Automatic mapping Lexical Resources: A Lexical Unit as a Keystone

Eduard Bejček, Václava Kettnerová and Markéta Lopatková

Task: Link two lexicons together.

Both are valency lexicons. Both are for Czech language.

An example of VALLEX lexicon:

obracet^{impf}, **obrátit**^{pf}

1 ≈ impf: **otáčet; měnit směr pohybu** pf: **otočit**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\uparrow\mathrm{DIR}^{typ}$
-example: impf: obracet auto / loď na bok / skříň ke zdi pf: obrátit seno
-rcp: ACT-PAT:

2 ≈ impf: **proměňovat** pf: **proměnit**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{EFF}_{k+3,na+4,v+4}^{obl}$
-example: impf: obracel nepřátele v prach pf: obrátil pohany na křesťanství
-rcp: ACT-PAT:

3 ≈ impf: **zaměřovat** pf: **zaměřit** (idiom)
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{k+3,na+4}^{obl}$ $\mathbf{DPHR}_{zřetel,zájem,pozornost}^{obl}$
-example: impf: obracet pozornost / zájem pf: obrátit zájem / zřetel

4 ≈ impf: **převracet** pf: **zpřevracet** (idiom)
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{MANN}^{obl}$
-example: impf: obracel vše vzhůru nohama pf: obrátil vše naruby
-usage in ČNK: impf: Svou propagační kampaň přirozeně obracím naruby.

5 ≈ impf: **měnit mínění** pf: **změnit mínění** (idiom)
-frame: $\mathbf{ACT}_1^{obl}$
-example: impf: Pavel najednou rychle obracel pf: Pavel najednou obrátil
-rfl: pass0: impf: rychle se obracelo pf: rychle se obrátí

6 ≈ impf: **měnit** pf: **změnit** (idiom)
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{EFF}_{v+4}^{opt}$
-example: impf: chabá koruna obracela vývoj obchodní bilance
pf: jeho slova obrátila diskusi v prudkou polemiku
-rfl: pass: impf: obracel se vývoj obchodní bilance

Very similar format.
Very similar content.

...and yet...

It is relatively easy to transform formats to a common one. But difficult to compare and link the content.

An example of PDT-Vallex lexicon:

obracet

obracet^1 *1x* **ACT**(1) **PAT**(4) **ADDR**(na+4)
- *obracela svůj vztek na dávného přítele*

obracet^2 1x, *2x* **ACT**(1) **PAT**(4) **EFF**(v+4)
(proměnit) • *obracejí naše výroky v pravý opak*

obracet^3 *1x* **ACT**(1) **PAT**(4)
(otáčet, převracet) • *obracet skříň; o. auto; ke zdi.DIR3*

obracet^4 **ACT**(1) **CPHR**({pozornost-1,...}.4) **DIR3**()
(obrátit, upřít) • *obracet pozornost jinam*

