Appendix A

Additional analyses on task performance

Interaction between hint and target condition on trial accuracy

Table A1

Experiment 1. Parameters *of the best-fitting mixed-effects logistic regressions for SWOW similarity (mHxT1p), containing both hint condition and target condition and an interaction. The reference categories are* AoA *(hint condition) and* Equal *(target condition). ICC = Intraclass Correlation Coefficient.*

Term	*b*	95% CI	*z*	*p*
Intercept	-0.87	[-1.07, -0.67]	-8.49	< .001
WF hint	-0.40	[-0.69, -0.11]	-2.67	.008
InStr hint	0.15	[-0.13, 0.42]	1.06	.291
AoA target	0.16	[-0.12, 0.43]	1.13	.260
WF target	-1.22	[-1.57, -0.88]	-6.94	< .001
InStr target	0.69	[0.42, 0.96]	5.04	< .001
WF hint × AoA target	0.23	[-0.17, 0.63]	1.12	.262
InStr hint × AoA target	-0.23	[-0.62, 0.16]	-1.14	.253
WF hint × WF target	1.43	[0.97, 1.89]	6.14	< .001
InStr hint × WF target	0.72	[0.27, 1.17]	3.15	.002
WF hint × InStr target	0.03	[-0.37, 0.42]	0.13	.899
InStr hint × InStr target	0.10	[-0.28, 0.48]	0.53	.598
ICC	0.036			

Table A1 shows the full results for the best-fitting mixed-effects logistic regression model for trial accuracy (SWOW similarity), including the effects of hint condition, target condition, and the effect of hint and target condition. The pattern of the results appears to

show that for the WF target words, the effect of WF and InStr hints giving higher accuracy was stronger than for the other target conditions, rather than the AoA and InStr hints being generally more effective.

Number of guesses

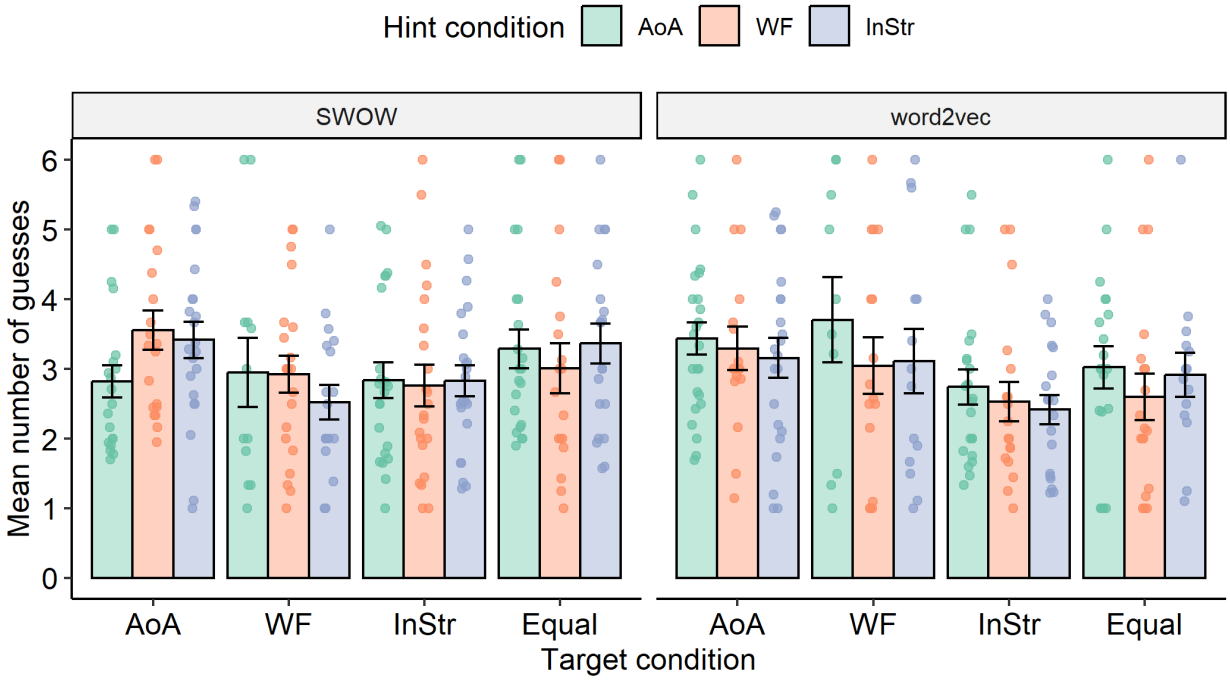

Figure A1

Experiment 1: Mean number of guesses *on successfully-guessed trials by hint and target condition. The results are displayed separately for each similarity type used to select hints. Each dot represents the mean number of guesses for a target word in that hint (colour) and target (x axis) condition. Overall, there is little systematic difference between the hint conditions or target words.*

Experiment 1

As Figure A1 shows, there was little systematicity in how many guesses it took people to guess correctly: regardless of the nature of the hint word, the target word, or the similarity type, people averaged 2.67 guesses, with the majority (54.5%) of trials being guessed after one or two attempts. Given the relatively low number of trials on which there

was a correct guess at all and the fact that this measure is conditioned on that, this may not be too surprising.

