COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13372

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13399

Predicting Hand Movements With Distributional Semantics: Evidence From Mouse-Tracking

Daniele Gatti,[a] Marco Marelli,[b,c] Luca Rinaldi[a,d]

[a]*Department of Brain and Behavioral Sciences, University of Pavia*
[b]*Department of Psychology, University of Milano-Bicocca*
[c]*NeuroMI, Milan Center for Neuroscience*
[d]*Cognitive Psychology Unit, IRCCS Mondino Foundation*

Received 13 January 2022; received in revised form 12 December 2023; accepted 19 December 2023

Abstract

Although mouse-tracking has been taken as a real-time window on different aspects of human decision-making processes, whether purely semantic information affects response conflict at the level of motor output as measured through mouse movements is still unknown. Here, across two experiments, we investigated the effects of semantic knowledge by predicting participants' performance in a standard keyboard task and in a mouse-tracking task through distributional semantics, a usage-based modeling approach to meaning. In Experiment 1, participants were shown word pairs and were required to perform a two-alternative forced choice task selecting either the more abstract or the more concrete word, using standard keyboard presses. In Experiment 2, participants performed the same task, yet this time response selection was achieved by moving the computer mouse. Results showed that the involvement of semantic components in the task at hand is observable using both standard reaction times (Experiment 1) as well as using indexes extracted from mouse trajectories (Experiment 2). In particular, mouse trajectories reflected the response conflict and its temporal evolution, with a larger deviation for increasing word semantic relatedness. These findings support the validity of mouse-tracking as a method to detect deep and implicit decision-making features. Additionally, by demonstrating that a usage-based model of meaning can account for the different degrees

Correspondence should be sent to Daniele Gatti, Department of Brain and Behavioral Sciences, University of Pavia, Piazza Botta 6, 27100 Pavia, Italy. E-mail: daniele.gatti@unipv.it

of cognitive conflict associated with task achievement, these findings testify the impact of the human semantic memory on decision-making processes.

Keywords: Semantic memory; Distributional semantic models; Mouse-tracking; Semantic relatedness; Hand movements

1. Introduction

In the last decade, mouse-tracking has seen enduring and rising popularity across several branches of cognitive science, mainly because of the methodological and theoretical advantages this approach carries (Stillman, Shen, & Ferguson, 2018). As compared to reaction times (RTs), indeed, mouse-tracking has been revealed to be particularly reliable in isolating the dynamics of response conflict, by quantifying the relative directness with which participants make their decision (Freeman, 2018). That is, using mouse-tracking, it is possible to quantify the stages by which the decision process unfolds (e.g., Kieslich, Henninger, Wulff, Haslbeck, & Schulte-Mecklenbeck, 2019). Accordingly, mouse-tracking has been successfully used across a large number of domains, such as language (Lins, & Schöner, 2019; Spivey, Grosjean, & Knoblich, 2005), social cognition (Freeman, Pauker, & Sanchez, 2016), and memory (Gatti, Rinaldi, Marelli, Mazzoni, & Vecchi, 2022a; Papesh, & Goldinger, 2012; Papesh, Hicks, & Guevara Pinto, 2019).

Briefly, in mouse-tracking tasks, participants are required to move from a starting point position (i.e., typically placed in the middle-bottom part of the screen) to select either of the alternatives presented in the two upper corners of the screen. On the assumption that mouse movements (i.e., hand movements) are executed in parallel with the decision that participants are required to make (e.g., Freeman, & Ambady, 2010), mouse-tracking packages allow to quantify the conflict of the choice and its evolution, which cannot be directly assessed using only RTs (Stillman et al., 2018). The magnitude of the decision conflict is generally quantified by computing the maximum deviation from the direct path (i.e., the furthest point on the actual trajectory from the idealized straight trajectory between the starting point and the selected stimulus), while the decision evolution is quantified by means of sample entropy, which measures the degree of movement irregularity and unpredictability (for a complete discussion on other possible indexes see, Freeman & Ambady, 2010; Stillman et al., 2018). More specifically, the greater the deviation from a straight path or the trajectory irregularity, the greater the decision conflict. In addition to this, mouse-tracking provides a precious real-time window into the temporal sequence of the processes subserving the conflict, being informative about the time at which the decision takes place.

Although mouse-tracking has been employed to investigate different cognitive functions, including the language domain, the decision conflict subserving word meaning processes are still largely unknown. Insofar, evidence for an effect of word meaning on response selection has been mainly observed in priming tasks, hence focusing mostly on RTs. For instance, in lexical priming tasks, the response to a target word (e.g., *dog*) has been systematically shown to vary as a function of the preceding linguistic context (i.e., the prime word; e.g., Meyer & Schvaneveldt, 1971). Typically, the target word elicits a faster response after a meaning-related prime word (e.g., *cat*) as compared to when it is presented after an unrelated word (e.g., *pen*). Similar evidence has been gathered from a previous mouse-tracking study, in which participants were asked to indicate which target word, among two possible alternatives, was more associated with a preceding cue word (Hindy, Hamilton, Houghtling, Coslett, & Thompson-Schill, 2009; but for other studies on lexical decision, see also: Barca & Pezzulo, 2012; 2015; Dale, Hindy, & Spivey, 2006, 2007).

