FREQUENCY AND FACT: LEARNING ABOUT THE WORLD THROUGH A CORPUS OF WORLD-ENGLISHES

Bryor Snefjella

FREQUENCY AND FACT: LEARNING ABOUT THE WORLD THROUGH A CORPUS OF WORLD-ENGLISHES

by BRYOR SNEFJELLA, B.A.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

McMaster University MASTER OF SCIENCE (2014) Hamilton, Ontario (Linguistics and Languages)

TITLE: FREQUENCY AND FACT: LEARNING ABOUT THE WORLD THROUGH A CORPUS OF WORLD-ENGLISHES

Bryor Snefjella, B.A. (Queen's University) SUPERVISOR: Professor Victor Kuperman

NUMBER OF PAGES: 59

DECLARATION OF STUDENT CONTRIBUTION

Chapter 2 was submitted to a journal, and is now under revision, with Victor Kuperman as first author and (myself) Bryor Snefjella as second author, conducted during the fall 2013 semester. My contribution was portions of the literature review, analysis of the historical relationships to the main finding, and some revisions following submission. Chapter 3 is entirely my work.

Abstract

Two studies are presented, linking word-frequency information within the Global Corpus of Web-based English to real world facts. The first study concerns how patterns of the use of place names reflect geospatial and geopolitical relationships of English-speaking nations. The second study concerns how the emotional connotation of words before place names reflects general well-being in that place. Taken together, these studies demonstrate that the surface structure of language, as embodied in word frequencies, is a useful source of information about the real world.

Acknowledgements

Thanks to the Sherman Centre for Digital Scholarship.
Thanks to Louis Schmidt for getting on board at the last minute, in an unfamiliar field.
Thanks to Victor Kuperman for his constant support and generosity.
Thanks to Anna Moro for her help at all steps of the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Scholars within the social sciences are increasingly selecting large repositories of texts, often derived from the internet, as sources of data. This shift towards "'big data" offers scholars the ability to perform analyses that, prior to the development of the internet, would have simply been impossible. Often, this "big data" is in the form of natural language texts which are processed with computers, but the process of linking patterns in texts to real-world phenomena is not straightforward. Linguistics is no exception to this trend towards larger and larger collections of texts. Linguistics is still a relatively young field, and there is still disagreement between linguists as to the appropriate sort of data for a linguist to be concerned with. So, from the outset, it is important to define precisely what sort of linguistics we are engaged in.

First, we should distinguish between E-language (linguistic behaviour) and I-language (the system of mental representations generating this behaviour). Linguists still have disagreements about which of these the linguist is supposed to explicate. Over the last 20 to 30 years, the focus on I-language that the Chomskian revolution initiated has been relaxing, and linguists now generally enjoy — though not entirely without controversy — a variety of sources of data, not restricted to introspective judgments of grammaticality (for recent examples of the debate regarding introspective judgements, see Gibson and Fedorenko (2010, 2013) and Culicover and Jackendoff (2010); Sprouse (2011)). Indeed, sociolinguistics, pho-

netics, psycholinguistics, and corpus linguistics are now disciplines in their own rights, with well developed and diverse methodologies and terminologies. This thesis is concerned with E-language, that is, with describing linguistic behaviour as it occurs.

The methodology employed to explore and describe E-language is corpus linguistics. In modern corpus linguistics, the most prototypical corpus (Gilquin and Gries, 2009; Gries, 2009) is one that is a machine readable, representative and balanced body of texts that is produced in a natural communicative setting. At its most basic, corpus linguistics is the counting of character strings. To become intellectually rigorous, a corpus linguist has to take these occurrence counts and use them to explain or describe language. This means that usually corpus linguists are interested in the distributional properties of character strings, since simple occurrence frequencies are rarely informative by themselves.

Corpus linguistics has been criticized on the basis of only being capable of sampling a portion of language. An oft quoted statement from Chomsky (1962, cited from McEnery (2001) reads:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the distribution [based on it] would be no more than a mere list.

As stated above most modern corpora are specifically and carefully designed, whenever possible, to be balanced and representative. The criticism that a corpus cannot ever be representative because of language's infinite nature is too dire. Scientists are always working within limited data, and insuring that conclusions drawn from this limited data generalize is part of the task of any scientist. A corpus linguist working within a finite selection of language is no different from a pharmaceutical scientist testing out the safety of a new drug within a limited group of participants. It should also be said here that since corpus linguistics is a methodology, it can and has been used fruitfully within a generative framework (e.g. generative phonology in Goldsmith and Riggle (2012), generative syntax in Aarts (1992); Haegeman (1987)).

A corpus linguist must be careful when asserting that word frequencies are indicative of some fact about the world. Taking word frequencies from a corpus and using them to make some inference about the world is not a trivial process. (Hunston, 2002, p.23) points out that the sequence *left-handed* occurs twice as often as *right-handed* in the Bank of English corpus. From frequencies alone, it may be appealing to therefore conclude that there are more left-handed than right-handed individuals. Obviously, this is not the case, and the greater frequency of *left-handed* has some other explanation (e.g. that it is salient because it is less common). Careful consideration of why a given string receives the frequency it does is critical. Within this thesis, there are two projects, both involving connecting language use as embodied in word-frequencies and their related statistics to phenomena in the real world.

The corpus used for this thesis is the Global Corpus of Web Based English (GloWbE) (Davies, M., 2013). GloWbE is composed of English-language text gathered from the internet. The composition of GloWbE is about an equal split between blogs and regular websites, with text assigned to one of twenty English-speaking countries depending on website URLs. GloWbE is certainly machine readable, but representativeness and balance deserve some commentary. The percentage of residents of these twenty countries with internet access does vary considerably, so GloWbE is best said to represent those individuals with internet access in these countries. There is also the possibility that websites with country specific URLs tend to be more formal or official (Davies, personal communication), which the inclusion of blogs is supposed to alleviate. Blogs can be about anything, and written by anyone with internet access. GloWbE is also a naturalistic communicative setting, though of course a written one.

Because we are able to link variations in GloWbE data with specific communities, we are doing a form of sociolinguistics. Sociolinguistics is generally concerned with the role of variation between speakers of a language. Labov (1996) summarizes the key insight of sociolinguistics as follows:

> The central finding of sociolinguistics is that the community is the stable

> and systematic unit, and that the behavior of individuals cannot be interpreted without a prior knowledge of the community pattern. At the level of the individual, language varies substantially, at the level of the community, language is stable.

In regards to any apparent antagonism between sociolinguists and those of a generative persuasion, it is important to note that the presence of variation between speakers does not invalidate the study of I-language. The purpose here is to explain the variations we observe. I believe the words of Weinreich et al. (1968) are an excellent description of the general goals of this thesis:

> The solution... lies in the direction of breaking down the identification of structuredness with homogeneity. The key to a rational conception of language change - indeed, of language itself - is the possibility of describing orderly differentiation in a language serving a community.

The particular linguistic features I am examining can be seen in the tradition of examining regional variations between communities.

There are two studies in this thesis, each focusing on a different topic. The first concerns the distribution of words denoting nations, and whether the frequencies of these words within a corpus bear a relationship to the geographical locations of nations. The second study concerns the distributions of words bearing different emotional connotations around words denoting nations. I believe a quote from Chambers and Schilling-Estes (2013) will cement the idea that we are in fact doing sociolinguistics.

> Studying language variation proceeds mainly by observing language use in natural social settings [a corpus of web-text] and categorizing the linguistic variants [emotional connotations of words, distributions of place-names], according to their social distribution [which nation they belong to].

To reiterate, because we are examining the behavioural output of the linguistic system, we are concerned with E-language, not I-language. Because we are counting character strings

within a machine readable body of language, intended to be balanced and representative, our methodology is corpus linguistics. Because we are examining orderly variation in language linked to specific communities, we are doing a type of sociolinguistics. The particular variations in language use we are observing are the affective connotations of words before place names and the frequency distributions of place names.

Chapter 2

National Spatial Relationships are Encoded in Text

2.1 Abstract

Distributional patterns of language use have recently been shown to encode spatial relations. Louwerse [Louwerse, M. M. (2010). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science, 3(2)*:273–302] poses two challenges to language-statistical approaches: how simple the computational algorithm can be and how little perceptual grounding is necessary to extract spatial semantics from texts. This paper addresses both challenges by demonstrating that comparison of the frequency distribution of references that one country makes to other countries of the world can be used to can approximate that country's cognitive map of the world, and that cognitive maps of countries show more similarity if the countries are geographically or historically related. We use relative entropy to evaluate semantic distances between countries. That geographic information can be distilled from basic statistical units of language (word occurrences, not co-occurrences) using a simple algorithm with only a modicum of perceptual grounding suggests that language encodes spatial relations even more robustly and more transparently than proposed by symbolic-cum-embodied cognition accounts.

2.2 Introduction

Previous decades have seen an upsurge of research on how cognitive maps – mental representations for spatial relations, distances and directions – are formed, stored and retrieved in a variety of cognitive tasks. The primary source of experience that informs cognitive maps is perceptual information, including objects and distances in the physical world. I know my computer screen is in front of my office wall because I have learned that when one object occludes another it is in front. A second source is the iconic representations of objects (e.g. maps, schemes, or diagrams; Richardson, Montello, & Heggarty, 1999). I have learned that Papua New Guinea is north of Australia through encountering maps of the world, despite that I have never been to Papua New Guinea. Cognitive maps can also benefit from symbolic information conveyed through texts, such as explicit descriptions of spatial relations (e.g., Avraamides, 2003; Taylor & Tversky, 1992). For example, I can help someone navigate my apartment by telling them that the kitchen is to their left when they enter the front door. Recently, an intriguing addition has been made to this repertoire of sources, as distributional statistics of language use have been shown to implicitly encode spatial information. More specifically, patterns of word co-occurrences with names of spatial objects in text corpora were analyzed with sophisticated statistical techniques to estimate semantic distances between spatial objects both on a large scale (cities in China, Middle East, the US, the UK, and the fictional world of Middle Earth) and a small scale (body parts). These distances showed moderate to strong correlations with physical distances between these spatial objects in the embodied world (Davies, C., 2013; Louwerse & Benesh, 2012; Louwerse, Hutchinson, & Cai, 2012; Louwerse, & Zwaan, 2009; Tillman, Datla, Hutchinson, & Louwerse, 2012). These findings corroborated the long-standing (though not uncontroversial) notion that language use and language structure are optimized for the purpose of communication, and that aspects of meaning can be gleaned from the distributional statistics of linguistic units (for an extensive discussion see Louwerse, 2010, and references therein). For instance, words that are relatively similar in meaning will tend to

either co-occur more often in a stretch of text (first-order co-occurrences), or be found in similar contexts more often than semantically distant words (higher-order co-occurrences). In the words of Firth (1957: 11): "You shall know a word by the company it keeps".

The apparent iconicity of language surface structure with respect to spatial relations raises the question of the interplay between symbolic and embodied sources of cognitive maps across tasks. Several theoretical accounts postulate that the perceptual and symbolic sources are not mutually exclusive and, while only perceptual experience is the necessary condition for forming and using cognitive maps, language users may rely on both (cf. Goldstone & Rogosky, 2002). The account that this paper comments on, the Symbol Interdependency Hypothesis (Louwerse, 2010), further claims that the relative contributions of perceptual and symbolic information to cognitive maps are largely contingent on the nature of the task. While a minimum of perceptual grounding for symbols is required, language is argued to offer a "communicative short-cut for language users" (Louwerse, 2010, page 7) in that it allows for bootstrapping encoded spatial relations from relations between linguistic symbols.

