

## Order and comparison asymmetry



Primacy effect: first mentioned target is more likely to capture the attention, is better remembered, is more likely to be perceived as the cause (e.g., Bettinsoli et al.; see also passive )



the partner possessing more stereotypically masculine traits is mentioned first (Hegarty et al. 2001).

Humanity -> central factor that gives rise to the different word orders found in the world's languages (Meier et al., 2017)



when men are presented as referent group (e.g., compared to males, females are ...), gender differences in status were perceived as larger and more legitimate (Bruckmüller et al. 2012)

# Semantic Networks: a definition

- WHAT graphical representations of knowledge based on meaningful relationships of written text, structured as a network of labeled nodes cognitively related to one another
- WHY GOAL: extract meanings
- HOW semantic networks connect words to words/hashtags/phrases, based on their co-occurrence
- WHO human and computerized methods, dealing with challenges such as co-reference resolution, synonym resolution, and ambiguity

## How good are the retrieved docs?



*Precision* : “purity” Fraction of retrieved docs that are relevant to the user’s information need (reject irrelevant)



*Recall* : “completeness” Fraction of relevant docs in collection that are retrieved (select relevant)

# CLEAN DATA

Pre-processing starts the text preparation into a more structured representation.



1) **Tokenization:** Tokenization is used to identify all words in a given text.

2) **Data Filtering:** People use a lot of casual language on twitter. To improve this and make words more similar to generic words, such sets of repeated letters are replaced by two occurrences.

haaaaappy -> haappy.



3) **Stop Word Removal:** Is used to eliminate that words that occurs frequently such as article, prepositions, conjunction and adverbs. These stop words depends on language of the text in questions. For example, words like the, and, before, while, and so on do not contribute to the sentiment.



4) **Stemming:** In information retrieval, stemming is the process of reducing a word to its root form.

walking, walker, walked -> walk

The process of "stemming" is removing these endings from words in a corpus. A "lemma" is a more linguistically informed version of a stem, such that "fight" is the lemma of "fought."

# Cleaning products

- Natural Language Toolkit (NLTK) in Python (Loper & Bird, 2002)
- Text Mining (tm) library in R (D. Meyer et al., 2008)

*Common Tokenization Implementations in Python and R Programming Languages*



	Description	Language/Library/function
Regular Expressions	Pattern matching algorithms that are fast and customizable.	R/tm/Regexp_Tokenizer <sup>4</sup>
CoreNLP Tokenizer	Rule-based, extensive support for many (human) languages.	Python/stanza <sup>5</sup>
Punctuation Tokenizers	Punctuation-sensitive, splits contractions	Python/nltk/word_tokenize <sup>6</sup>
Tokenizers for Social Media	Twitter and other media have specific conventions that need to be parsed by dedicated approaches	R/tokenizers/tokenize_tweets <sup>7</sup>
Model-specific tokenizers	Methods in NLP use tokenizers that automatically identify lemmas, contractions, and some parts of speech	Python/transformers/tokenizer <sup>8</sup>

**THE BEST= SPACY—  
later in the lab**

<sup>4</sup><https://rdrr.io/rforge/tm/man/tokenizer.html>

<sup>5</sup><https://github.com/stanfordnlp/stanza>

<sup>6</sup><https://www.nltk.org/api/nltk.tokenize.html>

<sup>7</sup><https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html>

<sup>8</sup><https://github.com/huggingface/tokenizers>

STEMMING □

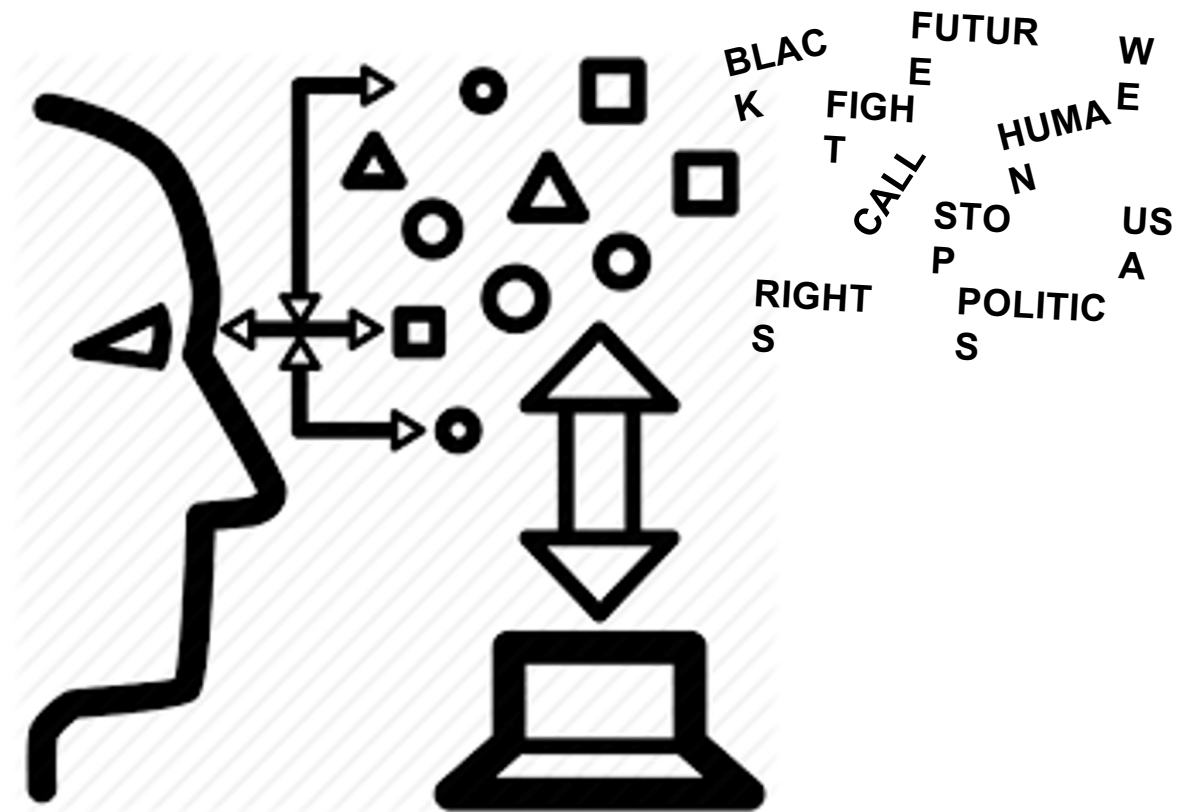
[https://www.rdocumentation.org/packages/corpus/versions/0.10.1/topics/stem\\_snowball](https://www.rdocumentation.org/packages/corpus/versions/0.10.1/topics/stem_snowball)