Results showed that the response conflict, as indexed by the maximum deviation, was smaller for more semantically associated correct cue-target pairs as compared to less semantically associated pairs (Hindy et al., 2009). This is compatible with the view of an automatic contextual binding of actions and verbal information (for a review: García & Ibáñez, 2016a), according to which various details of an action's unfolding are revealed by kinematic variables that cannot be thus simply accounted by classical reaction-time tasks (García & Ibáñez, 2016b; but see also: Afonso et al., 2019; García-Marco et al., 2019). Yet, in the study by Hindy et al. (2009), semantic relatedness was operationalized on a dichotomous basis (i.e., by comparing weakly associated with strongly associated words) and participants' performance was directly predicted by this very same dimension (i.e., participants had to select the target word most semantically related to the cue). In contrast, here we aimed at predicting mouse-tracking performance based on a continuous semantic index.

Here, building upon this evidence, we take advantage from distributional models (DSMs) to explore semantic involvement in decision-making using mouse-tracking. DSMs induce word meanings from large databases of natural language data, representing them as high-dimensional numerical vectors. More critically, DSMs are thought to well capture the structure of semantic memory, in which word meanings would be conceived as distributed representations populating a continuous mental space (Günther, Rinaldi, & Marelli, 2019; Jones, Willits, Dennis, & Jones, 2015b). On these grounds, DSMs may be considered the ideal tool to probe whether word meanings affect not only the processes driving RTs, but also the programming and execution of (hand) movements. Accordingly, previous studies have shown that word relatedness measures extracted from DSMs can predict lexical priming effects at the item-level (Günther, Dudschig, & Kaup, 2016; see also: Jones, Kintsch, & Mewhort, 2006; Mandera, Keuleers, & Brysbaert, 2017). In these studies, word meaning was indeed operationalized as a quantification of a word distribution over linguistic contexts from large corpora of natural text.

In the present study, participants were, therefore, shown several word pairs and were required to indicate the more abstract one (or the more concrete one, depending on the experimental condition). In Experiment 1, participants answered using the keyboard (i.e., in a standard keyboard task), while in Experiment 2, a different sample of participants was asked to perform the task by moving the mouse. We then predicted participants' responses in a two-alternative forced choice task using the degree of semantic relatedness between the two words as extracted from DSMs. Hence, although word relatedness was not a primary dimension for performing the task, we nevertheless expected this dimension to influence participants' performance.

Building on a seminal body of work, the present study extends it in four different ways. First, in this study, we predicted participants' behavior in a mouse-tracking paradigm on a continuous scale (i.e., semantic relatedness extracted from DSMs), while previous works only employed categorical predictors (e.g., weakly associated vs. strongly associated words; see Dale et al., 2006, 2007; Hindy et al., 2009), with our approach more reliably reflecting recent models of semantic memory conceiving word meanings as distributed representations populating a continuous mental space (Jones et al., 2015b). Second, and differently from previous studies, the task is not explicitly focusing on semantic similarity, but rather on concreteness judgments, thus investigating the (semantic) processes involved at an implicit level. Third, it should be noted that the semantic relatedness metric employed here is not derived from human-based ratings (which in turn were employed in previous research), but it is an independent-source index extracted from natural language regularities, with the architecture of word-embeddings being consistent with relatively simple, psychologically grounded associative learning mechanisms (Günther et al., 2019; Rinaldi & Marelli, 2020). Supporting this choice, recently it has been indeed argued that, in order to improve our knowledge of the cognitive processes underlying an experimental task (and to avoid explanatory circularity), studies should strive to employ predictors built from sources independent from human-based ratings (Westbury, 2016; and for an example of this effect on semantic fluency see: Jones, Hills, & Todd, 2015a). Consistent with this, our approach allows to predict participants' behavior in the task at hand starting from independent models that replicate the structure of semantic memory by applying a learning mechanism to environmental regularities (i.e., word co-occurrences). Finally, by performing the same task with two different procedures (i.e., keyboard vs. mouse), the present study could provide evidence regarding how different methods allow to distinctively detect deep and implicit features of human behavior. In particular, we predict that the higher the semantic relatedness between the words to be judged, the higher the response conflict. This should be reflected in the specific profiles of hand movement trajectories, with a higher maximum deviation from the direct path as a function of increasing semantic relatedness between the two words presented.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Forty-seven right-handed students (7 males, M age = 24.10 years, SD = 3.58) participated in the study for course credits. All participants were native Italian speakers, had normal or corrected to normal vision, and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of the University of Pavia (Italy) and participants were treated in accordance with the Declaration of Helsinki.

2.1.2. Distributional semantic model

Vector representations for the words used in this study were extracted from a DSM obtained via the application of neural networks, and in particular the Continuous Bag of Words (CBOW) method, an approach originally proposed by Mikolov, Chen, Corrado, and Dean

(2013): when using CBOW, the obtained vector dimensions capture the extent to which a target word is reliably predicted by the contexts in which it appears. The model, released by Marelli (2017), was trained on itWaC, a free Italian text corpus based on web-collected data and consisting of about 1.9 billion tokens (semantic relatedness values are available online via the SNAUT database: http://meshugga.ugent.be/snaut-italian-2/). The model is set on the following parameters: *5-word co-occurrence window*, *400-dimension vectors*, negative sampling with $k = 10$, subsampling with $t = 1e^{-5}$. This set of parameters defines the learning procedure used to induce word vectors (Mikolov et al., 2013). Co-occurrence window size indicates how large the considered lexical contexts are; in our case, *a 5-word window* indicates that we estimated predictions concerning two words on the left and two words on the right of the target word. The obtained vectors were used to estimate the degree of semantic relatedness (SRel) between word pairs by simply considering the cosine of the angle formed by them: the higher the SRel value (i.e., the closer the vectors in the multidimensional space), the more semantically related the words are expected to be.

