

DIACRAN II

Ondřej Herman, Miloš Jakubíček



NLP Centre, Masaryk University,
Brno, CZ

Portoroz, May 23, 2016



Outline

1 DIACRAN I

2 DIACRAN II

3 Conclusions

Diachronic analysis

Main goal: neologism finding → lexicography

Neologisms:

- new lexemes
- new senses



Diachronic analysis

Main goal: neologism finding → lexicography

Neologisms:

- new lexemes – easy bits
- new senses – hard bits



Diachronic analysis

Needed:

- data
- algorithms

Output:

- trend
- significance

Neologisms: data

Corpora with accurate time annotation are a scarce resource

- COCA (© Mark Davies)
- BNC (but ...)
- in-house data (e.g. OEC)
- FeedsCorpus (2008–2014)
 - RSS feeds
 - so far English only, others to follow
- JSI Feeds Corpus (thanks to Simon Krek)

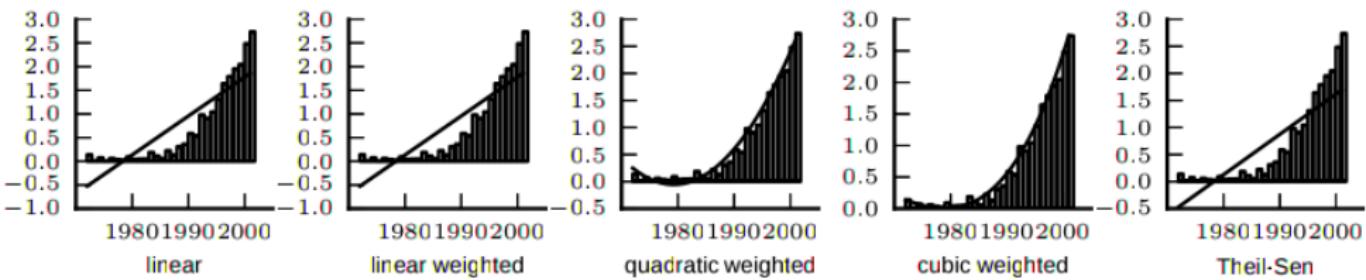
Neologisms: algorithms

- linear regression (and its variations)
- Mann-Kendall / Theil-Sen

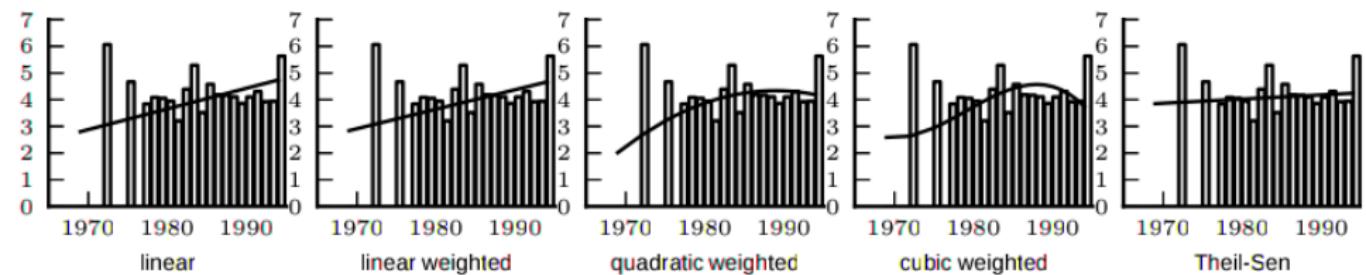
Data much more important than algorithms.



Google ngrams: *globalization*



British National Corpus: *tree*



new configuration directive: DIACHRONIC "doc.year, doc.month"

Top 1,000 trends sorted by trend

Corpus: **Corpus of Contemporary American English (COCA)** Method: **Mann-Kendall, Theil-Sen (all)**

Show 25 ▾ items

Word	Trend	p-value	Freq	Graph
blog	3.4874 +	0.000000	1.976	
blogged	3.4874 +	0.000002	53	
SODIUM	3.2708 +	0.000002	993	
website	3.2708 +	0.000000	4.460	
dcor	3.2708 +	0.000002	180	
bloggers	3.2708 +	0.000002	402	
jalapeo	3.2708 +	0.000001	275	
earbuds	3.2708 +	0.000001	94	
Googling	3.2708 +	0.000001	54	
blogger	3.2708 +	0.000000	366	
Googled	3.2708 +	0.000004	79	
Crme	3.2708 +	0.000001	132	
modele	-3.0776 -	0.000007	18	

Evaluation

- work in progress
- neologism data obtained from major UK publishing houses



Current work

- neologisms: new lexical items vs. new senses
- so far: new lexical items
- to be continued with: new senses
- new senses = new contexts \Rightarrow word sketches as input to regression



