SOFTWARE AND TOOLS FOR CORPUS PATTERN ANALYSIS

Vít Baisa, Ismaïl El Maarouf, Adam Rambousek, Pavel Rychlý

OUTLINE

- Corpus Pattern Analysis
- Annotation in Sketch Engine
- CPA editor
- Public access
- SemEval 2015
- LEMON API

INTRODUCTION

- tools and datasets
- to support Pattern Dictionary of English Verbs (PDEV)
- since 2006
- DVC project 2012–2015

CPA

- associating word meaning with word use by an analysis of phraseological patterns and collocations
- meaning is associated with prototypical sentence contexts
- concordance lines are grouped into semantically motivated syntagmatic patterns
- hard problem: granularity
- Subj, Obj, Complement, Adverbial, Indirect Obj
- British National Corpus (written part)

CPA II

- determiners: *take place* vs. *take his place*
- semantic types: `build [[Machine]]` vs. `build [[Relationship]]`
- contextual roles: `[[Human = Director]] shoot` vs. `[[Human = Sports Player]] shoot`
- lexical sets: `reap {the whirlwind}` vs. `reap {the harvest}`

ANNOTATION IN SKETCH ENGINE

- not so well-known feature of SkE
- no paper published (until now)
- not documented :)

```
commit 7b682e7473d935b14f48b7b5352838318bfda523
Author: pary
Date:   Sat Sep 2 22:34:10 2006 +0000

    [bonito2 @ 2006-09-02 22:34:10 by pary]
    added linegroup/annotconc
```

Info Sort by label Finish | Filter Wsketch Reveal Bootstrap | Batch annotation: selection (0 lines selected)

Query **abandon-v** **250** > Random sample **250** (4.6 per million)

Page 1 of 13 Go Next | Last

the high-minded Virginsky who ` will never , never	abandon	2	these bright hopes ' (my italics) , and another	C
Labour did not agree that Britain could or should	abandon	1	development , either for itself or for the developing	C
minority tribes into the governance of the state . He has	abandoned	4.f	much of his Marxist baggage and , so far , set his	C
morning for a briefing session . They say they will	abandon	1	the dispute over the Government 's refusal to raise	C
review . After discreet soundings , they prudently	abandoned	2	the idea , which would have involved a major encroachment	C
assure Janet Daley that these children have not been	abandoned	6	to a type of educational apartheid . It is views	C
and Excise rates by 1993 , the ministers agreed to	abandon	1	key provisions for revising VAT collection arrangements	C
quality of British filmmaking nosedived . UK filmmakers	abandoned	1	their innovations with film narrative , producing	C
oneself , and not selling out to Hollywood , really mean	abandoning	u	melodrama for realism , showmanship for seriousness	C
had come to demand . Dr Clark was now seen hastily	abandoning	4	all those notes , containing all the furious denunciations	C
former scourge of the dozy British motorist , has not	abandoned	1	his life 's work just because Her Indoors shunted	C
£11BILLION a year Lloyd 's of London insurance market is	abandoning	1	rules forcing underwriters to specialise in particular	C
another to impose controls on everything . Mr Gonzalez	abandoned	1	another long-running tradition by which Argentine	C
. Only a fortnight ago , the Lithuanian parliament	abandoned	2	the constitutional guarantee of the party 's leading	C
1987 that Gen Noriega was negotiating with the US to	abandon	1.a	his command for a comfortable exile , sent him a	C
1989 in protest against the government 's decision to	abandon	1	the British PWR programme . The IAEA and safeguards	C
weapons if the Soviet Union broke up , was abruptly	abandoned	1	last year . Too political . The distinction between	C
mean taking up an idea first mooted last year , but	abandoned	1	as too risky : a full coalition between more cautious	C
which will radically affect their lives . It means	abandoning	6	the working class to sectarian politics and violence	C
fight against fascism , and eventually the need to	abandon	2	its emphasis upon political independence . It was	C

Page 1 of 13 Go Next | Last

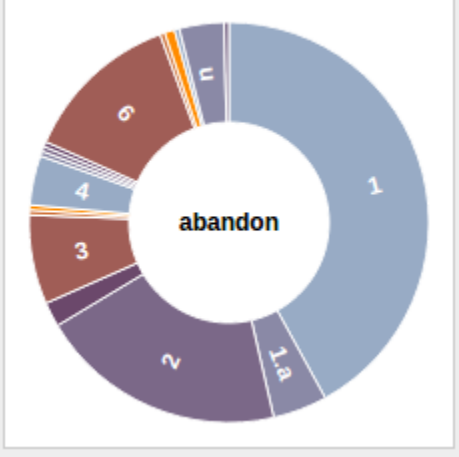
ANNOTATION IN SKETCH ENGINE II

- features for lexicographers
- annotation with word sketches
- bootstrapping of partial annotation
- automatic patterns
- training mode
- multi-line labelling
- custom labels