2.1.3. Stimuli

The stimuli used in the present study were built starting from the words included in the Italian ANEW database (Montefinese, Ambrosini, Fairfield, & Mammarella, 2014). In this study, a large number of participants had been asked to rate Italian words on different scales, including concreteness (ranging from 1, abstract, to 9, concrete).

First, we retrieved from the ANEW database all the words included and generated all possible pairs, obtaining for each pair the difference in concreteness (i.e., by subtracting the word with the lower value of concreteness from the word with the higher value). Then, in order to allow participants in our study to easily discriminate the more concrete/abstract of the two words, we removed pairs with small difference in concreteness (i.e., concreteness difference < 2.5). For each of the remaining word pairs, we extracted an SRel index based on the cosine of the angle formed by their corresponding vectors (see the previous section for a description of the DSM used). Finally, we selected 120 word pairs guaranteeing almost uniformly distributed SRel values (see Fig. 1A for SRel distribution and Fig. 1B for the distribution of concreteness difference for the word pairs selected).

In particular, in the final set of 120 word pairs, the concreteness difference ranged from 2.6 for the pair "poesia–penna" ("poetry–pen"), respectively, with concreteness $= 5.85$ and $= 8.45$; to 5.4 for the pair "spirito–specchio" ("spirit–mirror"), respectively, with concreteness $= 3.15$ and $= 8.55$. SRel varied from very low (SRel $= .007$, e.g., "idrante–intelletto"; "hydrant–intellect") to relatively high values (SRel $= .39$, e.g., "musica–violino"; "music–violin"). To have an idea of how SRel increases, while keeping constant and high the concreteness difference (to allow participants to detect the more concrete or abstract item easily), in Table 1, we report several examples of word pairs (translated in English) included in each .1 step of SRel. SRel and concreteness difference showed a small correlation, $r = -.17$. In Table 2, we report the psycholinguistic characteristics of the word pairs included.

2.1.4. Procedure

Participants were tested online using Psychopy (Peirce, 2007, 2009; Peirce & MacAskill, 2018; Peirce et al., 2019) through the online platform Pavlovia (https://pavlovia.org/).

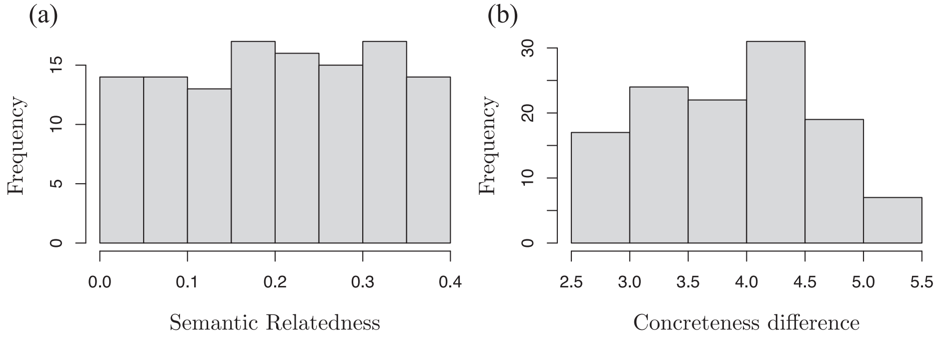

Fig. 1. Frequency distributions of the Semantic Relatedness values (A) and of concreteness difference (B) of the word pairs included in this study. Note that for concreteness difference, the distribution starts from a high value, since in order to include only word pairs in which the correct answer was easily discriminable, we did not include word pairs with concreteness difference < 2.5.

Table 1
Examples of word pairs included in the present study (translated in English)

SRel step	Word pair
.007−.1	intellect−hydrant; health−garden; justice−glass
.1−.2	brutal−honey; fast−wasp; smell−milk
.2−.3	grateful−hug; passion−bike; tired−couch
.3−.39	poetry−pen; cold−coat; christmas−tree

Note. The more abstract word on the pair is presented in this table always as first, the more concrete always as second.

Participants were shown two words (half of the abstract words composing the word pair were located on the left side and the other half on the right side) and were instructed to indicate the more concrete/abstract one (counterbalanced across participants). Participants were instructed to respond as fast and accurately as possible by pressing the left/right key (A and L) using either hand, in order to indicate whether the word placed on the left side or the one on the right side. Trials were shown in random order.

Each trial started with a central fixation cross (presented for 500 ms) followed by a word pair (with the two words being completely in the two different halves of the screen), after participants' response, the trial moved to a black screen (presented for 1000 ms) which ended the trial.

