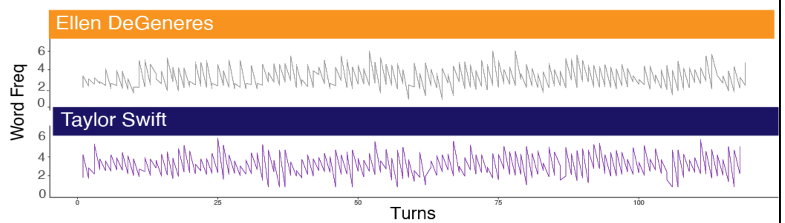
Quantifying Semantic and Affective Flow in Natural Language: New tools, New Applications

Jamie Reilly, PhD[1,2]
[1]Eleanor M. Saffran Center for Cognitive Neuroscience
[2]Department of Communication Sciences & Disorders
Temple University, Philadelphia Pennsylvania

Part 1: Dyad Alignment

```
ellen    Justin Timberlake is your favorite?
tswift   Yea.
ellen    Justin!
tswift  " That’s the best surprise ever, this
ellen    Yea.
ellen    Finish the statement  I am Taylor bl
tswift  " I think, just birth certificate wis
ellen   " Like I am Ellen blank, I am Ellen D
Portia de Rossi. I am Taylor Swift and I am d
tswift  " Nobody. That’s, that’s true though.'
```



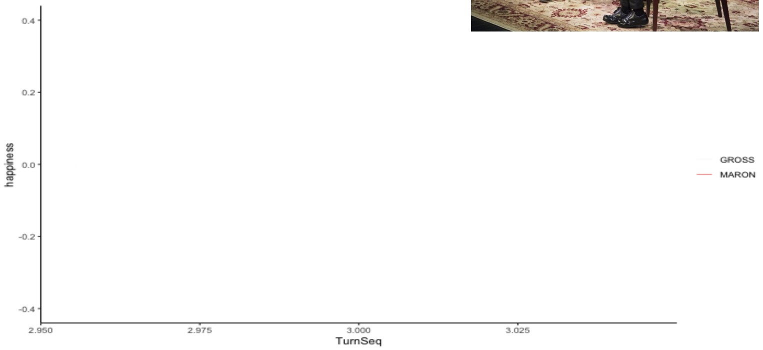
Dyadic Alignment

- Both interlocutors calibrate their own production to match each other (formality, complexity, affect)
  - Gricean Maxims (quantity, quality, relevance)

- Challenging / Demanding Exchanges
  - Intergenerational communication
  - Interdialectal and intercultural exchanges
  - Communicative disorders

ConversationAlign:

- Fresh Air (May 2015)
- Converting words to time series data



How it works…. Start with a transcript

```
ellen   " We have had a lot of fun over the years with our next guest, take a look."
ellen   " Your musical crush, someone in the business?"
tswift  " Oh, Justin Timberlake."
ellen    Justin Timberlake is your favorite?
tswift   Yea.
ellen    Justin!
tswift  " That's the best surprise ever, this is the best day ever!"
ellen    Yea.
ellen    Finish the statement  I am Taylor blank.
tswift  " I think, just birth certificate wise, it is Swift."
ellen   " Like I am Ellen blank, I am Ellen DeGeneres. I am Ellen DeGeneres and I am
and I am dating blank."
tswift  " Nobody. That's, that's true though."
ellen    I am Taylor Swift and my publicists told me to say blank.
tswift   My publicists told me not to answer any personal questions.
ellen    Now we're getting somewhere.
ellen    Kitty corner is a show for people who love cats.
tswift   I will answer your questions about cats.
ellen    Yea.
tswift   You can call us with your questions.
ellen    We can call it 'cat calls'.
tswift   Or 'look what the cat called in'.
tswift  " Every single I come on this show, it's really weird, really weird"
```

Text Cleaning

raw language transcript

P1	The cat is drinking the milk.
P2	I just love cats!

lowercase

P1	the cat is drinking the milk.
P2	i just love cats!

omit non-alphabetic

P1	the cat is drinking the milk
P2	i just love cats

Text Cleaning

stopwords

P1	cat drinking milk
P2	love cats

lemmatize

P1	cat drink milk
P2	love cat

squish

P1	cat drink milk
P2	love cat

Text Format: Vectorize

PID	Group	Turn#	Word
P1	young	1	cat
P1	young	1	drink
P1	young	1	milk
P2	old	2	cat
P2	old	2	love

