Treebanking à la carte[1]

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

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

Dubrovnik, 12th – 14th September 2011

[1]This paper was financially supported by the Grant Agency of the Czech Republic No. P406/10/0434

Outline of the talk

Outline of the talk

Different views of syntactic structure

- Syntax is a discipline of many theories
- Syntactically annotated corpus runs the risk of a theoretical bias
- Theory-specific representations have different appearances but share a large part of content
- Treebank offering different views of a single syntactic annotation is a realistic and appropriate goal

From scratch? A bad idea!

- Enormous manual efforts went into building treebanks already [Hajič(2006), Hajič et al.(1998), Skut et al.(1997), i.a.]
- Scaling-up possible by automatic tools.

Improvements over existing annotation schemes

- Potentially underspecified morphological and syntactic core
- Multiple interaction shells, customisable in shape and detail according to the preferences of humans or computer applications
- Accessible to lay users and satisfying experts at the same time

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 heads as special functions)
- 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 *být*, prepositions, subordinate conjunctions, numerals

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)

Syntactic phenomena

Annotation of:

- **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

Outline of the talk

Representation layers:

- **text**
- **morphology**
- **syntactic structure**

Lexicons:

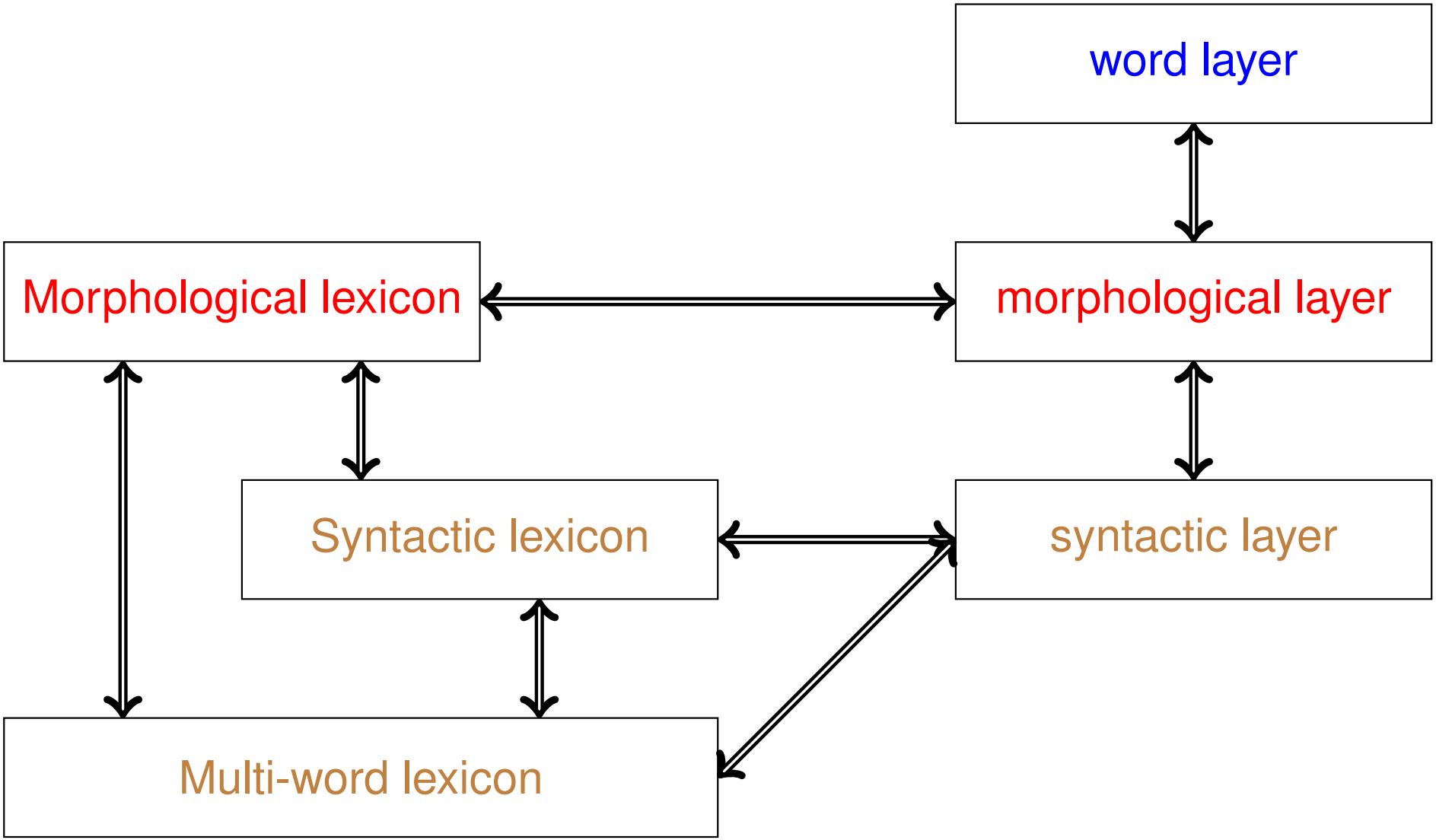
- **morphological**
- **syntactic**
- **multi-word expressions**

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

Syntactic structure is represented by a constituency-based tree where:

- each nonterminal node is assigned a type & a syntactic function
- each terminal node is assigned a syntactic function

Hierarchy of types

- **TypeHeaded**
- **TypeUnHeaded**
  - **Coord** – coordination
  - **Adord** – adordination
  - **Unspec** – unspecified (for collocations and other)

Syntactic functions

Special functions for TypeHeaded:

- **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** & **DeepHead**

Syntactic functions – continued

Other functions for TypeHeaded:

- **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...)

