Advanced Searching in the Valency Lexicons Using PML-TQ Search Engine

Eduard Bejček, Václava Kettnerová, and Markéta Lopatková
Charles University in Prague, ÚFAL MFF

Text, Speech and Dialogue 2010
Brno, Czech Republic
September 8th, 2010

Outline

- Motivation
  - searching in complex lexicons
- Valency lexicons for a demonstration
- PML format
  - Viewing – TrEd
  - Searching – PML-TQ
  - Interlinking

Motivation

- Electronic lexical resource (an example):
  - Simple bilingual dictionary

- What we need to find?
  - search the lemma *look*

1. *(gaze, stare)* regarder; ***2.*** ...

Motivation

- Electronic lexical resource (an example):
  - Simple bilingual dictionary

- What we need to find?
  - search the lemma *look*
  - fulltext search *"looking forward to"*

1. *(gaze, stare)* regarder; ***2.*** ...

Motivation

- Electronic lexical resource (an example):
  - Simple bilingual dictionary

- What we need to find?
  - search the lemma *look*
  - fulltext search *"looking forward to"*
  - regular expressions `/look.* to/`

1. *(gaze, stare)* regarder; ***2.*** ...

... ***c.*** she's **looking forward to** going on holiday → elle a hâte de partir en vacances ...

Motivation

- Electronic lexical resource (an example):
  - Simple bilingual dictionary

- What we need to find?
  - search the lemma — *look*
  - fulltext search — *"looking forward to"*
  - regular expressions — `/look.* to/`
  - advanced search? — *"a word with more than five translations"*

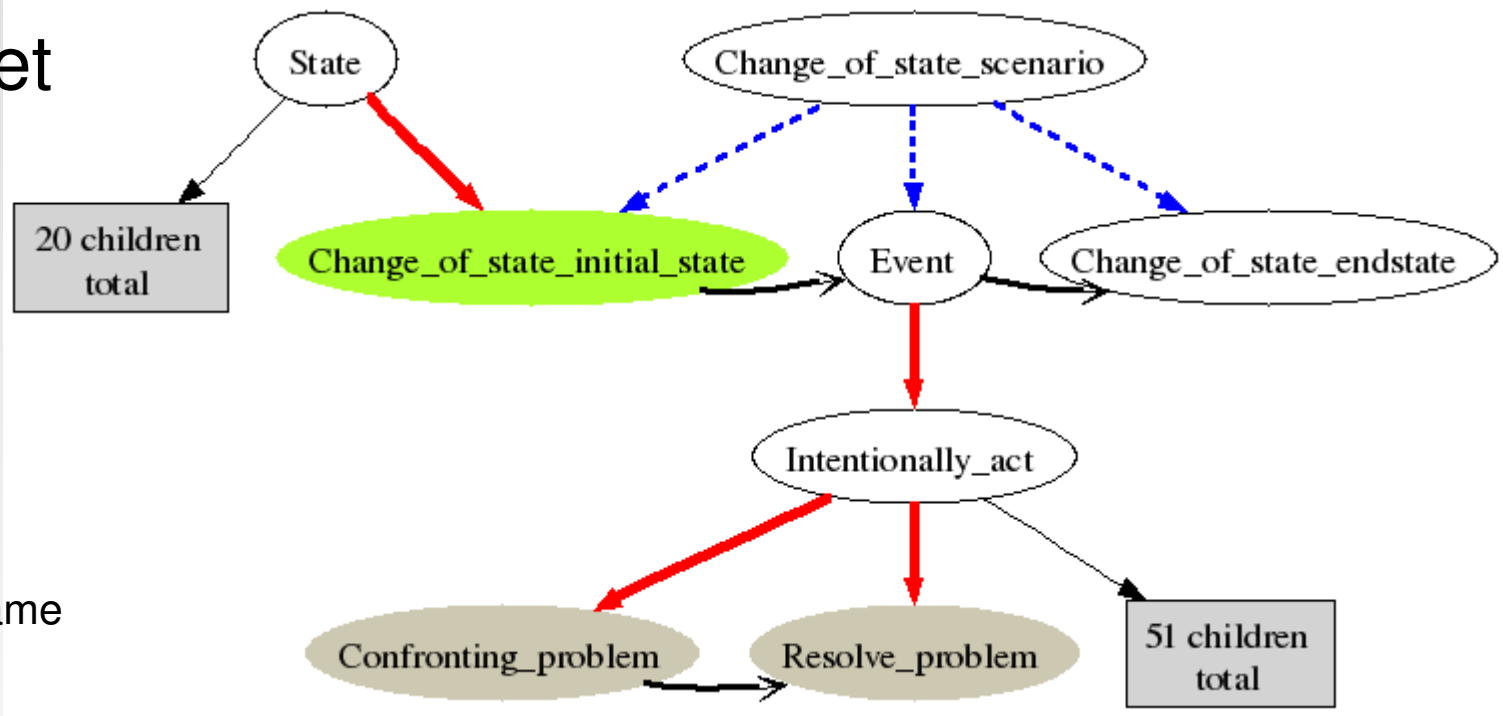
1. *(gaze, stare)* regarder; ***2.*** ...

... ***c.*** she's **looking forward to** going on holiday → elle a hâte de partir en vacances ...

stand → ***1.*** se lever; ***2.*** tenir debout; ***3.*** se trouver; ***4.*** marcher sur; ***5.*** rester valable; ***6.*** se présenter; ...

Motivation (2)

- Lexicons with complex hierarchical structure (1)
  - FrameNet



Searching for:

- lemma
- name of the frame

Motivation (3)

- Lexicons with complex hierarchical structure (2)
  - VerbNet

Searching for:

- lemma

ROLES REF

- AGENT [+ANIMATE | +ORGANIZATION]
- THEME
- SOURCE [+ANIMATE | [+LOCATION & -REGION]]
- BENEFICIARY [+ANIMATE]

FRAMES REF KEY

NP V NP

EXAMPLE	"The thief stole the paint."
SYNTAX	AGENT V THEME
SEMANTICS	MANNER(DURING(E), ILLEGAL, AGENT) HAS_POSSESSION(START(E), ?SOURCE, THEME) HAS_POSSESSION(END(E), AGENT, THEME) NOT(HAS_POSSESSION(END(E), ?SOURCE, THEME)) CAUSE(AGENT, E)

NP V NP PP.SOURCE

EXAMPLE	"The thief stole the paint from the museum."
SYNTAX	AGENT V THEME {{+SRC}} SOURCE
SEMANTICS	MANNER(DURING(E), ILLEGAL, AGENT) HAS_POSSESSION(START(E), SOURCE, THEME) HAS_POSSESSION(END(E), AGENT, THEME) NOT(HAS_POSSESSION(END(E), SOURCE, THEME)) CAUSE(AGENT, E)

NP V NP PP.BENEFICIARY

EXAMPLE	"The thief stole the paint for Mary."
SYNTAX	AGENT V THEME {FOR} BENEFICIARY
SEMANTICS	MANNER(DURING(E), ILLEGAL, AGENT) HAS_POSSESSION(START(E), ?SOURCE, THEME) HAS_POSSESSION(END(E), BENEFICIARY, THEME) NOT(HAS_POSSESSION(END(E), ?SOURCE, THEME)) CAUSE(AGENT, E)

Motivation (4)

- Lexicons with complex hierarchical structure (3)
  - Corpus Pattern Analysis

Last

No.	%	Pattern / Implicature
1	86%	**[[Eventuality | Physical Object | Psych]] last [NO OBJ] [Adv[Time]]**
		[[Eventuality | Physical Object | Psych]] continues (to exist) for a particular length of time expressed by [Adv[Time]]
2	13%	**[[Eventuality | Physical Object]] last [NO OBJ] [NO ADVL]**
		[[Eventuality | Physical Object]] continues to exist for a long time
3	2%	**[[Physical Object]] last [[Human]] [Adv[Time]]**
		[[Physical Object]] does not run out and is enough for [[Human]] for [Adv[Time]]

Searching for:

- lemma

Motivation (5)

