results from round 1, we thought there were some dramatic spatial distortion during war-time. We were puzzled by the appearance that all countries became more friendly during war time. However, after round 2, some war-time patterns disappear. And after round 3, there is no

discernible global war-time patterns that can be observed from the time series. Figure 16-18 are the cosine similarities between "ami (friend)," "ennemi (enemy)," "nous (we)," "ils (they)" as well as "il (he)," "elle (she)," "riches (rich)" and "pauvres (poor)" and all major countries in the world. Each line is the similarity of a country to a chosen word. It seems that all patterns that seemingly happened during war time get eliminated in round 3. When the corpus size is small, word vectors tend to be all closer to each other.

Examining the Qualities of the Embeddings

In computational content analysis, word-embedding is usually used as a unsupervised method for revealing interesting patterns. There hasn't been a clear standard in terms of telling whether a model is good or bad. One way to evaluate how well a model fits the data is to stick with the log likelihood function (which is just the negative of its loss function) and examine whether it is maximized. However, people almost never do it, and the feature is not even properly implemented in the latest version of *gensim*, which is the most widely used package for training Mikolov's skipgram models.

There is a reason why the log likelihood function is somewhat obsolete. The skip-gram models have two implementations for training its word vectors: hierarchical softmax and negative sampling. Neither implementation ever uses the original softmax function. In negative sampling, which is slightly more popular due to its efficiency, the model even completely alters the prediction task from a multi-class problem to a binary classification problem. What it seeks to model is no longer the conditional probability $p(w_{i+j}|w_i)$ but

$$p(y = 1|w_i, w_j) = \text{sigmoid}(\vec{u}_{w_i}^\top \vec{v}_{w_j}) \tag{2}$$

, where y is a binary variable indicating whether w_i and w_j co-occur in the training corpus. The sigmoid function is just the link function of a logistic regression. What it does is given a pair of words, take the dot product of the input and output vector representations of

Figure 14: Dimensional loadings of selected Western nations in the friend-enemy dimension from round 1 (top) to 3 (bottom)

