Supplementary Material

1 SUPPLEMENTARY MATERIALS

1.1 Further details of data preprocessing

To ensure the quality of data for the analysis of semantic change and obtaining phrase-BERT embeddings of the senses, we filtered DatSemShift database by including semantic shifts that satisfy the following three criteria: 1) the shifts are unidirectional, fall under either the type synchronic polysemy or semantic evolution, and contain source-target senses of the same word in the same language (namely, strictly within-language semantic change with no overt change in word form). We chose to work with synchronic polysemy or semantic evolution, because we did not want to consider instances where the morphology of a word changed. Additionally, in some shifts, the senses are marked as the broader semantic category instead of the shift itself (for example, ⟨animal⟩ representing the sense of a specific animal, and not the overall concept of animal). We only consider shifts in which the source and target senses are specific concepts, and not those broader semantic categories, since they constitute more than one concept.

To further clean the data and prepare for high-quality semantic embedding, we manually replaced all instances of British spellings with American ones in cases where they differ (e.g., replace `grey` with `gray`), replaced all cases of scientific naming of animals and plants with common names, and replaced any obscure words (e.g., `swearword` was replaced with `swear word`). We defined a word as obscure if it was not contained in the original corpus where semantic embeddings were trained. These replacements were made to generate phrases that more closely matched those that phrase-BERT was trained on.

There were some shifts for which the meaning of a word was explained within the shift. For these cases, we removed the word and just used the given definition - for example, we replaced `aer (veil covering vessels in the orthodox Church)` with `veil covering vessels in the orthodox Church`, or `smut (plant disease)` with `plant disease`. In other cases, certain senses had tags that clarified the meaning of a word, such as `mind (n.)`, indicating that `mind` is a noun. We removed these part-of-speech tags.

There were also a few instances where contrasting words were given to clarify a word's meaning, such as `land (vs. sea)`. In these cases, we removed all pairs of parentheses containing `vs.`, since phrase-BERT might struggle with interpreting that `(vs. sea)` is meant to clarify the meaning of `land`.

Finally, we replaced all instances of commas and semicolons with `or`, as `or` more directly demonstrates that the multiple definitions of a sense are being given. We replaced all hyphens with a space and removed all cases of parentheses.

The full list of string replacements used to clean the dataset can be found in Tables S1, S2, S3.

1.2 Alternative data samples for directionality inference

When we performed directionality inference, our method for assigning concreteness, valence, and frequency values excluded many of the shifts in the dataset. To evaluate whether our results are robust to a large set of data samples, we also tried averaging the variable values of every word (excluding function words) in the sense glosses — for example, this procedure would assign a concreteness rating to "to hold

(in hands)", instead of discarding it as a data point because "hold (in hands)" was not available in the concreteness database. We include these results here for completeness.

Figure S1 shows that our basic findings hold in this larger data sample. Concreteness is the most accurate predictor of directionality, followed by valence, then frequency. These models are less accurate than the conservative directionality tests we reported in the main text, but the relative performance of all the three predictors and the combined model holds the same.

1.3 Alternative semantic representations for target inference

For target inference, we used phrase-BERT embeddings as the primary semantic representation for word senses, and took the difference between source embedding and target embedding to represent a sense pair. To assess the robustness of our findings with respect to the choice of semantic representation, we also considered alternative semantic representations including sentence-BERT and word2vec. Since word2vec only has embeddings for individual words, we averaged vectors across words if a sense is annotated as a phrase in order to assign a vector to that phrase.

Our motivation for subtracting source and target embeddings was that this difference potentially captures the relationship between source and target, as opposed to isolated information about the source and the target. Our motivation for using phrase-BERT over sentence-BERT or word2vec was that the senses in DatSemShift were typically phrases, but not sentences or individual words.

To evaluate these different design choices empirically, we considered using different embeddings on the target inference task. We took the predicted target to be the one whose vector formed the smallest angle with the source vector (as a measure of similarity). The task samples were generated from taking targets randomly from DatSemShift.

The results for this analysis are summarized in Figure S2. We observed that overall, all three semantic representations yield good predictive accuracy in target inference, but phrase-BERT offers the most superior accuracy.

2 SUPPLEMENTARY TABLES AND FIGURES

Original String	Replaced With
vapour	vapor
honour	honor
organisation	organization
harbour	harbor
odour	odor
centre	center
analyse	analyze
theatre	theater
colour	color
rumour	rumor
behaviour	behavior
armour	armor
grey	gray
mould	mold
neighbour	neighbor
axe	ax
moustache	mustache
plough	plow
mandarine	mandarin
adj	
gipsy	
albumen	
campanula	
boletus edulis	penny bun fungus
ursus	
swearword	swear word
adj.	
coleus	
n.	
OK	ok
typha	cattail
pacifica	peaceful
mustella	
smail	
one's	
spurflower	perennial plant
sabre	
equus	
etc.	
ciconia	
aër	
panthera	panther
erinaceus	
e.g.	
centaurea	thistle
moschiferus	
apterus	
pyrrhocoris	
smn.	
pritchardia	
100	one hundred

Table S1. List of string replacements used to clean DatSemShift (part 1). Empty entries in replaced words correspond to deletions of the original string.

