Mining cultural insights from online texts

NSF Specialists Meeting
August 12, 2015

Rob Malouf
San Diego State University

Language diversity

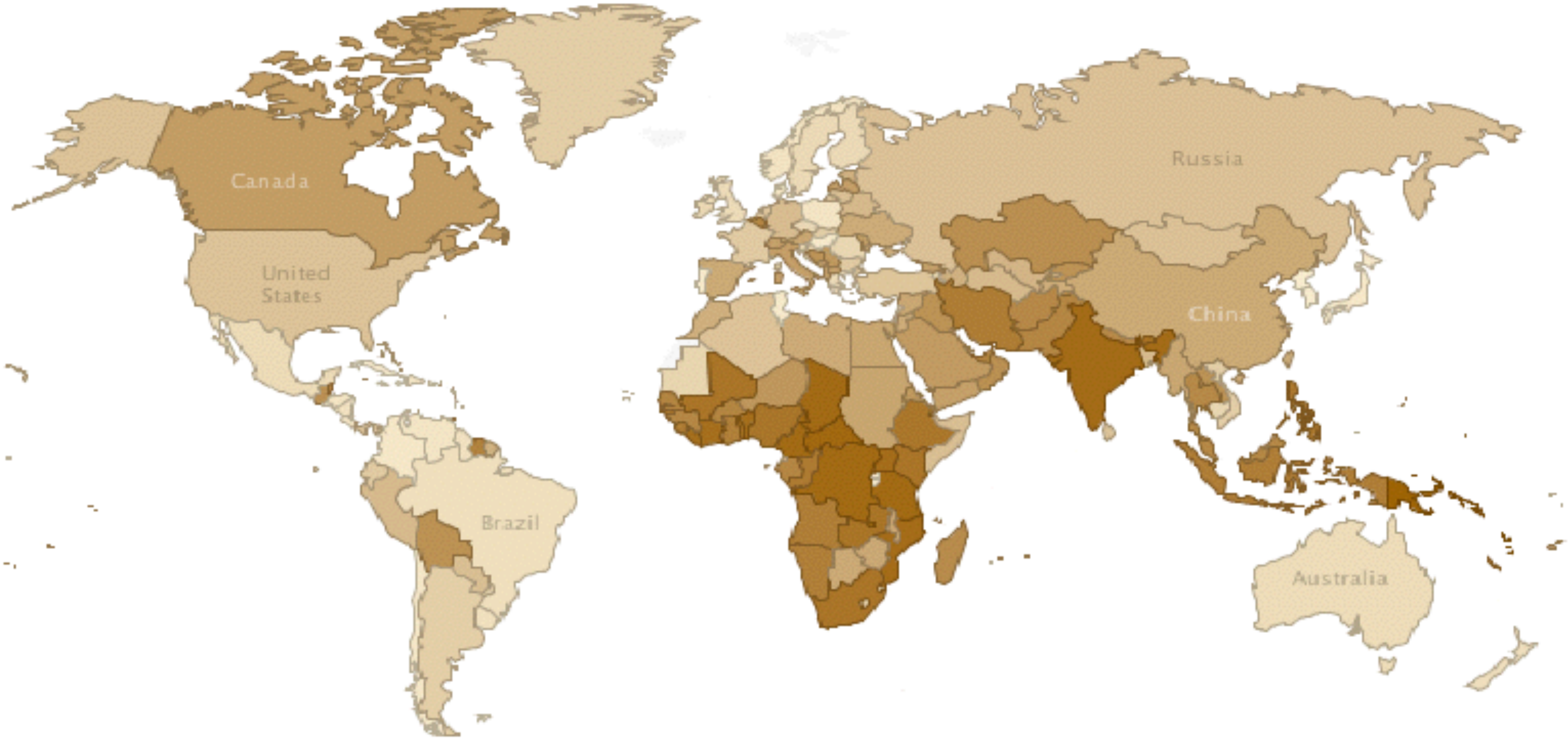
Language variation is a constant

Variation in space

- Twitter ‘supports’ 33 languages
- *Ethnologue* lists 7,102 languages
- ISO 639-3 codes for 8,140 languages
- Countless regional, ethnic, professional dialects / varieties

Variation in time

Language diversity



Fæder ūre, þū þe eart on heofonum; Sīe þīn nama gehālgod, tō becume þīn rīce, gewurþe þīn willa, on eorðan swā swā on heofonum. Ūrne gedæghwamlican hlāf sele ūs tōdæg, and forgif ūs ūre gyltas, swā swā wē forgifaþ ūrum gyltendum, and ne gelǣd þū ūs on costnunge, ac ālīes ūs of yfele, sōþlīce.