The robust findings that spatial relations are retrievable from textual sources, even for such large-scale objects as country maps, give two challenges as to the boundary conditions of the Symbol Interdependency Hypothesis (Louwerse, 2010). One is identifying how little perceptual information is minimally required for conceptual grounding of linguistic symbols. Louwerse et al. (2012, page 699) define the minimum of perceptual grounding warranted for their task as follows:

> [...] if a language user knows the location of the city [Ürümqi], and knows only that the other Chinese words are Chinese city names, the language user can bootstrap the geographical locations of these other cities on a country map of China.

One of our goals is finding out whether spatially meaningful information can be obtained with even more limited perceptual experience.

Another challenge is finding the simplest, psychologically plausible algorithm capable

of gleaning spatial relations from observable, surface-level distributional statistics of language. Most demonstrations of the validity of a symbolic approach to spatial relations (see references above) made use of Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Landauer et al., 1998), an advanced statistical technique that exploits higher-order word co-occurrence statistics and thus bases its evaluation of semantic distances between linguistic units on latent non-observable semantic structures in a high dimensional space. As noted by Louwerse (2010), the demonstrable utility of LSA may suggest that it is the algorithm itself, rather than language structure, that enables the extraction of meaning from texts. Identifying an algorithm that directly utilizes observable patterns of language – i.e. first-order co-occurrences of words or, even more ambitiously, mere word occurrences – is then critical for the proposed role of language as a tool for, and the product of, information encoding. It is equally critical for validity of the symbolic approach to spatial cognition. Several studies have reported a similar performance of LSA and other statistical techniques that operate on first-order (rather than higher-order) word co-occurrences in a variety of meaning-extraction tasks (e.g., Louwerse & Zwaan, 2009; Louwerse, 2010; Recchia & Jones, 2009), including the relationship between place names in web-pages (Liu et al., 2014), thus corroborating the role of language surface structure as a vehicle of meaning. Yet establishing the lower bound of what humans can extract from language through the simplest of algorithms is still an open challenge.

Before moving into our hypothesis, it is important to note that whether we should call our text-based spatial relations "cognitive maps" is not clear. Our method gives less detailed spatial representations than Louwerse and Zwaan (2009); Louwerse and Benesh (2012); Louwerse et al. (2012). It may be better to call our results "distance inferences" or "very simple cognitive maps." We leave more precise classification of our results to later work.

This paper proposes an alternative approach to extracting large-scale geographic maps from textual information. The approach demands less perceptual grounding (without discarding it completely) and uses an algorithm based on simpler units of statistical informa-

tion about language, i.e. word occurrences, not co-occurrences. Specifically, we argue that a cognitive map of the world that a given country possesses can be characterized by the frequency distribution of the references that country makes to other countries. Moreover, we hypothesize that countries that are close geographically or historically will have more similar cognitive maps, as reflected in similar frequency distributions of their references to the outside world. We further argue that distances can be fruitfully used for identifying geopolitically and historically meaningful clusters of countries, solely based on textual information. To anticipate the results, this method gives rise to non-random clusters of countries that show an above-chance correspondence to regions of the world.

2.3 Methods

We extracted demonyms, or adjectives denoting nations (e.g. *Russian, Chinese, Irish*), from the 1.9 billion-token Corpus of Global Web-Based English (Davies, M., 2013). The corpus is based on 1.8 million English-language web pages collected from 20 different English-speaking countries in December 2012: we dub these *source* countries. The countries were selected to represent both the Inner Circle of World Englishes, i.e. countries in which "English is used as a mother tongue for all administrative and most social purposes", and the Outer Circle, i.e. countries where "English is regarded as a second language and where much of the administration of the country is achieved through the medium of English" (Crystal, 2003): see Table 2.1 for the list of source countries, their regions and affiliation with circles of World English. Attribution of web pages to countries in the corpus was based on their domain names: the number of tokens varied by country, with a maximum of 388 million tokens (Great Britain) and a minimum of 35 million tokens (Tanzania).

A comprehensive list of demonyms was obtained from Wikipedia (2013). All demonyms were assigned the code of the country they denoted: we dubbed these countries *target countries*. The International Organization for Standardization's (2013) two-letter country codes were used. Each search for a demonym in the GloWbE corpus yielded a vector

of 20 frequency counts, reflecting occurrences of that demonym in each source country's subcorpora. In a number of cases, multiple demonyms referred to a single country: such demonyms were labeled with the same country code and the respective frequency counts were summed per source. For example, *afghan* has a frequency of 194 and *afghani* a frequency of 4 in Singaporean websites: both demonyms were mapped to Afghanistan (code AF) and the total frequency of 198 was considered in further calculations. In the case of Hong Kong, there is not a well-established demonym, so the string *Hong Kong* and its respective frequencies in 20 subcorpora were used instead. Furthermore, an inspection of references to *Korean* revealed that unless further qualified they mostly implied *South Korean*. We attributed references to *Korean* not preceded by the adjective *North* to South Korea; and the rest to North Korea. The demonym *Congolese* is used for both the Democratic Republic of the Congo (CD), and The Republic of the Congo (CG), all uses of this demonym were treated as referring to the Democratic Republic of the Congo, the larger of the two countries.

The full list of demonyms represented 163 target countries. Each was associated with a 20-element frequency vector quantifying how often the 20 source countries referenced target countries. We removed 12 (or 7%) countries which had fewer than 400 occurrences of their associated adjectives in GloWbE. The resulting matrix of frequencies was the size of 20 (the number of source countries) x 151 (the number of target countries).

2.4 Results and discussion

Our analysis of text-based cognitive maps hinges on the hypothesis that the cognitive map of a source country can be adequately approximated by how frequently those countries reference other countries. That is, we hypothesize that the frequency distributions of these references are indicative of the strengths of geographical, political, cultural or historical associations that a source country has with nations of the world. An inspection of the 20 frequency distributions revealed certain regularities that supported our intuition. First, the most frequent reference for the vast majority of source countries was to itself (except

Tanzania and Bangladesh who both referred to *British* more frequently than to *Tanzanian* and *Bangladeshi* respectively). Second, most source countries preferentially mentioned their neighbors and only rarely demonyms from other major regions of the world (details not reported in the interest of space).

Our stronger claim was that distance estimates based on our frequency distributions will be closer if countries that show a stronger geo-political, cultural or historical proximity. In other words, we assumed that frequency distributions of demonyms enable both a sufficient discrimination between countries and sufficient sensitivity to detect similarities between them. To test this claim, we estimated the distance between each pair of source countries, or more precisely between each pair of 151-element vectors of demonym occurrences attested in the corpora associated with the two source countries. Information theory provides a useful metric that quantifies a divergence between two probability distributions p and q: i.e. relative entropy, or the Kullback-Leibler divergence. Relative entropy is a non-negative quantity that amounts to zero if the two probability distributions under comparison are identical. To operationalize the divergence in probability distributions of two countries p and q, we used the symmetric formulation of Jensen-Shannon divergence D, defined as:

$$D = -1/2(\Sigma_x p(x) \log m(x) + \Sigma_x p(x) \log p(x)) - 1/2(\Sigma_x q(x) \log m(x) + \Sigma_x q(x) \log q(x)),$$

where $m = 1/2(p + q)$ and each distribution has x elements.

Relative entropy can be construed as a measure of inefficiency (or information loss) of assuming that the true distribution is q while it is p. This is the inefficiency of approximating one country's frequency distribution by another country's frequency distribution. We applied relative entropy as a distance metric to evaluate similarities between source countries and, separately, between target countries.

2.4.1 Source countries

Relative entropy operates on pairs of probability distributions. To obtain those for each pair of source countries p and q, we first removed frequencies of references to countries p

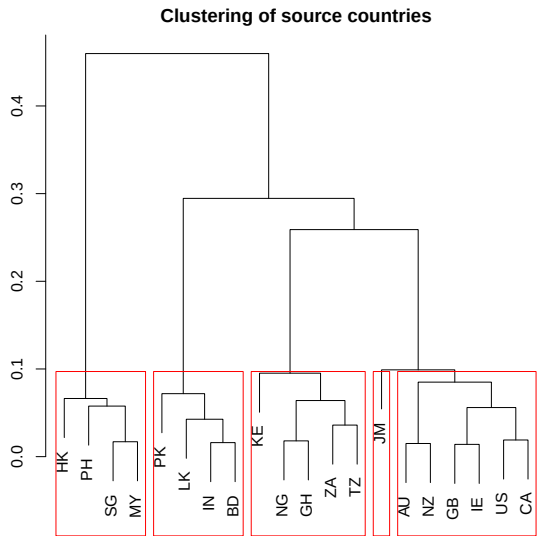

Figure 2.1: Hierarchical cluster analysis of source countries
Hierarchical cluster analysis of source countries by probability distributions of their references to target countries, with the Euclidean distance metric and the Ward agglomeration method. For country codes see Table 2.1. Rectangles indicate clusters used to estimate the similarity with geo-historical regions.

and q from both frequency distributions: that is, we removed self-references and references to the counterpart country. This was done to ensure that the comparison is not distorted by the self-image of either source country nor by their mutual relationship. The resulting pair of frequency distributions contained 151 - 2 = 149 elements each. Frequency distributions were re-normed by dividing by the sum frequency to obtain probability distributions. We further calculated the Jensen-Shannon divergence for each pair of source countries based on those probability distributions to yield a 20 x 20 distance matrix.

We further conducted hierarchical cluster analysis to identify structure in the text-derived data, using the Euclidean distance metric and Ward's linkage criterion. The dendrogram of the cluster analysis for 20 sources countries is reported as Figure 2.1, while Figure 2.2 plots five resulting clusters on the world map. The clusters bear a remarkable resemblance to either geo-political or historical unions and divisions in the world. Clustered together are African countries (Kenya, Nigeria, Ghana, South Africa and Tanzania), countries of the Indian subcontinent (Pakistan, Sri Lanka, India and Bangladesh), East Asian

countries (Hong Kong, Philippines, Singapore, and Malaysia), and countries of the Inner Circle, that is, Great Britain and its neighbors (Ireland) and former colonies (Australia, New Zealand, USA and Canada); Jamaica is a cluster of its own. Accurate clusters are found even at the lowest levels of agglomeration: Great Britain and Ireland are closer to each other than to other Inner Circle countries, and so are USA and Canada, and Australia and New Zealand, true to the geographical fact. It is worth noting that the cluster analysis sheds light on the sociolinguistic debate about the status of South Africa as a member of the Inner or the Outer Circle of World Englishes (Bruthiaux, 2003; Crystal, 2003, cf.). In the present analysis, South Africa clusters with its geographic neighbors representing the Outer Circle rather than with other countries of the Inner Circle.

