Cognitive Science 33 (2009) 51–73

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2008.01003.x

Language Encodes Geographical Information

Max M. Louwerse,[a] Rolf A. Zwaan[b]

[a]*University of Memphis*
[b]*Erasmus University Rotterdam*

Received 31 October 2007; received in revised form 21 March 2008; accepted 18 April 2008

Abstract

Population counts and longitude and latitude coordinates were estimated for the 50 largest cities in the United States by computational linguistic techniques and by human participants. The mathematical technique Latent Semantic Analysis applied to newspaper texts produced similarity ratings between the 50 cities that allowed for a multidimensional scaling (MDS) of these cities. MDS coordinates correlated with the actual longitude and latitude of these cities, showing that cities that are located together share similar semantic contexts. This finding was replicated using a first-order co-occurrence algorithm. The computational estimates of geographical location as well as population were akin to human estimates. These findings show that language encodes geographical information that language users in turn may use in their understanding of language and the world.

Keywords: Spatial cognition; Geography; Latent semantic analysis; Computational linguistics; Corpus linguistics; Geographical coordinates; Word frequency; Multidimensional scaling; Semantic representations

1. Introduction

Humans can acquire geographical knowledge about their environment in various ways (Montello & Freundschuh, 1995). First, they can acquire spatial information about the environment experientially, for example via locomotion and stationary viewing. Second, they can acquire information through static pictorial representations, such as diagrams, paintings and photos, provided on a map. Third, they can acquire information via dynamic pictorial representations, including animations, movies and videos. Finally, they can acquire geographical knowledge via verbal descriptions. In real life, usually a combination of these four

Correspondence should be sent to Max M. Louwerse, Department of Psychology/Institute for Intelligent Systems, University of Memphis, 202 Psychology Building, Memphis, TN 38152. E-mail: mlouwerse@memphis.edu

sources of information will be used. However, given the limitations of acquiring experiential information about large geographical expanses and the importance of the written word in our modern culture, it is plausible to assume that a significant portion of geographical knowledge is also acquired via the vast amounts of textual information that one has been exposed to. Several studies have investigated the effect of language on the formation of spatial relations (Ferguson & Hegarty, 1994; Franklin & Tversky, 1990; Perrig & Kintsch, 1985; Taylor & Tversky, 1992a). For instance, Taylor and Tversky (1992a) found evidence that comprehenders construct spatial maps from verbal descriptions (''Sitting on a shelf directly to your right …'') or explicit routes (''Driving from … to …, you pass …'') equally well as from actual maps. In general, the research investigating the effect of language on cognitive maps focuses on ''local'' environmental distances and perceptual distances (Baird, 1970; Canter & Tagg, 1975), to which route directions apply. Examples of these maps are amusement parks (Taylor & Tversky, 1992b), convention centers (Taylor & Tversky, 1992a), neighborhoods (Tversky, Lee, & Mainwaring, 1999), doll houses (Ehrich & Koster, 1983), households (Ehrlich & Johnson-Laird, 1982), and research centers (Freundschuh & Mercer, 1995; Morrow, Greenspan, & Bower, 1987). No studies have investigated the role of language in ''global'' geographical knowledge, that is, cases where locations are typically difficult to estimate by route directions.

There is, however, a considerable amount of research available on how subjective representations of geography are formed. Stevens and Coupe (1978) for instance found that participants made category errors when judging whether Reno, Nevada, was further west than San Diego, California. Similar mistakes were made in judging whether Windsor, Ontario, was further north than Detroit, Michigan. One explanation is that participants group cities by region (e.g., states and countries) and make judgments on the basis of these categories (Friedman & Montello, 2006). Another (complementary) explanation is that participants use a rotation and alignment heuristic (Tversky, 1981, 1997) to map relative positions more vertically or horizontally and more lined up than they really are. Replicating Tversky's (1981) finding, Friedman, Kerkman, and Brown (2002) found that participants have a bias to locate European cities south of their U.S. counterparts even if they have similar latitudes, and argued for categorical grouping rather than a heuristic. This categorical bias has been acquired over time (Kerkman, Friedman, Brown, Stea, & Carmichael, 2003).

The consequence of the geographical bias varies depending on where participants live. Tobler (1970) for instance proposed a proximity hypothesis stating that a participant's estimation bias should increase with the increasing physical distance from the participant's home town. More recently, Friedman et al. (2002) tested this hypothesis by comparing latitude estimates of participants in Alberta, Canada, and Texas. Their findings did not support the proximity hypothesis: Texans had a considerably greater bias in their estimates of Mexican locations than did the Albertans. Friedman et al.'s (2002) explanations for this bias ranged from cognitively based beliefs, geopolitically based beliefs, to socioculturally based beliefs.

Akin to the question whether subjective representations of geographical locations could be acquired via textual information is the question whether subjective representations of population size could be acquired via textual information. Goldstein and Gigerenzer (1999, 2002) conducted a number of experiments identifying what heuristic participants use in

estimating the population size of cities. Their experiments defied common belief that performance in estimating population size is a product of knowledge. Goldstein and Gigerenzer (2002) found that U.S. participants were slightly better at estimating the population size of cities in Germany than cities in their own country when they were given pairs of cities and were asked to choose which city had a higher population count. They explained these results by participants using a *recognition heuristic*. This heuristic is an ignorance-based reasoning that infers that recognized objects have higher values than objects that are not recognized. Whether the availability (Tversky & Kahneman, 1974) and familiarity (Griggs & Cox, 1982) heuristics are cognitively very different from the recognition heuristic—Goldstein and Gigerenzer (2002) argue they are because contrary to the other two a recognition heuristic is an all-or-nothing approach that does not rely on recall—falls outside the scope of this paper. Goldstein and Gigerenzer (2002) describe the recognition heuristic as follows: An inaccessible criterion (e.g., population size) is reflected by a mediator variable (e.g., the frequency a city is mentioned in the news), and the mediator influences the probability of recognition. The mediator could be static or dynamic pictorial representations (e.g., diagrams) or language.

In sum, research on geographical judgments has found that participants are able to locate cities on a map, but they have geographical belief biases that are a function of where participants live. How participants have acquired ''global'' geographical knowledge is not entirely clear. At least it involves direct environmental experience, static pictorial representations, and dynamic pictorial representations. But whether participants could have acquired this knowledge through language remains a research question, because all research that has investigated the assessment of spatial representation using text has focused on ''local'' perceptual information and local route descriptions.

Furthermore, studies that investigated participants' estimates of population size of cities asked participants to choose the largest member from pairs of cities. Though these results show a correlation between frequency of mention and actual population, as well as a correlation between frequency of mention and recognition, it is unclear how participant population estimates (rather than choice) relate to actual scores and to mediator scores (e.g., verbal descriptions).

The goal of the present paper was to determine to what extent geographical information can be extracted from a body of texts, even though these texts themselves are not necessarily spatial descriptions. For example, it could be hypothesized that larger geographical entities (e.g., cities) are mentioned more often in texts than smaller geographical entities. Thus, one would expect the frequency of mention for Chicago to be substantially higher than that for, say, Memphis. Another hypothesis pertains to proximity. It might be expected that the names of geographical entities that are close in space co-occur more often in texts than the names of cities that are far apart in geographical distance. Thus, we might expect the word pair *Memphis–Nashville* to show a higher co-occurrence rate than the word pair *Memphis–Chicago*. These co-occurrences can then be transformed into locations in a two-dimensional space, where axes represent latitude and longitude.

