Identifying Differences in News Coverage Between Cultural/Ethnic Groups

CHARLES B. WARD, MIKHAIL BAUTIN, and STEVEN SKIENA[1]
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400

1. INTRODUCTION

Interactions between distinct ethnic/cultural groups comprise one of the dominant social forces shaping our world. Accurately quantifying the nature of these interactions provides essential data for social science research, in fields as diverse as history, political science, and international relations.

In this paper, we report on an effort to use computational analysis of long-scale text streams (newspapers) to measure differences in the reference frequency and sentiment associated with distinct ethnic/cultural groups. Our methodology is as follows. Using two distinct and orthogonal classification methods (ethnic name analysis and co-reference association), we identify likely ethnicities and nationalities for each person mentioned in the news. By aggregating reference/sentiment counts across all members of a cultural/ethnic group, we obtain strong-enough signals to detect and measure interesting trends.

In particular, we report on the following contributions:

—*Methods for Ethnicity and Nationality Detection* – We develop a method to associate primary nationalities for distinct named entities (people) based on a statistical analysis of geographic collocations. We also briefly review our methods for name ethnicity detection based on surname / given name frequencies and k-mer analysis, which are more thoroughly presented in [Ambekar et al. 2009]. Here we validate both methodologies, and demonstrate how they can work together to measure ethnic divisions in regional and national contexts.

—*Geographic News Analysis of Cultural Groups* – We demonstrate that our methods coupling ethnicity/nationality detection with large-scale news analysis produce striking and insightful distributions of ethnicity and nationality. In particular, we produce a series of cartograms (skewed data maps) reporting the ethnic distribution of 13 cultural/ethnic (CEL) groups across the world. We present accurate frequency distribution and sentiment maps [Mehler et al. 2006] for CEL groups within the United States, and demonstrate that they accurately reflect survey data collected by the U.S. Census in 2000.

—*Time-Series Trends in CEL Group Frequency and Sentiment* – Though analysis of two extensive news corpora, we detect interesting trends and periodicities in news coverage. Our "dailies" corpus covers roughly 500 U.S. newspapers on a daily basis over the past four years, while our analysis of *The New York Times* spans 27 years, from 1981 to 2008. Interesting trends of where different CEL groups appear within sections of the newspaper (e.g. business, sports, entrainment) are revealed. Our sentiment analysis measures the half-life of major events such as 9/11 in public discourse, as well as the ebbs and flows of smaller events.

—*Intra-Group Interactions and Sentiment* – Our methods generalize to measure the frequency and sentiment co-locations spanning CEL groups, with natural geographic and political associations being revealed through the news data. We can quantify the tenor of interactions between different groups and how they change over time.

Of course, our methods can be applied to text streams such as blogs as well, but we limit our attention here to news data, because of its applicability over greater time spans and less controvertible geographic resolution. We anticipate that our data/analysis will be of intense interest to social scientists, and are working to make it readily available to scholars.

This paper is organized as follows. We begin with a quick review of related studies of ethnic bias in news coverage, and the *Lydia* text analysis system used to perform named entity recognition and sentiment analysis. We then present

[1]Corresponding author (skiena@cs.sunysb.edu). SS also serves as Chief Scientist at General Sentiment LLC.

our orthogonal methods for ethnicity and nationality detection, with evaluation results. Finally, we report on our observed temporal, geospatial, and association trends for all CEL groups.

2. PREVIOUS WORK

2.1 Ethnic Biases in Newspaper Coverage

Media bias is of considerable interest within the social sciences. Representative studies include [Taylor and Sorenson 2002], who examined the *Los Angeles Times* coverage of 1241 homicides over a 5-year period, and found that neither the ethnicity of perpetrator nor the victim was associated with the nature of news coverage. However, in related work reviewing the same news coverage, [Sorenson et al. 1998] found that the ethnicity of the victim did affect the volume of news coverage. Dixon, et al. found that Whites were over-represented in news coverage of crime, both as perpetrators and as victims [Dixon et al. 2003]. In a study of U.S. newspapers, Johnson [Johnson 1997] found that the coverage of Mexico is influenced by the ethnic makeup of the newspaper's circulation region.

2.2 News Analysis Infrastructure

The *Lydia* text analysis system [Lloyd et al. 2005] performs named entity recognition and analysis over text corpora. Although this paper is primarily concerned with the application of the system to newspaper data, the system can be applied to any other time-dependent text stream, such as blogs, patent records, or supreme court decisions. The system assumes only that its text input is quantized into dated articles from some number of different sources. *Lydia* consists of four primary components:

—Data collection: *Lydia* collects text from over 500 U.S. daily newspapers (as well as a significant number of foreign English-language newspapers) on a daily basis. This data collection has been ongoing since November of 2004, and we have accumulated a corpus of approximately one terabyte of text comprised of more than one-hundred million news articles (henceforth, the "Dailies" corpus).

—Natural Language Processing: *Lydia* performs a series of NLP tasks beginning from the raw text. These tasks include part-of-speech tagging, named entity extraction, named entity classification, coreference (pronoun) resolution, geographic normalization, and sentiment analysis [Bautin et al. 2008; Godbole et al. 2007]. The results of these tasks are passed to the final stage of the system as XML markup.

—Data Analysis: the final processing stage of the *Lydia* system takes the NLP marked-up text and generates statistics suitable to answer research questions in areas ranging from market research to social science. These statistics are typically time-series data, and range from frequency of mention, to polarity of sentiment, to frequency of entity juxtapositions.

—Visualization: the statistics computed by the analysis engine can be visualized in a number of ways. The most common of these is a timeseries for some entity on some statistic, such as reference frequency or sentiment polarity. At present, we support only relatively simple visualizations of time-series data, but richer visualizations such as those provided by ThemeRiver [Havre et al. 2002] could be easily supported from our data analysis component. We also create static or animated "heatmaps" (e.g., figure 6) to visualize the geographic and temporal flux of these statistics.

The data analysis phase is particularly challenging with large text corpora. Performing analysis of our one-terabyte Dailies corpus requires a system which is fast and scalable. Using a newly implemented Map-Reduce-based framework for statistical analysis, the Lydia system is now capable of generating interesting statistics for over 74 million named entities spanning this corpus using a 24-node cluster for 7 days. Among the results of this process are reference time series, sentiment analysis, and juxtaposition relationships.

For the purposes of this paper, the Lydia system has been extended to efficiently compute statistics over large groups of entities such as ethnic groups and synsets [Lloyd et al. 2006]. Like the remainder of the new system, this aggregation is done using Map-Reduce. As a result, we are able to aggregate hundreds of gigabytes of statistics in 10 hours on the same 24-node cluster.

