Classic Approaches in Quantitative Text Analysis

Étienne Ollion

Introduction

- Reading, and more « distant reading »
- An old endeavor: from the Bible’s index to content analysis
- With a long tradition in the human and social sciences

Introduction

An overview that is necessarily

- Subjective
- Incomplete (hundreds of works!)

Possible organizing principles

- Historical?
- Disciplinary?
- By use (text as a mean, text as an end)
- …

Introduction

Today: An Organization by ‘Families of Methods’

- Cuts Across Disciplines
- Highlights Strengths and Limitations
- One commonality: the DTM

Introduction

Why learning about ‘old’ techniques ?

- Knowing where the field comes from
- Still used, and still useful at times
- Progress, and forgotten pathways…

Outline

Families

- Lexical statistics
- Dictionary-based methods
- Multidimensional Lexical Statistics
- Topic Models
- Stylistic Analysis

In practice

- Mention intuition
- Present a few seminal texts, evoke use cases
- Present known limitations & current developments

Introduction

Builds on a few central reviews:

- Grimmer & Stewart, 2013 [PoliSci]
- Evans & Aceves, 2017 [Soc]
- Gentzkow, Kelly & Taddy, 2017 [Econ]
- Cointet & Parasie, 2018 [Soc]

A useful book: *Text as data*, Grimmer, Robert, Stewart 2022

Prologue: A ‘Shocking’ Idea

Prologue: A ‘Shocking’ Idea

A ‘Shocking’ Idea (Grimmer and Stewart, ‘12)

Treating **‘text as data’** (Laver et al., 2003)

In order to treat ‘text as data’:

- We should discard the order of words
- And we will score the ‘most important’ words

Shocking, but not without success

Prologue: A ‘Shocking’ Idea

In Practice: a **DTM** (Document-Term Matrix)

- d1: ‘*NLP is fun to learn*’
- d2: ‘*It is fun to teach NLP*’

Rows = documents, columns= terms

	NLP	**is**	**fun**	**to**	**learn**	**it**	**teach**
d1	1	1	1	1	1	0	0
d2	1	1	1	1	0	1	1

Prologue: A ‘Shocking’ Idea

All operations are then premised on the DTM

Relies on **‘bag of words’** hypothesis

- An original, numeric, representation of language
- A major loss of information
- A very sparse matrix (lots of 0s)
- Yet, many accomplishments

Lexical Statistics

Lexical Statistics

Oldest endeavor, different flavors

- **Berelson & Lazarsfeld (1948)**, and before them Weber 1913.
  - Role of the media in the shaping of mentalities
  - Focusing on (sometimes counting) words to capture meaning
  - The most ideographic approach

Lexical Statistics

Oldest endeavor, different flavors

- **BankSpeak** (Pestre & Moretti, 2015).

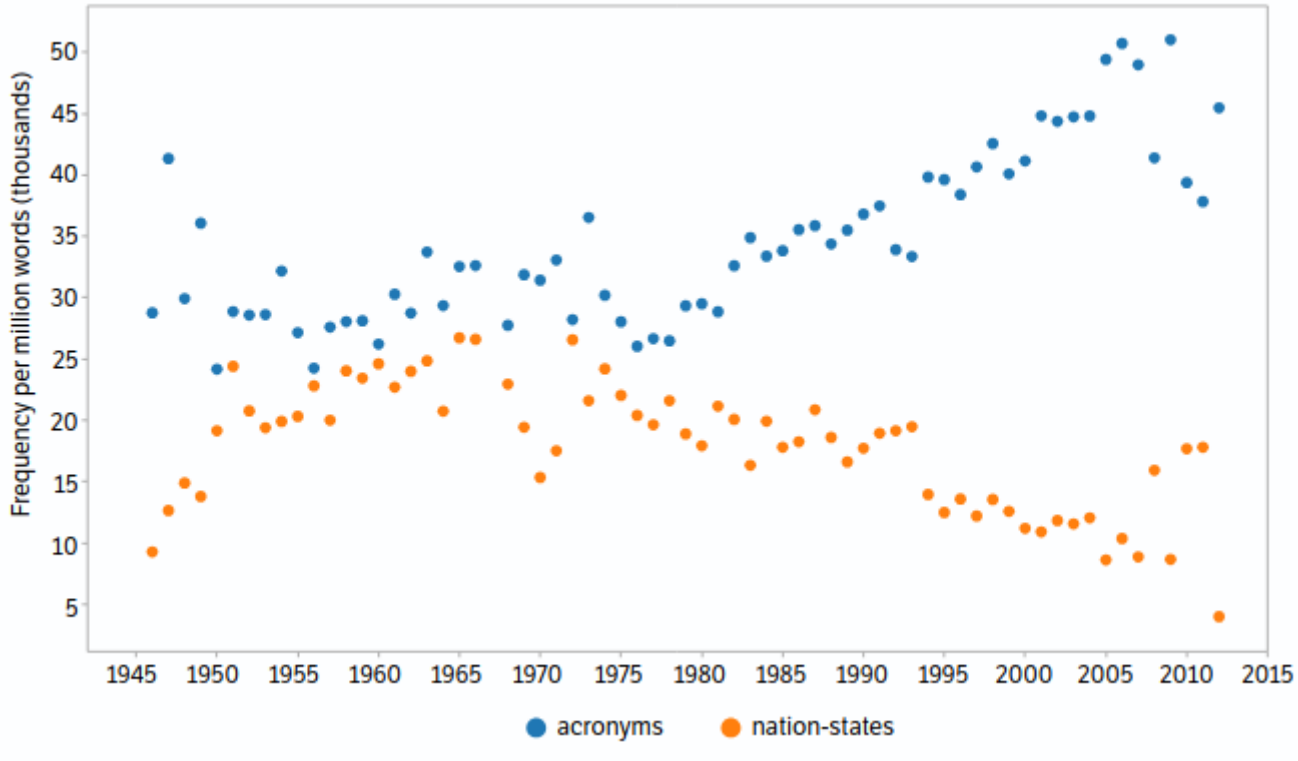
– An analysis of the langage in World Bank Reports over 40 years.