In the first study reported in this paper, we tested two hypotheses regarding geographical patterns in large bodies of text. According to the first hypothesis, text co-occurrence scores

between pairs of cities should correspond to the distance between them: ''cities that are located together are debated together.'' According to the second hypothesis, the frequency with which city names occur in text corpora should correspond to their population sizes: ''cities that are populated more are debated more.'' To address these two hypotheses, we used computational linguistic techniques to produce location and population estimates based on word frequency and word co-occurrences in text corpora.

In the second study, we asked to what extent these text-based representations corresponded to estimates generated by human participants. To address this question, we asked participants to estimate the location of 50 U.S. cities as well as their population size.

2. Study 1a: Corpus-based geographical estimates

Word frequencies for cities in a particular corpus of texts are easy to compute. However, determining whether *Nashville* is more likely to appear in the same document with *Memphis* than with *Chicago* is more difficult to compute, because of a sparsity problem. *Memphis* and *Nashville* may in fact never co-occur in a paragraph of a corpus, although they *could* have if the corpus had been large enough. One way to solve this sparsity problem is to not rely on the co-occurrences of the words per se, but on the co-occurrences of the neighbors of these words (and the neighbors of those neighbors, as well as their neighbors, etc.). In other words, the sparsity problem in co-occurrence analyses can be solved by examining whether the words occur in similar contexts. Computing semantic similarities of higher-order relationships between words is the strength of the statistical technique behind Latent Semantic Analysis (Landauer, McNamara, Dennis, & Kintsch, 2007).

Latent semantic analysis (LSA) is a statistical, corpus based technique for representing world knowledge that estimates semantic similarities on a scale of −1 to 1 between the latent semantic representation of terms and texts (Landauer et al., 2007). In the current study the input to LSA was a sample of newspaper articles (e.g., *New York Times*) segmented into paragraphs. Mathematical transformations created a large term-document matrix from the input. For example, if there are m terms in n paragraphs, a matrix of $A = (f_{ij} \times G(j) \times L(i,j))_{m \times n}$ is obtained, in the case of the *New York Times* corpus used in this study m = 53,263 and n = 35,299. The value of f_{ij} is a function of the integer that represents the number of times term i appears in document j: $L(i; j)$ is a local weighting of term i in document j; and $G(j)$ is the global weighting for term j. Such a weighting function is used to differentially treat terms and documents to reflect knowledge that is beyond the collection of the documents. As in most LSA studies (Dumais, 2007; Martin & Berry, 2007), we used natural log as the local weight and log entropy as the global weight in the current analyses. The large matrix of A has, however, lots of redundant information, for instance because not every word occurs in every paragraph. Singular Value Decomposition (SVD) reduces this noise by decomposing the matrix A into three matrices $A = U\Sigma V'$; where U is an m by m and V is an n by n square matrix, with Σ being an m by n diagonal matrix with singular values on the diagonal. By removing dimensions corresponding to smaller singular values and keeping the dimensions corresponding to larger singular values, the representation of each

word is reduced as a smaller vector with only 300 dimensions. The new representation for the words (the reduced *U* matrix) is no longer orthogonal, but the advantage of this is that only the most important dimensions that correspond to larger singular values are kept. Each word now becomes a weighted vector on 300 dimensions, with only the most important dimensions that correspond to larger singular values being kept (Louwerse, Cai, Hu, Ventura, & Jeuniaux, 2006). The number of dimensions can be determined ad hoc, but we followed the trend set by most LSA studies and used 300 factors (Landauer & Dumais, 1997). Each term has now become a vector of dimensions. The semantic relationship between words can be estimated by taking the cosine between two vectors. What is so special about LSA is that the semantic relatedness is not (only) determined by the relation between words but also by the words that accompany a word (Landauer & Dumais, 1997). Landauer et al. (2007) present an extensive overview of studies showing that semantic similarity ratings by LSA are akin to human ratings.

In Study 1 we used LSA for geographical location estimates and word frequency for population estimates for 50 cities in the United States.

2.1. Materials

We selected the 50 largest cities of the United States and determined their longitude, latitude, and population by using the Census 2000 data from the U.S. Census Bureau (http://www.census.gov). These cities, with the corresponding states and Census regions, as well as their latitude, longitude, and populations, are listed in Table 1.

We used three newspaper corpora for the analyses: the *Wall Street Journal* (July 1994–December 1996), the *New York Times* (July 1994–December 1996) and the *Los Angeles Times* (May 1994–August 1996). Because of the large size of these corpora (in some cases over 500 MB), a random sample was taken and cleaned for tags and white spaces. Paragraphs with less than 100 words were removed from the corpus to provide LSA with sufficient contextual information. Details of the three corpora are presented in Table 2. Corpora were comparable in size, number of paragraphs, and news broadcasted coverage (1994–1996), but they differed in place of publication (New York City for the *Wall Street Journal* and the *New York Times* and Los Angeles for the *Los Angeles Times*).

2.2. Results and discussion

Results are presented in three separate sections: absolute estimates of geographical location, relative estimates of geographical location, and estimates of population size.

2.2.1. Absolute estimate

Latent semantic analysis spaces were created for each of the three corpora and cosine values were computed for each of the city pairs resulting in a 50×50 cosine matrix. This matrix was next submitted to an MDS analysis using the ALSCAL algorithm (SPSS 15.0.1 MDS procedure). Because cosine values indicate similarities rather than dissimilarities, distances were created from the data using the Euclidean distance measure. Default criteria