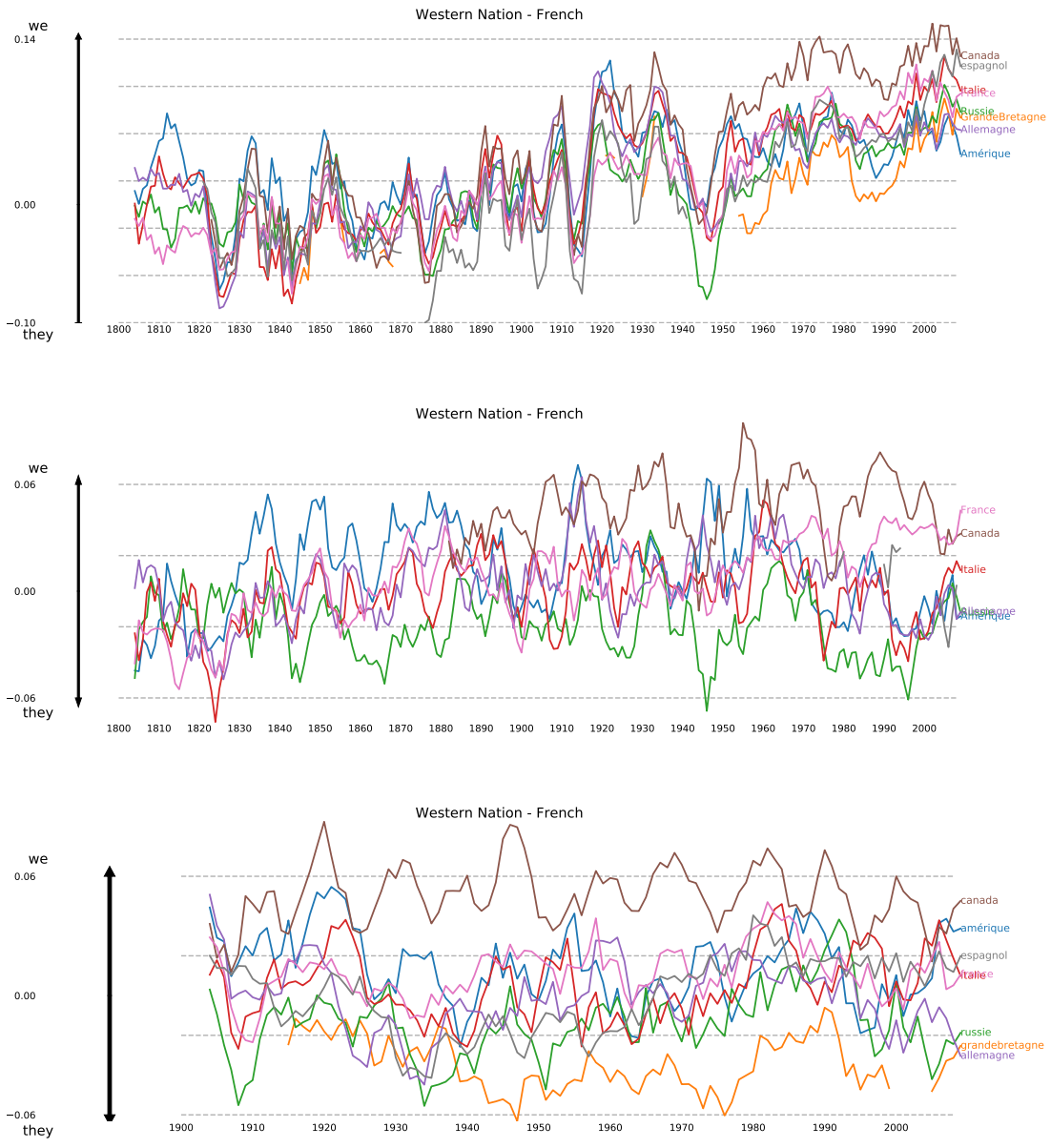

Figure 15: Dimensional loadings of selected Western nations in the we-they dimension from round 1 (top) to 3 (bottom)