Oure fadir that art in heuenes halowid be this name, thi kyngdom come to, be thi wille don in erthe es in heuene, yeue to us this day oure bread ouir other substance, & forgeue to us oure dettis, as we forgeuen to oure dettouris, & lede us not in to temptacion: but delyuer us from yuel, amen.

Our father which art in heaven, hallowed be thy Name. Thy kingdome come. Thy will be done, in earth, as it is in heaven. Giue vs this day our dayly bread. And forgiue vs our debts, as we forgiue our debters. And leade vs not into temptation, but deliuer vs from euill: For thine is the kingdome, and the power, and the glory, for euer, Amen.

Our Father, who is in heaven, may your name be kept holy. May your kingdom come into being. May your will be followed on earth, just as it is in heaven. Give us this day our food for the day. And forgive us our offenses, just as we forgive those who have offended us. And do not bring us to the test. But free us from evil. For the kingdom, and the power, and the glory are yours forever. Amen.

Meaning

reversible vs. *irreversible*

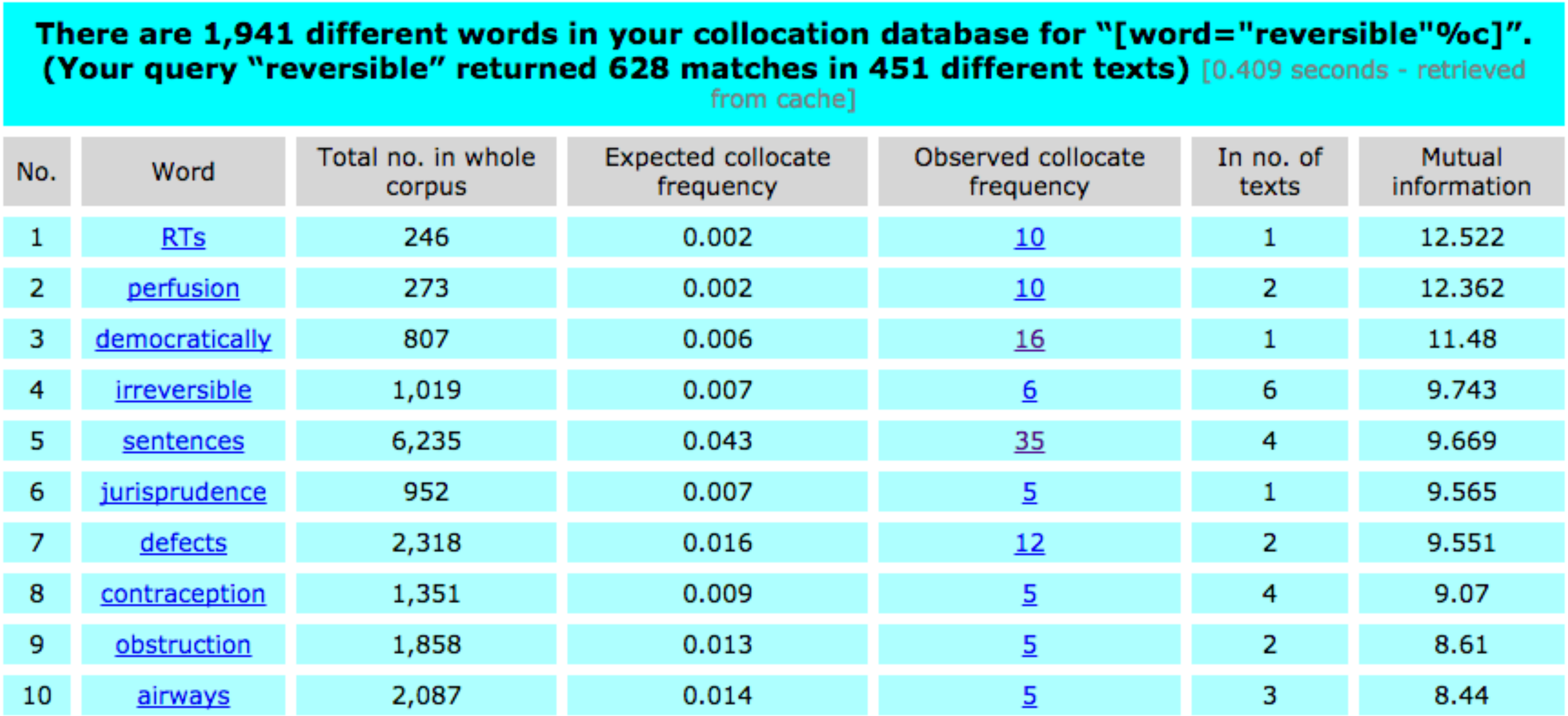
In Standard American English (COCA), *reversible* has many senses, but *irreversible* is fairly restricted

Biomedical contexts:

> Following a careful screening by a physician for concurrent medical illnesses and **reversible** causes of pain, persons with CBP undergo evaluation by physical and exercise therapists .

