

Legal Terms and Word Sketches: a Case Study

Eva Mráková Karel Pala

NLP Lab, FI MU
Brno

Raslan 2010

Introduction

- Goal:
offer supplementary tools for building dictionary of Czech law terms
- Method:
 - 1 VaDis partial parser: complex nominal constructions
 - 2 Clustering according to headwords
 - 3 Word Sketches for the headwords

Recognition of the Complex Nominal Constructions

- Tool: VaDis partial parser
- Source: the Penal Code of Czech Republic (36,000 word forms)
- Result: complex noun groups
- Further processing: clustering groups according to their headwords
- Headwords of big clusters: often essential legal terms like *čin (act)*, *trest (punishment)*, *pachatel (offender)*, *zákon (law)*, *sazba (penalty)*, *vazba (detention)*, *soud (court)*, *předpis (regulation)*, *opatření (measure)*, *následek (consequence)*, *škoda (damage)*, etc.

Example: a Part of the Cluster for the Headword *čin* (act)

complex nominal group	English equivalent
pachatel trestného činu	who committed a criminal act
spáchání trestného činu	committing a criminal act
pokus trestného činu	an attempt to commit a criminal act
znaky trestného činu	attributes of the criminal act
způsob provedení činu	way of the committing a criminal act
dokonání trestného činu	completing a criminal act
účastník trestného činu	participant of the criminal act
trestnost pokusu trestného činu	punishability of the attempt
doba spáchání činu	time of the committing a criminal act
povaha spáchaného činu	nature of the committed criminal act
stupeň nebezpečnosti činu	degree of the dangerousness of crim. a.

Obtaining Legal Terms through WSE

- at the moment: no corpus of Czech legal texts
- instead we used big corpora of Czech common texts: SYN2000 (110 million tokens), Czes (1.2 billion tokens)
- we explored WS for the headwords
- results:
 - we obtained several hundreds relevant legal terms (despite of common texts)
 - some WS tables form natural groups of legal terms (e.g. `gen_1` contains a reasonable classification of criminal acts)
 - we also obtained verbs with legal meaning, including candidates for further extension of VerbaLex (e.g. *promlčet* (*be time-barred*), *překvalifikovat* (*change qualification*), *prošetřovat* (*investigate*), *zpochybňovat* (*question*),...)

Example: Legal Verbs from WS Tables

spáchat (commit)	dopouštět se (perpetrate)
páchat (commit)	odsoudit (condemn, sentence)
vyšetřovat (investigate)	potrestat (punish)
překvalifikovat (change qualification)	přiznat (confess)
prošetřovat (investigate)	postihovat (affect)
vykonat (perform)	prokázat (prove)
zmařit (thwart)	ohlásit (announce)
ospravedlňovat (justify)	uprchnout (escape)
stíhat (prosecute)	zodpovídat (be responsible)
objasňovat (explain)	zadržet (arrest, detain)
promlčet (be time-barred)	napravit (amend)
zpochybňovat (question)	litovat (regret)

Case Study: WS for the Headword *čin* (*act*)

- WS with mainly legal terminology (despite its common meaning)
- *čin*: 13,400 occurrences in SYN2000, 88,500 ones in Czes
- WS tables: more than one hundred words in SYN2000, more than two hundreds words in Czes
- results:
 - gen_1 table containing a reasonable classification of criminal acts
 - tables containing verbs with legal meaning
 - terminological legal adjectives, e.g. *úmyslný* (*deliberate*), *nedbalostní* (*caused by negligence*), *násilný* (*violent*), *protiprávní* (*illegal*)

Example: Sketch Table gen_1 for *čin*

zpronevěra (defalcation)	padělání (forgery)
krádež (larceny)	hanobení (defamation)
poškození (damaging)	podílnictví (shareholding)
loupež (robbery)	vražda (murder)
maření (obstruction)	zneužití (misappropriation)
porušování (violation)	ohrožování (threatening)
zneužívání (misappropriating)	zvýhodňování (privileging)
výtržnictví (disorderly behaviour)	ohrožení (emergency)
vydírání (extortion)	znásilnění (rape)
pomluva (slander)	podplácení (bribery)
zkrácení (reduction)	týrání (abuse)
zanedbání (negligence)	

