

Word spaces and temperature words

Magnus Sahlgren

March 20, 2010

Word usage

How are words **really** used?



Distributional semantics

The distributional hypothesis: Words with similar distributional properties have similar meanings

(cf. Zellig Harris, Wittgenstein, Firth)

Distributional semantics

But which distributional properties and what kind of meanings?

- Words that tend to co-occur have a **syntagmatic** relation
- Words that tend to co-occur with the same *other* words have a **paradigmatic** relation
- Words that tend to occur in the same texts have an **associative** relation

Word spaces

How do we collect, represent, and compare distributional profiles?

Word spaces

The co-occurrence matrix:

	context ₁	context ₂	context ₃	...	context _n
word ₁	freq ₁	freq ₂	freq ₃	...	freq _n
word ₂	freq ₁	freq ₂	freq ₃	...	freq _n
word ₃	freq ₁	freq ₂	freq ₃	...	freq _n
...
word _m	freq ₁	freq ₂	freq ₃	...	freq _n

Word spaces

The co-occurrence matrix:

	context ₁	context ₂	context ₃	...	context _n
word ₁	freq ₁	freq ₂	freq ₃	...	freq _n
word ₂	freq ₁	freq ₂	freq ₃	...	freq _n
word ₃	freq ₁	freq ₂	freq ₃	...	freq _n
...
word _m	freq ₁	freq ₂	freq ₃	...	freq _n

Each row is a **context vector**

Word spaces

Word spaces enable us to **compute** similarity between words

The choice of context determines the type of relations

Word spaces

The co-occurrence matrix:

	document ₁	document ₂	document ₃	...	document _n
word ₁	freq ₁	freq ₂	freq ₃	...	freq _n
word ₂	freq ₁	freq ₂	freq ₃	...	freq _n
word ₃	freq ₁	freq ₂	freq ₃	...	freq _n
...
word _m	freq ₁	freq ₂	freq ₃	...	freq _n

Associative relations if the contexts are documents

Word spaces

The co-occurrence matrix:

	word ₁	word ₂	word ₃	...	word _n
word ₁	freq ₁	freq ₂	freq ₃	...	freq _n
word ₂	freq ₁	freq ₂	freq ₃	...	freq _n
word ₃	freq ₁	freq ₂	freq ₃	...	freq _n
...
word _m	freq ₁	freq ₂	freq ₃	...	freq _n

Paradigmatic relations if the contexts are words

Word spaces

The co-occurrence matrix:

	word ₁	word ₂	word ₃	...	word _n
word ₁	freq ₁	freq ₂	freq ₃	...	freq _n
word ₂	freq ₁	freq ₂	freq ₃	...	freq _n
word ₃	freq ₁	freq ₂	freq ₃	...	freq _n
...
word _m	freq ₁	freq ₂	freq ₃	...	freq _n

Syntagmatic relations if we look at the individual frequency counts

Word spaces

Descriptive — only what is **really there** in the data

Automatic, scalable, efficient

What is the alternative?

The alternative



Word spaces for corpus analysis

Local vs. global analysis

Paradigmatic, associative, syntagmatic

Comparing genres

Are temperature words used differently in different text genres?

Data

BNC: balanced, 44M words

Reuters: newswire, 214M words

Spinn3r: blogs, 473M words

Paradigmatic similarity

	BNC	Reuters	Spinn3r
hot	boiling	warm	castoff
	distilled	inclement	bombsight
	brackish	wintry	warm
	drinking	changeable	scald
	cold	mild	bottled
	semer	cool	steamy
	soapy	dry	soapy
	warm	unreasonable	muddy
	shallow	balmy	serendipity
	tepid	unseasonally	shop

Paradigmatic similarity

	BNC	Reuters	Spinn3r
cold	hot	inclement	cream
	franco-prussian	mild	cube
	boer	warm	rink
	iran-iraq	wintry	floe
	napoleonic	changeable	skating
	outbreak	cool	berg
	russo-japanese	dry	lolly
	soapy	waging	sundae
	warm	balmy	icepack
	punic	frigid	cone

Paradigmatic similarity

	BNC	Reuters	Spinn3r
chilly	warm	warm	balmy
	cold	cool	stormy
	cool	mild	tomorrow
	frosty	inclement	wintry
	hot	frigid	warm
	chill	changeable	dreary
	balmy	wintry	drizzly
	stomry	balmy	rainy
	wintry	dry	fateful
	foggy	seasonable	blustery

Paradigmatic similarity

	BNC	Reuters	Spinn3r
cool	warm	warm	neat
	hot	mild	darn
	clean	frigid	awesome
	soft	inclement	boring
	cold	changeable	much
	fresh	wintry	excite
	calm	balmy	nice
	gulp	chilly	nifty
	quiet	dry	cute
	chilly	seasonable	scary

Associative similarity

	BNC	Reuters	Spinn3r
hot	heat	h	sexy
	water	creditcapital	sex
	cold	b	porn
	boiler	rolled	videos
	cook	aramark	nude
	temperature	amb	movie
	butter	daew	clips
	stealthily	briquetted	adult
	gravity	guj	ranks
	go	sski	klipler

Associative similarity

	BNC	Reuters	Spinn3r
cold	glass	weather	wound
	goldberg	temperature	bone
	harsnet	somar	actinomyces
	boxing-match	winter	bones
	oculist	meteorologist	limb
	write	heating	adherent
	actaeon	damaging	fomentations
	slimness	fahrenheit	granulations
	ovid	forecaster	phagocytes
	kafka	war	gangrene

Syntagmatic similarity

	BNC	Reuters	Spinn3r
_ hot	too -	red -	shop hydrophobic
hot _	water cold air summer weather -	weather dry weight rolled carcass topic	water dog tub spot - -

Syntagmatic similarity

	BNC	Reuters	Spinn3r
_ cold	hot bitterly	- -	ice blow
cold _	war water weather air wet politique	weather war - - - -	war air weather - - -

Syntagmatic similarity

	BNC	Reuters	Spinn3r
_ cool	keep allow stay -	- - - -	pretty liquid really very
cool _	air water drink ground down dark breeze	weather temperature - - - - -	stuff modern stroke - - - -

Word spaces and temperature

- Domain specific similarities (“hot”/“castoff”)
- Constructions (“hot carcass weight”)
- Clear differences between genres