Syntactic functions – continued

Special function for TypeUnHeaded structures:

- Memb – member of a TypeUnHeaded structure

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**.

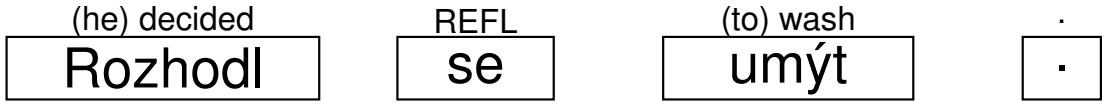
Syntactic lexicon

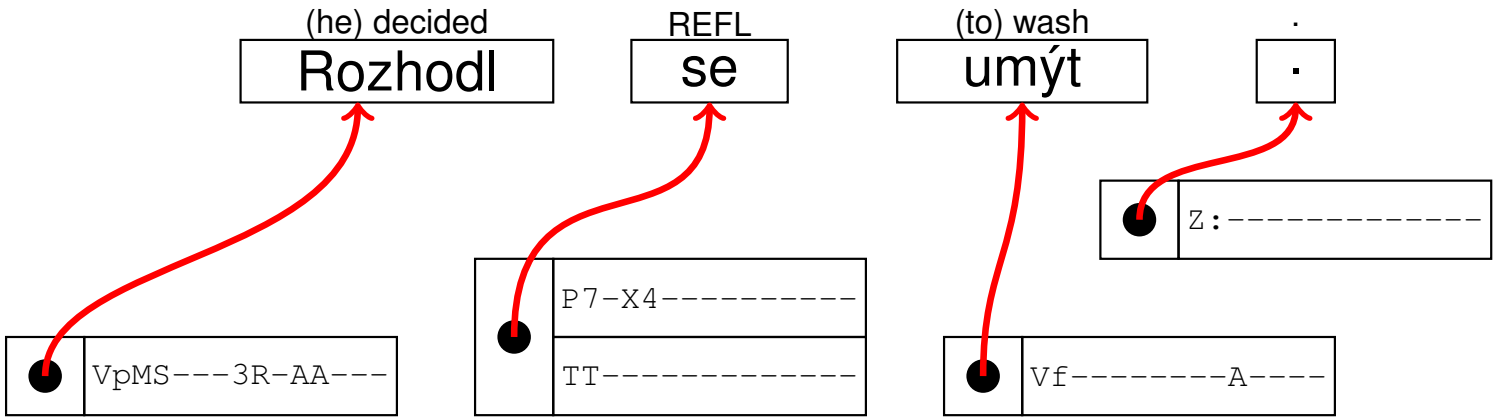
- list of lemmas with their syntactic properties
- inherent reflexives have separate entries
- *rozhodnout* and *rozhodnout se* ‘decide’ are two separate entries
- *vidět* ‘see’ is one entry
- **valency frame** entries
- different valency frames listed under one lemma
- ...