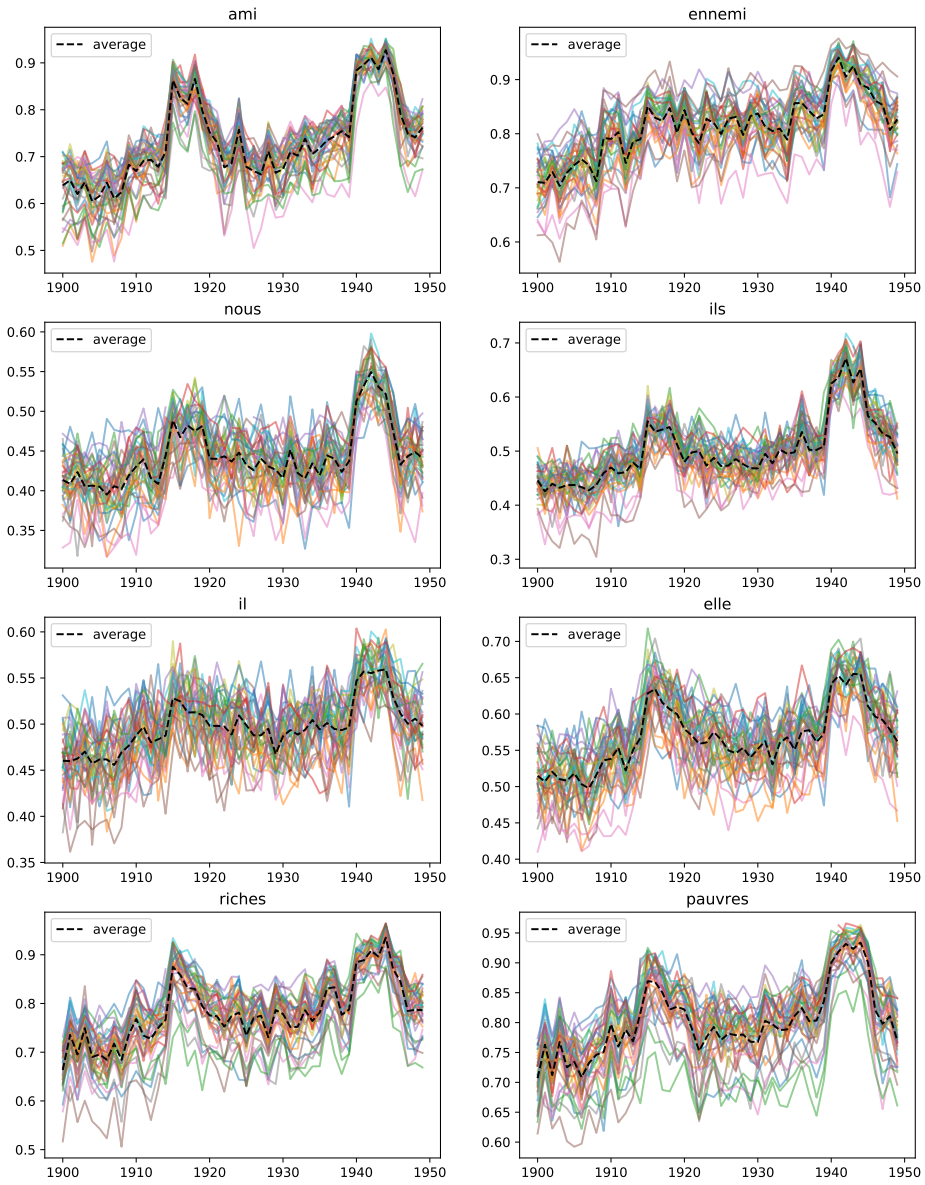

Figure 16: Country-identity similarities, round 1

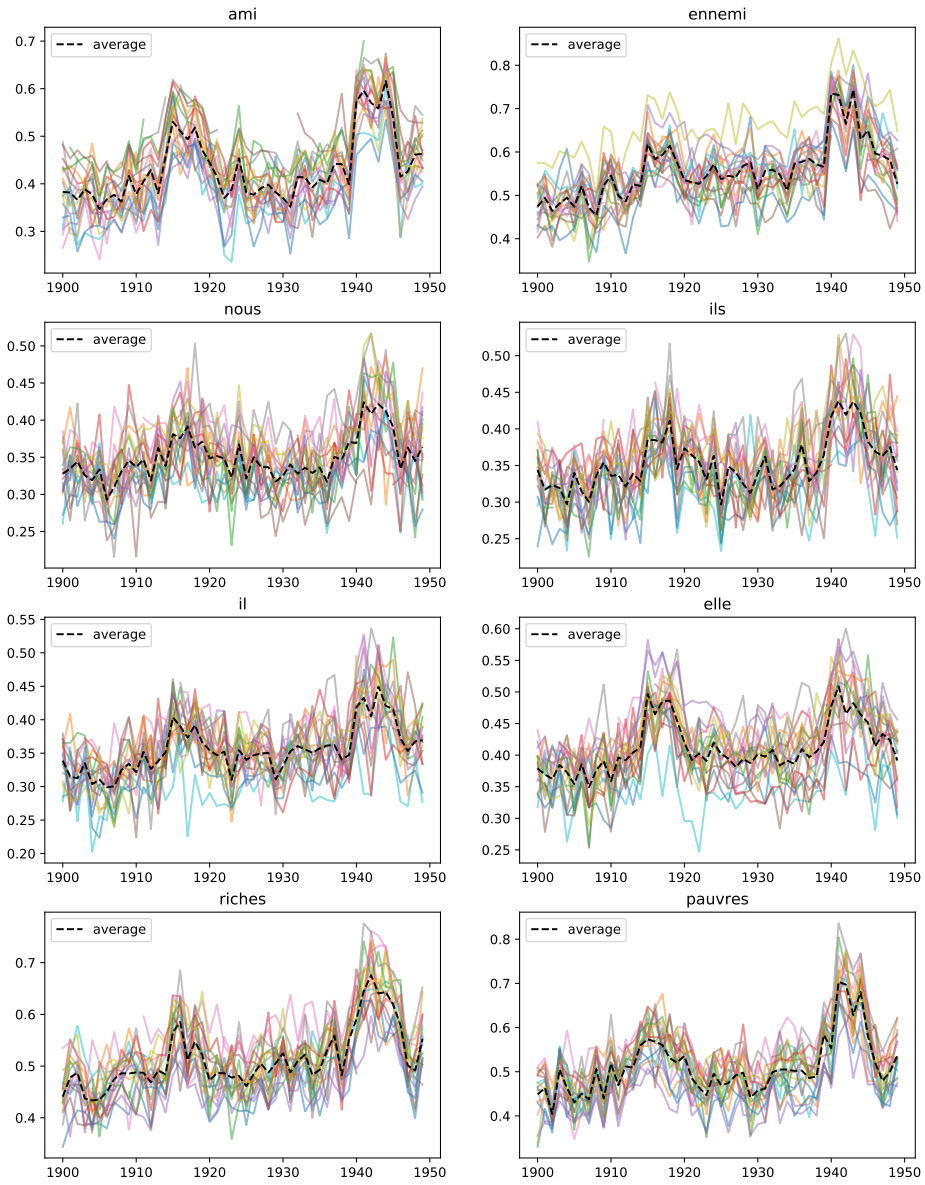

Figure 17: Country-identity similarities, round 2

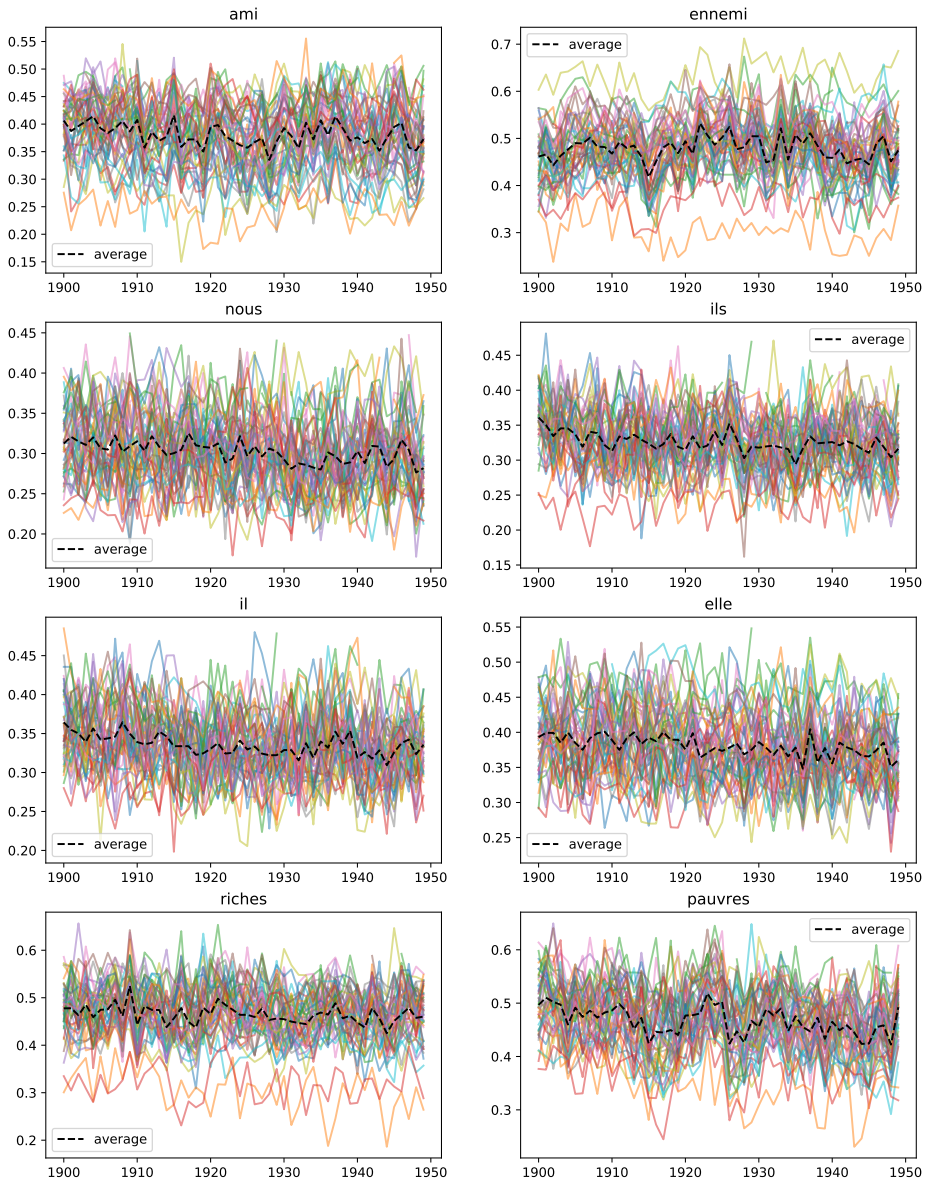

Figure 18: Country-identity similarities, round 3

the words into a logistic regression to predict whether the pair would actually occur in the training corpus. It has trivial solutions if the length of the input and output vectors are arbitrarily large. So for every positive instance, it adds to the objective function a fixed number of negative instances where the output words are drawn from a noise distribution. Given any pair of words, the logistic regression tells whether the word pair would occur as a true pair in the corpus or simply as a noise pair by chance.

As Mikolov, Sutskever, et al. (2013) explain in their paper, there is no guarantee that vectors trained from negative sampling would maximize the softmax function. However, that is also not the goal of their model. Their goal is to obtain high-quality single-layer vector representation of words. The prediction task is only a means for obtaining word embeddings. One way to simply tell whether a model is good or bad is to use it to obtain similar words that occur in similar contexts, and the results usually intuitively make sense. However, Mikolov also claims that their word embedding models are able to learn some global cultural/linguistic directions such as in the example of "king - queen $\approx$ man - woman." They found that their models can perform well in solving