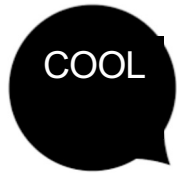
<https://www.kite.com/python/docs/nltk.SnowballStemmer>

WordNet Lemmatizer: <https://pythonprogramming.net/lemmatizing-nltk-tutorial/>

## PROCESS DATA

Dealing with textual data: from text to numbers





# Words or Hashtags



- Top down semantic/sentiment classification: bag of words
  - Bottom up semantic/sentiment classification: human coding
  - Meta-semantic classification: pronouns, nouns, verbs, adjectives
  - Meta-semantic structural properties: word order, dropping
  - Semantic & grammar: future/past/present tense
- topical signifier : shared conversation marker
  - can also represent the context of a tweet
  - flag an individual's community membership
  - indicate shared interests

# Dealing with textual data: from text to numbers



Theory  
Driven



Human Coding



Automatized  
Coding

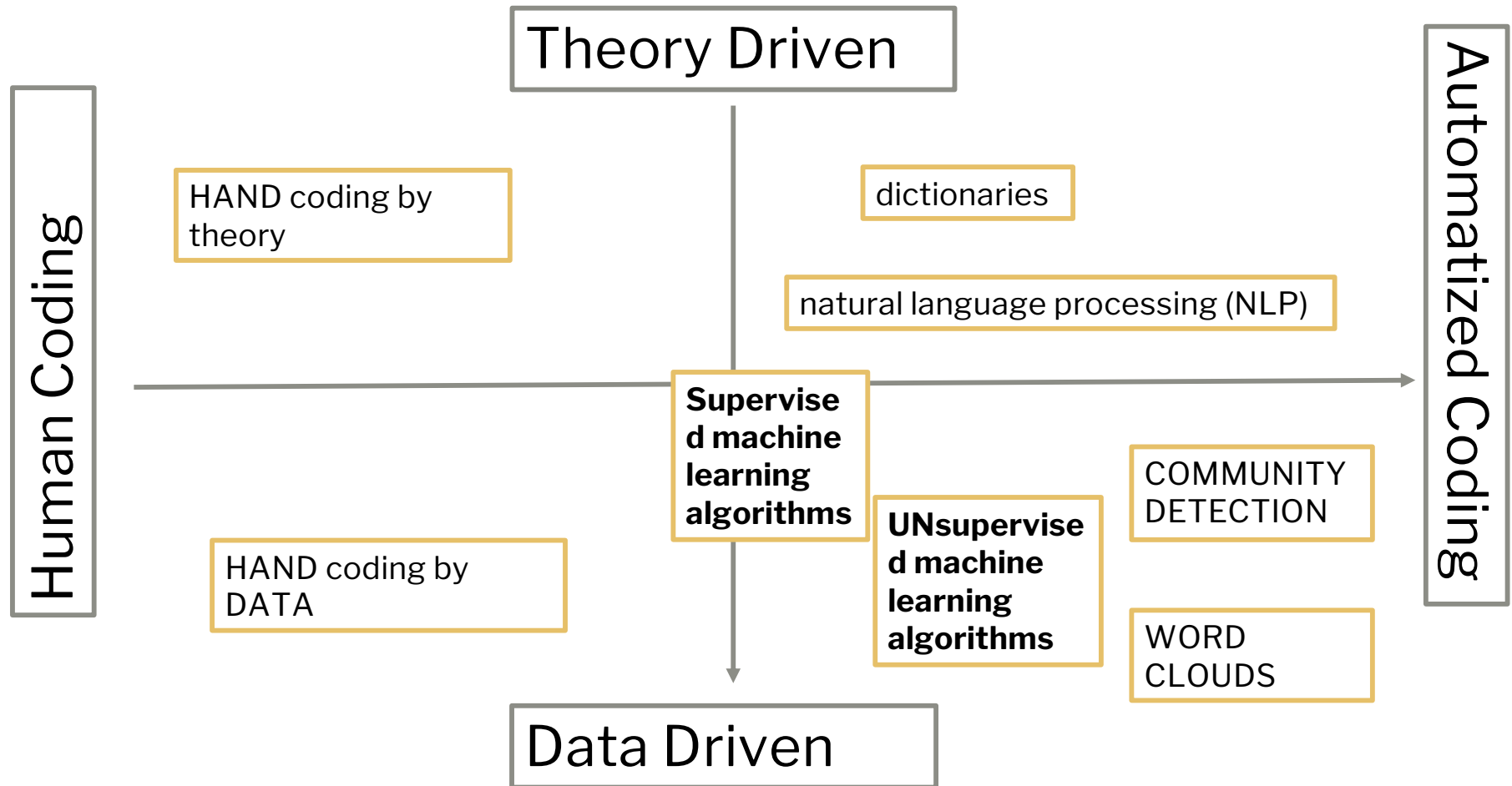


Data Driven



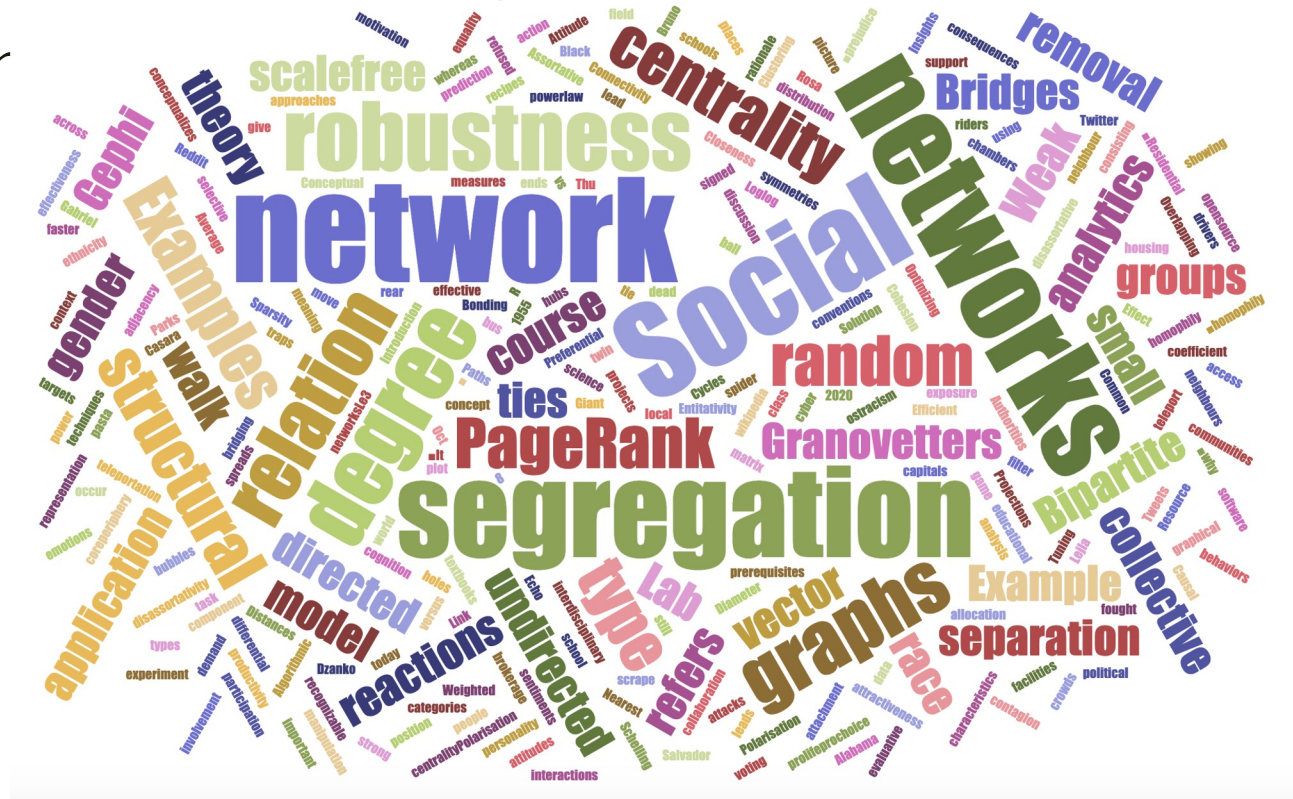


# Dealing with textual data: from text to numbers



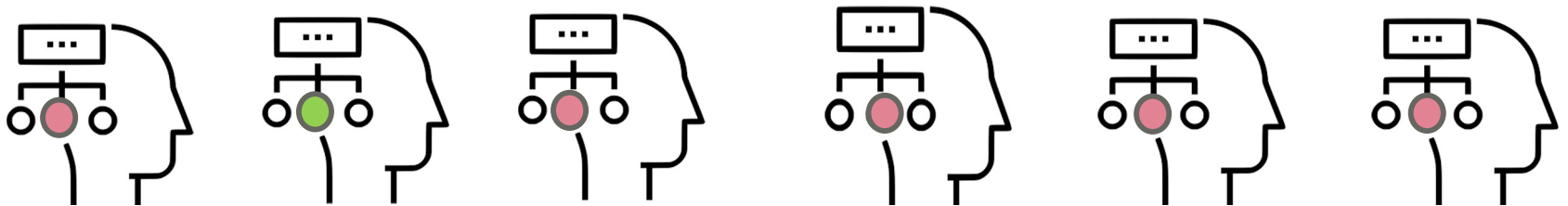
\_\_\_\_\_

- Based on word count
- Bigger words = more frequent
- bottom



# Human Coding

- *top down (coding by theory)*: initial coding scheme developed from the from pre-existing theory or assumptions
- *bottom up (grounded theory)*: initial coding scheme developed from the data
- *THE SUBJECTIVITY ISSUE: intercoder & intracoder reliability*
  - a classification procedure is reliable when it is consistent:  
Different people should code the same text in the same way



# Dictionaries

- A **sentiment analysis dictionary** contains information about the emotions or polarity expressed by words, phrases, or concepts. In practice, a **dictionary** usually provides one or more scores for each word. We can then use them to compute the overall **sentiment** of an input sentence based on individual words.