Table 2
Psycholinguistics characteristics of the word pairs used in the present study

	SRel	Concreteness difference	Log(frequency) difference	Length difference
Mean	.20	3.93	.77	1.93
SD	.11	.70	.58	1.35
Min – Max	.007−.39	2.60−5.40	.01−3.29	0−7

2.2. *Data analysis and results*

Participants' log-transformed RTs in accurate trials was the dependent variable of interest. The statistical analyses were performed using *R-Studio* (RStudio Team, 2015). We estimated a linear mixed model (LMM) having RTs as dependent variable, with the *lme4 R* package (Bates, Maechler, Bolker, & Walker, 2015). *Pseudo-*R^2 were computed using the *summ* function from the *jtools R* package (Long, 2020). SRel, concreteness difference (i.e., the difference between the concreteness of the two words presented, as extracted from the Italian ANEW database; Montefinese et al., 2014), frequency difference (i.e., the absolute difference between the log-transformed frequencies of the two words presented, as extracted from the Italian SUBTLEX; http://crr.ugent.be/subtlex-it/), and length difference (i.e., the absolute difference of the number of letters comprising each word) were included as continuous predictors, while subjects and items were included as random intercepts in all the models estimated. In particular, in *lme4* syntax, the model tested was:

$$RTs \sim SRel + delta\_concreteness + delta\_(frequency) + delta\_lenght$$
$$+\ (1|Participant) + (1|Item)$$

Nonaccurate trials (302 trials on a total of 5640 trials) and trials in which overall RTs were faster than 300 ms or slower than 3000 ms (185 additional trials on a total of 5640 trials) were excluded from the analysis. Overall participants' accuracy was very high ($M = .94$, $SD = .03$).

The model had *Pseudo-*R^2 (total) $= .47$, *Pseudo-*R^2 (fixed effects) $= .004$. Results showed that the effect of SRel was significant, $t(113) = 2.35$, $p = .02$, $b = .19$, while the effects of concreteness, $t(113) = -.78$, $p = .44$, $b = -.01$, frequency, $t(114) = -.85$, $p = .40$, $b = -.01$, and length, $t(114) = -.44$, $p = .66$, $b < .01$, were not significant.

3. Experiment 2

3.1. *Methods*

3.1.1. *Participants*

Fifty-five right-handed students participated in the study for course credits (none of the participants had participated to Experiment 1). Fifteen participants were a-priori excluded as they used the trackpad during the experiment (see below the Section *Procedure*). The final sample was composed of 40 participants (8 males, M age $= 22.15$ years, $SD = 2.90$). All participants were native Italian speakers, had normal or corrected to normal vision, and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of the University of Pavia (Italy) and participants were treated in accordance with the Declaration of Helsinki.

3.1.2. *Distributional semantic model*

The distributional semantic model used in Experiment 2 was identical to Experiment 1.

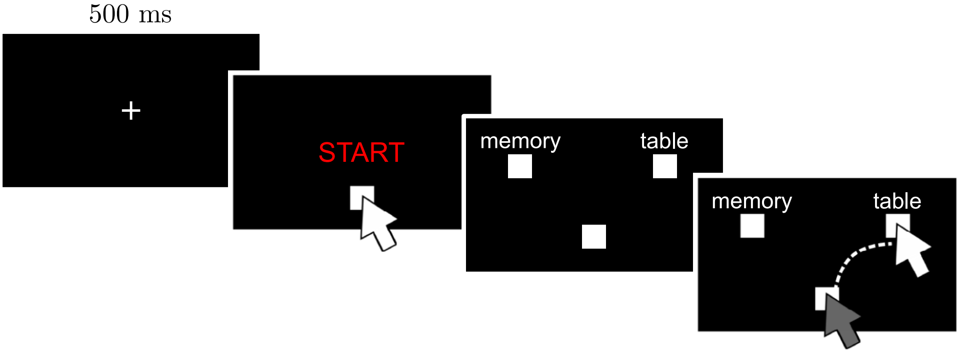

Fig. 2. The timeline of the task. Participants were asked to indicate the more concrete/abstract of the two words shown using their mouse. In particular, after clicking on the square presented below the "START" signal, they were shown two words and were required to make their judgment by moving their mouse and clicking on the square below the selected word.

3.1.3. Stimuli

The stimuli used in Experiment 2 were identical to Experiment 1.

3.1.4. Procedure

Participants were tested online using Psychopy (Peirce, 2007, 2009; Peirce & MacAskill, 2018; Peirce et al., 2019) through the online platform Pavlovia (https://pavlovia.org/).

At the beginning of the experiment, participants were asked to indicate the kind of tool with which they were answering (i.e., external mouse or trackpad). Participants were also informed that the use of the trackpad would not have affected the earning of the course credits. This allowed us to identify those participants who did not use the external mouse, who were subsequently excluded from the analyses since the reliability of mouse-tracking data collected with the trackpad is not yet established in the scientific literature.[1]

Then, participants were instructed on how they were supposed to use their mouse during the task, and, after five practice trials, (without receiving feedback on correct/wrong responses), they were asked to perform the experimental task. Participants were shown two words (half of the abstract words composing the word pair were located on the left side and the other half on the right side) and were instructed to indicate the more concrete/abstract one (half of the participants were asked to indicate the more concrete word, the other half the more abstract one) using their mouse: participants had to first click on a square presented below the "START" label by pressing the mouse left button, and next to move to the selected option, again pressing the left button to make their decision. Each trial began with a fixation cross presented at the center of the screen (500 ms), then a red "START" button appeared at the bottom-center of the screen (Arial font; screen coordinates: $x = 0$, $y = -.35$); after the participants clicked on it, the two words were shown at the upper corners of the screen (Arial font; screen coordinates: $x = \pm .50$, $y = .25$), along with two squares that participants used as buttons (Fig. 2). Participants answered to each trial by moving the mouse from the "START" button to the square associated with the chosen word and by clicking on this latter.

3.1.5. Data analysis

Participants' performance was explored through five dependent variables. Two dependent variables were related to spatial measures, namely, the maximum deviation from direct path (MD) and sample entropy; while three dependent variables were related to temporal measures, namely, the maximum deviation from direct path time (MDtime) and RTs computed at different stages: initiation RTs and overall RTs.