Table 1
Fifty U.S. cities used in Study 1–3

City	State	Census Regions	Latitude	Longitude	Population
Albuquerque	NM	West	35.1172	−106.6246	384,736
Arlington	TX	South	32.6945	−97.1275	261,721
Atlanta	GA	South	33.7629	−84.4226	394,017
Austin	TX	South	30.3059	−97.7505	465,622
Baltimore	MD	South	39.3008	−76.6106	736,014
Boston	MA	Northeast	42.3360	−71.0179	574,283
Charlotte	NC	South	35.1976	−80.8345	395,934
Chicago	IL	Midwest	41.8371	−87.6850	2,783,726
Cleveland	OH	Midwest	41.4797	−81.6785	505,616
Colorado Springs	CO	West	38.8632	−104.7599	281,140
Columbus	OH	Midwest	39.9889	−82.9874	632,910
Dallas	TX	South	32.7942	−96.7652	1,006,877
Denver	CO	West	39.7680	−104.8727	467,610
Detroit	MI	Midwest	42.3831	−83.1022	1,027,974
El Paso	TX	South	31.8493	−106.4375	515,342
Fort Worth	TX	South	32.7539	−97.3362	447,619
Fresno	CA	West	36.7806	−119.7929	354,202
Honolulu	HI	West	21.3173	−157.8042	365,272
Houston	TX	South	29.7687	−95.3867	1,630,553
Indianapolis	IN	Midwest	39.7764	−86.1462	731,327
Jacksonville	FL	South	30.3346	−81.6577	635,230
Kansas City	KS	Midwest	39.1223	−94.5520	584,913
Las Vegas	NV	West	36.2058	−115.2228	258,295
Long Beach	CA	West	33.7889	−118.1598	429,433
Los Angeles	CA	West	34.1121	−118.4112	3,485,398
Louisville	KY	South	38.2248	−85.7412	269,063
Memphis	TN	South	35.1056	−90.0070	610,337
Mesa	AZ	West	33.4177	−111.7403	288,091
Miami	FL	South	25.7757	−80.2108	358,548
Milwaukee	WI	Midwest	43.0634	−87.9666	628,088
Minneapolis	MN	Midwest	44.9619	−93.2668	368,383
Nashville	TN	South	36.1716	−86.7848	488,374
New Orleans	LA	South	30.0658	−89.9314	496,938
New York	NY	Northeast	40.6698	−73.9438	7,322,564
Oakland	CA	West	37.7715	−122.2246	372,242
Oklahoma City	OK	South	35.4671	−97.5135	444,719
Omaha	NE	Midwest	41.2639	−96.0117	335,795
Philadelphia	PA	Northeast	40.0068	−75.1347	1,585,577
Phoenix	AZ	West	33.5426	−112.0714	983,403
Portland	OR	West	45.5383	−122.6565	437,319
Sacramento	CA	West	38.5669	−121.4674	369,365
San Antonio	TX	South	29.4577	−98.5054	935,933
San Diego	CA	West	32.8150	−117.1358	1,110,549
San Francisco	CA	West	37.7933	−122.5548	723,959
San Jose	CA	West	37.3040	−121.8498	782,248
Seattle	WA	West	47.6218	−122.3503	516,259

Table 1
(Continued)

City	State	Census Regions	Latitude	Longitude	Population
Tucson	AZ	West	32.1958	−110.8917	405,390
Tulsa	OK	South	36.1278	−95.9164	367,302
Virginia Beach	VA	South	36.7394	−76.0437	393,069
Washington	DC	South	38.9051	−77.0162	606,900

Erroneously, Colorado Springs (ranked as the 55th largest city in the United States), Mesa (ranked 54), and Louisville (ranked 59) were included in the list of 50 largest cities, but San Juan (ranked 33), Pittsburgh (ranked 41), and Toledo (ranked 50) were not.

Table 2
Type, token, and paragraph counts of the three corpora

	Wall Street Journal	*New York Times*	*Los Angeles Times*
Token count	3,019,335	4,932,193	3,878,005
Type count	37,835	53,263	44,940
Paragraph count	42,738	35,299	30,979

were used with an S-stress convergence = .001, minimum stress value = .005, and maximum iterations 30. That is, the algorithm stopped iterating when the difference between stress values across iterations was less than the criterion, the stress value itself was less than the criterion, or when the maximum number of iterations was reached.

Following Borg and Groenen (1997) among others, we chose a low dimensionality in order to cancel out over- and underestimation errors in the proximities. The fitting of the data was acceptable with a two-dimensional scaling for all three corpora, as presented in Table 3. For all three corpora, Young S-stress improvement was less than .001 at six iterations.

Stimulus coordinates of all 50 cities were compared with the actual longitude and latitude. Consistently in all three corpora coordinates on Dimension 1 correlated with actual longitude, whereas Dimension 2 of the MDS loading correlated with actual latitude. Correlation coefficients between the corpora are presented in Table 4 and a scatter plot for the *Wall Street Journal* results in Figs. 1 and 2.

Although these correlations may help in understanding relations on separate one-dimensional axes, ideally we want to assess the relations between the bidimensional LSA-based

Table 3
Stress and R^2 scores for two-dimensional fitting of the 50 × 50 matrices of each of the three corpora

	Stress	R^2
Wall Street Journal	.321	.621
New York Times	.353	.512
Los Angeles Times	.344	.522

Table 4
Correlations between computational estimates and (1) latitude scores and (2) longitude scores by newspaper corpus

	Longitude and LSA-based Estimate	Latitude and LSA-based Estimate
Wall Street Journal	.607**	.322*
New York Times	.286*	.403**
Los Angeles Times	.324*	.437**

Note: $^{**}p < .01$, $^{*}p < .05$.

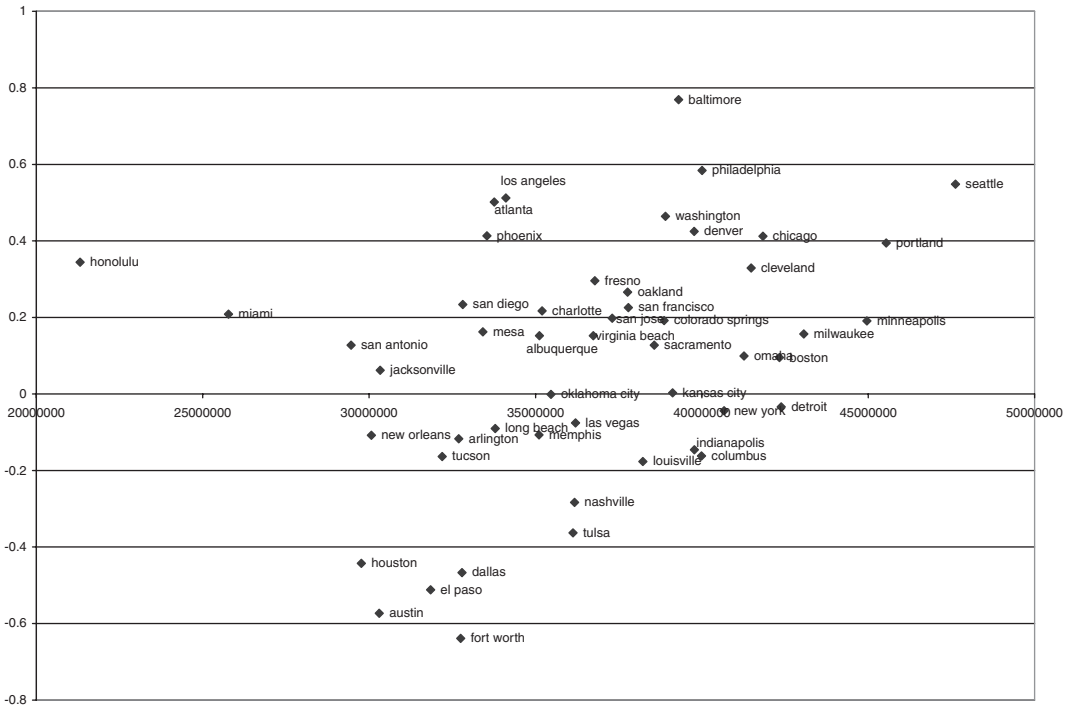

Fig. 1. Scatterplot longitude (X-axis) and MDS coordinates of LSA cosines for *Wall Street Journal* Corpus (Y-axis).

scores and the actual bidimensional maps. Friedman and Kohler (2003) and Tobler (1994) proposed a bidimensional regression method to perform such an analysis. In the case of a unidimensional regression an estimate is made of the regression of one variable on another variable using a regression line, whereby each value is shifted by slope (steepness of the line) and intercept (y value of the line when x equals zero). In a bidimensional regression each value of the dependent variable and the predicted value of the dependent variable are presented by a point in space, whereby vectors represent intercept and slope.