obracet^5 1x **ACT**(1) **DPHR**(naruby) **PAT**(4)
- *obracet vše naruby*

VALLEX lexicon

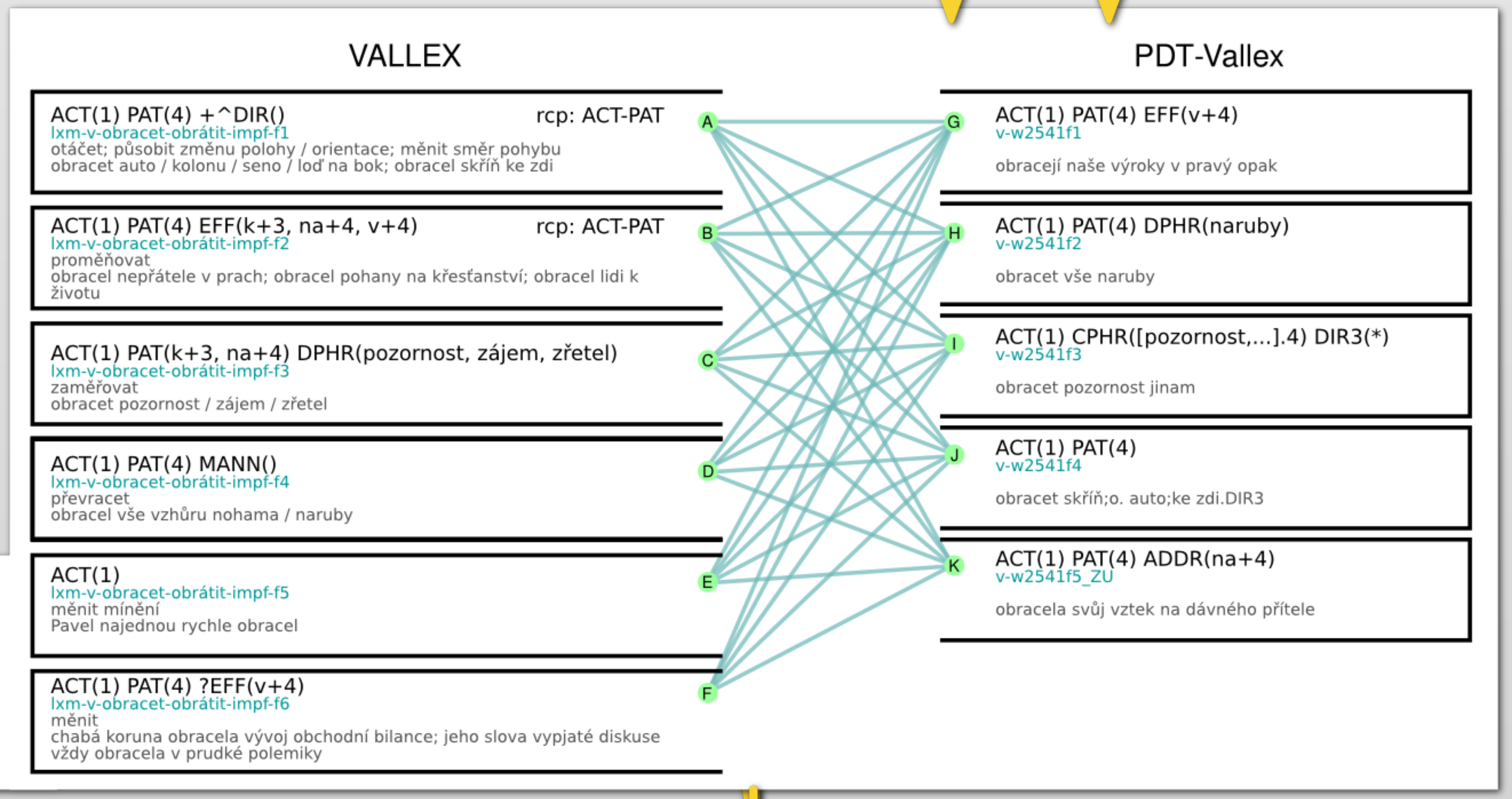
- Complex description of selected verbs.
- Verbs selected according to their frequency.
- Some supplement information (missing in PDT-Vallex).

Common format

There was no need to use one of universal formats.
A plain extension of original format was used:
- identical attributes are represented in the same way
- there is also an option to represent attributes from one lexicon only.

PDT-Vallex lexicon

- Description of only selected lexical units.
- Verbs and lexical units selected from PDT corpora.
- Each LU is provided with a corpus evidence.



VALLEX

A — ACT(1) PAT(4) +^DIR() rcp: ACT-PAT
lxm-v-obracet-obrátit-impf-f1
otáčet; působit změnu polohy / orientace; měnit směr pohybu
obracet auto / kolonu / seno / loď na bok; obracel skříň ke zdi

B — ACT(1) PAT(4) EFF(k+3, na+4, v+4) rcp: ACT-PAT
lxm-v-obracet-obrátit-impf-f2
proměňovat
obracel nepřátele v prach; obracel pohany na křesťanství; obracel lidi k životu

C — ACT(1) PAT(k+3, na+4) DPHR(pozornost, zájem, zřetel)
lxm-v-obracet-obrátit-impf-f3
zaměřovat
obracet pozornost / zájem / zřetel

D — ACT(1) PAT(4) MANN()
lxm-v-obracet-obrátit-impf-f4
převracet
obracel vše vzhůru nohama / naruby

E — ACT(1)
lxm-v-obracet-obrátit-impf-f5
měnit mínění
Pavel najednou rychle obracel

F — ACT(1) PAT(4) ?EFF(v+4)
lxm-v-obracet-obrátit-impf-f6
měnit
chabá koruna obracela vývoj obchodní bilance; jeho slova vypjaté diskuse vždy obracela v prudké polemiky

PDT-Vallex

G — ACT(1) PAT(4) EFF(v+4)
v-w2541f1
obracejí naše výroky v pravý opak

H — ACT(1) PAT(4) DPHR(naruby)
v-w2541f2
obracet vše naruby

I — ACT(1) CPHR([pozornost,...].4) DIR3(*)
v-w2541f3
obracet pozornost jinam

J — ACT(1) PAT(4)
v-w2541f4
obracet skříň;o. auto;ke zdi.DIR3

K — ACT(1) PAT(4) ADDR(na+4)
v-w2541f5_ZU
obracela svůj vztek na dávného přítele

All possible mappings are shown.
All of them will be assigned a *Score* according to lemmas, valency frames and reciprocity.