MD is defined as the furthest point on the actual trajectory from the idealized straight trajectory and is thought to quantify the conflict in the choice. Sample entropy is defined as the degree of irregularity and unpredictability in movement across the *x*-axis and is thought to measure the evolution of the choice (i.e., with higher values indicating higher irregularity) and is computed by comparing windows of a fixed size across all recorded positions. Sample entropy is thus obtained computing the negative natural logarithm of the conditional probability that this window remains similar across the trial (Hehman, Stolier, & Freeman, 2015). Initiation RTs is computed as the time elapsed between the click on the START button and the first hand-movement, while MDtime indexes the time at which the trajectory reaches the MD. These two dependent variables are considered to measure two different time windows of decision-making, with the former (initiation RTs) indicating the time at which the decisional process starts and the latter (MDtime) indicating the time at which a decision is finally achieved (see for a complete discussion regarding mouse-tracking variables: Stillman et al., 2018). Finally, overall RT is computed as the time elapsed between the click on the START button and the final click on the selected stimulus. This variable is, therefore, necessarily influenced by MDtime, in those trials in which the decision occurs later (i.e., higher MDtime) will also result in higher overall RTs.

All the dependent variables were computed only on the trials in which participants answered correctly, using the *mousetrap* package (Kieslich et al., 2019). All trajectories were normalized into 101 time steps and remapped symmetrically in order to allow for direct comparison of trajectories which differed in duration and number of data points. Initiation RTs, MDtime, and overall RTs were all log-transformed.

The statistical analyses were performed using *R-Studio* (RStudio Team, 2015). We estimated five separated LMMs, one for each dependent variable, with the *lme4 R* package (Bates et al., 2015). *Pseudo-*R^2 were computed using the *summ* function from the *jtools R* package (Long, 2020).

As for Experiment 1, for each LMM, we included SRel, concreteness difference, frequency difference, and length difference as continuous predictors, while subjects and items were included as random intercepts in all the models estimated. To account for multiple comparisons, since we estimated five models with as dependent variable indexes emerging from the same set of behaviors, we set $\alpha = .01$ (i.e., $\alpha = .05$/number of indexes tested).

3.2. Results

Nonaccurate trials (151 trials on a total of 4800 trials) and trials in which overall RTs were faster than 300 ms or slower than 3000 ms (306 additional trials on a total of 4800 trials) were excluded from the analysis. Aberrant movements (i.e., trials in which the *mousetrap R*

Table 3
Descriptive statistics for RTs analyzed in Experiment 1 and for the five dependent variables analyzed in Experiment 2 (aggregated across participants)

	Experiment 1	Experiment 2				
	Overall RTs	MD	Sample entropy	Initiation RTs	MDtime	Overall RTs
Mean	1273	.32	.12	320	870	1660
SD	293	.17	.04	220	220	260
Min – Max	738−1950	−.02 to .70	.05−.20	80−1180	520−1550	1150−2170

Note. MD values are reported as a function of normalized windows and range from −1 to +1 in both *x* and *y* axes (with 0,0 corresponding to the center of the screen). The descriptive statistics of the three variables expressing temporal processes (initation RTs, MDtime, overall RTs) are reported in milliseconds.

Table 4
Correlation matrix between the five dependent variables included in the current study

	MD	Sample entropy	Initiation RTs	MDtime	Overall RTs
MD	1				
Sample entropy	.61	1			
Initiation RTs	−.31	−.29	1		
MDtime	−.04	.05	.46	1	
Overall RTs	.27	.34	.20	.67	1

Note. The correlation values ranged from very low to high (4105 degrees of freedom, all *p*s < .01).

package was unable to compute the trajectories or in which MD was ± 3SD from the mean of the participants) were detected in 236 of the 4343 remaining trials and were discarded. Overall participants' accuracy was very high ($M = .97$, $SD = .03$). Descriptive statistics for the five variables are reported in Table 3.

The correlation matrix between the five dependent variables is reported in Table 4. Overall, the correlations ranged from very low, as the one between MD and MDtime ($r = -.04$), to high, as the one between overall RTs and MDtime ($r = .67$). As discussed previously, this latter correlation corroborates the fact that overall RTs is necessarily influenced by MDtime, since trials in which the decision occurs later (i.e., higher MDtime) will also result in higher overall RTs. Note that the correlations were computed on raw data, hence on a total of 4107 trials.

The model on MD had *Pseudo-*R^2 (total) = .26, *Pseudo-*R^2 (fixed effects) = .005. Results showed that the effect of SRel was significant, $t(112) = 2.56$, $p = .01$, $b = .17$, while the effects of concreteness, $t(112) = -1.06$, $p = .28$, $b = -.01$, frequency, $t(112) = -.39$, $p = .69$, $b < -.01$, and length, $t(111) = 1.45$, $p = .14$, $b < .01$, were not significant. See Fig. 3 for a graphical representation of the actual trajectories collapsed as a function of three different levels of SRel.

The model on sample entropy had *Pseudo-*R^2 (total) = .23, *Pseudo-*R^2 (fixed effects) = .003. Results showed that the effect of SRel was significant, $t(110) = 2.63$, $p = .009$, $b = .03$, while the effects of concreteness, $t(111) = -1.43$, $p = .15$, $b < -.01$, frequency, $t(110) = -1.09$, $p = .27$, $b < -.01$, and length, $t(110) = .76$, $p = .44$, $b < .01$, were not significant.