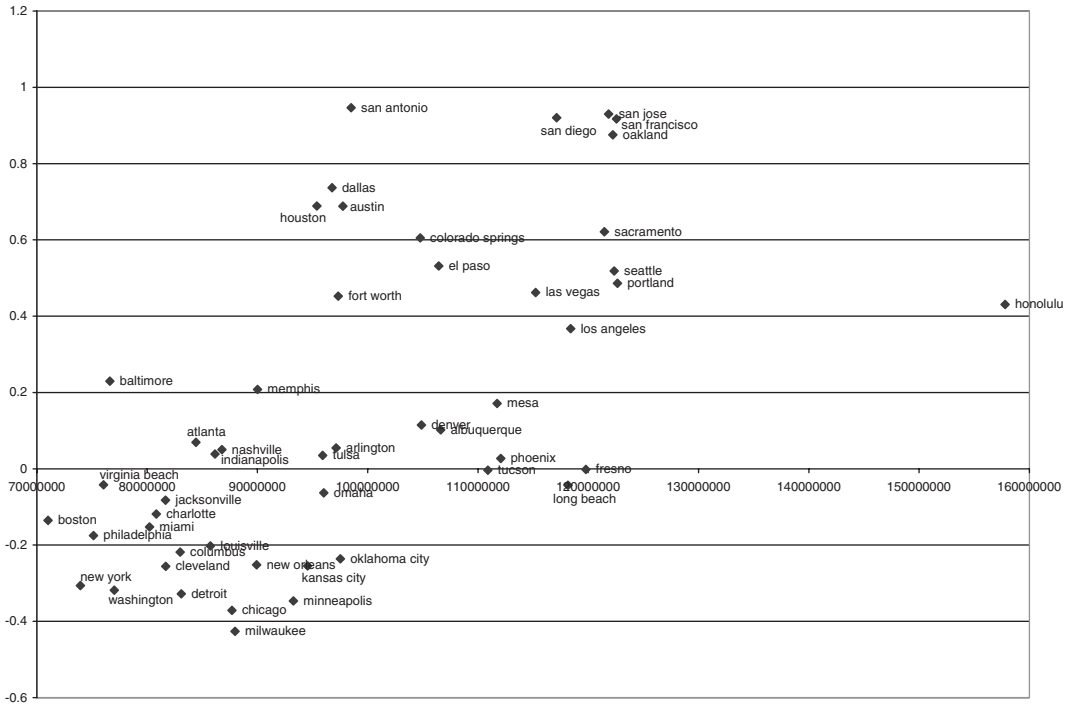

Fig. 2. Scatterplot latitude (X-axis) and MDS coordinates of LSA cosines for *Wall Street Journal* Corpus (Y-axis).

Bidimensional regressions were computed between the coordinates estimated by LSA and the actual coordinates. All three bidimensional regression analyses showed significant correlations for the three newspaper corpora between the computational findings and the actual coordinates (*Wall Street Journal*: $r = .529$, $p < .01$, $n = 50$; *New York Times*: $r = .277$, $p < .05$, $n = 50$; *Los Angeles Times*: $r = .427$, $p < .01$, $n = 50$). These results show that the three corpora provide estimates of geographical locations that correspond reliably with latitude and longitude of 50 largest cities in the United States.

2.2.2. Relative estimate

Earlier, we mentioned Tobler's (1970) proximity hypothesis stating that a participant's estimation bias should increase with the increasing physical distance from the participant's home town and Friedman et al.'s (2002) opposite findings that participant's estimation bias was highest for the area close to their residency. The current data also allow for testing the proximity hypothesis, because the newspapers come both from the Southwest and the Northeast Coast. The map of the United States was divided into four Census regions, as defined by the Census Bureau: West, Midwest, Northeast, and South. Table 1 gives the region the 50 cities belong to.

The prediction would then be that (1) *Wall Street Journal* and *New York Times* have similar estimates but differ from the *Los Angeles Times*; (2) based on the results in Friedman

et al. (2002) results, absolute errors for *Wall Street Journal* and *New York Times* should be higher in Midwest and Northeast regions than in West regions, whereas an opposite pattern is expected to be found for the West region.

Actual coordinates and computational coordinates were normalized by determining the relative position of the coordinates on a 100-unit scale. Absolute errors were computed by averaging the absolute values of the signed errors, which in turn were obtained from subtracting the LSA-estimated x- and y-coordinates from the actual x- and y-coordinates. An ANOVA on the absolute errors revealed a significant interaction between region and corpus ($F(6,138) = 3.12$, $p < .01$, $MSE = 517.42$, $\eta^2_p = .119$). As Fig. 3 shows, absolute errors were lowest in the Northwest and Midwest regions for the *Los Angeles Times* compared to the *Wall Street Journal* and *New York Times*. For the West region however the reverse pattern was found, with highest absolute errors for the *Los Angeles Times* compared to *Wall Street Journal* and *New York Times*. Furthermore, absolute errors were reliably different from zero for *Wall Street Journal* and *New York Times* but not for the *Los Angeles Times* in Midwest regions ($t(8) = 9.15$, $p < .001$ and $t(8) = 4.54$, $p < .002$ respectively). On the other hand, absolute errors were reliably different from zero for the *Los Angeles Times*, but not for the *Wall Street Journal* or the *New York Times* in the West region ($t(17) = 5.73$, $p < .001$). These results contradict the proximity hypothesis, but instead support the Friedman et al. (2002) hypothesis. While newspapers overall reliably predict longitude and latitude, they have an estimation bias for those areas in which these newspapers are published.

2.2.3. Population

The final hypothesis tested whether ''cities that are rated more are debated more.'' That is, to estimate the population of the cities, the word frequency of each of the cities was computed in each of the three corpora. Word frequencies are presented later in this paper (Table 9). Correlations between the word frequencies and actual population counts for these cities were high in all three corpora (*Wall Street Journal*: $r = .847$, $p < .001$, $n = 50$;

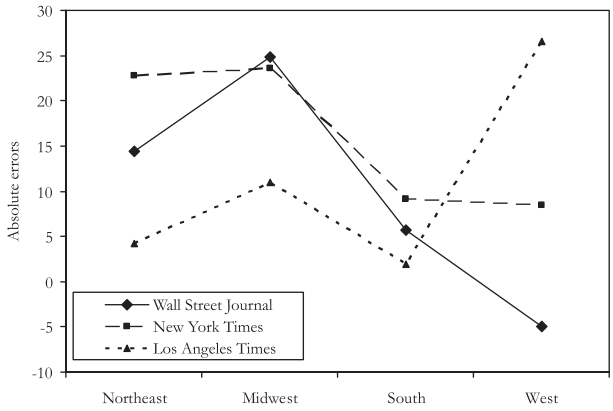

Fig. 3. Absolute errors (difference normalized latitude and MSD loadings from LSA findings) for the three newspaper corpora by region.

New York Times: $r = .839$, $p < .001$, $n = 50$; *Los Angeles Times*: $r = .612$, $p < .001$, $n = 50$). In sum, both the location hypothesis (in terms of absolute estimate and relative estimate) and the population hypothesis were supported for all three corpora.