Lemmas

D | H
obracel vše vzhůru nohama / naruby | obracet vše naruby ACT(1) PAT(4) DPHR(naruby)

⇓ lemmatization ⇓
the verb removed
autosemantic words removed

~~obracet všechen~~ vzhůru noha naruby | ~~obracet všechen~~ naruby

one out of three = $^1/_3$ | $^1/_1$ = one out of one

⇓

$(^1/_3 + {}^1/_1) / 2 = {}^2/_3$

Valency Frames

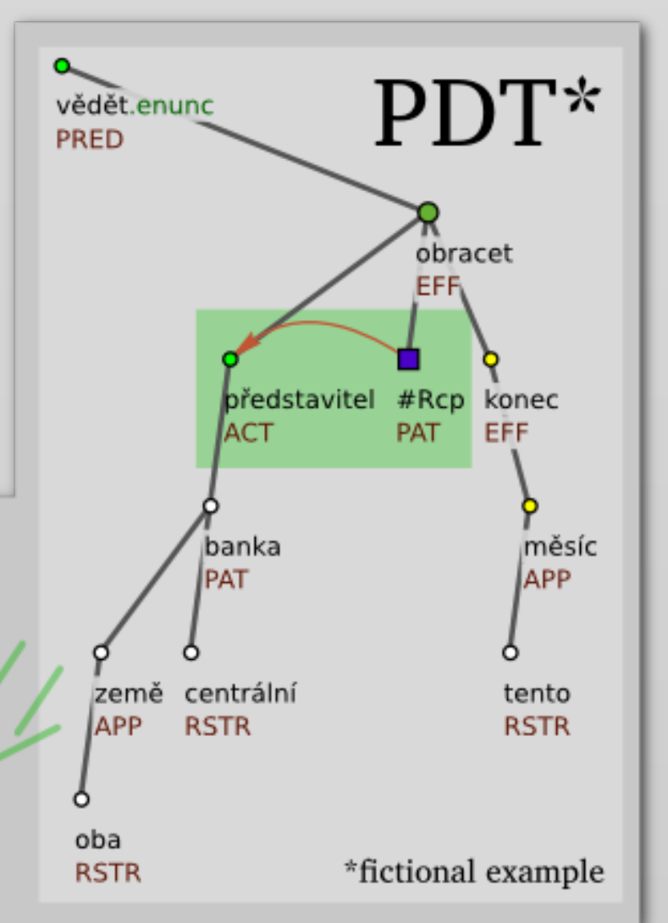
B ACT(1) PAT(4) EFF(k+3, na+4, v+4)
G ACT(1) PAT(4) EFF(v+4)

rule #14: ACT + PAT + EFF(with missing forms)