> In 1993 , the gene that determines the occurrence of Huntington’s disease, an **irreversible** degeneration of the nervous system , was discovered .

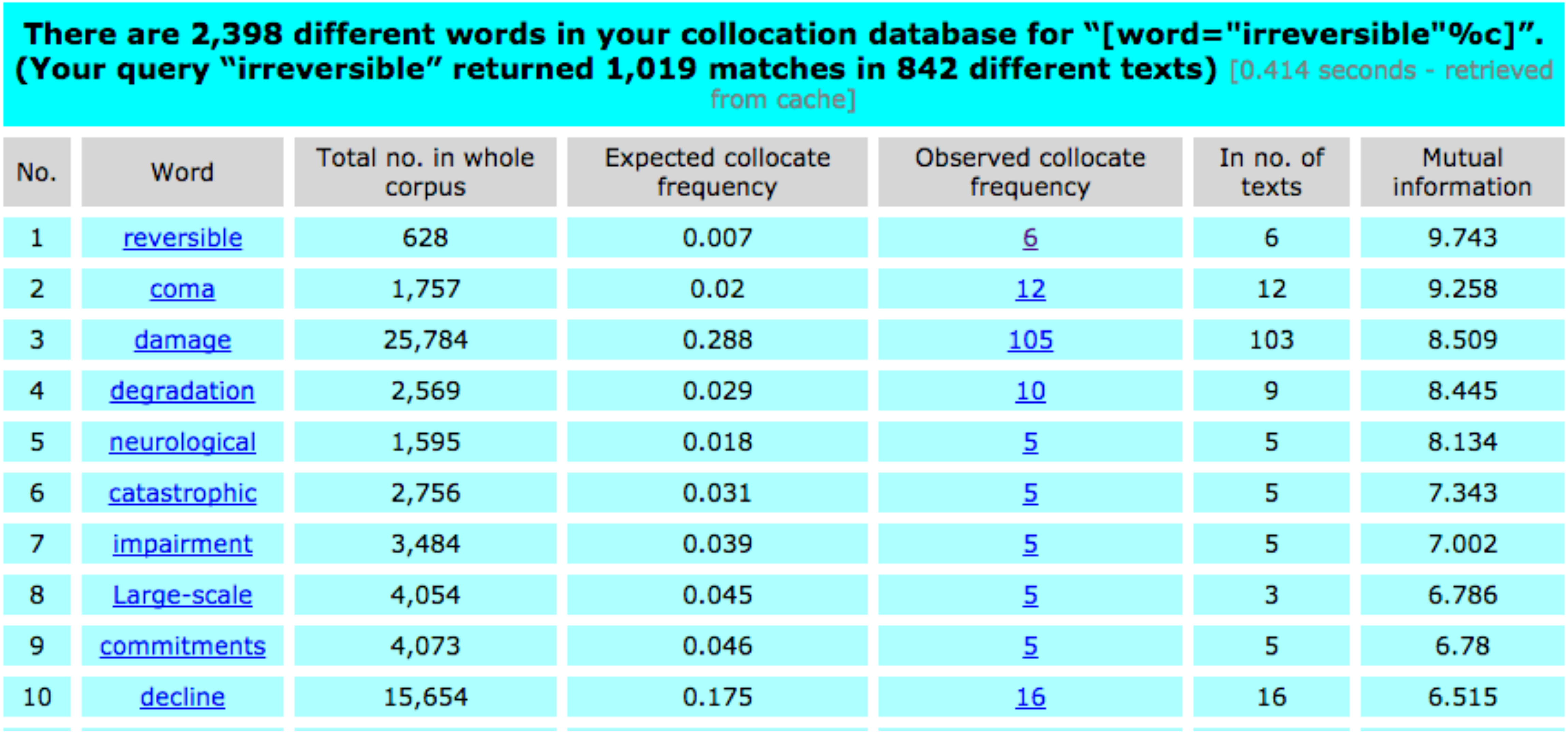
Meaning



There are 1,941 different words in your collocation database for "[word="reversible"%c]". (Your query "reversible" returned 628 matches in 451 different texts) [0.409 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Mutual information
1	RTs	246	0.002	10	1	12.522
2	perfusion	273	0.002	10	2	12.362
3	democratically	807	0.006	16	1	11.48
4	irreversible	1,019	0.007	6	6	9.743
5	sentences	6,235	0.043	35	4	9.669
6	jurisprudence	952	0.007	5	1	9.565
7	defects	2,318	0.016	12	2	9.551
8	contraception	1,351	0.009	5	4	9.07
9	obstruction	1,858	0.013	5	2	8.61
10	airways	2,087	0.014	5	3	8.44

