A treebank for everyone

Milena Hnátková, Petr Jäger, Tomáš Jelínek,
Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová

Ústav teoretické a komputační lingvistiky
Filozofická fakulta Univerzity Karlovy v Praze

Institute of Theoretical and Computational Linguistics
Faculty of Arts, Charles University, Prague

Zespół Lingwistyki Korpusowej Języków Słowiańskich
Instytut Slawistyki Zachodniej i Południowej
Uniwersytet Warszawski
19 czerwca 2012

Outline of the talk

Outline of the talk

About treebanks

- Treebank = a corpus annotated with syntactic structure
- Probably first major project: Penn Treebank (release v. 0.5, 1992)
- Now: 74 treebanks in about 40 languages (Wikipedia)
- Different in:
  - size
  - linguistic background
  - format
  - level of detail
  - depth of analysis
  - ways they are built
- Parallel treebanks, semantic annotation, ...

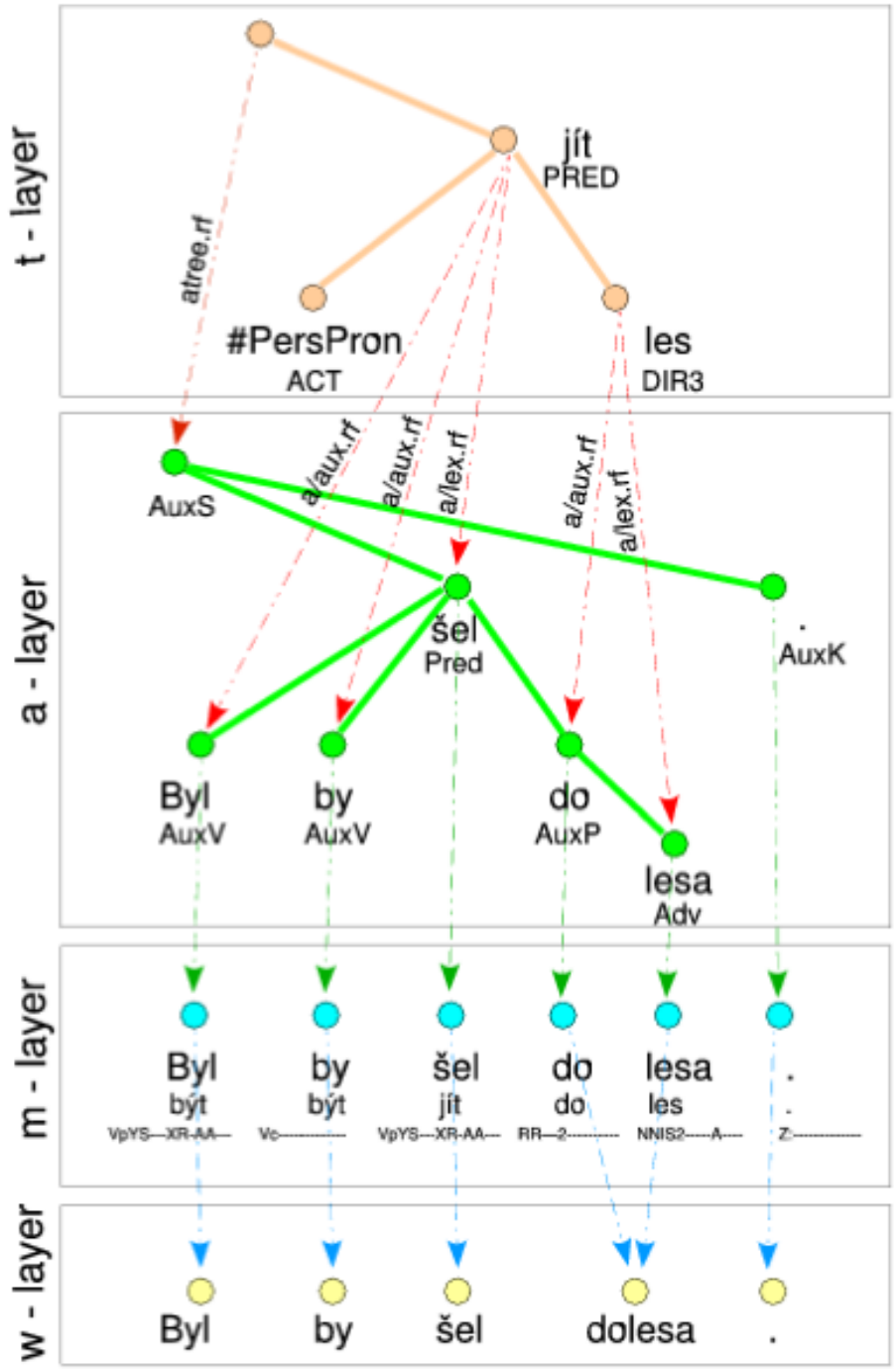
Why are treebanks useful?

- Explicit markup of syntactic relations (constituents; heads and dependents)
- – easier to identify semantic relations (predicates/functors and arguments)
- – simplifies some queries
- – simplifies extraction of lexical properties (valency) or syntactic rules
- – support for grammar development

Treebanks of Czech

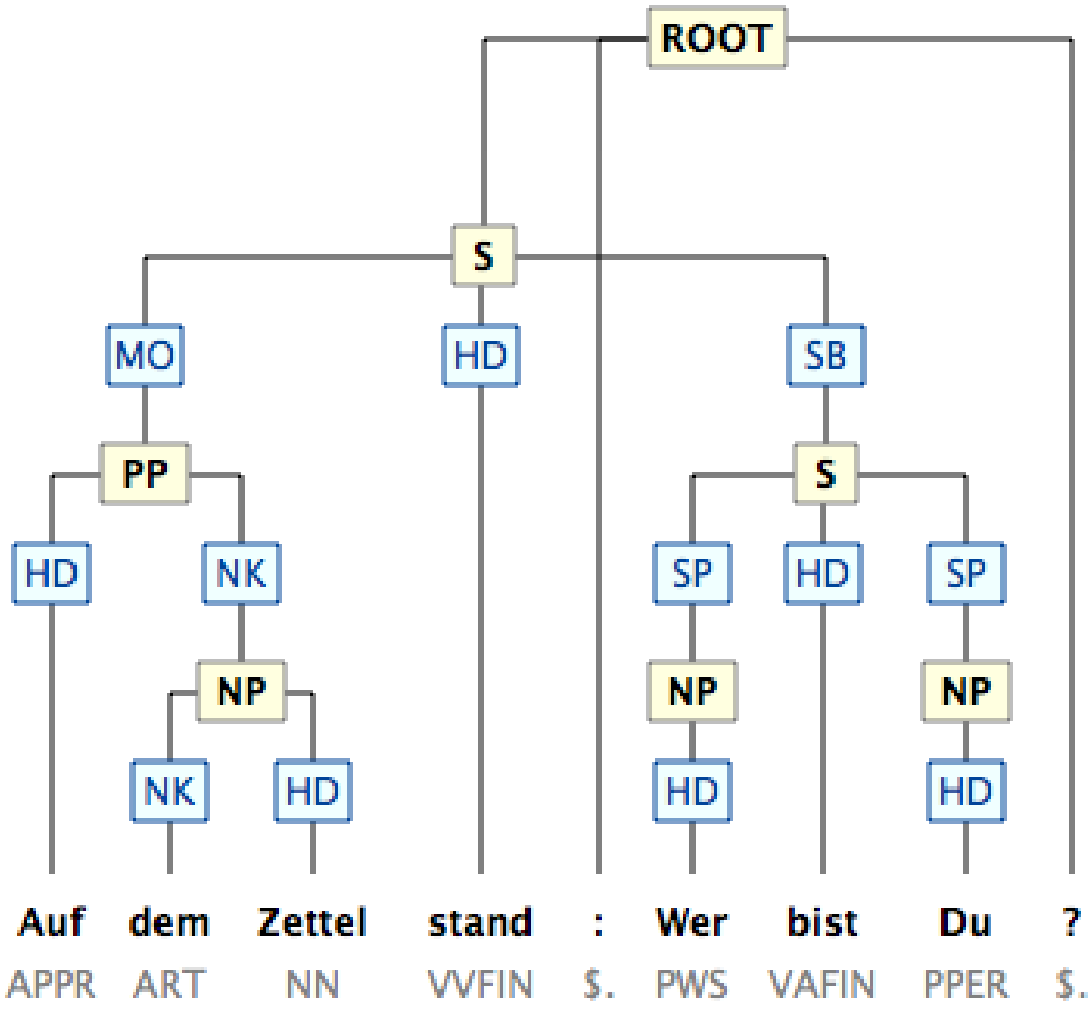
Prague Dependency Treebank

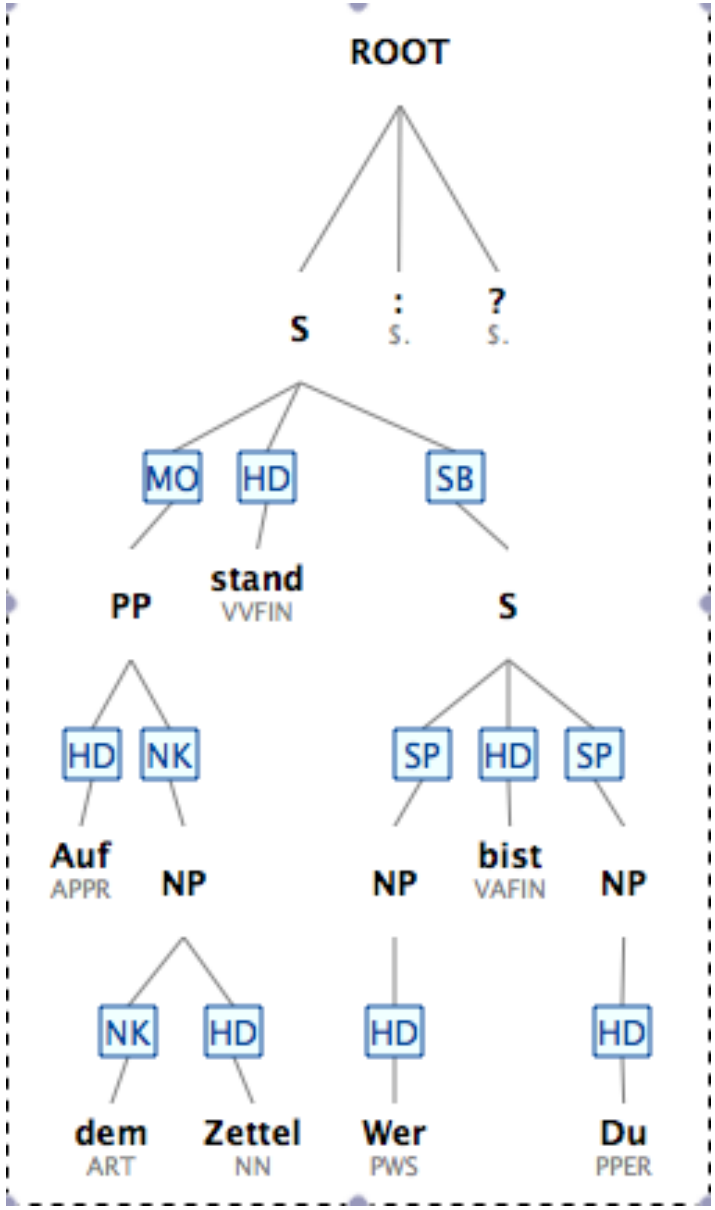
- Dependency syntax, close to the Prague theory of Functional Generative Description [Sgall et al.(1986)]
- 3 annotation levels: morphology, surface syntax, deep syntax
- PDT 0.5 – 1998, 0.5M tokens
- PDT 1 – 2000, 1.5M tokens
- PDT 2 – 2004, deep syntax

