

First International Sketch Grammar Workshop

Ljubljana

3-4 February 2010

Workshop goals

- (as I see them)
- Share grammar-writing experience
- Feedback to LCL
- LCL tells you
 - Other possibilities
 - What is in the pipeline

Apologies

□ Masha Kholkova, Carole Tiberius

LCL projects and plans

- Corpora
 - Corpus Factory
 - English: bigger and better
- Corpus NLP with remote corpora
 - Web-API use of SkE
- Far horizons
 - From text towards meaning
- **Tomorrow**
 - SkE Interface, extra functionality
 - Formalism (Pavel)

Corpus Factory

☐ Goal

- All medium-large world lgs
 - ☐ All EU languages
 - ☐ About 100
- 100m word web corpus

☐ Hyderabad team

☐ Done

- Dutch
- Thai
- Vietnamese
- Hindi
 - ☐ but Indians mainly use English on web

☐ Earlier projects

- Greek
- Japanese

☐ Next

- Swedish
- Norwegian
- Korean

☐ Collab: WaCKY

- Bologna (Marco Baroni
 - ☐ German
 - ☐ Italian
- Leeds (Serge Sharoff)
 - ☐ Arabic Chinese Polish Russian French Spanish ...

BootCat method

- ❑ Wikipedia for the lg
 - Word freq list
 - ❑ Mid-freq words: seeds
 - ❑ Highest-freq words: use for filtering
- ❑ Queries of n words to search engine
- ❑ Clean, dedupe, filter
- ❑ Tokenise, POS-tag, lemmatise
- ❑ Load in SkE
 - Word Sketches

English

☐ Bigger

☐ Better

Bigger

□ Motivation

- Ample data for rare phenomena
- Big subcorpora
- For language modelling

□ More like Google-scale

- but without Google disadvantages
 - See *Googleology is Bad Science*, CL 2007

Better

- ❑ Less noise
- ❑ Fewer duplicates
- ❑ Richer markup
 - At word, sentence level
 - At document level (text type, subcorpora)

Divide and rule

- Bigger (+ cleaning + deduplication)
 - Big Web Corpus (BiWeC)
 - Currently 5.5b fully processed
 - Target 20b words
 - Jan Pomikalek, Pavel Rychly
- Better
 - New Model Corpus

New Model Corpus

□ model

1. small version: *model train*
2. design: *data model*

□ New Model Corpus

- 1:100 scale model
- To replace BNC as design model

BNC design model

- Most often used
 - Eg for other languages
- pre-web
 - $f(\text{blog})=0$
- Corpora now bigger, far quicker, far cheaper, different issues
- *BNC design model past its sell-by*
 - Kilgarriff Atkins Rundell, Corpus Lg 2007

New model

- ☐ Data
- ☐ Markup

Data

- From the web
- 100m words
- Small sample size
 - Copyright
 - ??Creative Commons Licence

Composition

<input type="checkbox"/> General crawl	50
<input type="checkbox"/> Targeted	
<input type="checkbox"/> Fiction	7
<input type="checkbox"/> Blog	7
<input type="checkbox"/> Newspaper (RSS feed)	7
<input type="checkbox"/> Speech	10
<input type="checkbox"/> Film transcripts, chatshow	
<input type="checkbox"/> Domain-specific	19
<input type="checkbox"/> Business, medical, law	

Markup

□ Collaborative

- We distribute data
- Anyone applies their tools
 - Pos-tagger, parser, co-ref resolution, domain classifier, WSD, semantic classifier, time phrases, named entities...
- We integrate, display in Sketch Engine
- Research potential from multiple markup

Recombine the two strands

- Apply methods with good accuracy (and ***fast***) to BiWeC
- Result will be
 - ***Bigger***
 - ***Better***

Corpus NLP with Remote Corpora/ NLP by web services?