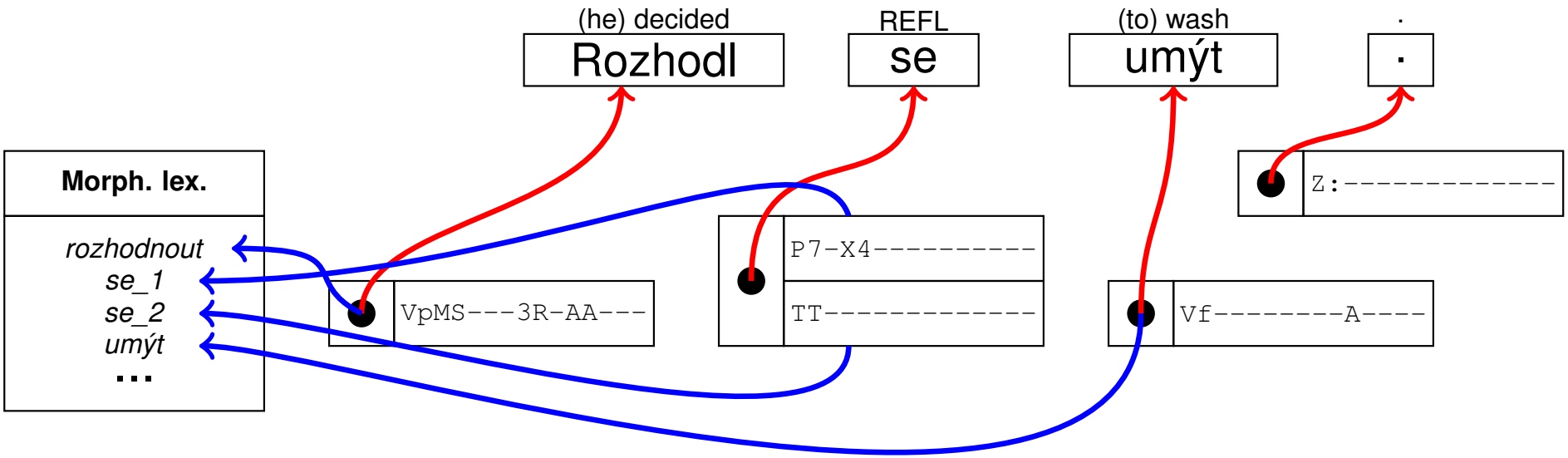
Multi-word lexicon

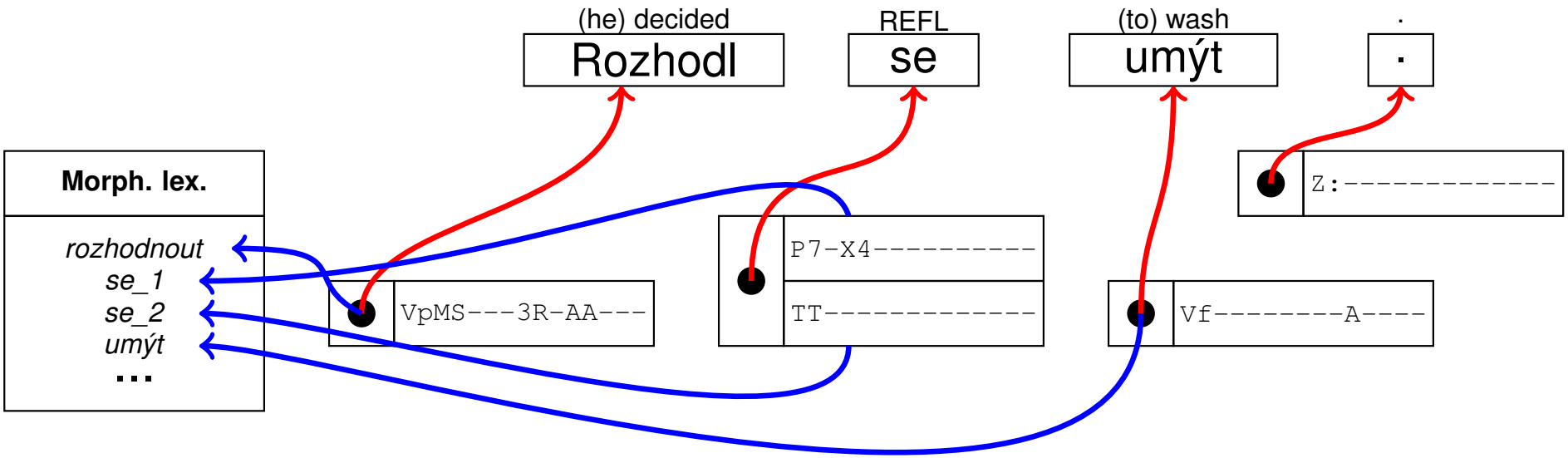
- collocations: *křížem krážem* “in all directions”, *nechat na holičkách* “leave in the lurch”
- types of analytical verb forms: *bych byl přišel* “(I) would have come”, *jsi přišel* “(you) have come”
- inherent reflexives: *usmíváš se* “(you) smile”; *rozhodne se* “(he) decides”
- (reflexive) deagentive 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”