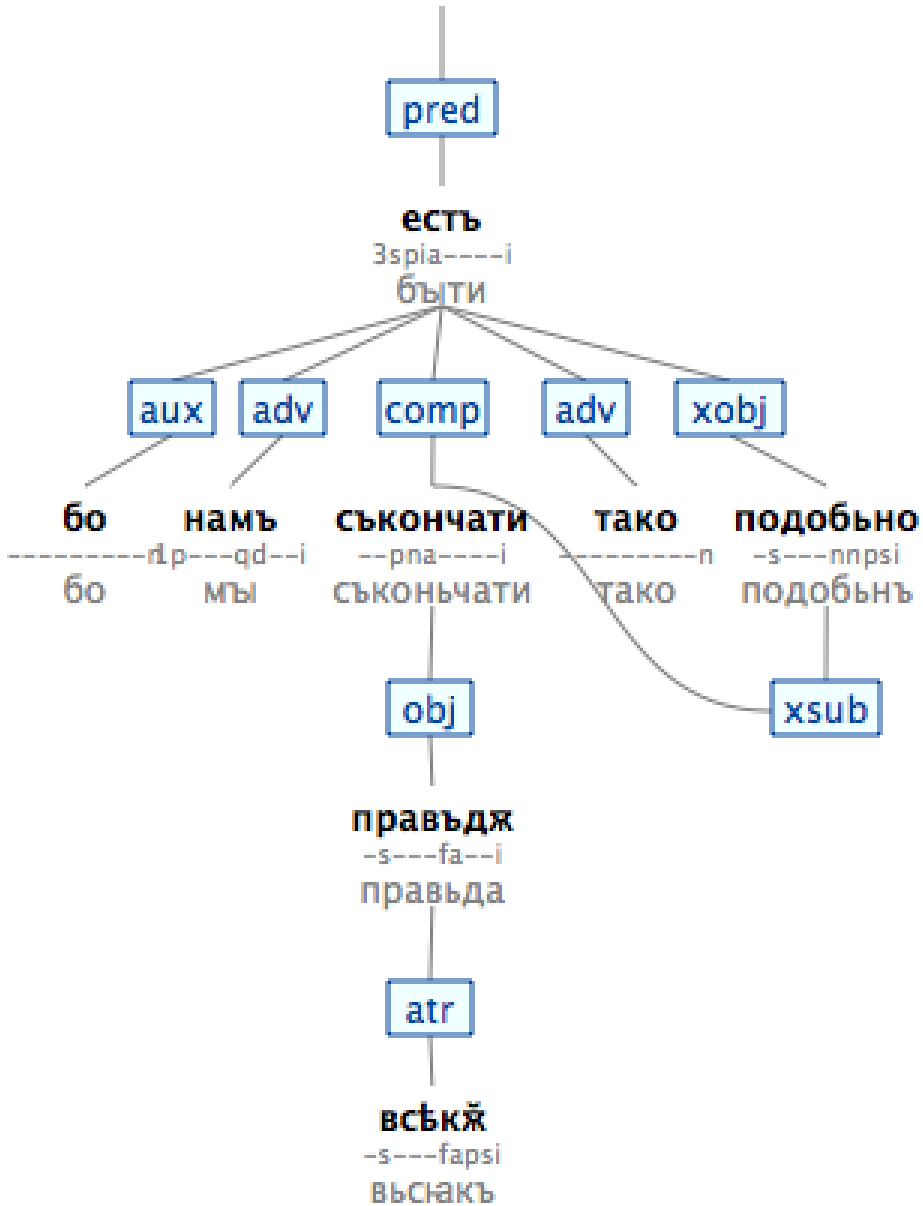

t - layer
atree.rf
jít
PRED
#PersPron
ACT
les
DIR3
a - layer
AuxS
a/aux.rf
a/aux.rf
a/lex.rf
a/aux.rf
a/lex.rf
šel
Pred
.
AuxK
Byl
AuxV
by
AuxV
do
AuxP
lesa
Adv
m - layer
Byl
být
VpYS---XR-AA---
by
být
Vc----------
šel
jít
VpYS---XR-AA---
do
do
RR--2----------
lesa
les
NNIS2-----A----
.
.
Z:----------
w - layer
Byl
by
šel
dolesa
.

Some other treebanks

- Polish Constituency Treebank
- BulTreeBank
- Iness Treebanking Infrastructure
- Lingo Redwoods
- Tiger
- LASSY
- ...


ROOT
S
MO
HD
SB
PP
S
HD
NK
SP
HD
SP
NP
NP
NP
NK
HD
HD
HD
Auf
dem
Zettel
stand
:
Wer
bist
Du
?
APPR
ART
NN
VVFIN
$.
PWS
VAFIN
PPER
$.


ROOT
S
:
S.
?
S.
MO
HD
SB
PP
stand
VVFIN
S
HD
NK
SP
HD
SP
Auf
APPR
NP
NP
bist
VAFIN
NP
NK
HD
HD
HD
dem
ART
Zettel
NN
Wer
PWS
Du
PPER


pred
есть
3spia----i
быти
aux
adv
comp
adv
xobj
бо
---------d1p---qd--i
бо
намъ
мы
съкончати
--pna----i
съконьчати
тако
--------n
тако
подобьно
-s---nnpsi
подобьнъ
obj
xsub
правъдж
-s---fa--i
правьда
atr
вс҃кѫ
-s---fapsi
вьсѧкъ

Why isn't PDT good enough?

- Too small for investigating less frequent forms and phenomena
- Theoretical bias — distracts some people used to constituency structure

How to build a large treebank?

- Treebanks built with human assistance,
  even semi-automatically, are too small.
- They can be built by a stochastic parser.
- Or by a rule-based parser,
  with the advantage of correspondence to the grammar

Can a single core annotation be viewed in different ways?

- Theory-specific representations have different appearances but share a large part of content: constituency/dependency, morphosyntactic categories, even the spirit of analyses of many phenomena
- A treebank offering different views of a sufficiently expressive annotation scheme is a realistic goal
- Additional benefit: relating linguistic theories

A larger treebank with customizable visualization?

A project (2010—2012) aiming at:

- Syntactic annotation of the Czech National Corpus (1.3 billion words) using a stochastic parser, followed by a rule-based correction module
- Robust and expressive core annotation format, potentially underspecified
- Customizable query, visualization and export interface, offering multiple options to view syntactic structure
- Accessible to lay users and satisfying experts at the same time

Hopefully a follow-up project (2013–2015) aiming at:

- Development of a corpus-based grammar
- Options for queries, visualization and export:
  - ready-made, tailored to specific theories, or
  - definable by the user
- Improvements of the correction module

A corpus grammar?

- to bridge the gap between the empirical and the theoretical
- for grammar development
- for checking consistency
- for adding more info
- for assisting the treebank user
- to help converting the data onto other formats more easily

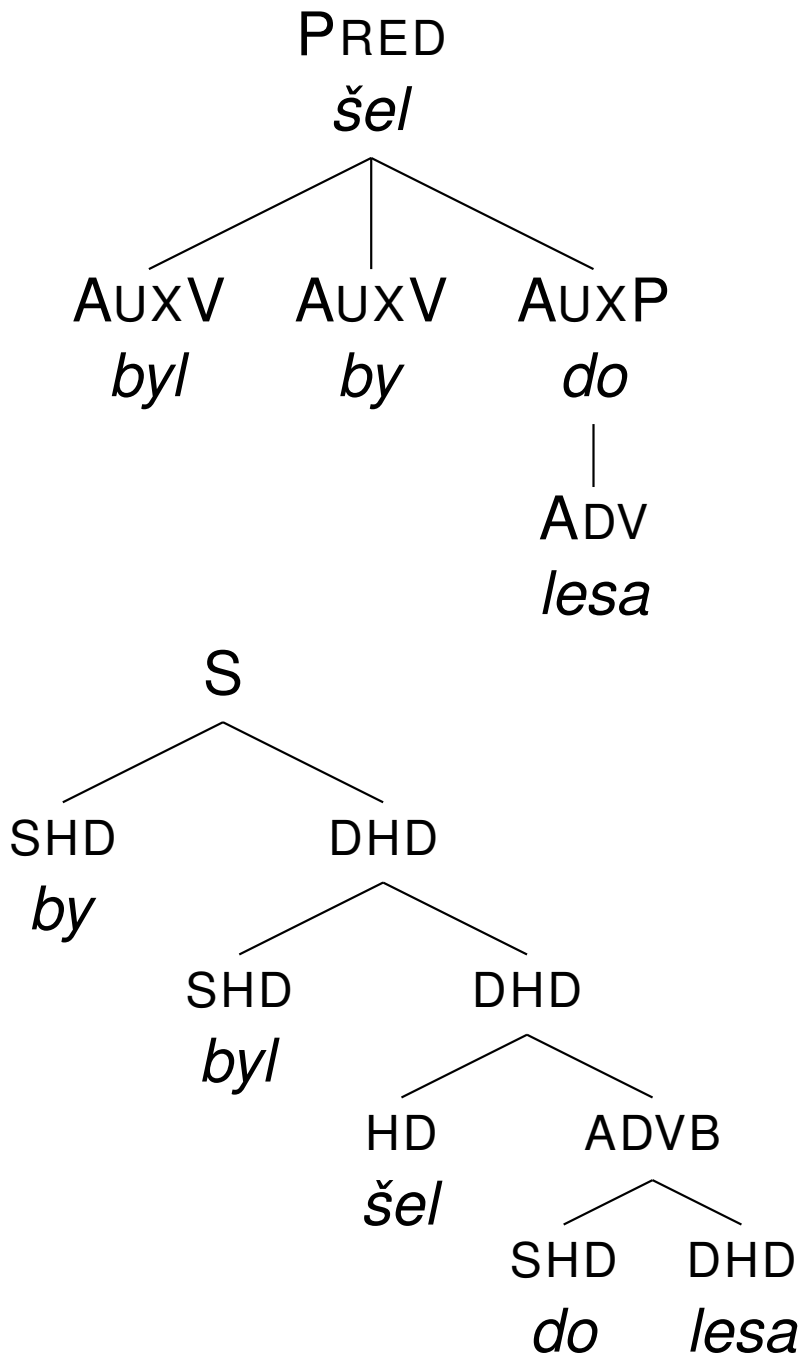
A corpus grammar can be:

- hand-crafted
- extracted from corpus
- hand-crafted but verified against the corpus data
  advantage: incremental development

Outline of the talk

Syntactic structure

- Internal skeleton structures: constituency-based, with a combination of **binary** and **flat** branching
- Interpretable as **constituency** or **dependency** trees, according to users' specification, visualized with an arbitrary amount of detail, not necessarily by tree graphs
- Surface and deep structure encoded within a single structure: constituents are labelled as **syntactic functions** including **head** as a special function
- Heads are further specified as **deep** or **surface**
  - **Deep head**: deep syntactic governor: *bylo by se to **povedlo***
  - **Surface head**: can be identical to the deep head or different: auxiliary, prepositions, subordinate conjunctions, numerals


PRED
šel
AUXV
byl
AUXV
by
AUXP
do
ADV
lesa
S
SHD
by
DHD
SHD
byl
DHD
HD
šel
ADVB
SHD
do
DHD
lesa

Three levels

- Word order and syntactic structure as distinct dimensions, each sentence is represented at three inter-linked levels:
  - **graphemics** (orthographic words, contractions)
  - **morphology** (syntactic words, including haplologized items)
  - **syntax** (trees, no nodes for pro-dropped subjects)

Annotation of syntactic phenomena

- Agreement of various types
- Compound periphrastic verbal forms (passives, conditional structures, future...)
- Grammatical co-reference (grammatical control, relative/reflexive pronouns, predicative complements)
- Multi-word units (collocations)

Expressive power

- Expressive enough to accommodate analyses of arbitrary granularity
- Ambiguous or undecidable phenomena represented by underspecification and distributive disjunction
- Annotation of any kind can be missing, a sentence may be a mere list of words

Specifications

- Annotation must be licensed by a formal grammar. Words and constituents have their appropriate (potentially underspecified) sets of features
- Lexicons are used to index forms, syntactic words and compound forms
- Customizable visualizations are enabled by formal definitions

Representation layers:

- text
- morphology
- syntactic structure

Lexicons:

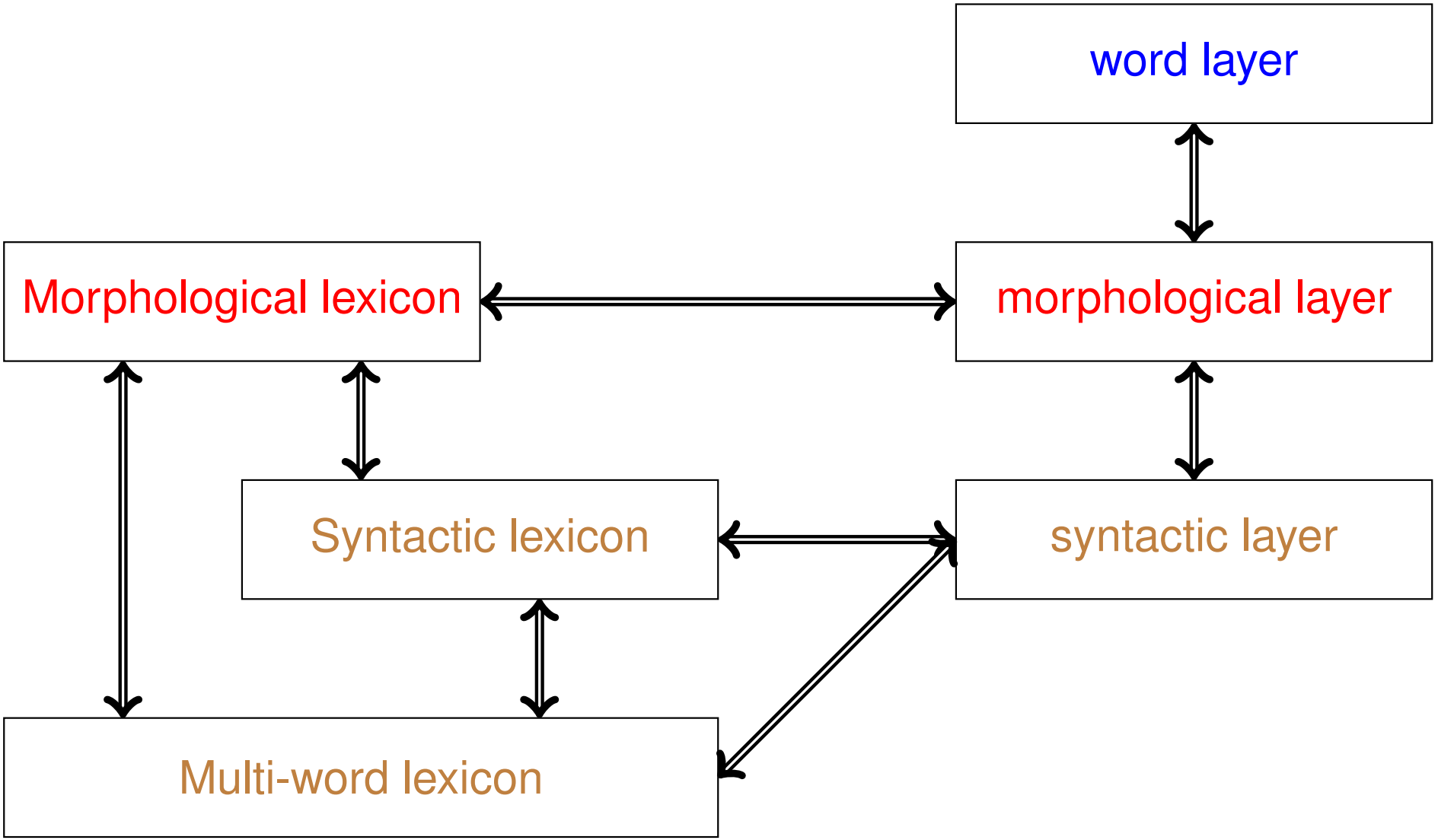
- morphological
- syntactic
- multi-word expressions

Two-way links between layers, and between layers and lexicons

- to link information across the layers
- to provide lexeme-specific information
- to identify multi-word expressions, including periphrastic forms

Links within a tree

- Agreement
- Compound (multi-word) verbal predicates
- Grammatical coreference
- ...


word layer
Morphological lexicon
morphological layer
Syntactic lexicon
syntactic layer
Multi-word lexicon

Word layer

- tokenized, including punctuation and MWE: *česko-slovenský*
- contractions left intact (not interpreted): *očs, ses*

Morphological layer

- morphological analysis and lemmatization of all forms
- contractions split (i.e. interpreted): *očs* → *o co jsi*
- some punctuation marks glued back with word forms: *atd.*, *česko-*

Syntactic layer

- **constituency-based structure**, representable according to user's options/specifications
- punctuation omitted

Syntactic structure

- each nonterminal node is assigned a construction type and a syntactic function
- each terminal node is assigned a syntactic function

Hierarchy of construction types

- **Headed**
- **UnHeaded**
  - **Coord** – coordination
  - **Adord** – adordination
  - **Unspec** – unspecified (for collocations and other)

Function for **UnHeaded** structures:

- **Memb** – a member

Syntactic functions for **Headed**

- **SurfHead** – surface head: auxiliary *být/bývat*, prepositions, subordinate conjunctions, numerals in quantified expressions: *pět dětí*
- **DeepHead** – in case it differs from SurfHead (head nouns in PPs, autosemantic verbs in analytical predicates...)
- **Head** – both **SurfHead** and **DeepHead**

Other syntactic functions for **Headed**

- **Subj** – subject
- **Attr** – attribute
- **Obj-Advb**
  - **Obj**
  - **Advb**
- **VbAttr** – predicative complement
- **ReflTant** – reflexive element (*si*, *se*) for inherent reflexives
- **Deagent** – deagentive reflexive
- **Apos** – apposition
- **InDep** – independent syntactic element (parenthesis, vocative syntactic noun...)

Morphological lexicon

- List of lemmas with inflection paradigms
- A lemma is introduced if two words differ in morphological paradigms – not only in syntactic properties or in semantics:
  - *travička* 'little grass'/'female poisoner' has only **one lemma**,
  - *člen* 'member' has two lemmas, as it is either **masculine animate** or **masculine inanimate**.