- top down
- Pro: transparency, objectiveness, replicability
- contro: limited amount of words, context not taken into account

# Dictionaries

- Descriptive dictionary: describe a target construct (extroversion: “gregarious,” “social,” and “approachable”)
  - Vs.
- Predictive dictionary: words that are usually used by extraverted individuals (e.g., “party,” “bar,” and “together”).
- create your own dictionary
  - Vs.
- Use a dictionary developed by other scientists
- LIWC, Bing (in R), WordNet (Miller, 1990)



■ STOP

# LIWC... *Psychometrics of Word Usage* <sup>\$109.74</sup>

---

## **The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods**

Journal of Language and Social Psychology  
29(1) 24–54  
© 2010 SAGE Publications  
DOI: 10.1177/0261927X09351676  
<http://jls.sagepub.com>  


**Yla R. Tausczik<sup>1</sup> and James W. Pennebaker<sup>1</sup>**

### **Abstract**

We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors. This article reviews several computerized text analysis methods and describes how Linguistic Inquiry and Word Count (LIWC) was created and validated. LIWC is a transparent text analysis program that counts words in psychologically meaningful categories. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.

[https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015\\_OperatorManual.pdf](https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf)

# LIWC

Summary Variable	Informal Speech	informal
Analytical Thinking	Swear words	swear
Clout	Netspeak	netspeak
Authentic	Assent	assent
Emotional Tone	Nonfluencies	nonfl
	Fillers	filler

With the exception of the summary variables and words per sentence, all LIWC2015 output variables are expressed as percentage of total words.

