



Published in final edited form as:

Psychol Sci. 2015 September ; 26(9): 1449–1460. doi:10.1177/0956797615591771.

Concreteness and Psychological Distance in Natural Language Use

Bryor Snefjella and

McMaster University, Canada

Victor Kuperman

McMaster University, Canada

Abstract

Existing evidence shows that more abstract mental representations are formed, and more abstract language is used, to characterize phenomena which are more distant from self. Yet the precise form of the functional relationship between distance and linguistic abstractness has been unknown. In four studies, we test whether more abstract language is used in textual references to more geographically distant cities (Study 1), times further into the past or future (Study 2), references to more socially distant people (Study 3), and references to a specific topic (Study 4). Using millions of linguistic productions from thousands of social media users, we determine that linguistic concreteness is a curvilinear function of the logarithm of distance and discuss psychological underpinnings of the mathematical properties of the relationship. We also demonstrate that gradient curvilinear effects of geographic and temporal distance on concreteness are near-identical, suggesting uniformity in representation of abstractness along multiple dimensions.

Keywords

psychological distance; construal-level theory; embodied cognition; social media; twitter; abstraction; concrete

Introduction

One of the fundamental and unique abilities of the human mind is to transcend the boundaries of here and now: to imagine distant times, far away places, and other people. The psychological mechanism of *abstraction* that underlies this mental ability is a matter of continuing debate (Paivio, 1990; Schwanenflugel et al., 1988; Boroditsky & Ramscar, 2002; Gallese & Lakoff, 2005; Barsalou, 2008; Fischer & Zwaan, 2008; Meteyard et al., 2012; Burgoon et al., 2013). Yet social psychologists have noted a positive correlation between the perceived distance of an object or event, and the level of abstraction at which that event is

Corresponding author: Bryor Snefjella, Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 626, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M2, Phone: 905-525-9140, x. 20384, snefjebn@mcmaster.ca.

Authorship Contributions

Bryor Snefjella developed the study concept. Both authors contributed to data collection, statistical analysis, and interpretation. Bryor Snefjella drafted the manuscript, with revisions by Victor Kuperman. All authors approved the final version of the manuscript for submission.

represented mentally. For instance, the influential construal-level theory of psychological distance (henceforth, CLT; Trope & Liberman, 2010, 2003) states that objects and events that are proximal (close to an egocentric self) are represented with rich, complex, concrete, contextual, and subordinate features. This is referred to as a low-level construal. A high-level construal is the tendency to represent distal objects and events abstractly, by their simple, invariant, superordinate characteristics. If I am preparing a lecture for tomorrow, I'll worry about which room to go to. When preparing a lecture for next month, I'll worry about its topic. Within the CLT, the distance-driven differences in construal arise because abstract representations and goals are more stable over time (the topic of my lecture remains the same, even if the location changes). Thus, abstraction leads to successful traversing across psychological distances (Trope & Liberman, 2010).

The relationship between abstraction and psychological distance is implicated in many personal and social phenomena, including the consistency of attitudes and evaluations in an individual (Ledgerwood et al., 2010), the actor-observer bias (Nisbett et al., 1973), moral judgments (Amit & Greene, 2012), politeness (Stephan et al., 2010), subjective judgments of truth (Hansen & Wanke, 2010) and consumer preferences (Fiedler, 2007). The hypothesized positive correlation between abstractness of mental representations and psychological distance has received support in a many experimental paradigms and measures, including action identification (Liviatan et al., 2008; Fujita et al., 2006), a "distance Stroop" task (Bar-Anan et al., 2007), and surveys and questionnaires (Wakslak et al., 2006; Eyal et al., 2004; Trope & Liberman, 2000): for a recent meta-analysis of research into the CLT see Soderberg et al. (2014). Of greater importance for this paper are studies that capitalized on the ability of language to reflect abstractness of mental representations through abstractness of expressed meanings (Paivio, 1990; Schwanenflugel et al., 1988). These tasks elicited linguistic productions from participants, while manipulating the distance of what they were prompted to write about. Psychological distance of described phenomena, typically conceptualized as their construal level, was then operationalized as relative abstractness of produced texts. A robust finding across the studies was that more abstract language is used to characterize phenomena which are more distant from the self temporally, spatially, or socially, or are more hypothetical (Trope & Liberman, 2010).

While valuable, the current methodology of using linguistic productions to study the link between abstraction and psychological distance is limited. A typical laboratory experiment prompts small groups of undergraduate participants to write about distant/near objects or events, often resulting in small samples of language obtained from individuals with a relatively homogenous age range and experience. The scale of data collection is further limited by labor-intensive manual coding of linguistic abstractness. For instance, Fujita et al. (2006), Stephan et al. (2010), and Gong & Medin (2012) coded the writing of participants using the Linguistic Categorization Model of Semin & Fiedler (1988). In a similar vein, Alter & Oppenheimer (2008) used three human coders to rate productions of participants as either being abstract, concrete, or both.

Even more drastically, instead of being treated as a continuum, abstractness is routinely binned into discrete categories. As the meta-analysis of CLT research further shows, distance is dichotomized into "close" and "far" categories in all 267 studies that met the

authors' inclusion criteria (Soderberg et al., 2014). Either categorization obscures the precise mathematical form of their functional relationship and prevents characterization of the effect of abstractness on psychological distance in a graded way.