such analogy tests and constructed a large test corpus (as shown below) including 5 categories of semantic tests and 9 categories of syntactic tests. Their analogy tests have become the standard test for evaluating the quality of word-embedding models. In this study, we also use the Mikolov Analogy Tests to evaluate the performance of our models. We found that negative sampling indeed outperforms hierarchical softmax in analogy tests. Therefore, we use negative sampling for training our models.

capital-common-countries: Athens Greece Baghdad Iraq

capital-world: Abuja Nigeria Accra Ghana

currency: Algeria dinar Angola kwanza

city-in-state: Chicago Illinois Houston Texas

family: boy girl brother sister

gram1-adjective-to-adverb: amazing amazingly apparent apparently

gram2-opposite: acceptable unacceptable aware unaware

gram3-comparative: bad worse big bigger

gram4-superlative: bad worst big biggest

gram5-present-participle: code coding dance dancing

gram6-nationality-adjective: Albania Albanian Argentina Argentinean

gram7-past-tense: dancing danced decreasing decreased

gram8-plural: banana bananas bird birds

gram9-plural-verbs: decrease decreases describe describes

In our experimentation, I came up with two ways to pre-process our ngram data. The original purpose of these tweaks was to save computing resources. In machine learning, there is a general consensus that the more data there is, the more accurate the model trained from the data is. Our results based on analogy tests suggest that adding more variety definitely helps, but reducing repeating information also helps.