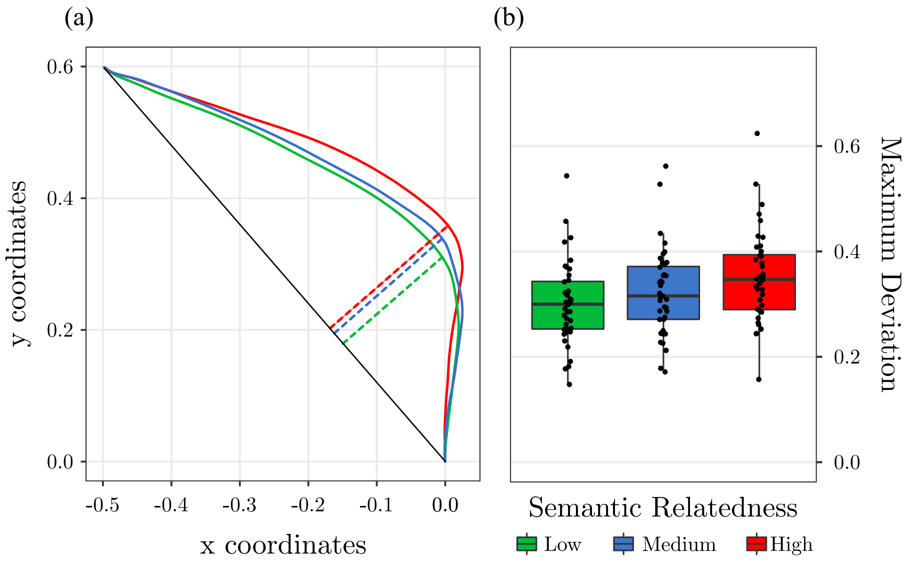

Fig. 3. Schematic representation of how the meaning distance (in dashed lines), computed from the direct path (black straight line) to the maximum deviation registered, varied across three levels of SRel. Note that in order to optimally represent as categorical the continuous predictor, we divided it in three arbitrary categories, each comprising 40 word pairs (A), but data were analyzed using a continuous predictor. A boxplot of the maximum deviation for each word pair across the three hypothetical levels of SRel (B).

The model on initiation RTs was singular, we thus refitted the model with a simplified random structure (i.e., removing the intercept of the item). The second model had *Pseudo-*R^2 (total) = .36, *Pseudo-*R^2 (fixed effects) < .001. Results showed that the effect of SRel, $t(4063) = 1.53$, $p = .12$, $b = .21$, concreteness, $t(4063) = -.37$, $p = .70$, $b = -.01$, frequency, $t(4063) = .95$, $p = .33$, $b = .01$, and length, $t(4063) = .63$, $p = .52$, $b < .01$, were not significant.

The model on MDtime had *Pseudo-*R^2 (total) = .39, *Pseudo-*R^2 (fixed effects) = .003. Results showed that the effect of SRel was significant, $t(111) = 2.65$, $p = .008$, $b = .17$, while the effects of concreteness, $t(112) = -.62$, $p = .53$, $b < -.01$, frequency, $t(111) = -1.03$, $p = .30$, $b < -.01$, and length, $t(111) = 1.22$, $p = .22$, $b < .01$, were not significant.

Finally, the model on overall RTs had *Pseudo-*R^2 (total) = .42, *Pseudo-*R^2 (fixed effects) = .008. Results showed that the effect of SRel was significant, $t(111) = 2.79$, $p = .006$, $b = .17$, while the effects of concreteness, $t(111) = -1.28$, $p = .20$, $b = -.01$, frequency, $t(111) = -.94$, $p = .34$, $b < -.01$, and length, $t(111) = 1.32$, $p = .18$, $b < .01$, were not significant.

4. Discussion

In the present study, we investigated the impact of word meaning on decision-making processes during a two-alternative forced choice concrete-abstract judgment task with. In Experiment 1, participants responded using standard keyboard presses, while in

Experiment 2, we adopted a mouse-tracking approach. In particular, we combined data from distributional semantic models and from a mouse-tracking task to shed light on the response conflict driving word meaning processing. In both experiments, participants were shown several word pairs and were required to indicate the more abstract word (or more concrete one) using the keyboard or the computer mouse. The semantic relatedness between the two words was manipulated in order to ideally test its possible influence on decision-making processes.

Across both experiments, our findings indicate that participants' judgments were affected by the semantic relatedness of the two words, although this was not a dimension directly involved in the task at hand. That is, in Experiment 1, the effect of semantic relatedness was significant, with slower response latencies for word pairs semantically more related. This was replicated and extended in Experiment 2 analyzing data extracted from mouse trajectories across several different dependent variables. These results indicate that semantic relatedness can affect human behavior well beyond processing speed, testifying its influence also on online response conflict at the level of motor control.

The present study extends previous findings showing that the association strength between a cue and a target word can affect participants' performance in a mouse-tracking task (Hindy et al., 2009). Yet, in this previous study, the semantic predictor was operationalized on a dichotomous basis (i.e., strongly associated vs. weakly associated pairs) and only limited to the relationship between the cue and the (correct) target word. In striking contrast, here the semantic relatedness between the two words was operationalized as a continuous variable. This represents a clear advantage, as such an operationalization would more reliably reflect the complex structure of the mental lexicon, especially considering that recent models of semantic memory conceive word meanings as distributed representations populating a continuous mental space (Jones et al., 2015b).