The present study addresses these drawbacks by using norming megastudies with ratings of semantic word properties as well as vast collections of language productions in text corpora. We use these resources to examine — on a larger scale and with a broader range of examples — abstractness of linguistic productions that describe objects or events positioned at various psychological distances. In our study, operationalization of construal level relies on a recent data set of concreteness ratings of 40,000 English words (Brysbaert et al., 2014). Ratings to words were given on a scale from 1 (abstract) to 5 (concrete) and averaged over 30 participants: resulting concreteness norms ranged from 1.04 (*essentialness*) to 5 (*pitbull*). Using this data set, we can measure abstractness in language without human coders, and with any number of productions.

By handing the drudgery of coding to a computer, we are able to study psychological distance at scales that were not previously possible, i.e. millions of observations from thousands of language speakers varying in age, socio-economic status, personality traits, or place of residence. Furthermore, we are analyzing natural language use, not language elicited in an experimental task; our study is a test of the ecological validity of the CLT and is complementary to experimental research. The meta-analysis by Soderberg et al. (2014) pointed to the inability of current studies of psychological distance to identify the precise mathematical form of the gradient functional relationship between distance and abstractness. Using clever analytical techniques, Soderberg et al. (2014) predicted the relationship to be curvilinear. Our approach charts the relationship along a continuum of psychological distances, and provides other psychologically meaningful interpretations of its mathematical properties. Limitations of our automated analysis are addressed in the General Discussion.

We present four studies, each examining one of the critical dimensions of the CLT and based on social media sources. Specifically, we test whether more abstract language is used in textual references to more geographically distant cities (Study 1), times further into the past or future (Study 2), and references to more socially distant people (Study 3). In study 4, we examine if a theme that it is commonly experienced across all times and distances – death and dying – follows the pattern observed in the aggregate data.

Study 1: Geographic distance

Methods

This study explores the role of geographic distance in explaining variability in the level of construal of US cities, where construal is operationalized as concreteness of language used in relation to the cities. We selected the 30 most populous US cities (with 600,000 inhabitants as an arbitrary lower threshold) using rankings of their population size in 2013 (http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population), with the exception of Washington, DC, as the city name is homonymous with a name of a geographically distant state. New York City and Oklahoma City, while homonymous with respective states, are embedded in those states and thus are expected to introduce less noise

in distance estimates. Social media — a source of millions of data points, a broad selection of geographic objects and a full range of possible distances — enables an expansion and refinement of prior experimental studies. Using the publicly available data stream of Twitter API at <https://dev.twitter.com>, we collected tweets that were (a) geo-tagged (i.e. had GIS coordinates of the location where the tweet was produced), (b) were sent from within the US, and (c) contained a name of one of these major US cities. To calculate the geographical distance between the location of the tweet production and the city of reference, we used the latitude and longitude coordinates of that tweet and of that city's center, as supplied by the `geocode()` function of the `ggmap` package (Kahle & Wickham, 2013) of the R statistical software v. 3.01 (R Core Team, 2013). We applied the Haversine formula to obtain the great circle-distance between these two points. To remove distributional skewness, all distances were log-transformed (base 10). All calculations and analyses in this and subsequent studies were made using the R software (R Core Team, 2013). It is possible that a more psychologically valid measure of geographic distance is the time it takes to commute from the tweet location to the respective city center.¹ We opted for the great circle-distance because estimates of the commute time (including e.g. waiting time in airports and traffic hours) are inherently variable.

A total of 712198 tweets satisfying our criteria were collected between March and May 2014. Twitter is trendy and dynamic; collecting tweets over a wide time frame helps prevent trending topics from influencing the results. A further trimming retained tweet messages that contained four or more words (excluding the city name) which had concreteness ratings available in Brybaert et al.'s data set: an average message contained 8.42 such words ($sd = 3.9$). This reduced the pool to 478920 geo-tagged messages. For each message, we calculated its mean concreteness based on words with available ratings ($M = 2.83$, median = 2.78, range [1.46, 4.92], $sd = 0.47$). Thus, the degree of construal was operationalized as a prevalence in the tweet message of words that were rated in Brysbaert et al.'s study as more concrete or more abstract when presented out of context.

Geographic distances between the point of origin of tweets and cities they refer to ranged from 0.1 km to 4220.9 km ($M = 516.7$, median = 24.3, $sd = 885.7$), see bottom of Figure 1A for the distribution of distances). To decrease noise in the raw data, we binned observations into percentiles of the log distance distribution and calculated mean concreteness of each bin. A cubic function ($y = 3.05 + 0.03 * x - 0.18 * x^2 + 0.04 * x^3$) provided an optimal polynomial fit to concreteness in a linear regression model with raw polynomials of log distance as predictors (adjusted $R^2 = 0.88$), see Figure 1A. Table 1 reports the goodness-of-fit of the cubic function and compares it to lower and higher order polynomial functions using hierarchical regression: successive models are compared using function `anova()` in the R statistical software. While a cubic function is the best *polynomial* fit to the data in Studies 1 and 2 (see below), other functional relationships might offer a similarly good fit (e.g. logistic curve). In the absence of a theoretically grounded expectation as to the form of the curvilinear dependency, we allow for the possibility of alternative functional forms.

¹We thank Emmanuel Keuleers for raising this point.

The pattern in Figure 1A indicates a substantial decrease in concreteness of tweets regarding cities with an increase in the distance of the tweeting person from that city: the drop in concreteness between the extreme distances is estimated at 0.5 points of concreteness or 12.5% of the concreteness scale. This pattern, based on nearly half a million of observation points and thirty cities, is perfectly in line with the predictions and experimental validations of the positive correlation between psychological and geographic distance. Moreover we confirm the predictions of Soderberg et al. (2014) that construal and distance have a curvilinear relation.