Meaning

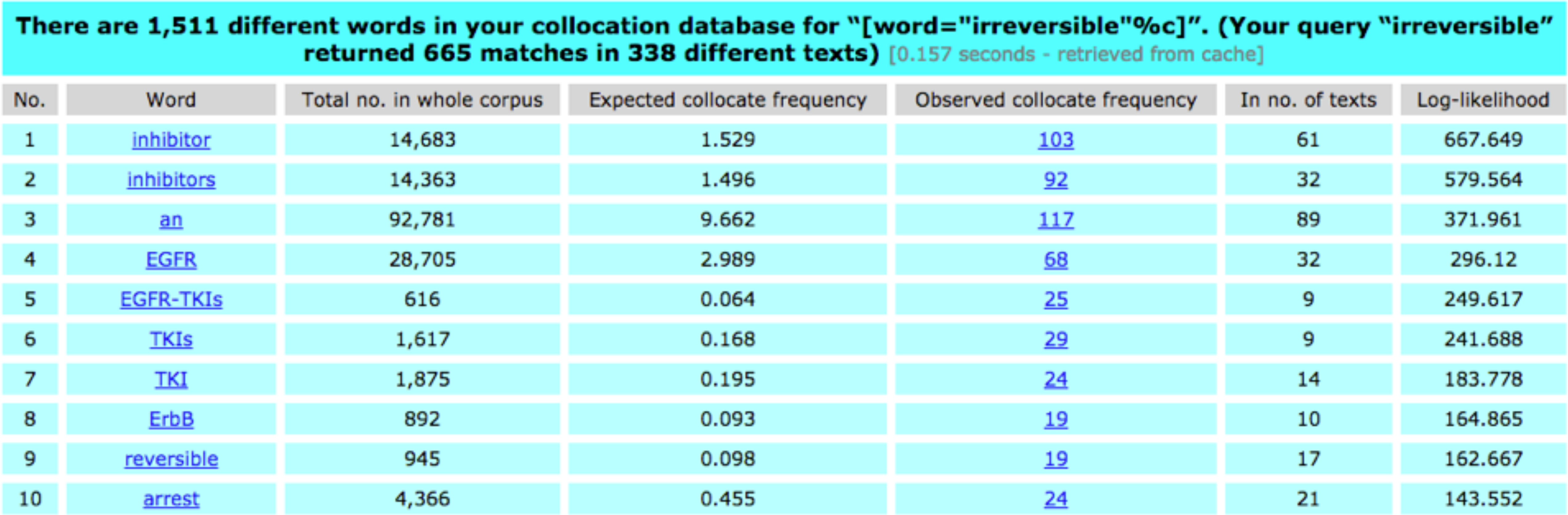


There are 2,398 different words in your collocation database for "[word="irreversible"%c]". (Your query "irreversible" returned 1,019 matches in 842 different texts) [0.414 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Mutual information
1	reversible	628	0.007	6	6	9.743
2	coma	1,757	0.02	12	12	9.258
3	damage	25,784	0.288	105	103	8.509
4	degradation	2,569	0.029	10	9	8.445
5	neurological	1,595	0.018	5	5	8.134
6	catastrophic	2,756	0.031	5	5	7.343
7	impairment	3,484	0.039	5	5	7.002
8	Large-scale	4,054	0.045	5	3	6.786
9	commitments	4,073	0.046	5	5	6.78
10	decline	15,654	0.175	16	16	6.515

Meaning

Specialized corpus: 7,317 papers (38,314,863 words) on *non-small cell lung cancer*



There are 1,511 different words in your collocation database for "[word="irreversible"%c]". (Your query "irreversible" returned 665 matches in 338 different texts) [0.157 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	inhibitor	14,683	1.529	103	61	667.649
2	inhibitors	14,363	1.496	92	32	579.564
3	an	92,781	9.662	117	89	371.961
4	EGFR	28,705	2.989	68	32	296.12
5	EGFR-TKIs	616	0.064	25	9	249.617
6	TKIs	1,617	0.168	29	9	241.688
7	TKI	1,875	0.195	24	14	183.778
8	ErbB	892	0.093	19	10	164.865
9	reversible	945	0.098	19	17	162.667
10	arrest	4,366	0.455	24	21	143.552

Afatinib is an **irreversible** EGFR/HER2 inhibitor developed by Boehringer Ingelheim [11] currently being clinically evaluated in NSCLC .