Original String	Replaced With
sabrefish	sabre carp
putorius	
adv.	
petromyzontidae	
botaurus	
standart	standard
leccinum	
sg.	
gemini	Gemini
tabanidae	
anagallis	
decorticate	stiff
albugo	
frangula	
sciurus	
scrofa	
relig.	
headstream	head stream
solanum	
anguilla	
anat.	
nectarinia	
ipomoea	
repaire	repair
vaccinium	
smth	
smth.	
bubo	
deflorate	remove flowers
tr.	
traveller	traveler
bubalis	
marmorata	
furuncul	
caballus	
microchiroptera	
urtica	
plumbum	
biol.	
intr.	
bubalus	
columba	
cucurbita	
goldcrest	small bird
melongena	
picea	
arvensis	
moschus	
psidium	
radiointerference	radio interference
owre	
ricinus	

Table S2. List of string replacements used to clean DatSemShift (part 2).

Original String	Replaced With
capricorn	goat zodiac sign
mustela	
pandion	
adj.of	
nomadize	become nomadic
smb.	
kneepit	knee pit
num.	
pl.	
extortioner	extortion doer
enculturate	assimilate
asquint	squint
uliginosum	
heteroptera	
abies	fir
stratiotes	
fiddlestick	violin bow
scabrum	
grus	bird
acarina	
guajava	
bitterling	freshwater fish
lycopersicum	
lutra	otter
plectranthus	
macereed	mace reed
24	twenty four
acris	
rotundifolius	common weed
gutturalis	
oxyeleotris	
geometrid	
citrullus	
lepus	
motacilla	
crake	bird
haliaëtus	
glasswort	herb
quinsy	throat abscess
shoulderblade	shoulder blade
spearthrower	spear thrower
ridgepole	ridge pole
pimpleface	pimple face
tumpline	backpack
cushma	clothing
curassow	tropical bird
banisterium	plant
paca	rodent
netbag	net bag
muntjacs	barking deer

Table S3. List of string replacements used to clean DatSemShift (part 3).