Table A2

Experiment 1 model comparison: Number of guesses. *We compare a set of pre-registered mixed-effects linear regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs word2vec). All models have number of guesses as the continuous outcome variable. Models vary in the presence of a random effect for participant, target word condition (`target`), hint condition (`hint`), and interaction(s). The best-fit model for SWOW similarity was the null model while the best-fit model for word2vec contained target word condition as a predictor.*

		BIC	
Model	**Description**	SWOW	word2vec
mNull	`guesses ~ 1`	**6721**	5120
mHxT	`guesses ~ hint * target`	6758	5123
m1p	`guesses ~ 1 + (1 | participant)`	6732	5132
mH1p	`guesses ~ hint + (1 | participant)`	6750	5144
mT1p	`guesses ~ target + (1 | participant)`	6738	**5109**
mHT1p	`guesses ~ hint + target + (1 | participant)`	6755	5119
mHxT1p	`guesses ~ hint * target + (1 | participant)`	6790	5155

As with accuracy, we first conducted pre-registered analyses (on only the trials that were guesses successfully) to assess whether the similarity metric that was used to generate the hints had an effect. Two 2×3 two-way repeated measures ANOVAs were conducted with number of guesses as the outcome measure and hint condition and similarity metric as the two predictors. There was no significant effect, but because separate analyses were conducted for SWOW and word2vec for accuracy, we did the same here.

Thus, for each similarity type, we conducted a 4×3 two-way mixed ANOVA with mean number of guesses as the outcome and target condition and hint condition as the predictors (see Figure A1). For the trials with SWOW hints, there were no significant effects. However, for word2vec trials, there was a significant effect of target condition, $F(3, 41) = 4.47$, $p = .008$, with post-hoc tests showing that InStr target words required significantly fewer guesses than AoA target words ($p = .02$); in other words, performance was better on InStr target words. There were no significant effects of hint condition, $F(1.61, 66.07) = 0.02$, $p = .969$, or the interaction, $F(4.83, 66.07) = 0.22$, $p = .949$.

As pre-registered, we also conducted mixed-effects models for the number of guesses, analogous to the models for accuracy. The results, shown in Table A2, show that for SWOW similarity, the best-fitting model was the null model, consistent with the ANOVA results.

For word2vec similarity, the best-fitting model was `mT1p`, which contained a fixed effect of target word condition. Relative to the reference category Equal, the InStr target words required significantly fewer guesses, $b = -0.26$ $[-0.47, -0.05]$, while AoA targets required significantly more, $b = 0.48$ $[0.26, 0.70]$.

Experiment 2

Similarly to Experiment 1, preliminary analyses assessing the effects of similarity metric showed that there was a significant effect on mean number of guesses, $F(1, 89) = 4.37$, $p = .039$, in which better performance was attained for SWOW hints compared to word2vec hints. There were no significant interactions between hint condition and similarity metric.

The ANOVA analyses on mean number of guesses (see Figure A2) showed differences between the target conditions for both the SWOW, $F(3, 91) = 3.47$, $p = .019$, and word2vec trials, $F(3, 86) = 6.34$, $p < .001$. Post-hoc tests showed that, for SWOW trials, InStr targets required significantly fewer guesses than AoA targets ($p = .014$), and for word2vec trials, InStr targets required significantly fewer guesses than both AoA

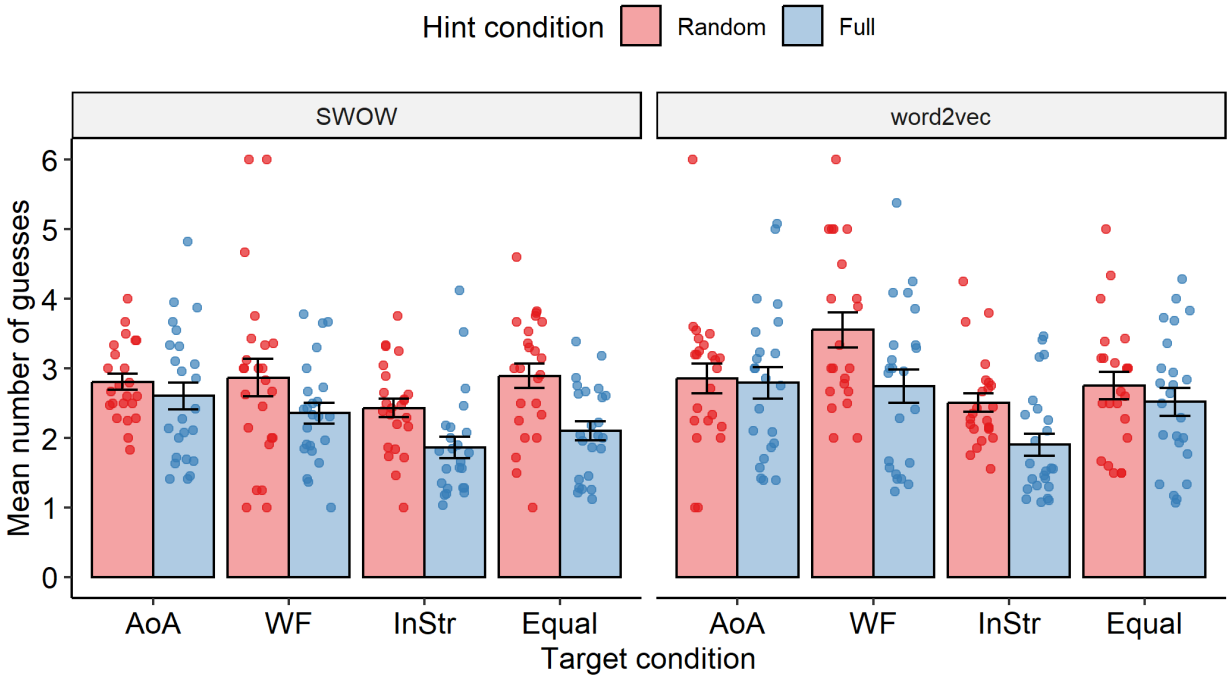

Figure A2

Experiment 2: Mean number of guesses *on successfully-guessed trials by hint and target condition. The results are displayed separately for each similarity type used to select hints. Each dot represents the mean number of guesses for a target word in that hint (colour) and target (x axis) condition. Overall, there is little systematic difference between the hint conditions or target words.*

($p = .015$) and WF targets ($p < .001$). This is consistent with a guessing advantage for InStr words, even conditioning only on correct trials.