– « Behind this façade of uniformity, a major metamorphosis has taken place »

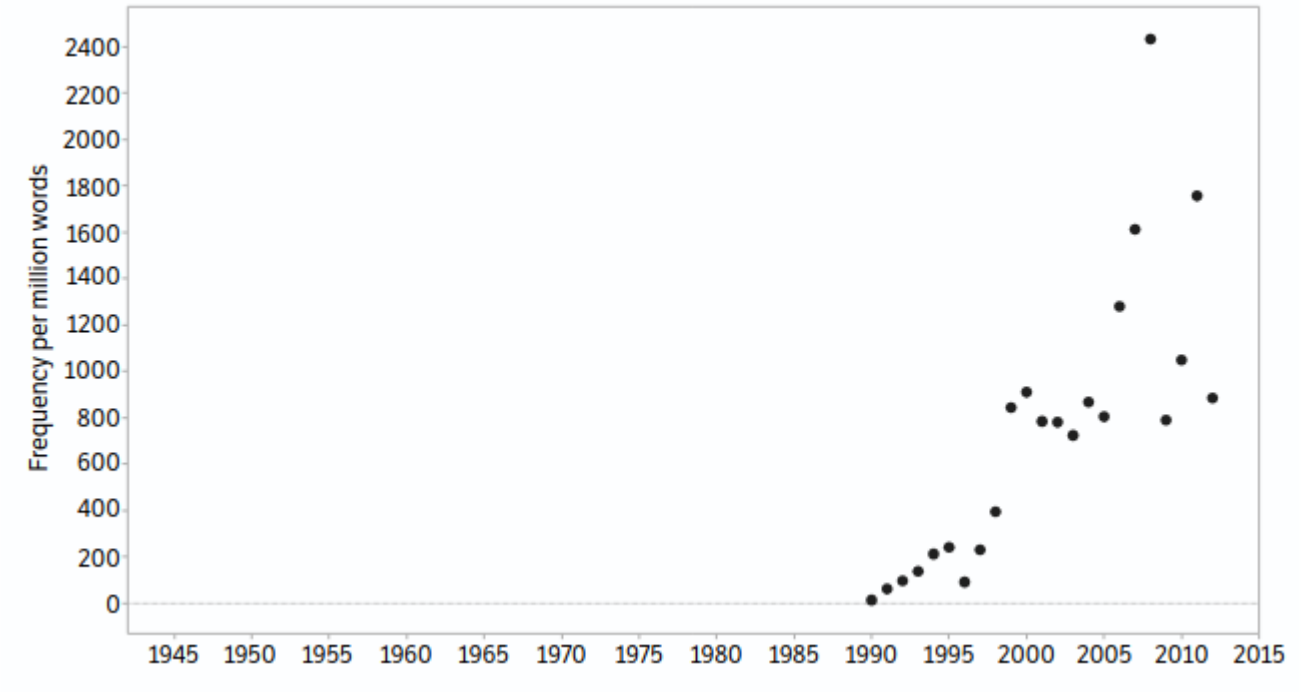
– Change in the semantics, from *plainspeak* to *bankspeak*.

– Matters because: language construes our representations (~ Sapir-Whorf hypothesis)

Lexical Statistics



Lexical Statistics



‘Governance’
(+ unlike government,
always positive)

Lexical Statistics

Oldest endeavor, very different options

Classic, but criticized

- Purely descriptive
- Massive reduction (⇒ need a lot of signal)
- Lack of structure, of context (Guerrini 2011)

Dictionary-based Methods

Dictionary-based Methods

Old idea too (Stone et al., 1966)

Revival in the 2000s, partly due to commercial interest (Pang *et al.*, 2000 ; Pang & Lee, 2008)

Not a focus on words, but on broader categories the word point towards.

Dictionary-based Methods

Most classic example: **sentiment analysis**

- Determine a sentiment score for a sentence/doc
- Based on certain pre-determined terms denoting positive or negative sentiments