The functional form of the fitted curve, and its excellent goodness-of-fit to the data, enables us to further interpret parameters of the cubic function. The inflection point, i.e. the point at which the second derivative of the function changes its sign, is estimated at 1.5 (in log 10 units) or around 30 km from the city center. This implies that as a typical tweeting person moves away from the city center to this point, the concreteness of her references to the city decreases at a relatively high rate. This decrease becomes less precipitous as distances increase beyond 30 km and, as the first derivative of the polynomial shows, there is little to no decrease in concreteness associated with distances above 100 km.

We further speculated that the inflection point is psychologically meaningful and demarcates a distinction between (a) being within a city, where concreteness of mental representation of that city is the highest (the level of construal the lowest) and decreases sharply from the city center to the outskirts, and (b) being outside of the city where the representation of the city is more abstract overall (the level of construal is higher) and is less affected by how far the Twitter user is from the city. To test this hypothesis of immediacy of experience, we calculated the radius of the city area for each of 30 cities, under a simplifying assumption that cities have a perfect circular shape: estimates of the city area were obtained from <http://www.citymayors.com/statistics/largest-cities-area-250.html>. Extreme radii were found for New York, NY ($r = 53$ km), and El Paso, TX ($r = 13$ km), and the mean radius was 25 km, close to the inflection point of the fitted curve at 30 km. While more sophisticated measurements of the urban territory will be necessary, the observed value is consistent with the notion that the construal level increases (and concreteness of language decreases) more drastically as the speaker loses the immediacy of the urban experience when moving from the city center to its outskirts: once outside the city, the construal level is more stable and high.

Study 2: Temporal distance

Methods

A robust finding in the literature on memory and prediction in relation to psychological distance is that remote events, whether in the past or the future, elicit a higher level of construal (Trope & Liberman, 2003, 2000). To test this dimension of psychological distance, we used the USENET corpus by Shaoul & Westbury (2013) consisting of over 7 billion word tokens of public USENET postings collected from 47860 English language news groups between Oct 2005 and Jan 2011. Several temporal terms were used to examine effects of temporal distance on concreteness of language in which past and future events are described. We explored distance both within specific time units (e.g. 10 years ago vs 100

years ago) and between units (e.g. days from now vs centuries from now; last week vs next week). In the remainder of this section, we describe the data collection procedure for each of the three types of time references separately, and present the results in the next section.

Study 2.1, “years ago”—Soderberg et al. (2014) predicted the curvilinear relationship between distance and construal on the basis of their meta-analysis, which placed different studies of temporal distance along an objective timeline from 1 to 365 days. We were unable to recreate this finding in the corpus, as USENET contributors almost never refer to distances over 10 days with the phrase *X days ago*. However, phrases such as *X years ago* yielded intriguing results, reported below.

We identified all occurrences in the corpus of the phrase *X years ago*. We further extracted 5 words to the left and the right of the critical phrase, i.e. italicized fragments in *wind generation of electricity 30 years ago and they were commonplace then*. The 10-word window around the target word was chosen to approximately equate the number of words in the context between Twitter (with its 140 character limit per tweet) and USENET emails. We leave for further study the question of which window size provides the best accuracy. Numerals preceding the target phrase (spelled either as thirty or 30) served as the metric of the temporal distance from the time of email submission to USENET. We restricted the time range to 1 through 999 years ago. Finally, we removed all contexts that contained fewer than 4 words with concreteness ratings available in Brysbaert et al.’s dataset.

Study 2.2a, “ago” vs “from now”—We were also interested in comparing abstractness with which people refer to distances to the past or future, as measured by different time units. In the corpus, we identified all occurrences of phrases *X ago* and *X from now* where X was a time unit: minute, hour, day, week, month, year, decade, or century. We further extracted 5 words to the left and the right of the critical phrase, i.e. italicized fragments in *situation as recent as two centuries ago when much academic instruction was*. Numerals (spelled either as two or 2) were removed from the preceding context window. The resulting scale of time units was then ordinal (a week ago is more in the past than a day ago), rather than continuous, as in Study 2.1. Finally, we removed all contexts that contained fewer than 4 words with concreteness ratings available in Brysbaert et al.’s data set.

Study 2.2b, Between time units, “last” vs “next”—To ensure that observed differences between time units are not an artifact of our choice of the language denoting temporal distance (time-units from now and time-units ago), we conducted an additional set of analyses using contexts surrounding phrases like *yesterday, tomorrow, last/next week, month*, etc. Contexts were defined as above, and the trimming procedures were the same as Study 2.2a.

Results

Study 2.1, within a time-unit: years ago—A total of 265859 occurrences containing *years ago* were identified in the USENET corpus. Due to skewness, temporal distances in years were log (base 10) transformed. Observations were binned by their temporal distance into 36 intervals with open left boundaries formed by numbers 1 to 19 (incremented by 1),

20 to 90 (incremented by 10), and 100 to 1000 (incremented by 100), and closed right boundaries. The histogram of the distribution of temporal distances is shown in the bottom of Figure 1B. Mean concreteness of contexts was calculated for each bin, and plotted against the log 10 of the numeral in the interval's left boundary, see Figure 1B.