All Punctuation <sup>s</sup>	Allpunc
Periods	Period
Commas	Comma
Colons	Colon
Semicolons	SemiC
Question marks	QMark
Exclamation marks	Exclam
Dashes	Dash
Quotation marks	Quote
Apostrophes	Apostro
Parentheses (pairs)	Parenth
Other punctuation	OtherP

Language Metrics	
Words per sentence <sup>1</sup>	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb
Conjunctions	conj
Negations	negate

Grammar Other	
Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

Word count: people who is lying use more words!!!  
 Hancock, Curry, Goorha, and Woodworth (2008)  
 Extrovert people use more words (Pennebaker & King, 1999)



# LIWC

Summary Variable
Analytical Thinking
Clout
Authentic
Emotional Tone

Informal Speech	informal
Swear words	swear
Netspeak	netspeak
Assent	assent
Non-fl	nonfl
Fillers	fi

With the exception of the summary variable, all LIWC2015 output variables are expressed as a percentage of total words.

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., & van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations*, 24(4), 606-623.

Language Metrics	
Words per sentence <sup>1</sup>	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
	you
	shehe
	they
	ipron
	article
	prep
	auxverb
	adverb
	conj
Negations	negate

All Punctuation <sup>5</sup>	Allpunc
Periods	Period
Commas	Comma
Colons	Colon
Semicolons	SemiC
Question marks	QMark
Exclamation marks	Exclam
Dashes	Dash
Quotation marks	Quote
Apostrophes	Apostro
Parentheses (pairs)	Parenth
Other punctuation	OtherP

Grammar Other	
Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

# LIWC

Summary Variable  
Analytical Thinking  
Clout  
Authentic  
Emotional Tone

Informal Speech  
Swear words  
Netspeak  
Assent  
Nonfluencies  
Fillers

informal  
swear  
netspeak  
assent  
nonfl  
filler

With the exception of the summary variables, per  
sentence, a  
percentage

All Punctuation<sup>s</sup>  
Periods  
Commas  
Colons  
Semicolons  
Question marks  
Exclamation marks  
Dashes  
Quotation marks  
Apostrophes  
Parentheses (pairs)  
Other punctuation

People who are experiencing physical or emotional pain tend to have their attention drawn to themselves and subsequently use more first-person singular pronouns (e.g., Rude, Gortner, & Pennebaker, 2004).  
When people sit in front of a mirror and complete a questionnaire, they use more words such as “I” and “me” than when the mirror is not present (Davis & Brock, 1975)

Language Metrics  
Words per sentence<sup>1</sup>  
Words>6 letters  
Dictionary words  
Function Words  
Total pronouns  
Personal pronouns  
1st pers singular  
1st pers plural  
2nd person  
3rd pers singular  
3rd pers plural  
Impersonal pronouns  
Articles  
Prepositions  
Auxiliary verbs  
Common adverbs  
Conjunctions  
Negations

WPS  
Sixltr  
Dic  
function  
pronoun  
ppron  
i  
we  
you  
shehe  
they  
ipron  
article  
prep  
auxverb  
adverb  
conj  
negate

Grammar Other  
Regular verbs  
Adjectives  
Comparatives  
Interrogatives  
Numbers  
Quantifiers

verb  
adj  
compare  
interrog  
number  
quant

# LIWC

Summary Variable
Analytical Thinking
Clout
Authentic
Emotional Tone

Informal Speech	informal
Swear words	swear
Netspeak	netspeak
Assent	assent
Nonfluencies	nonfl
Fillers	filler

All Punctuation <sup>5</sup>
Periods
Commas
Colons
Semicolons
Question marks
Exclamation marks
Dashes
Quotation marks
Apostrophes
Parentheses (pairs)
Other punctuation

With the exception of the summary variables a per sentence, percentage

STATUS Across five studies in which status was either experimentally manipulated, determined by partner ratings, or based on existing titles, increased use of first-person plural was a good predictor of higher status, and in four of the studies increased use of first-person singular was a good predictor of lower status (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2009)

Language Metrics	
Words per sentence <sup>1</sup>	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb
Conjunctions	conj
Negations	negate

Grammar Other	
Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

# LIWC

Summary Variable  
Analytical Thinking  
Clout  
Authentic  
Emotional Tone

Informal Speech  
Swear words  
Netspeak  
Assent  
Nonfluencies  
Fillers

informal  
swear  
netspeak  
assent  
nonfl  
filler

With the exception of the summary variables and words per sentence, LIWC scores are calculated as a percentage of the total word count.

All Punctuation<sup>5</sup>  
Periods  
Commas  
Colons  
Semicolons  
Question marks  
Exclamation marks  
Dashes  
Quotation marks  
Apostrophes  
Parentheses (pairs)  
Other punctuation

## relationship quality

first-person plural (“we”) has not been found to be related to higher relationship quality, instead use of second person (“you”) is more important in predicting lower-quality relationships.  
Simmons, Chambless, and Gordon (2008)

Language Metrics  
Words per sentence<sup>1</sup>  
Words>6 letters  
Dictionary words  
Function Words  
Total pronouns  
Personal pronouns  
1st pers singular  
1st pers plural  
2nd person  
3rd pers singular  
3rd pers plural  
Impersonal pronouns  
Articles  
Prepositions  
Auxiliary verbs  
Common adverbs  
Conjunctions  
Negations

WPS  
Sixltr  
Dic  
function  
pronoun  
ppron  
i  
we  
you  
shehe  
they  
ipron  
article  
prep  
auxverb  
adverb  
conj  
negate

Grammar Other  
Regular verbs  
Adjectives  
Comparatives  
Interrogatives  
Numbers  
Quantifiers

verb  
adj  
compare  
interrog  
number  
quant

# LIWC

Summary Variable
Analytical Thinking
Clout
Authentic
Emotional Tone

Informal Speech	informal
Swear words	swear
Netspeak	netspeak
Assent	assent
Nonfluencies	nonfl
Fillers	filler

With the exception of the summary variables and words per sentence, LIWC2015 calculates 150 word-level variables as percentages.