- Lexicons with complex hierarchical structure (4)
  - VerbaLex

vyčkat$^{pf}_{1}$
vyčkávat$^{impf}_{1}$ **počkat**$^{pf}_{2}$ **čekat**$^{impf}_{5}$

definition: *odkládat něco na pozdější dobu*
class: linger-53.1
passive: no

1 čekat$_5$, vyčkat$_1$, vyčkávat$_1$, počkat$_2$ ≈
-frame: **AG**<person:1>$^{obl}_{kdo1}$ **VERB**^{obl} **ACT**<act:2>$^{obl}_{s+čím7}$
-example: *počkal s nákupy* ***(pf)***
-example: *vyčkal s odjezdem* ***(pf)***
-example: *čekal s přípitkem* ***(impf)***
-synonym:
-use: prim
-reflexivity: no

2 čekat$_5$, vyčkat$_1$, vyčkávat$_1$, počkat$_2$ ≈
-frame: **AG**<person:1>$^{obl}_{kdo1}$ **VERB**^{obl} **ACT**<act:2>$^{obl}_{s+čím7}$ **TIME**<point:6>$^{obl}_{do+čeho2}$
-example: *počkali se zahájením porady do čtyř hodin* ***(pf)***
-synonym:
-use: prim
-reflexivity: no

Searching for:

- lemma

Lists for

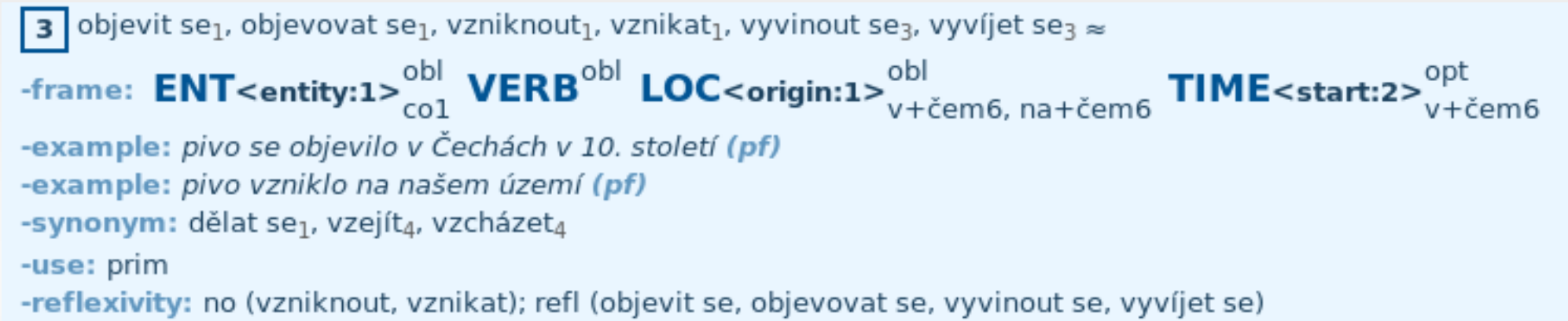
- verb roles
- morphemic roles
- verb classes, etc.

Motivation (6)

- Example query
  - find a reflexive verb
    with optional TIME complement
    realized in other than prepositional genitive case
    that has more than two synonyms
    - hardly by a regular expression

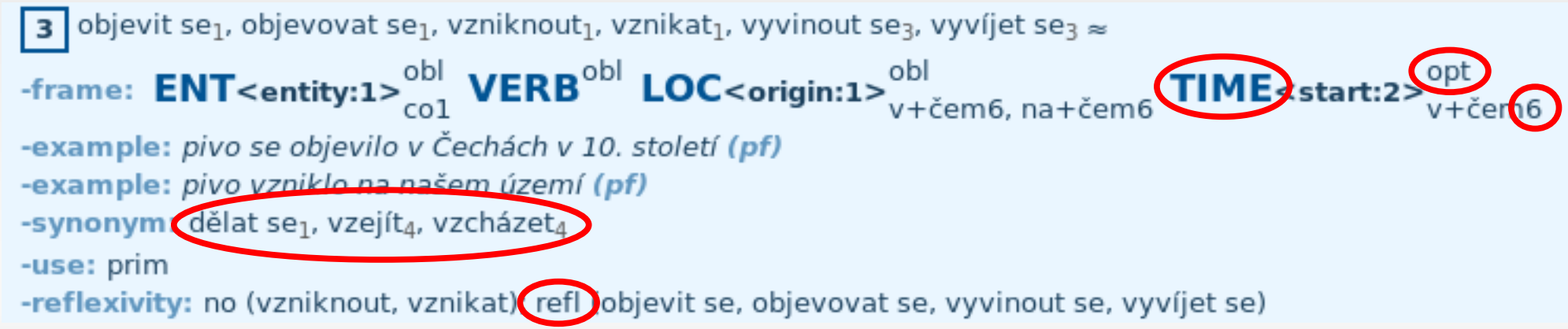
Motivation (6)

- Example query
  - find a reflexive verb
    with optional TIME complement
    realized in other than prepositional genitive case
    that has more than two synonyms
    - hardly by a regular expression



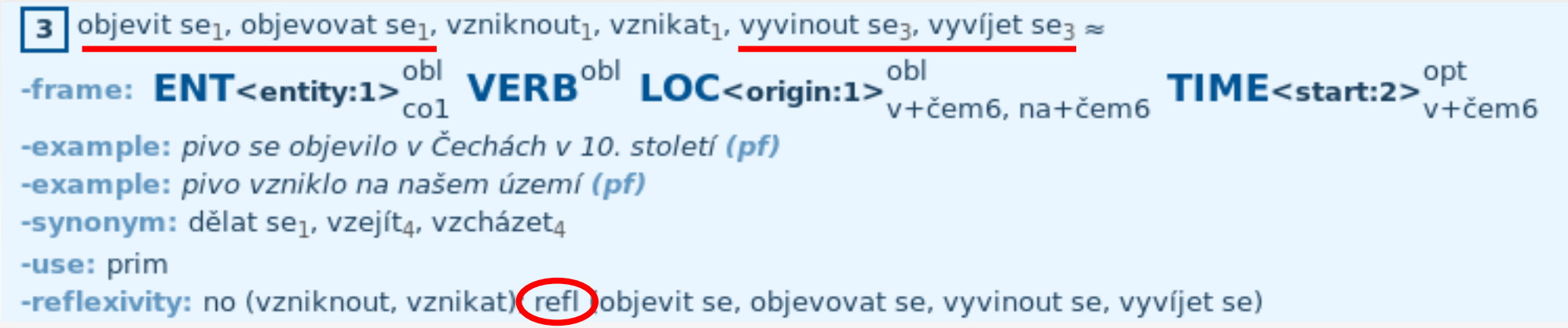
Motivation (6)

- Example query
  - find a reflexive verb
    with optional TIME complement
    realized in other than prepositional genitive case
    that has more than two synonyms
    - hardly by a regular expression



Motivation (6)

- Example query
  - find a reflexive verb
    with optional TIME complement
    realized in other than prepositional genitive case
    that has more than two synonyms
    - hardly by a regular expression



Motivation (7)

- What (and how) we want to search in lexicons?
- How we want to display it?