3. ETHNICITY DETECTION

Name-based ethnicity classification has important applications to a number of fields, including biomedical research [Berchard et al. 2003] and sociology [Aries and Moorehead 1989], and has seen a good deal of research. However, the lack of high-quality, multi-ethnic, freely-available classifiers has, for the purposes of this paper, forced

us to create our own classification engine, described in more detail in [Ambekar et al. 2009]. In this section, we discuss prior work in the field as well as our own approach to the problem.

3.1 Previous Work in Ethnicity Detection

Many ethnic surname classification systems perform only binary classification. Buechley's Generally Useful Ethnicity Search System (GUESS) uses Spanish names to determine Hispanic ethnicity [Buechley 1976]. Coldman, Braun, and Gallagher perform binary classification of names as either Chinese or non-Chinese [Coldman et al. 1988]. Similarly, Harding et al. and Nanchahal et al. both built binary classifiers to identify South Asian names.

Binary classifiers can be very useful in particular domains. For example, Stewart et al. analyzes the utility of using the GUESS system to estimating relative cancer rates in the Hispanic population [Stewart et al. 1999]. However, for the purposes of our analysis, we require a multi-faceted classification system, capable of discriminating among a relatively large number of ethnic groups. As noted by Mateos [Mateos 2007], there has been relatively little work done on this type of system.

3.2 Methodology

The primary method used in ethnicity classification is comparison to known surname lists. Deriving accurate name lists for use in a classification engine is non-trivial, however. For example, Lauderdale and Kestenbaum describe the development of surname lists for six Asian ethnicity groups [Lauderdale and Kestenbaum 2000]. Somewhat more complex probabilistic methods have also been used [Coldman et al. 1988]. In this section, we will describe the compilation of our name lists, our classification methods which utilize them, and the accuracy of the resulting system.

3.3 Ethnicity Detection

Name-based ethnicity determination has important implications in a number of fields, including biomedical research and sociology, and has seen a good deal of research [Buechley 1976; Coldman et al. 1988; Stewart et al. 1999]. However, the lack of high-quality, multi-ethnic, freely-available classifiers has, for the purposes of this paper, forced us to create our own classification engine, described in more detail in [Ambekar et al. 2009].

Mateos, Webber, and Longley describe a methodology for ethnic classification based on the concept of CEL (Cultural, Ethnic, and Linguistic) groups [Mateos et al. 2007]. They combine 185 CEL types (e.g., Danish, Egyptian, etc.) into 13 CEL groups (e.g., Muslim, East Asian, etc.). These groupings are very useful, and heavily informed the choice of groupings in our classification engine. Moreover, the mapping from CEL types to CEL groups provided us with a convenient mapping between nationality of origin and CEL group, which proved important in compiling name lists for the system.

The primary source of names for our system was the Wikipedia categorizations of personal national origin. We extracted approximately $150,000$ names from Wikipedia, which were split into CEL groups according to national origin broadly by the basis of the guidelines suggested by Mateos et al. [Mateos et al. 2007] Although the Wikipedia data is somewhat noisy (and inaccurate in certain cases) this approach provides two advantages not utilized in many approaches:

—In addition to surnames, we can also derive ethnic information about given names.

—We can roughly estimate the extent to which different names are ethnic. That is, we have a rough distribution of the frequency of both surnames and given names in each population.

We broadly follow the CEL groups given by Mateos et al., though we depart in several minor ways so as to allow for a more natural hierarchical categorization of these groups. The reason for this is two-fold. First, this hierarchical categorization allowed us to analyze the accuracy of our classification engine at multiple levels of granularity as well as identify name groups which were indistinct. Secondly, efficiency issues which arise in dealing with statistics over our large text corpus force us to precalculate statistics over all groups of interest. The hierarchical classification provides us with reasonable aggregation levels for this purpose, allowing us to vary the trade-off between granularity and accuracy in our analyzes. Figure 3.3 shows our hierarchical categorization of CEL groups.

3.3.1 *Classification Engine.* Our classification engine consists of a series of sub-classifiers at each level of the CEL group hierarchy. Each classifier is essentially a Hidden Markov Model style automaton, with states corresponding to the position within a name and observations corresponding to substrings from the name. For a classifier discriminating between K ethnicities, the states of the classifier constitute a directed acyclic graph with K parallel paths,

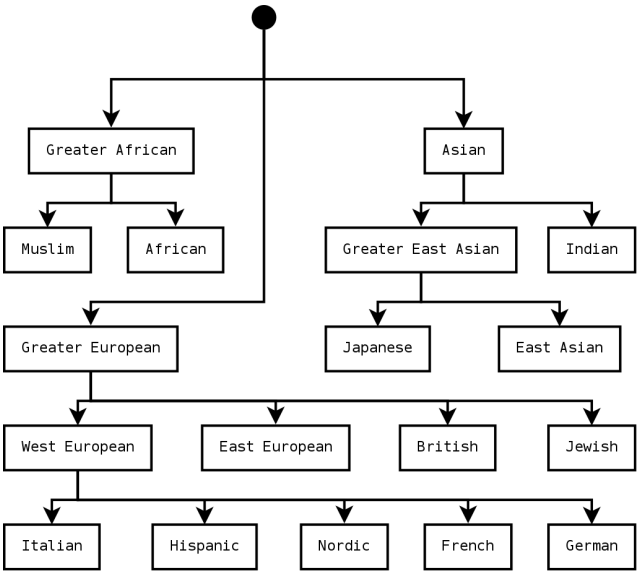


Fig. 1. Hierarchical categorization of CEL groups

each corresponding to an ethnicity. Each of these paths has edges with observations consisting of strings of letters determined by the strings' relative frequency within the training data for that ethnic group.

Observations along edges consist both of full names and of substrings of substrings. These observations, as well as the transition probabilities along these edges, determine the likelihood of a particular name being generated by the portion of the automaton corresponding to each ethnicity. Transition probabilities along multiple possible paths were manually determined, and biased towards paths which consume larger numbers of letters. That is, the system will prefer paths which allow it to process an entire name over those which it assembles from prefixes and suffixes, all other things being equal.

The automaton is also constructed to allow a path which "blends" ethnicities. That is, the given name and the surname of a name may indicate two separate ethnicities. The automaton allows this by allowing a path to move between ethnicities between first/middle/last name boundaries, at some cost. Varying this cost allows us to vary the relative weighting of the first, middle, and last names in the final analysis. Finally, the final state of the path with the highest probability based on this automaton yields the ethnicity classification.

3.3.2 *Results.* To assess the accuracy of our classification engine, we constructed automatons for each level of classification using training sets comprising 70% of the data extracted from Wikipedia and tested against the remaining 30%. Figure 3.3.2 shows the precision, recall, and F-scores for each ethnicity.