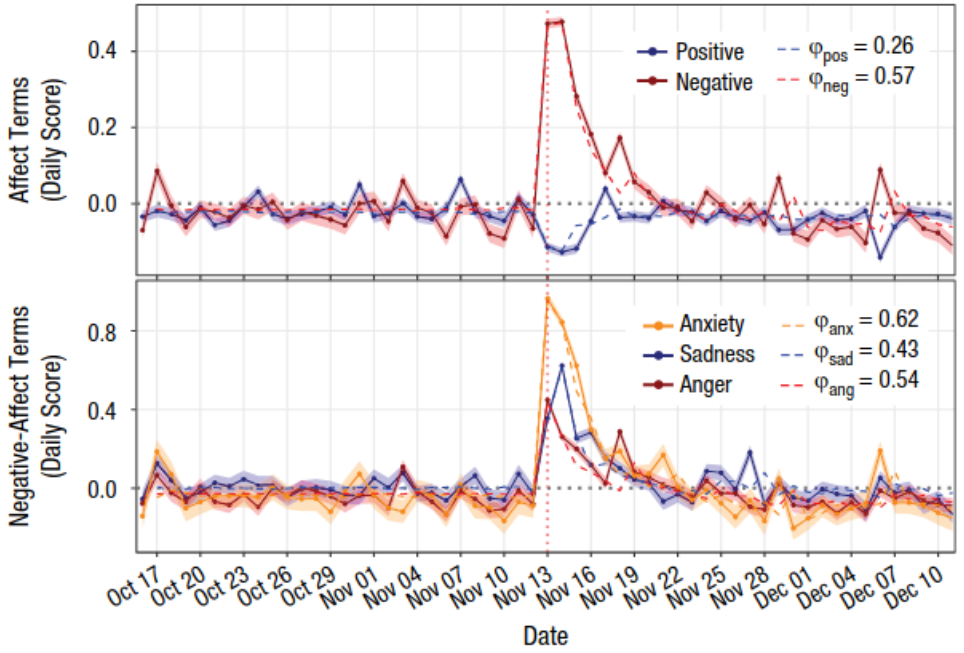


Fig. 2. Daily scores of positive affect and negative affect (top) and anxiety, sadness, and anger (bottom). Points show median estimates over 10,000 bootstrapped samples, and shaded areas show their 95% confidence intervals. Dashed lines show model fits, along with the estimated values of each φ. The dotted vertical line marks the day of the attacks.

Garcia and Rimé, 2019

Dictionary-based Methods

Most classic example: **sentiment analysis**

- Flores, Anti-Immigrant Sentiment, *AJS* 2017
  - Does the passing of the law influence public opinion, and if yes, how ?
  - Tweets in Arizona in 2010 after the passing of a restrictive law. Control with Nevada.
  - Advanced Sentiment Analysis
    - Scores gradually (from -4 to 4)
    - Distinguishes subject of message
    - Controls for # of active twitter accounts

      → Feeds into regression models

Dictionary-based Methods

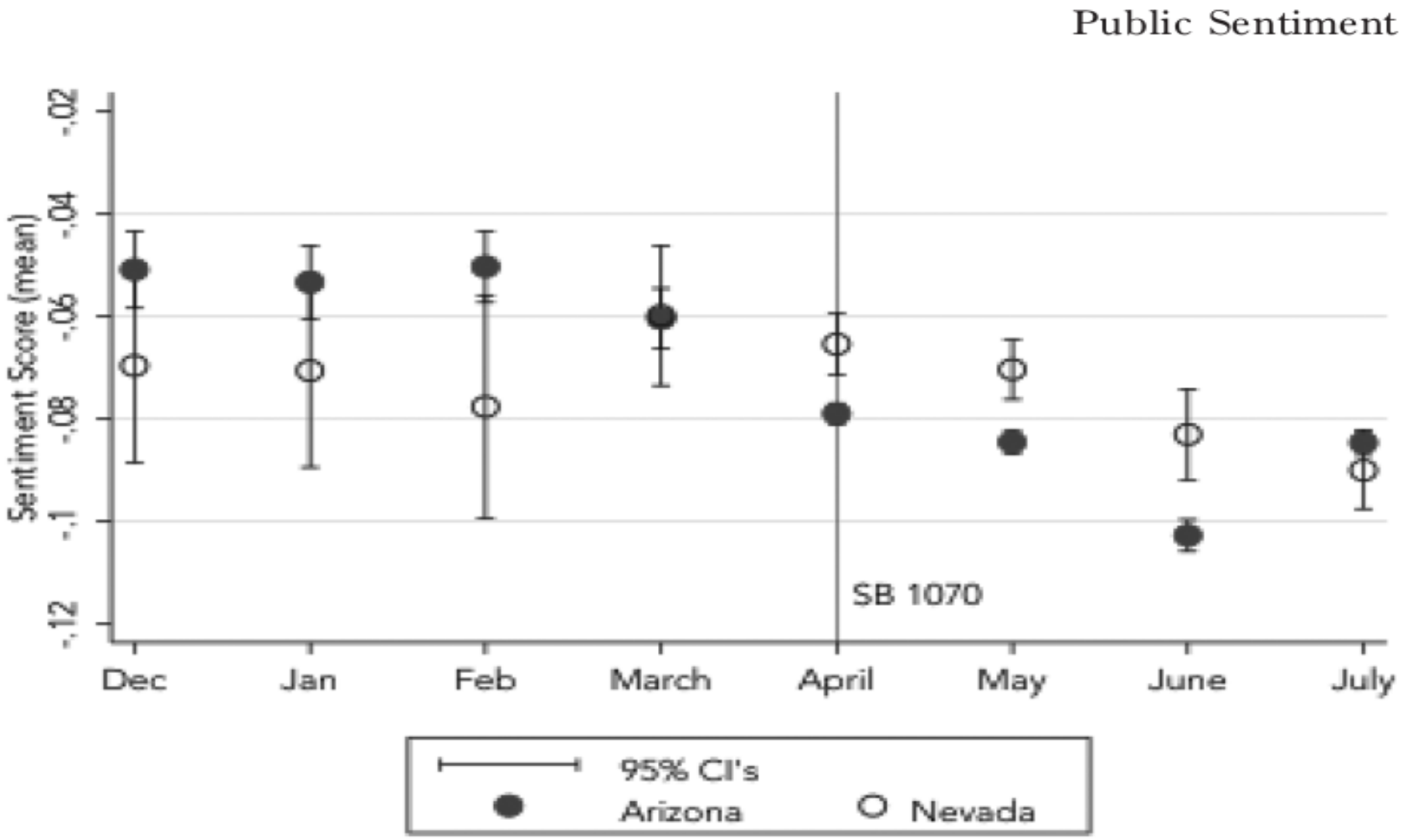


FIG. 3.—Average sentiment score of tweets about immigrants. The vertical lines represent 95% confidence intervals. The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

Dictionary-based Methods

Known issues

- Does not deal well with negation, with irony, second degree, metaphors (Bosco et al., 2013)
- What about multiple referents ? (Ruiz & Poibeau 2015)
- Dichotomous categories (positive, negative)

Multidimensional Lexical Stats

- A long (French) tradition (Lebart and Salem, ‘94)
- Using word co-occurences
- Multiple uses: associations, clustering, author detection, etc...