Unlike in Experiment 1 but consistent with our accuracy results for Experiment 2, we found significant effects of hint condition for both the SWOW, $F(1, 91) = 20.57$, $p < .001$, and word2vec trials, $F(1, 86) = 10.29$, $p = .002$. The Full vocabulary hints required fewer guesses on average, but there was no significant interaction effect in either the SWOW, $F(3, 91) = 0.96$, $p = .414$, or word2vec trials, $F(3, 86) = 2.57$, $p = .059$.

As before, we also compared mixed-effects models with number of guesses as the outcome variable and participant as the random effect. Table A3 shows the results, which

Table A3

Experiment 2 model comparison: Number of guesses. *We compare a set of pre-registered mixed-effects linear regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs word2vec). All models have number of guesses as an outcome variable (`guesses`). Models vary in the presence of a random effect for participant, target word condition (`target`), hint condition (`hint`), and interaction(s). For both similarity metrics, the best-fit model contained no random effect of participant but both target word condition and hint condition as predictors and an interaction between them.*

		BIC	
Model	**Description**	SWOW	word2vec
mNull	`guesses ~ 1`	11214	9661
mHxT	`guesses ~ hint * target`	**11097**	**9561**
m1p	`guesses ~ 1 + (1 | participant)`	11217	9673
mH1p	`guesses ~ hint + (1 | participant)`	11154	9640
mT1p	`guesses ~ target + (1 | participant)`	11195	9624
mHT1p	`guesses ~ hint + target + (1 | participant)`	11113	9574
mHxT1p	`guesses ~ hint * target + (1 | participant)`	11128	9592

indicate that the best-fitting models for both similarity types contained hint condition and target condition but no random effects. The parameters of the best-fit models, shown in Table A4, indicate that for both similarity types the InStr target words required significantly fewer guesses than the reference condition (Equal) while the WF target words required significantly more. The Full vocabulary hints also required significantly fewer guesses than the Random vocabulary hints in both.

Table A4

Experiment 2: Parameters *of the best-fit linear regression models (mHxT) with number of guesses as the outcome and hint and target condition as predictors. The reference categories are* Random *(hint condition) and* Equal *(target condition).*

	SWOW				word2vec			
Term	*b*	95% CI	*t*	*p*	*b*	95% CI	*t*	*p*
Intercept	3.02	[2.81, 3.22]	29.50	$< .001$	2.81	[2.55, 3.06]	21.49	$< .001$
Full hint	-0.99	[-1.22, -0.76]	-8.36	$< .001$	-0.44	[-0.73, -0.15]	-3.01	.003
AoA target	-0.24	[-0.55, 0.07]	-1.54	.125	0.01	[-0.35, 0.37]	0.06	.949
WF target	-0.29	[-0.60, 0.01]	-1.87	.062	0.49	[0.06, 0.92]	2.22	.026
InStr target	-0.66	[-0.91, -0.42]	-5.27	$< .001$	-0.35	[-0.66, -0.05]	-2.25	.024

Appendix B

Hint similarity and interrelatedness

Experiment 2

Table B1

Experiment 2: Model comparisons for semantic similarity *analyses. We compare a set of mixed-effects linear regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs word2vec). All models have a random effect for target word (`targetWord`) and use mean `accuracy` of each target word for a given hint type as the outcome variable. Models vary in the presence of predictors for average hint-target similarity (`hintSim`) and hint interrelatedness (`hintRel`) for that word. Both best-fitting models contain both predictors, but for SWOW there is also an interaction between hint similarity hint and relatedness.*

		BIC	
Model	**Description**	SWOW	word2vec
mNull	`accuracy ~ 1 + (1 | targetWord)`	111	130
mS	`accuracy ~ hintSim + (1 | targetWord)`	-30	-20
mR	`accuracy ~ hintRel + (1 | targetWord)`	8	33
mRS	`accuracy ~ hintSim + hintRel + (1 | targetWord)`	-29	**-25**
mRxS	`accuracy ~ hintSim * hintRel + (1 | targetWord)`	**-54**	**-25**

In order to analyse the effects of hint-target similarity and hint relatedness, we performed model comparison between a set of mixed-effects models which all had target word as a random effect and an outcome variable of mean accuracy for each target word in a given hint condition and similarity type. Both predictors were mean-centered. The results, shown in Table B1, reveal that for both similarity types, accuracy was best predicted by both hint-target similarity and hint interrelatedness. Hint-target similarity

had a significant positive relationship with mean accuracy on SWOW trials, $b = 2.23$ $[1.60, 2.85]$, as well as word2vec ones, $b = 2.86$ $[2.12, 3.61]$. This indicates that, as one would expect, targets with more semantically-related hints were guessed more easily. As before, hint interrelatedness had a significant negative effect on mean accuracy on word2vec trials, $b = -0.78$ $[-1.53, -0.03]$, indicating that more semantically distinct hints were better, but this effect was not significant for SWOW trials, $b = -0.24$ $[-0.86, 0.383]$. This was qualified by a significant interaction effect for SWOW trials, $b = -4.36$ $[-5.85, -2.87]$, which again indicated that hints that were more diverse benefited from being more similar to the target.

Appendix C

Relationship between InStr and semantic similarity

Figure C1 shows the semantic similarity between each hint and target for the core-word hints in Experiment 1. Despite being computed on the SWOW data, hints selected using the SWOW similarity metric did not have the highest similarity to the InStr targets. InStr targets also did not have the highest hint-target similarity when word2vec was used. Thus, neither similarity metric favoured the InStr target condition. Regardless of this, the results reported in the main paper showed that InStr targets were still guessed more accurately even when taking semantic similarity into account. Figure C2 visualises this result, which corresponds closely to the analyses reported in the paper involving hint-target similarity, only that hint interrelatedness was also included in those models.