Downsampling

The approach I used for speeding up training is downsampling frequent 5-grams. Just like distributions of words, 5-grams also follow a very skewed distribution with a few 5-grams appearing way more times than other 5-grams. Cutting repeated information could help reduce training time. The skipgram model has a built-in downsampling feature that randomly discard high-frequency words during training. Mikolov, Sutskever, et al. (2013) even claim that their downsampling approach could help increase performance in analogy tests in some cases. I tried it. Although it helps to speed up training, it in general worsens performance in analogy tests. I came up with an alternative downsampling scheme. For 5-grams that appear multiple times in a year's corpus, I applied a scaling function to their frequencies such that

$$n'_{ngram_i} = {n_{ngram_i}}^{\beta} \tag{3}$$

, where n_{ngram_i} is the count of $ngram_i$ in a yearly sample, and n'_{ngram_i} is the count I used in the training corpus. I varied β from 0 to 1. When $\beta = 1$, there is no scaling. When $\beta = 0$, all counts information is discarded, and every 5-gram is treated as occurring once. When β is between 0 and 1, it heavily downsamples the most frequent 5-gram and lightly downsamples rare 5-grams. The intuition is that in negative sampling, all the model cares is whether a pair of words would co-occur or not in the corpus, repeatedly feeding the model with the same pairs would probably not help with training and could potentially lead to overfitting.