Reciprocity

B rcp: ACT-PAT
F rcp: ---

G in the data:
#Rcp: ACT-PAT



*fictional example

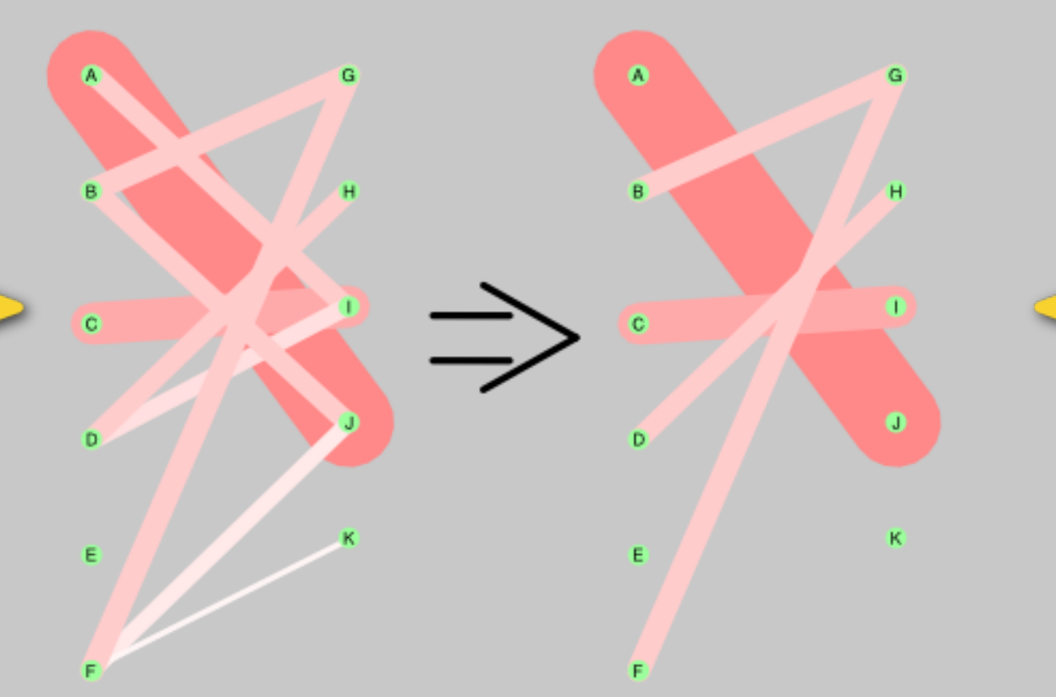
Linking statistics

	VALLEX	PDT-Vallex
Verb lemmas covered by both lexicons	3,541	3,541
LUs represented by the given verb lemmas	8,816	7,674
Average number of LUs per verb lemma	2.5	2.2
LUs with no link	2,245	1,622
LUs with just one link	5,537	4,670
LUs with more than one link	1,034	1,382

Comparision with manual linking

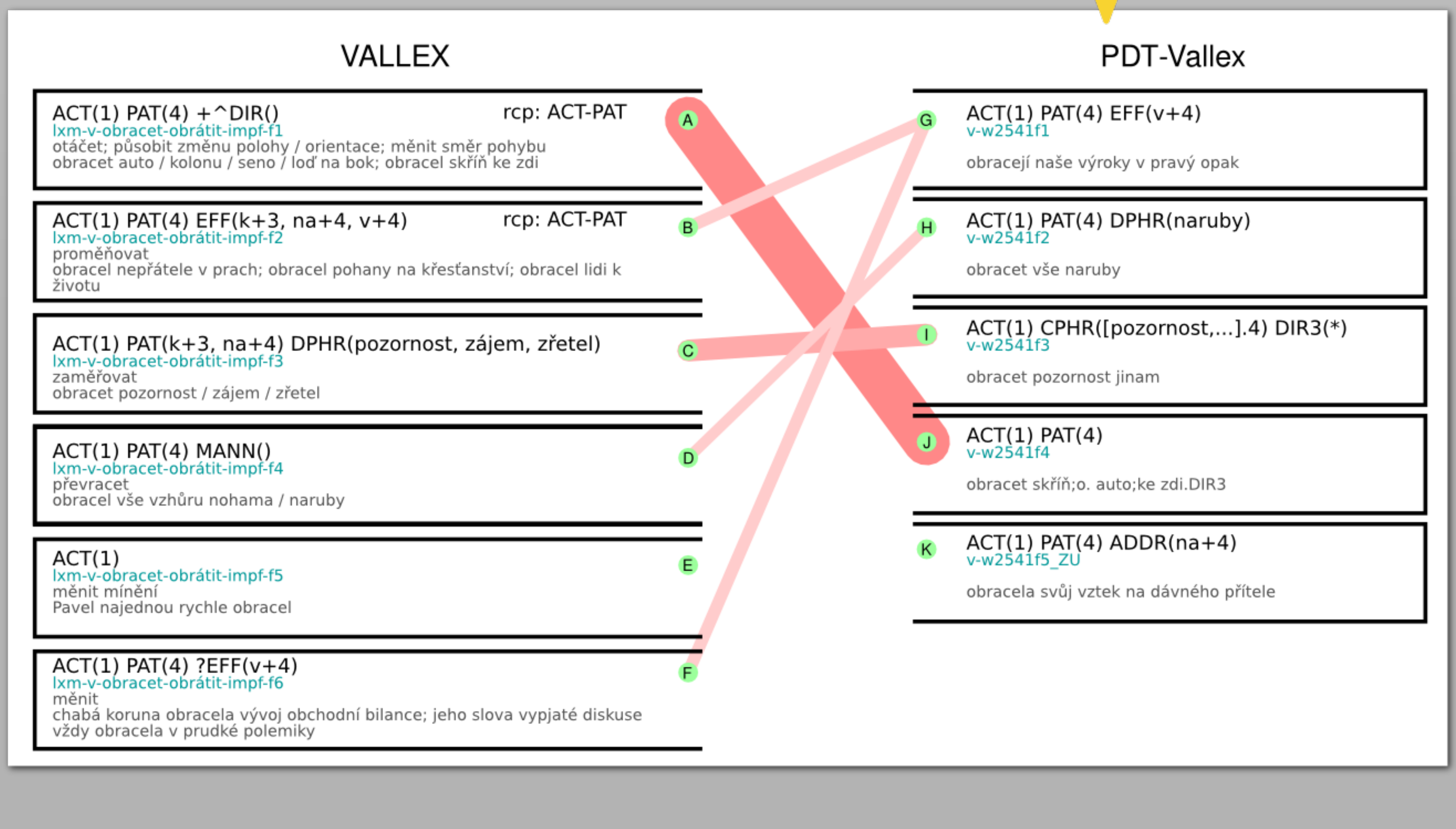
	VALLEX			PDT-Vallex		
Verb lemmas selected for annotation	200			200		
LUs represented by the given verb lemmas	716			528		
Average number of LUs per verb lemma	3.6			2.6		
	A	B	auto	A	B	auto
LUs with no link	249	280	**175**	61	72	93
LUs with just one link	415	386	464	417	422	312
LUs with more than one link	52	50	77	50	34	**123**