Selects variable(s) for lookup db

Affective			
anger	anxiety	boredom	closeness
confusion	dominance	doubt	empathy
encouragement	excitement	guilt	happiness
hope	hostility	politeness	sadness
stress	surprise	trust	valence

Lexical-Semantic-Phonological	
age of acquisition	word length (n-letters)
n-senses (polysemy)	word frequency (lg10)
n-morphemes	semantic diversity
concreteness	arousal
prevalence	semantic neighbors

Text Formatting: Merge lookup database

PID	Grp	Turn #	Word	freq	hostility	metadata (neuropsy)
P1	yng	1	cat	?	?	?
P1	yng	1	drink	?	?	?
P1	yng	1	milk	?	?	?
P2	old	2	cat	?	?	?
P2	old	2	love	?	?	?

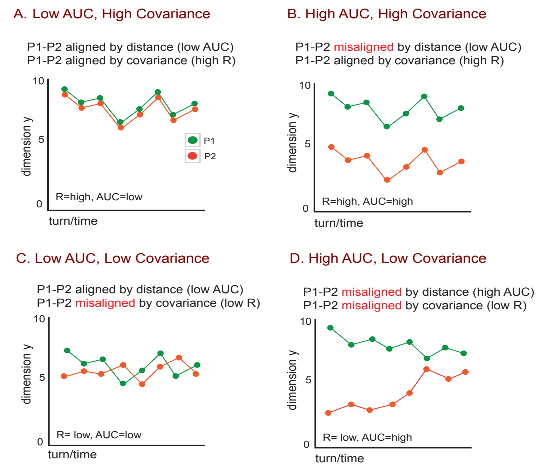
These transformations give us….

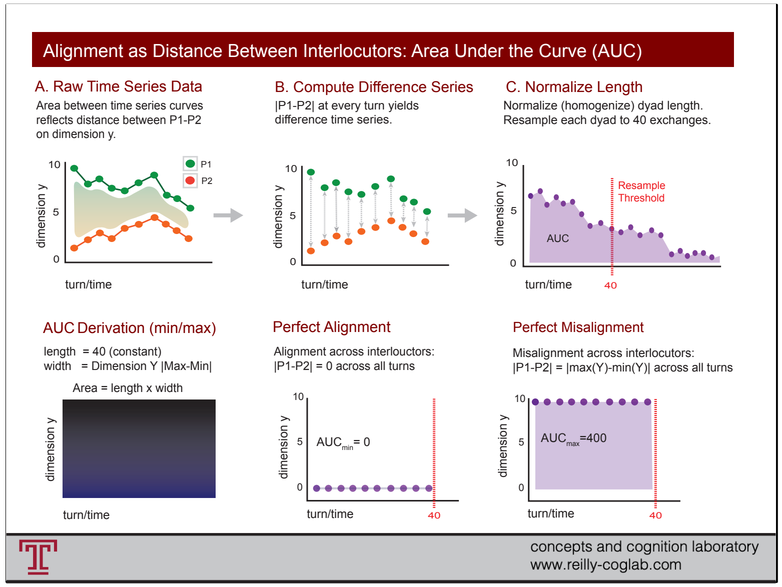
- Each dyad is transformed into two continuous simultaneous time series delineated by interlocutor and turn

- Treating language as time series data opens a whole world of causal modeling
  - Causal modeling, Granger Causality

- Analytical approaches
  - Nested data and linear mixed effects
    - word, dyad, person, group – etc.
  - Two different dimensions for computing alignment
    - Interlocutor covariance (do people move together?)
    - How distant are interlocutors on dimension X?