In general, the accuracy at the coarsest level of granularity is quite good, while at the finest it is comparable to other published ethnic classification systems. For example, the performance of our system in classifying Hispanic names is roughly equivalent to the accuracy of GUESS (as assessed by Stewart et al. [Stewart et al. 1999]). The most significant failing of the classifier is in the classification of Jews; the classifier here has quite low accuracy, owing largely to a comparatively poor training set, resulting in the Jewish group being artificially inflated. Table I gives the total entity counts and breakdown by percentage of each ethnicity as determined by our classifier; the final column of Table I gives the agreement with entities geographically associated to countries whose population largely falls within the CEL group.

4. ENTITY GEOGRAPHICAL ASSOCIATIONS

Beyond the groups formed by our ethnicity classification, another interesting type of culturally defined entity group can be composed from those entities which are closely associated with some particular country. That is, we would like to automatically classify large groups of news entities into groups by national identity. In this section we will discuss both how we do this, and the results obtained using our method.

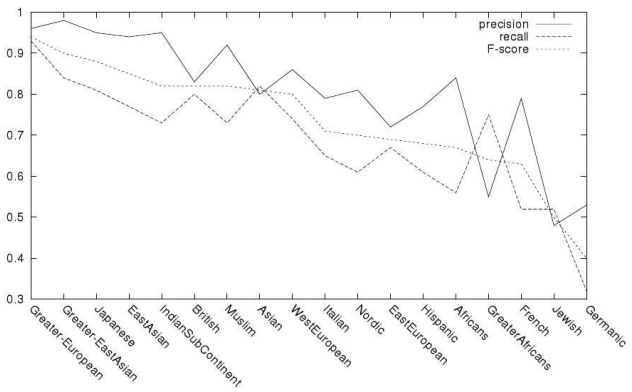


Fig. 2. Precision, recall and F-score curves for all the CEL groups in the hierarchy, with ethnicities sorted by F-score.

CEL Group	Entity Count	% of Names	% of Refs	% Geo. Associated
African	498185	2.2%	2.0%	27%
British	11354975	51.1%	56.7%	61%
E. Asian	508293	2.3%	1.9%	17%
E. Euro.	785609	3.5%	2.9%	16%
French	1105933	5.0%	3.8%	14%
German	555810	2.5%	2.0%	6%
Hispanic	1160509	5.2%	5.0%	26%
Italian	1252638	5.6%	5.0%	17%
Indian	712934	3.2%	2.1%	42%
Japanese	354765	1.6%	1.2%	19%
Jewish	2954538	13.3%	13.4%	40%
Muslim	515007	2.3%	2.4%	27%
Nordic	420626	1.8%	1.6%	10%

Table I. Distribution of entities by classified ethnicity.

4.1 Geographic Association

A juxtaposition relationship occurs between a pair of entities which co-appear in sentences within news articles. The strength of a juxtaposition relationship can be assessed simply by the frequency of appearance (juxtaposition count), or by a normalized juxtaposition score based on the relative frequency of both entities. Given two entities in an N sentence corpus which appear in N_1 and N_2 sentences and co-appear in F sentences, the juxtaposition score for these two entities is calculated as:

$$-\log\left(\left(\frac{e^{\frac{FN}{N_1N_2}-1}}{\left(\frac{FN}{N_1N_2}\right)^{\left(\frac{FN}{N_1N_2}\right)}}\right)\frac{N_1N_2}{N}\right)$$

In order to associate an entity with a country, we consider the list and strength of juxtaposition relationships associated with the entity, as reported in [Lloyd et al. 2005].

For convenience, the set of interesting geographic juxtapositions we use for this task are the set of countries, national adjectives (e.g., Russian), and international cities already recognized by the geographical normalization engine in the *Lydia* entity recognition pipeline. Table II shows two world leaders and their highest ranking geographical juxtapositions by juxtaposition score. Because the sources in our corpus are predominately U.S. daily newspapers (with some additional major Canadian and British sources), we expect a large skew towards associations with places within the United States.

Because of this, we first compute the mean and standard deviation of the juxtaposition frequency rate to each country's set of geographic juxtapositions and use this for normalization. Thus, out of the set of all geographic

Vladimir Putin		Nicolas Sarkozy	
Russia	652,743	France	322,273
Russian	335,563	French	265,513
Moscow, RUS	234,006	Paris, FRA	126,910
U.S.	100,705	China	51,744
Iran	98,324	Russia	53,238
Ukraine	83,574	Iran	47,706
Georgia	70,550	Afghanistan	44,505

Table II. Top geographic juxtapositions (with scores) for two important world leaders.

	Heads of State	Heads of Gov.
Total	177	192
Correct	116	112
Incorrect (Total)	36	32
(US)	5	2
(Same Region)	16	10
No data	25	48
Precision	0.76	0.78
Recall	0.66	0.58

Table III. Agreement between nationality and dominant geographic association for heads of state/government.

juxtapositions for a given entity, the entity will be assigned the nationality with the most statistically improbable set of juxtapositions, based on the frequencies precalculated for each nationality over the given corpus.

4.2 Results and Comparison with Ethnicity Data

Even following normalization, the dominant country of association for 80% of all entities is the United States, which should be expected in the analysis of U.S. newspapers. As validation that the remaining entities are associated correctly, we perform two assessments. First, we determined the country association of the Head of State and the Head of Government for 192 countries. Table III gives the results of this classification, establishing a high precision and recall.

As a second method of validating the utility of our geographically-associated entities, we analyze their ethnicities using our name-based classifier. Figure 3 shows cartograms for various CEL groups where the cartogram density of each country is determined by the percentage of associated entities classified as within the ethnic group. That is, for a particular ethnic chart, the size of a country, relative to its original geographical size, is determined by the fraction of individuals which were determined to have that nationality which were of that ethnic group. For example, a very proportion of those entities assigned to the Indian nationality were classified as ethnically Indian, yielding an increase in the size of India in the Indian map. Contrarily, a very small proportion of individuals associated with the Ivory Coast were classified as ethnically Indian, yielding a decrease in the size of the Ivory Coast in the Indian map.

These graphs demonstrate a number of interesting trends with strong demographic justification:

—The Hispanic CEL group is very well represented throughout the Americas, Spain, the Phillipines and parts of Indonesia, and the Portuguese colonial holdings in Africa.

—The effect of colonialism in Africa is somewhat visible, particularly in the abundance of French names throughout regions of northern Africa.

—Several CEL groups are quite global in their reach (primarily, the Western Europeans), while others are extremely regional (Eastern Europeans, Eastern Asians, and particularly Muslims).