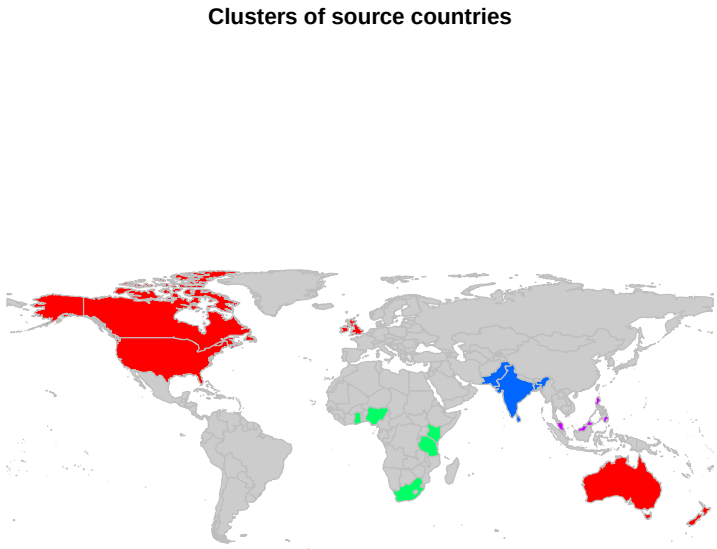

Figure 2.2: World map with text-based clusters of source countries.

While the qualitative match of the similarity pattern in Figure 2.1 to geo-political and historical regions in the English-speaking world appears strong, we tested whether compa-

rable results could result from a random clustering of source countries. We assigned 20 source countries to a total of 5 "embodied" geo-historical regions based on our knowledge of the world: the regions are reported in Table 2.1. One region was geographically diverse and comprised countries of the Inner Circle of World Englishes, i.e. Great Britain, Ireland, US, Canada, Australia and New Zealand. Remaining four regions were strictly geographical and incorporated five African countries, four East Asian countries, 4 countries of the Indian subcontinent, and, separately, Jamaica. To match the number of regions, five distinct "symbolic" text-derived clusters were identified in the dendrogram (see rectangles in Figure 2.1), and each source country was attributed to a cluster. While we required the number of clusters and regions to match, the number and identity of countries in each cluster could differ from the one observed in any of the regions.

Country	CountryCode	Region	Circle	SettlementYear	DistanceToGB	PercentL1
Australia	AU	1	Inner	1770	3	79.00
Canada	CA	1	Inner	1763	2	63.29
Ireland	IE	1	Inner	1607	1	97.40
New Zealand	NZ	1	Inner	1790	3	95.76
United Kingdom	GB	1	Inner	0	0	97.50
USA	US	1	Inner	1607	2	80.92
Jamaica	JM	2	Outer	1665	4	97.56
Bangladesh	BD	3	Outer	1858	6	0.00
India	IN	3	Outer	1858	6	0.03
Pakistan	PK	3	Outer	1858	6	0.00
Sri Lanka	LK	3	Outer	1796	6	0.05
Ghana	GH	4	Outer	1874	5	0.00
Kenya	KE	4	Outer	1920	5	0.00
Nigeria	NG	4	Outer	1861	5	0.00
South Africa	ZA	4	Outer	1806	5	8.49
Tanzania	TZ	4	Outer	1890	5	0.00
Hong Kong	HK	5	Outer	1842	7	2.08
Malaysia	MY	5	Outer	1874	7	1.71
Philippines	PH	5	Outer	1898	7	0.02
Singapore	SG	5	Outer	1826	7	8.14

Table 2.1: Source Countries

Source countries: names, country codes and geo-historical regions to which each country was attributed. Also reported are the distance from Great Britain in the dendrogram, and the percent of L1 speakers of English.

For each pair of source countries a similarity metric was calculated. If both countries belonged to the same United Nations defined region and the same cluster, or both belonged

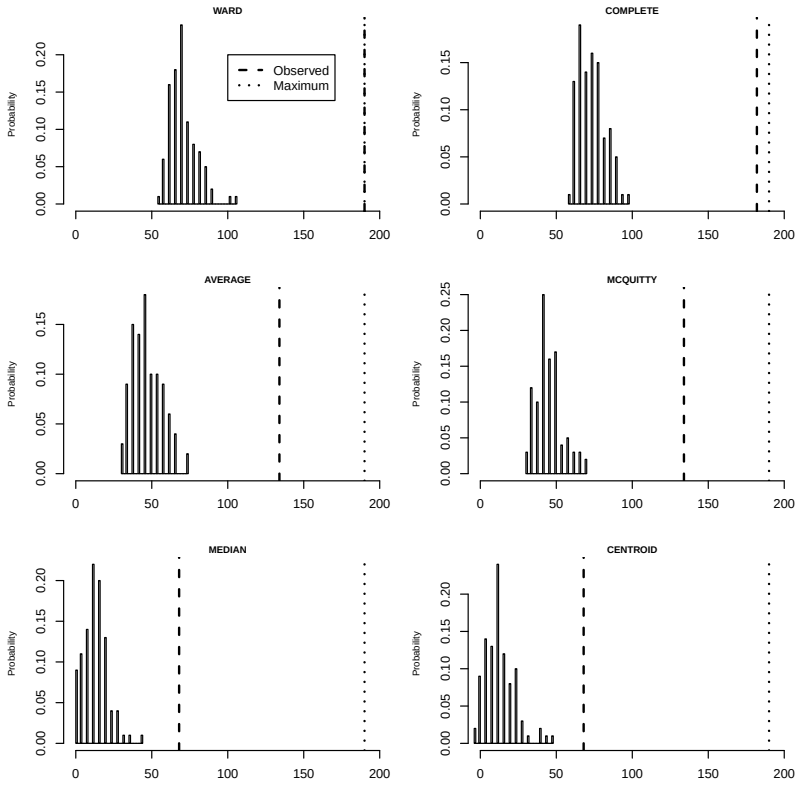

Figure 2.3: Simulations
Distribution of values of the similarity metric for 1,000 random cluster assignments for source countries, along with the theoretic maximum (text-based clustering perfectly matches geographic clustering, dotted line) and the observed value of the similarity metric (dashed line). Panels represent different clustering methods: the theoretic maximum and the observed value overlap for the Ward method.

to different regions and different clusters (true positives and true negatives), the count that served as a similarity metric was increased by one. The count was decreased by one if countries belonged to the same/different region but to different/same clusters (false positives and false negatives). When all pairwise comparisons were completed, the resulting count was a measure of the similarity of our cluster analysis to actual geo-historical regions. The upper boundary, or the theoretic maximum, of the similarity count is achieved if text-based clusters are identical to geo-historical regions: this optimal case would yield the similarity count of 190. This identity between clusters and regions was indeed observed in the cluster analysis using the Ward's minimum variance method, which yielded the maximum similarity count

of 190. We also ran 1,000 simulations in which 20 source countries were randomly assigned to 5 clusters. The number of countries in each cluster (though not the country identities) were preserved as determined in Figure 2.1. The similarity counts obtained in simulations are summarized in a histogram in the top left panel of Figure 2.3, shown along with the (here, overlapping) dotted line representing the theoretic maximum count and the dashed line representing the similarity count obtained from the cluster analysis. The maximum similarity count obtained in 1,000 simulations was 106. As the histogram demonstrates, the convergence between the geo-historical regions and clusters obtained from text sources using the Ward clustering method is perfect, and very unlikely to arise by chance ($p < 0.001$). Remaining panels of Figure 2.3 represent several other clustering methods, reporting the theoretic maximum of the similarity count, the observed similarity count in textual data, and distributions of similarity counts in 1000 simulations. While the accuracy of clustering methods varies (with the maximum similarity score shown by the Ward method and the minimum by the median and centroid methods), all sets of text-based clusters approximate the geo-political and historical structure of the world significantly better than could be expected by chance (all ps < 0.001). Non-randomness of similarity patterns of text-derived distances was further confirmed with varying definitions of geo-historical regions, different methods of cluster analysis, and different requirements for randomization (all ps < 0.001).

Besides observing that most clusters were geographically motivated, we found that the ordering of clusters might also be *historically* motivated. Specifically, there was an apparent association between the succession of splits in the dendrogram and the history of world colonization spreading out of Great Britain: the higher the split in the dendrogram, the more distant the clusters are. With Great Britain as a focal country, the lowest splits in the dendrogram obtained with the Ward's method (Figure 2.1) roughly corresponded to the territories that became British colonies first (e.g. the split between Great Britain and Ireland) and the highest splits indicated territories where the significant British presence was established relatively late (Great Britain vs East Asia cluster). To formally test this intuition, we ranked distances from Great Britain to all other source countries in our den-

drogram (rank 1: Ireland, rank 2: US and Canada, rank 3: Australia and New Zealand; rank 4: Jamaica; rank 5: African countries; rank 6: countries of the Indian sub-continent; rank 7: East Asian countries). These ranks were correlated with the year that a British settlement was established in respective countries, as documented in Crystal (2003)[1]. The Spearman's correlation between the distance rank from Great Britain and the time of colonization was strong and positive: $\rho = 0.51$, p = 0.027: (groups of) countries that were colonized by Great Britain earlier tend to have frequency distributions of demonyms that are closer to that of Great Britain. This provides circumstantial evidence that the ordering of countries within clusters, and that of clusters among other clusters is non-random, and is in fact is co-determined by the history of the spread of the English language and the colonizing nation that spoke it. Given present data, we cannot rule out a possibility of a third, mediating, factor that underlies the observed correlation. For instance, countries that were colonized earlier tend to have a larger number of speakers of English as their first language (L1): data on the linguistic profile of countries was obtained from Crystal (2003). The country's distance from Great Britain in the cluster analysis correlated negatively with the percent of L1 speakers of English in the country (ρ = -0.6, p = 0.007) and positively with the percent of L2 speakers of English ($\rho = 0.63$, p = 0.004). We leave more accurate historical characterization of similarities between country-specific demonym frequency distributions to future research.

2.4.2 Target countries

So far we have interpreted frequency distributions of references to target countries as an approximation of the cognitive maps of source countries. These distributions, however, can be fruitfully used to focus on target countries and evaluate, for instance, whether France and Germany would be considered similar by the English-speaking world. Using the same procedure as described above, we calculated the Jenson-Shannon divergence as a dissimilar-

[1]Chronology of colonization, and particularly, the precise dating of the British settlements or colonies in respective countries is a contentious issue in the historical literature. In our discussion, we selected only one of the possible settlement dates for each country, and we acknowledge that substantial variation in those estimates is possible depending on the chosen historical account.

ity measure for each pair of countries from the 131 countries that we had frequency data for and that were not source countries. We further identified 20 major clusters in the hierarchical cluster analysis, to match the number of geographic subregions defined by the United Nations Statistics Division (2013). Figure 2.4 plots the resulting clusters against the world map. The data patterns are less clear cut then one we observed for source countries. Some clusters were overly inclusive or geographically unmotivated. We connect this poorer performance of the method to the fact that source countries were characterized by a distribution of 149 frequency counts, while the distribution for target countries only contained 20 elements, thus yielding a weaker discriminatory power. We note however that many text-derived clusters *were* geographically plausible: cf. Middle Eastern countries and Afghanistan; China and Far East; Mexico and Central America; Argentina and neighboring countries; as well as regions in Central and South Africa. Critically, our simulations confirmed again that the similarity between the clustering observed in Figure 2.4 and geographic regions defined by the United Nations Statistics Division (2013) is very unlikely to be due to chance ($p < 0.001$; histograms not shown).