Multidimensional Lexical Stats

- A long (French) tradition
  (Lebart, Salem, Berry, ‘98, see Beaudoin 2019 for a history)
- Using word co-occurences
- Multiple uses: associations, clustering, author detection, etc…

Multidimensional Lexical Stats

Mapping words

- Analyzing open-ended questions, speeches
- Showing proximities and distance in speech

Multidimensional Lexical Stats

Mapping words

- Showing proximities and distance in speech

What would make you hesitate to have children ?

Exemples de réponses à la question *Enfants*

Ind=01	*les difficultés financières et matérielles*
Ind=02	*les problèmes matériels, une certaine angoisse vis à vis de l'avenir*
Ind=03	*la peur du futur, la souffrance, la mort, le manque d'argent*
Ind=04	*l'avenir incertain, les problèmes financiers*
Ind=05	*les difficultés financières*
Ind=06	*les raisons matérielles et l'avenir qui les attend*
Ind=07	*des problèmes financiers*
Ind=08	*l'avenir difficile qui se prepare, la peur du chomage*
Ind=09	*l'insecurite de l'avenir*
Ind=10	*le manque d'argent*
Ind=11	*la guerre éventuelle*
Ind=12	*la charge financière que ca représente, la responsabilité morale aussi*
Ind=13	*la situation économique, quand le couple ou la femme n'est pas psychologiquement prêt pour accueillir un enfant*
Ind=14	*raisons éthiques*

Lebart, Salem, Berry 1998

Multidimensional Lexical Stats

Mapping words

- Showing proximities and distance in speech
- Factorial analysis as the tool of choice