2.3. Discussion

The results presented in this study support the ''cities that are located together are debated together'' and ''cities that are rated more are debated more'' hypotheses. The corpus analysis shows that absolute estimates of latitude and longitude correlate with actual longitude and latitude of the 50 cities, relative estimates show patterns similar to Friedman et al.'s (2002) findings and frequency of mention correlates with population counts.

It can be argued that what we have called *absolute estimates* are in fact *relative estimates* in the sense that MDS finds dimensions, not concepts like North and South, West and East. Consequently, the current method cannot distinguish actual locations from ones in which North and South or West and East were reversed. However, longitude and latitude estimates presented here approximate absolute locations, whereas distance estimates are relative and do not, in principle, require knowledge of the absolute locations (Friedman & Montello, 2006), warranting the current terminology. Conversely, in earlier work we compared distances between cities and LSA values (Louwerse et al., 2006) (relative estimates) but not their absolute location (absolute estimates).

The current results obtained using LSA in combination with MDS raise a number of important questions regarding the methods used.

First, in this analysis we used MDS, even though other analyses like Principal Component Analysis are available. There are a number of reasons for choosing MDS. Borg and Groenen (1997) pointed out that one advantage of MDS over PCA is that MDS establishes a direct relationship between the similarity measures and the geometric distance. Furthermore, we successfully used MDS with LSA cosine matrices in previous studies (Louwerse, 2007; Louwerse et al., 2006). Besides, a different data reduction method does not affect the analysis. When a PCA was applied on the cosine matrices from the three corpora, correlations were obtained that were similar to those found using MDS (Table 5).

Second, the argument can be made that there is redundancy in the current analysis. The word-by-paragraph matrix is reduced to a matrix with 300 dimensions using SVD. These 300 dimensions are then reduced to cosine values between pairs of cities, which are next placed into a two-dimensional space. The question can therefore be raised whether the

Table 5
Correlations between PCA and longitude and latitude

	Latitude—LSA	Longitude—LSA
Wall Street Journal	.322*	.607**
New York Times	.437**	.324*
Los Angeles Times	.442**	.320*

Note: $^{**}p < .01$, $^{*}p < .05$.

longitude and latitude dimensions obtained from MDS are not already present in the 300-dimensional space. To investigate this question, we looked at the 300 dimensions of each city name in the *U* matrix, one of the three matrices obtained after SVD from the original word-paragraph matrix *A*, as described earlier. The 50×300 values were compared with the actual longitude and latitude values in each of the three corpora. First, the values on the first three dimensions were compared with longitude and latitude, yielding correlational values considerably lower than the ones obtained from MDS (Table 7). Next, the correlations were conducted between each one of the 300 dimensions and the actual longitude and latitude. The dimensions that had the highest Pearson *r* values with the actual longitude and latitude were selected. Clearly, this is an undesirable approach because it cannot be determined a priori which dimension may best correlate with either one of the geographic coordinate measurements. Indeed, the highest Pearson *r* values were found for different dimensions across the three corpora ruling out that one dimension account for longitude or latitude[1] (Table 6). More importantly, the selected dimensions had Pearson *r* values that were at best comparable with the ones obtained using the MDS approach.

A third question regarding the results obtained in Study 1a concerns the correlational results in the light of word frequency. City names with a higher frequency in the corpus have more contextual information, which could either make the geographical location more reliable (more specific information) or less reliable (more noise). Like most linguistic data (Zipf, 1949), the frequency distribution of city names follows a power law. It does therefore not suffice to consider performance for *n* percentiles. Instead, we grouped the cities in three groups: those with low frequencies (about 70% of the data in each corpus), those with intermediate frequencies (about 20% of the data in each corpus), and those with high frequencies (about 10% of the data in each corpus). Overall, the results suggest cities that are mentioned infrequently in the corpus have lower correlations than those mentioned frequently in the corpus (Table 7).

In sum, evidence for the ''cities that are located together are debated together'' hypothesis can be obtained through LSA in combination with MDS. Similar results can be found with other data reduction techniques like PCA. Best results are obtained by using these data

Table 6
Correlations between LSA dimensions and longitude and latitude

	Dim. 1	Dim. 2	Dim. 3	Highest *r* Across 300 Dimensions	Dim. No. (1–300) With the Highest *r*
New York Times					
Latitude	.004	.168	.014	.368**	292
Longitude	.091	.242	.160	.413**	111
Wall Street Journal					
Latitude	.008	.140	.004	.273*	210
Longitude	.129	.195	.127	.489**	104
Los Angeles Times					
Latitude	.049	.033	.103	.291*	197
Longitude	.063	.154	.142	.468**	96

Note: $^{**}p < .01$, $^{*}p < .05$.

Table 7
Bidimensional correlations between LSA dimensions and longitude and latitude as a function of frequency of mention

	New York Times	*Wall Street Journal*	*Los Angeles Times*
Low frequency	.239 (33)	.493 (40)	.204 (37)
Medium frequency	.267 (10)	.632 (5)	.579 (10)
High frequency	.587 (7)	.657 (5)	.407 (5)

Note: Number of cities between parentheses.

reduction techniques on the cosines of all city pairings, rather than by considering the city names in a 300-dimensional space. Moreover, better results are obtained for cities that have an average to high frequency in the corpus. This further supports the ''cities that are located together are debated together'' hypothesis, in the sense that cities that are debated together more, allow for a more reliable determining of geographical location.

3. Study 1b: More corpus-based geographical estimates

The findings reported in Study 1a might also be explained by two confounding variables that have not been considered so far.

First, the correlations between longitude and latitude and the semantic information came from newspaper articles. Although Louwerse et al. (2006) have shown similar results based on textbook corpora, the argument could be made that newspaper corpora have a specific format that benefits LSA to produce cosine matches which ultimately correlate with longitude and latitude. That is, newspaper articles tend to start with the city source of the story immediately followed by the state name. Because LSA picks up on co-occurrences, the shared state names could explain why, for instance, *Memphis, TN* and *Nashville, TN* yield high cosine values. Evidence for such an explanation would of course not invalidate the current results, but would nevertheless weaken the evidence for the hypothesis that cities that are located together are debated together. Whereas the advantage of using newspaper articles is that geographical bias can be investigated, the disadvantage is that the newspaper format might bias results.

Secondly, LSA uses the cosines from higher order vectors, which makes the semantic analysis *latent*. That is, LSA may use state names or other semantic information to compute semantic similarities, but what the exact co-occurrence information is remains hidden in the 300 dimensions from which the cosine values are derived. Whereas the advantage of the latent higher-order co-occurrence method is that relatively small corpora can be analyzed with relatively powerful statistical techniques, the disadvantage is that the analysis remains somewhat opaque.

In short, what is desirable for additional support to the ''cities that are located together are debated together'' hypothesis is a transparent semantic analysis on a heterogeneous corpus.

3.1. Materials

In Study 3 we computed first-order co-occurrences of the 50 cities from Study 1 and 2 using the *Web 1T 5-gram* corpus (Brants & Franz, 2006). The corpus consists of 1 trillion word tokens (13,588,391 word types) from 95,119,665,584 sentences. This corpus is about 250,000 times larger than the newspaper corpora used in Study 1, allowing for first-order co-occurrences rather than higher-order co-occurrences as is the case in LSA.