As with geographic distance, the best polynomial fit to concreteness was obtained with a cubic curve ($y = 2.44 + 0.11 * x - 0.12 * x^2 + 0.03 * x^3$) and showed an excellent fit in a linear regression model with raw polynomials of log temporal distance (adjusted $R^2 = 0.80$), see Table 1 for model comparison. Verbal descriptions of past events were the more abstract (i.e. construed at a higher level) the more years have passed since the described event. The drop in concreteness between the extremes of the temporal range was fairly small and amounts to approximately 0.1 units of concreteness, or 2.5% of the concreteness scale. Again, the observed pattern converged with the experimental evidence of the construal-level theory of psychological distance, where construal is operationalized as concreteness of verbal description of the event. This also shows that the predicted curvilinear relationship (Soderberg et al., 2014) between construal and distance holds for multiple dimensions of psychological distance.

A further inspection of the functional curve pointed to a faster drop in concreteness in verbal representations of events up to the inflection point of 1.57 log units (or 37 years ago), a slower decrease in concreteness when more distant events past the inflection point, and virtually no change in concreteness of contexts associated with events taking place 200 to 1000 years ago. In analogy to geographic distance, we speculated that the inflection point demarcated a change in the immediacy of one's experience with events that happened during one's life time and those that preceded it. As the literature on collective and generational memory demonstrates, critical social events (wars, natural disasters, political regime changes) are more salient in mental representations of the past of those individuals who had an exposure to the event as it happened (Pennebaker et al., 2013). If true, we would expect the inflection point of the functional curve (37 years of age) to be close to the age of a typical contributor to the USENET corpus. The median age of the US population is 37.6 years (Central Intelligence Agency, 2014) and although we could not find age data on USENET users, the available statistics on internet users did not diverge from this number (e.g. average age of social media users in 2012 was 37.9, Pingdom, 2014, see also Eisenstein, 2015). Again, the functional form of the effect of temporal distance on concreteness of language production suggests immediacy of one's experience with events as an important factor in the construal level of mental representation and linguistic expression of those events.

Study 2.2a, between time units: ago vs from now—A total of 767842 critical phrases and surrounding contexts were identified with time units (minutes to centuries) followed by *ago* or *from now*. After trimming, the data pool contained 698391 contexts. Mean concreteness was calculated for each context and plotted against respective time units. Figure 1C summarizes the functional relationship between temporal distance from *now* (the time of writing of the posting) and the level of construal of the temporally marked event.

Figure 1C indicated a near-linear decrease in concreteness for events that are further away from the present on the ordinal scale of time units, in convergence with the hypothesized relationship between temporal and psychological distance. The maximum contrast between time units (hours ago and centuries ago) was 0.2 units of concreteness, corresponding to 5% of the concreteness scale. Regression models further indicated large effect sizes ($R^2 = 0.89$ and 0.55 for past and future, respectively). The patterns also showed a preference for talking about past rather than future events, as evidenced in the circle sizes, proportional to log frequency of the phrase occurrence in the corpus. The intercepts of the regression lines further suggested that overall past experiences are represented in more detail (higher concreteness) than events that are envisioned in the future in line with experimental research into the mental simulation of past and future events (DArgembeau & Van der Linden, 2004; Johnson et al., 1988).

Study 2.2b, between time units: last vs next—There were a total of 1025121 contexts surrounding phrases like *last month vs next month*. Figure 1D summarizes the results for all time units. We again noted a decrease in concreteness as temporal distance from the present increased. However, with these key phrases, the past and future appear to be (almost) mirror images of each other, similar both in the slopes of the regression lines ($\beta = 0.08$ for the past, and $\beta = -0.08$ for the future events), the amount of explained variance ($R^2 = 0.89$ and 0.91 respectively), and (log10) frequencies of occurrence of respective phrases, shown as circle sizes in 1D. Also, the contrast between maximally different time units (*yesterday/today* and *last/next century*) was much larger than in the comparison above (*days ago/from now vs centuries ago/from now*) and amounted to 0.4 units of concreteness, or 10% of the concreteness scale.

Study 3: Social distance

Methods

A series of experiments (Liviatan et al., 2008; Stephan et al., 2010) have shown that psychological distance between individuals is perceived to be larger, and the level of construal higher, if a social relationship between those individuals is more distant. To operationalize closeness of social relationships in a corpus, we took as a point of departure the Bogardus social distance scale (Bogardus, 1922), see also Parrillo and Donoghue (2005). The scale evaluates the degree of willingness to establish social contacts with representatives of a racial, ethnic, socio-economic, occupational or other social group. The scale identifies closeness as the individual's willingness to accept the group representatives on a seven-point scale: as (1) potential partners in marriage, (2) close friends, (3) neighbors on the same street, (4) co-workers in the same occupation, (5) citizens in the same country, (6) only visitors to his/her country, or (7) people to be excluded from his/her country. To adapt the scale to the observational data at hand, we converted the scale from a cumulative one (i.e. agreement with a higher degree of closeness implies agreement with all lower-degree categories) to a discrete ordinal one by identifying terms belonging to each of the scale's categories: e.g. friend, ally, confidant, pal, chum, buddy for category (2) and compatriot, countryman, countrywoman for category (6). The full list of 39 terms – created

using our linguistic intuitions and the Merriam-Webster thesaurus (<http://www.merriam-webster.com/>) – is reported in Table 2.