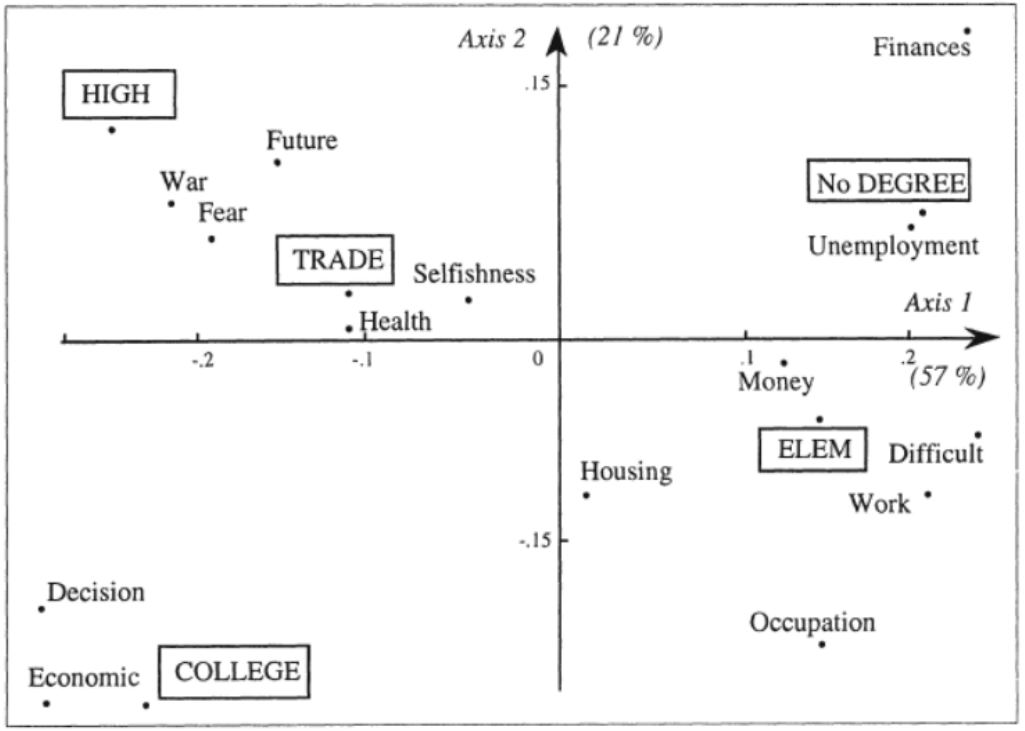


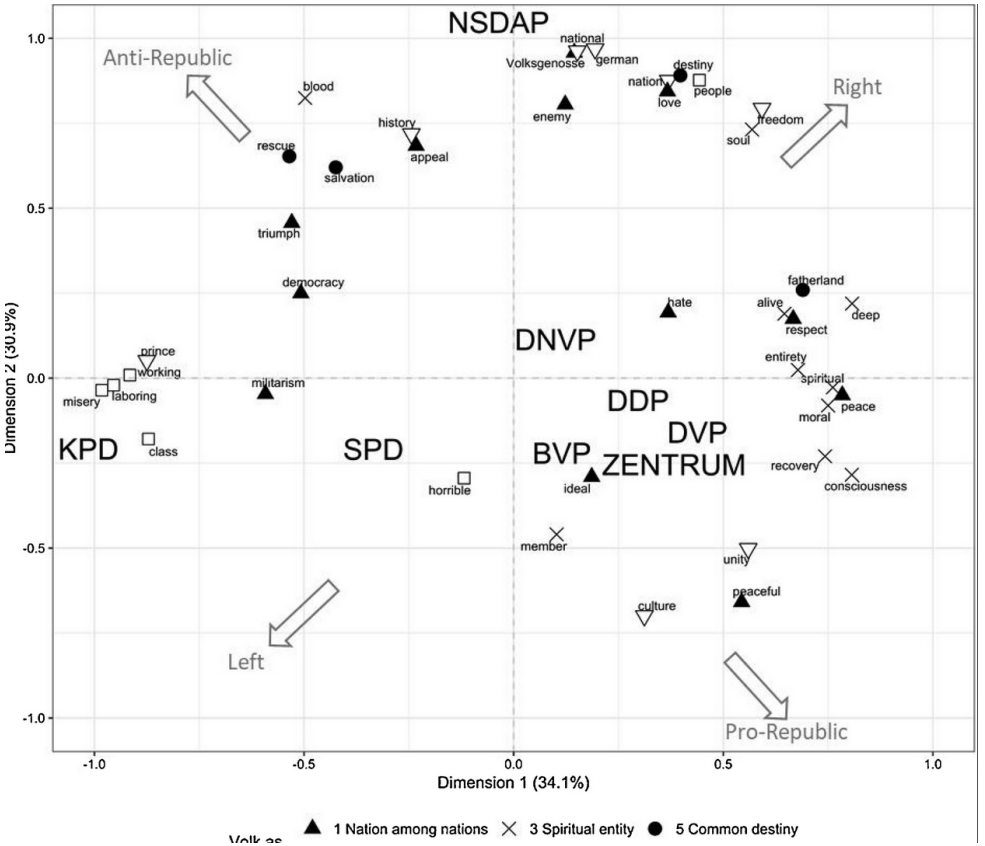
Figure 2. Proximities among words and among educational level
(Lebart et al., 1998, p. 52)

Lebart, Salem, Berry 1998

Multidimensional Lexical Stats

Mapping words

- Showing proximities and distance in speech



Fuhse et al, 2021

Multidimensional Lexical Stats

Grouping, Separating Texts

Multidimensional Lexical Stats

Grouping, Separating Texts

- For instance to attribute authorship (see *stylometry*)
- Premised on words (= form)

Multidimensional Lexical Stats

Grouping, Separating Texts

Did Shakespeare write a given poem ?

1. Word frequencies in Shakespeare's

Gamme des fréquences dans l'oeuvre de Shakespeare (tableau partiel)

Fréq. f_i	*Effectif V_i*	*Fréq. f_i*	*Effectif V_i*
1	14376	*8*	519
2	4343	*9*	430
3	2202	*10*	364
4	1463	*11*	305
5	1043	*12*	259
6	837	.....	
7	638	*>12*	846
Nombre des formes	$= \Sigma_i V_i$	31 534	
Nombre des occurrences	$= \Sigma_i f_i V_i$	884 647	

(Efron et Thisted , 1976)

Multidimensional Lexical Stats

Grouping, Separating Texts