3.1.1. Absolute estimate

Because some of the 50 cities consist of word pairs (e.g., *New York*), we only analyzed the frequency of co-occurrences of the 50 cities within four- and five-grams (window span of four and five words) and disregarded unigrams, bigrams, and trigram frequencies. The result, a 50×50 matrix of raw frequencies of co-occurrences was supplied to an ALSCAL algorithm to derive a Multidimensional Scaling (MDS) representation of the stimuli. Values were standardized on a -1 to 1 scale before the computation of proximities. The fitting of the data was poor (Stress = .402, R^2 = .232). Nevertheless, stimulus coordinates of all 50 cities were compared with the actual longitude and latitude. As in Study 1a, Dimension 1 correlated with actual longitude (r = .36, p = .01, n = 50), whereas Dimension 2 of the MDS loading correlated with actual latitude (r = .32, p = .02, n = 50). Bidimensional regressions showed similar results (r = .32, $p < .01$, n = 50).

These results are comparable, to those obtained using LSA (longitude: r = .32 vs. .29–.61 vs. .32; latitude: .32–.44 vs. .32; bidimensional regression: .28–.52 vs. .32), with one important difference, which is the strength of higher-order co-occurrence algorithms over a first-order co-occurrence algorithms: the corpus used in the current analysis was 250,000 times larger than the corpora used in the LSA analyses.

These results of this analysis provide additional support for the hypothesis that cities that are located together are talked about together by demonstrating that the findings in Study 1a cannot simply be attributed to the newspaper format or the LSA technique used.

4. Study 2: Human geographical estimates

Study 2 was conducted to address four questions. First, are human participants able to estimate city locations and population sizes? Second, do they show specific biases in their estimates? Third, how do their estimates compare to the corpus-based estimates? And fourth, to what extent can human estimates be predicted based on the corpus estimates?

4.1. Method

4.1.1. Participants

Twenty-eight University of Memphis undergraduate students (18 females, 10 males) participated in this experiment for course credit. All participants were natives of the United States all living in the city of Memphis or its suburbs.

4.1.2. Materials

The same 50 cities were used as in Study 1.

4.1.3. Procedure

Participants were asked to mark the location of all 50 cities used in Study 1 on a sheet of paper. For their convenience, an x-axis and y-axis were provided to mark meridian (line of longitude) and parallel (line of latitude) with the square through which the meridian and parallel was drawn representing the United States. No other information in terms of borders or states was given, though participants were allowed to draw an outline of the country on paper. After participants marked the location, they were asked to estimate the population for each city. Participants were told that the aim of the experiment was not to see whether they would receive a passing grade in geography, but to see how well people are able to estimate the relative coordinates and population size.

4.2. Results and discussion

4.2.1. Absolute estimate

All 28 participant-marked sheets were scanned into an electronic format. Software was written to manually identify each of the 50 points and automatically compute coordinates for these points. The average coordinates for the participant data correlated significantly with the actual longitude and latitude of the cities (longitude: $r = .806$, $p < .001$, $n = 50$; latitude: $r = .829$, $p < .001$, $n = 50$). Moreover, these results correlated with the results obtained from the newspaper corpora, as presented in Table 8, with significant correlations for longitude and for latitude.

As in Study 1, we conducted a bidimensional regression analysis to account for coordinates on a two-dimensional plane. Not surprisingly, the bidimensional regression again showed a significant relation between the coordinates of the 50 U.S. cities drawn by the participants and the actual longitude and latitude coordinates ($r = .562$, $p < .001$, $n = 50$), showing that human participants can reliably determine the location of these cities. Human bidimensional regressions also correlated with the computational findings, as shown in Table 8.

Table 8
Correlations between human and computational longitude and latitude estimates (by newspaper corpus)

	Human Estimates (Longitude)	Human Estimates (Latitude)	Human Estimates (Bidimensional Regressions)
Wall Street Journal	.588**	.400**	.497**
New York Times	.415**	.590**	.337**
Los Angeles Times	.411**	.456**	.427**

Note: **$p < .01$.

4.2.2. Relative estimate

In Study 1 we found evidence for Friedman et al.'s (2002) proximity hypothesis finding that an estimation bias decrease with the increasing physical distance from the source (i.e., participant's home town or source of newspaper). All participants in Study 2 were residents of Memphis, Tennessee, or its suburbs. According to the proximity hypothesis participants should therefore be more accurate in their estimates for these areas in the Midsouth than for areas elsewhere in the country. Alternatively, according to Friedman et al.'s (2002) findings and the findings in Study 1, estimates for the areas in the Midsouth are less accurate than those further away.

As in Study 1 coordinates were normalized on a 100-unit scale and absolute errors were computed by averaging the absolute values of the signed errors (actual coordinates minus the estimates). Differences were found between the four regions for latitude ($F(3,46) = 8.80$, $p < .001$, $MSE = 114.84$, $\eta^2_p = .365$). A Scheffé post-hoc analysis revealed that West and Midwest regions differed from South and Northeast regions, with only the latter two being reliably different than zero (South: $t(19) = 4.05$, $p < .001$; West: $t(17) = 9.35$, $p < .001$). These findings suggest that participants tend to overestimate latitude for the cities in South and West regions, but not for the Midwest and Northeast regions. Estimate differences between the four regions were also found for longitude ($F(3,46) = 22.478$, $p < .001$, $MSE = 95.252$, $\eta^2_p = .594$). Scheffé post-hoc comparisons revealed that cities in the West region differed from all other regions ($p < .05$), but no differences were found between these other three regions. Three regions reliably differed from zero (Midwest: $t(8) = 33.53$, $p < .001$; South: $t(19) = 3.42$, $p < .003$; West: $t(17) = 31.14$, $p < .001$), with participants underestimating longitude coordinates. The one exception, which approximated significance, was the Northeast region notably because of the small number of data points ($t(2) = 3.02$, $p < .09$).

As with Friedman et al.'s (2002) results and the results in Study 1, absolute errors for latitude suggest that participants' belief bias cannot be explained by a proximity hypothesis. On the contrary, estimation biases seem to be high for the areas close to participants' residence. Friedman et al.'s (2002) explained this bias by referring to cognitively based beliefs, geopolitically based beliefs, and socioculturally based beliefs. These beliefs may tie in with the recognition heuristic discussed earlier (Goldstein & Gigerenzer, 2002).

4.2.3. Population

Participants' population estimates correlated with the actual population estimates according to those reported by the Census Bureau ($r = .507$, $p < .01$, $n = 50$) as well as with the newspaper estimates (*Wall Street Journal*: $r = .667$, $p < .001$, $n = 50$; *New York Times*: $r = .685$, $p < .001$, $n = 50$; *Los Angeles Times*: $r = .687$, $p < .001$, $n = 50$), suggesting that participants can give an acceptable estimate of the population size of the 50 cities. Estimates are presented in Table 9.