While the impact of semantic relatedness as extracted from DSMs on RTs has been already established (e.g., Gatti, Marelli, & Rinaldi, 2022b; Günther et al., 2016; see also: Jones et al., 2006; Mandera et al., 2017), our study is the first to report its effect also on mouse-tracking variables. Indeed, previous studies have shown that semantic relatedness facilitates participants' RTs, with faster responses occurring for more semantically related word pairs (Günther et al., 2016). Here, we found a similar pattern for the overall RTs, an index that takes into account the overall time window (i.e., from the movement initiation to the selection of the target word), hence also being possibly affected by the conflict arising during the evaluation process. Indeed, the other variables considered (namely, MD and MDtime) were particularly informative of decision-making processes. That is, the results of the maximum deviation index suggest that the conflict increased with increasing semantic relatedness, as the furthest point on the actual trajectory was farther from the idealized straight trajectory for larger semantic relatedness values (Freeman, 2018). Overall, these findings can be interpreted by assuming that participants' choice is affected by the overlap between the positions of the two to-be-compared word-vectors (which account for the distributional pattern of these words in linguistic contexts and, thus, their meaning). Accordingly, participants would rely on their semantic memory structure to solve the task at hand, and the more the vectors representing the two words would overlap (i.e., the more they would be semantically similar), the larger the cognitive conflict (i.e., the more the decision to be reached would be difficult).

Notably, regarding the time course of participants' decisions, our findings also show that the effects of semantic relatedness were relevant only in the online component of the choice (i.e., on MDtime) and in the overall RTs. Indeed, MDtime results suggest that the time at which a decision occurred was delayed for more semantically similar words. Such an effect may have likely determined also the difference observed in terms of overall RTs, a variable that is necessarily influenced by MDtime (i.e., trials in which the decision occurs later will also result in higher overall RTs), as also observed in the moderate correlation between them. On the contrary, no effect was detectable on initiation RTs: this indicates that, at this initial stage, the participants did not yet elaborate the two stimuli in terms of semantic relatedness. A possible explanation could be that the conflict observed between the meanings of the two words in the task at hand emerges at later stages of the decision process, thus allowing to fully observe it with higher reliability when analyzing data originating from mouse trajectories.

Our findings also contribute to the debate on decision-making processes and, more specifically, on the possible influences between action and decision. Indeed, seminal theories proposed a serial view of the decision process (Newell & Simon, 1972). That is, according to these serial models, the decision would be first taken (i.e., the accumulation of relevant information to guide the choice for the task at end would reach a certain threshold) and, only later, the action (i.e., pressing a button to express the choice) would be executed. Yet, other theories supported by recent mouse-tracking findings rather proposed a parallel view, in which the decision would be continuously revised (e.g., Freeman & Ambady, 2010; Resulaj, Kiani, Wolpert, & Shadlen, 2009), hence conceiving it as an online process that can be constantly updated even during action execution. This is supported by several studies showing that in mouse-tracking tasks participants typically start the response movement before the completion of the decision process: for this very reason, the action's spatial and temporal dynamics would be paramount in unveiling the stages of the decision process (Freeman, Dale, & Farmer, 2011; Selen, Shadlen, & Wolpert, 2012; Spivey et al., 2005). Our data, with the effect of semantic relatedness on the online spatial parameter of the movement (i.e., MD), thus provide further support to the parallel view and suggest an online exploration of semantic memory contents while solving the task at hand.

Although DSMs are known to encode concreteness, up to a certain level (Hollis & Westbury, 2016), this piece of information does not overlap with semantic relatedness: related words, in fact, do not necessarily manifest the same degree of concreteness (church vs. religion, flag vs. patriotism). Indeed, we purposely selected the stimuli included in this study in order to allow for an easy detection of the target stimulus, as also testified by participants' accuracy being higher than 94% in both experiments. This could have consequently limited the impact of this dimension in the behavioral measures analyzed. Accordingly, in our analyses (i) we observe a very low correlation between concreteness difference and semantic relatedness, and (ii) concreteness difference is dropped from model selection, for all the dependent variables tested. The results of the present study, therefore, seem to genuinely depend on the relatedness between word meaning, in a way that is largely independent from their concreteness level (notwithstanding participants were asked to perform a concreteness-judgment task): the more the competitor word will be related to the target word, the more the response will be affected in terms of decision conflict. However, it should be noted that the results reported here could be partially dependent on the type of DSM used. That is, here we employed a DSM obtained via the application of neural networks, and in particular, the CBOW method. Other methods, as the Skip-Gram or the GloVe, or other parameter settings could in principle lead to slightly different results.

Taken together, our study has relevant implications from both theoretical and methodological points of view. On a theoretical level, our data clarify the participation of semantic memory when the performance is measured through fine-grained hand movements. This adds to previous evidence showing that DSMs are able to predict with considerable accuracy human performance across a wide range of semantic tasks, such as multiple-choice tests (Bullinaria & Levy, 2007), word categorization (Baroni & Lenci, 2010), word relatedness ratings (Bruni, Tran, & Baroni, 2014), word naming and lexical decision (Marelli & Amenta, 2018; Gatti et al., 2022b), as well as recognition memory (Gatti et al., 2022a, Gatti, Marelli, Mazzoni, Vecchi, & Rinaldi, 2022c) and spatial judgments, (Gatti, Marelli, Vecchi, & Rinaldi, 2022d), providing in turn further support to the idea that DSMs are indeed extremely efficient in capturing the structure of human semantic memory (Günther et al., 2019). On a methodological level, the fact that DSMs paired with mouse-tracking can be used to investigate deep decision-making features of human behavior, open new avenues for probing the detailed processes subserving the mechanisms of semantic memory.