Pruning



Evaluation against manual annotation

Number of LUs in VALLEX	Precision	Recall	F-measure
v_1	95	77	85
v_2	84	72	77
v_3	69	82	75
v_4	66	75	70
v_5	57	88	69
v_6	47	83	60
v_7	45	68	54
v_8	40	73	52
v_9	54	76	63
Average weighted over all 200 verbs	81	77	79
Average weighted for annotators	93	92	92



VALLEX

A — ACT(1) PAT(4) +^DIR() rcp: ACT-PAT
lxm-v-obracet-obrátit-impf-f1
otáčet; působit změnu polohy / orientace; měnit směr pohybu
obracet auto / kolonu / seno / loď na bok; obracel skříň ke zdi

B — ACT(1) PAT(4) EFF(k+3, na+4, v+4) rcp: ACT-PAT
lxm-v-obracet-obrátit-impf-f2
proměňovat
obracel nepřátele v prach; obracel pohany na křesťanství; obracel lidi k životu

C — ACT(1) PAT(k+3, na+4) DPHR(pozornost, zájem, zřetel)
lxm-v-obracet-obrátit-impf-f3
zaměřovat
obracet pozornost / zájem / zřetel

D — ACT(1) PAT(4) MANN()
lxm-v-obracet-obrátit-impf-f4
převracet
obracel vše vzhůru nohama / naruby

E — ACT(1)
lxm-v-obracet-obrátit-impf-f5
měnit mínění
Pavel najednou rychle obracel

F — ACT(1) PAT(4) ?EFF(v+4)
lxm-v-obracet-obrátit-impf-f6
měnit
chabá koruna obracela vývoj obchodní bilance; jeho slova vypjaté diskuse vždy obracela v prudké polemiky

PDT-Vallex

G — ACT(1) PAT(4) EFF(v+4)
v-w2541f1
obracejí naše výroky v pravý opak

H — ACT(1) PAT(4) DPHR(naruby)
v-w2541f2
obracet vše naruby

I — ACT(1) CPHR([pozornost,...].4) DIR3(*)
v-w2541f3
obracet pozornost jinam

J — ACT(1) PAT(4)
v-w2541f4
obracet skříň;o. auto;ke zdi.DIR3

K — ACT(1) PAT(4) ADDR(na+4)
v-w2541f5_ZU
obracela svůj vztek na dávného přítele

Conclusion

If there is
- a different granularity,
- a border between lexical units set differently,
- missing lexical unit or
- not enough comparable information,

it is very difficult to automatically link the verb.
(No matter which format we use for the data.)

This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

This research has been supported by the Czech Science Foundation GA ČR, grant No. GA P406/12/0557 and by grant COST CZ LD14117 of the Ministry of Education, Youth and Sports of the Czech Republic.