Did Shakespeare write a given poem?

2. Comparisons between texts

Distribution des formes dans les huit poèmes, selon leur fréquence d'apparition dans l'oeuvre de Shakespeare.

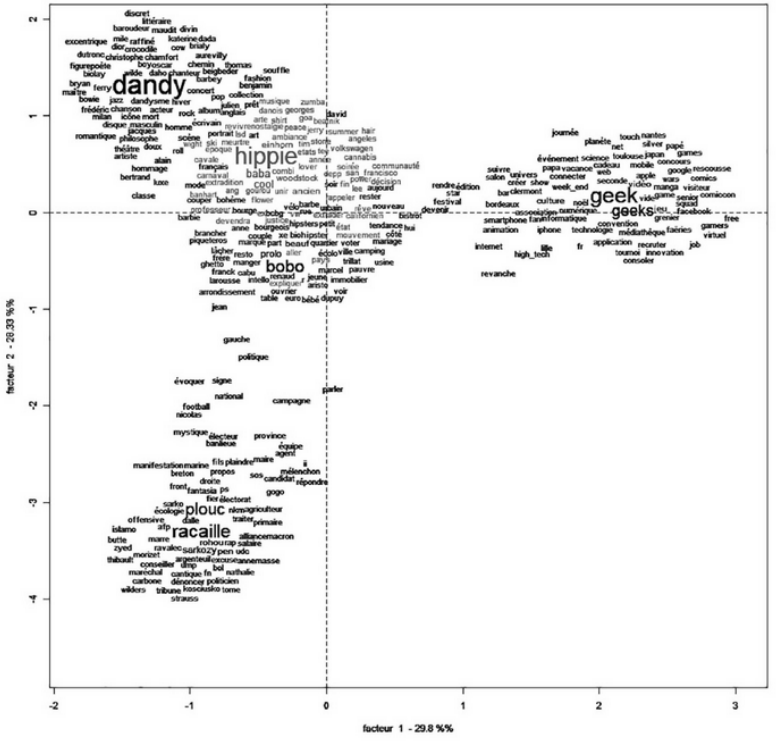
Freq.	BJon	Marl	Donn	Cymb	Mids	Phoe	Sonn	Tayl	*Total*
0	8	10	17	7	1	14	7	9	73
1	2	8	5	4	4	5	8	7	43
2	1	8	6	3	0	5	1	5	29
3-4	6	16	5	5	3	9	5	8	57
5-9	9	22	12	13	9	8	16	11	100
10-19	9	20	17	17	6	18	14	10	111
20-29	12	13	14	9	9	13	12	21	103
30-39	12	9	6	12	4	7	13	16	79
40-59	13	14	12	17	5	13	12	18	104
60-79	10	9	3	4	9	8	13	8	64
80-99	13	5	10	4	3	5	8	5	53
+100	148	138	145	120	103	111	155	140	1060
Total	243	272	252	215	156	216	264	258	1876

Valeurs théoriques et observées pour le poème *Taylor*

Fréquences	*Taylor observé*	*Taylor théorique*
0	9	6.97
1	7	4.21
2	5	3.33
3-4	8	5.36
5-9	11	10.24
10-19	10	13.96
20-29	21	10.77
30-39	16	8.87
40-59	18	13.77
60-79	8	9.99
80-99	5	7.48

Multidimensional Lexical Stats

- Capturing the « worlds of the words »
- Clustering
- Yet: « words, words, words »



(Roquebert, 2018)

Topic models

How to classify themes over a large number of texts ?