Alignment: Covariance vs Distance between Interlocutors

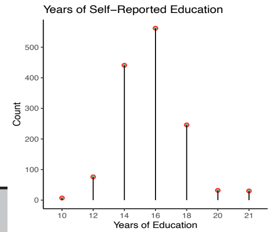


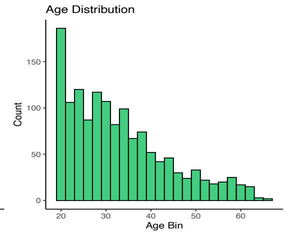

Alignment as Distance Between Interlocutors: Area Under the Curve (AUC)
A. Raw Time Series Data
Area between time series curves reflects distance between P1-P2 on dimension y.
P1
P2
dimension y
turn/time
B. Compute Difference Series
|P1-P2| at every turn yields difference time series.
C. Normalize Length
Normalize (homogenize) dyad length. Resample each dyad to 40 exchanges.
Resample Threshold
AUC
40
AUC Derivation (min/max)
length = 40 (constant)
width = Dimension Y |Max-Min|
Area = length x width
Perfect Alignment
Alignment across interlouctors: |P1-P2| = 0 across all turns
AUCmin = 0
Perfect Misalignment
Misalignment across interlocutors: |P1-P2| = |max(Y)-min(Y)| across all turns
AUCmax =400
concepts and cognition laboratory
www.reilly-coglab.com

ConversationAlign in Action

- Why is it so difficult to talk to older (and younger) people?
  - Older on younger: self-absorbed, ill-informed
  - Younger on older: repetitive, tangential, hyper-fixated

- We analyzed conversation transcripts (N=XX) from the Candor Corpus (Reece et al, 2023)
  - Unscripted conversations American English
  - Adult age range 18-66



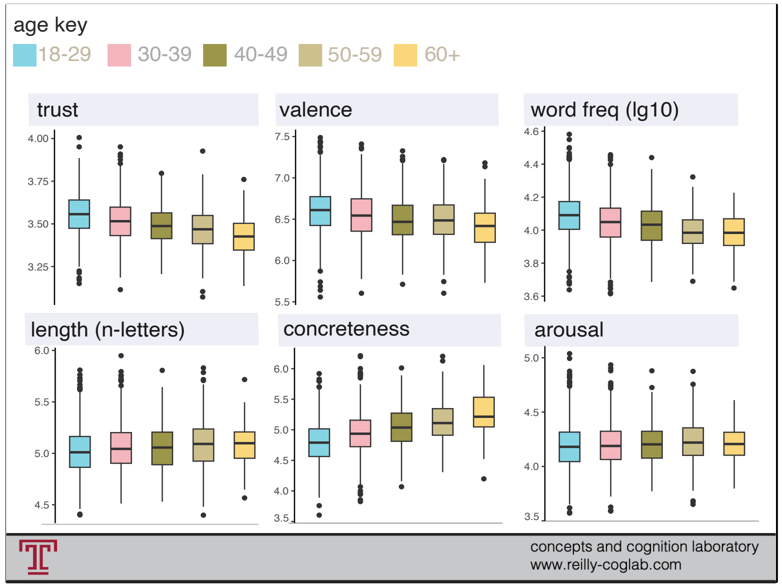


Methods (Using ConversationAlign)

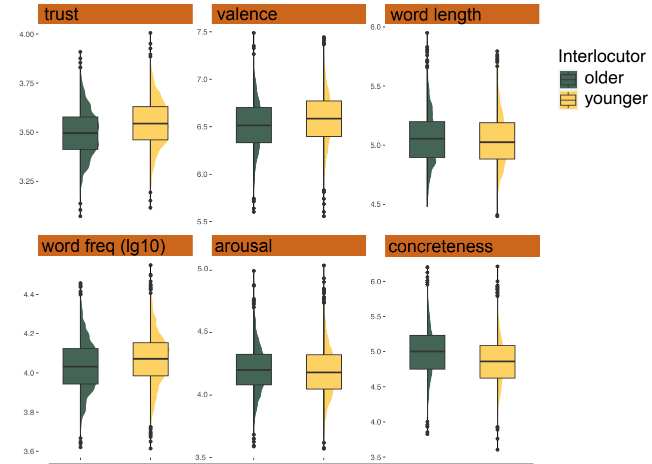
- Read data into R
- Retained dyads with ages (No NAs) and >40 exchanges
- Cleaned and formatted data
- Yoked lookup values to each word:
  - Concreteness, Word Frequency (lg10), Word Length (letters),
  - Trust, Valence, Arousal

code & results

https://reilly-lab.github.io/Candor_Stats_Analysis_v17.html


age key
18-29
30-39
40-49
50-59
60+
trust
valence
word freq (lg10)
length (n-letters)
concreteness
arousal
concepts and cognition laboratory
www.reilly-coglab.com