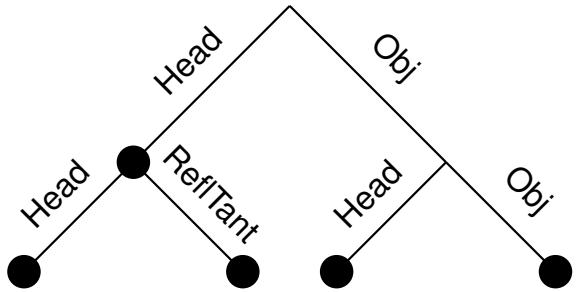
Outline of the talk

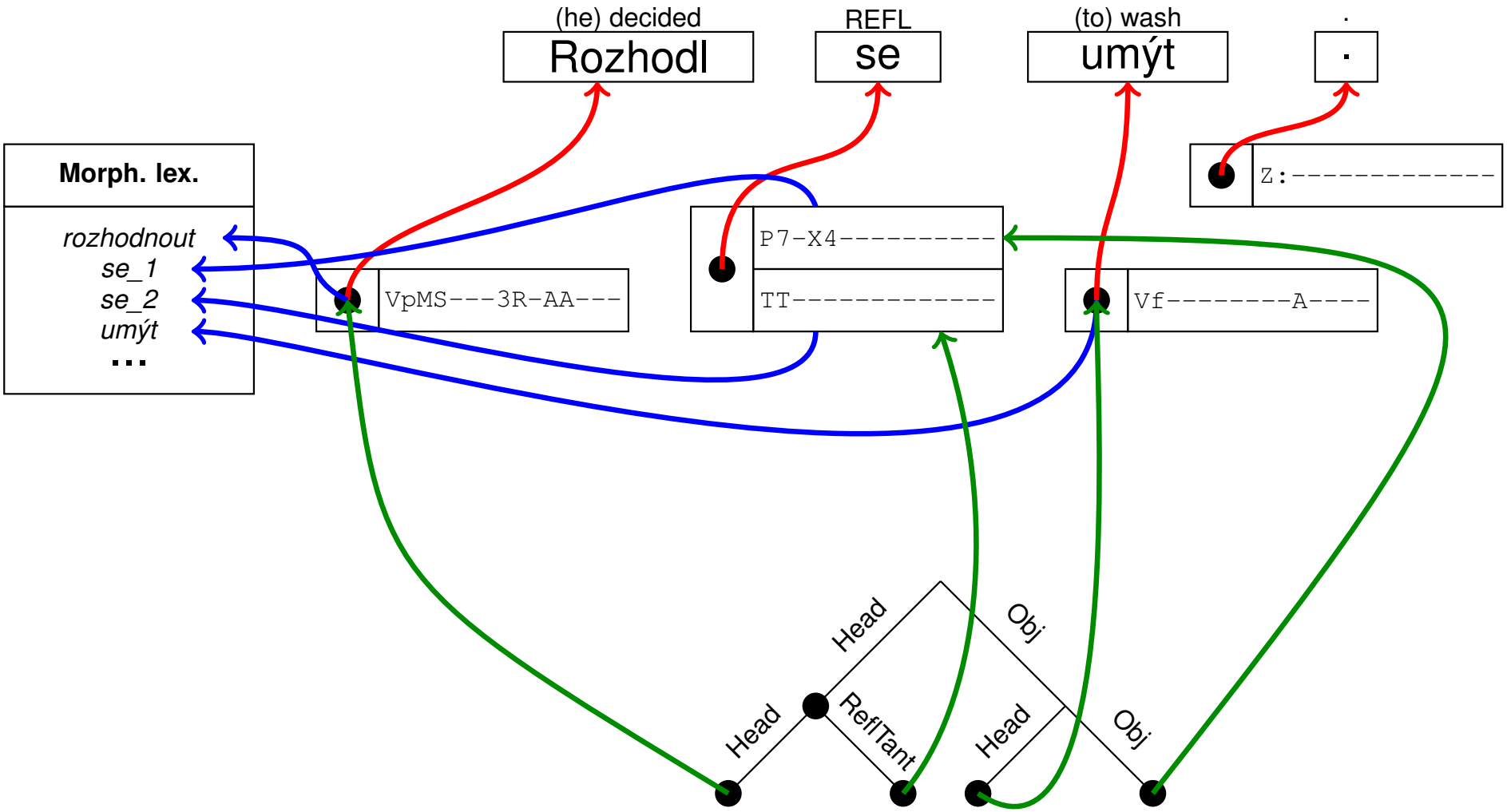

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.