- Dictionary-based methods are an option
- Topic models is their **unsupervised counterpart**
  - Unsupervised: opposed to supervised
  - When there is no established coding scheme, nor can we know how to do the classification.

Topic models

How to classify themes over a large number of texts ?

- Topic Models (LDA): Blei, 2000s
  > « uncover underlying semantic regularities in a set of documents by mapping recurring relationships between words ».
  > 100s (1000s?) of works in the Social Sciences

Topic models

How to classify themes over a large number of texts ?

- Topic Models (LDA): Blei, 2000s
  In practice:
  - Assumes that texts are made of topics
  - Assumes that topics are made of words
  - Probabilist retrieval of said topics (k, predetermined)

Topic models

How to classify themes over a large number of texts ?

- Countless examples in the social sciences (2010s)
  > Fligstein *et al.*, *ASR*, 2017
  Why did the Fed fail to foresee the 2008 crisis?
  ⇒ (macro) frames and confirmation bias

  72 FOMC meetings, basic LDA on those documents

Topic models

How to classify themes over a large number of texts ?

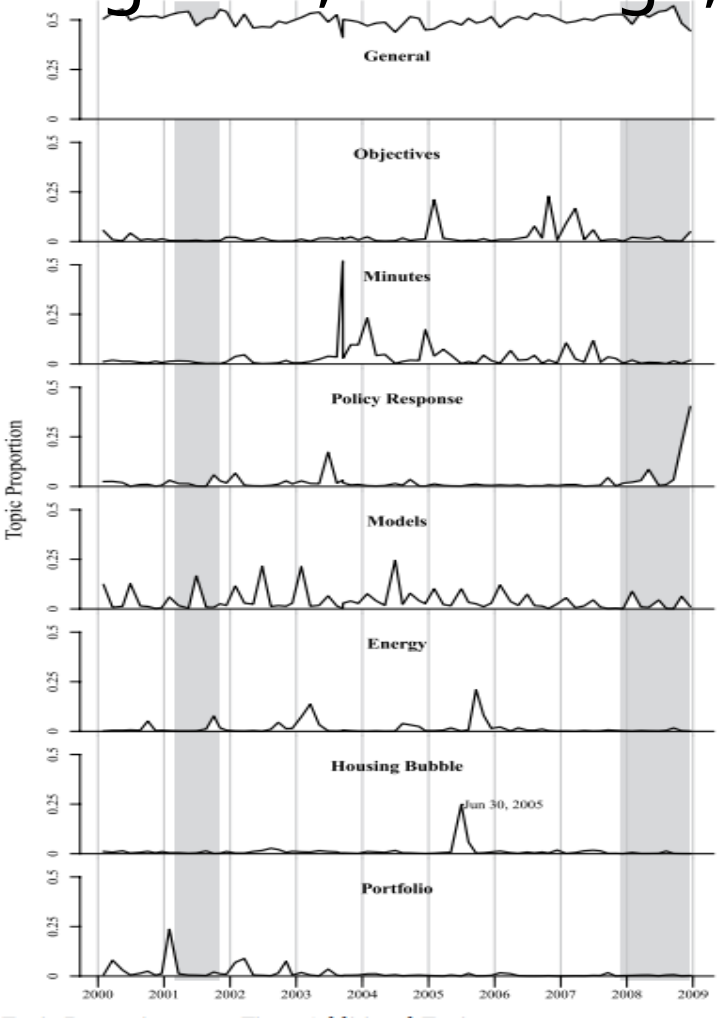
- Fligstein, Brundage, Schultz 2017

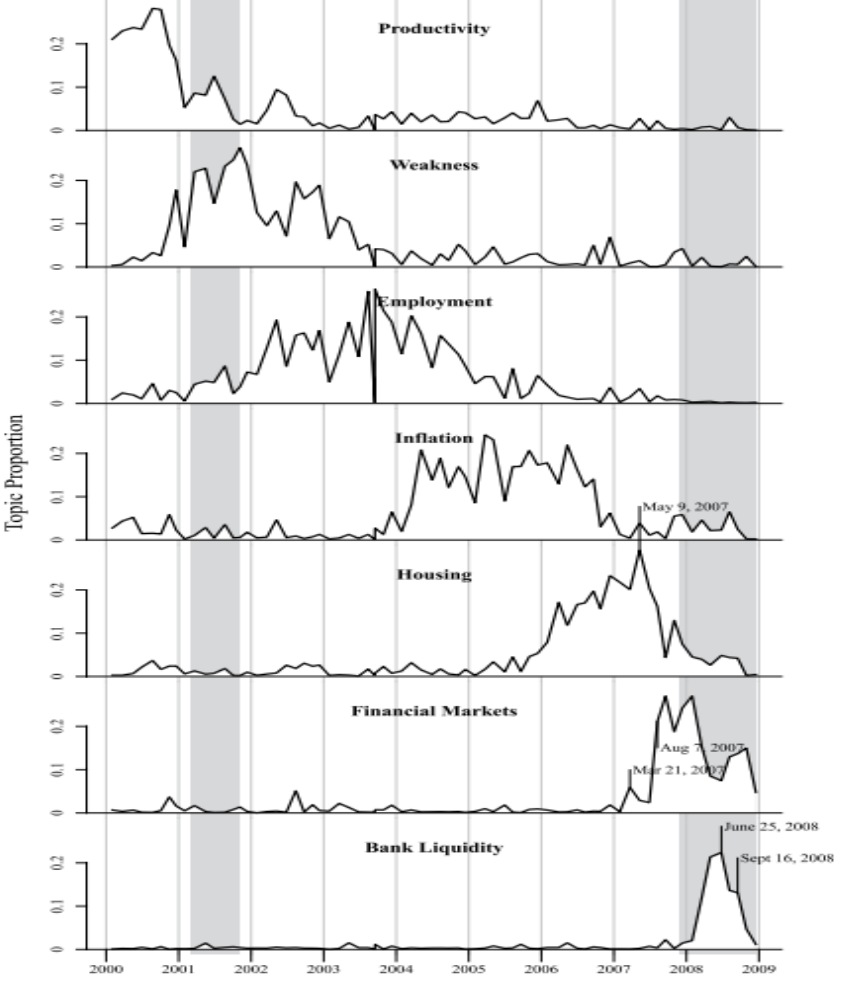
Portfolio	Housing Bubble	Energy	Models	Policy Response	Minutes	Objectives	General
contingency	arms	disruptions	depreciation	treasury	formulaic	Congress	see
Mae	lenders	gas	red	banks	think	horizon	like
window	loans	inertia	present	facility	announce-ment	adopt	economy
collateral	constant	storm	year	guarantee	meeting	achieve	well
outright	quality	impact	dollar	interest	expediting	anchored	don
sovereign	value	refinery	exports	purchases	vote	public	may
issue	afford	ports	top	ceiling	information	benefits	much
discount	family	barrel	foreign	quantity	process	committee	can
system	ofheo	crude	simulations	effect	decision	run	even
liquidity	percentile	energy	variables	deflation	view	definition	get
Lombard	nonmarket	heating	account	tools	memo	regime	chairman
Freddie	appreciation	effect	bars	excess	oni	prices	say
debt	bond	Venezuela	Taylor	money	give	defined	risk
operations	component	million	structural	policy	issue	specific	because
tally	Francisco	aftermath	rate	fomc	editing	think	look
gnmas	misalloca-tion	stagflation	unemploy-ment	size	use	cpi	come
diversified	shown	damage	productivity	monetary	transparency	consensus	know
disclose	bias	inertial	different	desk	press	diversity	next
Fannie	city	coast	show	alternative	convey	transparency	policy

Topic models

How to classify themes over a large number of texts ?

- Fligstein, Brundage, Schultz, *ASR*, 2017





Topic models

How to classify themes over a large number of texts ?

- Many examples in the social sciences
- Classic criticisms:
  - Brittle (Shadrova, 2021)
  - ‘Exploratory analysis’ (see Grimmer & Stewart 2013).
  - Change in the meaning of words
- Refinements
  - Over time (dynamic topic models)
  - Conditional on certain terms

Stylistic analysis

Stylistic analysis

Not so frequent but full of potential

- Idea: focus on the « style » (use of langage, deviations from norms) to investigate formality, complexity, politeness, etc.