5. TRENDS IN GROUP COVERAGE

With entity statistics now aggregated by ethnic group, we can examine trends in news volume and sentiment across various ethnic groups. This section provides charts detailing four primary methods of examining CEL group news coverage:

—*Reference volume time series* – the fraction of sentences which contained a person entity which belonged to this ethnic group.

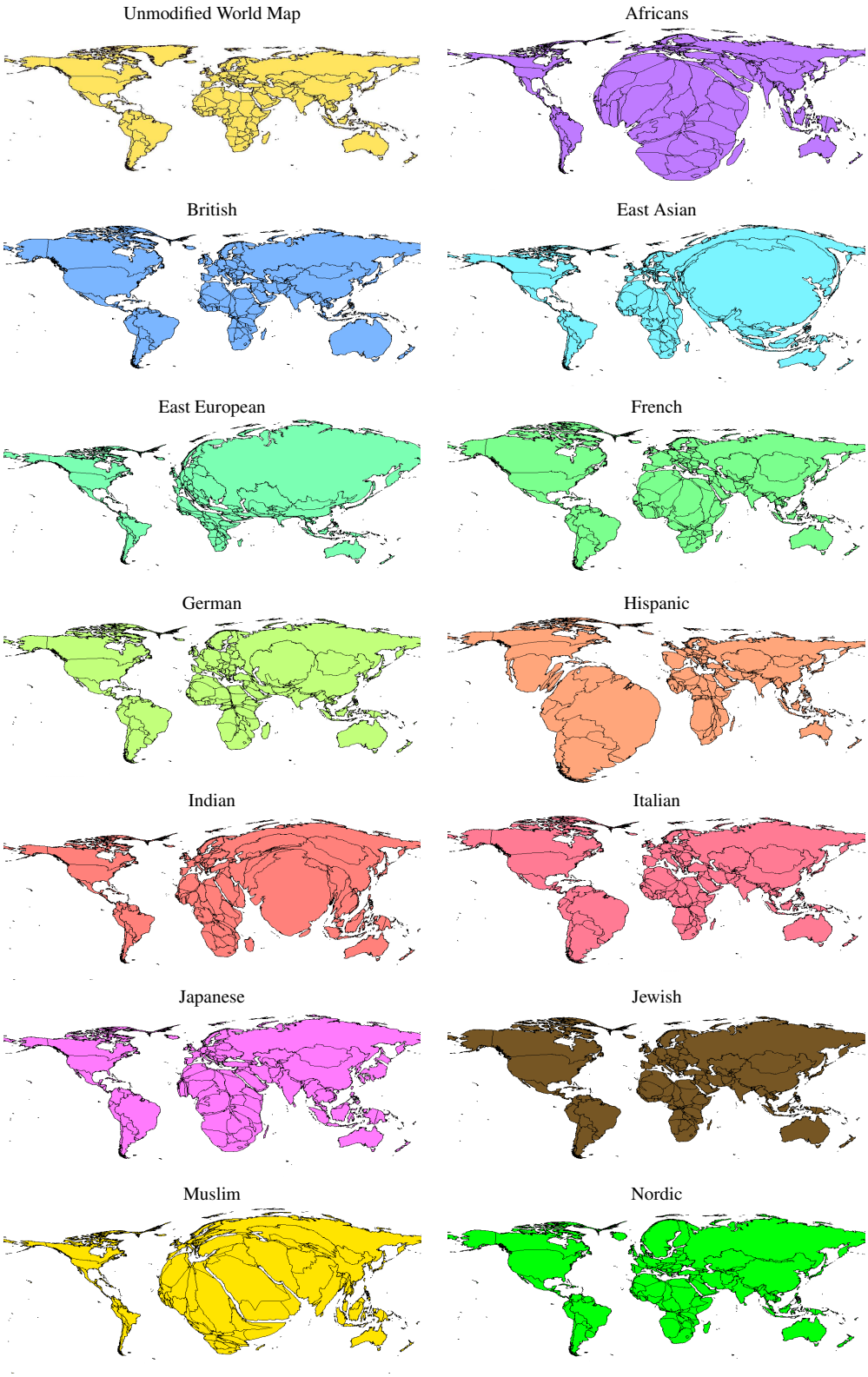


Fig. 3. CEL groups by nationality cartograms. Enlarged countries reflect higher concentrations of the given CEL group.

—*Sentiment score time series* – the aggregate sentiment score for all entities of the ethnic group. Sentiment score is a measure of how positively or negatively entities are being discussed, and is based on predefined dictionaries of positive and negative sentiment words. The score for a given entity is simply calculated as:

$$\frac{pos\_sent\_words - neg\_sent\_words}{pos\_sent\_words + neg\_sent\_words}$$

—*Geographic distribution* – the distribution of news volume of a CEL group by geographic region.

—*Juxtaposition relationships* – the aggregate juxtaposition score between two CEL groups, as computed by the formula given in the section on Geographic Associations.

We will discuss each of these in turn.

5.1 News Volume

Figure 4 shows aggregate news volume for all CEL groups across four years of U.S. daily newspaper coverage and the *New York Times* (NYT) coverage from 1981–2008, respectively. Interesting features include:

—Coverage of Muslim entities spikes dramatically during the first Gulf War and after September 11th.

—Coverage of Italian entities is consistently larger (8% on average) in the NYT than the national average, consistent with New York State's large Italian population. However, this increase is not proportionate to the size of the Italian population in New York, which is nearly double the national average.

—Coverage of Hispanics in the New York Times rose by only a third from 1980 to 2007, while over the same time period the country's Hispanic population roughly doubled as a share of the U.S. population. The Hispanic share of U.S. daily news coverage actually *fell* each of the past four years, for a total of over 6% in the period. This trend was split somewhat across regions, with the Southwest seeing a slight (3%) increase in Hispanic share, and other states with large Hispanic populations, such as Florida, remaining relatively stable.

One interesting trend immediately visible from the dailies data is the cyclic nature of coverage of Hispanic entities. The volume of coverage of Hispanics cycles between approximately 4% and 6% of total coverage of entities classified as people, peaking around July and falling to its lowest ebb around December. The cause becomes clear when we break down news volume by news article type and aggregate by month. The seasonal trend in Hispanic news volume results from the disproportionate number of baseball players who are Hispanic. It is also interesting to note that this trend is mostly washed out in the Southwest U.S., where Hispanic sports coverage does not dominate Hispanic coverage in general.

Table IV presents a breakdown by article type for each CEL group. Several ethnicities (e.g. Muslims, Jews, and East Asians) are dramatically underrepresented in sports, while Muslims and Africans are underrepresented in entertainment. Many groups show significantly different representation between articles classified as news and business, though the fraction of news articles is quite consistent across ethnicities.