Vector Space Models

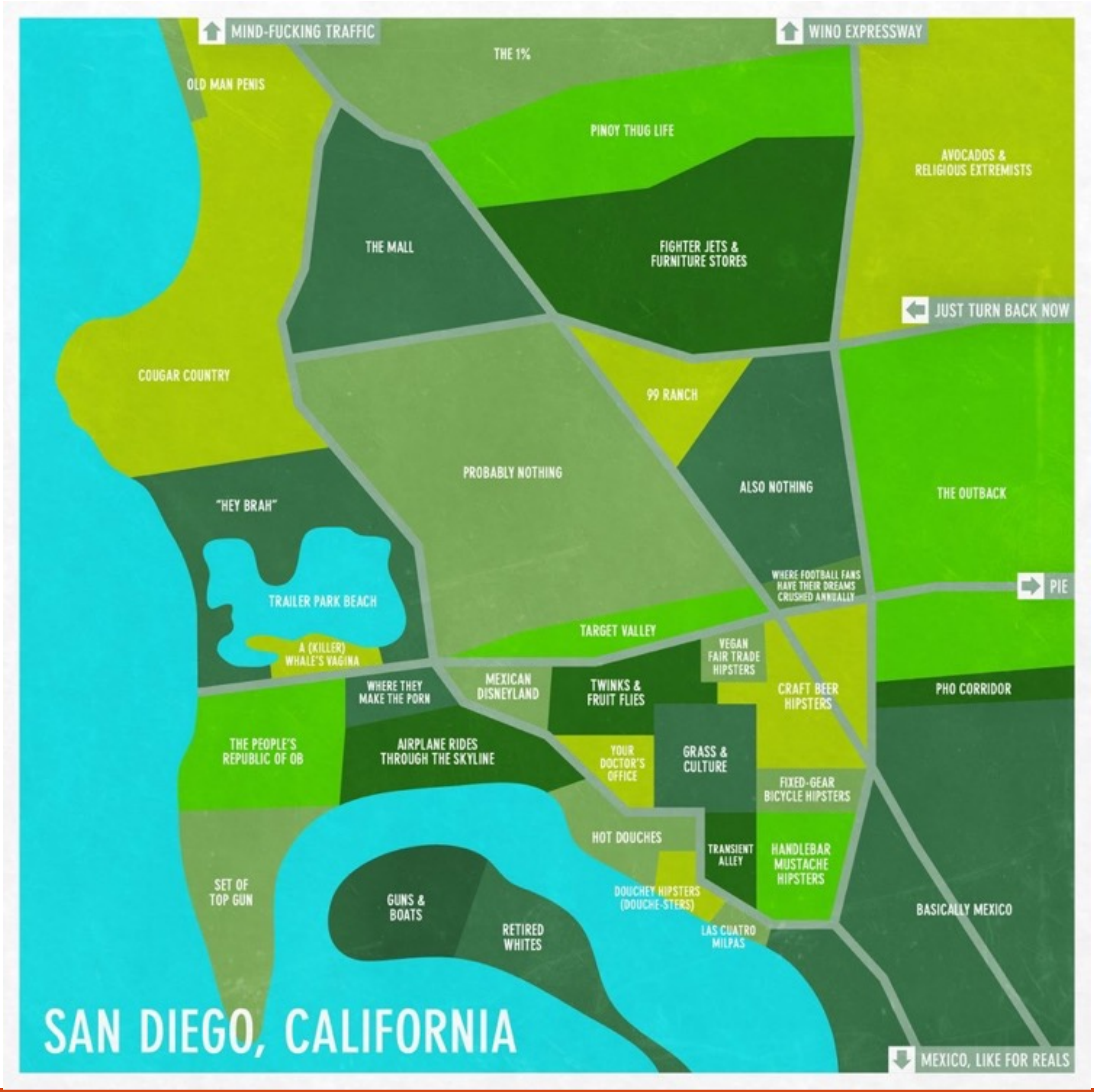
Vector Space Models are one way to operationalize a distributional notion of meaning

Results from shallow methods can only be as good as the input (representativeness)

Corpus of reading material for K-12 students

- **bicycle**: pedals, handlebars, bicycles, pedaling, bike, starley, highwheeler, boneshaker, mede, lallement, gearwheels, gearwheel, drais, bikers, bikes, wheels, wheel, bicycled, pedal

- **patriot**: 1775, patriots, lexington, concord, loyalist, loyalists, 1777, bunker, minutemen, hancock, 1776, redcoats, ticonderoga, sniping, framingham, edgel, revere, cornwallis, saratoga