2.5 General Discussion

This study points to a new source of geographic information that is implicitly encoded in language statistics, i.e. distributional patterns of referring to countries of the world as attested in text corpora. We show that (a) this symbolic information can serve as a way of approximating the cognitive maps of countries, and that these distance estimates obtained from multiple textual sources (b) are smaller if those countries are related to one other geographically or historically and (c) form clusters that are a good-to-excellent approximation of geo-political regions of the world. The major point of novelty in our approach departs is that we do not use word-occurrences of any kind unlike methods favored in prior computational work on embodied language, using either first or higher order word co-occurrences. We use the occurrence of a critical word (i.e. demonym, or nationality-word)

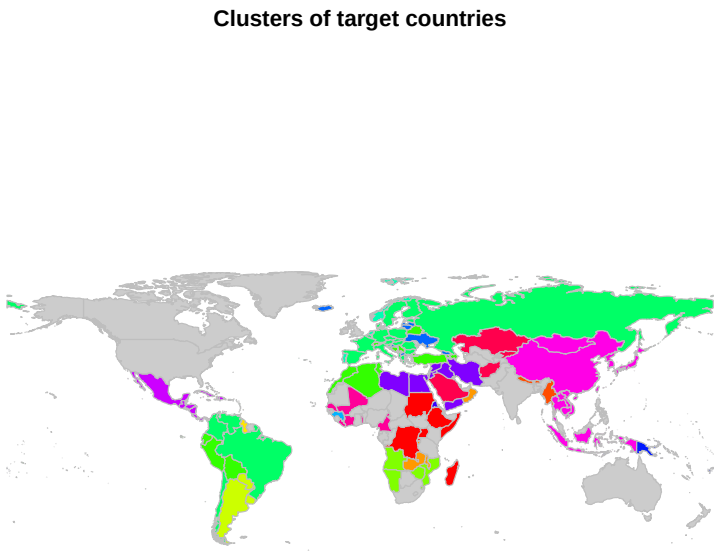

Figure 2.4: World map with text-based clusters of target countries.

as a unit of distributional information, rather than the co-occurrence statistics of words neighboring that critical word. We also use a simpler method of obtaining frequency (or probability) vectors to represent geographic objects than the one used in LSA (see below). Furthermore, we employ a different distance metric to quantify the distance between vectors in a multidimensional space: the symmetric version of relative entropy vs the cosine value. Finally, our approach hinges on the use of multiple geographically-defined same-language corpora, whereas a single corpus suffices LSA applications. The present method of evaluating embodied relations through symbolic relations pushes further the boundary conditions of symbolic approaches summarized in Louwerse (2010) and Louwerse et al. (2012): i.e. the minimal amount of perceptual information necessary for conceptual grounding of symbols, and simplicity and psychological plausibility of the statistical algorithm. In what follows we spell out our contribution and discuss implications of our findings for theories of embodied

and symbolic representations in language.

2.5.1 Perceptual grounding

A requirement that we pose, which is arguably not of a perceptual nature, is that language users are aware that distributional statistics of demonyms comes from different sources, each source representing a different geographic entity, e.g. a country. Again, no knowledge of what these countries are and where any of them is located is obligatory. To place these minimal conditions in the venerable tradition of the Chinese Room argument (Louwerse et al., 2012; Searle et al., 1980), we invite the reader to imagine a person who is brought to identical hotel rooms in 20 different countries, and has no other knowledge of her whereabouts except for knowing that these rooms are in different (unspecified) countries. Suppose further that the only stream of information that the person accesses in each room are the English-language reading materials produced in this country: the materials are stripped of the country's identity. Sensitivity to frequencies with which countries of the world are mentioned in those materials would enable the person to identify which rooms are likely to be situated in geographically or historically related countries, and whether or not moving to another room meant moving to a different geo-political region. In other words, this sensitivity is enough to give rise to intuitions such as: "they sure talk more about Africa and less about Far East in this place than in the previous two: now I must have crossed the ocean".

The fact that meaningful geographic patterns can be obtained with a smaller modicum of perceptual grounding than Louwerse and colleagues allow for in their studies (Louwerse & Benesh, 2012; Louwerse, Hutchinson, & Cai, 2012; Louwerse, & Zwaan, 2009) follows directly from a simpler nature of our task. We merely evaluate which geographic objects tend to form coherent larger units, rather than calculate and compare parametric distances between geographic objects in the symbolic and embodied worlds. It is likely that the structural complexity of spatial relations that one can infer from symbolic representations depends on richness of perceptual experiences that those representations are rooted in.

We believe, however, that the present study is a valuable illustration of how scant these perceptual experiences can be – amounting to knowing that certain strings denote countries, and are used more or less often – to give rise to simple yet non-random generalizations about the structure of the world.

2.5.2 Statistical algorithms

Our use of relative entropy as a distance metric substantially simplifies the method of extracting frequency vectors representing an object, and evaluating semantic distances between those vectors. First, the basic unit of the distributional information that the algorithm operates on is an occurrence of a word in a corpus: it is a zero-order co-occurrence and not a first- or higher-order co-occurrence of words which are operational units in LSA and most similar techniques (cf. Hyperspace Analogue to Language (Lund & Burgess, 1996), HiDEX (Shaoul & Westbury, 2006)). To reiterate, a vector representing a source country merely consists of frequency counts of demonyms occurring in the corpus produced by that country: no windows are defined around the demonyms (unless one considers an entire corpus a window) and no neighbor words are considered. This utility of occurrence counts, for at least some computational tasks, has profound theoretic repercussions: we discuss these below.

Second, once the frequency vectors are obtained and a trivial operation is performed to convert them to probability vectors, the calculation of relative entropy can performed in less than five lines of program code. This stands in stark contrast to the considerable computational complexity of LSA, which makes use – among other steps – of singular value decomposition or similar dimension reduction techniques due to the sparseness of co-occurrence matrices: for a detailed explication of the LSA method see e.g. Landauer & Dumais (1997). We relegate to future research a direct comparison between relative entropy as a computational algorithm and LSA (or any other technique that uses the first-order or higher-order word co-occurrences). Here, we confine ourselves to advocating relative entropy and its use of word occurrences as a candidate for the lower-bound algorithm that can extract meaning from the most basic units of language statistics using the simplest

means (Louwerse, 2010).

Importantly, a statistical algorithm - however advantageous it is in simplicity and efficiency - needs to be justified in terms of psychological plausibility and interpretability: for discussion of plausibility of LSA see Landauer et al. (1998), Louwerse (2010). As one of the basic statements of information theory (Shannon, 1948), relative entropy is also in use in information retrieval, sentence parsers, machine translation, and other computational-linguistic applications (cf. e.g., Manning & Schuetze, 1999). Moreover, a range of studies shows that sensitivity to differences in probability distributions, gauged by relative entropy, is indeed characteristic of language users and demonstrably affects language processing. Behavioral effects of relative entropy have been observed in studies exploring probability distributions associated with morphological case paradigms (Balling and Baayen, 2012; Kuperman et al., 2010; Milin et al., 2009a,b) and continuations of sentences given the present context (Hale, 2001; Levy, 2008), among others. Besides its relevance for psychological processes, relative entropy is also readily interpretable in terms of processing efficiency: it measures the loss of information that comes approximating the true probability distribution (e.g. one specifying one country's cognitive map) by a different probability distribution (e.g. another country's cognitive map). We argue that relative entropy stands the tests of computational simplicity, psychological plausibility and interpretability.

2.6 Conclusion

To sum up, the spatial patterns reported here are obtained using an uncomplicated set of mathematical operations on word occurrences. Our nationality-words or demonyms do not inherit any perceptual information from a surrounding linguistic context. The words need keep no company for the algorithm to work, and for the cognitive maps to emerge. Differences in the amount of occurrences between symbols from one corpus to another supply sufficient information to make inferences about those symbols, as long as one knows a little bit about those symbols to begin with. In this case, that "little bit" is that our symbols are

demonyms. Taken together, these findings point to the distributional patterns of language surface structure as a faithful representation of spatial relations in a more compelling way than prior demonstrations. They also rule out the possibility that the retrieval of geographic patterns from corpora is due to a particularly sophisticated meaning extraction algorithm that captures latent semantic representations in language. The present study subjects the contribution of language to a strong test and observes that – even under stricter boundary conditions – it cannot be fully ascribed to either the influence of embodied information or specifics of the mathematical apparatus. The substantial role of language, with an undeniable impact of perceptual information, lends support to both theories of symbolic cognition and to accounts that tie together embodied and symbolic spatial representations.

Chapter 3

Patterns in Affective Connotative Meaning at a Global Scale

3.1 Abstract

Using the GloWbE corpus of world Englishes, we perform a simple but psychologically plausible analysis of 20 English-speaking countries' sentiments towards around 130 countries of the world. We present evidence that affective information in language, occurring in the context of a place name, is indicative of the general well-being of people in that place, correlating with generally accepted indexes of well-being, including the Human Development Index, life expectancy at birth, and GDP per capita. We also show that affective information in adjectives occurring before place names corresponds to the relationship between the text's country of origin and the country that text references, correlating with public polling data about international relations. Together, these findings highlight that the affective connotative meaning of language reflects how we conceive of others, and other places.

3.2 Introduction

An increasing body of evidence calls into question the notion that cognition and emotion are separable (for reviews see (Pessoa, 2008; Phelps, 2006)). Although emotion may seem an odd interest for a language researcher, there is mounting evidence that language and emotion are also interrelated. Behavioural and neurophysiological experiments have shown that the emotional connotations of words affect experimental results. There is also considerable interest in using computers to extract sentiments from texts. Texts, the primary tool for the corpus linguist, are not just repositories of grammatical markers, constructions, and orthographic sequences; they also contain information about the distribution of affective information in naturalistic language use. The aim of this paper is to combine both theories of emotion with psychological reality and sentiment analysis of a corpus, to show that dimensions of emotion that are relevant in psycholinguistic experimentation can provide additional information that less psychologically valid sentiment analysis methods would miss.

The distinction between denotation and connotation is meant to capture that although words may refer to the same entity, they do not imply the same things about that entity. A common English example is that of words for olfactory experiences, where *smell* is neutral/negative, *stink* is negative, and *odour* is neutral. Or take words for combustion, such as *fire*, *inferno*, *blaze*, where the latter two imply more intense combustion. Our aim in this study is to examine the distribution of connotative meaning, without any particular reference to denotation. In particular, we are interested in emotional component of connoative meaning.

We approach emotion, in this paper, using a dimensional theory. Dimensional theories of emotion treat emotional states as a combination of continuous variables (for a review see (Mauss and Robinson, 2009)). There are three generally accepted dimensions to emotion, under dimensional theories; valence, arousal, and dominance. Valence refers to the pleasantness of a stimulus, while arousal refers to the intensity of emotion evoked by the stimulus,

and dominance refers to being in control or being controlled by the stimulus. So, anger is a composition of unpleasant, excited feelings, along with a desire to approach the stimulus, whereas fear is a composition of an unpleasant and excited feelings, along with a desire to withdraw from the stimulus. Many sentiment analysis methods include only valence. This paper will show that arousal can also be informative.

The relationship between language and emotion can be approached two ways, either that emotion affects language, or that language reflects emotion. That emotion affects language has been demonstrated with a variety of experimental methods. Numerous neurophysiological experiments have demonstrated the effects of emotion on event-related potentials during language processing studies (Kissler et al., 2009; Herbert et al., 2008; Kissler et al., 2007; Scott et al., 2009; Bayer et al., 2012; Ortigue et al., 2004; Bernat et al., 2001). Behavioural experiments also shown effects based on the emotional connotation of words; Herbert et al. (2006) showed that startle responses (blinking) are increased when subjects are exposed to emotionally arousing words. Indeed, response times in lexical decision tasks are responsive to the valence of a word (Kousta et al., 2009), and there is also evidence that exciting words tend to be recognized faster than calming words (Estes and Adelman, 2008). However, the precise contribution and structure of valence and arousal in word recognition remains controversial (Larsen et al., 2008; Kuperman et al., 2014).