Finally, although the InStr measure and SWOW similarity use the same data,

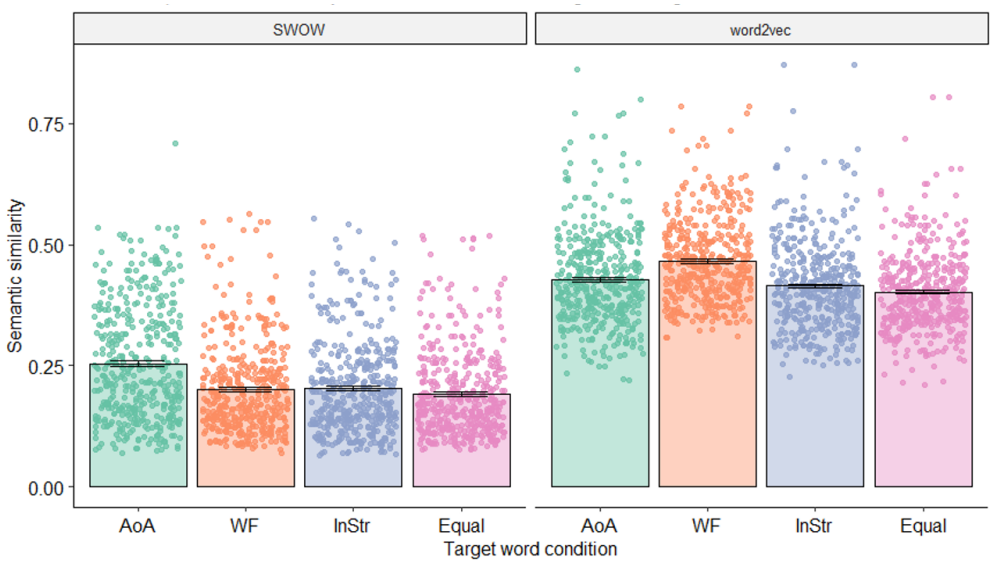

Figure C1

Experiment 1: Semantic similarity *between hints and targets in each target condition. Each dot represents the similarity between a hint and target using SWOW or word2vec in each target condition (x axis). Overall, AoA target words have the highest similarity for the SWOW hints, and the WF target words have the highest similarity for word2vec hints.*

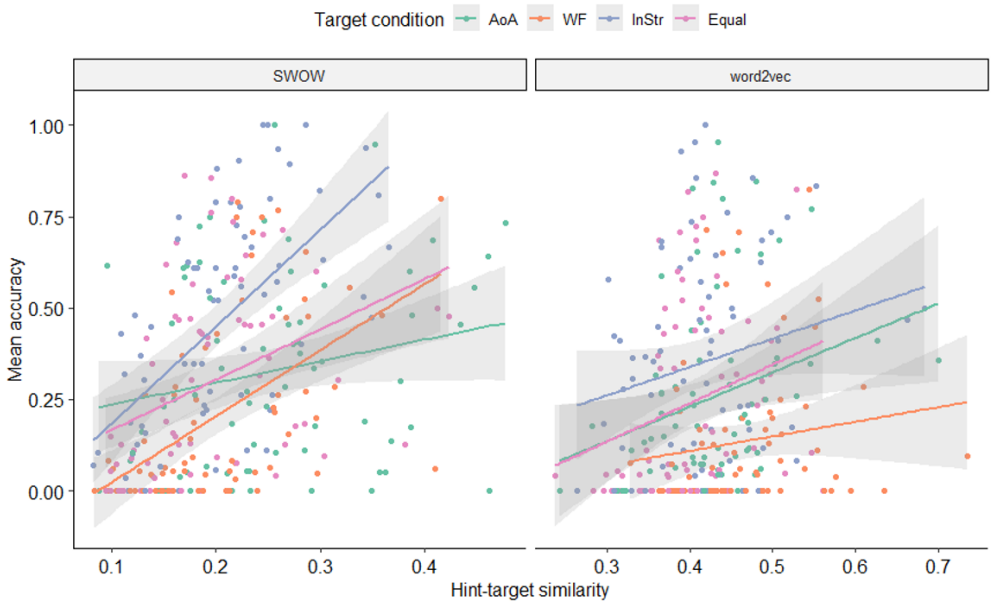

Figure C2

Experiment 1: Mean accuracy *by hint-target similarity and target condition. The results are displayed separately for each similarity type used to select hints. Each dot represents the mean accuracy for a target word in a given hint condition. The similarity between the target and each of its hints in that condition is averaged and shown on the x-axis. For both SWOW and word2vec conditions,* InStr *targets remained more accurately guessed than the other target conditions after accounting for semantic similarity.*

they are computed in very different ways: the former is based on association strength over all words in the dataset, whereas the latter is based on similarity to a limited set of core words. Similarity, the extent to which words have the same meaning or fulfill the same function, and association strength, measuring words that are related in meaning, are psychologically different constructs, and they are only weakly related, as Figure C3 shows. The same is true of word2vec similarity (also shown in Figure C3), indicating that association and similarity are distinct measures.