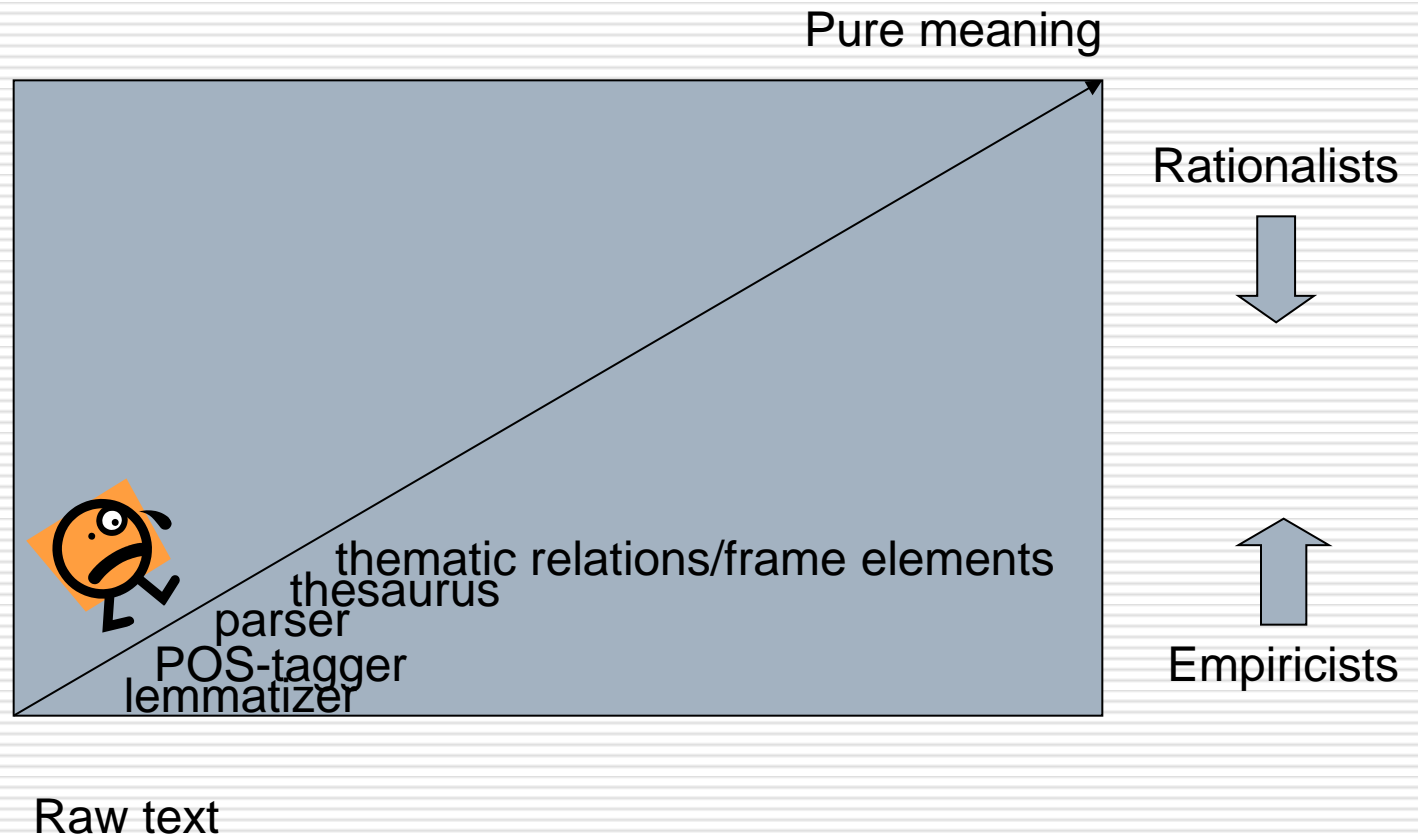
- Big corpora
 - big to hold, hard to access fast
- Sketch Engine: corpus specialist
- Web API
 - FrameNet
 - TEDDCLOG: Taiwan English Data Driven Cloze (test sentence) Generation
- All welcome

Practicalities

- Free trial accounts
- Collaborators, innovative users
 - free longer-term accounts
 - *Wikinomics*, Tapscott and Williams
- API
 - Details under 'help' on SkE home page
- New Model Corpus
 - Available soon: watch *Corpora*

Far horizons

The long journey from text towards meaning



Next steps

☐ Semantic tagging

- Extra positional attribute
- Use in Sketch Grammar patterns
 - ☐ English: Lancaster system
 - ☐ Russian: ABBYY system

☐ Learn

- Hanks: Corpus Pattern Analysis
- Melcuk Lexical Functions
- Frame semantics: frames

-- and WSD

- Semi-Automatic Dictionary Drafting
- SADD
 - Builds on WASPS
 - Shares CPA technology
- Senses as clusters of instances
 - 'one sense per collocate'
 - Shortcut: clusters of collocates
- In pictures

Clustered word sketch

<u>object</u>	<u>58698</u>	4.0
food <u>4972</u>	<u>11512</u>	8.22
fish <u>1156</u> anything <u>790</u> everything <u>271</u> animal <u>304</u> heart <u>293</u> plant <u>298</u> something <u>448</u> variety <u>238</u> nothing <u>247</u> pattern <u>189</u> word <u>217</u> thing <u>389</u> place <u>392</u> quality <u>213</u> product <u>224</u> day <u>367</u> way <u>270</u> area <u>234</u>		
disorder <u>2361</u>	<u>4752</u>	9.0
diet <u>1385</u> habit <u>1006</u>		
meal <u>1783</u>	<u>4334</u>	8.32
lunch <u>1046</u> breakfast <u>886</u> dinner <u>619</u>		

Init annotation

Home Concordance Word List Word Sketch Thesaurus Sketch-Diff View

Corpus: b
[conc desc](#)