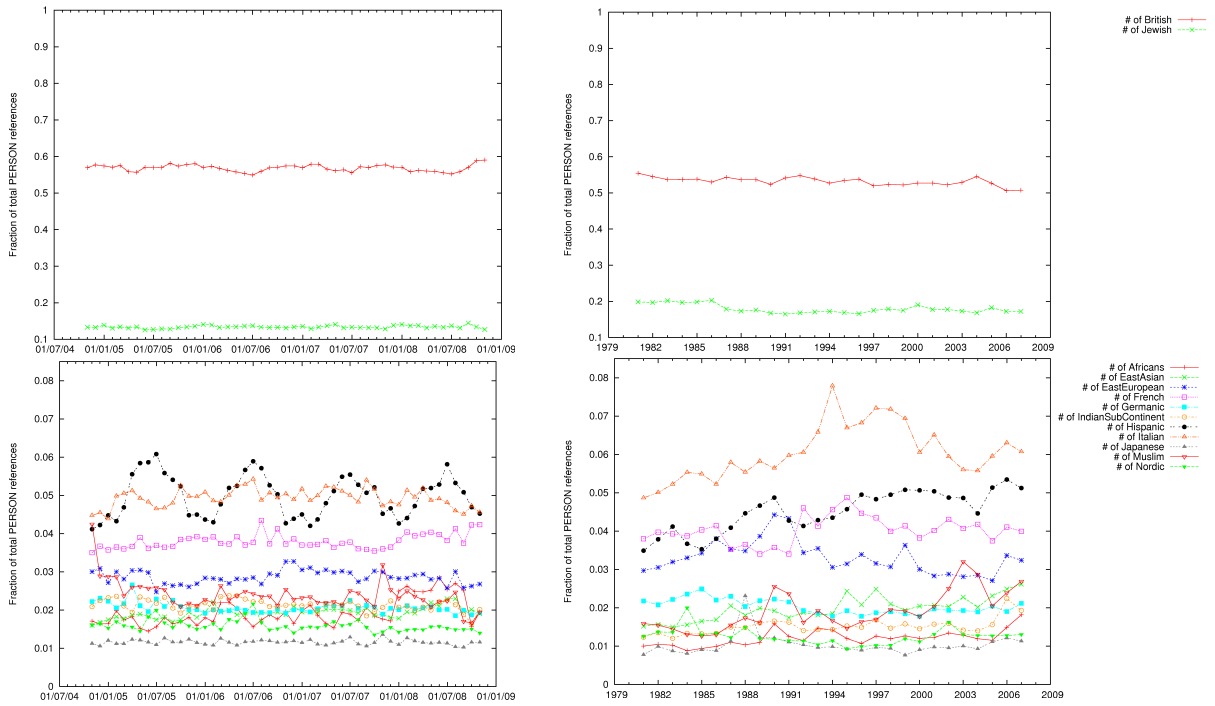
CEL group	Afri.	Brit.	E. As.	E. Eur.	Fren.	Germ.	Hisp.	Ital.	Ind.	Jap.	Jew.	Mus.	Nor.
News	17.4	18.5	17.3	16.5	16.4	17.3	20.6	19.3	17.1	14.7	18.9	17.7	16.8
Business	36.8	19.0	28.4	28.7	23.0	18.4	19.1	20.9	28.5	36.7	20.6	48.3	17.6
Entertainment	19.3	28.3	29.0	25.9	31.5	34.9	21.3	27.7	23.7	24.5	34.1	16.0	31.0
Sports	16.3	23.4	14.7	20.2	19.4	18.8	30.1	20.2	18.7	14.0	13.2	8.5	23.4
Other	10.1	10.9	10.6	8.6	9.7	10.7	8.9	11.8	11.9	10.1	13.2	9.4	11.1

Table IV. Type of news coverage by ethnicity

5.2 News Sentiment

Sentiment analysis broadly measures the tone of text pertaining to news entities. The *Lydia* sentiment analysis system is described in [Godbole et al. 2007]. Figure 5.1 shows time series of sentiment scores for each CEL group as computed by the system.

It is interesting to note that the daily sentiment scores for ethnicities are approximately normally distributed over the time range, though with differing means and variances. This is particularly true when we normalize to a difference



(a) US Dailies (b) *New York Times*

Fig. 4. Newspaper references to CEL groups.

from global sentiment (average sentiment for all people) for each date. In general, the predominant variation from normality is a thick negative tail, sometimes including a small second negative mode. In the following discussion, a "significant" deviation, is one which is significant at the 0.01 level, under the assumption that the distributions of daily sentiment for individual groups are normally distributed.

Several interesting trends are quickly observable:

—The majority of ethnic groups do not greatly differ in average sentiment from the baseline sentiment of the British CEL group. The French and Eastern European groups, in fact, do not significantly differ from British sentiment, but the majority of these differences, though small, are significant.

—Over the 27 year dataset of the NYT, the variance in sentiment from this baseline tends to narrow – perhaps reflecting improving sensitivities.