However, what if there is still something about these metrics that favours InStr for some reason? The strongest test of this concern would be to use a totally different

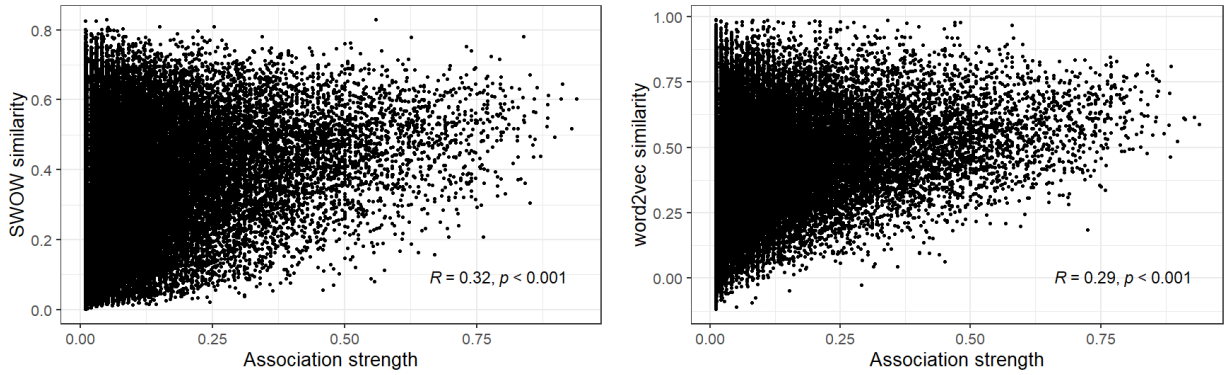

Figure C3

Relationship between the association strength *of word pairs in SWOW and their semantic similarity according to SWOW (left) and word2vec (right). Association and similarity are only weakly related, even when both are computed on the same data (SWOW). Pearson's r and associated p value are displayed for each.*

similarity metric and see if the results still hold. Another well-known measure of similarity is human similarity ratings, which aim to explicitly measure similarity in the sense of words that have the same meaning (Hill, Reichart, & Korhonen, 2015). There is a practical concern to using similarity ratings, however, as even the most extensive similarity datasets, such as SimVerb-3500 (Gerz, Vulić, Hill, Reichart, & Korhonen, 2016), have ratings for only 3,500 word pairs, whereas 86,400 hint-target similarity values were required to perform hint selection in Experiment 1, making large-scale datasets like SWOW and word2vec ideal for this task. That said, we perform a re-analysis of the key result (shown in Figure C1), replacing the SWOW and word2vec similarity with human similarity ratings from both SimLex-999 (Hill et al., 2015) and SimVerb-3500 (Gerz et al., 2016). Because only 241 out of the 3,456 hint-target pairs in Experiment 1 were found in the SimLex databases, our analyses only include those pairs. Nevertheless, as Figure C4 shows, the pattern of results is the same: the InStr target words achieved higher accuracy than the other target conditions, even after controlling for hint-target similarity based on human judgments. This suggests that even if the hints could have been selected based on human similarity

ratings, our qualitative findings would be the same.

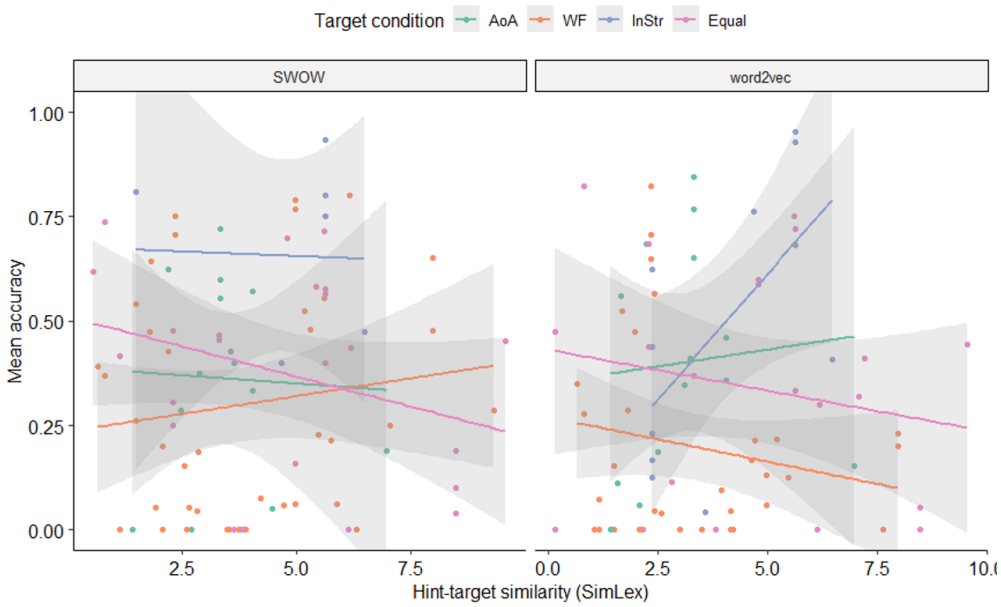

Figure C4

Experiment 1: Relationship between mean accuracy *and mean hint-target similarity (using human similarity judgments from SimLex and SimVerb) by target condition in each similarity metric condition. Each dot represents one target word in a given hint condition. The hint-target similarity based on human judgments, shown on the x-axis, was calculated for each of the hint-target pairs for which similarity ratings data were available (241 out of 3,456 total word pairs).*

Appendix D

Alternative measures of DSM coreness

In order to define alternative definitions of distributional coreness to word frequency, we leveraged word-word co-occurrence data from the Corpus of Contemporary American English (COCA; Davies, 2008-) and word embeddings from word2vec. These can be considered more principled definitions of coreness in DSMs compared to word frequency, because they take into account the relationships between words in text (rather than simple unigram counts), and leverage representations of words learned by DSMs (in the case of word2vec).

We defined and investigated 7 measures of DSM coreness:

- COCA: Word frequency computed by summing across COCA co-occurrence counts for each word, serving as a reliability comparison against the SUBTLEX data

- CCCen: Co-occurrence strength centrality: Strength centrality calculated over co-occurrence data, in an analogous way to InStrength for the word association data. Strengths for each word were calculated by dividing each co-occurrence count by the frequency of the context word, and then strength centrality was computed by summing over strengths.

