Studio Mind Oy

CEO, Pasi Karhu

Suurten tekstimassojen hyödyntäminen kontekstitietokantojen luonnissa
Utilizing large text masses (corpora) in creation of context databases

Kites Symposium 30.10.2014

What is Context?

- Merriam-Webster online dictionary:

  : the words that are used with a certain word or phrase and that help to explain its meaning

  : the situation in which something happens
  : the group of conditions that exist where and when something happens

- Linguistic context
- Social context

What is Context?

- My pragmatic approach in computer linguistics:

  : the words that are used with a certain word or phrase

- Statistical analysis of large bodies of text capture this
- Applying this statistics cleverly we can get computers to help to explain its meaning
- It is plausible to think that our brain does this kind of linguistical “statistics” (gathered from all our past experiences)

Why should computers know about context?

- Without context awareness computers are stuck with the simpler but often not adequate understanding through keywords, synonym lists and thesauri
- Lexical disambiguation of polysemy, homonymy and synonymy need context awareness
- Proper context awareness opens road to advanced application like natural language communication with computers, more intelligent searches and better language translations

Does Google do context?

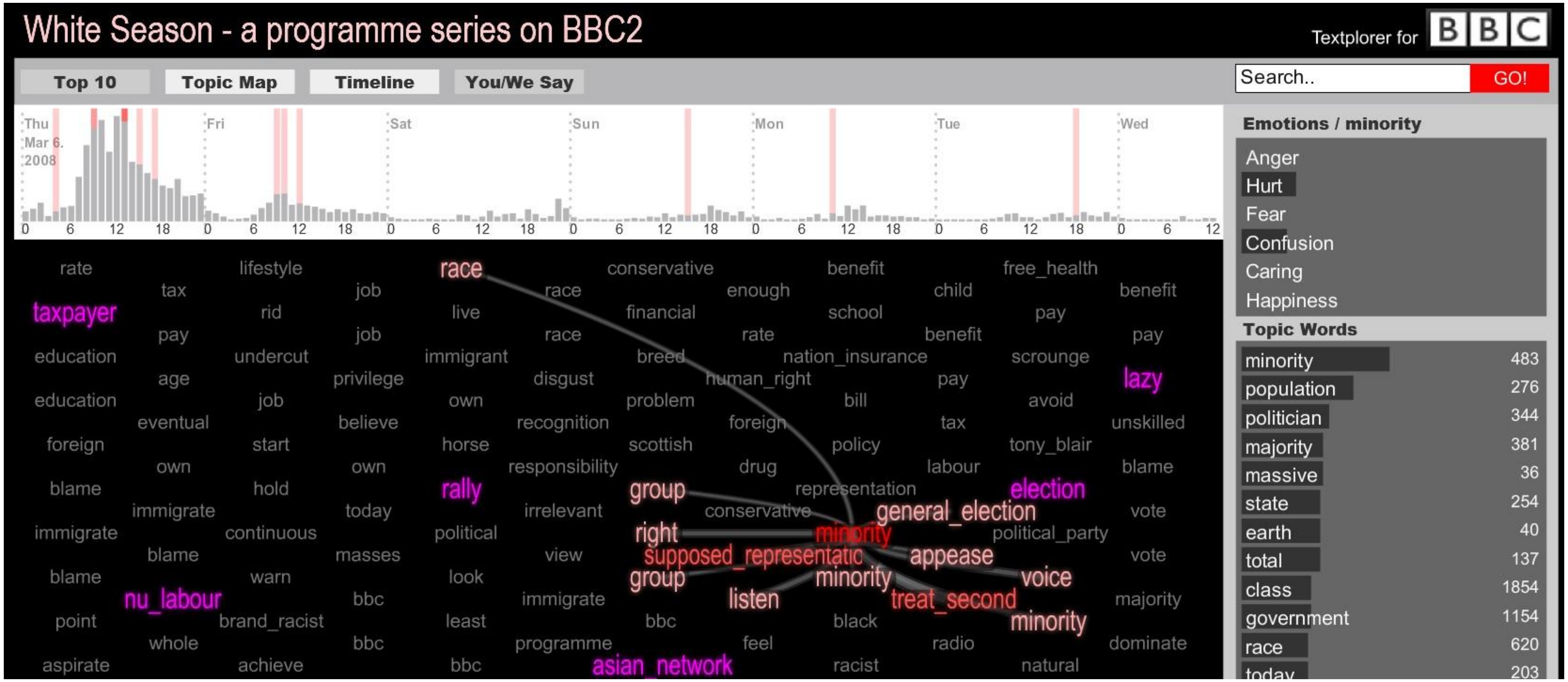
- They have huge textual (and research) resources!
- Yet they don’t offer contextual search
- What about Google translate:
  “First World War shell explodes at former Ypres battlefield”
  => “Ensimmäisen maailmansodan kuori räjähtää entisillä Ypres taistelukenttä”
  ”Kiltti on miehen hame Skotlannissa”
  => ”Kind is a man's skirt in Scotland”
  “Kiltti hame Skotlannissa”
  => “Kilt Skirt in Scotland”