Differences between Older/Younger Interlocutors



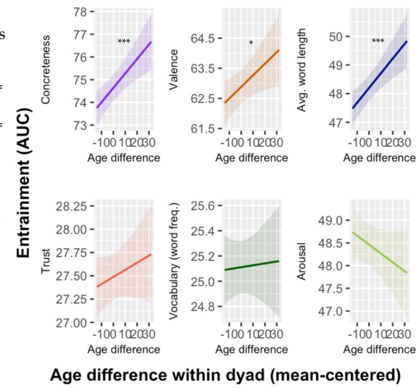
When CANDOR dyad members had a *greater* age difference between them, they tended to be *less* aligned on concreteness, valence, and their average word length.

In multilevel regression models accounting for covariates of sex, race, and education differences between dyad members in conversation, **age difference predicted entrainment in**…

- **Concreteness**, $b = 0.06$, $SE = 0.02$, $t(2320.08) = 3.65$, $p < .001$
- **Valence**, $b = 0.04$, $SE = 0.02$, $t(2301) = 2.27$, $p = .02$
- **Average word length**, $b = 0.05$, $SE = 0.01$, $t(262.75) = 3.82$, $p < .001$

Age difference did ***not*** predict entrainment in…

- **Trust**, $b = 0.008$, $SE = 0.003$, $t(2037) = 1.03$, $p = .304$
- **Vocabulary (word freq.)**, $b = 0.002$, $SE = 0.006$, $t(1743) = 0.244$, $p = .807$
- **Arousal**, $b = -0.02$, $SE = 0.01$, $t(1643.33) = -1.44$, $p = .149$



The Big Effect: AUC grows with age differences

Part II: Bigram Semantic Distance

concepts and cognition laboratory
www.reilly-coglab.com

Part 1: Bigram Semantic Distance

- Conceptual similarity (or dissimilarity) between two or more concepts within an n-dimensional semantic space

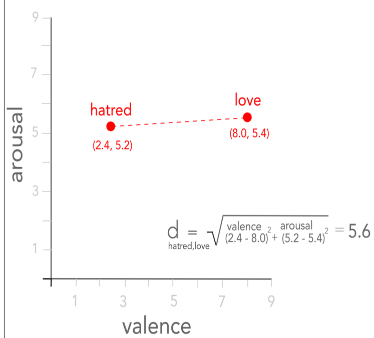
dog:cat

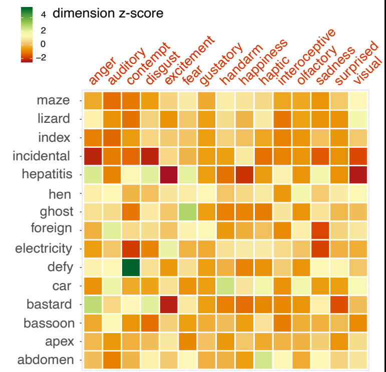
Taxonomic
Categorical
Feature-Based
Linnean

dog:leash

Thematic
Association
Contextual
Co-occurrence

Semantic spaces increasing in dimensionality





An Application of Semantic Distance to Narrative

The quick brown fox jumps over the lazy dog.

1. Clean, lemmatize, vectorize

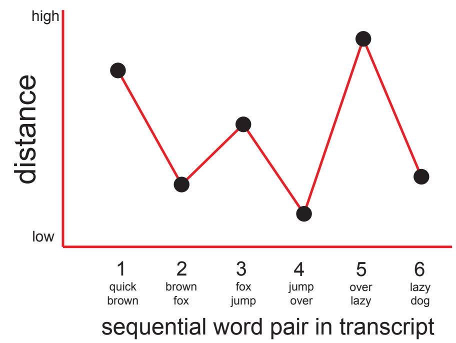
   quick brown fox jump over lazy dog

2. Compute distance for every sequential pair of words

word	word+1	$d_{(word,\ word+1)}$
quick	brown	3.5
brown	fox	5.2
fox	jump	1.8
jump	over	0.6
over	lazy	7.2
lazy	dog	1.3