Syntactic lexicon

- list of lemmas with their syntactic properties
- inherent reflexives have separate entries – *rozhodnout* and *rozhodnout se* 'decide',
  vidět 'see' is one entry
- **valency frame** entries
- different valency frames listed under one lemma

Multi-word lexicon

- collocations: *křížem krážem* ‘in all directions’
  nechat na holičkách ‘leave in the lurch’
- types of analytical verb forms: *bych byl přišel* ‘(I) would have come’, *jsi přišel* ‘(you) have come’
- inherent reflexives: *usmíváš se* ‘(you) smile’;
  rozhodne se ‘(he) decides’
- (reflexive) deagentive construction: *jde se* ‘let’s go’
- reflexive passives: *bábovka se peče* ‘the cake is baking/being baked’
- analytical passives: *je čten* ‘is read’
- nominal predicates: *je velký* ‘is big’
- agreement patterns:
  Lucie ho viděla opilého ‘Lucie saw him drunk’

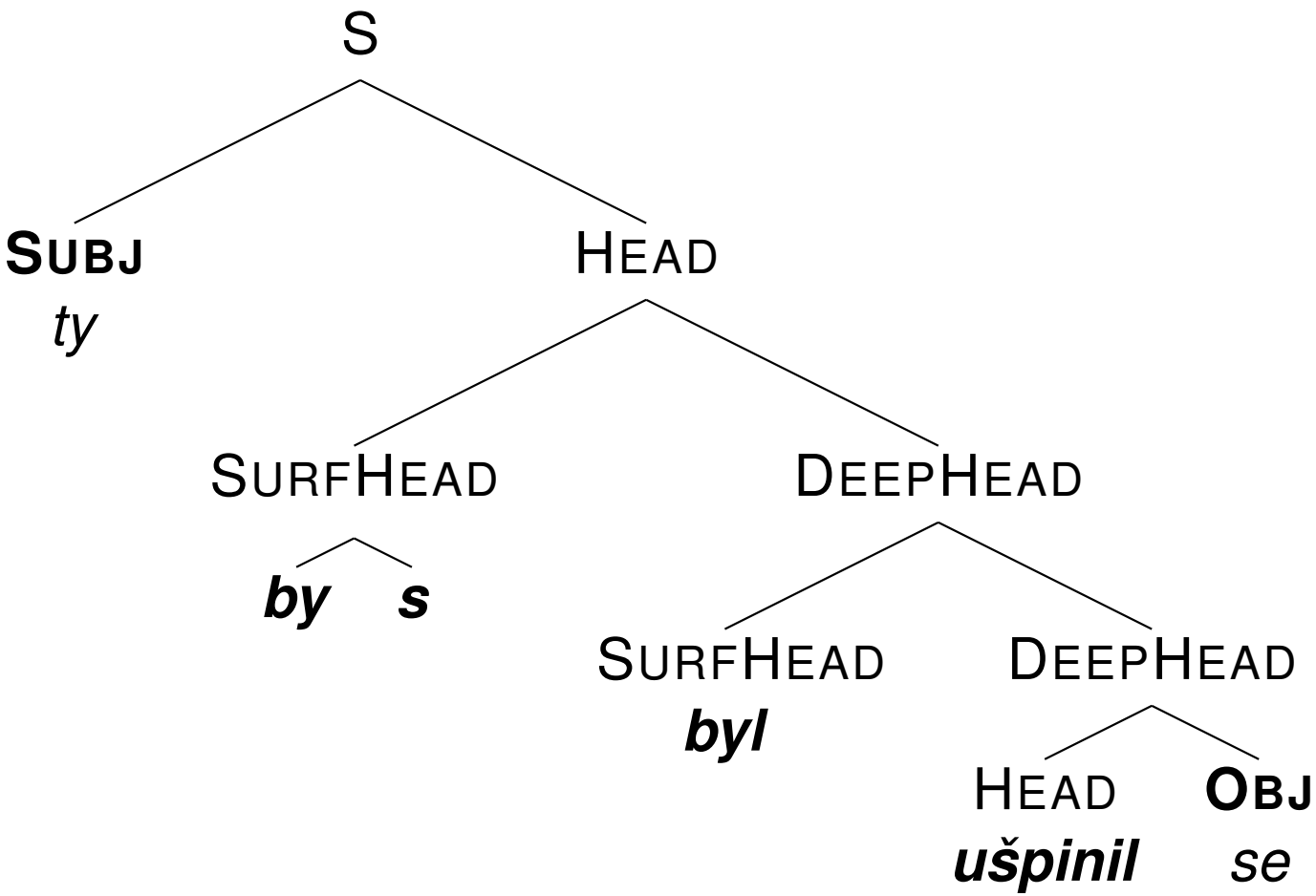
Outline of the talk

Treating contractions

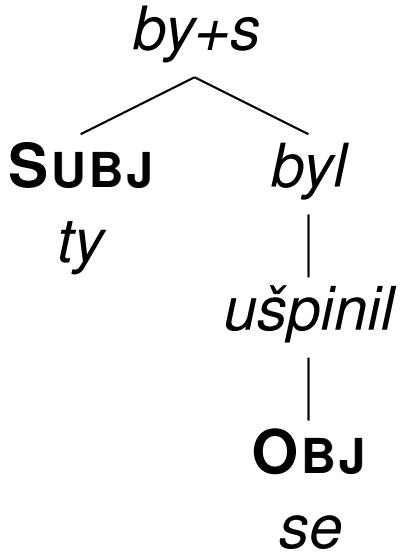
(1) **Ty** by *se*s byl ušpinil.
you would REFL+AUX$_{2nd,sg}$ be$_{pple}$ get dirty$_{pple}$
'You would have got dirty.'

Ty by ses byl ušpinil.

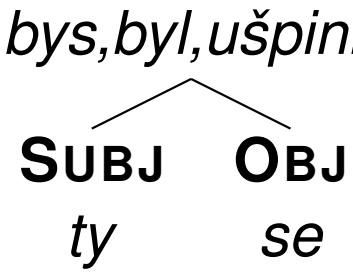
(2)



(3) **Surface dependency structure** derived from (2)



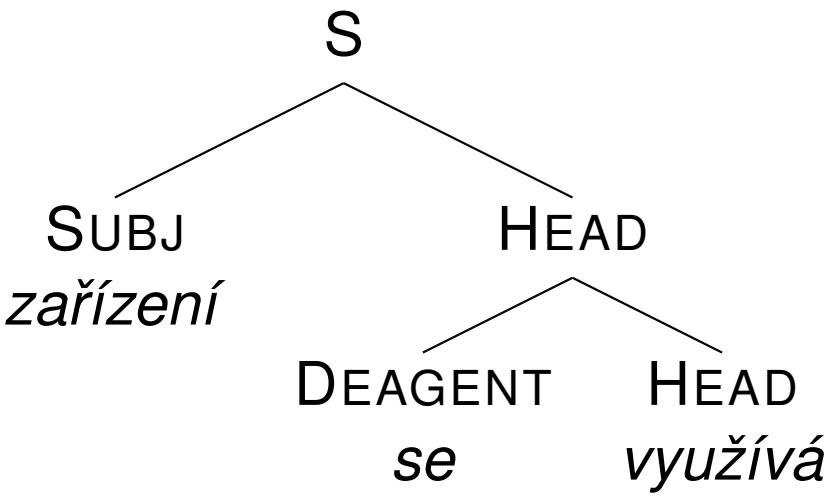
(4) **Deep dependency structure** derived from (2)

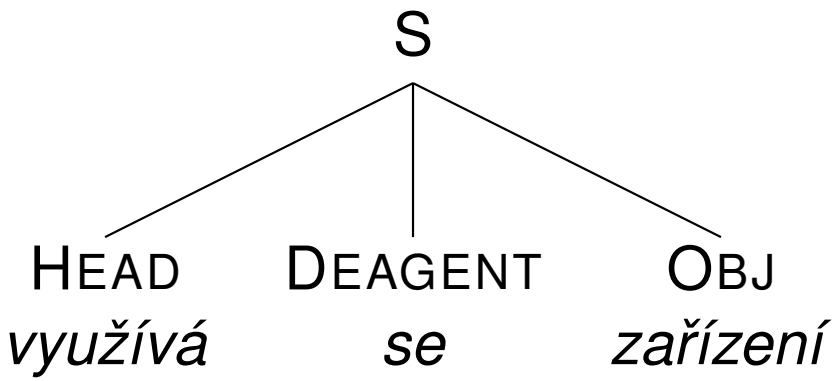


Subject/object ambiguity

Reflexive passive:

(5) Zařízení$_{Nom/Gen}$ se využívá.
device REFL uses
'The device is being used.'