All Punctuation <sup>5</sup>
Periods
Commas
Colons
Semicolons
Question marks
Exclamation marks
Dashes
Quotation marks
Apostrophes
Parentheses (pairs)
Other punctuation

**COHERENCE**  
Conjunctions (e.g., and, also, although) join multiple thoughts together and are important for creating a coherent narrative (Graesser, McNamara, Louwerse, & Cai, 2004).

Language Metrics	
Words per sentence <sup>1</sup>	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb
Conjunctions	conj
Negations	negate

Grammar Other	
Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

# LIWC

People experiencing physical or emotional pain tend to use more first-person singular pronouns (Rude, Gortner, & Pennebaker, 2004).

Emotion Depressed patients are more likely to use more first-person singular and more negative emotion words than participants who have never been depressed in emotional writings (Rude et al., 2004) or negative emotionality more broadly (Tackman et al., 2019) — and susceptibility to suicide (Stirman & Pennebaker, 2001)

When people sit in front of a mirror use more words such as “I” and “me” than when the mirror is not present (Davis & Brock, 1975)

Semicolons	SemiC
Question marks	QMark
Exclamation marks	Exclam
Dashes	Dash
Quotation marks	Quote

Language Metrics	
Words per sentence <sup>1</sup>	WPS
Words > 6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb
Conjunctions	conj

Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

“we” can signal a sense of group identity, such as when couples are asked to evaluate their marriages to an interviewer, the more the participants use “we,” the better their marriage (Simmons, Gordon, & Chambless, 2005)

[www.secretlifeofpronouns.com](http://www.secretlifeofpronouns.com)

# Psycho-social index

Social Words	social
Family	family
Friends	friend
Female referents	female
Male referents	male

Positive political ads used more present and future tense verbs, and negative ads used more past tense verbs (Gunsch et al., 2000). From the tense of the verbs and the personal pronouns used, we can infer that negative ads focus on past actions of the opponent, and positive ads focus on the present and future acts of the candidate.

<b>Core Drives and Needs</b>	drives
Affiliation	affiliation
Achievement	achieve
Power	power
Reward focus	reward
Risk/prevention focus	risk
<b>Time Orientation<sup>a</sup></b>	
Past focus	focuspast
Present focus	focuspresent
Future focus	focusfuture
Relativity	relativ
Motion	motion
Space	space
Time	time

Affect Words	affect
Positive emotion	posemo
Negative emotion	negemo
Anxiety	anx
Anger	anger
Sadness	sad

<b>Personal Concerns</b>	
Work	work
Leisure	leisure
Home	home
Money	money
Religion	relig
Death	death

# Psycho-social index

Affect Words	affect
Positive emotion	posemo
Negative emotion	negemo
Anxiety	anx
Anger	anger
Sadness	sad

Social Words	social
Family	family
Friends	
Female	
Male re	

Depressed and suicidal individuals are more self-focused, express more negative emotion and sometime use more death-related words.

. Depressed patients are more likely to use more first-person singular and more negative emotion words than participants who have never been depressed in emotional writings (Rude et al., 2004

Core Drives and	drives
Affiliation	affiliation
Achievement	achieve
Power	power
Focus	reward
Prevention focus	risk
Time Orientation	
Past focus	focuspast
Present focus	focuspresent
Future focus	focusfuture
Relativity	relativ
Motion	motion
Space	space
Time	time

Personal Concerns	
Work	work
Leisure	leisure
Home	home
Money	money
Religion	relig
Death	death



# Psycho-social index

Social Words	social
Family	family
Friends	friend
Female referents	female
Male referents	male

<b>Core Drives and Needs</b>	drives
Affiliation	affiliation
Achievement	achieve
Power	power
Reward focus	reward
Risk/prevention focus	risk
<b>Time Orientation<sup>a</sup></b>	
Past focus	focuspast
Present focus	focuspresent
Future focus	focusfuture
Relativity	relativ
Motion	motion
Space	space
Time	time

Affect Words	affect
Positive emotion	posemo
Negative emotion	negemo
Anxiety	anx
Anger	anger
Sadness	sad

Pasupathi, 2007

Participants were asked to either recall an event that they had discussed with someone else, or an undisclosed event past tense in discussing a disclosed event and greater present tense in discussing an undisclosed event.

<b>Personal Concerns</b>	
Work	work
Leisure	leisure
Home	home
Money	money
Religion	relig
Death	death

# Cognition & perception

LANGUAGE AMBIGUITY (insight, tentat, Roos et al.'s (2020) is related to dogmatism (Fast & Horvitz, 2016) and politeness (Li et al., 2020).

Cognitive Processes <sup>2</sup>	cogproc
Insight	insight
Cause	cause
Discrepancies	discrep
Tentativeness	tentat
Certainty	certain
Differentiation <sup>3</sup>	differ
Perpetual Processes	percept
Seeing	see
Hearing	hear
Feeling	feel
Biological Processes	bio
Body	body
Health/illness	health
Sexuality	sexual
Ingesting	ingest

Cognitive Processes <sup>2</sup>	cogproc
Insight	insight
Cause	cause
Discrepancies	discrep
Tentativeness	tentat
Certainty	certain
Differentiation <sup>3</sup>	differ

Prepositions (e.g., to, with, above), cognitive mechanisms (e.g., cause, know, ought), and words greater than six letters are all also indicative of more complex language.