Annotating: charge-v

Not assigned 6050 [P](#) / [N](#)

modifier: negatively, electrically, emotionally, *pp_with-p:* conspiracy, assault, *object:* vat, battery, *part_intrans:* in, *pp_for-p:* service, *part_trans:* up,
[word sketch](#)

crime 265 [P](#) / [N](#)

pp_with-p: murder, theft, offence, *and/or:* arrest, *subject:* magistrate, *modifier:* yesterday, today, *object:* magistrate, custody, court, [word sketch](#)

electric 44 [P](#) / [N](#)

modifier: positively, *object:* nucleus, proton, atom, [word sketch](#)

money 444 [P](#) / [N](#)

object: fee, sum, person, *pp_with-p:* offence, crime, *subject:* Girobank, *pp_to-p:* profit, *pp_at-p:* rate, *modifier:* only, also, [word sketch](#)

Minscore: Mindiff:

Annotating: **charge-v** New label: Add [Info](#) [Finish](#)

object	2755	4.5	subject	685	3.5	modifier	1043	4.1	and/or	62	-1.4	pp_with-p
vat money	27	7.13	indictment	8	7.53	negatively	34	9.65	release	4	3.4	manslaughter
battery electric	26	7.08	Land	9	6.93	electrically	23	9.1	try	5	1.66	conspiracy
premium money	22	6.74	restructuring	9	6.92	emotionally	24	8.87				assault
defendant crime	25	6.52	lender money	8	6.86	highly	73	7.89	part_trans	54	3.0	treason
rent money	25	6.51	turbo electric	4	6.75	jointly	12	7.55	over	7	2.9	misconduct
magistrate crime	20	6.38	capacitor electric	4	6.74	politically	13	7.44	up	26	2.41	theft
price money	120	6.36	prosecutor crime	5	6.5	yesterday	47	7.43	down	9	2.09	burglary
atmosphere electric	27	6.29	Fa	6	6.04	formally	15	7.33				sedition
depreciation money	9	6.21	retailer money	5	5.81	up to	16	7.07	part_intrans	137	3.5	kidnapping
registry	10	6.2	bull	4	5.66	oppositely	5	7.07	around	9	4.44	rape
capacitor electric	8	6.14	critic	6	4.84	subsequently	15	7.05	through	5	4.29	robbery
suspect crime	10	6.08	landlord	4	4.55	headlong	5	6.98	in	26	4.11	fraud
particle electric	15	6.07	solicitor	6	4.33	today	23	6.85	off	14	3.44	indecenty
call money	36	5.96	police crime	8	4.01	incorrectly	4	6.55	down	18	3.09	arson
commission money	45	5.91	bank electric	9	3.98	fully	23	6.37	up	28	2.51	kidnap
ion electric	9	5.9	offence money	5	3.96	normally	14	6.31	on	11	2.47	incitement
shilling money	7	5.73	court u	7	3.96	later	20	6.16	back	9	2.18	obstruction
rate money	107	5.65	fee x	5	3.83	separately	5	6.15	out	7	0.74	harm
fare money	9	5.62	person Add next	1	2.92	under	7	6.09				possession
electron electric	8	5.58	horse None	4	2.8	specifically	7	5.94	unary_rels			count
toll money	7	5.5	firm	5	2.51	repeatedly	4	5.87	prep_ing	645	8.3	corruption

reasons in Rwanda. </p><p> They have not been charged		or tried and may be imprisoned because of
but for Amnesty members these events are charged		with significance. The Cold War was at
al-Nahda (Renaissance). </p><p> Hamadi Jebali was charged		with defamation of a judicial institution
Kosovo arrested in September 1988. They were charged		with having formed a 'hostile' organization
short books -- objects physically slight but charged		cavalry confusions. They are marvellously
grace notes, the pointing of sentences, charged		crime and buoyant, intelligent wit.
Zuckerman's father -- after which Henry charges		electric with killing their parent by writing
. The last three months particularly are charged		money icity -- and you are suddenly
Goldberg called again the day after that, charged		u ind out. What do you mean abandoned
VAT. </p><p> The back dating of all VAT not charged		x acked lunches could lead to hefty bills
once again come under fire over the amounts they charge		Add next one calls. </p><p> An English Tourist
reports of a visitor from the USA being charged		None e than £200 for a 45-minute telephone
are now regular customers. </p><p> We don't charge		extra for coffee, service or bread, which
the supplier's own premises are typically charged		on a daily delegate rate structured like
delivery note to check that Brown is being charged		for the right goods. </p><p> As each item
others say widely differing prices are being charged		for exactly the same product and that suppliers
bistro food in designer surroundings -- and charging		the earth for it. </p><p> How many customers
management of the ingredients budget -- it simply charges	money	a management fee for the entire contract
glared at Conroy without seeing him and charged		back into No. 4, slamming the door. </p>
the big one. Ironically, the hostel was charging		so much in rent that while I did have to

LCL projects and plans

- Corpora
 - Many languages
 - English: bigger and better
- Corpus NLP with remote corpora
 - Web-API use of SkE
- Far horizons
 - From text towards meaning
- **Tomorrow**
 - SkE Interface, extra functionality
 - Subcorpora / text types
 - Formalism (Pavel)