Ontologies vs Statistical Context Databases?

- Ontologies are laborious to create and maintain
- Ontologies may have human bias
- Context info from large texts can be automatically created and maintained
- Statistical analysis of large text masses of interest captures all contextual relations of words as they are actually used in the texts (no human bias)
- Ontologies are more formal and precise
- Statistical context database are fuzzy by nature
- Maybe best results by combining both approaches!?

How much text do you need?

- Some results already with 3.5 MB – example from 7000 comments online (BBC – Have Your Say)



White Season - a programme series on BBC2

Textplorer for BBC

Emotions / minority

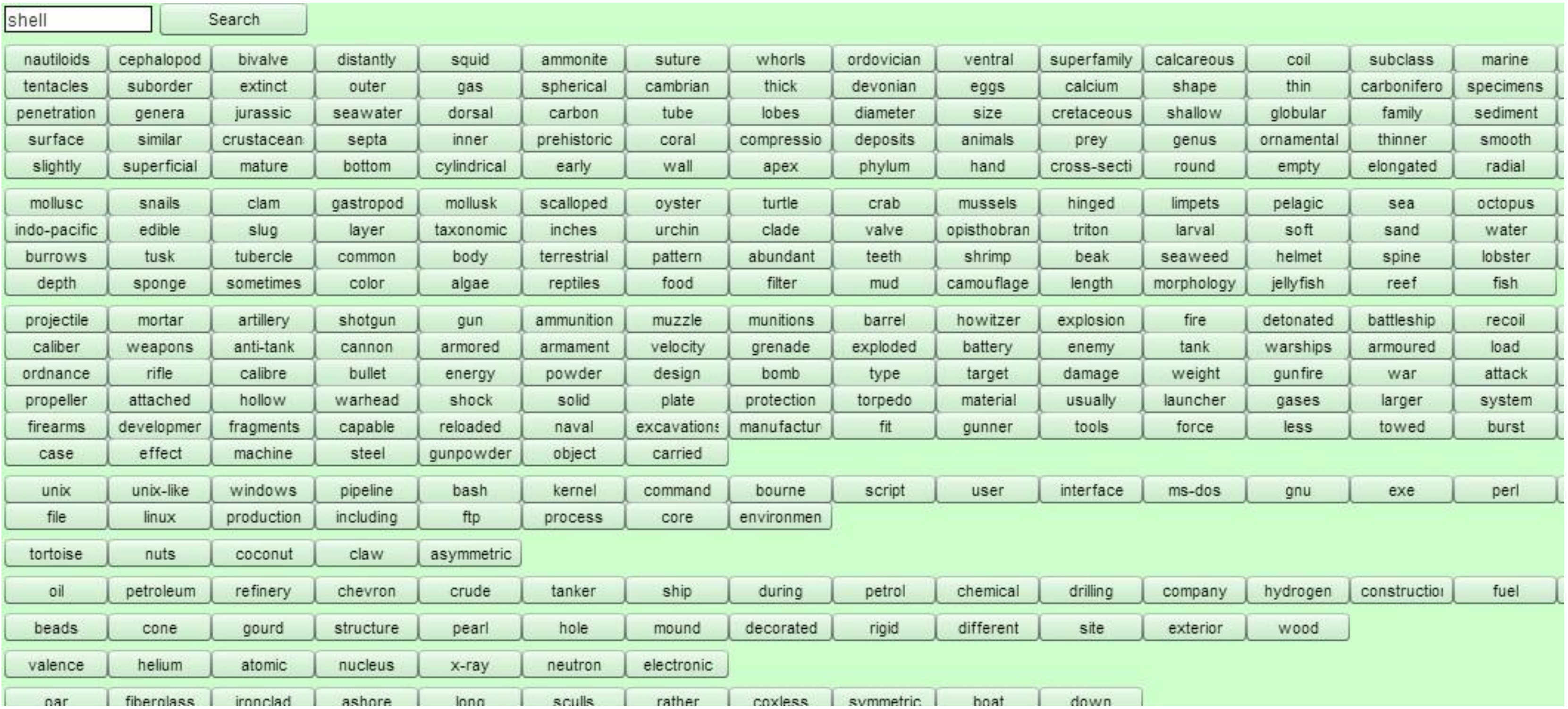
Anger
Hurt
Fear
Confusion
Caring
Happiness