- PPMI: Sum over PPMI-transformed co-occurrence counts for each word. The resulting metric can be thought of as the words which give the most information about all other words.

- w2vSum: word2vec summed distance: Summed Euclidean distance of each word from every other word. This measure indicates how far each word is situated from all other words in word2vec vector space, in terms of Euclidean distance. Lower summed distance indicates words which are closer to other words on the whole and thus greater coreness.

- W2VCEN: word2vec distance strength centrality: Strength centrality computed over a network constructed over word2vec representations. Words that were closer than 1 in terms of Euclidean distance in word2vec vector space were linked, with the strength of the links weighted proportionally to their distance. Strength centrality was computed by summing over strengths. This measure is in principle different to W2VSUM (since only words within a certain distance are connected up in the network), but the results were near-identical (see Figure D1), no matter what distance cutoff was used.

- W2VCOSSUM: word2vec summed cosine similarity: Summed cosine similarities of each word to every other word. Higher summed cosine similarity indicates words which are more similar to other words overall and thus greater coreness.

- W2VCOSCEN: word2vec cosine similarity strength centrality: Strength centrality computed over a network constructed over word2vec representations. Words that were above .5 in cosine similarity were linked, with the strength of the links equal to their cosine similarity. Strength centrality was computed by summing over strengths.

Just as with the original coreness measures, function words were removed and words were standardised to US spelling. Words in the COCA data were lemmatised, however, because embedding representations for different inflectional forms of the same word are different in principle, words in the word2vec data were not lemmatised. With the exception of W2VCOSSUM, all other measures were log10-transformed. Coreness measures were normalised in the same way as described in the main paper, with 0 indicating maximum coreness on each measure.

The top 10 core words as defined by each of the measures are shown in Table D1. Some of these, such as COCA and CCCEN, produce very similar core words to WF, although this is not surprising since the measures are computed in very similar ways. Others produce core words of a somewhat surprising nature; for example, the word2vec measures tend to produce names of people and adverbs which serve discourse functions

Table D1

Top 10 core words as defined by the DSM coreness measures.

	COCA	CCCen	PPMI	w2vSum	w2vCen	w2vCosSum	w2vCosCen
1	like	like	refer	characteristically	characteristically	seeming	jesse
2	know	know	include	aforementioned	aforementioned	seemingly	ryan
3	good	say	write	uncharacteristically	straightforwardly	merely	jamie
4	think	come	begin	unfortunately	uncharacteristically	perversely	laura
5	time	think	speak	consequently	consequently	ostensibly	steven
6	come	go	appear	straightforwardly	notwithstanding	ihey	jason
7	say	good	draw	notwithstanding	simultaneously	evidently	aaron
8	go	time	travel	simultaneously	unfortunately	paradoxically	jeff
9	want	want	feature	indistinguishable	institutionalization	sort	idiotic
10	people	look	claim	incomprehensible	counterproductive	overwrought	rebecca

(e.g., unfortunately, consequently, notwithstanding). One possible explanation for this is that these are words which may be used versatilely across lots of different linguistic contexts, and are therefore central in a word2vec sense; however, it is not clear that these are words that are intuitively core in the sense of being central to semantic representations or useful for communication. This highlights that operationalising coreness in a distributional model is not necessarily straightforward. Figure D1 shows the correlations between the DSM coreness measures. It shows that some measures are systematically related (e.g., COCA and CCCen) whereas other measures are much more distinct.

Target word conditions

Figure D2 shows a re-analysis of the target word condition design when, instead of frequency, the distributional coreness measure is substituted for each of the DSM coreness measures listed above. Due to the different scales of the variables, the coreness normalisation process produces very different coreness values for some measures compared to others (e.g., w2vCen); this implies that the target words, which had the status of being core words in the original experiment, were *not* as core under certain of these measures –