That language reflects emotion has been shown by linking distributional patterns of language use to known phenomena. The field of sentiment analysis has expanded in recent years due to increased size and availability of corpora, the rise of the internet, and increasing computational power (for reviews see (Pang and Lee, 2008; Liu, 2012)). Sentiment analysis methods vary depending on the type of information the analyst wishes to extract, the operational definitions of sentiment, and psychological plausibility/sophistication of the algorithm. Often, these methods treat valence as binary (a word or phrase is either positive or negative), and do not take arousal into account at all (the aforementioned review papers contain no mentions of arousal). However, dimensional theories of emotion have been previously employed in sentiment analysis (eg. Naveed et al. (2011); Brooke (2009)). Previous

work has shown that sentiment gathered from the microblogging website Twitter is capable of replicating data gleaned through traditional public polling. For example, O'Connor et al. (2010) found that the positivity of tweets containing the words *job, jobs, economy* positively correlated with the Consumer Confidence Index, a poll reflecting general public belief in the health of the economy. Indeed, there has been a large amount of interest in using Twitter to predict a wide variety of phenomena: the Dow Jones Industrial Average (Bollen et al., 2011), influenza outbreaks (Paul and Dredze, 2011), and political elections (Tumasjan et al., 2011). Though our study utilizes a static corpus, it is similar to Twitter in that it contains geographical information about the text, in this case its national origin. The method we employ in the paper is computationally very simple but psychologically plausible, and we show that valuable information would be missed if arousal was not taken into consideration.

For the study we utilized Davies' (2013) Global Corpus of Web-Based English (GloWbE), a 1.9 billion word corpus composed of web-text drawn from twenty English speaking countries. Given the global scale of this corpus, we chose relationships between nations as our object of interest in this study. In particular, we explored the affective qualities of adjectives preceding demonyms, i.e. words referring to nationality (*affluent Afghani, burly Belorussian* etc.). We took these qualities to reflect the attitude of the country of origin towards the country the demonym refers to. To anticipate the results, we find that the overall valence, arousal, and frequency of references to a country by the English speaking world reflects the general well-being of people in that country. Moreover, the overall valence in text originating in a single English speaking country, referring to other countries, correlates with public opinion polls conducted in that country. Countries in these polls that were said to have a more positive influence on the world tended to be discussed with more positive language in GloWbE.

3.3 Methods

The GloWbE corpus is composed of English language web-text gathered from twenty English-speaking countries. The text is attributed to one of these countries on the basis of URL, and is composed of an approximately equal mix of blogs and other types of websites. Subcorpora ranged in size from 35 million (Tanzania) to 387 million words (United Kingdom). A comprehensive list of demonyms was acquired from Wikipedia (2013). The corpus was queried for instances of these demonyms, immediately preceded by words part-of-speech tagged as adjectives. In the case of Hong Kong, there is not a well-established demonym, so the string *Hong Kong* and its respective frequencies in the 20 sub-corpora were used instead. Furthermore, an inspection of references to *Korean* revealed that unless further qualified they mostly implied *South Korean*. We attributed references to *Korean* not preceded by the adjective *North* to South Korea; and the rest to North Korea. The demonym *Congolese* is used for both the Democratic Republic on the Congo, and The Republic of the Congo, all uses of this demonym were treated as referring to the Democratic Republic of the Congo, the larger of the two countries.

To extract sentiments from GloWbE, we utilized a lexicon of words rated along three generally accepted emotional dimensions: valence, arousal, and dominance. This lexicon contains 13,915 English lemmas rated on a scale from 1 to 9 on each of these three emotional dimensions (Warriner et al., 2013). Words in this lexicon were rated by those with an American IP address and stated they were first language speakers of English. Speakers of different varieties of English may have different emotional connotations for some words than those contain in this lexicon. Nonetheless, we find interesting patterns in connotative meanings both treating the English speaking world as a whole, and within each GloWbE country's sub-corpus. A total of 367720 adjective + demonym tokens were extracted. We removed tokens that originated in the same place they referred to (*American* from an American website), and that did not match a word in our emotion lexicon. An informal inspection of these discarded adjectives showed that they were mostly composed of low-frequency ad-

jectives (eg. *majoritarian*) or hyphenated constructions (eg. *hyper-aggressive.*) We removed 18 countries with 30 or fewer tokens, as these had too few words associated with them to be a reliable indication of the sentiments of the English-speaking world. We conducted two analyses, the first treating the English speaking world as a whole, the second examining the sentiments of English speaking countries individually.

3.3.1 The English Speaking World as a Whole

Our first analysis concerned the aggregate sentiments of the English speaking world towards all countries of the world. All 20 English speaking countries' sub-corpora were included in this analysis, and no distinction was made between texts based on country of origin.

For each demonym, we calculated a mean valence and mean arousal of the adjectives that preceded it, weighting the valence and arousal scores of each adjective by their frequency of occurrence before that demonym. This weighting ensured that the emotionality of adjectives used with higher frequency in the text is a stronger determiner of the overall score. This answers the question "overall, is this place discussed with more high valence/low valence or high arousal/low arousal language?"

The critical task of sentiment analysis is to identify language that carries sentiments from language that does not. Intuitively, not every occurence of a positive word is necessarily indicative of a positive opinion. One example from the GloWbE data is the sequence *free syrian*, which occurs with high frequency. Although *free* is a high valence word, its occurences before *syrian* do not represent a positive evaluation of Syria; the high frequency of *free syrian* is due to it being part of the name of a Syrian organization, the Free Syrian Army. Sentiment analysis frequently uses measures of the strength of association between positive and negative words and an item of interest (Liu, 2012, p. 28-29). In regards to the GloWbE data, if a positive word, such as *Traditional*, is used frequently before all demonyms, does that word represent a positive evaluation of all those demonyms? It is possible that demonyms simply attract the word *traditional*, as in *traditional Indian food* or *traditional American values.* In this case, counting each occurrence of *traditional* as a

positive evaluation by the English-speaking world might make many places appear to be more positively evaluated than they really are. Thus, we need a tool that can find which particular adjective + demonym sequences we should treat as carrying the sentiments of the English-speaking world and which we should not.

In corpus linguistics, there are a wide variety of association measures employed to determine the strength of association between linguistic units (Pecina (2010) gives 82 such measures.) These association measures are used for a wide variety of purposes aiming to discover an above-chance level of association between linguistic units that share some context. To name just a few applications, association scores have been used to identify complex lexical items such as phrasal verbs *eg. put off, got up* (Church and Hanks, 1990), in collostructural analysis (Gries and Stefanowitsch, 2004; Stefanowitsch and Gries, 2003), and analyzing stereotyping in language (Stubbs, 1996). Returning to our example of *traditional*, we can use association scores to determine if *traditional* has an above-chance level of occuring with some demonyms, or if *traditional* occurs with demonyms in an essentially random fashion (and thus we can discount it as indicating a positive evaluation by the English speaking world.)

We opted for ΔP, an association score derived from research into associative learning (Ellis, 2006). This choice was due to the curiosity of the researcher, and because ΔP ranges between -1 and 1, and therefore is less likely to overestimate the importance of a single data point when used in a weighted average. In corpus linguistics, ΔP is the probability of a first word occuring, given a second word, minus the probability of the first word occuring in the absence of the second word. Like most association measures, we construct a contingency table based on the joint and individual frequencies of the words in a collocation.

	$word_2$ present	$word_2$ absent
$word_1$ present	a	b
$word_1$ absent	c	d

For example, the sequence *worthwhile Canadian* occurs 7 times in our data, *worthwhile* occurs 11 times, *Canadian* occurs 14239 times, and there are 367456 adjective + demonym

sequences that contain neither *worthwhile* nor *Canadian*. Therefore, the contingency table for *worthwhile canadian* would be:

	Canadian present	Canadian absent
worthwhile present	7	4
worthwhile absent	14232	353224

Many association measures assume the independence of the words in a collocation (that is, they return a single value representing the strength of association between words). This is problematic with natural language. For example, take the collocation *of course*. Most association measures would return a single value inicating that these words are attracted to eachother. However, although *course* is often preceded by *of*, *of* is followed by many words. Therefore, it may be useful to discuss direction of the association between words. ΔP is a just such a directional association measure. Thus, in our case, we can find the probability of a demonym given an adjective, and the probability of an adjective given a demonym. So we can generate separate association scores for how probable *worthwhile* is to occur before *Canadian* and how probable *Canadian* is to follow *worthwhile*. This allows us to separately ask the questions "when this adjective occurs, what demonyms are likely to follow it?" and "when this demonym occurs, what adjectives are likely to preceed it?"

ΔP is calculated as

$$\Delta P_{2|1} = p(word_2|word_1 = present) - p(word_2|word_1 = absent) = \frac{a}{a+b} - \frac{c}{c+d}$$
$$\Delta P_{1|2} = p(word_1|word_2 = present) - p(word_1|word_2 = absent) = \frac{a}{a+c} - \frac{b}{b+d}$$

ΔP ranges from -1 to 1, with -1 indicating complete repulsion, and 1 indicating complete attraction, and 0 indicating no attraction or repulsion. For the sequence *worthwhile Canadian*, ΔP of *Canadian* given *worthwhile*, and *worthwhile* given *Canadian* is

$$\Delta P_{Canadian|worthwhile} = \frac{7}{7+4} - \frac{14232}{14232+353224} = 0.600$$
$$\Delta P_{worthwhile|Canadian} = \frac{7}{7+14232} - \frac{4}{4+353224} = 0.0004$$

This indicates that although *worthwhile* has a high probability of preceeding *Canadian*, *Canadian* has a low probability of following *worthwhile*.

Returning to our example of *traditional*, *traditional* occurs frequently before many demonyms, and our frequency weighted valence and arousal means treat all occurrences of *traditional* equally. By comparison, ΔP scores for the word *traditional* range from -0.1 to 0.2, with the vast majority very close to zero, indicating that *traditional* has no particular attraction or repulsion to any of our demonyms. To further illustrate, Table 3.1 shows the top 10 adjectives preceding *Iranian*, ranked by word frequency and by ΔP of a demonym, given an adjective. Note that when ranked by frequency, several high valence words like *new, young*, and *possible* are large contributors to the mean valence score for Iran, whereas when ranked by ΔP, the words tend to be lower valence.

Top 10 by frequency	Valence	Top 10 by $\Delta P_2\|1$	Valence
national	5.68	blocked	4.42
senior	4.79	blacklisted	3.45
other	5.45	coveting	4.74
current	5.53	invalid	3.45
young	6.36	offered	5.94
new	7.95	preposterous	5.10
alleged	3.45	sought	6.37
top	5.80	traversing	5.16
official	5.79	withholding	3.11
possible	7.10	assassinate	3.24

Table 3.1: Top 10 adjectives preceding *Iranian* ranked by word frequency and ΔP

We computed ΔP for all adjectives given demonyms, and all demonyms given adjectives. We then took these two ΔP scores and weighted mean valence and arousal scores for each country by them. Thus, we generated four separate scores, two each for valence and arousal. In other words, we generated scores representing whether when a demonym occurs, it is likely to be followed by more or less positive or more or less exciting adjectives and we generated scores representing when more or less positive adjectives occur, are they likely to be followed by particular demonyms. The fact that ΔP is directional allows us to separately assess whether a demonym shows a stronger attraction to more positive/more exciting adjectives, or whether more positive/more exciting adjectives shows a stronger attraction to particular demonyms.