Trending Word Sketch triples, $p < 0.1$

head	rel	coll	slope	p	tot
Katrina-e	X of-i N*	handling-n	0.417	0.011	31
Parnassus-e	X of-i N*	Imaginarium-e	0.415	0.010	35
al-Maliki-e	X mod N*	prime-j	0.413	0.019	65
Parish-e	X mod N*	Bernard-e	0.403	0.026	140
levee-n	N* subj V	fail-v	0.402	0.011	36
Imaginarium-e	N* PREP	of-i	0.393	0.010	38
Adonis-e	X mod N*	Lord-e	0.392	0.011	88
Alphonso-e	X* mod N	Jackson-e	0.390	0.040	36
Cornwall-e	X of-i N*	Duchess-e	0.389	0.026	324
Sunni-e	X* mod N	party-n	0.389	0.032	102

Trending Word Sketch triples, named entities filtered out

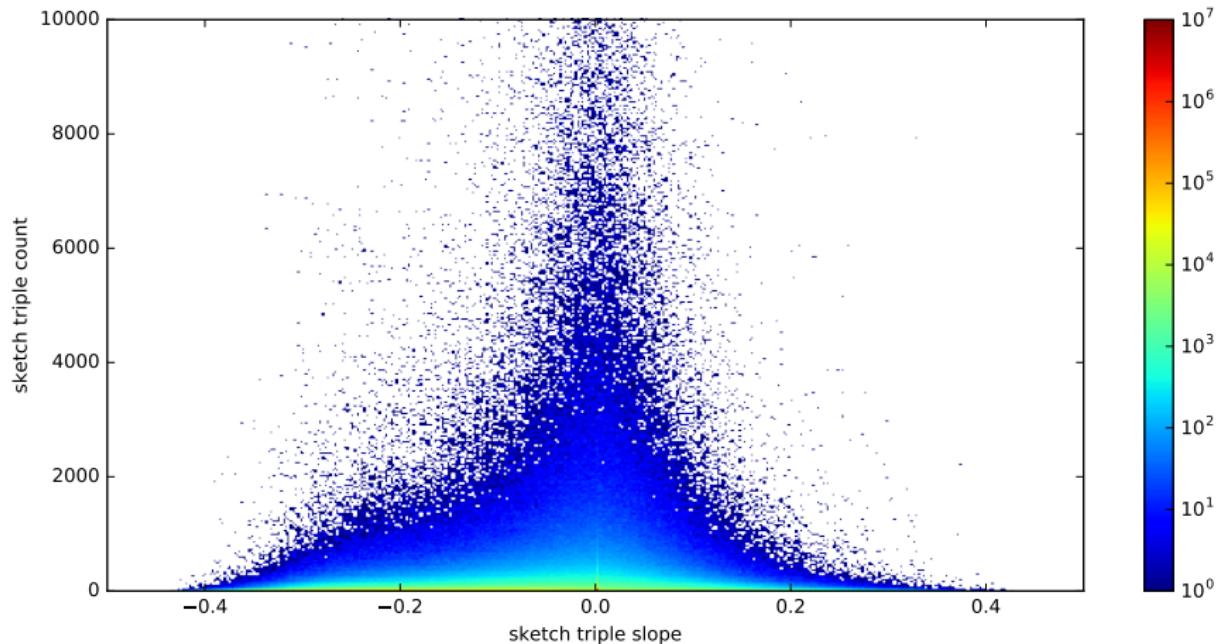
head	rel	coll	slope	p
levee-n	N* subj V	fail-v	0.402	0.011
phone-n	X* mod N	evidence-n	0.388	0.054
deadly-j	X* mod N	school-n	0.383	0.063
Syrian-j	X* mod N	security-n	0.377	0.034
war-n	X mod N*	gangland-n	0.372	0.041
oil-for-food-j	X* mod N	scandal-n	0.372	0.032
century-n	X mod N*	st-n	0.369	0.012
pandemic-n	X of-i N*	fear-n	0.368	0.007
root-v	V* N out-r	insurgent-n	0.367	0.034
black-j	X* mod N	abaya-n	0.366	0.008
evidence-n	X mod N*	phone-n	0.365	0.054
break-v	X* mod N	immigration-n	0.364	0.005
tribal-j	X* mod N	sheik-n	0.363	0.034
strategy-n	X mod N*	counterterrorism-n	0.361	0.095
search-n	X* mod N	giant-n	0.357	0.039