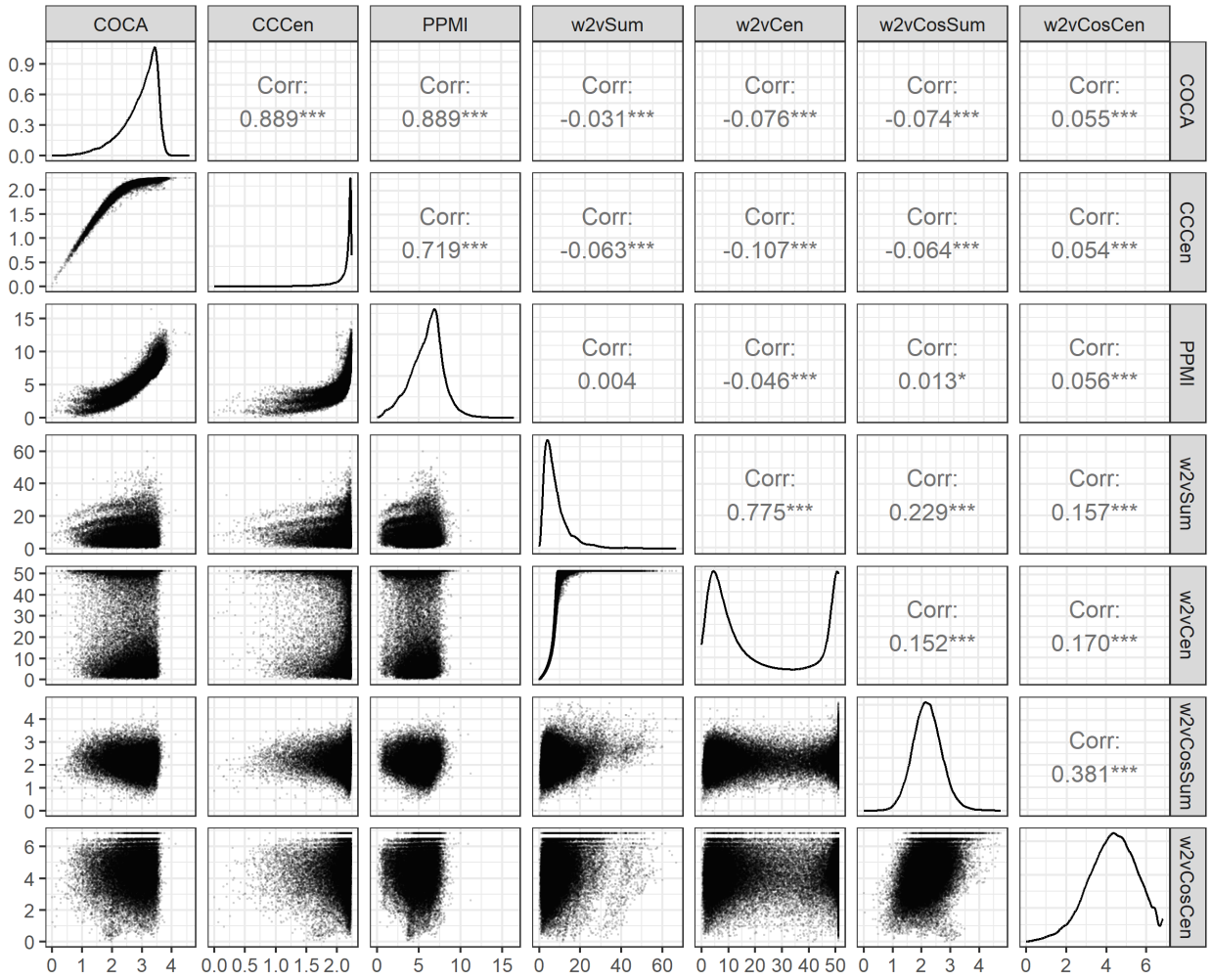

Figure D1

Scatterplot matrix showing the relationships between the alternative measures of DSM coreness (COCA, CCCen, PPMI, w2vSum, w2vCen, w2vCosSum, and w2vCosCen). All measures are defined in-text. Some measures are very closely related (e.g., COCA and CCCen), whereas others are highly distinct.

instead, these measures produce very different kinds of core words, as seen in Table D1.

The re-analysis of target conditions shows that in general, the WF (distributionally core) target words still have greater coreness on the corresponding distributional coreness measure compared to the other target word conditions. The notable exception is the

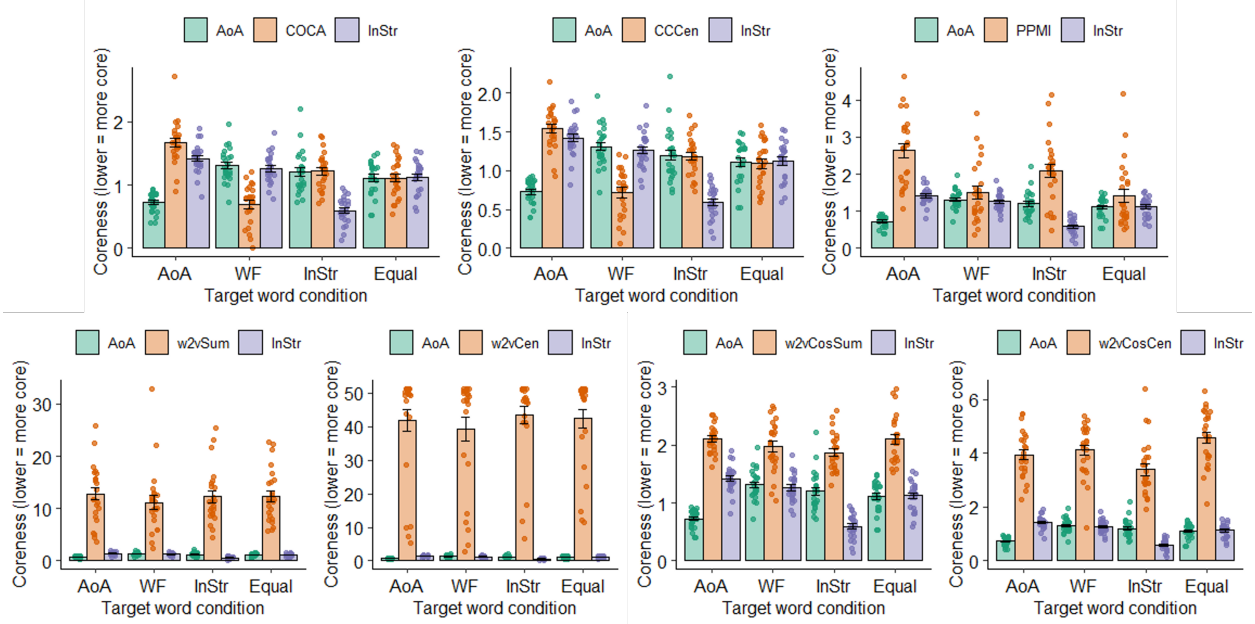

Figure D2

Re-analysis of the coreness of each target word in the four target conditions (AoA, WF, InStr, and Equal) when WF coreness is replaced by each alternative distributional coreness measure (COCA, CCCen, PPMI, w2vSum, w2vCen, w2vCosSum, and w2vCosCen). Coreness values closer to 0 represent words that are more core. In general, the WF targets were more core on the distributional coreness measures compared to the other target conditions, except for the measures based on word2vec cosine similarity (w2vCosSum and w2vCosCen) which have greater coreness for the InStr target condition.

word2vec coreness measures based on cosine similarity, which show that the InStr target words are greatest in distributional coreness. This shows that to some extent, the experimental target condition design that aimed to select target words that were more core on a given coreness measure and less core on the others (originally based on WF) generalises to these alternative measures of DSM coreness, and that the resulting findings for target words generalise to these measures too.