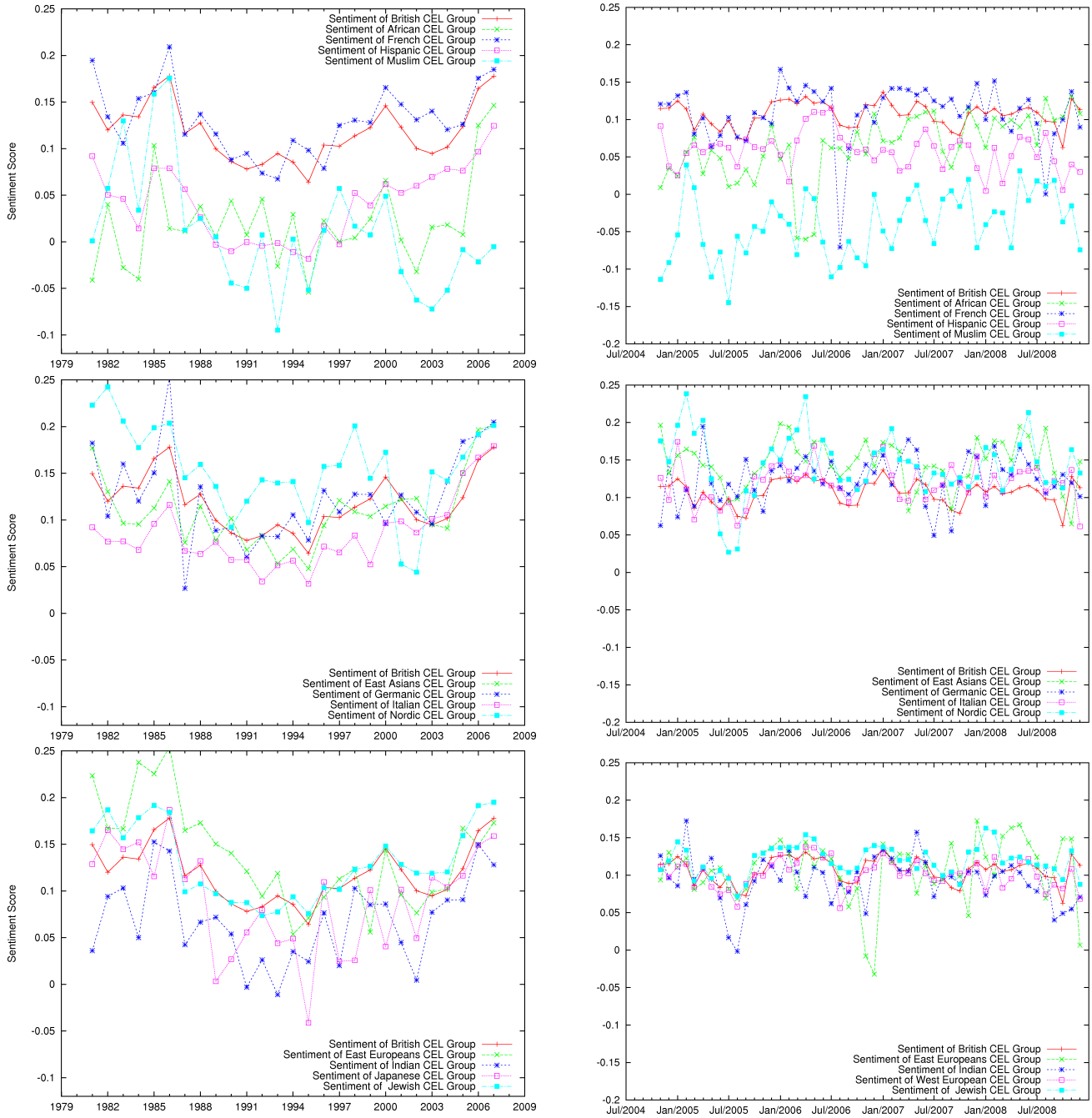
—The sentiment of Muslim entities is by far the lowest of any CEL group, with Muslim sentiment being particularly low during the Gulf War, after the World Trade Center Bombing, and for two full years following the September 11th attacks in 2001. Of all CEL groups, only the coverage of the Muslim CEL group is *more negative than positive*, a trend which shows no sign of reversing.

—Hispanic and African sentiment scores are also substantially and significantly lower than the baseline. This gap seems to narrow for the African group, though this is substantially due in 2008 to the favorable coverage for Barack Obama. The gap remains relatively constant for Hispanics.

—The Nordic and East Asian groups have the highest average sentiment scores of all CEL groups, both being significantly higher than all other groups.

5.3 Geographic Biases in News Coverage

Figure 6 shows "heatmaps" which illustrate the U.S. geographic biases in news volume and sentiment (respectively) for all CEL groups. In each of these maps, each source creates a "heat bloom" surrounding it, the strength of which is determined by the volume of news coverage for that ethnicity and the overall circulation of the newspaper. In these maps, oranges and reds reflect high volumes of coverage, while yellows and greens reflect less volume of coverage (these maps are obviously best viewed in the color electronic version of this paper). Note that the color scales on these



(a) *New York Times*, 1981–2008.

(b) US dailies, 2004-2008.

Fig. 5. Sentiment of CEL groups

maps are not comparable, as the color scales were adjusted for each ethnicity to maximize the ability to distinguish geographic disparities. Our heatmap visualizations are discussed in more detail in [Mehler et al. 2006].

Figure 7 provides a similar view on the geographic biases of sentiment for each ethnic group. These maps are computed in a similar way to the heatmaps in figure 6, but averaged over entire states. This was done because to provide more easily interpretable sentiment maps.