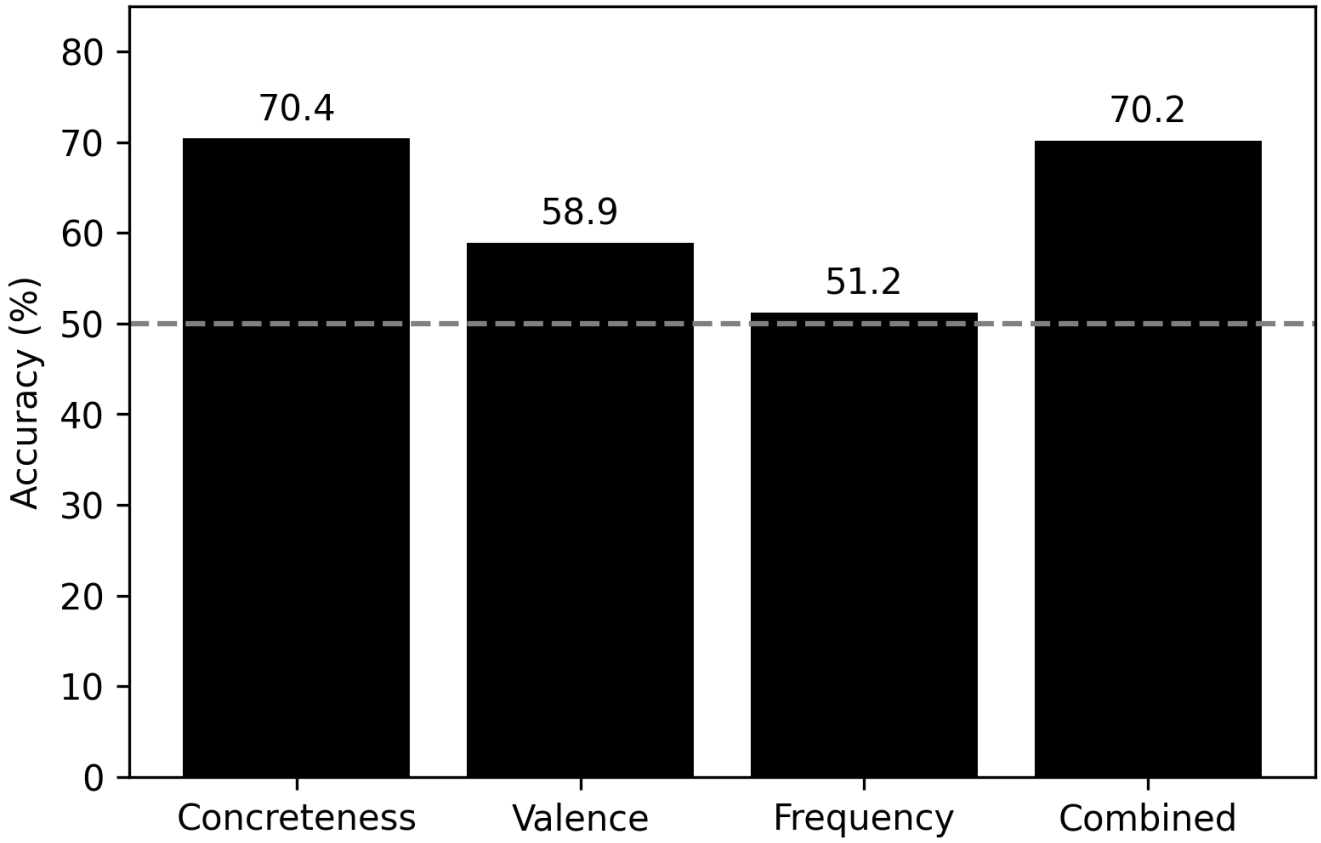

Figure S1. Predictive accuracy of concreteness, frequency, and valence in inferring directionality of semantic change in a larger sample of data. "Combined" refers to the logistic regression model that combines the three variables. Dashed line indicates chance accuracy (50%).

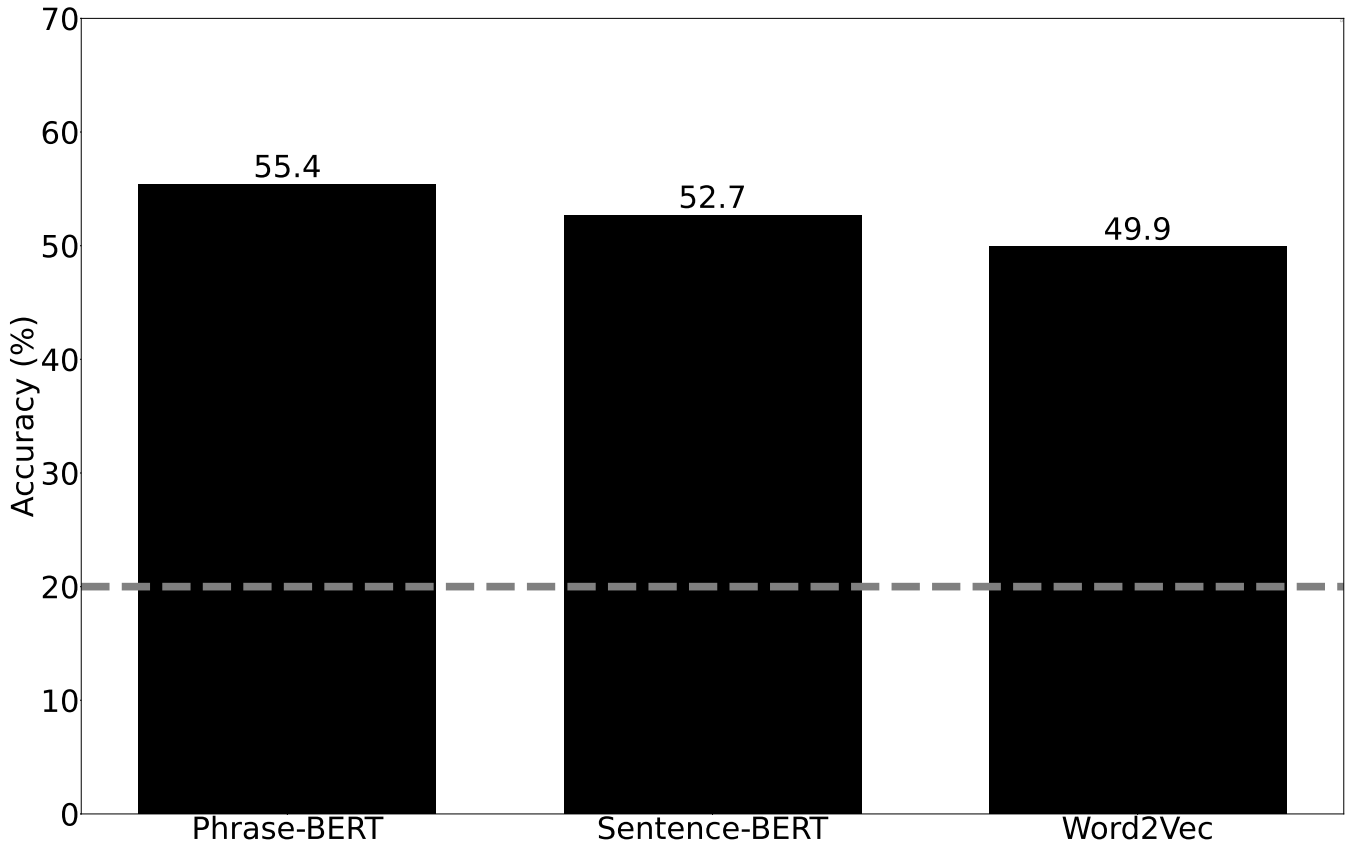

Figure S2. Predictive accuracy with randomly generated targets for different semantic embeddings. The dashed line indicates chance accuracy (20%).