ANNOTATION IN SKETCH ENGINE III

- synchronization with CPA editor
- basic statistics

Label	Occurs	Filter
1	105	P / N
1.a	11	P / N
2	50	P / N
2.a	5	P / N
3	18	P / N
3.a	1	P / N
3.f	1	P / N
4	10	P / N
4.a	1	P / N
4.f	1	P / N
5.s	1	P / N
6	32	P / N
6.a	1	P / N



CPA EDITOR

- JavaScript (jQuery), standalone
- connected to SkE and DEB server
- creating and managing PDEV entries, pattern, ontology
- code used by Ken Litkowski for PDEP

Pattern Dictionary of English Verbs

Expand SkE Print patterns Unselect Ontology Listing Report a bug Print query

Filter: all Show/Hide columns: Verb Status Patterns Sample BNC50 BNC OEC FN Links Created by Created Last editor Modified Print

Verb	Status	Patterns	Sample	BNC50	BNC	OEC	FN Links	Created by	Created	Last editor	Modified	Print
break	ready	83	2000	8297	18603	186711	13	patrick	2009/06-01	saramoze	2015-08-25	
blow	complete	62	1516	1516	4796	55320	6	patrick	2009/06-01	patrick	2015-01-26	
throw	complete	61	1000	3710	10919	143403	7	patrick	2007/06-01	ymaarouf	2015-03-18	
lose	ready	57	1000	11868	26605	301942	2	jane	2014/12-08	saramoze	2015-01-25	
take	WIP	56	1000	75872	173412	1733310	5	patrick	2009/11-26	saramoze	2015-05-13	
open	ready	56	1000	8695	22394	268691	0	cpa04	2006/11-01	jane	2014-12-15	
go	G	50	54872	54872	226268	2127417	7	patrick	2008/06-20	jane	2014-12-05	
live	ready	43	1000	15402	31991	316892	1	patrick	2009/06-01	jane	2014-12-05	
set	G	38	20542	20542	38838	315361	0	patrick	2008/05-05	jane	2014-11-18	
hit	WIP	38	1000	3706	10344	173794	0	patrick	2006/09-30	jane	2014-12-15	
hang	ready	38	500	2242	8659	96907	11	patrick	2009/06-01	patrick	2014-12-20	
beat	ready	37	1000	2224	7859	101890	2	patrick	2007/07-19	imaarouf	2015-09-18	
call	complete	36	1000	24439	51912	591606	12	patrick	2006/09-30	patrick	2014-12-05	
dig	ready	30	845	845	2623	27870	0	jezek	2007/05-26	jane	2014-12-05	

Showing 1 to 5,601 of 5,601 entries

owe Add pattern Stretch Shrink more Concordance Ontology Renumber Save Save&Close Close

Sample size 300 (out of *2026*) Semantic class Status ready Difficulty 5 Compilation time Erlangen

#	%	Pattern & primary implicature
1.	25.00%	[[Human 1 | Institution 1]] owe [[Human 2 | Institution 2]] [[Money]] (for [[Physical_Object]] | for [[Asset]]) [[Human 1 | Institution 1]] is under obligation to repay [[Money]] borrowed from [[Human 2 | Institution 2]]
2.	19.67%	[[Human 1]] owe [[Human 2 | Institution]] [[Obligation]] [[Human 1]] is morally and/or legally bound to honour [[Obligation]] to [[Human 2]]
3.	17.00%	[[Entity]] owe [[REFLDET Privilege | REFLDET Property | REFLDET {Eventuality = Desirable}]] {to [[Anything]]} [[Entity]] is able to gain [[Privilege | Property]], or have [[Eventuality = Desirable]] happen to them because of [[Anything]]
4.	13.67%	[[Entity]] owe {much | little | ...} {to [[Anything]]} [[Human | Institution | Concept]] is able to develop intellectually, culturally, economically or otherwise because of [[Anything]]
5.	15.00%	[[Eventuality 1]] owe {much | something | ...} {to [[Eventuality 2]]} [[Eventuality 1]] is, to a certain extent, caused or affected by [[Eventuality 2]] [[Anything]] either contributed to, or partially caused [[Eventuality]] to take place
6.	2.33%	[[Human 1]] owe {it} {to [[Self]] | to [[Human 2]]} ({to/INF [V]}) [[Human 1]] feels morally obligated to do something for [[Self | Human 2]]
7.	2.33%	[[Human 1]] owe [[Human 2]] {a debt (of gratitude)} idiom [[Human 1]] feels thankful to [[Human 2]]
8.	1.00%	[[Human 1]] owe [[Human 2]] {apology} [[Human 1]] needs to apologize to [[Human 2]]

Pattern gnaw 4 In corpus: bnc50 ▼ Concordance ▼ Insert Merge into: none ▼ Copy Delete Save Save&Close Close

Subject Animal N Role Rodent Lexset

Verb gnaw no object ☑ no adverbial

Adverbial Opt Prep|Part through Physical_Object N Opt Role Wood Lexset Opt

Primary implicature [[Animal = Rodent]] makes a hole in [[Physical_Object = Wood]] by means of rapid, small bites with the teeth

Generate

idiom pv

Show: Sub. conjunction Indirect object Object Complement Clausals Clausals objects Secondary implicature Domain & Register Framenet Comment Sem. Class