Results

In the same procedure as one used in Study 2, we extracted from the USENET corpus contexts of five words on each side of a target word. The resulting pool of contexts with four or more words with available concreteness ratings contained 422553 data points. Figure 1E plots mean concreteness of the contexts, grouped by social distance categories, against the ordinal scale of social distance. While the number of data points and confidence intervals vary across categories, the overall trend is in agreement with the hypothesized link. Groups of individuals that are considered more distant socially are also construed in less concrete terms, with a maximum contrast of 0.15 points (about 4% of the available concreteness scale) between categories (1) family members and (7) foreigners.

Study 4: Concreteness of the theme of death over time and geographic distance

Methods

Two points of criticism can be raised with regard to Studies 1–3. First, our currency is in aggregate measures of the concreteness of verbal contexts, which gloss over a multitude of phenomena and a diversity of personal and collective experiences, and thus might lead to ecological fallacy (Robinson, 1950). Second, there are alternative explanations to why a person might choose more abstract over more concrete words when describing a remote phenomenon. It might be due to the linguistically faithful reflection of the distance-driven change in one's mental representation, which would be consistent with the CLT premises. Conversely, it might take place because one does not have a direct experience with the phenomenon and only has access to its gist-like representation through language and thus can only describe it in abstract terms: this relative abstractness is not expected to vary with the distance, only with the amount of experience². This study considers construal of events related to death as a function of geographic and temporal distance. Death is a concept that is acquired early, is salient and memorable as an event, and occurs to all living beings, at all times and all locations, giving on average an equal probability to directly or indirectly experience (somebody else's) death at all distances from the self. Finding a curvilinear relationship between concreteness of texts constrained to a familiar, ubiquitous event and distance from the self would make a step towards ensuring that the aggregate patterns are made of converging individual patterns, and that – at least in some cases – predictions of the CLT are due to the change in distance, and not only to the change in the strength of personal experience.

To address these issues, we extracted 354 Twitter messages containing words *died*, *dead*, or *death* from the data pool of Study 1, and 3735 contexts from USENET emails containing same words and a target phrase *years ago* in Study 2.1. Mean concreteness of those tweets and those contexts was calculated for each bin of log geographic distance (in km) and log

²We are indebted to Emmanuel Keuleers for raising this point.

temporal distance (in years), with bins defined as in Study 1 and Study 2.1 respectively. A similar study of social distance was not feasible as some of the categories (e.g. compatriots) did not offer a sufficient sample size to allow comparison.

Results

Concreteness and log geographic distance of texts related to death demonstrated a curvilinear relationship, which was well approximated by a cubic polynomial function ($2.96 - 0.05 * x - 0.1 * x^2 + 0.03 * x^3$, $R^2 = 0.18$). Figure F(a) both reports the relationship for the tweets referencing death (solid line) and – for the reference – replicates at a different scale the curve from Figure 1A (dotted line) that summarizes the trend in all tweets about major US cities in Study 1. The thematically constrained subset of tweets showed a familiar pattern even if a slightly flatter one than the overall trend. Tweets about death sent from the city center were maximally concrete, their concreteness dropped dramatically when outside of the city and levelled off at distances above 100 km, with a slight increase in concreteness at very remote distances.

Similarly, concreteness of USENET contexts containing the phrase *years ago* and death-related words was a sigmoid function of log temporal distance, which was well approximated by a cubic polynomial ($2.53 + 0.41 * x - 0.40 * x^2 + 0.08 * x^3$, $R^2 = 0.48$). Figure 1F(b) plots the curve estimated for death-related messages (solid line) and the overall trend for all messages (dotted line), which replicates, with a correction for scale, the curve in Figure 1B. Death-related contexts were generally more concrete than the thematically unconstrained contexts, but much like the overall trend in Figure 1B they showed the maximum of concreteness for deaths that occurred very recently, a drastic decrease in concreteness as the past became less recent and finally a levelled-off pattern after some three decades from the time of writing the message. To sum up, the curvilinear relationship between language concreteness and log (geographic and temporal) distance was confirmed even with a constraint that focused on one class of phenomena, i.e. those related to death. Thus, phenomena that are likely to be part of individual experience, and have a similar probability of occurring in an individual life recently or a long time ago, close or far, are construed with a similar level of detail at different distances as the entirety of phenomena that our method captures in a language corpus.

General Discussion

We present a new method of examining an aspect of embodied cognition (Gallese & Lakoff, 2005; Barsalou, 2008; Fischer & Zwaan, 2008; Meteyard et al., 2012): the interplay between perceived and objective distance, abstraction as a mental faculty and abstractness as a property of language. We identify words or phrases that denote an entity or event for which we have information about distance: this information could be encoded in the phrase (*spouse* vs *coworker* for social distance) or explicitly stated as numerals (*twenty years from now*). We then measure the concreteness of language that co-occurs with that word or phrase in texts, and correlate this concreteness with distance. The utility of our method is demonstrated in studies of three critical dimensions of the CLT: spatial, temporal and social distance. In all four studies, the predictions of the CLT held. Tweets containing the names of

an American city become more abstract as the geographic distance between the person sending the tweet and that city increases. Similarly, verbal contexts of times further into the past or future tend to be more abstract, as do verbal descriptions of more socially distant people.