Cognitive complexity can be thought of as a richness of two components of reasoning: the extent to which someone differentiates between multiple competing solutions and the extent to which someone integrates among solutions (Tetlock 1981)



# Incivility score in LIWC

- Addition of Swear, Anger, and Negative Emotions (based on previous research, see Ksiazek et al., 2015; Stoll et al., 2020)

# Various dictionaries....

- cognitive processes (Pennebaker et al., 2015),
- moral values (Graham et al., 2009),
- psychological motivations (Stone et al., 1966),
- well-being (Ratner et al., 2019),
- regulatory focus (Kanze et al., 2019),
- markers of suicidal ideation (Thomas & Duszynski, 1985)
- brand personality (Opoku et al., 2008).
- Concreteness (Brysbaert 2014)

# Sentiment /emotion tools

- vader\_df function of the VADER package (version 0.2.1, Roehrick, 2020). VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. Sensitive to negation. Sentence level analyses <https://github.com/cjhutto/vaderSentiment>
- Lexicoder Sentiment Dictionary which was developed to capture the emotional tone of political communications. Young and Soroka (2012)
- integrative complexity Conway, Conway, & Houck, 2020
- EmoLex, ANEW, SentiWordNet are designed to analyze larger sets of emotional categories
- General Inquirer (GI) human curated dictionary that operates over a broader set of topics (e.g., power, weakness)
- Empath allows researchers to perform text analyses over a broader set of topical and emotional categories than existing tools, and also to create and validate new categories on demand  
(PDF) *Empath: Understanding Topic Signals in Large-Scale Text*. Available from: [https://www.researchgate.net/publication/301872654\\_Empath\\_Understanding\\_Topic\\_Signals\\_in\\_Large-Scale\\_Text](https://www.researchgate.net/publication/301872654_Empath_Understanding_Topic_Signals_in_Large-Scale_Text) [accessed Nov 08 2023]. deceptive reviews convey stronger sentiment across both positively and negatively charged categories, and tend towards exaggerated language

# WORD Net & creativity

- Word Association nets: <https://wordassociations.net/en>
- WordNet, a comprehensive lexical database of English words, provides a list of the cognitive synonyms associated with any concept and information regarding the links between the various words (Miller, 1998).
- Can be used to expand a dictionary
- Using such semantic networks, it is possible to approximate the **semantic distance between any two words or sentences** (although such distance measures are subject to the accuracy and completeness of WordNet itself)
- One set of studies measured creativity from the semantic distance of participant-generated ideas from a common prompt initially given to all participants (M. L. Meyer et al., 2019). Ideas that diverged more semantically from the given prompt were scored as more original and creative.

# Natural language processing (NLP)

**Natural language processing (NLP)** is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

- tokenization
- grammatical role POS (part of speech) tagging (sbj, obj..)
- stemming
- thesauri
- shallow parsing : identifies constituent parts of sentences (nouns, verbs, adjectives, etc.)
  - the hand-coding of a set of rules, coupled with a dictionary lookup

# Machine learning



**Supervised machine learning algorithms** apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

**Unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.



# Content Analysis

- Detect systematic patterns in communication
  - -> *topic identification*



opinions

refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify

# Sentiment Analysis


- extract, quantify, and study **affective states** and subjective information



attitudes

# Topic modeling /community detection... interpretation

- “extrinsic evaluation” relies on outside sources (such as demographic variables, psychological outcomes, or annotated labels) of information to decide which topics are relevant (H. A. Schwartz et al., 2013)
- The **word intrusion task** gives testers six randomly ordered words, five of which are from a given topic. If the five words are coherent, then the intruder is easily identified Chang et al., (2009).
- **topic intrusion task** measures the model’s overall ability to describe the corpus. Given a document title, a snippet of the document, and four topics (represented by the top 8 words in the topic), the tester must identify which of the four topics was low probability for the document, versus the three high probability topics for the document.
- Or... a glance by a human over the topics! ;-)



## “Open Vocabulary” approach to language analysis (H. A. Schwartz et al., 2013)

patterns in language are discovered in a bottom-up manner

- correlated with dimensions of personality (Park et al., 2015),
- symptoms of depression (Guntuku et al., 2017)
- markers of schizophrenia (Mitchell et al., 2015),
- geographic distribution of variables such as “well-being” (H. A. Schwartz et al., 2013).

# ANALYSE DATA

- -> frequency
- -> correlations/regressions/mediations
- -> source comparison (t-test, Anova)
- -> networks: centrality measures, community detection etc

- Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In S. Hai-Jew (Ed.), *Data Analytics in Digital Humanities* (pp. 161–189). Springer International Publishing. [https://doi.org/10.1007/978-3-319-54499-1\\_7](https://doi.org/10.1007/978-3-319-54499-1_7)
- Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>

Article

*Group Processes &  
Intergroup Relations* | G  
P  
I  
R

# The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter

*Group Processes & Intergroup Relations*

2021, Vol. 24(4) 606–623

© The Author(s) 2021





Article reuse guidelines:


[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/1368430220987596

[journals.sagepub.com/home/gpi](https://journals.sagepub.com/home/gpi)



Amos Fong,<sup>1</sup> Jon Roozenbeek,<sup>1</sup> Danielle Goldwert,<sup>2</sup>   
Steven Rathje<sup>1</sup> and Sander van der Linden<sup>1</sup> 

- 
- Social media has transformed information dissemination, often leading to echo chambers.
  - **Echo Chambers:** Polarized online communities that foster conspiracy ideation.
  - **Objective:** Analyze language use on Twitter by conspiracy theorists and science advocates to identify psycholinguistic patterns.

# Ingroup and Outgroup Dynamics

- **Social Identity Theory:** This theory explains that people identify strongly with groups that share their beliefs, which helps to build a positive self-concept. For conspiracy theorists, the ingroup includes others who believe in similar conspiracies, and the outgroup consists of the "establishment" (scientists, governments, or other perceived authority figures).
- **Us vs. Them Mentality:** Conspiracy theorists often portray the outgroup as a dangerous, deceitful enemy that needs to be exposed or resisted. This antagonistic framing strengthens group solidarity and commitment to shared beliefs.





# Psychological Needs Fulfilled by Conspiracies

- **Uncertainty and Need for Closure:** Many people turn to conspiracies as a way to simplify complex issues, especially in times of uncertainty. The need for cognitive closure—an individual's preference for certainty and definitive answers—can make simple, absolute explanations (like conspiracies) more appealing.
- **Compensatory Control Theory:** This theory suggests that when people feel a lack of control, they may seek out patterns or explanations in chaotic situations, which conspiracy theories often provide. For example, attributing complex events (e.g., pandemics) to a deliberate plan by powerful elites restores a sense of predictability and control.