Trending Word Sketch triples, named entities filtered out

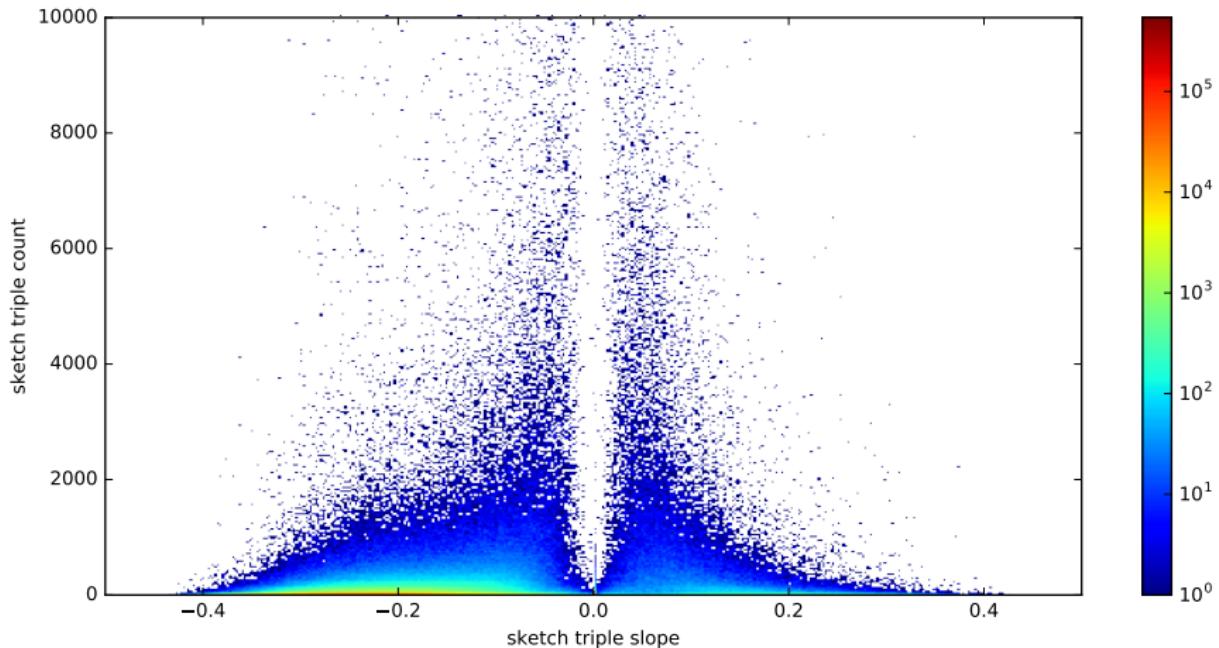
Trending triples with $p < 0.1$ and adjective-noun phrases

head	rel	coll	slope	p
deadly-j	X* mod N	school-n	0.383	0.063
Syrian-j	X* mod N	security-n	0.377	0.034
oil-for-food-j	X* mod N	scandal-n	0.372	0.032
black-j	X* mod N	abaya-n	0.366	0.008
tribal-j	X* mod N	sheik-n	0.363	0.034
video-sharing-j	X* mod N	website-n	0.349	0.024
big-j	X* mod N	reveal-n	0.348	0.054
self-governing-j	X* mod N	island-n	0.346	0.034
adjacent-j	X* mod N	lot-n	0.341	0.012
standard-j	X* mod N	mileage-n	0.334	0.063
extra-j	X* mod N	growth-n	0.332	0.059
gold-j	X* mod N	bull-n	0.329	0.095
trillion-dollar-j	X* mod N	deficit-n	0.327	0.054
artificial-j	X* mod N	deadline-n	0.327	0.041