Our use of multiple linguistic expressions of distance and massive amounts of linguistic productions in corpora allowed us to go beyond validation of prior experimental findings against new empirical materials. One theoretical point raised by Soderberg et al. (2014) was whether distance increased the processing of abstract information, decreased the processing of concrete information, or both. We note that in all our studies, greater distance led to more overall abstractness in language, but at every distance the linguistic productions showed gradience in how abstract versus concrete they are. Specifically, our regression analyses of two continuous metrics of spatial (kilometers from the city) and temporal (years before writing) distance revealed that the relationship between log distance and abstractness of language is curvilinear, and is well approximated by a cubic polynomial curve. Language used in relation to cities and events is at its most concrete (construal is at its lowest) when the experience of that city or that event is most immediate, e.g. being in the city center, or having an event in the very recent past (cf. Hirst et al., in press). Tweets become abstract more rapidly between a city center and its suburbs than between the city boundary and any other location in the country. Time references become abstract more rapidly between the present and the timepoint in the past that indicates a typical lifespan than between events in the distant and very distant past. This is also true when texts were thematically constrained to refer to death-related phenomena. Thus, we both confirm and specify the curvilinear relationship between distance and abstraction predicted by Soderberg et al. (2014). Moreover, the similarity of effects that physical and temporal distance have on linguistic concreteness – displayed over all relevant contexts or only a thematically constrained set – corroborates the long-standing observation that language often expresses temporal relations via metaphors of space (Boroditsky & Ramscar, 2002; Boroditsky, 2000, 2001). Finally, symmetrical effects of past and future temporal distance on concreteness suggests analogous cognitive processes involved in remembering past events and imagining future ones (e.g., Schacter, Addis, & Buckner, 2007).

As with any method, a corpus-based approach has limitations. We are unable to explore the construal of entities that do not correspond to a word or phrase, or do not occur in corpora with sufficient frequency. There is little doubt that noise is introduced into the data from homography and polysemy: *bank* as a financial institution and *bank* of a river; *work* as a noun and a verb; *Chicago* as a city and a musical. Also, we used concreteness ratings for words presented out of context to calculate the average concreteness of sequences of words that occur in context, missing out on metaphoric word use and other context-driven changes in word meaning. It is improbable, however, that our patterns arise due to a systematic bias in our operationalization of context concreteness, as this would require contexts not only to be consistent in how they change word concreteness but also modulate this amount and direction of change as a function of distance. Many of these limitations are addressable: by carefully restricting the searched linguistic materials and their contexts (exemplified partly

in Study 4), restricting the age, gender or place of residence of contributors (as self-reported in several social media sources), or taking temporal cross-sectional snapshots of the data.

The utility of an observational approach based on corpora as a complement to experimental studies outweighs its limitations. It has the advantages of (a) ecological validity through observation of psychological distance in texts produced in natural communicative settings, (b) automatized ability to track psychological distance in vast spans of language created by heterogeneous, large populations, and (c) ability to investigate a very broad, or a very focused, range of entities or events. For instance, we chose American cities as our geographical objects of interest. Any object for which we have a name, and latitude and longitude coordinates can be explored for the effect of geographic distance on construal with the method presented in this paper. Equally, choosing one theme or a specific time slice in a corpus enables one to break down the aggregate trends demonstrated here into any level of granularity. Importantly, corpora and social media free researchers to study psychological distance outside of the laboratory.

Acknowledgements

We thank the Sherman Centre for Digital Scholarship and the Research High-Performance Computing Support group at McMaster University for technical support. This research was supported by the SSHRC Insight Development grant 430-2012-0488, the NSERC Discovery grant 402395-2012, the NIH R01 HD 073288 (PI Julie A. Van Dyke), and the Early Researcher Award from the Ontario Research Fund to the second author. Thanks are due to Emmanuel Keuleers and two anonymous reviewers, as well as the audience of the Psychonomic Society's 55th Annual Meeting and the 2015 Annual Meeting of AAAS for providing valuable feedback.

References

- Alter AL, Oppenheimer DM. Effects of fluency on psychological distance and mental construal (or why New York is a large city, but *New York* is a civilized jungle). *Psychological Science*. 2008; 19(2):161–167. [PubMed: 18271864]
- Amit E, Greene JD. You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological science*. 2012; 23(8):861–868. [PubMed: 22745347]
- Bar-Anan Y, Liberman N, Trope Y, Algom D. Automatic processing of psychological distance: Evidence from a Stroop task. *Journal of Experimental Psychology: General*. 2007; 136(4):610–622. [PubMed: 17999574]
- Barsalou LW. Grounded cognition. *Annual Review Psychology*. 2008; 59:617–645.
- Bogardus ES. A social distance scale. *Sociology & Social Research*. 1933; 17:265–271.
- Boroditsky L. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*. 2000; 75(1):1–28. [PubMed: 10815775]
- Boroditsky L. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*. 2001; 43(1):1–22. [PubMed: 11487292]
- Boroditsky L, Ramscar M. The roles of body and mind in abstract thought. *Psychological Science*. 2002; 13(2):185–189. [PubMed: 11934006]
- Brysbaert M, Warriner AB, Kuperman V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*. 2014; 46(3):904–911. [PubMed: 24142837]
- Burgoon EM, Henderson MD, Markman AB. There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science*. 2013; 8(5):501–520. [PubMed: 26173209]
- Central Intelligence Agency. The World Factbook: North America::United States. 2014. Retrieved from <https://www.cia.gov/library/publications/the-world-factbook/geos/us.html>