# Negative Emotions and Motivated Reasoning

- **Role of Negative Emotions:** Conspiracy beliefs are strongly associated with negative emotions, particularly anger and anxiety. These emotions are often directed at the perceived threats posed by the outgroup, whether it's scientists or government officials.
- **Motivated Reasoning:** Conspiratorial thinking is often reinforced by motivated reasoning, where people selectively interpret information in a way that supports their pre-existing beliefs. This helps maintain the consistency of their worldview, particularly when it aligns with strong emotions like anger or fear.



# Research Goals and hypotheses

- Compare linguistic patterns in tweets by conspiracy influencers, science influencers, and their followers.
- Hypothesized psychological themes in conspiratorial speech: negativity, anger, ingroup/outgroup language, themes of power, religion, and death.

# Methodology

- **Data Collection:** 16,290 tweets from influencers and 160,949 tweets from followers over 5 days July 21, 2017 to July 25, 2017 posted by conspiracy theorists and science communicators who are active on Twitter (objective indicators of popularity (i.e., the highest number of followers) + random sample of followers
- **Analysis Tool:** Linguistic Inquiry and Word Count (LIWC) software to measure categories like negative emotion, cognitive processes, and thematic language.
- **Groups:** Conspiracy theorists vs. Science advocates.

# Key Findings: from liwc

## ■ Negative Emotions

- **Higher Negative Emotion:** Conspiracy tweets showed more anger and anxiety.
- **Statistics:** Conspiracy influencers scored significantly higher than science influencers in expressions of anger and anxiety.

## ■ Cognitive Processes

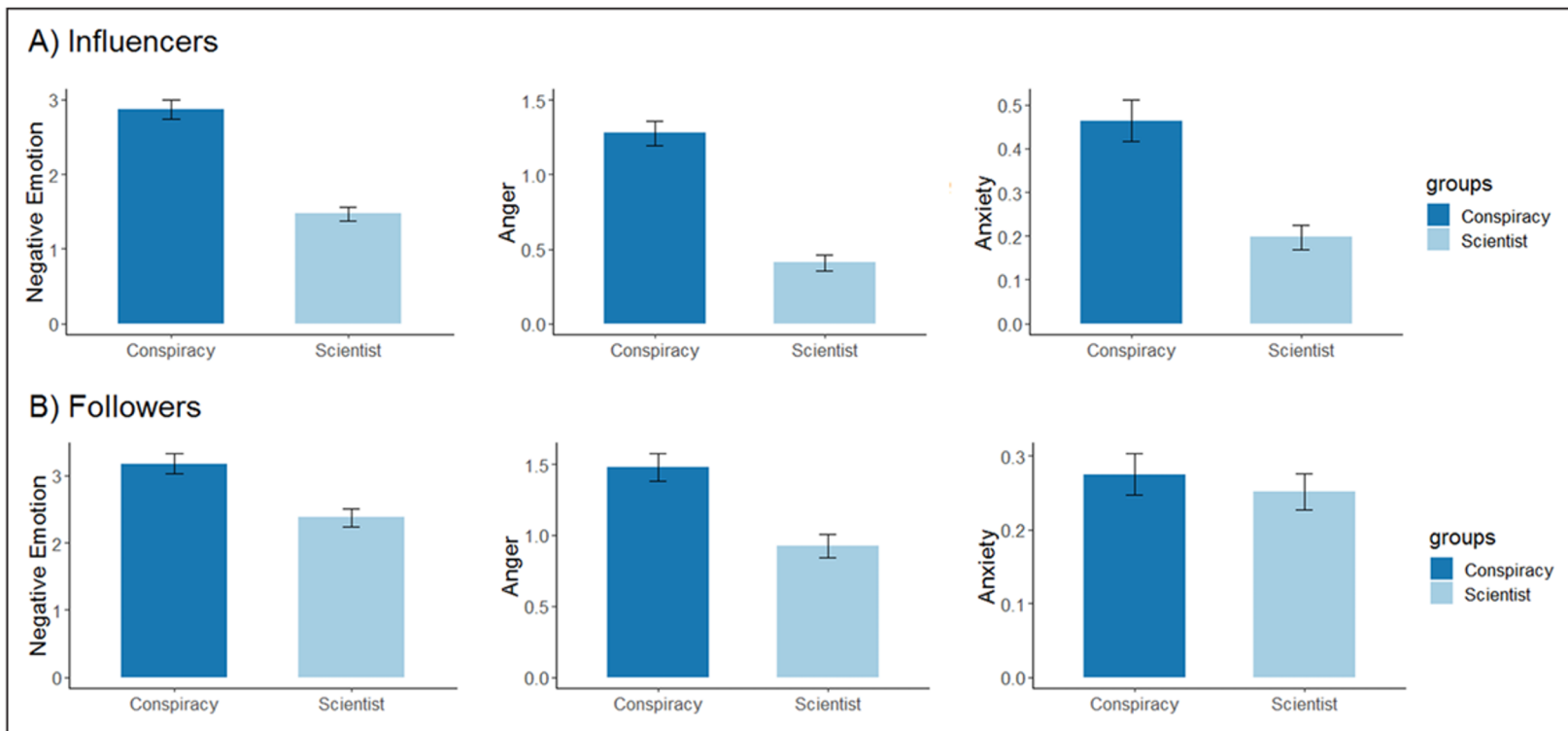
- Conspiracy followers displayed a higher need for certainty and past orientation compared to science followers.
- **Unexpected:** Conspiracy influencers showed less certainty in their language compared to science influencers.