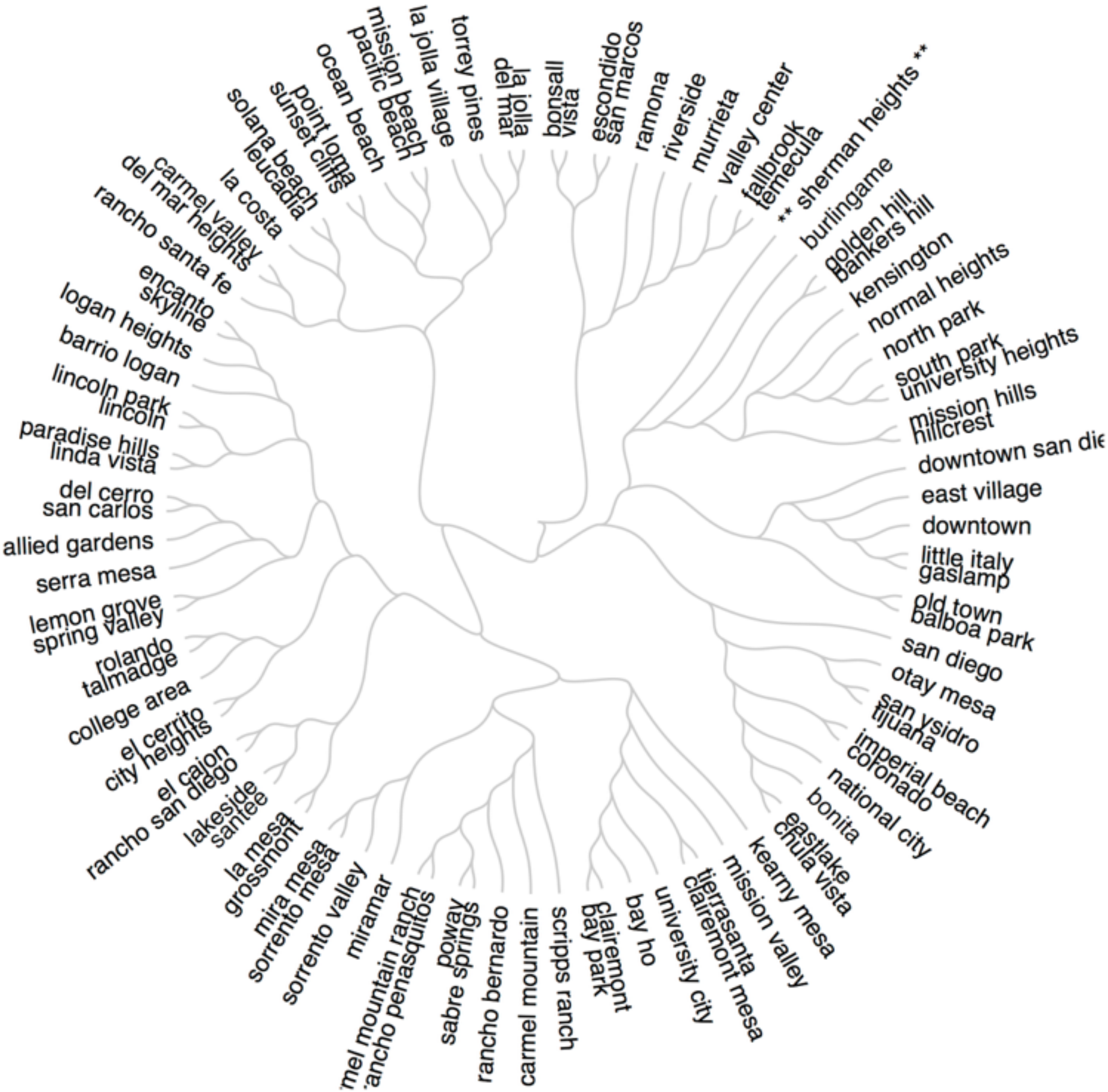

MIND-FUCKING TRAFFIC
WINO EXPRESSWAY
THE 1%
OLD MAN PENIS
PINOY THUG LIFE
AVOCADOS & RELIGIOUS EXTREMISTS
THE MALL
FIGHTER JETS & FURNITURE STORES
JUST TURN BACK NOW
COUGAR COUNTRY
99 RANCH
PROBABLY NOTHING
ALSO NOTHING
THE OUTBACK
"HEY BRAH"
WHERE FOOTBALL FANS HAVE THEIR DREAMS CRUSHED ANNUALLY
PIE
TRAILER PARK BEACH
TARGET VALLEY
A (KILLER) WHALE'S VAGINA
VEGAN FAIR TRADE HIPSTERS
WHERE THEY MAKE THE PORN
MEXICAN DISNEYLAND
TWINKS & FRUIT FLIES
CRAFT BEER HIPSTERS
PHO CORRIDOR
THE PEOPLE'S REPUBLIC OF OB
AIRPLANE RIDES THROUGH THE SKYLINE
YOUR DOCTOR'S OFFICE
GRASS & CULTURE
FIXED-GEAR BICYCLE HIPSTERS
HOT DOUCHES
TRANSIENT ALLEY
HANDLEBAR MUSTACHE HIPSTERS
SET OF TOP GUN
GUNS & BOATS
DOUCHEY HIPSTERS (DOUCHE-STERS)
BASICALLY MEXICO
RETIRED WHITES
LAS CUATRO MILPAS
SAN DIEGO, CALIFORNIA
MEXICO, LIKE FOR REALS

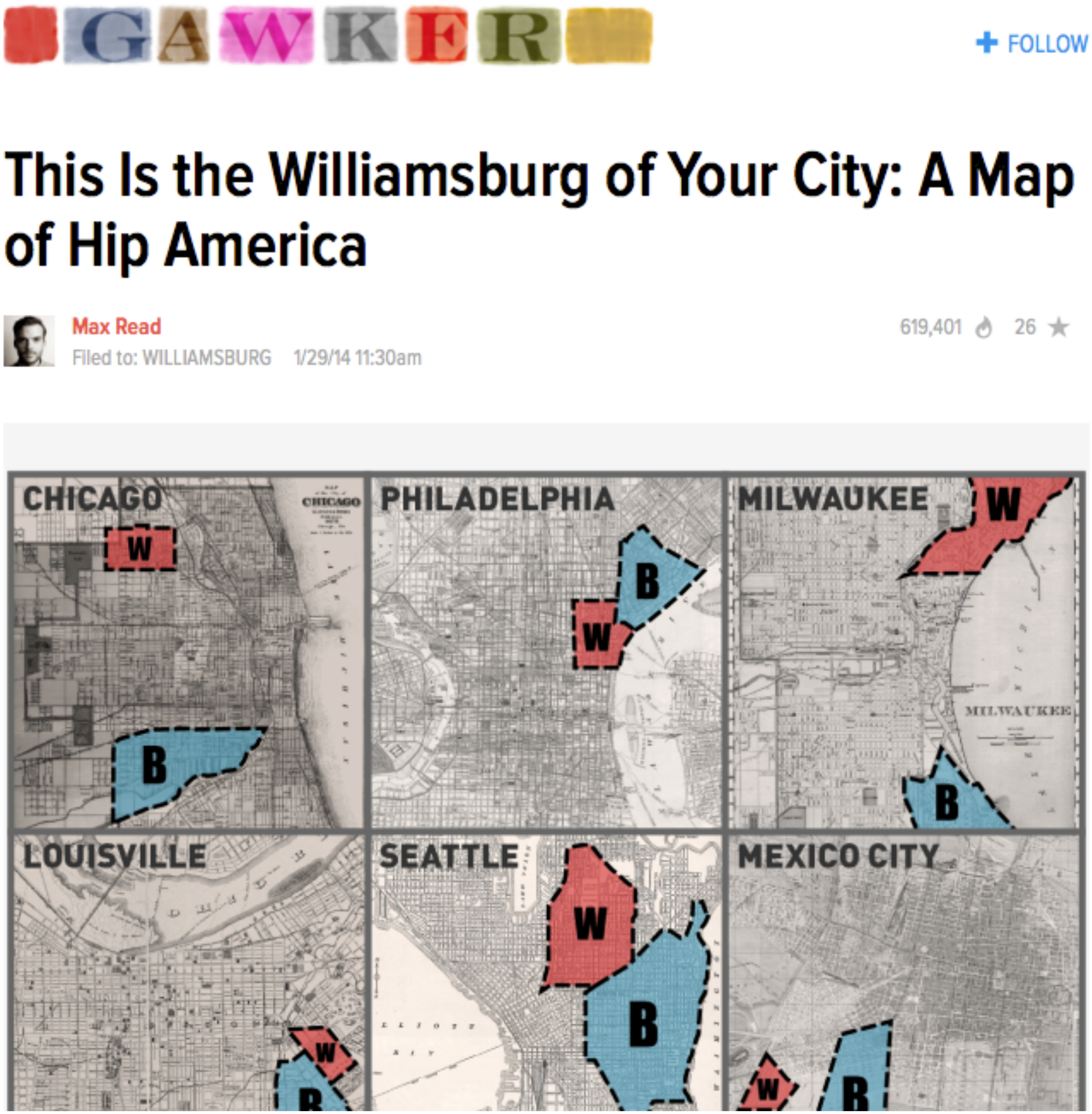
north park

```
hipster (0.201) craftsman (0.367) gentrified (0.381) flight path
(0.385) coffee shops (0.385) funky (0.402) artsy (0.418) cottage
(0.431) housing stock (0.456) iffy (0.458) walkable (0.459) gritty
(0.459) bungalow (0.462) urban (0.468) hip (0.482) main drag
(0.489) hipsters (0.521) mansions (0.529) apartment buildings
(0.532) pubs (0.538) charm (0.541) trendy (0.556) great
neighborhood (0.562) antique (0.569) walkability (0.576) great
areas (0.581) character (0.586) parts (0.595) eclectic (0.606) gay
(0.607) bars (0.607) tattoo (0.613) urban areas (0.618) pricier
(0.618) shops (0.621) gentrification (0.626) counts (0.627)
sketchy (0.628) cottages (0.630) particularly (0.634) coffee shop
(0.634) beach communities (0.635) upscale (0.636) blocks (0.638)
urban neighborhoods (0.650) congested (0.652) bungalows (0.654)
small city (0.655) vibe (0.662) charming (0.664) neighboring
(0.666) reached (0.670) northern part (0.673) hill (0.673) urban
core (0.674) downside (0.675) rental budget (0.675)
```