Topic Words

minority	483
population	276
politician	344
majority	381
massive	36
state	254
earth	40
total	137
class	1854
government	1154
race	620
today	203

How much text do you need?

- With Wikipedia content (many GB) you can build a large generic context database (demo: 22 000 words)



shell

nautiloids	cephalopod	bivalve	distantly	squid	ammonite	suture	whorls	ordovician	ventral	superfamily	calcareous	coil	subclass	marine
tentacles	suborder	extinct	outer	gas	spherical	cambrian	thick	devonian	eggs	calcium	shape	thin	carbonifero	specimens
penetration	genera	jurassic	seawater	dorsal	carbon	tube	lobes	diameter	size	cretaceous	shallow	globular	family	sediment
surface	similar	crustacean	septa	inner	prehistoric	coral	compressio	deposits	animals	prey	genus	ornamental	thinner	smooth
slightly	superficial	mature	bottom	cylindrical	early	wall	apex	phylum	hand	cross-secti	round	empty	elongated	radial

mollusc	snails	clam	gastropod	mollusk	scalloped	oyster	turtle	crab	mussels	hinged	limpets	pelagic	sea	octopus
indo-pacific	edible	slug	layer	taxonomic	inches	urchin	clade	valve	opisthobran	triton	larval	soft	sand	water
burrows	tusk	tubercle	common	body	terrestrial	pattern	abundant	teeth	shrimp	beak	seaweed	helmet	spine	lobster
depth	sponge	sometimes	color	algae	reptiles	food	filter	mud	camouflage	length	morphology	jellyfish	reef	fish

projectile	mortar	artillery	shotgun	gun	ammunition	muzzle	munitions	barrel	howitzer	explosion	fire	detonated	battleship	recoil
caliber	weapons	anti-tank	cannon	armored	armament	velocity	grenade	exploded	battery	enemy	tank	warships	armoured	load
ordnance	rifle	calibre	bullet	energy	powder	design	bomb	type	target	damage	weight	gunfire	war	attack
propeller	attached	hollow	warhead	shock	solid	plate	protection	torpedo	material	usually	launcher	gases	larger	system
firearms	developmer	fragments	capable	reloaded	naval	excavations	manufactur	fit	gunner	tools	force	less	towed	burst
case	effect	machine	steel	gunpowder	object	carried								

unix	unix-like	windows	pipeline	bash	kernel	command	bourne	script	user	interface	ms-dos	gnu	exe	perl
file	linux	production	including	ftp	process	core	environmen							

tortoise	nuts	coconut	claw	asymmetric

oil	petroleum	refinery	chevron	crude	tanker	ship	during	petrol	chemical	drilling	company	hydrogen	constructio	fuel

beads	cone	gourd	structure	pearl	hole	mound	decorated	rigid	different	site	exterior	wood

valence	helium	atomic	nucleus	x-ray	neutron	electronic

oar	fiberglass	ironclad	ashore	long	sculls	rather	coxless	symmetric	boat	down

Some things you can do with a context database:

- Identify most important words, sentences and paragraphs in document
  => automatic tagging (e.g. for Semantic Web)
  => creating headlines and “readers digest”
- Expand search terms
- Find correct senses of words (e.g. for translations)
- Detect different topics in a document collection or inside one document
- Classify, group and help visualize documents according to their context
- …

Do we need more than linguistic context?

Shakespeare’s Sonnet 18:

“ Shall I Compare Thee to a Summer's Day? ”

THANK YOU FOR YOUR ATTENTION!
QUESTIONS?

Should you need context awareness in your projects, you can reach me at:

pasi.karhu@studiomind.fi

+358 50 377 4875

Cooperation with other language technology companies also welcomed