- a sophisticated way to search lexicons with complex structure

Goal: Universal advanced searching

- As a case study: PDT-VALLEX and VALLEX 2.5
  - similar approaches, different formats and information
- What we want to do with them?
  - interlink lemmas and their lexical units
  - facilitate searching and browsing in linked lexicons
- Tools
  - search engine and viewer
  - universal, adaptable for other lexicons

Two Czech lexicons

Lexicons	PDT-VALLEX	VALLEX 2.5
PoS	verbs, some nouns and adjectives	verbs
lemmas	at least once in PDT	most frequent ones and their aspectual counterparts
frames	at least once in PDT	~ all
frame information - core	**frame**, gloss, example	
- further	syntactic information on functors, frame frequency in data	reciprocity, reflexivity, grammatical control, syntactico-semantic class
- linking	linked with the occurences in PDT	no linking
lexicon unit	lemma	lexeme

Unit: Lemma vs. Lexeme

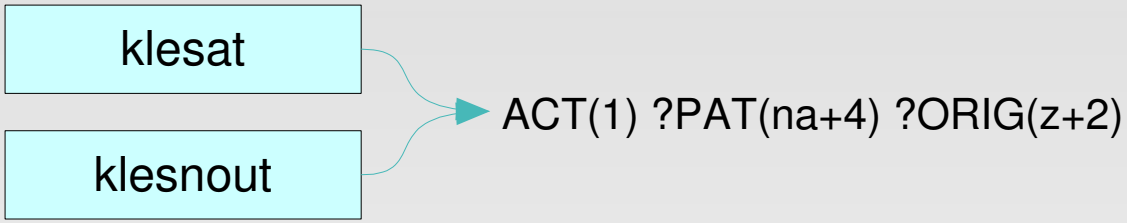
PDT-VALLEX

klesat

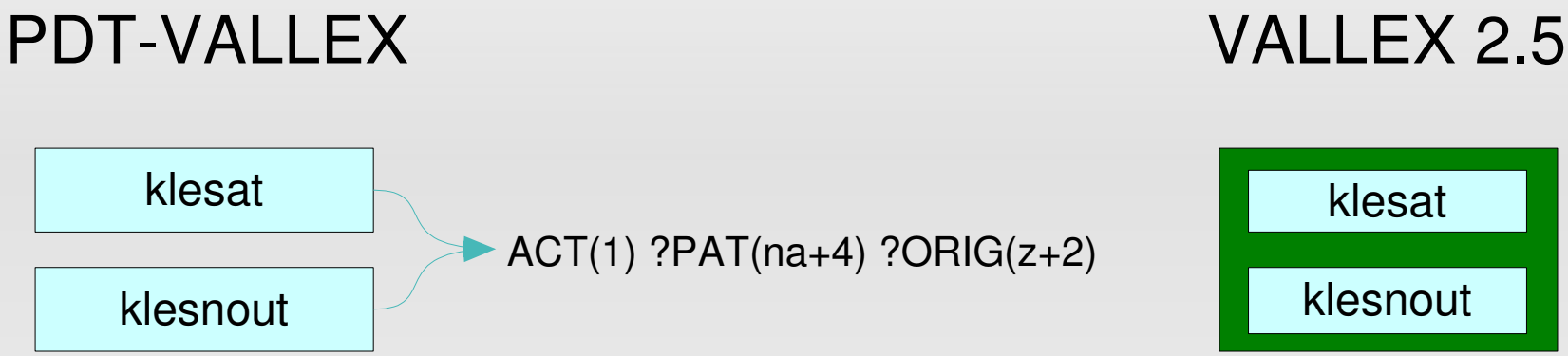
klesnout

Unit: Lemma vs. Lexeme

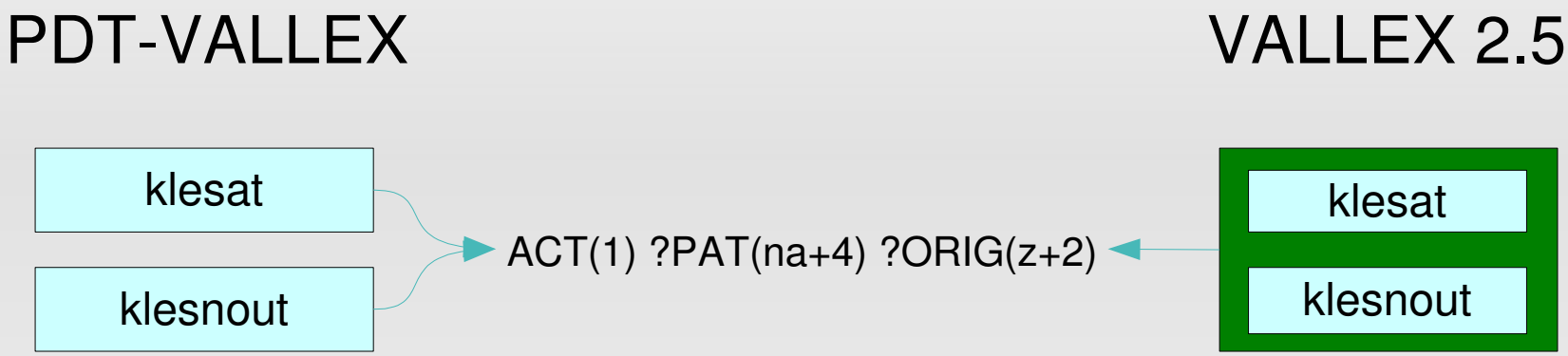
PDT-VALLEX



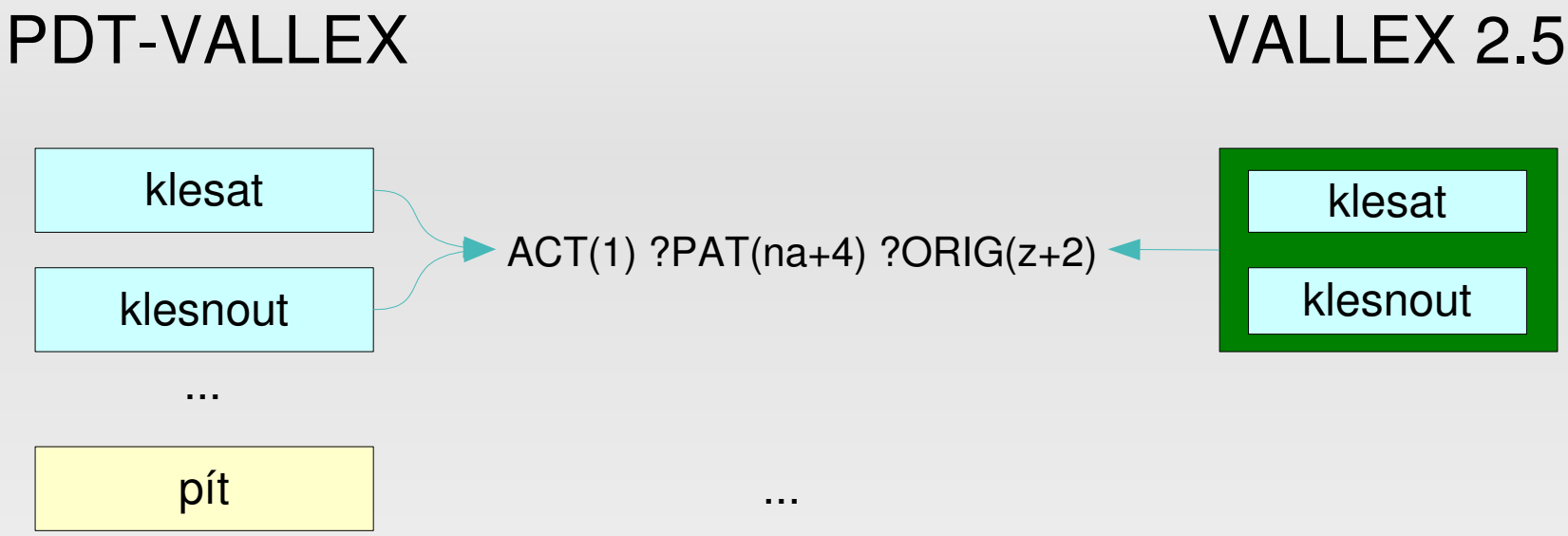