A Novel Application of Semantic Distance to Narrative

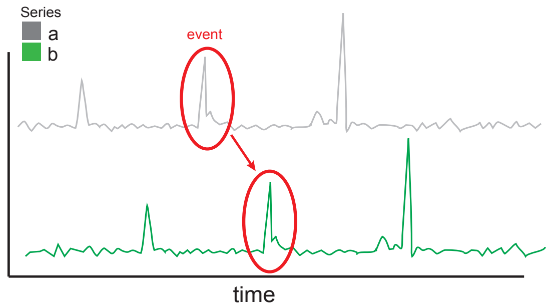
Plot distances as a time series



concepts and cognition laboratory
www.reilly-coglab.com

Advantages and applications of time series modeling

- Causal modeling
  - Analogous to modeling climate change
  - △ semantic distance cause △ in word frequency
  - △ semantic distance cause △ in pupil size, BOLD response, etc.



Advantages and applications of time series modeling

- Word-to-word semantic cohesion in discourse
  - Large shifts can signal topic changes
- Quantifying semantic development and/or decline
  - semantic acquisition (childhood)
  - semantic degradation (dementia)
  - semantic sophistication (skilled writers)
  - Semantic disorganization (psychosis)

Making it happen….

- 'semdistflow' R package

  https://github.com/Reilly-ConceptsCognitionLab/semdistflow

  – Free and open source takes any text as input
  – Cleans and formats text via hundreds of regular expressions
  – Outputs two metrics of semantic distance for every running pair of words
    - Embedding, Experiential

lemma_pair1	lemma	cosine.dist.sem	id	cosine.dist.glove	flipped_sem	flipped_glove
day	break	-0.494145934	2	0.281721607	1.494145934	0.718278393
break	cold	0.402349996	3	0.287982853	0.597650004	0.712017147
cold	grey	-0.595313892	4	0.151005454	1.595313892	0.848994546
grey	exceedingly	-0.213277662	5	0.101885343	1.213277662	0.898114657
exceedingly	cold	0.07948292	6	0.056358878	0.92051708	0.943641122
cold	grey	-0.595313892	7	0.151005454	1.595313892	0.848994546
grey	man	-0.205700992	8	0.295941045	1.205700992	0.704058955
man	turn	-0.160094173	9	0.30571489	1.160094173	0.69428511
turn	main	-0.118117782	10	0.25981348	1.118117782	0.74018652

Taxonomic Distance: Semdist15

- 15 feature dimensions, 70k+ words
- Features drawn from two sources:
  - Lancaster Sensorimotor Norms (Lynott et al, 2019) (n=8)
  - Affectvec (Raji et al, 2021) (n=7)
- Sensorimotor:
  - visual, auditory, gustatory, haptic, interoceptive, olfactory, hand-arm
- Affective
  - excitement, surprise, happiness, fear, anger, contempt, disgust, sadness

1	word	auditory	gustatory	haptic	interoceptive	olfactory	visual	handarm	excitement	surprised	happiness	fear	anger	contempt	disgust	sadness
2	abnormal	0.189	0.126	0.200	0.205	0.153	0.300	0.076	-0.240	0.008	-0.262	0.125	0.176	0.122	0.207	0.081
3	abnormality	0.142	0.068	0.184	0.153	0.079	0.305	0.111	-0.213	-0.064	-0.229	0.003	0.134	0.162	0.137	0.105
4	abnormally	0.106	0.044	0.111	0.239	0.044	0.256	0.118	-0.008	0.106	-0.040	-0.004	-0.019	-0.069	-0.075	-0.032
5	aboard	0.119	0.000	0.071	0.071	0.024	0.286	0.217	0.075	0.004	0.021	-0.024	-0.129	-0.042	-0.102	-0.048

Thematic Semantic Distance

- "A man is known by the company he keeps"
  ~ Aesop (620 – 564 BCE)

- Co-occurrence in language and the world (e.g., dogs, collars)
  - p(collar | dog): co-occur in event schemas
  - p(collar | dog): co-occur in linguistic contexts

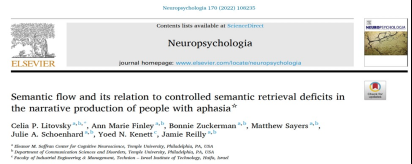
- Associative semantic networks scaffold inferences about word meaning through contextual *embeddings (e.g., funerals)*
  - Funerals are so sad.
  - The bereaved widow was crying at the funeral
  - The funeral was the worst day of my life.