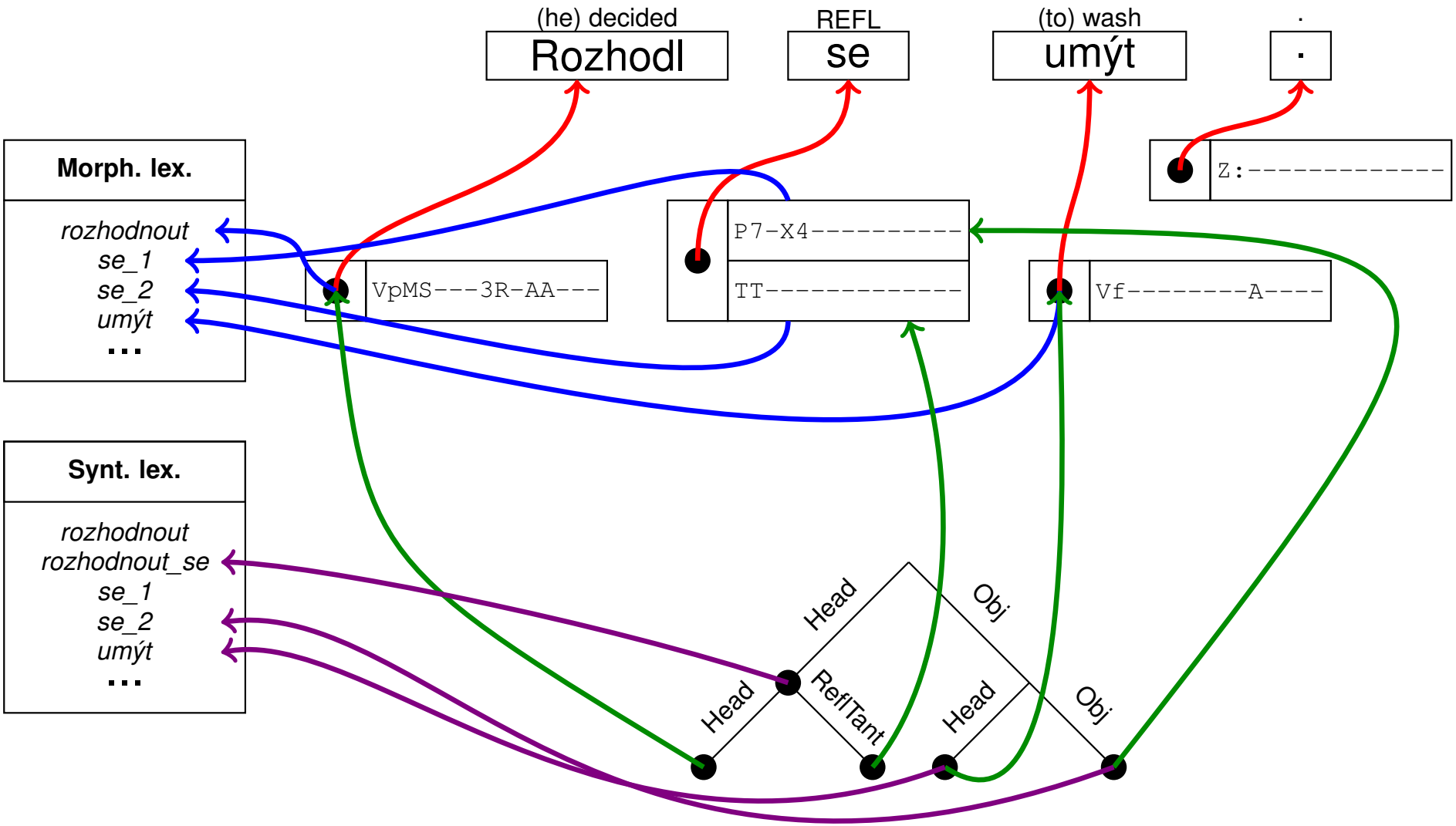

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Z:----------
P7-X4----------
TT----------
VpMS---3R-AA---
Vf--------A----

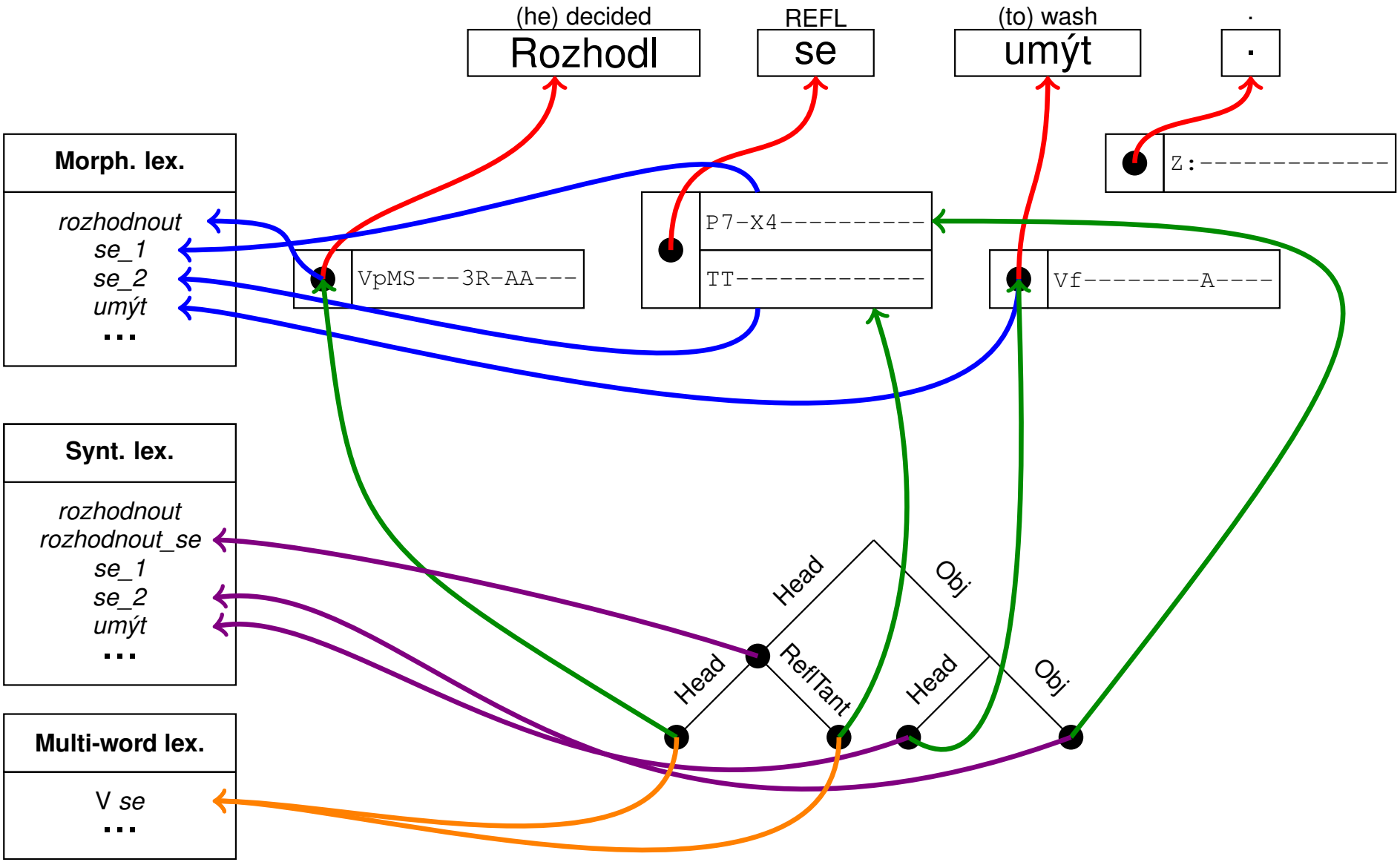

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT-------------
Vf--------A----
Z:-------------