## ■ Ingroup vs. Outgroup Language

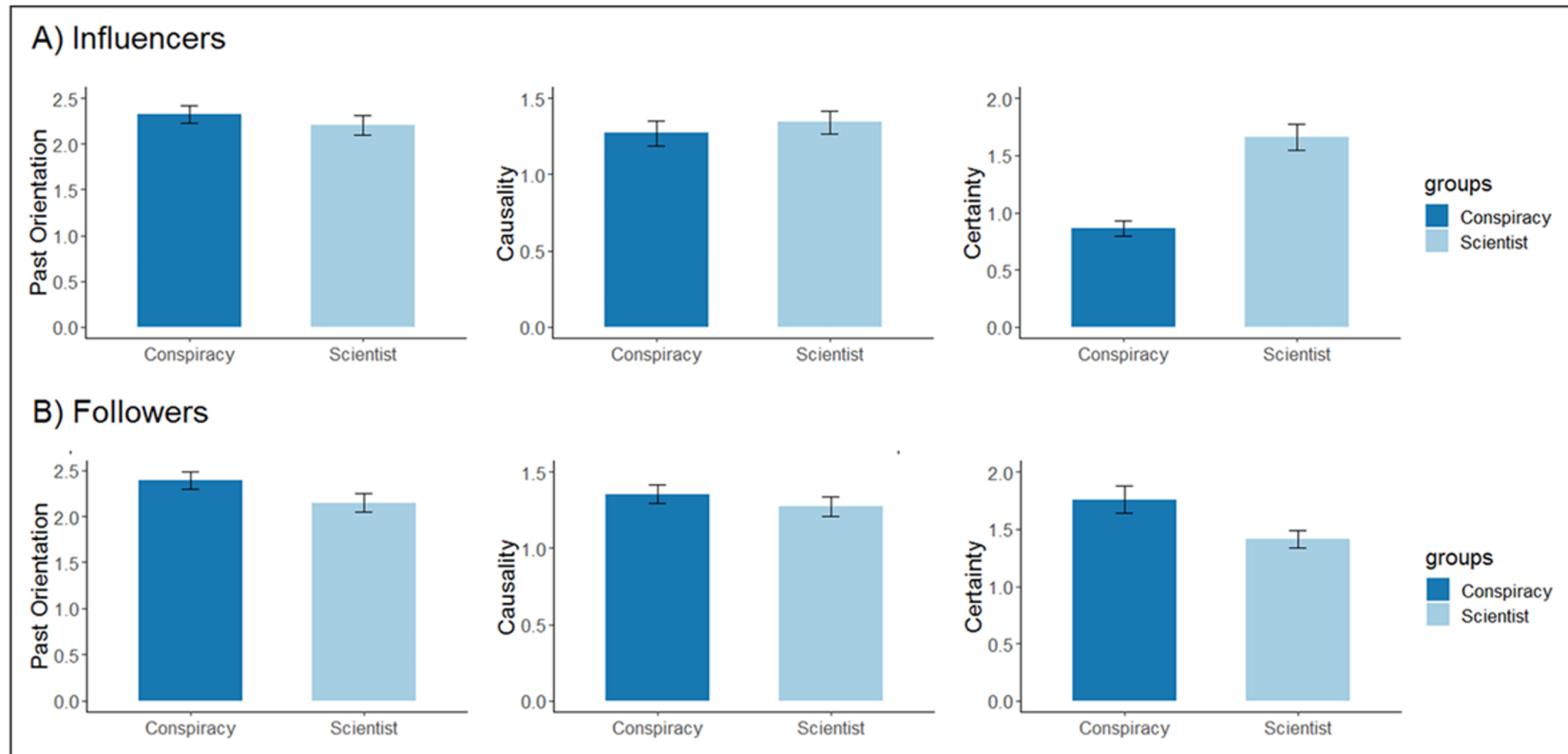
- **Ingroup Focus:** Conspiracy language reflects an “us vs. them” mentality.
- **Follower Comparison:** Conspiracy followers used more outgroup language than science followers.

## ■ Themes of Power, Death, and Religion

- **Power:** Conspiracy tweets frequently mentioned government, military, and elite figures.
- **Death:** Words like “war” and “killed” were more prevalent.
- **Religion:** Conspiracy language often referenced religious terms, linking to beliefs of a “New World Order.”



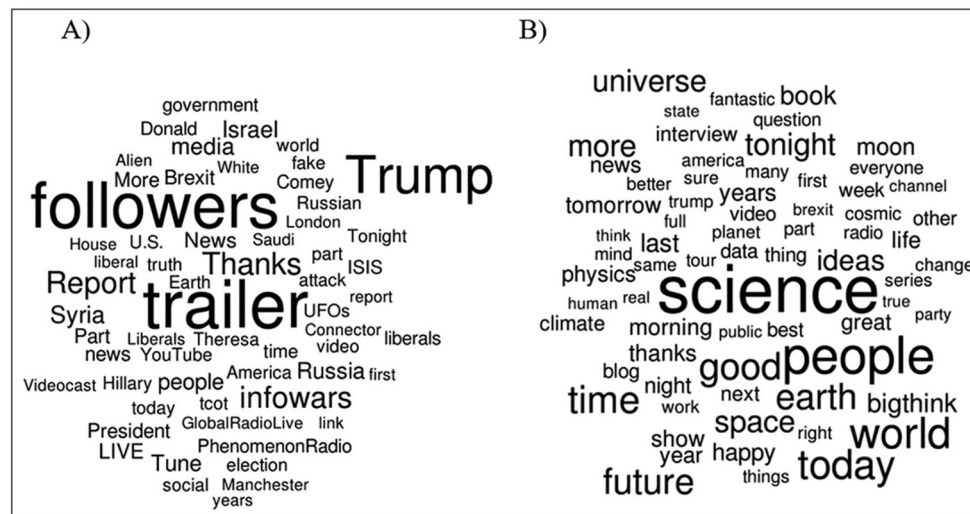
Note: Error bars represent 95% confidence intervals



*Note.* Error bars represent 95% confidence intervals.

# Key Findings

- Visual representation of follower networks showed clear separation between science and conspiracy followers, reinforcing echo chamber effects.



*Note.* Word cloud visualizing the most commonly used nouns and adjectives for the top 10 conspiracy and science influencers. Bigger and bolder representation indicates that the words appeared more frequently in the source text.



# Conclusion

- Language reflects psychological states and supports an “us vs. them” framework in conspiratorial thinking.
- **Impact on Public Discourse:** Increased polarization due to conspiracy-themed language.
- **Counteracting Conspiracies:** Highlighting potential for inoculation against misinformation.
- Conspiratorial language differs significantly from scientific language, reflecting anger, uncertainty, and themes of power and death.
- Results inform strategies for combating misinformation and fostering more inclusive, fact-based discourse.