- DArgembeau A, Van der Linden M. Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition*. 2004; 13(4):844–858. [PubMed: 15522635]
- Eisenstein, J. Written dialect variation in online social media. In: Boberg, C.; Nerbonne, J.; Watt, D., editors. *Handbook of Dialectology*. Wiley; 2015.
- Eyal T, Liberman N, Trope Y, Walther E. The pros and cons of temporally near and distant action. *Journal of Personality and Social Psychology*. 2004; 86(6):781–795. [PubMed: 15149255]
- Fiedler K. Construal level theory as an integrative framework for behavioral decision making research and consumer psychology. *Journal of Consumer Psychology*. 2007; 17(2):101–106.
- Fischer MH, Zwaan RA. Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*. 2008; 61(6):825–850. [PubMed: 18470815]
- Fujita K, Henderson MD, Eng J, Trope Y, Liberman N. Spatial distance and mental construal of social events. *Psychological Science*. 2006; 17(4):278–282. [PubMed: 16623682]
- Gallese V, Lakoff G. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*. 2005; 22(3–4):455–479. [PubMed: 21038261]
- Gong H, Medin DL. Construal levels and moral judgment: Some complications. *Judgment and Decision Making*. 2012; 7(5):628–638.
- Hansen J, Wanke M. Truth from language and truth from fit: The impact of linguistic concreteness and level of construal on subjective truth. *Personality and Social Psychology Bulletin*. 2010; 36(11):1576–1588. [PubMed: 20947772]
- Hirst W, Phelps EA, Meksin R, Vaidya CJ, Johnson MK, Mitchell KJ, Olsson A. A Ten-Year Follow-Up of a Study of Memory for the Attack of September 11, 2001: Flashbulb Memories and Memories for Flashbulb Events. *Journal of Experimental Psychology: General*. in press
- Johnson MK, Foley MA, Suengas AG, Raye CL. Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*. 1988; 117(4):371–376. [PubMed: 2974863]
- Kahle D, Wickham H. ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. [Computer software] R package version 2.3. 2013
- Ledgerwood A, Trope Y, Chaiken S. Flexibility now, consistency later: Psychological distance and construal shape evaluative responding. *Journal of Personality and Social Psychology*. 2010; 99(1):32–51. [PubMed: 20565184]
- Liviatan I, Trope Y, Liberman N. Interpersonal similarity as a social distance dimension: Implications for perception of others actions. *Journal of Experimental Social Psychology*. 2008; 44(5):1256–1269. [PubMed: 19352440]
- Meteyard L, Cuadrado SR, Bahrami B, Vigliocco G. Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*. 2012; 48(7):788–804. [PubMed: 21163473]
- Nisbett RE, Caputo C, Legant P, Marecek J. Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology*. 1973; 27(2):154–164.
- Paivio, A. *Mental representations: A dual coding approach*. New York: Oxford University Press; 1990.
- Parrillo VN, Donoghue C. Updating the Bogardus social distance studies: A new national survey. *The Social Science Journal*. 2005; 42(2):257–271.
- Pennebaker, JW.; Paez, D.; Rim, B. *Collective memory of political events: Social psychological perspectives*. New York: Psychology Press; 2013.
- Pingdom. Report: Social network demographics in 2012. 2014. Retrieved from <http://royal.pingdom.com/2012/08/21/report-social-network-demographics-in-2012/>
- R Core Team. R Foundation for Statistical Computing. Vienna; Austria: 2013. R: A language and environment for statistical computing. Available from <http://www.R-project.org/>
- Robinson WS. Ecological correlations and the behavior of individuals. *American Sociological Review*. 1950; 15:351–357.
- Schacter DL, Addis DR, Buckner RL. Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*. 2007; 8(9):657–661. [PubMed: 17700624]

- Schwanenflugel PJ, Harnishfeger KK, Stowe RW. Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*. 1988; 27(5):499–520.
- Semin GR, Fiedler K. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*. 1988; 54(4):558–568.
- Shaoul, C.; Westbury, C. 2013. A reduced redundancy USENET corpus (2005–2011) Edmonton, AB: University of Alberta (downloaded from <http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html>)
- Soderberg CK, Callahan SP, Kochersberger AO, Amit E, Ledgerwood A. The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin*. in press
- Stephan E, Liberman N, Trope Y. Politeness and psychological distance: a construal level perspective. *Journal of Personality and Social Psychology*. 2010; 98(2):268–280. [PubMed: 20085400]
- Trope Y, Liberman N. Temporal construal and time-dependent changes in preference. *Journal of Personality and Social Psychology*. 2000; 79(6):876–889. [PubMed: 11138758]
- Trope Y, Liberman N. Temporal construal. *Psychological review*. 2003; 110(3):403–421. [PubMed: 12885109]
- Trope Y, Liberman N. Construal-level theory of psychological distance. *Psychological Review*. 2010; 117(2):440–463. [PubMed: 20438233]
- Wakslak CJ, Trope Y, Liberman N, Alony R. Seeing the forest when entry is unlikely: Probability and the mental representation of events. *Journal of Experimental Psychology: General*. 2006; 135(4): 641–653. [PubMed: 17087578]