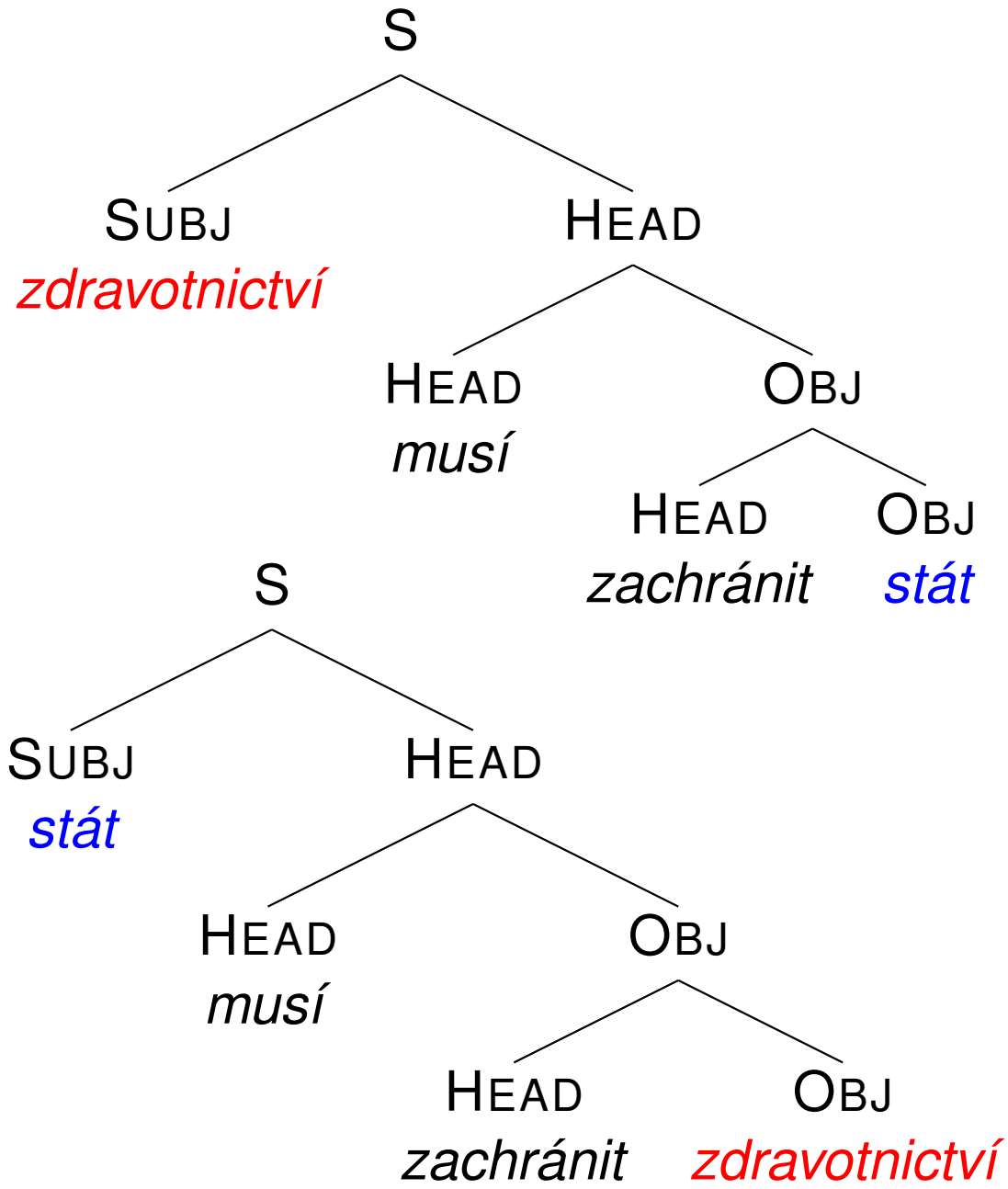


Another type of subject/object ambiguity

(6) Zdravotnictví musí zachránit stát.
health service$_{nom/acc}$ must save state$_{nom/acc}$

Two different readings:

- #1 Health service must save the State.
- #2 Health service must be saved by the government.


S
SUBJ
zdravotnictví
HEAD
HEAD
musí
OBJ
HEAD
zachránit
OBJ
stát
S
SUBJ
stát
HEAD
HEAD
musí
OBJ
HEAD
zachránit
OBJ
zdravotnictví

Morphological analysis of (6) with some values unspecified:

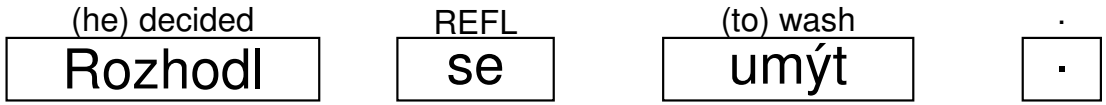
1	zdravotnictví	*noun*, CASE=*X*, NUM=*sg*, GEND=*n*
2	musí	*verbfin*, PERS=*3*, NUM=*sg*
3	zachránit	*verbinf*
4	stát	*noun*, CASE=*Y*, NUM=*sg*, GEND=*m*

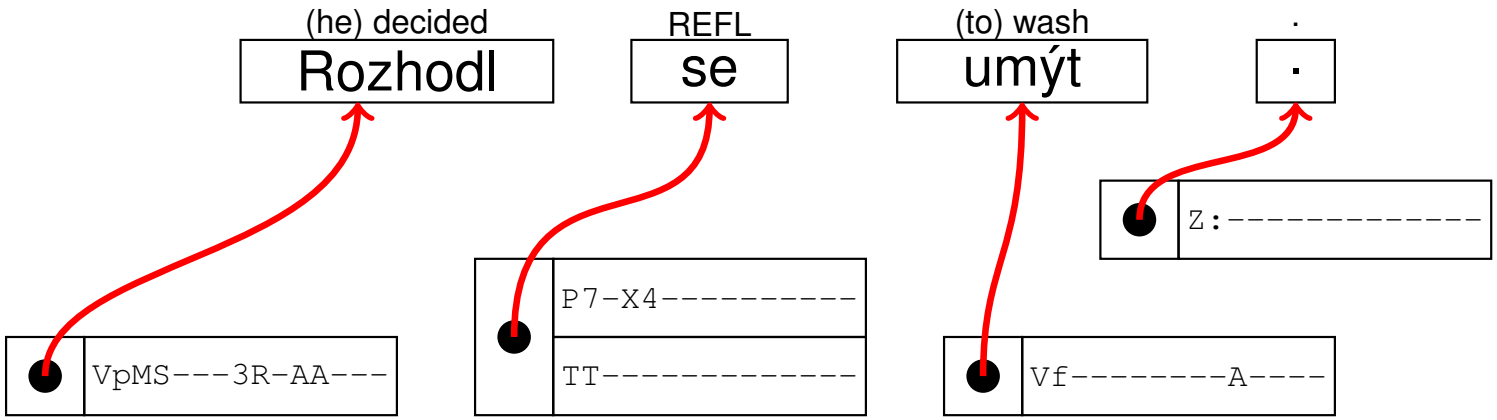
Constituents in one of the two possible syntactic structures of (6), some boxed numbers refer to the forms above:

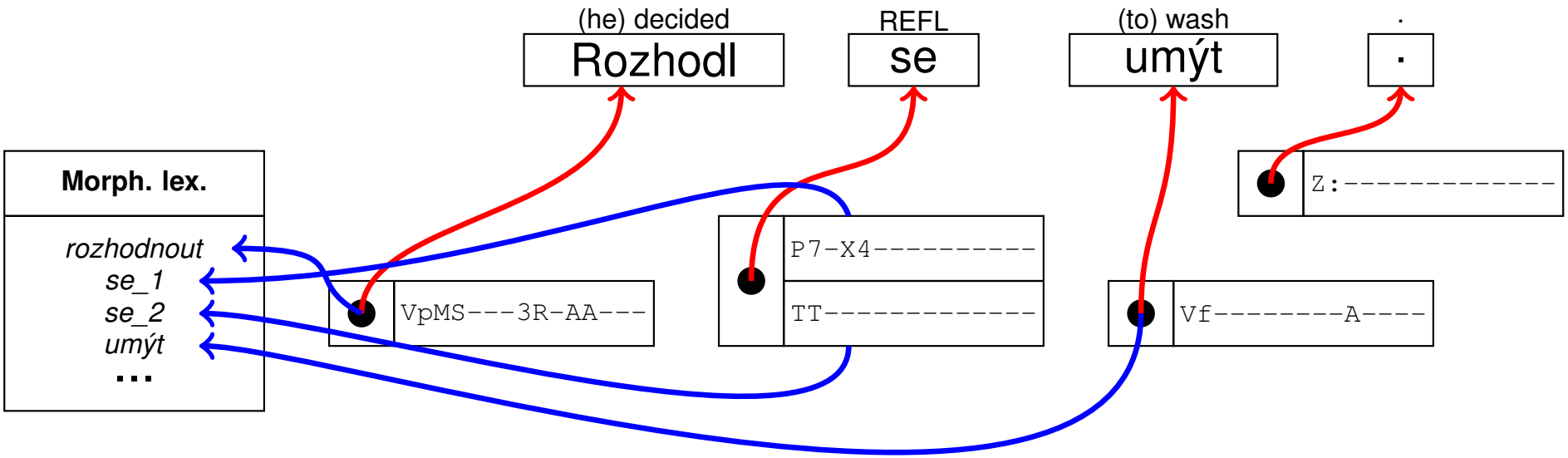
5 [3zachránit 4stát]
6 [2musí 5]
7 [1zdravotnictví 6]

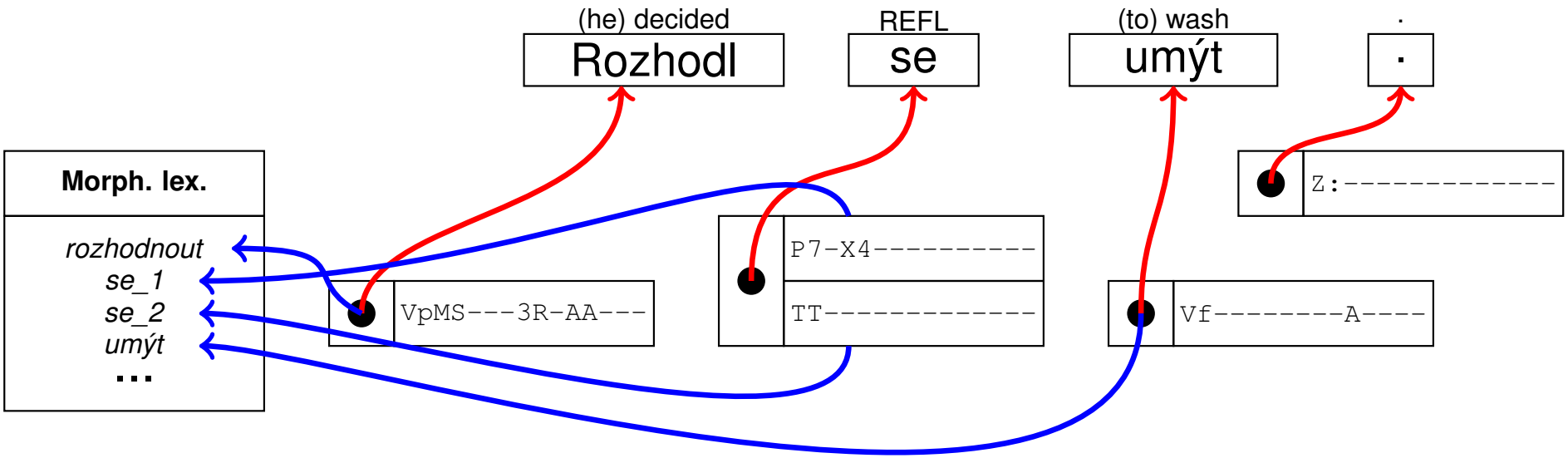
Two possible structures with constraints on category values and overriding clauses:

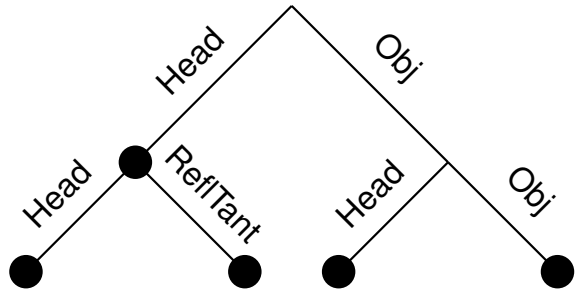
#1 = 7, *X*=*nom*, *Y*=*acc*
#2 = 7, *X*=*acc*, *Y*=*nom*, 1 → 4, 4 → 1

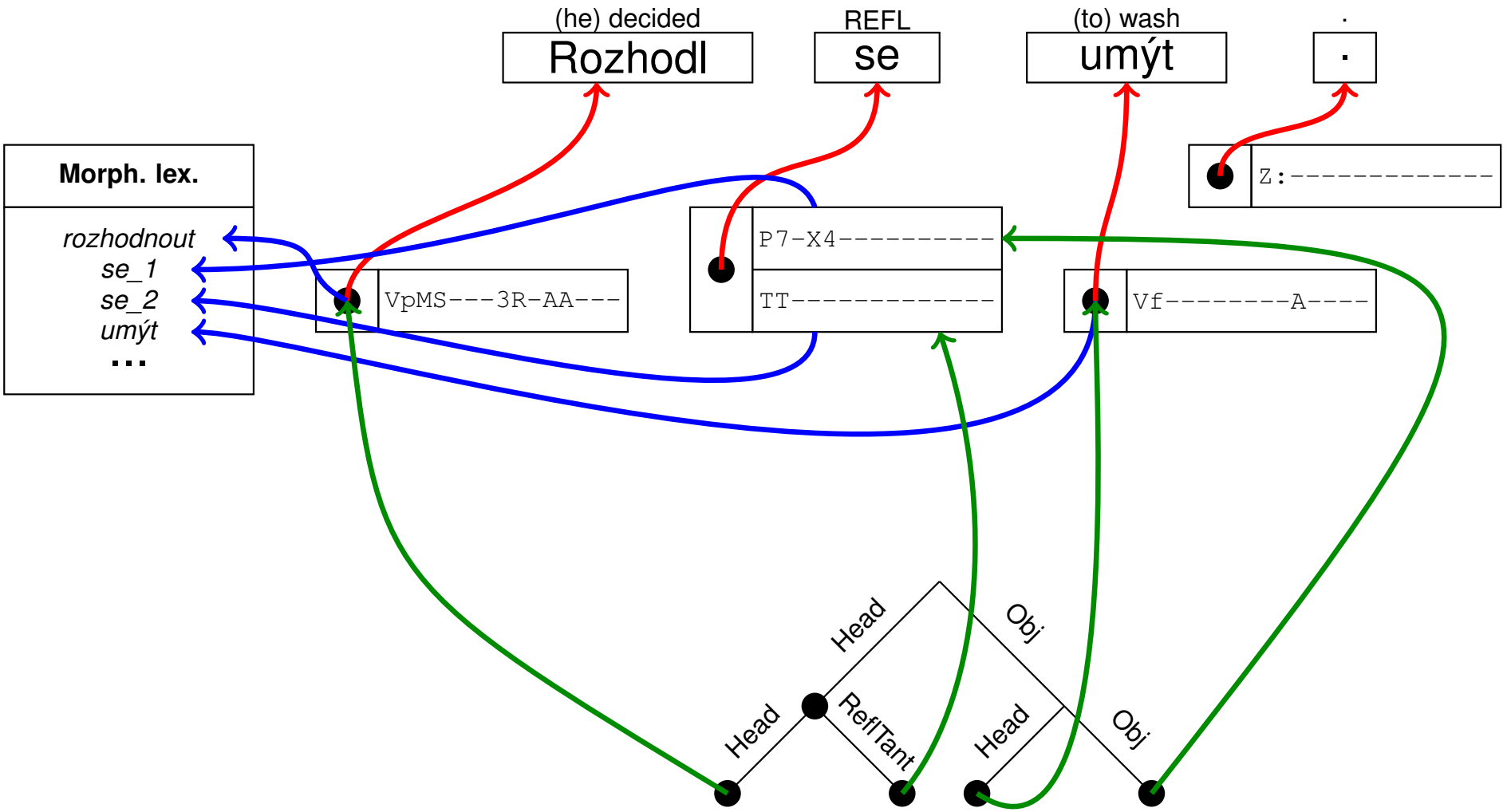

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.

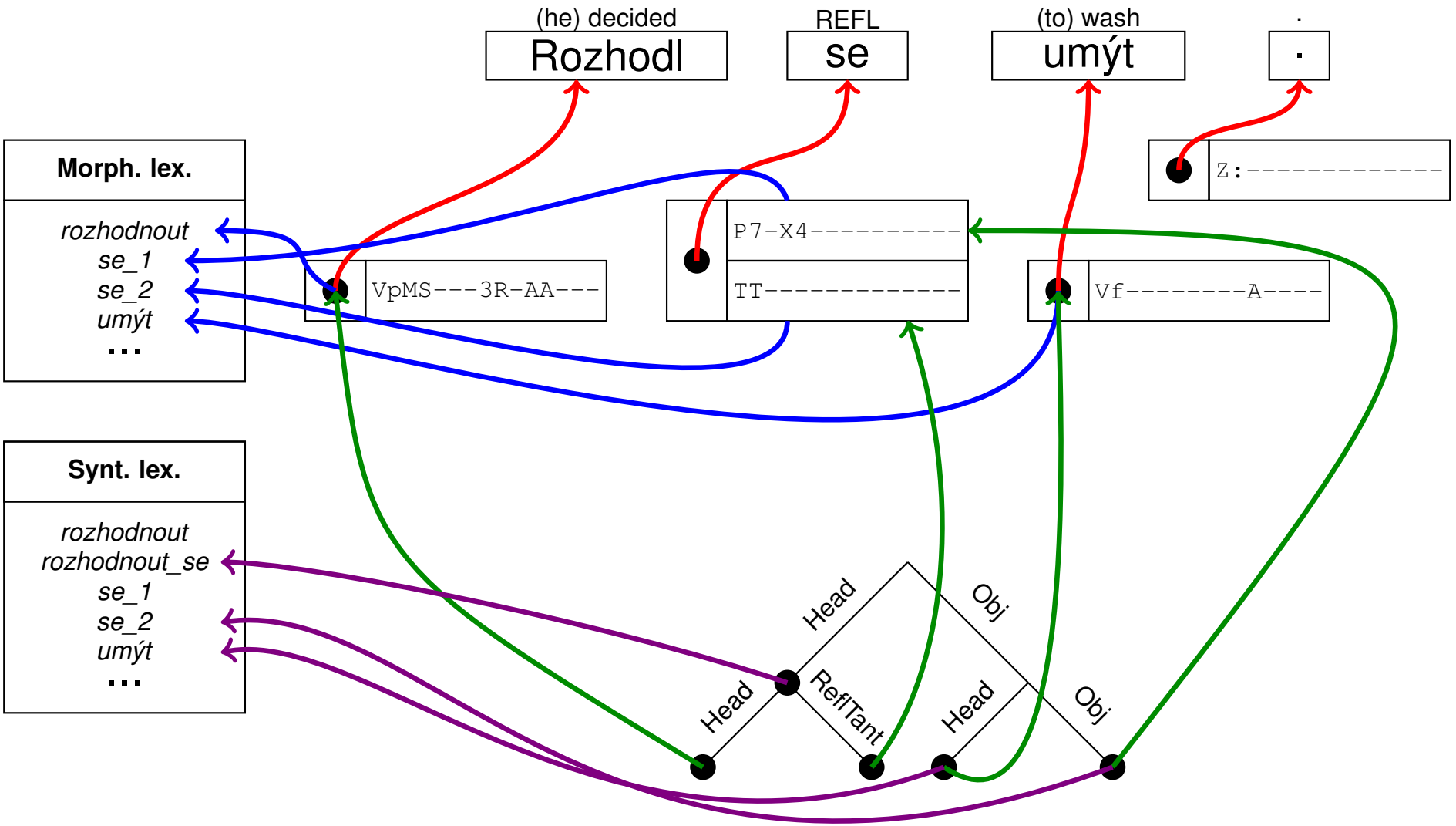

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Z:-------------
P7-X4----------
TT-------------
VpMS---3R-AA---
Vf--------A----