The results of our experiments are in Figure 19. Overall, there seems to be convincing evidence to suggest that scaling helps. Scale = 0.33 outperforms all other scales in terms of total accuracy. Running the same test on different years of corpus yields similar results. We do not understand why the cubic root is the magic transformation. The breakdown results also suggests that there is no scale that consistently wins in all categories.

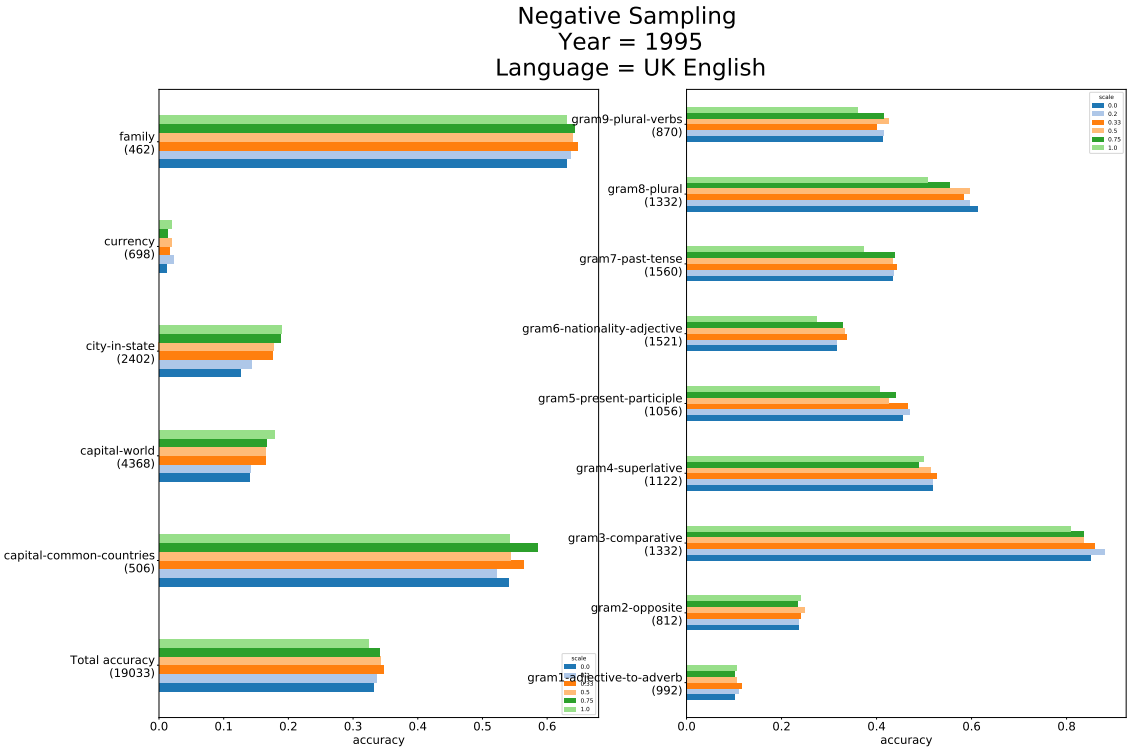

Figure 19: Breakdown analogy test results in categories

Patterns become clearly after all the analogy tests are divided into frequency bins. I calculated the minimum and median of the frequencies of the words in each analogy test. Then I took log base 10 and rounded them into the nearest integers. The bars in Figure 20 show the average accuracy in each frequency bin. Overall, tests that contain high-frequency words are easier. Scaling also helps in most frequency ranges except for the lowest ones.