Stylistic analysis

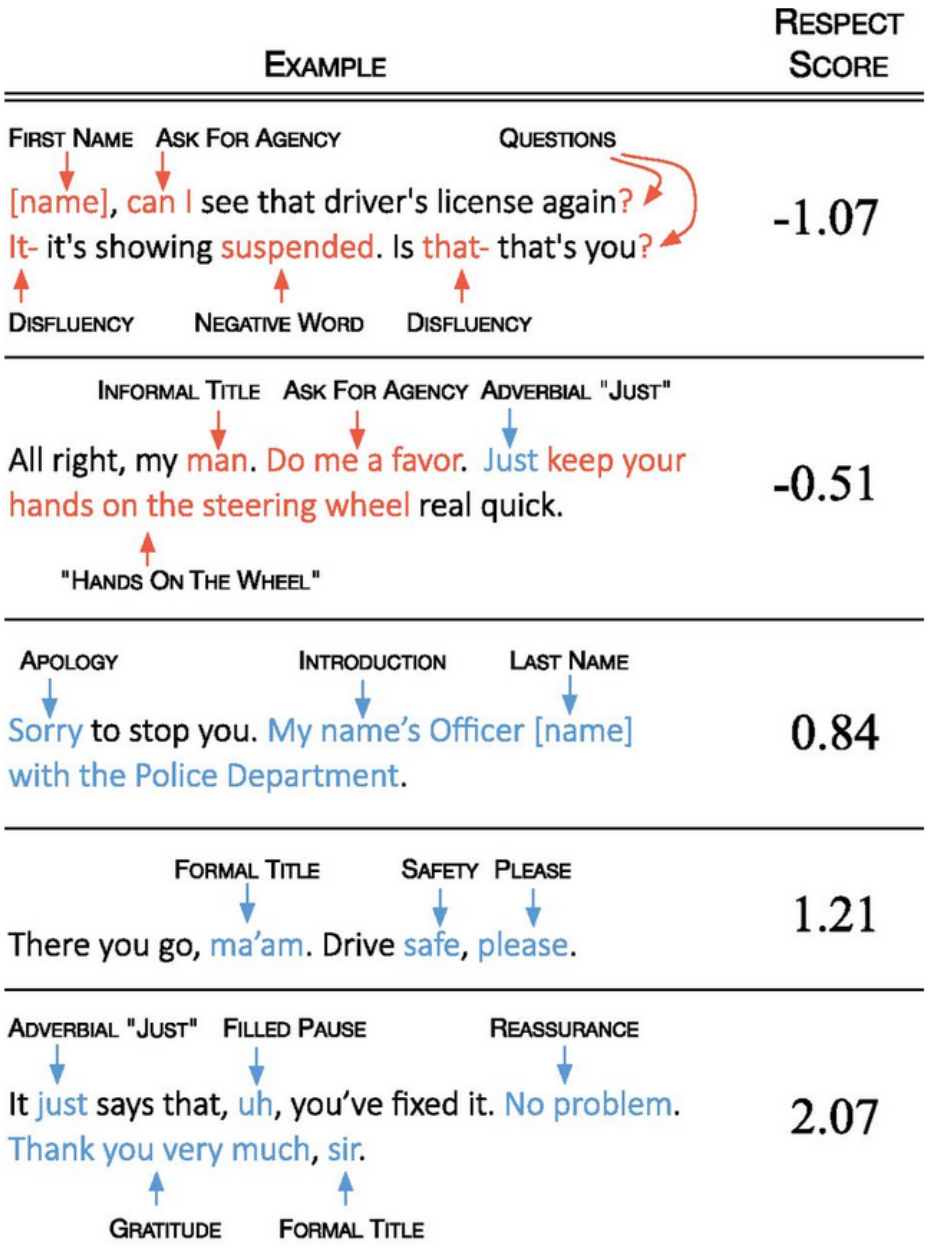
Not so frequent but full of potential

- **Voigt et al.**, *PNAS*, 2017
  - Are police/citizen interactions racialized?
  - Using information from body camera footages.

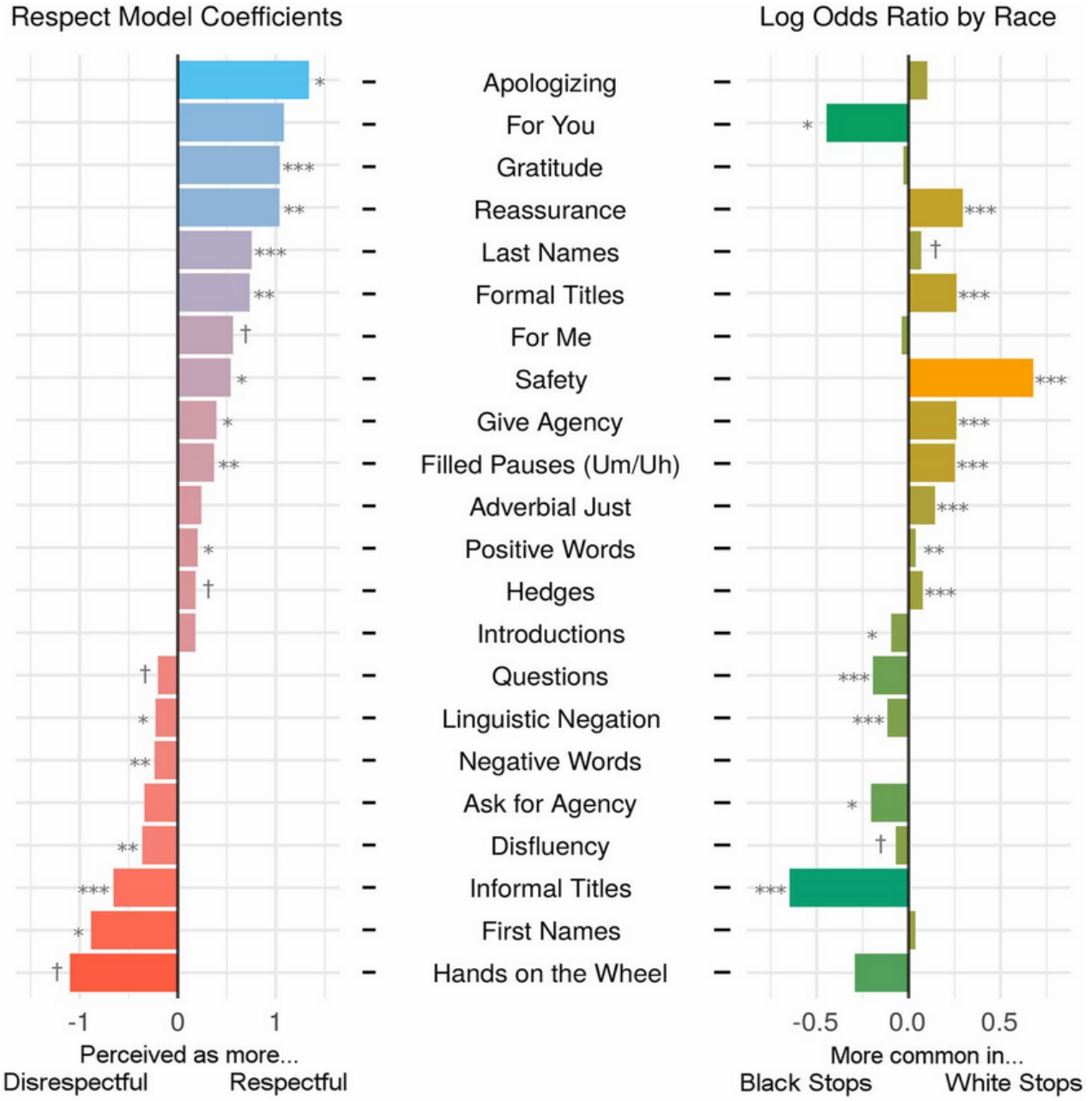
    > Analyzing officers' language during vehicle stops of white and black community members.

    > Controls by place, race of officer, type of suspected infraction, time of the day.

Stylistic analysis



Stylistic analysis



Conclusion

- Variety of methods, of approaches within a family
- Text analysis as an end, text analysis as a means

Conclusion

- One common assumption: B.O.W. hypothesis
  - Very unnatural assumption
  - Lots of information lost
  - Requires processing (lemmatization, selection)

⇒ Towards a more realistic representation of language