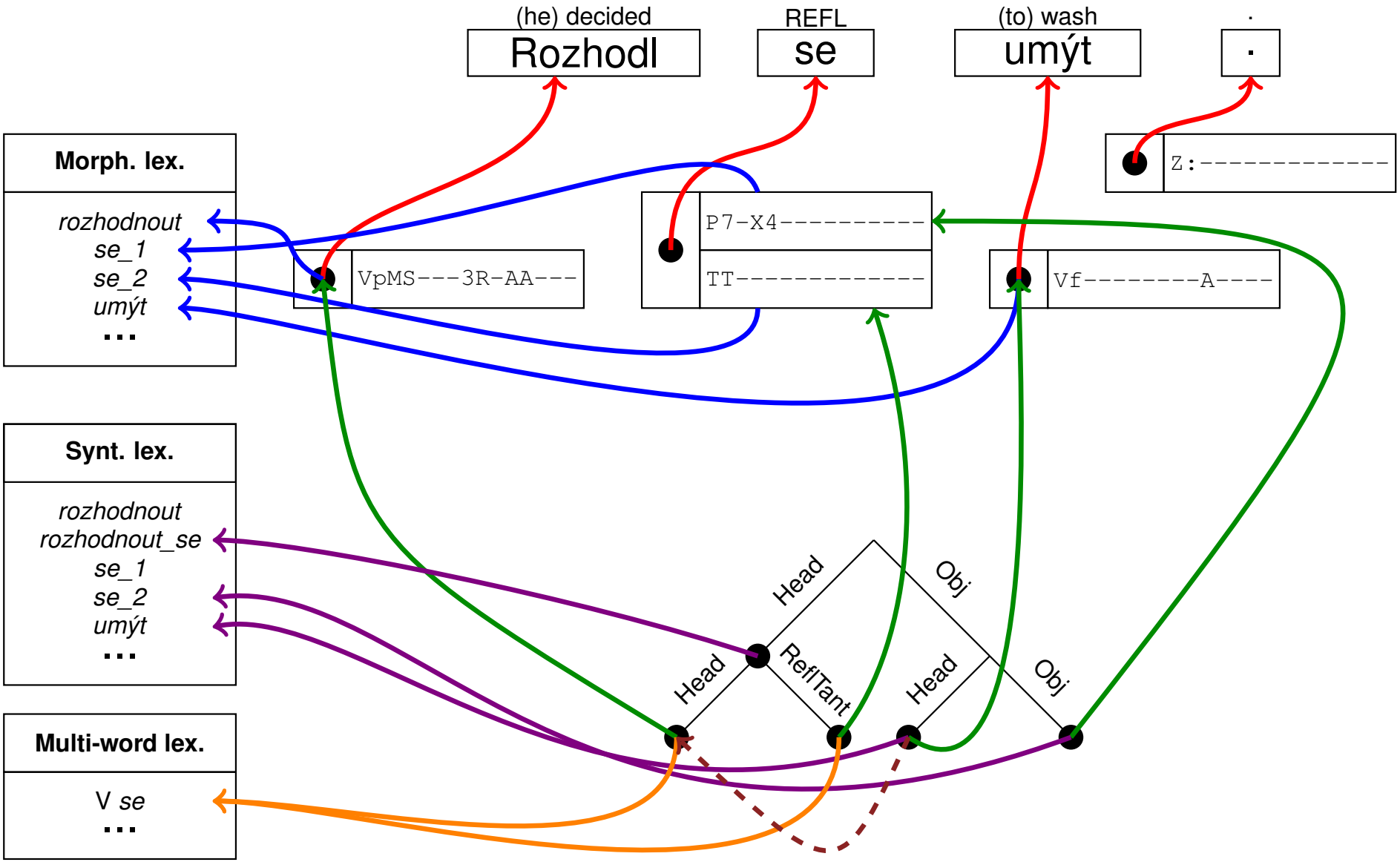

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT-------------
Vf--------A----
Z:-------------


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-------------
Z:-------------
Vf--------A----
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:-------------
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----------
Vf--------A----
Z:----------
Synt. lex.
rozhodnout
rozhodnout_se
se_1
se_2
umýt
...
Head
Obj
Head
ReflTant
Head
Obj
Multi-word lex.
V se
...