Based on the results of the experiments shown in this section, I decide to apply Scale = 0.33 in prepossessing our training data.

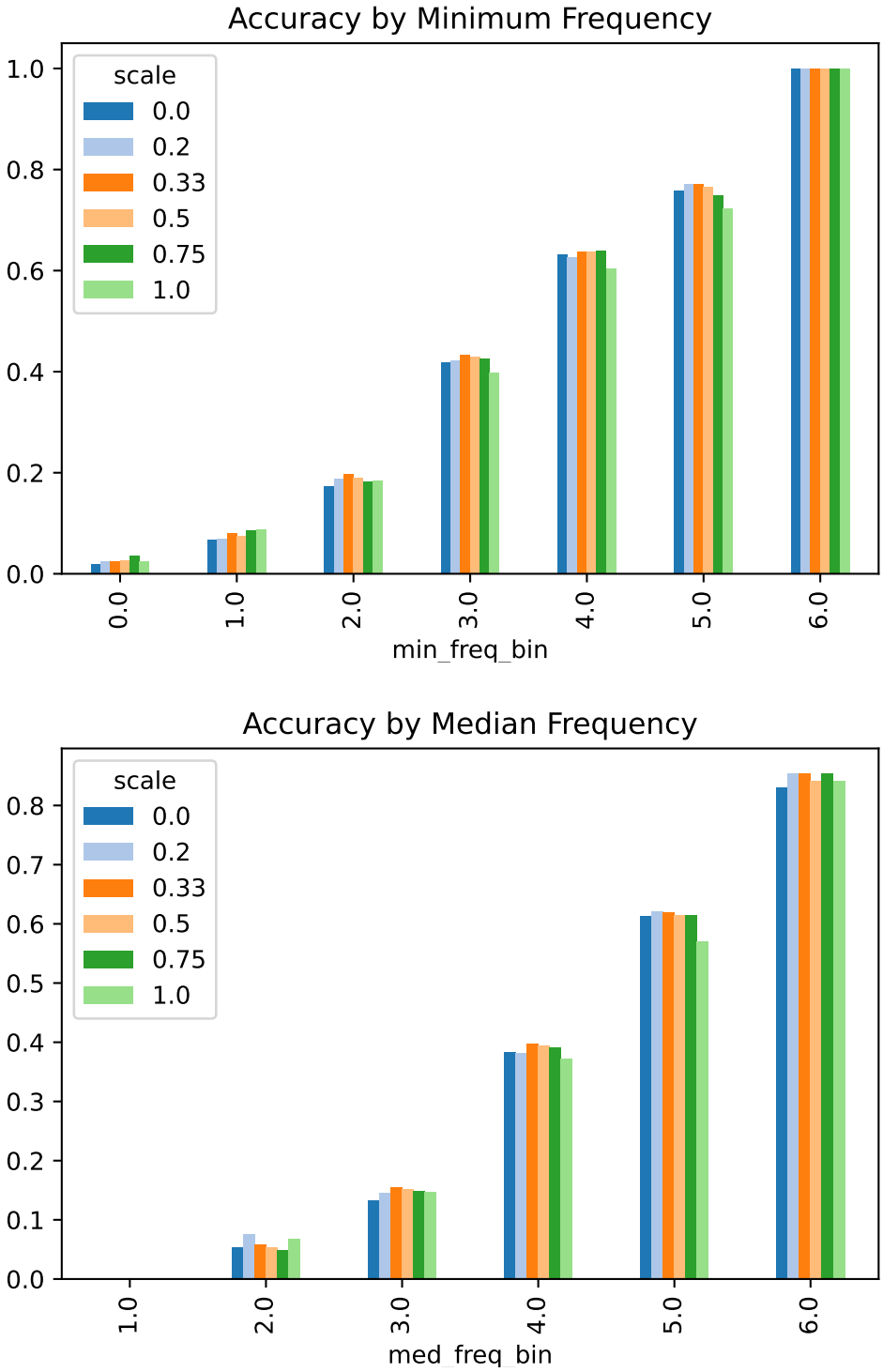

Figure 20: Breakdown analogy test results in frequency bins