  - sad, bereaved, crying, worst, widow

Thematic Semantic Knowledge

- People acquire meaning of unknown words by bootstrapping semantic properties of known associates
  - Latent semantic analysis
- A foundational assumption of NLP semantic approaches is:
  - Words that regularly appear in close proximity to each other are semantically related.
- GloVe Word Embedding Model (Pennington et al., 2014)
  - Trained on CoCA (Davies, 2009)
    - >900,000,000 words

bigram distance in aphasia


Neuropsychologia 170 (2022) 108235

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

ELSEVIER

Semantic flow and its relation to controlled semantic retrieval deficits in the narrative production of people with aphasia☆

Celia P. Litovsky[a,b,*], Ann Marie Finley[a,b], Bonnie Zuckerman[a,b], Matthew Sayers[a,b], Julie A. Schoenhard[a,b], Yoed N. Kenett[c], Jamie Reilly[a,b]

[a] Eleanor M. Saffran Center for Cognitive Neuroscience, Temple University, Philadelphia, PA, USA
[b] Department of Communication Sciences and Disorders, Temple University, Philadelphia, PA, USA
[c] Faculty of Industrial Engineering & Management, Technion – Israel Institute of Technology, Haifa, Israel

- Isolated all narratives from AphasiaBank
  - Aphasia (N=259), Controls (N=203)
  - Mean age 64.5 years, sd = 16.6
  - Cleaned/lemmatized
  - Segmented into bigrams, computed tax and thematic distances
- Derived a composite offline semantic measure
  - Boston Naming Test (Kaplan et al., 2001);
  - Verb Naming Test (Thompson, 2012);
  - WAB: Auditory Word Recognition Subtest
  - WAB: Sentence Completion subtest
  - WAB: Responsive Speech subtest

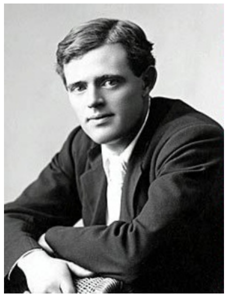
Results

- PWA showed lower:
  - thematic distance (0.723 ± 0.079) relative to controls
  - lower taxonomic distances (0.697 ± 0.097) relative to controls

- fixed effect of semantic impairment (as measured by the semantic composite score) significantly predicted participants' thematic ($p < .001$) and taxonomic ($p < .001$) semantic distances, indicating that better semantic ability in aphasia was associated with greater semantic distances.

A Demonstration Case:
To Build a Fire (Jack London, 1908)

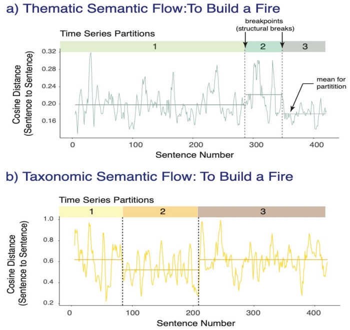
- Story…. 7125 words
- Imported into R and cleaned using 'semdistflow'



London, Jack (August 1908). "To Build a Fire". *The Century Magazine*. Vol. 76. pp. 525–534.

Semantic Drift across To Build a Fire

- Computed a semantic vector for the first ten words in the story
- Computed distance for every sentence relative to the first block
- Tested for stationarity
  - Rising or falling distance
- Tested for structural discontinuities
  - 'strucchange' package
  - Determines breakpoints



Other Applications:
Automated Semantic Fluency

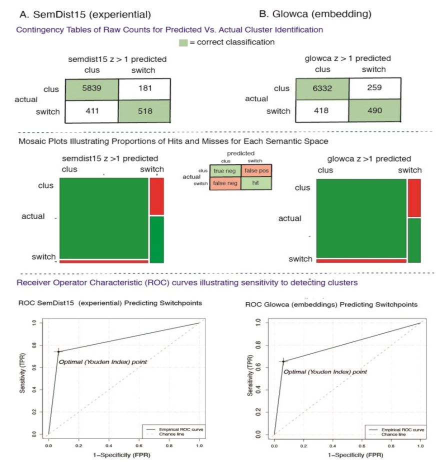
- Verbal fluency: Tell me as many animals as you can in 60s
  - Clusters and switches (ratio) is meaningful