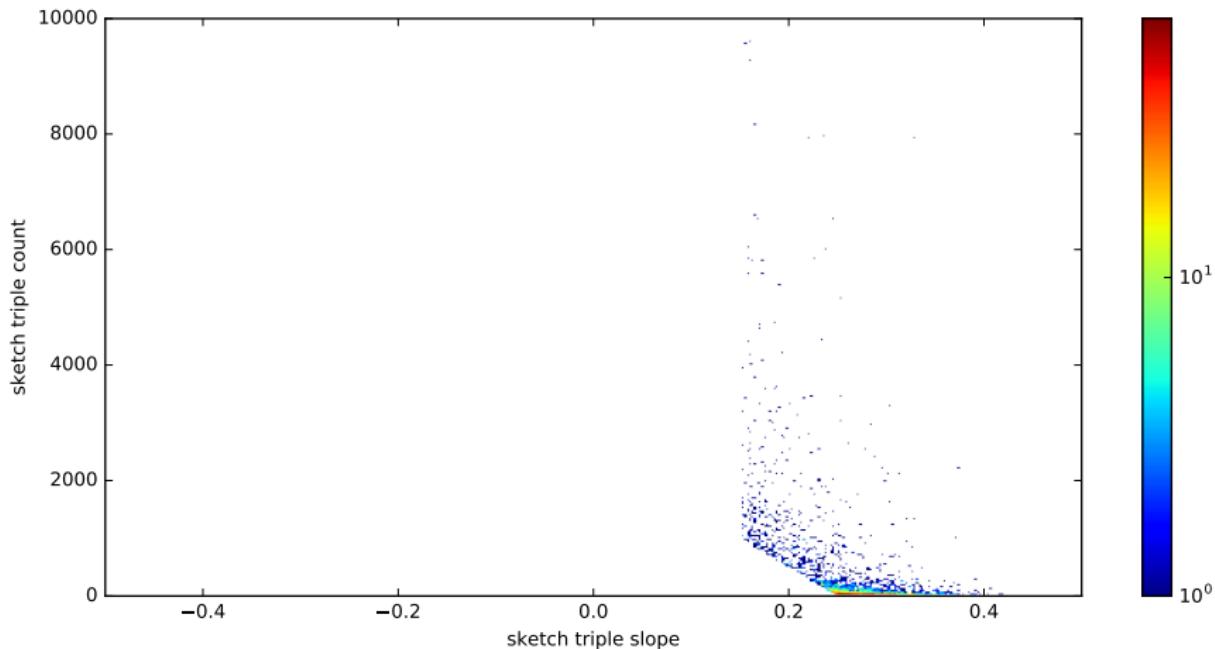
Distribution of word sketch triple count by slope



Distribution of word sketch triple count by slope, $p < 0.1$



Selected word sketch triples



head	rel	coll	slope
oil-for-food-j	X* mod N	scandal-n	0.372
century-n	X mod N*	st-n	0.369
tribal-j	X* mod N	sheik-n	0.363
strategy-n	X mod N*	counterterrorism-n	0.361
search-n	X* mod N	giant-n	0.357
record-n	X on-i N*	level-n	0.353
oil-n	X mod N*	record-n	0.352
insurgency-n	V obj N*	fuel-v	0.351
video-sharing-j	X* mod N	website-n	0.349
big-j	X* mod N	reveal-n	0.348
download-n	X* mod N	chart-n	0.347
join-v	V* obj N	insurgency-n	0.344
immigration-n	X* mod N	enforcement-n	0.342
ban-n	X mod N*	same-sex-j	0.340
author-n	X* of-i N	blog-n	0.337
rent-v	V* N out-r	home-n	0.337
dot-n	X mod N*	gmail-n	0.335
condone.v	V* obj N	torture-n	0.334

Merged with triples – rising trend

head	rel	coll	slope	
hi-def-n	0.299	PREP N*	in-i	0.299
one-bed-j	0.259	X* mod N	flat-n	0.239
banshee-n	0.246	PREP N*	like-i	0.145
micro-n	0.235	and/or	macro-n	0.159
micro-n	0.235	X* mod N	loan-n	0.205
nom_de_plume-n	0.229	PREP N*	under-i	0.124
hydrogen-powered-j	0.227	X* mod N	car-n	0.282
dumbness-n	0.227	PREP N*	of-i	0.155
sweptback-j	0.224	X* mod N	hair-n	0.181
constructionist-n	0.221	X mod N*	strict-j	0.183
roadless-j	0.219	X* mod N	rule-n	0.315
stop-work-n	0.215	X* mod N	order-n	0.159
walk-off-n	0.214	and/or	home-n	0.236
walk-off-n	0.214	X* mod N	home-n	0.236
walk-off-n	0.214	X* mod N	run-n	0.236

Merged triples – falling trend

head	rel	coll	slope	
on-topic-j	0.346	and/or	informative-j	-0.343
on-topic-j	0.346	V ADJ*	stay-v	-0.348
social-science-n	0.332	X* mod N	research-n	-0.332
chuan-n	0.326	X mod N*	chi-n	-0.326
derivatized-v	0.324	V* PREP	with-i	-0.324
non-directive-j	0.323	X* mod N	counseling-n	-0.344
map-based-j	0.321	X* mod N	cloning-n	-0.347
relicensure-n	0.321	PREP N*	for-i	-0.307
relicensure-n	0.321	X for-i N*	activity-n	-0.337
whitehead-n	0.318	and/or	blackhead-n	-0.318
work-place-n	0.313	PREP N*	in-i	-0.299
kohanga-n	0.309	X* mod N	reo-n	-0.354
alt.-j	0.309	X* mod N	country-n	-0.309
dyspeptic-j	0.307	X* mod N	symptom-n	-0.399
dyspeptic-j	0.307	X* mod N	patient-n	-0.231

Conclusions

- diachronic analysis to become part of Sketch Engine
- data more important than algorithms
- ongoing work on new sense detection