4.2.4. Comparison between human and corpus-based estimates

To compare the human-actual and corpus-actual correlation coefficients for latitude and longitude estimates, we used the method as described in Blalock (1972). Longitude and

Table 9
Population estimates for the 50 cities

City	Population	*NYT*	*WSJ*	*LAT*	Estimate Subjects
Albuquerque	384,736	47	17	27	314,607.1
Arlington	261,721	63	22	34	404,428.6
Atlanta	394,017	927	220	164	1,265,357.1
Austin	465,622	73	125	38	896,214.3
Baltimore	736,014	191	105	195	719,821.4
Boston	574,283	1,679	504	237	1,671,214.3
Charlotte	395,934	170	55	42	446,642.9
Chicago	2,783,726	856	658	379	2,569,821.4
Cleveland	505,616	278	92	77	767,535.7
Colorado Springs	281,140	52	12	3	319,714.3
Columbus	632,910	102	57	45	462,714.3
Dallas	1,006,877	515	276	98	1,165,928.6
Denver	467,610	307	148	63	826,678.6
Detroit	1,027,974	317	131	83	1,359,928.6
El Paso	515,342	14	35	12	384,285.7
Fort Worth	447,619	67	46	12	467,857.1
Fresno	354,202	15	6	4	386,259.3
Honolulu	365,272	11	10	25	492,464.3
Houston	1,630,553	353	325	125	1,520,357.1
Indianapolis	731,327	101	38	17	531,428.6
Jacksonville	635,230	69	20	9	538,642.9
Kansas City	584,913	296	121	31	492,642.9
Las Vegas	258,295	209	75	113	1,510,785.7
Long Beach	429,433	46	9	18	417,678.6
Los Angeles	3,485,398	932	366	948	5,940,535.7
Louisville	269,063	43	45	24	517,142.9
Memphis	610,337	62	33	41	945,035.7
Mesa	288,091	34	31	18	269,500
Miami	358,548	488	134	201	1,197,500
Milwaukee	628,088	108	45	25	821,035.7
Minneapolis	368,383	92	137	23	757,321.4
Nashville	488,374	88	33	49	739,178.6
New Orleans	496,938	154	38	106	584,642.9
New York	7,322,564	4,586	2,403	1,844	4,099,107.1
Oakland	372,242	188	34	34	761,851.9
Oklahoma City	444,719	83	18	133	459,464.3
Omaha	335,795	16	23	17	287,357.1
Philadelphia	1,585,577	356	165	123	1,502,500
Phoenix	983,403	303	57	38	1,050,142.9
Portland	437,319	144	46	43	639,142.9
Sacramento	369,365	89	17	24	874,678.6
San Antonio	935,933	127	105	27	797,777.8
San Diego	1,110,549	201	94	185	970,178.6
San Francisco	723,959	458	270	265	1,151,607.1
San Jose	782,248	70	52	23	457,285.7
Seattle	516,259	304	138	116	125,175

Table 9
(Continued)

City	Population	*NYT*	*WSJ*	*LAT*	Estimate Subjects
Tucson	405,390	25	14	15	419,178.6
Tulsa	367,302	26	34	8	319,642.9
Virginia Beach	393,069	2	2	10	744,464.3
Washington	606,900	1,563	1,136	2,404	2,354,464.3

latitude correlations were significantly higher for humans than for the *Wall Street Journal*, *New York Times*, and *Los Angeles Times* (longitude: $z = 1.978, 3.933, 3.720$; latitude: 4.118, 3.672, 3.496), respectively. All differences were at the $p < .01$ level except for the *Wall Street Journal* correlations on longitude. For the bidimensional regressions no differences were found in the correlations between the human estimates and the three corpus estimates ($z = .274, 1.725, .897$ respectively). For population, corpus estimates were significantly higher than human estimates for the *Wall Street Journal* and the *New York Times*, though not for the *Los Angeles Times* ($z = 2.571, 3.097, .773$, respectively).

However, these results only concern correlations. In addition to these correlational results, it would be desirable to know more about the accuracy of the computational and human data. In order to compute absolute errors, as in the analyses reported earlier, data needed to be normalized by transforming actual latitude and longitude, the dimension loadings of the newspaper corpora, and the human data transformed to a 0–100 scale.

Absolute errors were significantly higher for latitude estimates than longitude estimates for the three newspapers (Wilcoxon $Z = -4.89$, $p < .001$, $n = 50$), but no such difference was found in participants. Less absolute errors were found in the newspaper corpora than in human participants for longitude (Mann–Whitney $U = 960$, $z = -1.99$, $p = .04$), but the opposite was true for latitude (Mann–Whitney $U = 1{,}193$, $z = -3.32$, $p = .001$). Results are presented in Table 10.

Accuracy for population estimates were computed by determining the absolute error of the participant performance and newspaper city frequencies compared to the actual population counts. Because the minimum frequency in the newspaper population data was 2, the minimum population estimate was 5,000, and the minimum city population 258,295, we multiplied newspaper frequencies by 1,000. Absolute errors for the three newspapers were lower than for the human estimates. This difference was significant (Mann–Whitney $U = 922$, $z = -2.261$, $p = .02$, $n = 100$) (See Table 10).[2]

Table 10
Mean (and *SD*) absolute errors newspaper and human estimates

	Newspaper	Participants
Population	648,636.3 (718,650.4)	923,899.8 (1,120,026)
Longitude	21.88 (8.03)	33.24 (20.77)
Latitude	44.53 (20.65)	30.72 (17.38)

4.3. Discussion

Study 2 was conducted to address four questions. The first question was to what extent human participants are able to locate the 50 largest cities in the U.S. cities on a map and estimate their population size. The results show that human participant estimates for longitude, latitude, and population size all correlated with the actual geographical data.

The second question was whether the human subjects showed specific biases in their estimates. The most consistent estimation biases were found for the West and South regions (in both latitude and longitude). Though the unequal group sizes increase Type I errors, no evidence was found for the proximity hypothesis that stated that residents would be more accurate in estimates closer to home. On the contrary, evidence was found for the less accurate estimates close to the residency of participants. Admittedly, the participants of this study came from a college population that is typically geographically heterogeneous. Nevertheless, the effects found for relative estimates were in line with Friedman et al.'s (2002) findings.

The third question asked how well the human participants fared in comparison with the corpora with regard to estimating longitude, latitude, and population. Our results indicate that the human estimates of longitude and latitude were significantly more highly correlated with actual locations than were the corpus-based estimates. On the other hand, two of the corpus-based sets of population estimates yielded higher correlations than did the human estimates. The overall results show human participants were more accurate in estimating the latitude of the 50 cities compared to the corpus data. However, for longitude the newspaper corpus estimates outperformed human estimates. The accuracy for population size was also higher for newspaper estimates compared to human estimates.

The fourth and final question was to what extent the corpus estimates predicted the human estimates. The results indicated that 16–35% of the latitude and 17–35% of the longitude variance in human location estimates was predicted by the corpus data and 44–47% of the variance in population estimates. These results suggest that human geographical estimates might be based in part on spatial information implicitly coded in language.

5. General discussion

Our results show that corpus-based analyses using word frequency and word co-occurrences provide estimates of relative geographical distances and population sizes that correspond reliably with population and latitude and longitude of 50 largest cities in the United States. These relative estimates also correlate with human estimates. Moreover, the corpus-based estimates exhibited the same estimation biases found in the human estimates. For humans, estimates were less precise for the region of residence than for other regions. Similarly, the newspaper corpora yielded more precise estimates for the West region by the *Wall Street Journal* and *New York Times* corpora than the *Los Angeles Times* and reverse patterns for the eastern regions.