clairemont mesa

centrally (0.389) mesa college (0.459) apartment complexes (0.472) shopping centers (0.493) single family homes (0.512) supermarkets (0.512) located (0.519) supermarket (0.523) easy access (0.528) shopping malls (0.551) near (0.561) zip (0.562) branch (0.572) min drive (0.582) albertsons (0.608) apartments (0.609) good neighborhoods (0.614) branches (0.617) military housing (0.618) home depot (0.619) close (0.638) only one (0.640) classifieds (0.640) quiet (0.650) campus (0.656) henry (0.658) short commute (0.658) nasty (0.661) newly (0.661) complex (0.662) rush hour (0.663) short drive (0.663) repair (0.663) shopping center (0.665) pricey (0.669) condo complex (0.671) apts (0.671) item (0.680) recommended (0.682) centers (0.682) congestion (0.683) roommate (0.685) nearby (0.685) stores (0.686) parkway (0.688) ins (0.690) hey everyone (0.690) good area (0.691) pricier (0.691) mcdonalds (0.693) nicest (0.693)


la jolla village
torrey pines
la jolla
del mar
bonsall
vista
escondido
san marcos
ramona
riverside
murrieta
valley center
fallbrook
temecula
** sherman heights **
burlingame
golden hill
bankers hill
kensington
normal heights
north park
south park
university heights
mission hills
hillcrest
downtown san die
east village
downtown
little italy
gaslamp
old town
balboa park
san diego
otay mesa
san ysidro
tijuana
imperial beach
coronado
national city
bonita
eastlake
chula vista
kearny mesa
mission valley
tierrasanta
clairemont mesa
university city
bay ho
clairemont
bay park
scripps ranch
carmel mountain
rancho bernardo
poway
sabre springs
mel mountain ranch
ancho penasquitos
miramar
sorrento valley
mira mesa
sorrento mesa
la mesa
grossmont
lakeside
santee
el cajon
rancho san diego
el cerrito
city heights
college area
rolando
talmadge
lemon grove
spring valley
serra mesa
allied gardens
del cerro
san carlos
paradise hills
linda vista
lincoln park
lincoln
barrio logan
logan heights
encanto
skyline
rancho santa fe
del mar heights
carmel valley
la costa
solana beach
leucadia
point loma
sunset cliffs
ocean beach
mission beach
pacific beach



GAWKER

+ FOLLOW

This Is the Williamsburg of Your City: A Map of Hip America

Max Read

Filed to: WILLIAMSBURG 1/29/14 11:30am

619,401 26

	San Diego	**San Francisco**
hipster	south park	mission
dim sum	kearny mesa	chinatown
affordable	(temecula)	gilroy
pricey	del mar	pleasant hill
dangerous	logan heights	tenderloin
tourist	old town	union square
condos	downtown	soma
gay	hillcrest	(castro)
good schools	carmel mountain ranch	san ramon
hiking	ocean beach	presidio
white collar	sorrento valley	walnut creek
flight path	bankers hill	(burlingame)

Prospects

Big data linguistic techniques applied to broad spectrum texts allow us to extract real-world intelligence from ‘unstructured’ data

When applied to more focused corpora, they yield insights about speakers that would not be accessible via traditional qualitative methods

Hybrid quantitative / qualitative methods