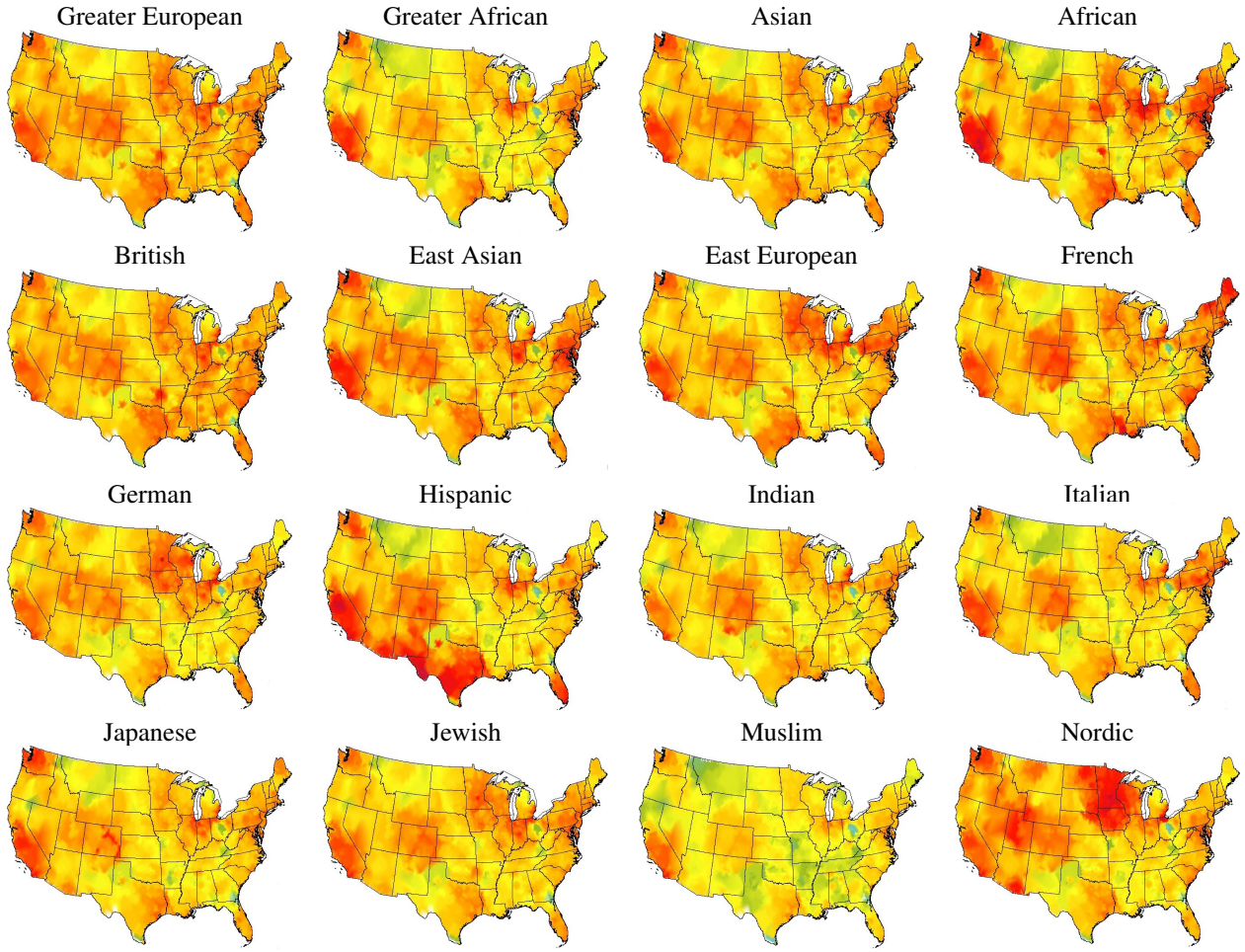


Fig. 6. Frequency Maps for CEL groups within the United States

In general, we note that the size of the local CEL group population heavily influences the news volume of entities from that group. The frequency maps in Figure 7 correlate extremely well with maps of ethnic ancestry generated by the U.S. Census [Census 2000], particularly with respect to Hispanics and Scandinavians. To summarize:

—Large Hispanic populations throughout the Southwest and Florida generate large volumes of local news coverage, which is disproportionally of negative sentiment.

—French populations emerge along the border with Quebec and historically French Louisiana.

—Scandinavian populations throughout Minnesota, Wisconsin, the Dakotas, Montana, and Utah are all reflected.

—The largest Italian populations throughout the Northeast and California are well-represented in news coverage, as are other significant Italian populations in the Midwest and Colorado.

5.4 Juxtaposition Relationships between CEL Groups

Table V reports the normalized strength of association between pairs of CEL groups derived from news analysis. Under the assumption of independence, the number of collocations between two groups should be proportional to the product of the group sizes. These values have been normalized so 1.0 corresponds to a number of juxtapositions between the two groups which would be expected by chance under the assumption of statistical independence. Large deviations from independence have been highlighted.

We note that most groups exhibit strong inter-group collocation. That is, entities of that ethnic group tend to be juxtaposed with other entities of the same ethnic group, over and above that which we would expect simply based on the relative frequency of mentions to that group. The notable exception to this trend is the British CEL group,

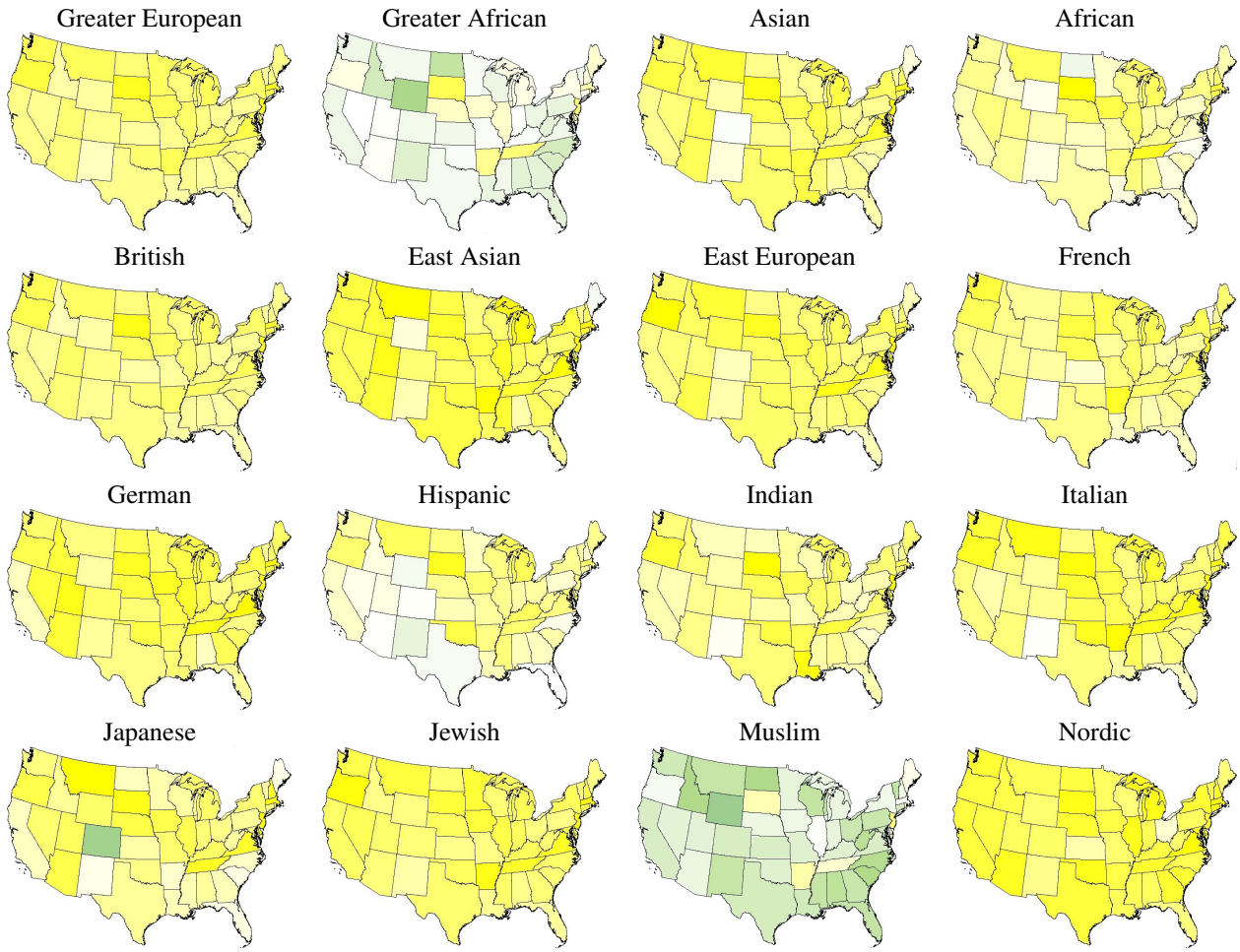


Fig. 7. Sentiment Maps for CEL groups within the United States.

which presumably reflects both the geographically widespread use of English names and the relatively high frequency of other CEL groups in the United States. The Muslim CEL group shows strikingly low rates of interactions with all but the African and Indian communities. Other sets of similar CEL groups, such as the Japanese CEL group and the East Asian CEL group (which includes Chinese names, Korean names, etc.), also show relatively high inter-group juxtaposition rates.

6. CONCLUSIONS

We have demonstrated that subtle spatial, temporal, and associative trends can be distinguished between cultural/ethnic groups on the basis of aggregate news analysis. We believe our results illustrate the power of our analytic techniques for serious research in several of the social sciences.