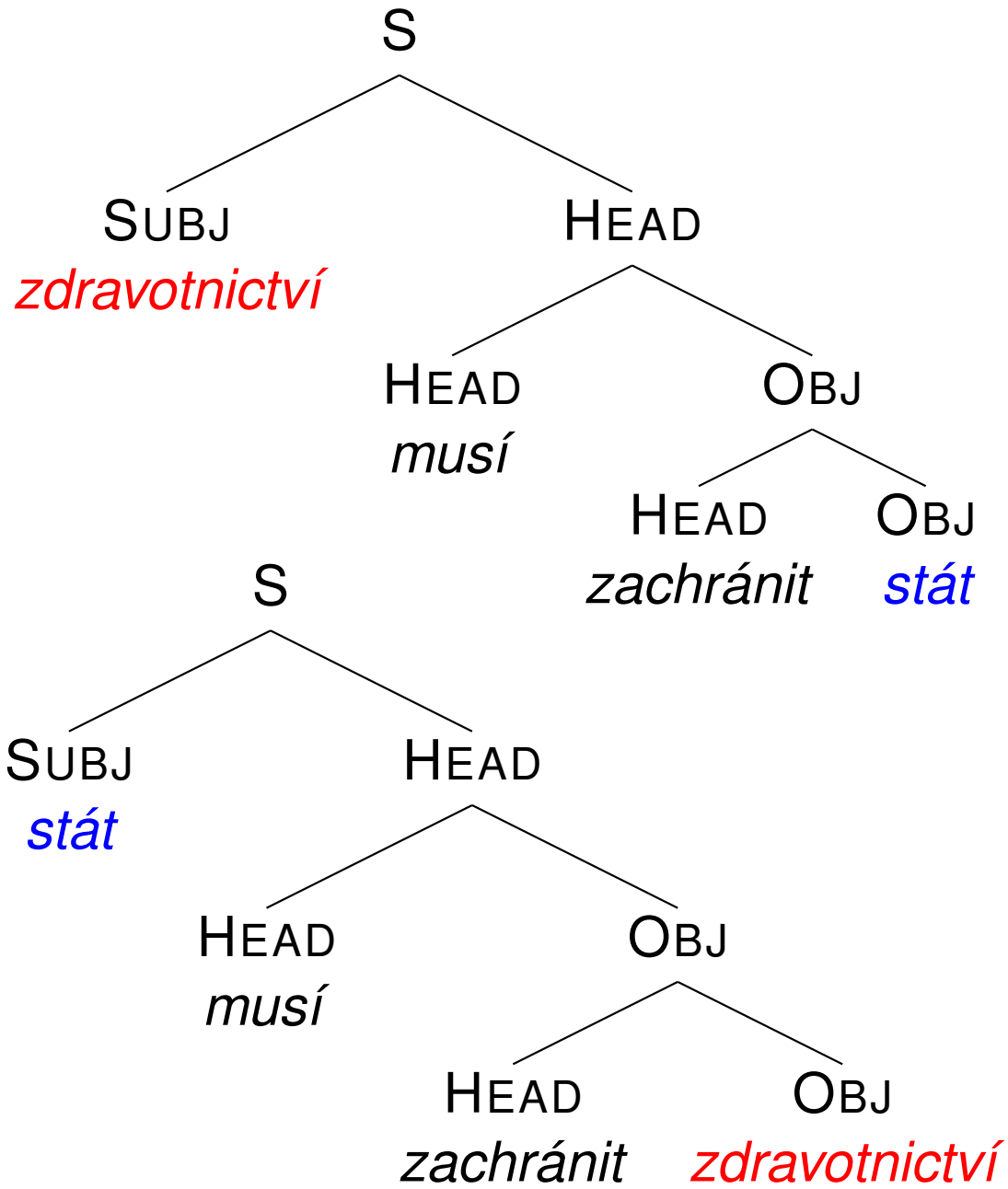

(he) decided
Rozhodl
REFL
se
(to) wash
umýt
.
.
Morph. lex.
rozhodnout
se_1
se_2
umýt
...
VpMS---3R-AA---
P7-X4----------
TT-------------
Vf--------A----
Z:-------------
Synt. lex.
rozhodnout
rozhodnout_se
se_1
se_2
umýt
...
Head
Obj
Head
ReflTant
Head
Obj
Multi-word lex.
V se
...

Subject/object ambiguity

(1) 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 (1) 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 (1), some boxed numbers refer to the forms above:

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

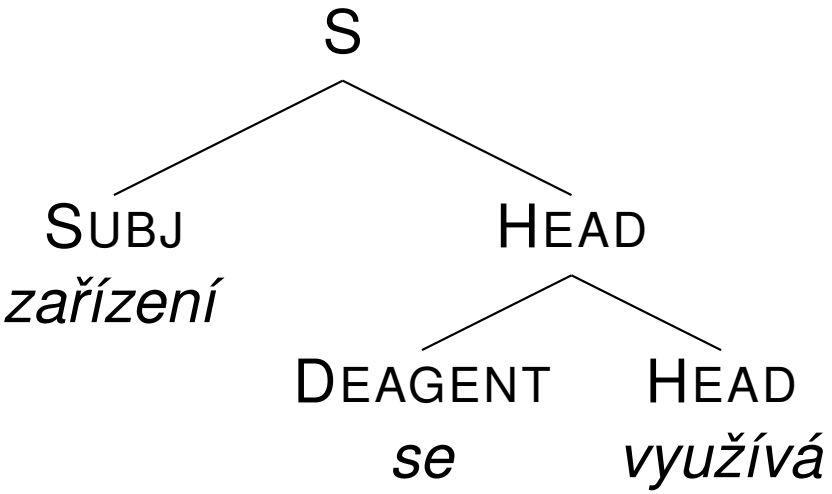
Two possible structures with constraints on category values and overriding clauses:

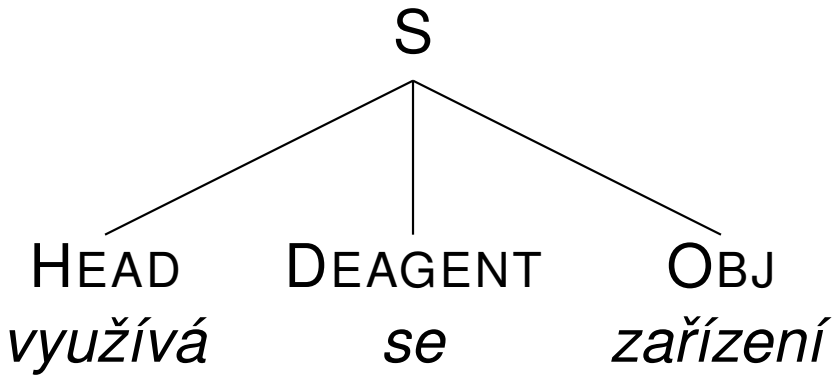
#1 = 7, *X*=*nom*, *Y*=*acc*
#2 = 7, *X*=*acc*, *Y*=*nom*, 1 → 4, 4 → 1