Ontology Expand all Shrink **version: 0.2.9**

- Anything *i.e. anything at all* edit details
  - Entity edit details
    - Abstract_Entity edit details
      - Concept *Must be a word meaning 'concept'; otherwise use Anything = Concept* edit details
        - Proposition edit details
          - Narrative edit details
        - Rule edit details
          - Permission edit details
        - Dispute edit details
        - Information edit details
      - Information_Source edit details
        - Document *[Information_Source, Artifact]* edit details
          - Agreement *[Speech_Act, Document]* edit details
        - Language edit details

Search for Semantic Types

Search

Search nouns

Search

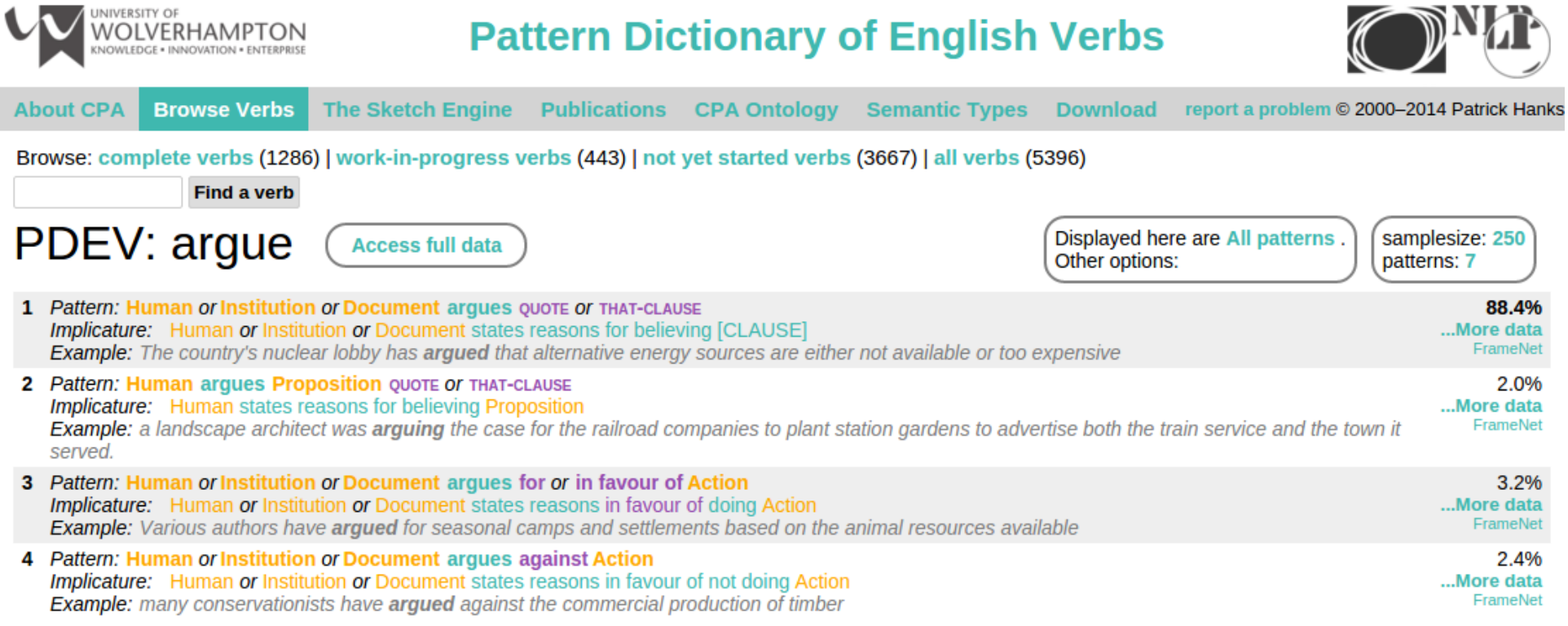
SEMANTIC TYPE: ATTITUDE Close

Verb	Pattern number	Freq
arouse	1,2	763
resent	1	457
applaud	3	158
awaken	3,4,5	78
brush	1,5	76
repeat	1	72
trigger	2	55
greet	2	50
admit	6	39

Nouns	S	O	Ad	∑
glory	0	0	15	15
intention	0	10	0	10
determination	0	9	0	9
glow	0	0	7	7
willingness	0	7	0	7
stance	0	6	0	6
acceptance	0	4	0	4
attitude	0	3	0	3
readiness	0	3	0	3

CPA PUBLIC ACCESS

- simplified interface
- live data, complete verbs



SEMEVAL DATASET

- NAACL 2015
- three tasks
  - CPA parsing
  - CPA clustering
  - CPA pattern editing
- Microcheck, Wingspread
- auto-cpa user

LEMON API

- an official release of PDEV as linked data
- RDF scheme, used by WordNet, DBpedia, ...
- 17,634 triples

CONCLUSION, FUTURE WORK

- a reference for future articles
- consolidation of code (merge)
- other projects are planned (CPA for nouns, adjectives)
- linking English, Italian, Spanish pattern dictionaries (EURALEX)
- full CPA bibliography in the proceedings