In conclusion, using distributional semantic models paired with mouse-tracking, we provided evidence for semantic memory involvement in decision-making as measured by hand movements. Our findings well complement previous findings accounting for participants' behavior in semantic tasks and, more generally, provide insights on the impact of semantic memory on cognitive operations.

Funding information

This work was supported by funding from the Italian Ministry of Health (Ricerca Corrente 2023) to LR.

Conflict of interest statement

The authors declare no conflict of interest.

Data availability statement

The data analyzed in this study and the full .txt file of the analyses are available at https://osf.io/948rp/

Note

1 Note that here: https://osf.io/948rp/ we report a .txt file including all the analyses performed, together with those performed on the entire sample in Experiment 2.

Specifically, we found comparable effects across all the dependent variables, except for sample entropy, for which we found an effect of SRel when including also trackpad users.

References

Afonso, O., Suárez-Coalla, P., Cuetos, F., Ibáñez, A., Sedeño, L., & García, A. (2019). The embodied penman: Effector-specific motor-language integration during handwriting. *Cognitive Science*, *43*, e12767.

Barca, L., & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLOS ONE*, *7*(4), e35932.

Barca, L., & Pezzulo, G. (2015). Tracking second thoughts: Continuous and discrete revision processes during visual lexical decision. *PLOS ONE*, *10*(2), e0116193.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*, 673–721.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Dale, R., Hindy, N. C., & Spivey, M. J. (2006). Feature-semantic gradients in lexical categorization revealed by graded manual responses. Proceedings of the Annual Meeting of the Cognitive Science Society.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*, 15–28.

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315–323.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*(4), 502–517.

García, A. M., & Ibáñez, A. (2016a). A touch with words: Dynamic synergies between manual actions and language. *Neuroscience and Biobehavioral Reviews*, *68*, 59–95.

García, A. M., & Ibáñez, A. (2016b). Hands typing what hands do: Action-semantic integration dynamics throughout written verb production. *Cognition*, *149*, 56–66.

García-Marco, E., Morera, Y., Beltrán, D., de Vega, M., Herrera, E., Sedeño, L., Ibáñez, A., & García, A. M. (2019). Negation markers inhibit motor routines during typing of manual action verbs. *Cognition*, *182*, 286–293.

Gatti, D., Rinaldi, L., Marelli, M., Mazzoni, G., & Vecchi, T. (2022a). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General*, *151*(2), 363–389.

Gatti, D., Marelli, M., & Rinaldi, L. (2022b). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*(3), 851–863.

Gatti, D., Marelli, M., Mazzoni, G., Vecchi, T., & Rinaldi, L. (2022c). Hands-on false memories: A combined study with distributional semantics and mouse-tracking. *Psychological Research*, *87*(4), 1129–1142.

Gatti, D., Marelli, M., Vecchi, T., & Rinaldi, L. (2022d). Spatial representations without spatial computations. *Psychological Science*, *33*(11), 1947–1958.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*, 626–653.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401.

Hindy, N. C., Hamilton, R., Houghtling, A. S., Coslett, H. B., & Thompson-Schill, S. L. (2009). Computer-mouse tracking reveals TMS disruptions of prefrontal function during semantic retrieval. *Journal of Neurophysiology*, *102*(6), 3405–3413.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015a). Hidden processes in structural representations: A reply to Abbott, Austerweil and Griffiths (2015). *Psychological Review*, *122*(3), 570–574.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015b). Models of semantic memory. In J. R. Busemeyer, Z. Wang, & J. T. Townsend, A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). Oxford University Press.

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods* (pp. 111–130). New York: Routledge.

Lins, J., & Schöner, G. (2019). Computer mouse tracking reveals motor signatures in a cognitive task of spatial language grounding. *Attention, Perception, & Psychophysics*, *81*(7), 2424–2460.

Long, J. A. (2020). _jtools: Analysis and Presentation of Social Scientific Data_. R package version 2.1.0. Retrieved from https://cran.r-project.org/package=jtools

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Marelli, M. (2017). Word-embeddings Italian Semantic Spaces: A semantic model for psycholinguistic research. *Psihologija*, *50*(4), 503–520.

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*, *50*(4), 1482–1495.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the affective norms for English words (ANEW) for Italian. *Behavior Research Methods*, *46*(3), 887–903.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Papesh, M. H., & Goldinger, S. D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychonomic Bulletin & Review*, *19*(5), 906–913.

Papesh, M. H., Hicks, J. L., & Guevara Pinto, J. D. (2019). Retrieval dynamics of recognition and rejection. *Quarterly Journal of Experimental Psychology*, *72*(9), 2328–2341.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*(10), 1–8.

Peirce, J. W., & MacAskill, M. R. (2018). *Building experiments in PsychoPy*. London: Sage.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Rinaldi, L., & Marelli, M. (2020). Maps and space are entangled with language experience. *Trends in Cognitive Sciences*, *24*, 853–855.

RStudio Team. (2015). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/
Selen, L. P., Shadlen, M. N., & Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*, *32*(7), 2276–2286.
Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.
Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*, *22*(6), 531–543.
Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *Mental Lexicon*, *11*(3), 350–374.