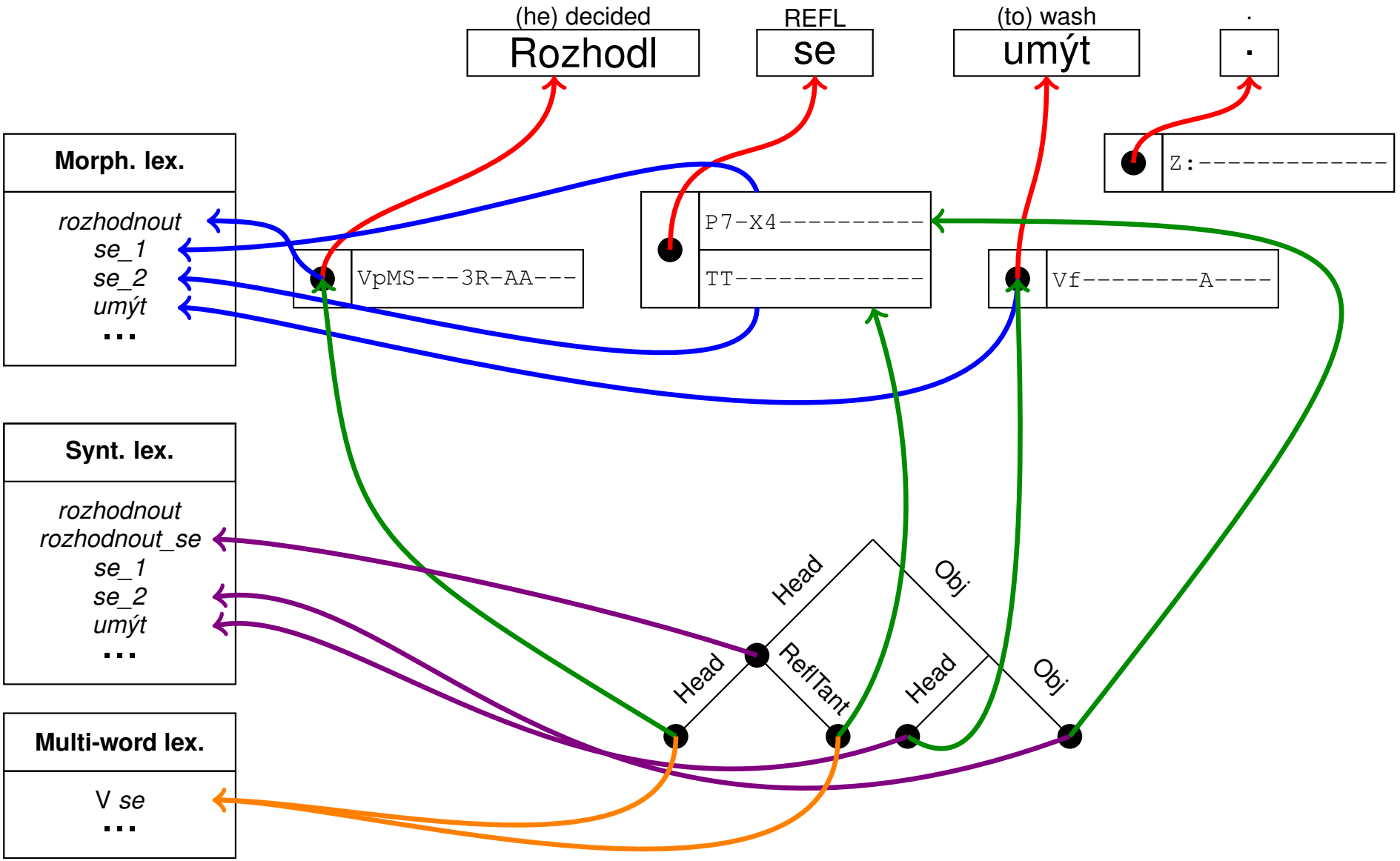

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT----------
Vf--------A----
Z:----------

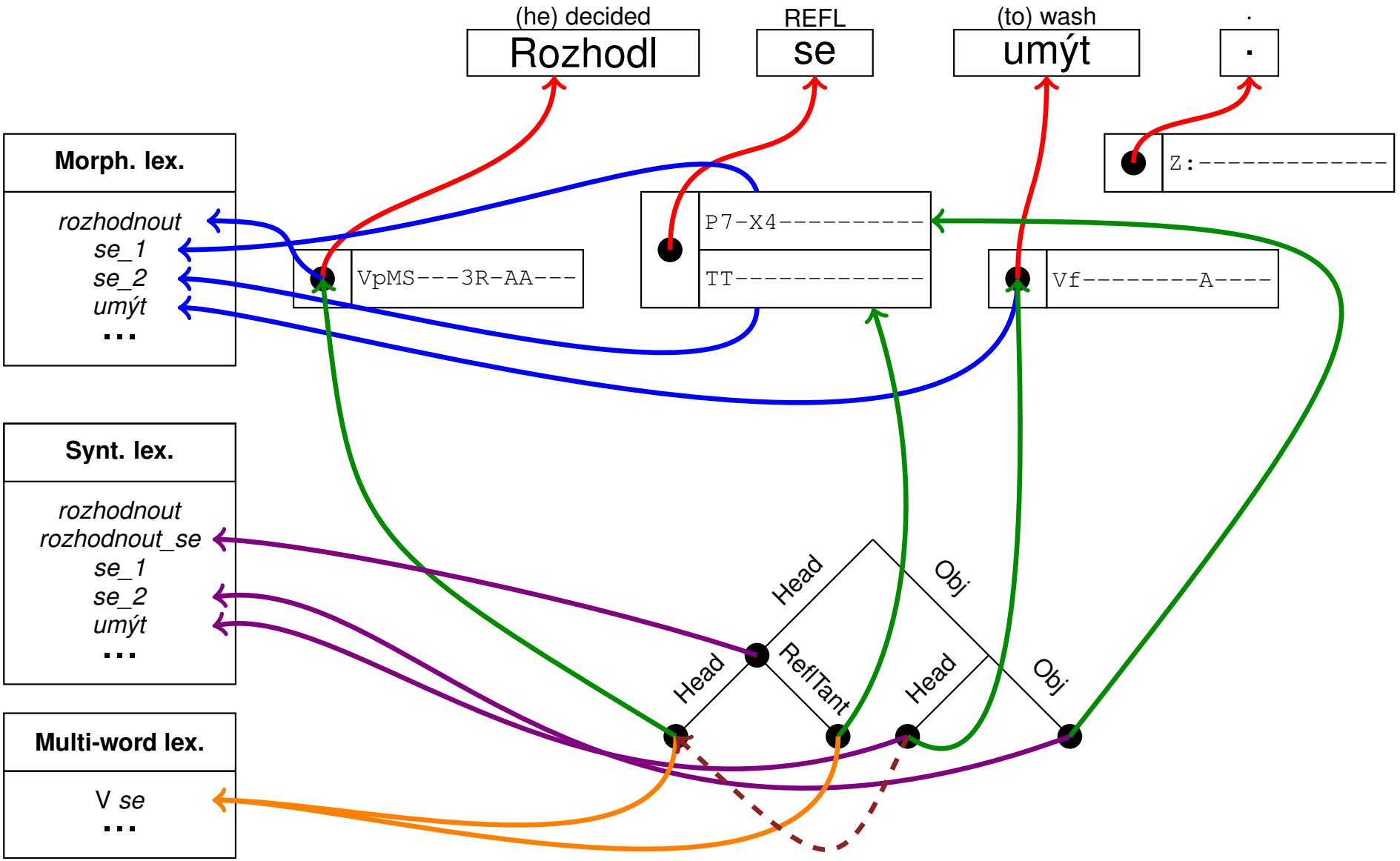

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT----------
Vf--------A----
Z:----------


Head
Obj
Head
ReflTant
Head
Obj


(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT------------
Vf--------A----
Z:------------
Head
Head
Refl Tant
Head
Obj
Obj


(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT------------
Z:------------
Vf--------A----
Synt. lex.
rozhodnout
rozhodnout_se
se_1
se_2
umýt
...
Head
Obj
Head
ReflTant
Head
Obj


(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT------------
Z:------------
Vf--------A----
Synt. lex.
rozhodnout
rozhodnout_se
se_1
se_2
umýt
...
Multi-word lex.
V se
...
Head
Obj
Head
ReflTant
Head
Obj


(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT------------
Z:------------
Vf--------A----
Synt. lex.
rozhodnout
rozhodnout_se
se_1
se_2
umýt
...
Head
Obj
Head
ReflTant
Head
Obj
Multi-word lex.
V se
...