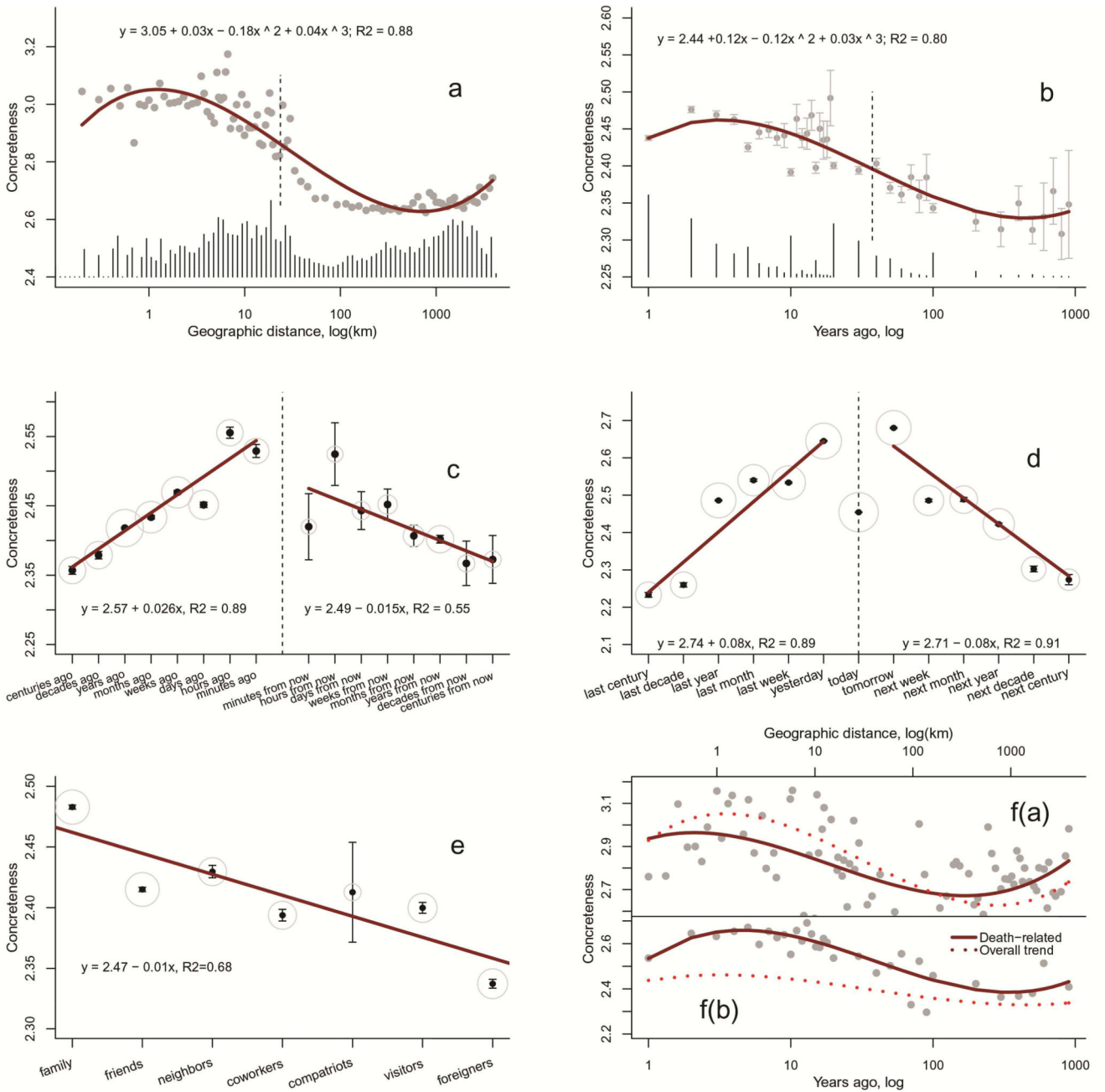


Figure 1. Panel A: Concreteness of Twitter messages regarding US cities as a function of log geographic distance (km). Panels B-E: Concreteness of USENET postings as (Panel B) a function of log temporal distance from 1 to 999 years; (Panel C) a function of time units in the past (units *ago*) and the future (units *from now*); (Panel D) a function of ordered time units in the past (*last* unit) and the future (*next* unit); and (Panel E) a function of social distance categories. In A and B, the dotted line represents the inflection point and the histogram of distances is presented in the bottom. In C-E, gray circle size is proportional to log 10 frequency of the term. In all panels, the best-fit regression line and equation are provided, and error bars (where visible) reflect the 95% confidence intervals. Panel F plots

(a) concreteness of death-related Twitter messages regarding US cities (solid line) and concreteness of all Twitter messages replicated from Panel A (dotted line); (b) concreteness of death-related USENET postings with *years ago* as a target phrase (solid line) and concreteness of all such USENET postings replicated from Panel B (dotted line).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Hierarchical regressions comparing models with polynomials of the 1st to 4th degree fitted to context concreteness as a function of log geographic distance from the city (Study 1) and log temporal distance from the event in the past (Study 2.1). R^2 is the amount of the model's explained variance, R^2 is the difference in explained variance from the model with a lower order polynomial, and p is the p-value of the comparison of successive models. The cubic polynomial provided the best fit.

Polynomial degree	Geographic distance (Study 1)		Temporal distance (Study 2.1)	
	R^2	p	R^2	p
<i>Linear</i>	0.775	–	0.745	–
<i>Quadratic</i>	0.776	> 0.5	0.749	0.004
<i>Cubic</i>	0.883	< 0.001	0.814	0.065
<i>Quartic</i>	0.885	> 0.5	0.819	0.005

Table 2

Terms used for the social distance groups as defined by Bogardus (1922), from the most proximal to the most distal.

Family	Friends	Neighbours	Coworkers	Compatriots	Visitors	Foreigners
husband	friend	neighbor	coworker	compatriot	visitor	immigrant
wife	ally	neighbours	co-worker	countryman	tourist	foreigner
spouse	confidant	peer	colleague	countrywoman	traveller	outsider
	confidante	homie	collaborator			stranger
	alter ego	homeboy	workmate			emigrant
	second self	homegirl				nonmember
	pal					noncitizen
	chum					newcomer
	buddy					alien