We are now working to expand our analysis to a variety of other text corpora and group phenomena (such as influence of gender). Of particular interest is the blogosphere (which should prove even more representative of local sentiment if the postings can be accurately geocoded) and longer-term news archives starting from the 1850's (providing insight into historical trends and cultural forces). Further work in computational methods will revolve around improvements to our ethnicity/nationality classifiers and other group identification techniques (e.g. [Mehler and Skiena 2009]).

Acknowledgements

We would like to thank Edward Fox, whose comments have improved the presentation of this paper.

CEL group	Afri.	Brit.	E. As.	E. Eur.	Fren.	Germ.	Hisp.	Ind.	Ital.	Jap.	Jew.	Mus.	Nor.
African	**3.01**	*0.75*	*0.75*	0.84	*0.74*	*0.71*	1.08	1.2	*0.79*	0.89	*0.73*	**1.6**	*0.7*
British	*0.75*	*0.64*	0.88	0.8	0.93	0.96	0.87	0.87	0.9	0.89	0.98	*0.64*	1.07
East Asian	*0.75*	0.88	**3.48**	*0.71*	0.8	*0.75*	0.86	1.13	0.82	**1.26**	0.83	*0.63*	0.9
East Euro.	0.84	0.8	*0.71*	**1.83**	1.1	0.99	*0.79*	0.81	0.8	0.85	0.89	*0.67*	1.16
French	*0.74*	0.93	0.8	1.1	**1.48**	1.15	1	0.96	1.19	*0.79*	0.94	*0.67*	1.18
German	*0.71*	0.96	*0.75*	0.99	1.15	**1.88**	1.01	0.86	0.93	*0.73*	1.17	*0.69*	**1.26**
Hispanic	1.08	0.87	0.86	*0.79*	1	1.01	**2.5**	0.87	**1.28**	1.1	*0.76*	*0.71*	1.09
Indian	1.2	0.87	1.13	0.81	0.96	0.86	0.87	**2.18**	*0.77*	0.99	0.91	**1.46**	1.05
Italian	*0.79*	0.9	0.82	0.8	1.19	0.93	**1.28**	*0.77*	1.16	0.84	0.95	*0.57*	0.97
Japanese	0.89	0.89	**1.26**	0.85	*0.79*	*0.73*	1.1	0.99	0.84	**3.15**	*0.77*	0.85	1.1
Jewish	*0.73*	0.98	0.83	0.89	0.94	1.17	*0.76*	0.91	0.95	*0.77*	0.85	*0.68*	1.07
Muslim	**1.6**	*0.64*	*0.63*	*0.67*	*0.67*	*0.69*	*0.71*	**1.46**	*0.57*	0.85	*0.68*	**2.98**	*0.58*
Nordic	*0.7*	1.07	0.9	1.16	1.18	**1.26**	1.09	1.05	0.97	1.1	1.07	*0.58*	**2.47**

Table V. Strength of juxtaposition relationships between CEL groups.

REFERENCES

AMBEKAR, A., WARD, C., MOHAMMED, J., REDDY, S., AND SKIENA, S. 2009. Name-ethnicity classification from open sources. In *ACM SIG-KDD 2009 (To Appear)*.

ARIES, E. AND MOOREHEAD, K. 1989. The importance of ethnicity in the development of identity of black adolescents. *Psychological Reports 65*, 75–82.

BAUTIN, M., VIJAYARENU, L., AND SKIENA, S. 2008. International sentiment analysis for news and blogs. In *Second Int. Conf. on Weblogs and Social Media (ICWSM 2008)*.

BERCHARD, E., ZIV, E., AND ET. AL. 2003. Importance of race and ethnic background in biomedical research and clinical practice. *The New England Journal of Medicine 348*, 1170–1175.

BUECHLEY, R. W. 1976. Generally useful ethnic search system, GUESS. *mimeographed paper, Cancer Research and Treatment Center, University of New Mexico*.

CENSUS, U. S. 2000. Maps of American ancestries. http://en.wikipedia.org/wiki/Maps_of_American_ancestries.

COLDMAN, A. J., BRAUN, T., AND GALLAGHER, R. P. 1988. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health 42*, 390–395.

DIXON, T. L., AZOCAR, C. L., AND CASAS, M. 2003. The portrayal of race and crime on television network news. *Journal of Broadcasting & Electronic Media 47,* 4, 498–523.

GODBOLE, N., SRINIVASAIAH, M., AND SKIENA, S. 2007. Large-Scale Sentiment Analysis for News and Blogs. In *Proc. First Int. Conf. on Weblogs and Social Media*. 219–222.

HAVRE, S., HETZLER, E., WHITNEY, P., AND NOWELL, L. 2002. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on 8,* 1 (Jan/Mar), 9–20.

JOHNSON, M. A. 1997. Predicting news flow from Mexico. *Journalism & mass communication quarterly 74*, 315.

LAUDERDALE, D. S. AND KESTENBAUM, B. 2000. Asian american ethnic identification by surname. *Population Research and Policy Review 19*, 283–300.

LLOYD, L., KECHAGIAS, D., AND SKIENA, S. 2005. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval (SPIRE 2005)*. 161–166.

LLOYD, L., MEHLER, A., AND SKIENA, S. 2006. Identifying co-referential names across large corpra. In *Proc. Combinatorial Pattern Matching (CPM 2006)*. Vol. LNCS 4009. 12–23.

MATEOS, P. 2007. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place 13,* 4, 243–263.

MATEOS, P., WEBBER, R., AND LONGLEY, P. 2007. The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. Tech. rep., CASA Working Papers 116, Centre for Advanced Spatial Analysis University College London. March.

MEHLER, A., BAO, Y., LI, X., WANG, Y., AND SKIENA, S. 2006. Spatial Analysis of News Sources. In *IEEE Trans. Vis. Comput. Graph.* Vol. 12. 765–772.

MEHLER, A. AND SKIENA, S. 2009. Expanding network communities from representative examples. *ACM Trans. Knowledge Discovery from Data (TKDD) 3,* 2.

SORENSON, S. B., MANZ, J. G., AND BERK, R. A. 1998. News media coverage and the epidemiology of homicide. *Am J Public Health 88,* 10, 1510–1514.

STEWART, S. L., SWALLEN, K. C., GLASER, S. L., HORN-ROSS, P. L., AND WEST, D. W. 1999. Comparison of Methods for Classifying Hispanic Ethnicity in a Population-based Cancer Registry. *Am. J. Epidemiol. 149,* 11, 1063–1071.

TAYLOR, C. A. AND SORENSON, S. B. 2002. The nature of newspaper coverage of homicide. *Inj Prev 8,* 2, 121–127.