3.3.2 Individual GloWbE Countries

It seems likely that the sentiments of the English speaking world towards countries of the world are heterogenous (e.g. Pakistan and the United States will have different sentiments towards Israel). Because GloWbE is composed of twenty sub-corpora with texts originating from different countries, we can also explore country specific relationships between sentiments in text and how nations of the world see each other. Therefore we also created weighted mean valence and arousal scores for all demonyms occurring in each sub-corpus. In other words, we calculated a valence and arousal score representing the pleasantness and excitedness of Canadian texts about Iran, British texts about China etc.) We weighted these means by frequency of occurrence before each demonym in each sub-corpus. Because of increasing data sparsity, calculating ΔP by each country would have led to many cases of perfect association (cases where a demonym and adjective only occur once), and would have skewed the results by treating very low frequency events as very informative, so only scores of frequency weighted arousal and valence were created.

3.4 Results

3.4.1 General Well-Being

We first assessed whether the overall sentiments expressed by the English-speaking world about other countries reflected national-level demographic statistics about other countries. We acquired national-level data on general well-being in countries through the United Nations Development Programme (2013), the World Bank (2013), and the International Monetary Fund (2012). These data sets included life expectancy at birth, per capita income, and rates of education, along with the Human Development Index (HDI), and Inequality Adjusted Human Development Index (IHDI). These statistics and indices are all intended to represent general well-being in a country. The Human Development Index is a composite index, calculated on the basis of life expectancy, levels of education, and income (United Nations Development Programme, 2014). There is a further index, the Inequality Adjusted Human Development Index, which is designed to overcome weaknesses in the HDI. Because the HDI represents *mean* levels of life expectancy, income, and education, it can fail to capture disparities in these factors between different populations in a country (such as very low life expectancy in a restricted segment of the population). To achieve high scores in the Inequality Adjusted Human Development Index, a country has to have a more uniform distribution of education, life expectancy, and income. The utility of these indices and statistics is a matter of some controversy (McGillivray, 1991; Hicks, 1997; Alkire and Foster, 2010), and our use of them is not intended as an attempt to suggest which is more useful or valid. We do note, however, that there is some variation in our ability to statistically predict these indicators using our text-derived data (see below), and it may be interesting that our best model predicts the Inequality Adjusted Human Development Index.

We performed a series of multiple least squares regressions, using the frequency weighted arousal and valence scores, the 4 ΔP weighted valence and arousal scores, and overall number of references to each country represented in our data as predictors. Arousal weighted by word frequency, valence weighted by the ΔP of a demonym given an adjective, and log frequency of

references to a country, emerged as significant predictors of a country's Inequality Adjusted Human Development Index ($R^2 = 0.24$, F(3, 101) = 12.06, p<0.001), Human Development Index ($R^2 = 0.22$, F(3, 130) = 13.57, p<0.001), and GDP per capita ($R^2 = 0.16$, F(3, 117) = 8.73, p<0.001). Table 3.2 presents summary statistics of each predictor in each model.

Specifically, the higher the valence score (weighted by ΔP of a demonym, given an adjective) of a country, the higher its IHDI score. In other words, if a demonym was likely to follow a more positive adjective, people tended to have greater general well-being in the country that demonym denotes. If a demonym was preceeded by adjectives that were overall more exciting, people tended to have greater general well-being in the country that demonym denotes. And if the English speaking world used a country's demonym more often, people in that country tended to have greater general well-being. This was true of all indicators of well-being, but interestingly the strongest effect size was found for the most inclusive indicator of general well-being. In other words, countries that had a stronger association with positive adjectives, that people modified with a predominance of calm adjectives, and that were dicussed more, tended to have higher levels of education, life expectancy, income, and less inequality, as measured in the Inequality Adjusted Human Development Index.

Figure 3.1 shows the partial effects of these predictors of a country's Inequality Adjusted Human Development Index score. For context, countries range from 0.183 (Democratic Republic of the Congo) to 0.894 (Norway), with a score of 1 representing high development with perfect equality.

One possible explanation for the observed correlation between arousal and these indicators of well-being is the general calmness bias in our emotional lexicon (Warriner et al., 2013). Indeed, there is evidence that there is a general calmness bias in the English language (Warriner & Kuperman, manuscript in preparation). The number of times each demonym is used in a corpus varies, with the demonyms of a few large and prominent countries taking up most of the references. The demonyms of smaller countries are used less frequently. Because highly frequent words tend to be calm, and smaller samples from within a corpus will usually contain more highly frequent words, it is likely that the average arousal of small

Model 1: Inequality Adjusted Human Development Index

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.7052	0.6085	-2.80	0.0061
Token Frequency Weighted Arousal	0.3891	0.1417	2.75	0.0071
$\Delta P_{demonym\|adjective}$ wt valence	0.0805	0.0310	2.60	0.0108
log frequency of references	0.0455	0.0100	4.53	0.0000

Model 2: Human Development Index

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.1690	0.4446	-2.63	0.0096
Token Frequency Weighted Arousal	0.3069	0.1027	2.99	0.0034
$\Delta P_{demonym\|adjective}$ wt valence	0.0787	0.0240	3.27	0.0014
log frequency of references	0.0354	0.0076	4.63	0.0000

Model 3: GDP per capita

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-162269.09	47563.59	-3.41	0.0009
Token Frequency Weighted Arousal	30862.86	10927.67	2.82	0.0056
$\Delta P_{demonym\|adjective}$ wt valence	7019.26	2537.66	2.77	0.0066
log frequency of references	2654.65	829.92	3.20	0.0018

Table 3.2: Summary statistics for regression models predicting measures of general well-being

samples will be generally lower than large samples. Therefore, our observed correlation between arousal and indicators of well-being would be an artefact of small countries being discussed less. To confirm that our observations are due to the sentiments of the English-speaking world, and not a potential calmness bias in the English language, we performed 1000 random simulations, sampling adjectives from the emotional lexicon weighted by token frequency in the SUBTLEX-US corpus (Brysbaert and New, 2009), while maintaining the size of each of our bins (the number of references to a country), and then correlating the weighted mean arousal of these sampled adjectives with our indicators of well-being. These random simulations only achieved a correlation of magnitude equal to or greater than our observed one in 2.7% of cases. Therefore, it is unlikely, but not impossible, that our observed correlation between arousal and indicators of general well-being are due to a calmness bias in the English language.

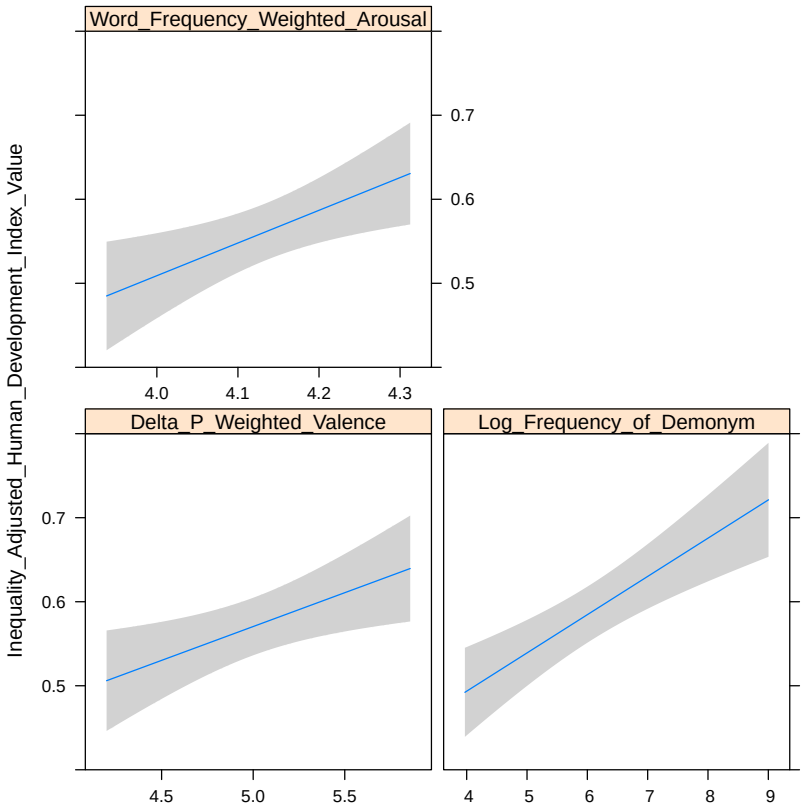

Figure 3.1: Partial effects of the predictors of the Inequality Adjusted Human Development Index

Together, these models show that patterns in the affective connotation of words used by the English speaking world around words denoting countries of the world reflect facts about those countries. Specifically, if a larger amount of words exciting in connotation occur near a place name, if a place name shows a stronger association with happy words, and the English speaking world uses a place name more in general, then people in the country corresponding to that place name tend to have greater general well-being. Places with lower general well-being tend to be associated with more unhappy words, more calm words are used to describe that place, and the English speaking world discusses that place less. Moreover, the strongest association between our text-derived data and measures of general well-being happens to be a multi-dimensional measure of well-being. This measure incorporates levels of education, life expectancy, income, and the distribution of these factors within a country. The weakest

association between text and indicators of well-being, although still a significant one, was with a unidimensional indicator of general well-being incorporating only income. This is suggestive that the English speaking world is concerned with more than national incomes; we care about levels of inequality, education, and life expectancy in other countries as well.

It should be noted that this finding is a correlation. It would be absurd to say that the connotative meanings expressed by the English speaking world *cause* people in other countries to have greater general well-being. And just because people in another country experience a certain level of well-being doesn't mean that we *must* use particular language when discussing that country. What this shows is that even at very high levels of aggregation (nationally, or even internationally) there is stability at the level of the community in terms of connotative meaning. Regardless of the complexities of international relations, all the reasons why people might come to have some opinion about a place, and all the ways they might choose to express those opinions, there are overall trends in the emotional connotation of language that correspond to real phenomena.

With regards to the frequency of references to a country as a predictor of well-being, it should be noted that the population of a country has a moderate to strong correlation with how often it is referenced by the English speaking world ($\rho = .61$, p = 0). However there is a good deal of variation in how often we talk about different countries and depending on their populations. Some very small countries (such as Israel, population 7 million) are discussed often (5453 times) and others (Uzbekistan, population 28 million) are discussed very little (61 times). It should come as no surprise that English-speaking countries discuss other countries for reasons other than their population size, but it is interesting that places where people are less well-off are generally discussed less.

3.4.2 World Public Opinion Polls

Our second analysis looked at trends in the emotional connotation of adjectives within texts belonging to one of the 20 individual English-speaking countries represented in GloWbE. Previous work has demonstrated that sentiment analysis can replicate traditional public

polls. We chose to compare the emotional connotation of texts from a "source" country (one of the twenty countries with data in GloWbE) to public polls conducted in that "source" country. The object of both the GloWbE text and the public poll was a "target country" (some other country in the world). For the study, we used the BBC World Service's World Public Opinion poll (2007; 2008; 2009; 2010; 2011; 2012; 2013). This poll asks people from around the world if a country's influence on the world is mostly positive or mostly negative. Polled individuals are also allowed to refuse to respond. These polls provide us with, for example, Australian polls respondents' opinions about Iran, which we can then compare to the affective quality of text about Iran that originates in Australia. The polling data is expressed as the percentage of respondents who said that a country had a 1. mostly positive influence on the world, 2. mostly negative influence on the world, or 3. did not respond. Table 3.3 shows the source and target countries included in the analysis (for which we had both text-derived data and polling data). There was some minor variation in both the source and target countries included in a given year of the polls.

