ConversationAlign Lookup Database
Variable Key & References

ConversationAlign works by yoking values for many possible dimensions (specified by the user) to each content word in a language transcript. This process is accomplished by joining each word in the transcript with its corresponding values for specific variables of interest within a custom lookup database embedded in the software package. This custom lookup database was created by merging other published psycholinguistic databases and rescaling values where possible to a common 0 to 10 range using min/max scaling implemented by the 'scales' package of R (Gao et al., 2022). Steps for merging and rescaling the original data along with graphical depictions of the distributions of the recalled data can be found at:

https://reilly-lab.github.io/ConversationAlign_LookupDatabaseCreation.html

ConversationAlign's lookup database currently spans 30 variables with partial coverage of 102682 words (and word fragments) as summarized in Table 1. Not every word in the lookup database has corresponding values for every possible dimension. Missing values are populated with NAs (no imputation). The lookup database itself (lookup_db) is freely available for inspection and use at https://osf.io/z7p5a.

Table 1. Variables populating the ConversationAlign lookup database

Variable Name (common terminology)	Description	Range	N	Source
word	All words or word fragments with values in any of the external linguistic databases indexed	n/a	102682	n/a
aff_anger (anger) **aff_anxiety** (anxiety) **aff_boredom** (boredom) **aff_closeness** (closness) **aff_confusion** (confusion) **aff_doubt** (doubt) **aff_empathy** (empathy) **aff_encouragement** (encouragement) **aff_excitement** (excitement)	Values derived from Affectvec dimensions represent pairwise semantic similarity/distance from each target word in a language transcript to the specified anchor word from Affectvec (e.g., dog:empathy) using a vector-based embedding approach. Pairwise similarity values in the original Affectvec database reflect cosine distance. In ConversationAlign, we rescaled these values from -1:1 to a 0:10.	*0-10	76427	Affectvec[1]

Table 1. Variables populating the ConversationAlign lookup database

Variable Name (common terminology)	Description	Range	N	Source
aff_guilt (guilt)				
aff_happiness (happiness)				
aff_hope (hope)				
aff_hostility (hostility)				
aff_politeness (politeness)				
aff_sadness (sadness)				
aff_stress (stress)				
aff_surprise (surprise)				
aff_trust (trust)				
aff_dominance (dominance)	connotation of a target word with dominance (least dominant to most pleasant)	*0-10	19971	NRC VAD[2]
aff_valence (valence)	connotation of a target word with pleasantness (most aversive to most pleasant)	*0-10	19971	
lex_age_acquisition (age of acquisition)	Subjective adult estimates of the age at which when one acquired a word rescaled from 0-10	*0-10	31104	Kuperman[3]
lex_letter_count (word length in letters)	orthographic length of each word (letters per word)	RAW	102682	Base R character count
lex_morphemecount (morphemes per word)	Total morphemes-per-word	RAW	51531	SCOPE[4] Morpholex[5]
lex_prevalence (prevalence of word knowledge)	Relative proportion of people who know the meaning of a particular word rescaled from 0-10 where 0 is least known.	*0-10	46237	Keullers[6]
lex_senses_polysemy (number of definitions)	Number of different senses for a given target word.	*0-10	36408	Wordnet[7]
lex_wordfreqlg10_raw (US Word Frequency)	Log transformed word frequency values (per million words)	RAW	60384	Subtlex-US[8]

Table 1. Variables populating the ConversationAlign lookup database

Variable Name (common terminology)	Description	Range	N	Source
	derived from US English subtitles			
sem_arousal	Extent to which a word evokes a heightened state of autonomic arousal.	*0-10	19971	NRC VAD
sem_concreteness	Human crowd-sourced ratings reflecting extent to a word can be experienced through the senses (e.g., seen, heard, touched)	*0-10	39576	Brysbaert[9]
sem_diversity	How variable the contexts a word might appear in as an index of semantic ambiguity and	*0-10	29613	SCOPE/ Hoffman[10]
sem_neighbors	Number of semantic neighbors based on distance in co-occurence space	*0-10	45871	SCOPE/ Shaoul & Westbury[11]

Note: * in the range column denotes that the values have been rescaled from their original range. 1) Affectvec (Raji & da Melo, 2020); 2) National Research Council Valence Arousal Dominance (NRC-VAD) Lexicon (Mohammad, 2018); 3) Kuperman norms = Age of acquisition norms reported by Kuperman et al (2012); 4) SCOPE = South Carolina Psycholinguistic Metabase (Gao et al., 2022); 5) Morpholex (Sánchez-Gutiérrez et al., 2018). 6) Prevalence norms from Keullers et al (2015); 7) Polysemy ratings reflecting number of word senses per target word from Wordnet (Miller, 1995); 8) Subtlex-US word frequency norms Log10 transformed per million words (Brysbaert & New, 2009); 9) Concreteness norms from Brysbaert and colleagues (Brysbaert et al., 2014) ; 10) Semantic diversity derivation and methods see Hoffman and colleagues (2013); 11) Semantic neighborhood density is a metric of how many recurring neighbors/neighborhoods a word appears in (Shaoul & Westbury, 2010)

References

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/41/4/977 [pii] 10.3758/BRM.41.4.977 [doi]

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Gao, C., Shinkareva, S. V., & Desai, R. H. (2022). SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods*, *55*(6), 2853–2884. https://doi.org/10.3758/s13428-022-01934-0

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10/f5d59g

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*. https://journals.sagepub.com/doi/full/10.1080/17470218.2015.1022560

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mohammad, S. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174–184. https://doi.org/10.18653/v1/P18-1017

Raji, S., & da Melo, G. (2020). What sparks joy: The AffectVec emotion database. *Proceedings of the Web Conference, ACM*.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, *50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413.