Figure: A part of the WS Table of *čin* in SYN2000

čin SYN2000c frekvence = 13398

a_modifier	10767	3.0	prec_misto/R	27	26.1	prec_za	294	6.1	prec_k	338	4.0	gen_1	2403	2.0
restný	5941	13.23	seriál	6	4.44	odsouzení	8	7.72	napomáhání	7	9.09	ublížení	168	10.52
násilný	259	9.37				odsoudit	17	6.87	odhodlat	8	7.89	krádež	164	9.82
spáchaný	132	8.59	is_obj2_of	357	8.9	stíhat	12	6.86	dohnat	8	7.61	výtržnictví	64	9.61
závažný	175	8.55	dopustit	245	10.96	zodpovědnost	7	6.29	odvaha	11	6.71	zneužívání	94	9.42
kriminální	129	8.35	dopouštět	48	9.87	trest	23	5.78	vůle	14	5.32	poškození	66	9.38
teroristický	120	8.31	týkat	10	4.2	odpovědnost	16	5.64	přejít	10	5.25	podvod	113	9.34
motivovaný	99	8.15				označit	9	4.66	přistoupit	6	4.97	porušování	88	9.33
hrdinský	95	8.13	prec_z	938	7.2	považovat	19	3.93	příprava	12	4.31	vydírání	65	9.26
dovolený	74	7.7	obvinít	436	10.84	cena	6	1.05	dojit	13	3.96	zpronevěra	52	9.15
úmyslný	58	7.39	obvinění	190	10.0				slovo	21	3.93	maření	47	9.14
hrůzný	53	7.24	obžalovat	34	9.52	prec_pro	329	5.1	rozhodnout	8	3.29	vlastizrada	46	9.08
tvůrčí	50	6.85	podezření	82	8.73	stíhat	111	10.06	pomoc	8	3.29	zneužití	72	9.08
konkrétní	72	6.64	vinit	12	7.19	stíhání	37	8.15	věst	10	2.72	hanobení	44	9.01
uvedený	67	6.64	vyšetřovatel	27	7.16	odsouzení	7	7.48				loupež	55	8.98
slavný	63	6.62	zodpovídat	8	7.15	obžaloba	9	7.41	post_proti	51	3.8	týrání	37	8.53
odvážný	35	6.56	obžaloba	9	6.97	odsoudit	24	7.36	lidskost	7	8.08	vražda	107	8.52
nedbalostní	31	6.55	police	11	3.47	obvinění	16	6.58				pomluva	34	8.51
zoufalý	35	6.52	muž	6	1.88	oznámení	11	6.03	prec_o	231	2.3	šíření	57	8.44
obecný	48	6.39				žaloba	6	5.54	jít	115	5.53	padělání	27	8.35
brutální	28	6.27							jednat	16	4.49	kuplíství	24	8.29
Palachův	22	6.04				prec_při	104	4.7	pokus	7	3.8	ohrožení	52	7.95

Verbs with Financial Meaning

- subgroup of verbs occurring in legal texts
- we inspected verbs with EXT(sum:1) argument in their complex valency frames
- then we explored their frequencies in Czes
- observation: less frequent verbs display specialized terminological meanings, more detailed evaluation would be needed

Table: Financial Verbs and Their Frequencies in Czes

alokovat (allocate)	670
realokovat (reallocate)	13
danit (tax)	45,374
zdanit (tax)	3,291
dodanit (pay up the tax)	117
dodaňovat (pay up the tax)	20
dlužít (owe, have a debt)	11,773
vydlužít (take on loan)	13
fakturovat (invoice)	755
vyfakturovat (invoice)	135
financovat (finance)	18,598
dofinancovat (finance up)	219
předfinancovat (prefinance)	19
počítat (calculate, compute)	116,673
spočítat (calculate, compute)	13,663
tarifikovat (tariff)	23
tarifovat (tariff)	12

Relation between VerbaLex Frames and WS

- assumption: semantic labels of the verb's arguments (e.g. EXT(sum:1)) match reasonably with the particular nouns in WS tables of the verb
- case study: verb *vy/fakturovat* (*invoice*)

fakturovat_{n1}, zaúčtovat_{n1}, zaúčtovávat_{n1}, naučtovat_{n1}, naučtovávat_{n1}

AG<person:1>^{obl}_{kdo1} VERB^{obl} REC<person:1|institut.:1>^{opt}_{komu3}
 ART<goods:1>|ACT<act:2>^{opt}_{za+co4} EXT<sum:1>^{obl}_{co4}

- result: nouns found in the respective corpus sentences semantically agree with what is predicted by the argument labels in the valency frame

Figure: WS of *fakturovat* (invoice) in Czes

fakturovat preloaded/czes frekvence = 755

has_obj3	79	59.6
odběratel	3	2.93
dealer	2	2.53
zákazník	56	2.2

post_od	2	2.6
duben	2	0.53

coord	32	1.8
prodavat	3	6.06
vyhodnocovat	10	5.26
inkasovat	2	2.93

has_obj4	96	6.8
caska	2	8.81
mlčení	2	3.7
provize	3	3.59
úrok	11	3.38
instalace	5	1.14
poradenství	2	0.97
montáž	2	0.97
nájem	2	0.35
pracoviště	2	0.29

post_v	13	1.8
přepočít	3	3.86

post_po	4	5.1
uskutečnění	2	2.94