It is far from obvious that computational linguistic techniques can predict geographical information. First, newspaper corpora do not explicitly discuss geographical information as

geography textbooks for instance may do. Hence, topic cannot account for the findings in this study. In fact, there is little that predicts a semantic similarity between cities. For instance, none of the 50 nearest LSA neighbors of the word *Seattle* happens to be among the nearest neighbors of the word *Portland*, suggesting that the shared semantic contexts between these two words are limited. Moreover, the geographical meaning of a word is at best ambiguous. For instance, there may be a semantic relation between the word *Washington* on the one hand and *Seattle* (LSA cos. = .39) and *Portland* (cos. = .13) on the other, but there is nothing that would distinguish the state *Washington* from the city *Washington*. A similar problem occurs with words like *New York*, which has a semantic relation with anything else that is novel, like a *new book* (cos. = .54), a *new house* (cos. = .57), a *new car* (cos. = .46), as well as with *New Mexico* (cos. = .52) and *New Orleans* (cos. = .88). Consequently, it is quite remarkable for even the slightest correlation to emerge between geographical coordinates and semantic similarity.

The argument could be made that the correlations between the MDS loadings based on the LSA cosine values and first-order co-occurrence frequencies can be attributed to a small number of cities that coincidentally have high co-occurrence matches. However, this argument is not supported by other research (Louwerse et al., 2006). To illustrate this further, 10 random samples of data points were taken from the *Wall Street Journal* estimates and the actual latitude and longitude. Bidimensional regressions on average showed a significant correlation between the sample of newspaper estimates and the actual coordinates (Mean $r = .68$, $SD = .12$).

More research is needed to determine the mechanisms behind the implicit encoding of geographical information in language and whether comprehenders use these language cues for their understanding of the geographical world. Goldstein and Gigerenzer (1999, 2002) argued that textual information fulfills a mediator function, particularly when factual information is not readily accessible. Our data suggest that between 16% and 35% of the latitude and longitude variance in human location estimates and 45% of the variance in population estimates can be attributed to linguistic coding. Elsewhere (Louwerse, 2008; Louwerse et al., 2006; Zwaan & Madden, 2005, pp. 227–230) we have argued that language is organized such that it reflects semantic relations in the physical world. Prelinguistic conceptual knowledge (e.g., geographical proximity) used when speakers formulate utterances gets translated in linguistic conceptualizations (collocations) (Levelt, 1989), so that as a function of language use, geographical information becomes encoded in language. Conversely, in going from text to mental representation, comprehenders form situation models (Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998), which in ideal cases correspond in essential ways to the mental representation that the speaker started out with. It is possible that our participants based their estimates in part on situation models. This would suggest that the language effect is mediated in part by conceptual representations. However, this account is clearly speculative.

Our results suggest that a remarkable amount of geographical information can be extracted from large corpora of newspaper articles texts that did not explicitly discuss such information. It is this aspect that makes the results relevant for the study of human spatial cognition, as it suggests that a significant amount of geographical

information can be acquired incidentally by reading a large number of texts that are about topics other than geographical distance or population size, because language encodes geographical information.

Notes

1. It is worth pointing out that the dimensions with the highest correlations for latitude were consistently higher (200 range) than the highest correlations for longitude (100 range), possibly suggesting a latent clustering of information that correlates with the geographical location.
2. Even when the newspaper frequencies were not multiplied by 1,000, absolute errors were still lower for newspaper frequencies than for human estimates (M = 810,893.8, SD = 1,110,237 vs. M = 923,899.8, SD = 1,120,026). This difference, however, was not significant (U = 1,115, z = −.931, p = .352, n = 100).

Acknowledgments

This research was supported by grant NSF-IIS-0416128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution.

References

Baird, J. C. (1970). *Psychophysical analysis of visual space*. Oxford, England: Pergamon Press.

Blalock, H. (1972). *Social statistics*. New York: McGraw-Hill.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer-Verlag.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

Canter, D., & Tagg, S. (1975). Distance estimation in cities. *Environment and Behavior*, *7*, 59–80.

Dumais, S. T. (2007). LSA and information retrieval: Getting back to basics. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 293–321). Mahwah, NJ: Erlbaum.

Ehrich, V., & Koster, C. (1983). Discourse organization and sentence form: The structure of room descriptions in Dutch. *Discourse Processes*, *6*, 169–195.

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, *21*, 296–306.

Ferguson, E. L., & Hegarty, M. (1994). Properties of cognitive maps constructed from text. *Memory & Cognition*, *22*, 455–473.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General*, *119*, 63–76.

Freundschuh, S. M., & Mercer, D. (1995). Spatial cognitive representations of story worlds acquired from maps and narrative. *Geographical Systems*, *2*, 217–233.

Friedman, A., Kerkman, D. D., & Brown, N. (2002). Spatial location judgments: A cross-national comparison of estimation bias in subjective North American geography. *Psychonomic Bulletin & Review*, *9*, 615–623.

Friedman, A., & Kohler, B. (2003). Bidimensional regression: A method for assessing the configural similarity of cognitive maps and other two-dimensional data. *Psychological Methods*, *8*, 468–491.

Friedman, A., & Montello, D. R. (2006). Global-scale location and distance estimates: Common representations and strategies in absolute and relative judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 333–346.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, *73*, 407–420.

Kerkman, D. D., Friedman, A., Brown, N., Stea, D., & Carmichael, A. (2003). Development of geographic categories and biases. *Journal of Experimental Child Psychology*, *84*, 265–285.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Levelt, W. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107–120). Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2008). Embodied representations are encoded in language. *Psychonomic Bulletin and Review*, *15*, 838–844.

Louwerse, M. M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively inspired natural-language based knowledge representations: Further explorations of latent semantic analysis. *International Journal of Artificial Intelligence Tools*, *15*, 1021–1039.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35–55). Mahwah, NJ: Erlbaum.

Montello, D. R., & Freundschuh, S. M. (1995). Sources of spatial knowledge and their implications for GIS: An introduction. *Geographical Systems*, *2*, 169–176.

Morrow, D. G., Greenspan, S. L., & Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *Journal of Memory and Language*, *26*, 165–187.

Perrig, W., & Kintsch, W. (1985). Propositional and situational representation of text. *Journal of Memory and Language*, *24*, 503–518.

Stevens, A., & Coupe, P. (1978). Distortions in judged spatial relation. *Cognitive Psychology*, *10*, 422–437.

Taylor, H. A., & Tversky, B. (1992a). Descriptions and depictions of environments. *Memory and Cognition*, *20*, 483–496.

Taylor, H. A., & Tversky, B. (1992b). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, *31*, 261–282.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography Supplement*, *46*, 234–240.

Tobler, W. (1994). Bidimensional regression. *Geographical Analysis*, *26*, 186–212.

Tversky, B. (1981). Distortions in memory for maps. *Cognitive Psychology*, *13*, 407–433.

Tversky, B. (1997). Spatial constructions. In N. Stein, P. Ornstein, B. Tversky, & C. Brainerd (Eds.), *Memory for emotion and everyday events* (pp. 181–208). Mahwah, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, B., Lee, P. U., & Mainwaring, S. (1999). Why speakers mix perspectives. *Journal of Spatial Cognition and Computation*, *1*, 399–412.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In D. Pecher, & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 224–245). Cambridge, England: Cambridge University Press.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185.