Unit: Lemma vs. Lexeme



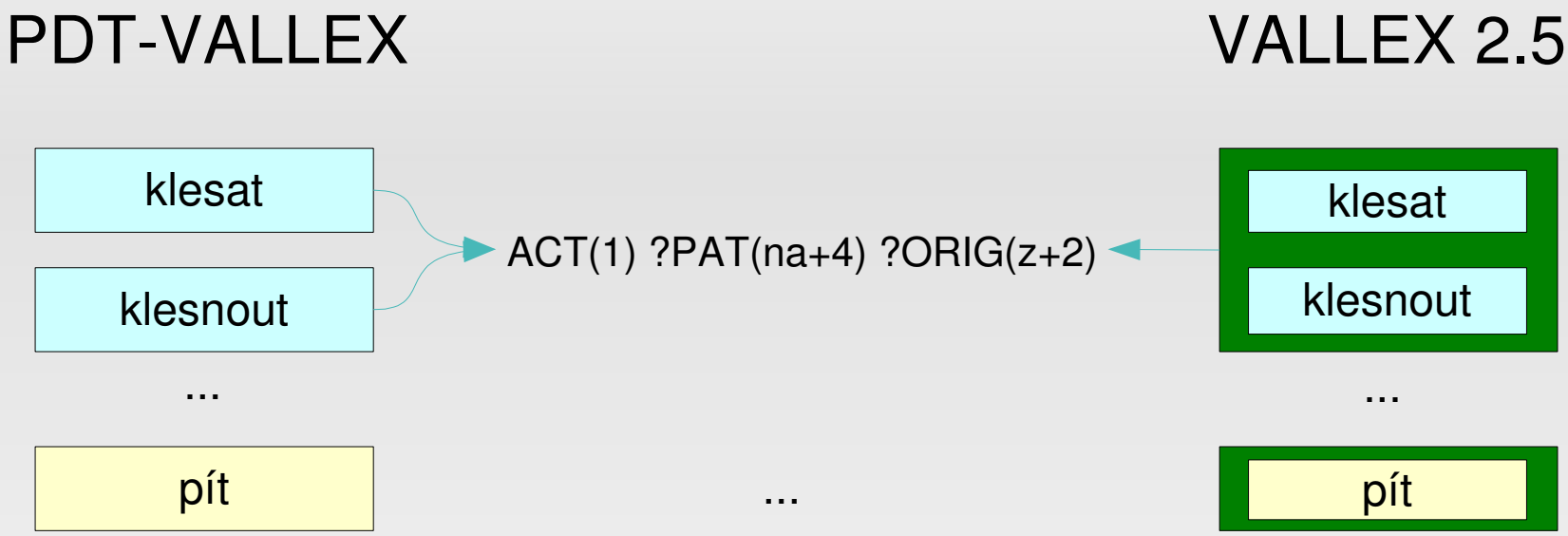
Unit: Lemma vs. Lexeme



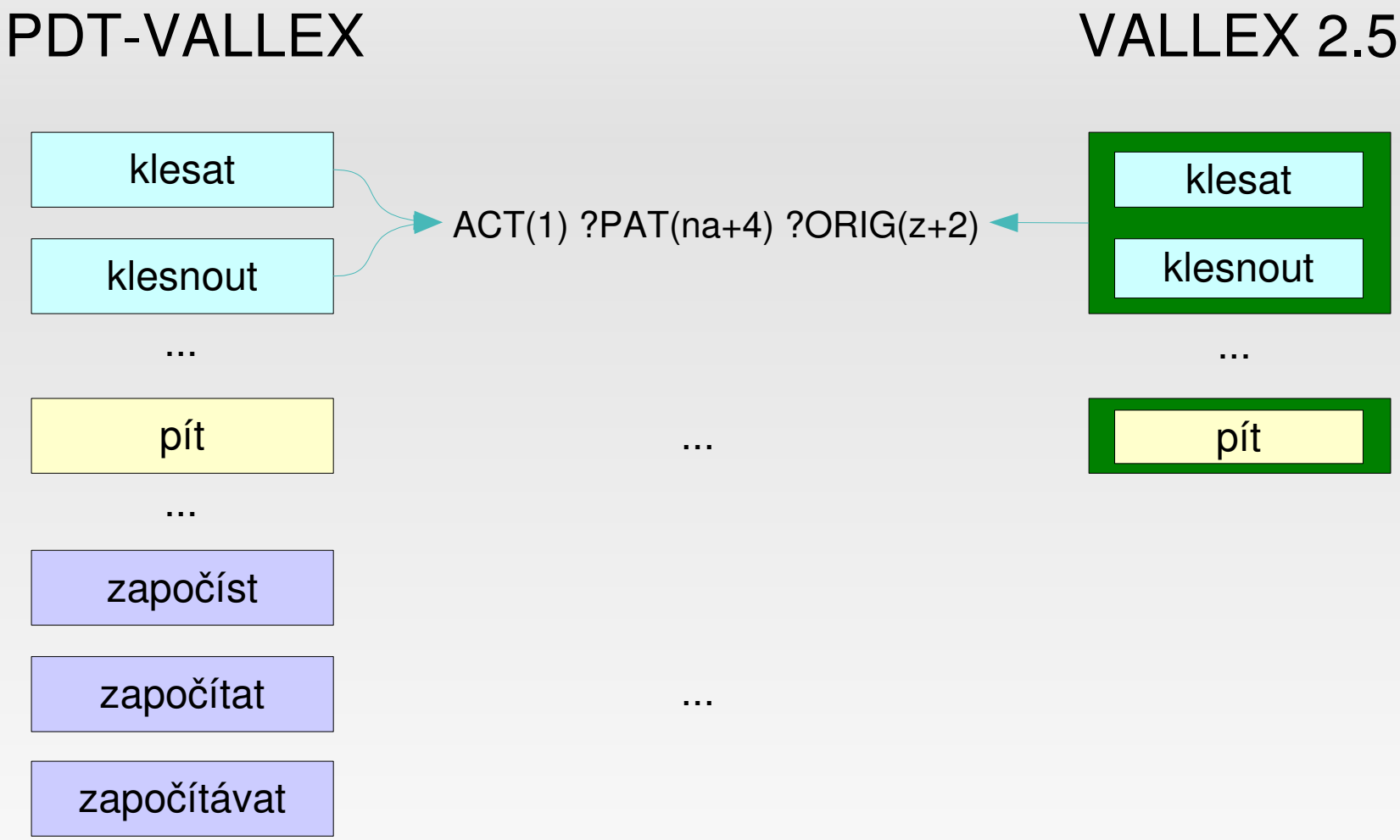
Unit: Lemma vs. Lexeme



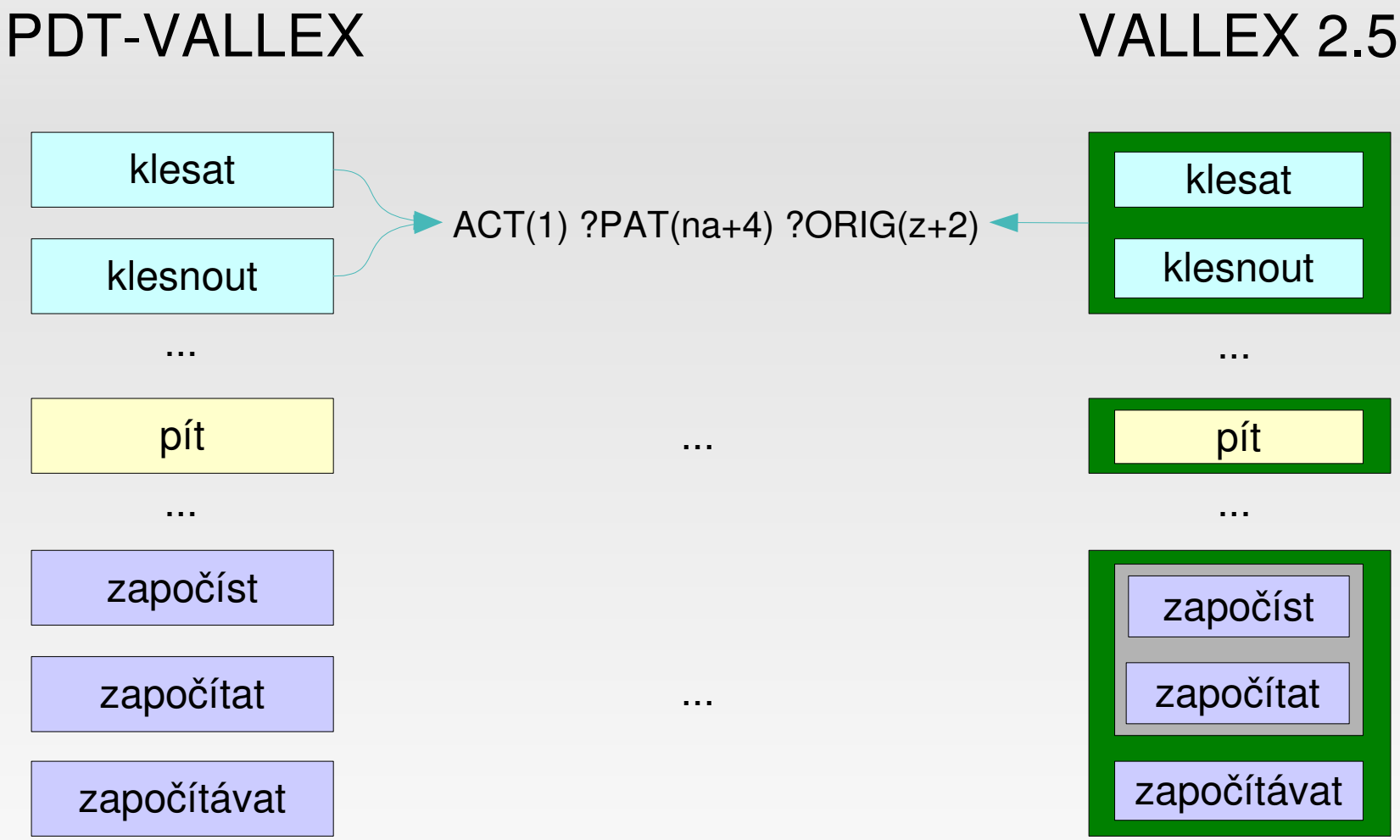
Unit: Lemma vs. Lexeme



Unit: Lemma vs. Lexeme

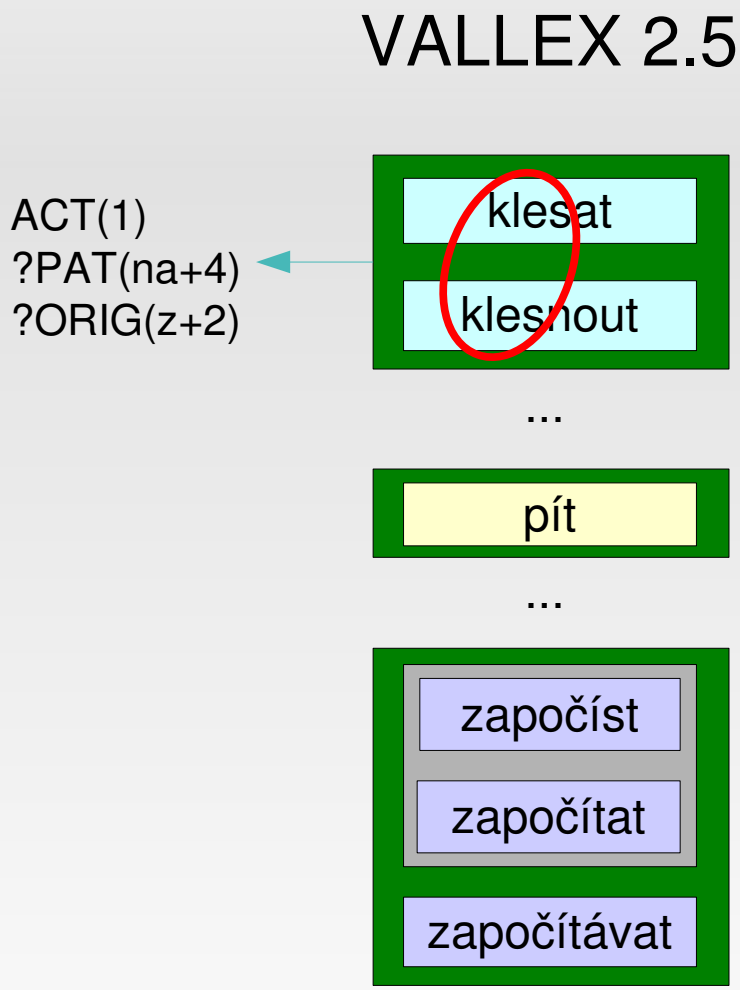


Unit: Lemma vs. Lexeme



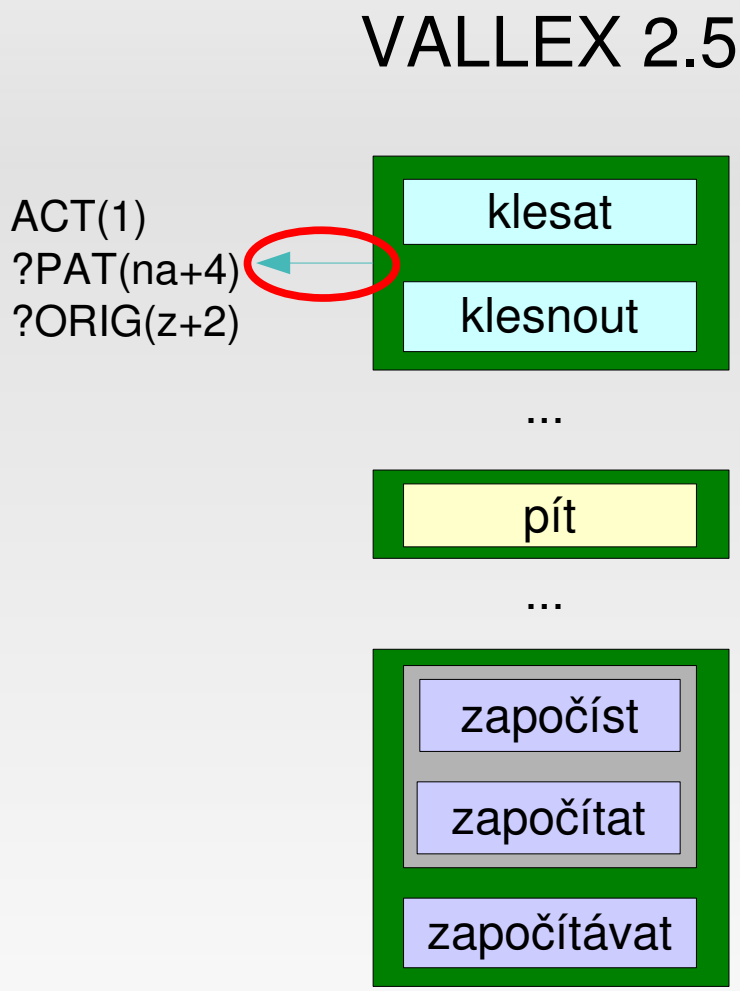
Why lexemes?

- Advantages
  - theoretical adequacy
  - compact representation
  - convenience for human users
- Disadvantages
  - complex format, hard to process



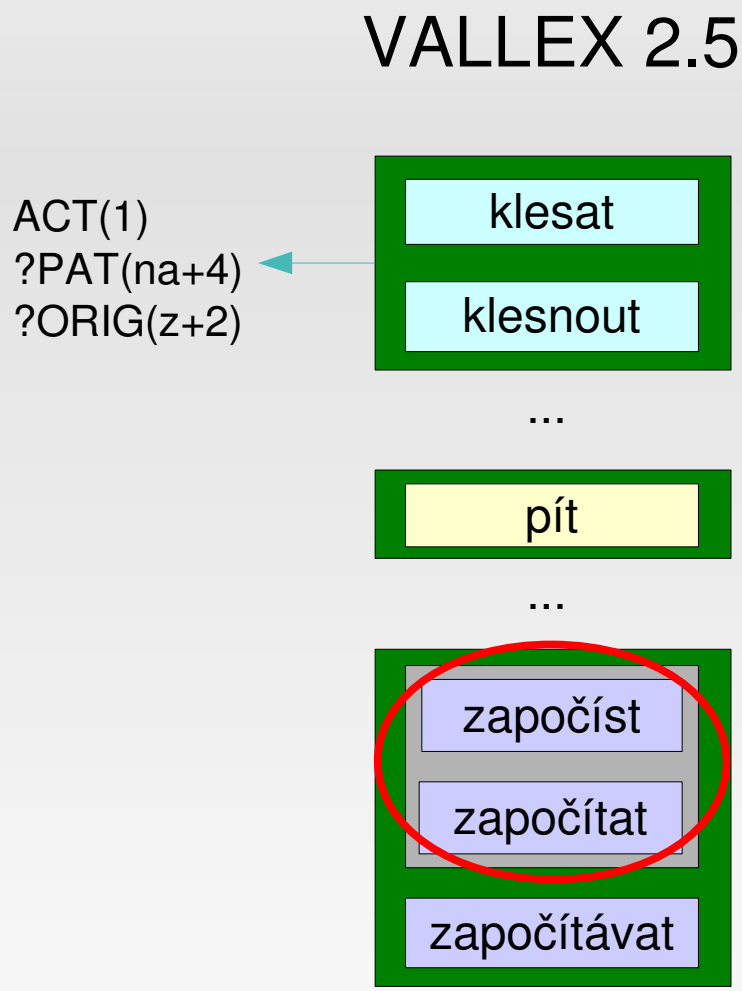
Why lexemes?