- Scoring and segmenting can be laborious. Can we automate it using semantic distance?
  - Assumption: distance is shorter within a cluster than between
  - Large 'jumps' in semantic distance signal a switch
    - e.g., dog – cat –shark – dolphin – whale - tiger – elephant – gazelle

- We simulated a continuous stream of VF data alternating 10-word blocks of animals, musical instruments, fruits sampled with replacement from fixed lists
  - 7500 word vector, 749 switches occurring every 10 words

Other Applications: Automated Semantic Fluency

Accuracy: >90% both



Other Applications: Sentence Boundaries

- Does semantic distance jump across sentence boundaries?
  - Hypothesis is that yes it does

- We isolated bigram distance for within-sentence word pairs relative to bigrams that crossed a sentence boundary marked by punctuation….
  - Cats drink milk. Dogs like bones.
  - cat-drink, drink-milk, milk-dog, dog-bone

Results

- Pretty much as you would expect....

Table 6

Embedding Distance for Bigrams Within Sentences Versus Crossing Sentence Boundaries

	Token counts		Bigram distance		
Source	Words	Sentences	Within	Between	t-Statistic
Prisoner of Azkaban	104,860	8,936	0.62	0.65	$t(7{,}514.9) = 7.92$, $p < .001$***
Little Women	194,059	9,266	0.61	0.63	$t(8{,}498) = 3.55$, $p < .001$***
Sherlock Holmes	107,372	7,065	0.60	0.61	$t(7{,}335) = 3.01$, $p = .002$**
Portrait of Dorian Gray	82,012	6,687	0.60	0.59	$t(7{,}150) = 3.77$, $p \leq .001$***
Pride and Prejudice	124,719	6,210	0.61	0.62	$t(6{,}569.6) = 2.33$, $p = .02$*
Room with a View	69,931	5,948	0.62	0.63	$t(6{,}192.3) = 2.79$, $p = .01$*
Sorcerer's Stone	77,536	6,474	0.60	0.63	$t(5{,}874.7) = 5.46$, $p < .001$***
Become an Engineer	21,072	1,466	0.64	0.68	$t(1{,}694.6) = 6.92$, $p < .001$***
Honey Bees	91,577	3,182	0.72	0.75	$t(3{,}061.3) = 8.97$ $p < .001$***
Prehistoric Villages	33,283	1,541	0.67	0.72	$t(1{,}521) = 7.05$, $p < .001$***

Note. Token counts derived using the Quanteda package of R (Benoit et al., 2018). Distances reflect 0–2 cosine rescaled and reverse scored (0 is identical). Texts queried: Harry Potter and the Prisoner of Azkaban (J.K. Rowling, 1999); Little Women (Louisa May Alcott, 1868); The Adventures of Sherlock Holmes (Arthur Conan Doyle, 1892); The Portrait of Dorian Gray (Oscar Wilde, 1890); Pride and Prejudice (Jane Austen, 1813); A Room with a View (E.M. Forster, 1908); Harry Potter and the Sorcerer's Stone (J.K. Rowling, 1998); How to Become an Engineer (Frank W. Doughty, 2014); The Honey Bee: Its Natural History, Physiology, and Management (Edward Bevans, 1873); Prehistoric Villages, Castles, and Towers of Southwestern Colorado (Jesse Fewkes, 1919).
*$p < .05$. **$p < .01$. ***$p < .001$.

Acknowledgements

Extramural Support

NIH/NIDCD R01 DC013063

PA Commonwealth Brain Initiative

Postdocs

Celia Litovsky, PhD

Graduate Students

Ann Marie Finley
Lucia Pattullo

Collaborators

Jonathan Peelle, PhD
Nadine Martin, PhD
Murray Grossman, MD
Ingrid Olson, PhD
Tania Giovannetti, PhD
Dan Mirman, PhD

Research Scientists & Staff

Allie Kelly

Undergraduates

Sam Sprague
Sarah Johnson
Sara Hoover
Dewitt Fortenberry

concepts and cognition laboratory
www.reilly-coglab.com

References

- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x