Source Countries	Target Countries
Australia	Brazil
Canada	Canada
United Kingdom	China
Ghana	France
India	United Kingdom
Kenya	Georgia
Nigeria	Israel
Philippines	India
Pakistan	Iran, Islamic Rep.
United States	Japan
South Africa	Korea, Dem. Rep.
	Korea, Rep.
	Pakistan
	Russian Federation
	United States
	Venezuela, RB
	South Africa

Table 3.3: Source and Target Countries with matching polling and GloWbE data

The BBC polls covered the years 2007 to 2013. For each year of the poll, we matched

each pair of source and target countries that present within both that poll and our token-frequency weighted valence and arousal scores. We then tested for correlations between our valence and arousal scores and the polling data. In other words, for each year of the poll, we tested if a source country using more positive and more exciting adjectives before a target country's demonym corresponded to more poll respondents from that source country saying that a target country had a more positive influence on the world.

We discovered a moderate to weak correlation between token frequency weighted valence scores and a country being said to have a more positive influence on the world, in all years of the poll. We found a moderate to weak correlation between token frequency weighted valence scores and a country being said to have a more negative influence on the world. Last, we found a weak correlation between the number of poll respondents who refused to respond to the poll, and token frequency weighted arousal scores (ps and Spearman's ρs for these correlations are reported in Table 3.4). In other words, when a source country used more high valence adjectives before a place name, poll respondents in that source country said that target countries had a more positive influence on the world. When a source country used more low valence adjectives before a place name, poll respondents in that source country said that target countries had a more negative influence on the world. When a source country used more high arousal adjectives before a place name, poll respondents in that source country were more likely to refuse to respond to the question,

Year		2007	2008	2009	2010	2011	2012	2013
Positive Influence on World	$\rho =$	.55	.46	.43	.41	.42	.38	.39
Negative Influence on World	$\rho =$	−.44	−.49	−.38	−.40	−.31	−.38	−.33
Did Not Respond	$\rho =$	.38	.36	.38	.35	.21	.32	.25

Table 3.4: Effect sizes our observed correlations, in each year of the BBC World Public Opinion Polls. All ps < 0.001

Interestingly, the GloWbE derived valence and arousal scores show a steady decrease over time in the magnitude of their correlation with the World Public Opinion Polls (Figure 3.2 displays this trend graphically). The strongest correlations are generally with the earlier polls (closer to 2007). Although precise dating of texts within GloWbE is difficult, we know

that the corpus was collected in 2012. This suggests that patterns in connotative meaning within GloWbE best represent opinions of the English speaking world several years before the present. Research comparing sentiments collected from Twitter to public polling data has shown that public polls generally represent a "snapshot" of opinions that lag behind the current trends (O'Connor et al., 2010). Our findings show that GloWbE is the opposite; the sentiments within it better reflect the past. Thus, although we find that at the level of communities there is stability in the connotative meanings expressed before place names, there is instability in these meanings over time. And depending on the nature of the text (Twitter versus GloWbE) the relationship to time changes.

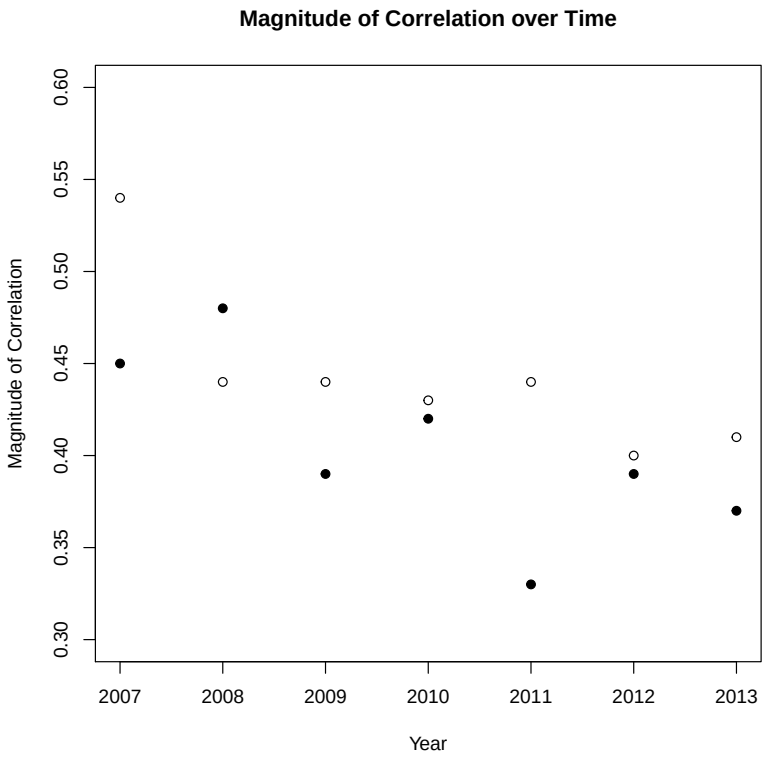

Figure 3.2: Correlations with Public Opinion Polls Over Time
Note that the black dots actually represent negative correlations between valence and the percentage of poll respondents who said a country had a negative influence on the world.

These results further demonstrate that sentiment analysis is capable of approximating the findings of a traditional public poll. Interestingly, the arousal scores of a target country

positively correlated with the percentage of polled individuals in a source country that did not respond, perhaps indicating that controversial countries receive more excited discussion in GloWbE text. Thus these findings also show the potential for arousal to play a separate and useful role in detecting sentiments.

3.5 Discussion

3.5.1 The GloWbE Corpus

One point of novelty within this study is the particular corpus we used, GloWbE. It appears that GloWbE is a good source of data about sentiments between English speaking countries. Twitter often has difficulty meeting the corpus-linguistics criteria of representativeness and balance. For example, users are generally young and use mobile phones (Lenhart and Fox, 2009). Indeed, GloWbE may be a good alternative to Twitter for studies involving geography, since it is static, and has been designed to be representative and balanced.

It is also worth noting the sheer size of the analysis we conducted. Indeed, it is remarkable that there is stability in the connotative meanings expressed at such large scales. In the introduction, I quoted from Labov (1996):

> The central finding of sociolinguistics is that the community is the stable and systematic unit, and that the behavior of individuals cannot be interpreted without a prior knowledge of the community pattern. At the level of the individual, language varies substantially, at the level of the community, language is stable.

Indeed, we find stability even at the level of nations, and most of the English-speaking world. It is worth noting that finding any consistency in the affective qualities of words at this scale is a surprise. There are many potential factors which might lead to noise in the data. Although we started with a list of demonyms, the data retrieved from GloWbE may denote many things. They may refer to languages (*French, English* etc), form collocations (*Russian Roulette*), or be adjectival (*Spanish cuisine, Chinese art*). Moreover, we have not

controlled for negation, or cases where words in combination form a phrase of different affective quality than the simple adjective + demonym collocation would suggest (e.g. *the great American myth of democracy*).

3.5.2 Public Opinion Polls and Sentiment Analysis

Labov's summary of sociolinguistics would perhaps be not all that surprising for someone who conducts public polls. After all, they are interested in generalizations about how, despite individual variations in opinions, there are generally held or more common opinions at the level of communities. What is interesting is the alignment of patterns of language use with generally help opinions in a community. Again, this is not a straightforward relationship. It may be useful to consider the following. What is the main difference between a public opinion poll and sentiment analysis? Public opinion polls generally get at their question of interest directly. Usually the party conducting the poll calls someone at random, and asks a direct and carefully worded question, and the response is coded according to some pre-arranged scheme. Sentiment analysis, on the other hand, involves no direct communication. Rather, a pre-existing text that has been deemed of interest is processed using a computer algorithm. Indeed, it is remarkable that it is even possible to get at someone's opinion without questioning them about it. Our results further cement the notion that sentiment analysis can provide information that previously would have required public polling. However, it should be noted that the correlation between the polls and our text-derived data is only moderate, and decreases over time. Furthermore, the particular poll we chose is quite general, since a positive or negative influence on the world could mean a good many things. It may be that polling questions of a more specific nature would be more difficult to re-create using text-derived data.

3.5.3 Psychologically Realistic Sentiment Analysis

We have shown that arousal contributes beyond valence in both major findings in this paper. The English speaking world doesn't just associate more negative words with places where

people are generally less well-off, it is also generally more calm about these places. This gives us a more specific picture of the sentiments of the English speaking world. Instead of places where people are less well-off being just "bad", the addition of arousal lets us talk about a combination of low valence and low arousal, the combination of which is generally described by emotion words like sad, bored, miserable, or depressed. We also know that the English speaking world discusses such places less in general. The second finding indicated that the presence of greater arousal may indicate indecision, controversy, or confusion, as the frequency weighted arousal means for each target country also correlated with being unwilling or unable to say whether that country had a negative or positive influence on the world. Without utilizing arousal, both observations would be missed or rendered less specific. Using an operational definition of emotion with psychological reality returns more nuanced information about community patterns of sentiment and language use.

Given that there is still controversy over how emotion is to be conceptualized, what exactly the features of a psychologically realistic sentiment analysis would look like remains an open question. Within research into the effect of emotion on word recognition, there is discussion of conceptualizing emotion in terms of danger and usefulness, rather than valence and arousal, where usefulness generally correlates with valence and danger correlates with arousal. Wurm (2007, p. 1218) points out that words can have both positive and negative connotations, such as "*electricity, elephant, pesticide, and syringe.*" In a series of experiments, (Wurm and Vakoch, 2000; Witherell et al., 2012; Wurm and Seaman, 2008) Wurm found that stimuli rated high both on both usefulness and danger slowed response times. This was interpreted to mean that items rated higher on both danger and usefulness engaged both approach and avoidance behaviours. Moreover, in an ERP experiment (Kryuchkova et al., 2012), danger and usefulness showed similar very early deflections to experiments using valence and arousal. The danger/usefulness account says that we monitor the environment for resources we can use, and threats, and this ancient adaptive behaviour has been carried over into how we process symbols. According to Wurm, this framework is better specified and makes more specific predictions than a framework of valence and

arousal, in that it gives us information about *why* we feel positively about some stimuli and negatively about others. He describes our responses to stimuli as being based in a system honed through natural selection to respond appropriately to different situations. Interestingly, in post-hoc analysis, Wurm found that danger and usefulness were better predictors of response times than valence and arousal (Wurm, 2007, p. 1123).