- Advantages
  - theoretical adequacy
  - compact representation
  - convenience for human users
- Disadvantages
  - complex format, hard to process



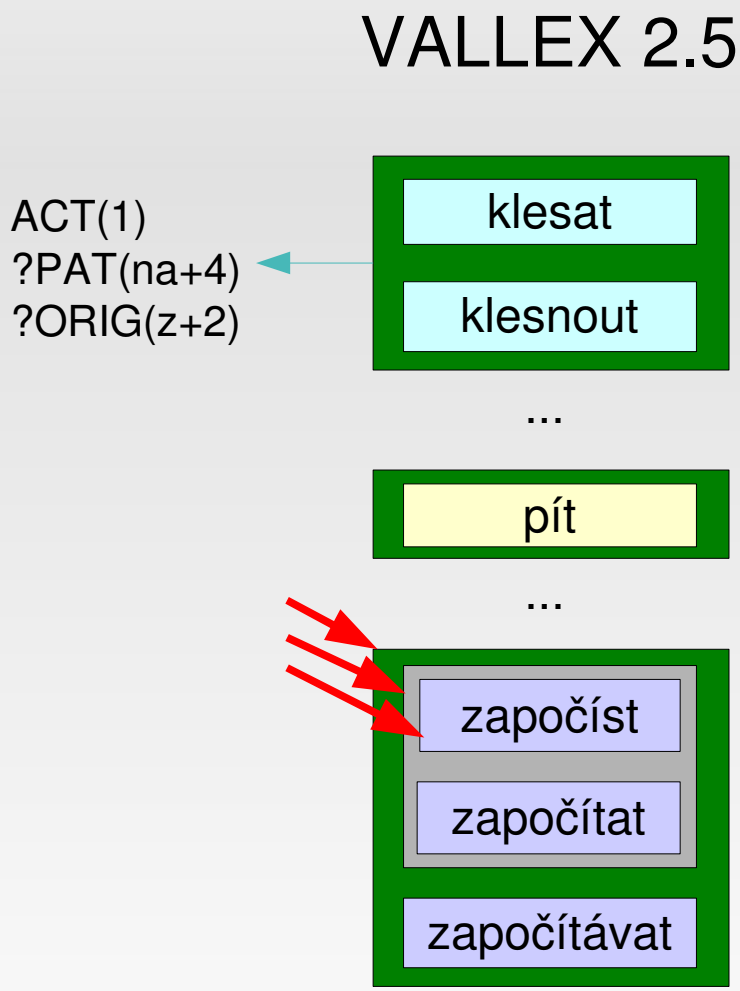
Why lexemes?

- Advantages
  - theoretical adequacy
  - compact representation
  - convenience for human users
- Disadvantages
  - complex format, hard to process

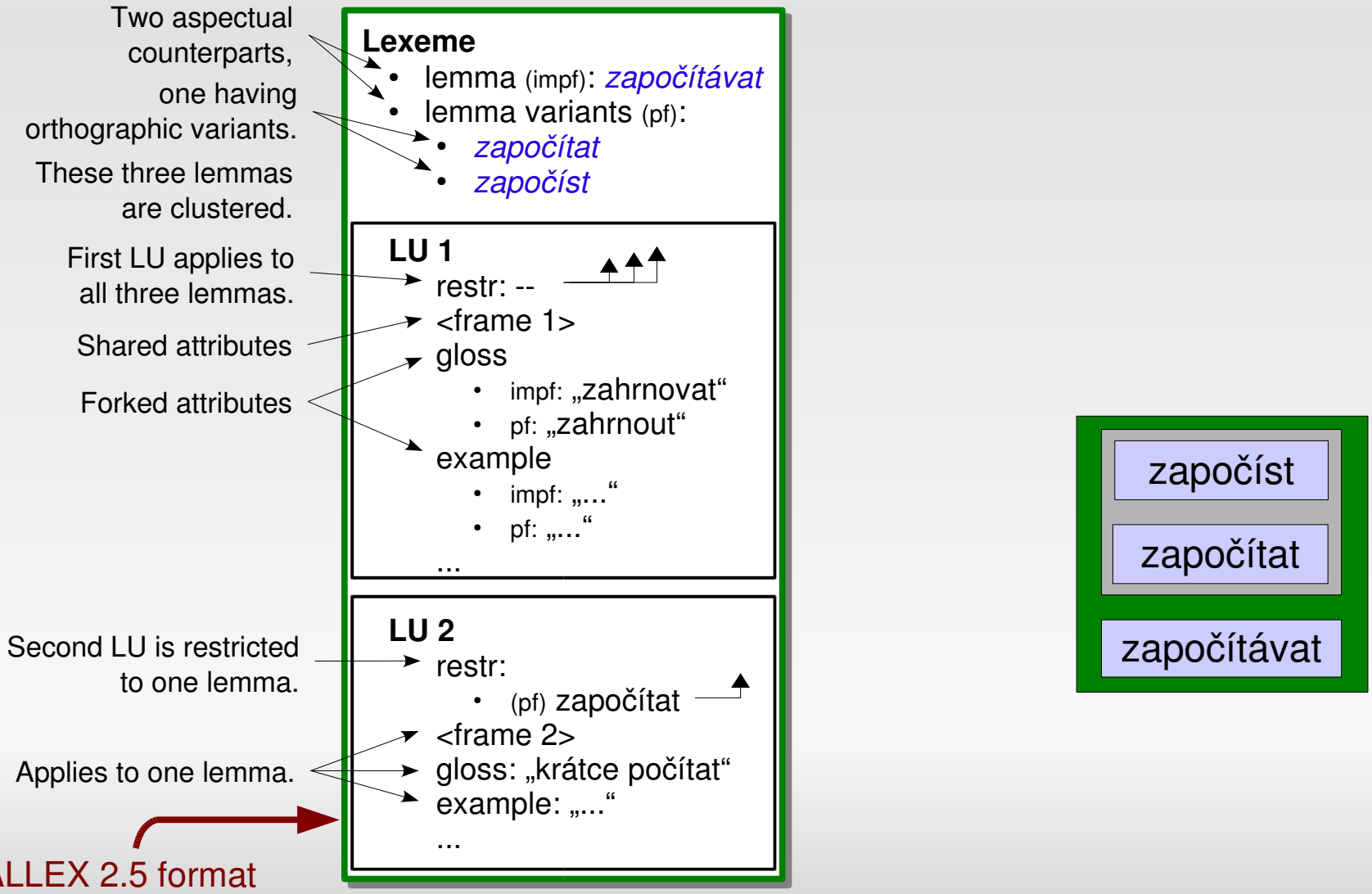


Why lexemes?