Target word accuracy

Experiment 1

Table D2 shows the results of linear regression models predicting target word accuracy from three coreness mesures. This provides a somewhat analogous analysis to comparing accuracy across target word conditions. In each linear regression model, target word accuracy is predicted by AoA, DSM, and InStr, where DSM is one of the DSM coreness measures defined above. As in the exploratory analyses in the main paper, target word accuracy was averaged across the three hint conditions.

Table D2

Experiment 1: Linear regression models *predicting target word accuracy from target word coreness. Each row represents one linear model predicting accuracy from AoA, DSM, and InStr coreness, where DSM coreness is defined by the measure in the left-most column. Coreness values closer to 0 represent words that are more core. Greater AoA and InStr coreness consistently and significantly predicts higher accuracy, whereas greater DSM coreness does not.*

Term	AoA			DSM			InStr		
	b	95% CI	*p*	*b*	95% CI	*p*	*b*	95% CI	*p*
COCA	-0.23	[-0.32, -0.13]	$< .001$	-0.01	[-0.08, 0.07]	.841	-0.28	[-0.37, -0.19]	$< .001$
CCCen	-0.23	[-0.32, -0.13]	$< .001$	-0.01	[-0.10, 0.08]	.794	-0.28	[-0.37, -0.19]	$< .001$
PPMI	-0.23	[-0.33, -0.13]	$< .001$	-0.01	[-0.04, 0.02]	.523	-0.28	[-0.37, -0.20]	$< .001$
w2vSum	-0.23	[-0.33, -0.13]	$< .001$	0.00	[-0.01, 0.01]	.758	-0.29	[-0.37, -0.20]	$< .001$
w2vCen	-0.22	[-0.32, -0.13]	$< .001$	0.00	[0.00, 0.00]	.947	-0.29	[-0.37, -0.20]	$< .001$
w2vCosSum	-0.22	[-0.32, -0.13]	$< .001$	0.11	[0.02, 0.20]	.013	-0.32	[-0.40, -0.23]	$< .001$
w2vCosCen	-0.22	[-0.32, -0.13]	$< .001$	-0.02	[-0.05, 0.01]	.260	-0.27	[-0.36, -0.19]	$< .001$

The results show a very clear pattern: in every model, both AoA and (word association) InStr significantly predict accuracy, with higher coreness on those measures

predicting greater accuracy (values closer to 0 indicating greater coreness, meaning that a negative coefficient shows this). By contrast, every single DSM coreness measure predicts no significant effect on target word accuracy, or predicts an effect in the opposite direction (in the case of word2vec summed cosine similarity), meaning that less core target words had greater accuracy. These findings corroborate what was found in the main paper in terms of the differences between the target conditions, and extend these findings to different ways of defining DSM coreness.

Experiment 2

Similarly, linear regression models predicting target word accuracy were conducted for Experiment 2. This time, accuracy was divided across the Random and Full vocabulary conditions, and the effect of hint condition (Full hint) was added to the model, alongside the three coreness measure predictors.

The effect of the Full over the Random vocabulary hint condition was always significant. (Word association) InStr remained a consistent significant predictor of target word accuracy in every model, with words that were more core in InStr achieving higher accuracy. However, this time, the effect of AoA was washed out, never significantly predicting accuracy. Once again, the effect of the DSM coreness measure predictors was either non-significant or in the opposite direction with more core words achieving lower accuracy. These results again show that the findings from the main analyses can be generalised to alternative measures of coreness in DSMs.

Table D3

Experiment 2: Linear regression models *predicting target word accuracy from target word coreness. Each row represents one linear model predicting accuracy from hint condition (*Full *compared to* Random*),* AoA*, DSM, and* InStr *coreness, where DSM coreness is defined by the measure in the left-most column. Coreness values closer to 0 represent words that are more core.* Full *vocabulary hints, as well as greater* InStr *coreness consistently and significantly predicts higher accuracy, whereas greater* AoA *and DSM coreness does not.*

Term	Full hint			AoA			DSM			InStr		
	b	95% CI	*p*	*b*	95% CI	*p*	*b*	95% CI	*p*	*b*	95% CI	*p*
COCA	0.45	[0.41, 0.49]	< .001	-0.02	[-0.08, 0.04]	.549	0.09	[0.04, 0.14]	< .001	-0.38	[-0.43, -0.32]	< .001
CCCen	0.45	[0.41, 0.49]	< .001	-0.02	[-0.08, 0.04]	.497	0.11	[0.06, 0.17]	< .001	-0.38	[-0.43, -0.32]	< .001
PPMI	0.45	[0.41, 0.49]	< .001	-0.04	[-0.11, 0.02]	.207	0.00	[-0.02, 0.03]	.729	-0.34	[-0.39, -0.28]	< .001
w2vSum	0.45	[0.41, 0.50]	< .001	-0.04	[-0.11, 0.02]	.177	0.00	[0.00, 0.00]	.976	-0.34	[-0.39, -0.28]	< .001
w2vCen	0.45	[0.41, 0.49]	< .001	-0.04	[-0.10, 0.03]	.240	0.00	[0.00, 0.00]	.274	-0.33	[-0.39, -0.28]	< .001
w2vCosSum	0.45	[0.41, 0.49]	< .001	-0.04	[-0.11, 0.02]	.175	0.07	[0.02, 0.13]	.014	-0.36	[-0.41, -0.30]	< .001
w2vCosCen	0.45	[0.41, 0.49]	< .001	-0.04	[-0.11, 0.02]	.180	-0.02	[-0.04, 0.00]	.082	-0.32	[-0.38, -0.26]	< .001

References

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA).* Available online at `https://www.english-corpora.org/coca/`.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.