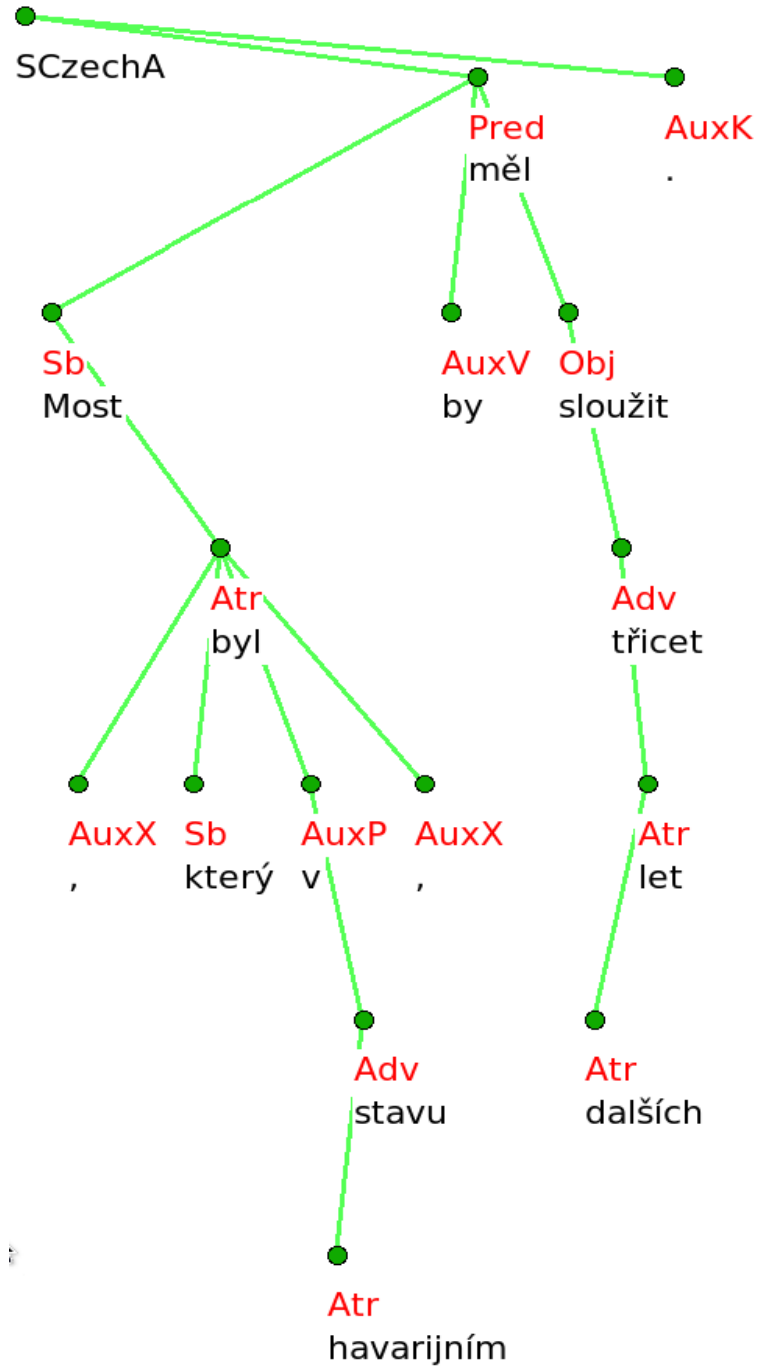
Outline of the talk

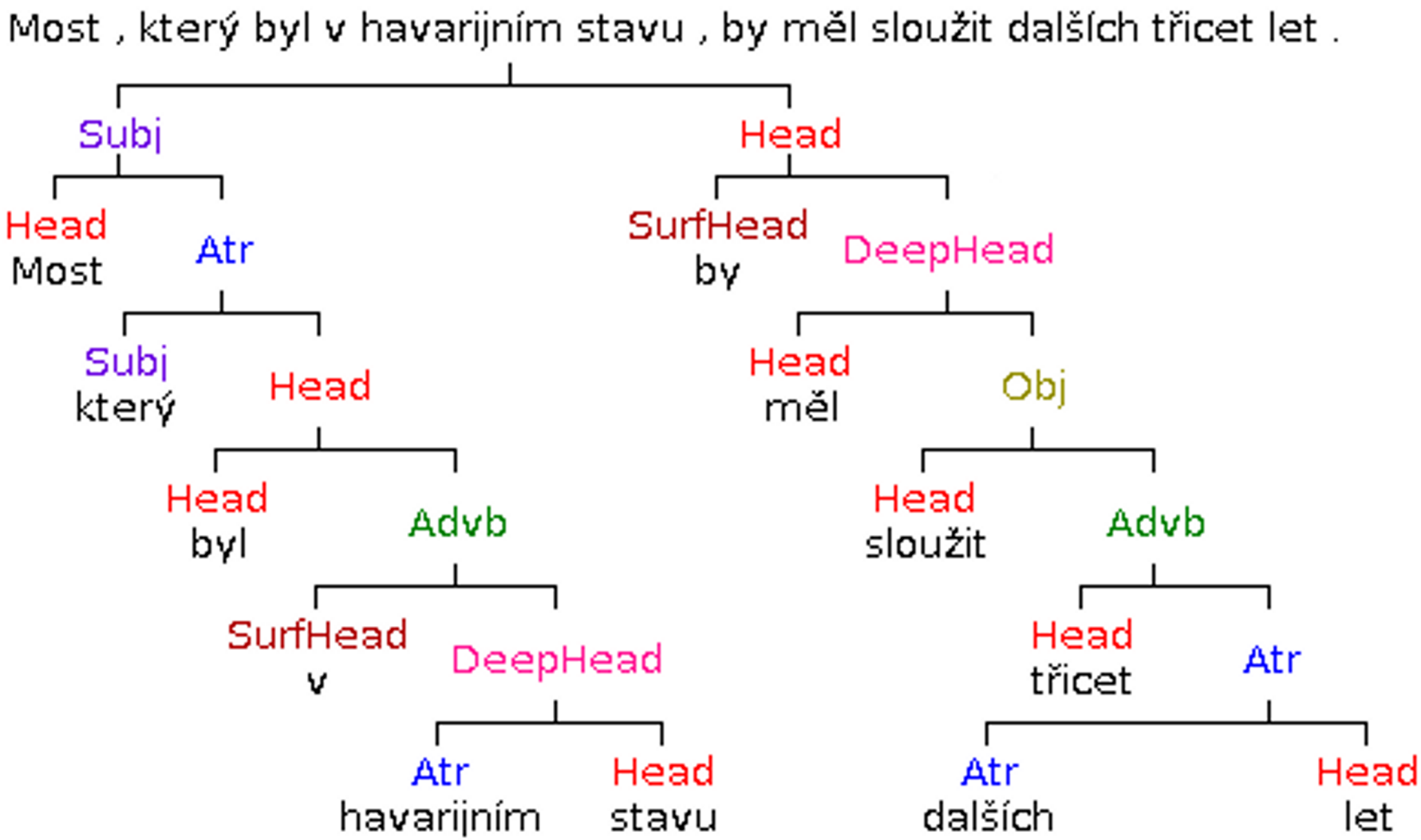
Processing of the input text:

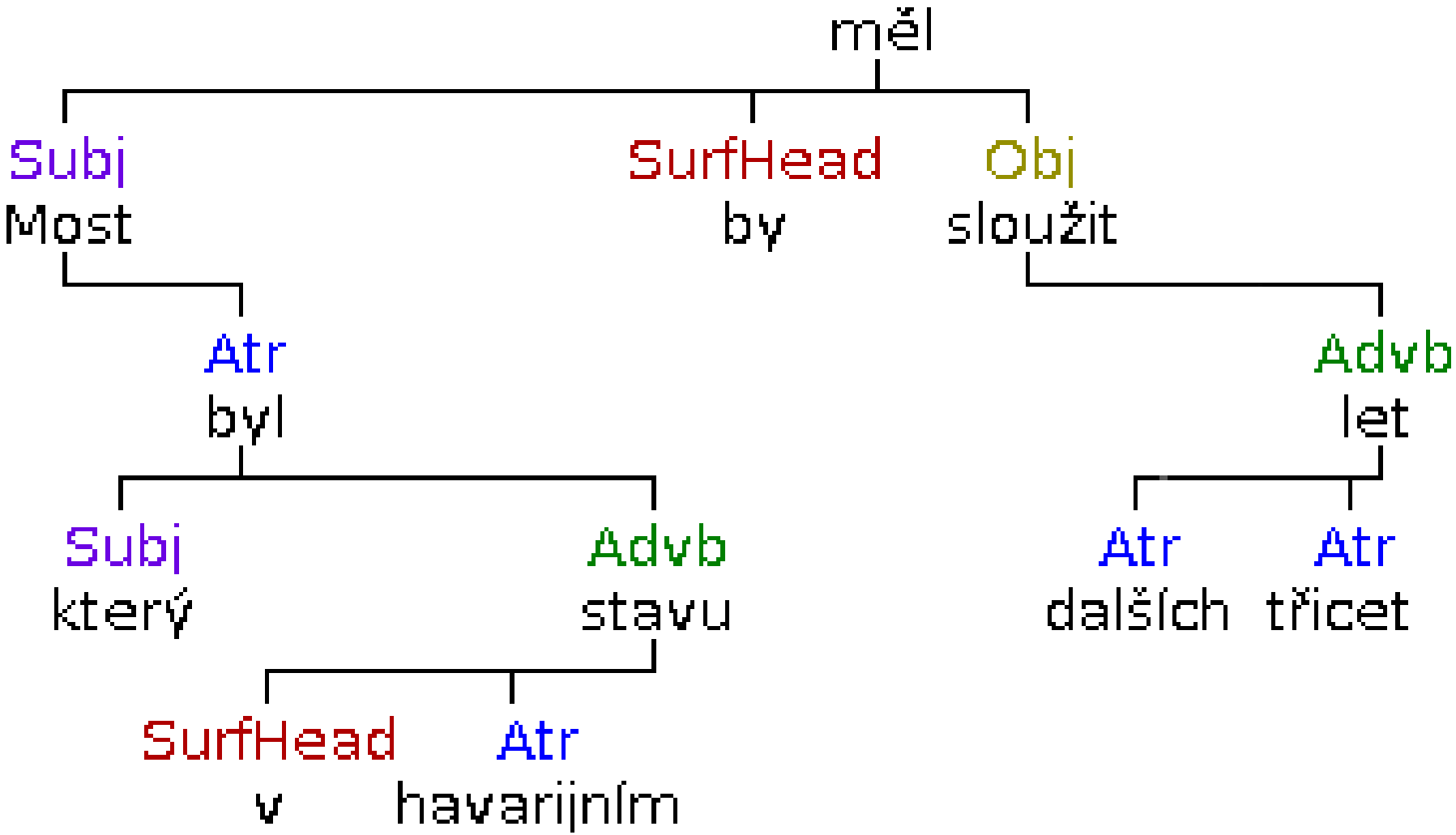
- Word and sentence segmentation → sentence on the **graphemic level**
- Morphological analysis and disambiguation (rule-based disambiguation, stochastic tagger, collocation module) → sentence on the **morphological level**
- Stochastic parser applied to the disambiguated sentence
- Automatic correction of the parse
- Conversion of the corrected parse + modifications:
  - phenomena that require arbitrary decisions in a dependency tree: constructions with function words, coordinated constructions, lists
  - **disjunction** accounting for structural ambiguities expressed by "combined functions" **AttrAdv**, **ObjAdv**

Syntactic tree in the PDT and the new format

(7) Most, který byl v havarijním stavu, by měl sloužit
Bridge which was in emergency state should have$_{modal}$ serve
dalších třicet let.
next thirty years.
'The bridge, which was ramshackle, should serve for another thirty years.'


SCzechA
Pred
měl
AuxK
.
Sb
Most
AuxV
by
Obj
sloužit
Atr
byl
Adv
třicet
AuxX
,
Sb
který
AuxP
v
AuxX
,
Atr
let
Adv
stavu
Atr
dalších
Atr
havarijním


Most , který byl v havarijním stavu , by měl sloužit dalších třicet let .
Subj
Head
Most
Atr
Subj
který
Head
Head
byl
Advb
SurfHead
v
DeepHead
Atr
havarijním
Head
stavu
Head
SurfHead
by
DeepHead
Head
měl
Obj
Head
sloužit
Advb
Head
třicet
Atr
Atr
dalších
Head
let


měl
Subj
Most
SurfHead
by
Obj
sloužit
Atr
byl
Advb
let
Subj
který
Advb
stavu
Atr
dalších
Atr
třicet
SurfHead
v
Atr
havarijním

Correction module

- 30 correction rules so far
- for more frequent errors which can be reliably corrected
- such as noun in accusative as subject

Success rate of the correction modules

	rules	dependency	label	total
Clauses	6	1688	774	1744
NP	8	819	2066	2625
PP	9	834	7160	7722
Other	5	412	1390	1802
Total (ppm)		3753	11390	13893
Total (%)		0.38%	1.14%	1.39%

Outline of the talk

Conclusions and plans

Results

- 200M corpus parsed and corrected
- First version of a viewer with three modes of representation

Further work

- Manual tagging of a 2M training corpus for improvement of the tagger
- Manual parsing of sentences for improvement of the parser
- Detection of more errors made by the parser and their correction
- Creating a corpus with improved tools
- Enabling more modes of viewing the syntactic structure
- Grammar development

Acknowledgment

Work on this project was supported
by the Grant Agency of the Czech Republic
as project No. P406/10/0434.

Thank you for your attention!


Sgall, P., Hajičová, E., & Panevová, J. (1986).
The Meaning of the Sentence in its Semantic and Pragmatic Aspects.
Reidel and Academia, Dordrecht and Praha.
Editor: Jacob Mey.