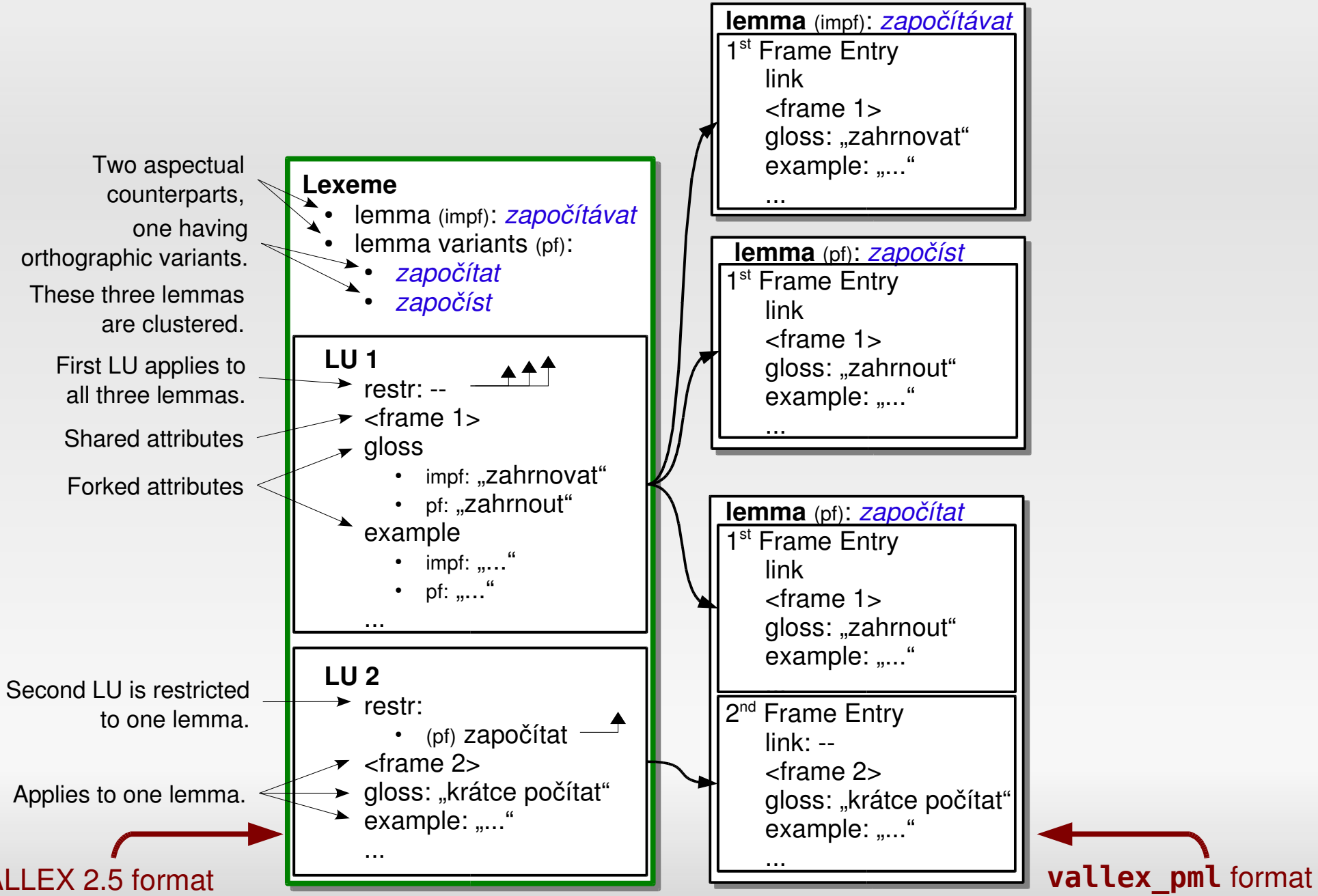
- Advantages
  - theoretical adequacy
  - compact representation
  - convenience for human users
- Disadvantages
  - complex format, hard to process