We have shown here that dimensions of emotion, in fact, the same data set of word ratings, explain variance in both experimental settings and within distributional patterns in a corpus. In other words, connotation seems to matter both on the millisecond timescale in word recognition, and across millions of utterances. So we have reason to think that observations within word recognition may be relevant to sentiment analysis. If danger and usefulness, not valence and arousal, are the critical dimensions of connotative meaning, what does this mean for sentiment analysis? Is it sensible to say that rather than viewing places with lower general well-being as calm and unpleasant we view them as neither useful nor dangerous? Moreover, if our theory of connotative meaning is grounded in what we know of approach and avoidance behaviours, how can this theory apply to things not in our immediate environment (such as other nations)? One direction for future research would be to collect a large body of danger and usefulness ratings, and apply them in sentiment analysis in comparison to valence and arousal. If considering danger and usefulness lead to better sentiment analysis, we may have evidence that they are the critical dimensions of connotative meaning. If they do not, we may have evidence that we need a framework that can account for both sets of connotative dimensions, and the relationship between them.

Although we have shown that arousal explains variance beyond valence, it should be noted that there are more sophisticated sentiment analysis methods, that control for negation and intensification. For example, GloWbE contains the strings *not a good American*, which we would have treated as positive, and *very successful Canadian*, which we treat the same as *successful Canadian*. Indeed, we only looked at demonyms, and only one part of their context with one Part of Speech, so perhaps larger contexts and other search-terms could reveal more than we found. It is also common for sentiment analysis methods to dis-

tinguish between "objective" (eg. *Compared to most of the world, the average Canadian is incredibly wealthy*) and "subjective" sentences *Every true Canadian is obsessed with [hockey]* (both strings from GloWbE), which are again treated equally by our method.

Chapter 4

General Discussion

This thesis considers whether

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the distribution [based on it] would be no more than a mere list.

Both studies presented in this thesis show that, far from being so skewed as to be rendered useless, patterns of language use within the GloWbE corpus contain a wealth of knowledge. Importantly, this is just not knowledge about natural language, but also knowledge about the real world. Tracking the occurrences of symbols provides us with information about both geography and how English-speaking countries see other countries of the world. The connotation of adjectives before place names bear a relationship to the general well-being of people in that place. Frequency distributions of place names are a sufficient basis to make inferences about geography. The GloWbE corpus is heterogeneous, and there is considerable variation from one utterance to the next, but this variation is structured. And this structure can be discovered if one knows what to look for and has the right statistical tools available to them.

These general insights into corpora and their value are not new, but they bear repeating. The social sciences is increasingly moving towards "big data" analysis, and the insights of

corpus linguists, and linguists in general, may be increasingly called upon as more and more scholars grapple with patterns of natural language use as a source of data. It seems that, within this form of interdisciplinary language-based social science, there are three main challenges. First, a technical one, including collecting and structuring the data, and choosing and applying appropriate algorithms/statistical methods. The second is careful consideration of the nature of language, and how language use patterns actually pertain to the topic of interest. The third is an understanding of human psychology and its relationship to linguistic behaviour. Thus, for the studies presented here, GloWbE had to be collected, constructed, and have access provided to researchers. We had to collect data from GloWbE, and apply appropriate corpus linguistic techniques to interrogate our questions. We needed to have data created by other scientists to compare to our own. We also had to have inspiration of what to look for, which came from investigations into human emotion and cognition.

The particular contributions of linguists to this process are in the latter two areas. As we have shown, even very simple analyses (in terms of computation or statistics) that have psychological and linguistic sophistication return valuable information. Simple frequency distributions of place names are sufficient information to build cognitive maps. The inclusion of one additional dimension of emotion provided a more nuanced picture of the sentiments of the English speaking world. For from being mere skewed lists, corpora are and will likely increasingly be valuable sources of data for a wide variety of scientists, with corpus linguists as important contributors to the endeavor.

Bibliography

Aarts, B. (1992). *Small clauses in English: The nonverbal types*, volume 8. Walter de Gruyter.

Alkire, S. and Foster, J. (2010). Designing the inequality-adjusted human development index.

Avraamides, M. N. (2003). Spatial updating of environments described in texts. *Cognitive Psychology*, 47(4):402–431.

Balling, L. W. and Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125(1):80–106.

Bayer, M., Sommer, W., and Schacht, A. (2012). P1 and beyond: Functional separation of multiple emotion effects in word recognition. *Psychophysiology*, 49(7):959–969.

BBC World Service (2007). Israel and Iran share most negative ratings in global poll. [Online; accessed 14-January-2014].

BBC World Service (2008). Global views of USA improve. [Online; accessed 14-January-2014].

BBC World Service (2009). Views of China and Russia decline in global poll. [Online; accessed 14-January-2014].

BBC World Service (2010). Global views of United States improve while other countries decline. [Online; accessed 14-January-2014].

BBC world service (2011). Views of US continue to improve in 2011 BBC country rating poll. [Online; accessed 14-January-2014].

BBC World Service (2012). Views of Europe slide sharply in global poll, while views of China improve. [Online; accessed 14-January-2014].

BBC World Service (2013). Views of China and India slide while UK's ratings climb: Global poll. [Online; accessed 14-January-2014].

Bernat, E., Bunce, S., and Shevrin, H. (2001). Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International Journal of Psychophysiology*, 42(1):11–34.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Brooke, J. (2009). *A semantic approach to automated text sentiment analysis*. PhD thesis, Simon Fraser University.

Bruthiaux, P. (2003). Squaring the circles: Issues in modeling English worldwide. *International Journal of Applied Linguistics*, 13(2):159–178.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Chambers, J. K. and Schilling-Estes, N. (2013). *The handbook of language variation and change*, volume 80. John Wiley & Sons.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Crystal, D. (2003). *English as a global language*. Cambridge University Press, Cambridge.

Culicover, P. W. and Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, 14(6):234–235.

Davies, C. (2013). Reading geography between the lines: Extracting local place knowledge from text. In *Spatial Information Theory*, pages 320–337. Springer.

Davies, M. (2013). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1):1–24.

Estes, Z. and Adelman, J. (2008). Automatic vigilance for negative words is categorical and general. *Emotion*, 8(4):453–457.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930 to 1955. *Studies in Linguistic Analysis: Special Issue of Philological Society*, pages 1–32.

Gibson, E. and Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6):233–234.

Gibson, E. and Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Gilquin, G. and Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Goldsmith, J. and Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

Goldstone, R. L. and Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3):295–320.

Gries, S. T. (2009). *Quantitative corpus linguistics with R: a practical introduction.* Routledge, New York, NY.

Gries, S. T. and Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective onalternations'. *International journal of corpus linguistics*, 9(1):97–129.

Haegeman, L. (1987). Register variation in english: some theoretical observations. *Journal of English Linguistics*, 20(2):230–248.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, volume 2, pages 159–166.

Herbert, C., Junghofer, M., and Kissler, J. (2008). Event related potentials to emotional adjectives during reading. *Psychophysiology*, 45(3):487–498.

Herbert, C., Kissler, J., Junghöfer, M., Peyk, P., and Rockstroh, B. (2006). Processing of emotional adjectives: Evidence from startle emg and erps. *Psychophysiology*, 43(2):197–206.

Hicks, D. A. (1997). The inequality-adjusted human development index: a constructive proposal. *World development*, 25(8):1283–1298.

Hunston, S. (2002). *Corpora in applied linguistics.* Ernst Klett Sprachen.

International Monetary Fund (2012). World economic and financial surveys world economic outlook database. [Online; accessed 5-Feb-2014].

International Organization for Standardization (2013). Online browsing platform — international organization for standardization. [Online; accessed 27-November-2013].

Kissler, J., Herbert, C., Peyk, P., and Junghofer, M. (2007). Buzzwords early cortical responses to emotional words during reading. *Psychological Science*, 18(6):475–480.

Kissler, J., Herbert, C., Winkler, I., and Junghofer, M. (2009). Emotion and attention in visual word processingan erp study. *Biological psychology*, 80(1):75–83.

Kousta, S., Vinson, D., and Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, 112(3):473–481.

Kryuchkova, T., Tucker, B. V., Wurm, L. H., and Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and language*, 122(2):81–91.

Kuperman, V., Bertram, R., and Baayen, R. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, 62:83–97.

Kuperman, V., Estes, Z., Brysbaert, M., and Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. 143(3):1065–1081.

Labov, W. (1996). When intuitions fail. In *32nd regional meeting of the Chicago Linguistics Society*, volume 32, pages 76–106.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Larsen, R., Mercer, K., Balota, D., and Strube, M. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion*, 8(4):4454–52.

Lenhart, A. and Fox, S. (2009). *Twitter and status updating*. Pew Internet & American Life Project Washington, DC.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, Y., Wang, F., Kang, C., Gao, Y., and Lu, Y. (2014). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS*, 18(1):89–107.

Louwerse, M. M. (2010). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

Louwerse, M. M. and Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive science*, 36(8):1556–1569.

Louwerse, M. M., Hutchinson, S., and Cai, Z. (2012). The chinese route argument: Predicting the longitude and latitude of cities in china and the middle east using statistical linguistic frequencies. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 695–700.

Louwerse, M. M. and Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive science*, 33(1):51–73.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, and Computers*, 28(2):203–208.

Manning, C. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

McEnery, T. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.

McGillivray, M. (1991). The human development index: yet another redundant composite development indicator? *World Development*, 19(10):1461–1468.

Milin, P., Filipović Đurđević, D., and Moscoso del Prado Martín, F. (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, 60(1):50–64.

Milin, P., Kuperman, V., Kostic, A., and Baayen, R. H. (2009b). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, pages 214–252.

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, page 8. ACM.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129.

Ortigue, S., Michel, C. M., Murray, M. M., Mohr, C., Carbonnel, S., and Landis, T. (2004). Electrical neuroimaging reveals early generator modulation to emotional words. *Neuroimage*, 21(4):1242–1251.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–158.

Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annual Reviews of Psychology*, 57:27–53.

Recchia, G. and Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41(3):647–656.

Richardson, A. E., Montello, D. R., and Hegarty, M. (1999). Spatial knowledge acquisition from maps and from navigation in real and virtual environments. *Memory & cognition*, 27(4):741–750.

Scott, G., O'Donnell, P., Leuthold, H., and Sereno, S. (2009). Early emotion word processing: Evidence from event-related potentials. *Biological psychology*, 80(1):95–104.

Searle, J. R. et al. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–457.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shaoul, C. and Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior research methods*, 38(2):190–195.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–167.

Stefanowitsch, A. and Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*.

Taylor, H. A. and Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, 31(2):261–292.

The World Bank (2013). Life expectancy at birth, total (years). [Online; accessed 14-January-2014].

Tillman, R., Datla, V., Hutchinson, S., and Louwerse, M. (2012). From head to toe: embodiment through statistical linguistic frequencies.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter how 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418.

United Nations Development Programme (2013). Human development reports. [Online; accessed 14-January-2014].

United Nations Development Programme (2014). Human development reports. [Online; accessed 19-Jun-2014].

United Nations Statistics Division (2013). Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. [Online; accessed 27-November-2013].

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change.

Wikipedia (2013). List of adjectival and demonymic forms for countries and nations — wikipedia, the free encyclopedia. [Online; accessed 27-November-2013].

Witherell, D., Wurm, L. H., Seaman, S. R., Brugnone, N. A., and Fulford, E. T. (2012). Danger and usefulness effects as a function of concept ancientness. *The Mental Lexicon*, 7(2):183–209.

Wurm, L. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic bulletin & review*, 14(6):1218–1225.

Wurm, L. H. and Seaman, S. R. (2008). Semantic effects in naming perceptual identification but not in delayed naming: implications for models and tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):381–398.

Wurm, L. H. and Vakoch, D. A. (2000). The adaptive value of lexical connotation in speech perception. *Cognition & Emotion*, 14(2):177–191.