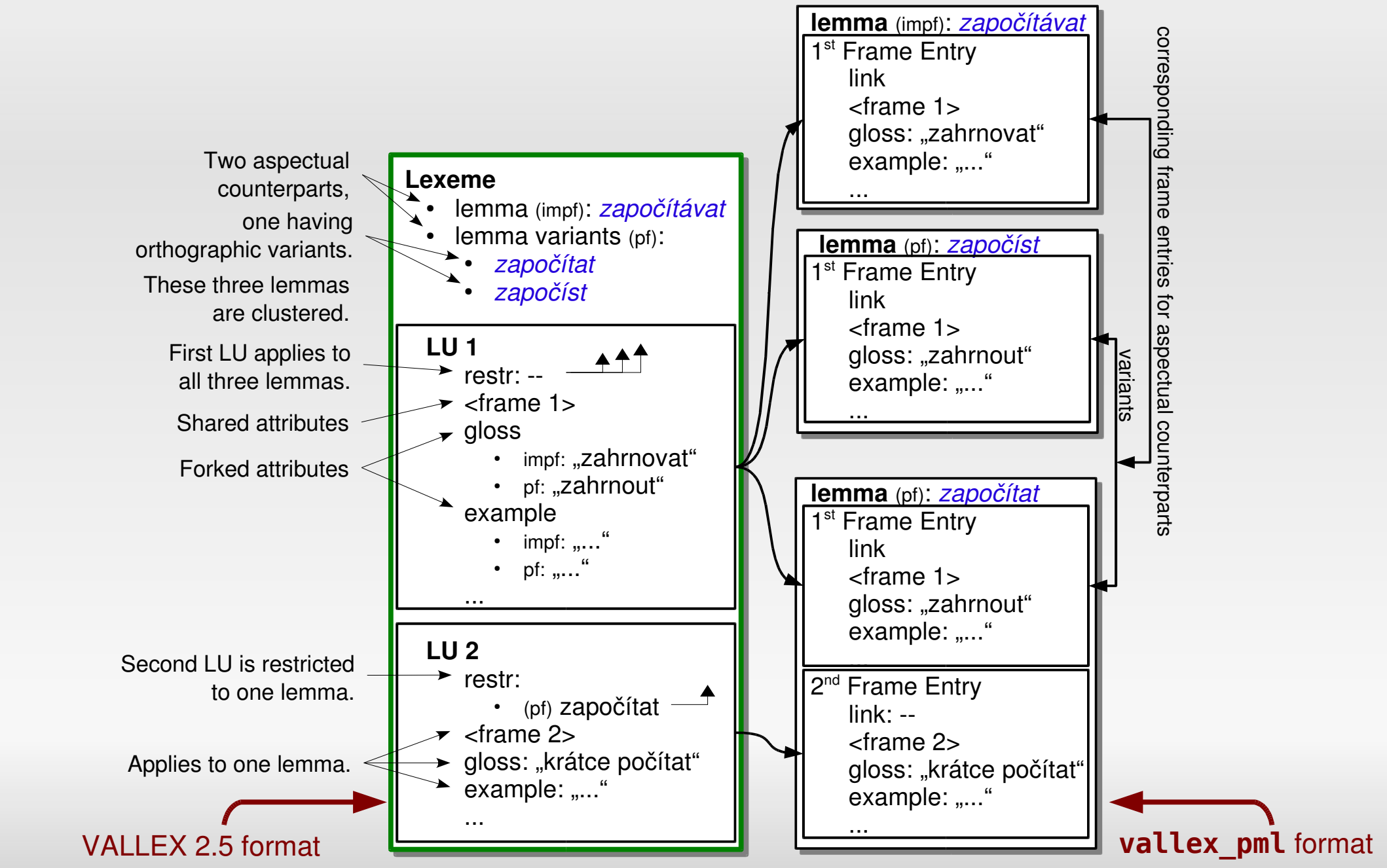
The common vallex_pml format



The common vallex_pml format



The common vallex_pml format



Prague Markup Language

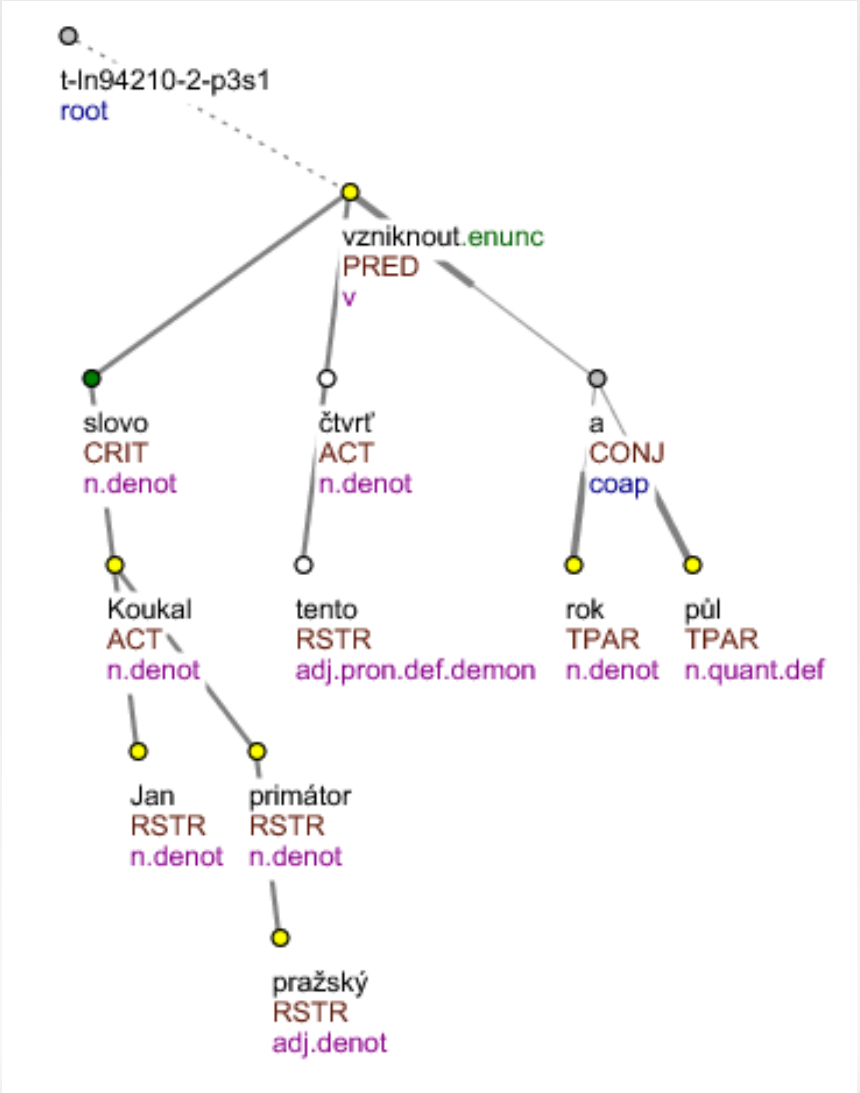
- Lexicon transformed into PML format
  - http://ufal.mff.cuni.cz/jazz/PML
  - XML-based
  - described by a PML-schema
    - automatically from DTD
    - PML roles (next slide)

Prague Markup Language (2)

- Reasons for PML
  - suitable common representation
  - can be displayed as a tree in the tree editor TrEd
    - PML roles (define nodes etc.)
  - and searched by PML-TQ

Prague Markup Language (2)

- Reasons for PML
  - suitable common representation
  - can be displayed as a tree in the tree editor TrEd
    - PML roles (define nodes etc.)
  - and searched by PML-TQ

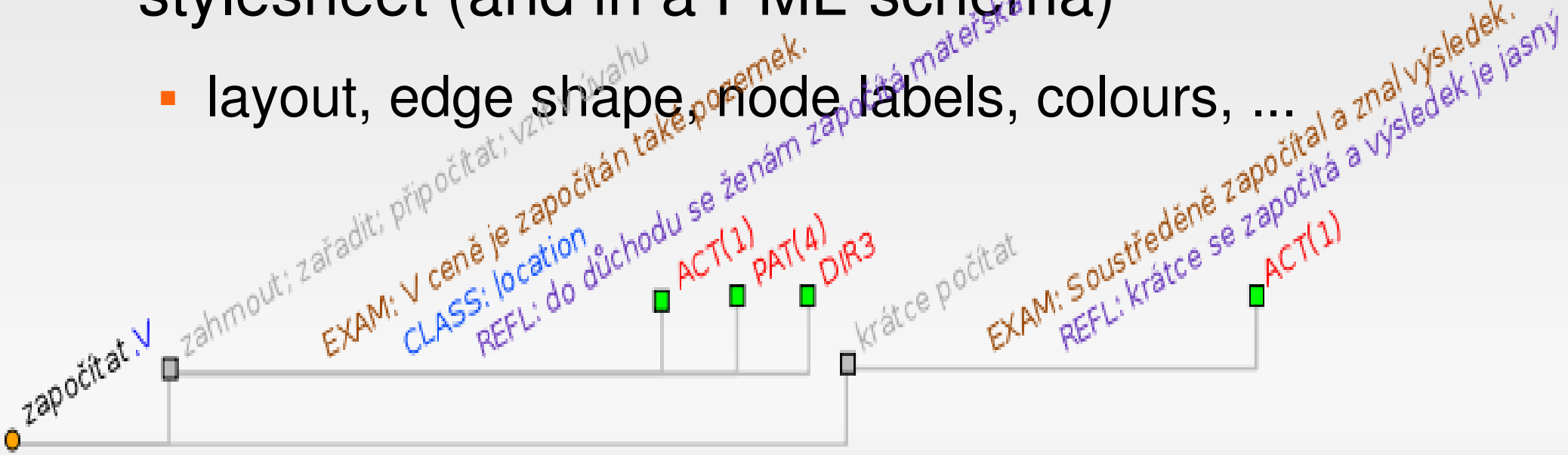


What can we do with lexicons?

- A) Display, go through
  - each lexicon entry as one "tree"
- B) Search, display the search
  - query in the form of a tree with constrains
- C) Interlink, display these links
  - links within a single lexicon
  - links to the other lexicon

A) Viewing in TrEd

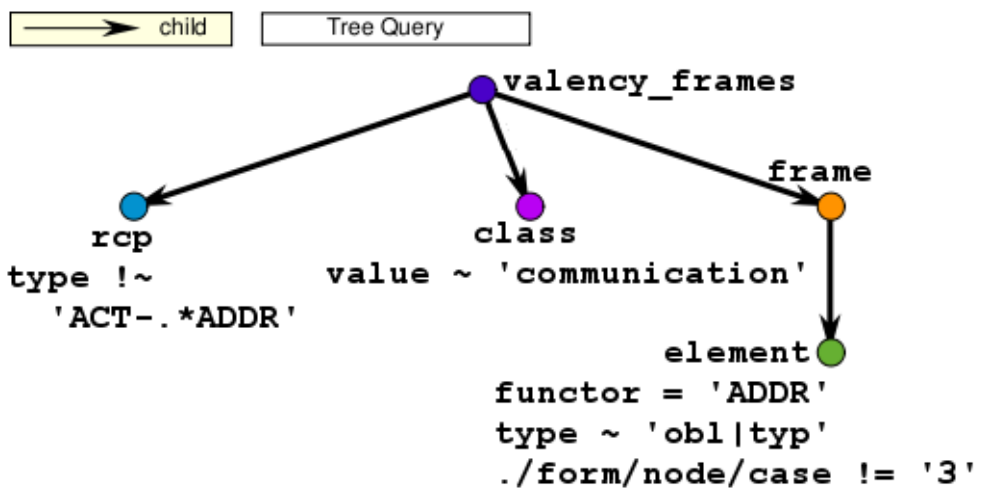
- TrEd http://ufal.mff.cuni.cz/~pajas/tred/
- lexicon entry (= a verb) as a tree
- the specific appearence is defined in a stylesheet (and in a PML-schema)
  - layout, edge shape, node labels, colours, ...



B) Searching by PML-TQ

- PML Tree Query
  http://ufal.mff.cuni.cz/~pajas/pmltq/
- a query created in a graphical mode – by drawing a subtree (with possible constraints) to search

- verbs from the class 'communication' with obligatory ADDR realized in other than dative case that cannot be in a reciprocity relation with ACT



B) Searching by PML-TQ (2)

- textual alternative to each query
- further processing ("filters")
  - frequency of found lemmas in indiviual verb classes
  - can be appended to any query
  - generate simple text table

C) Interlinking

- Three types
  - aspectual counterpart
  - lemma variant
  - same lemma/frame in the other lexicon (future work)
- view the linked lemma/frame in TrEd

Conclusion

- format for linking valency lexicons
- universal method, adaptable for other lexicons
- any tree-like structure (XML based lexicon)
- necessary modifications
  1. PML-schema from DTD
  2. assignment of PML roles
  3. stylesheet
  4. (transfer into the database)

Thank you.

This work has been supported by projects MSM 0021620838 and LC536 of the Ministry of Education and grants GAUK 4200/2009 and partially GAUK 52408/2008 of the Grant Agency of Charles University in Prague.