Another type of subject/object ambiguity

Reflexive passive:

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



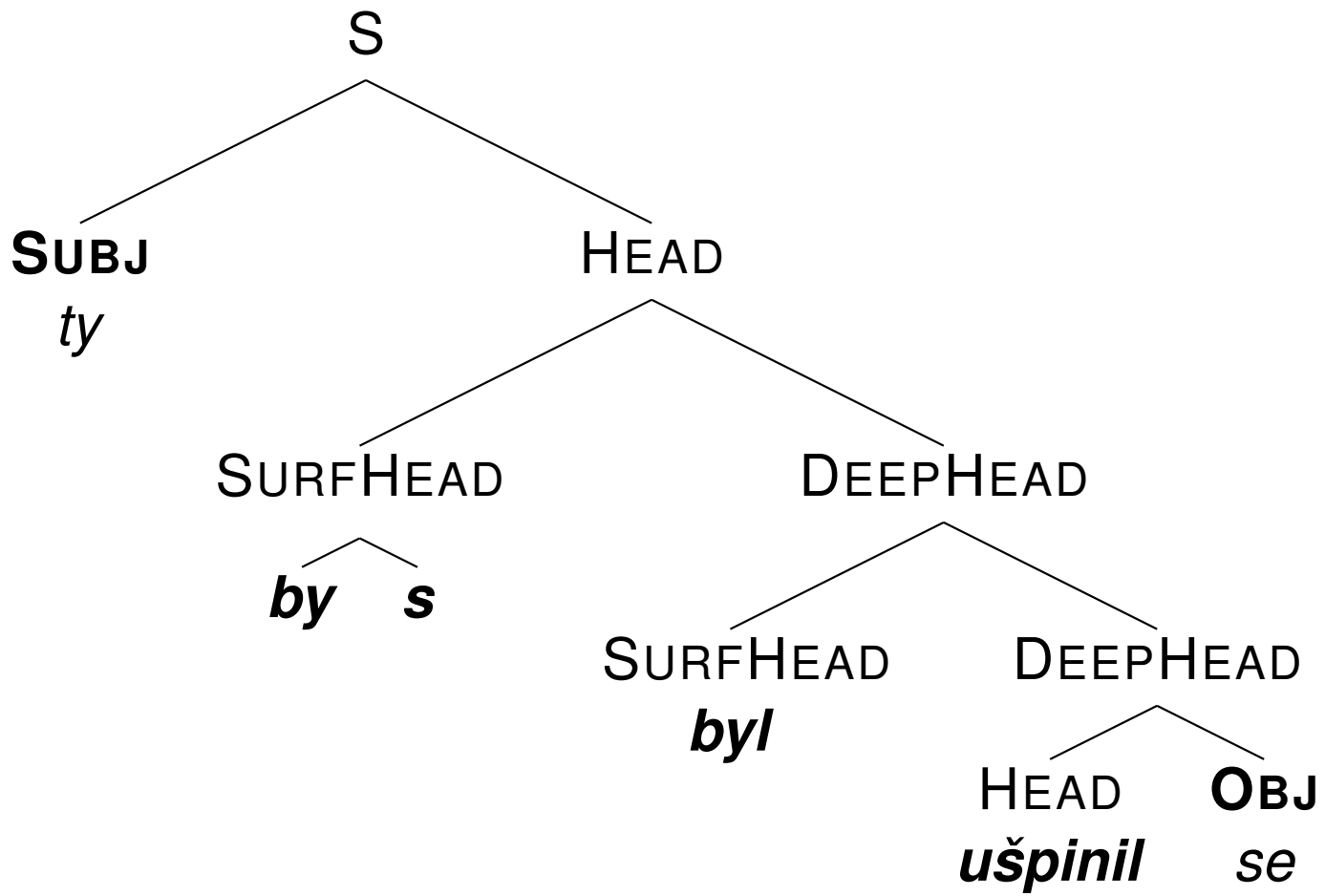


Treating contractions

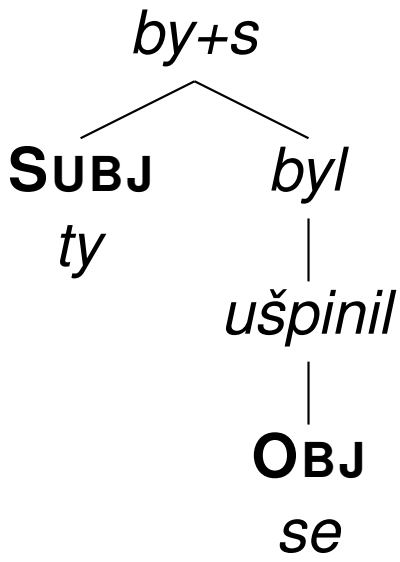
(3) **Ty** by *se*s byl ušpinil.
you would REFL+AUX$_{2nd,sg}$ be$_{pple}$ get dirty$_{pple}$
'You would have got dirty.'

Ty by *se*s byl ušpinil.

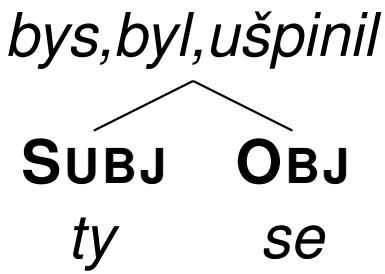
(4)



(5) **Surface dependency structure** derived from (4)



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



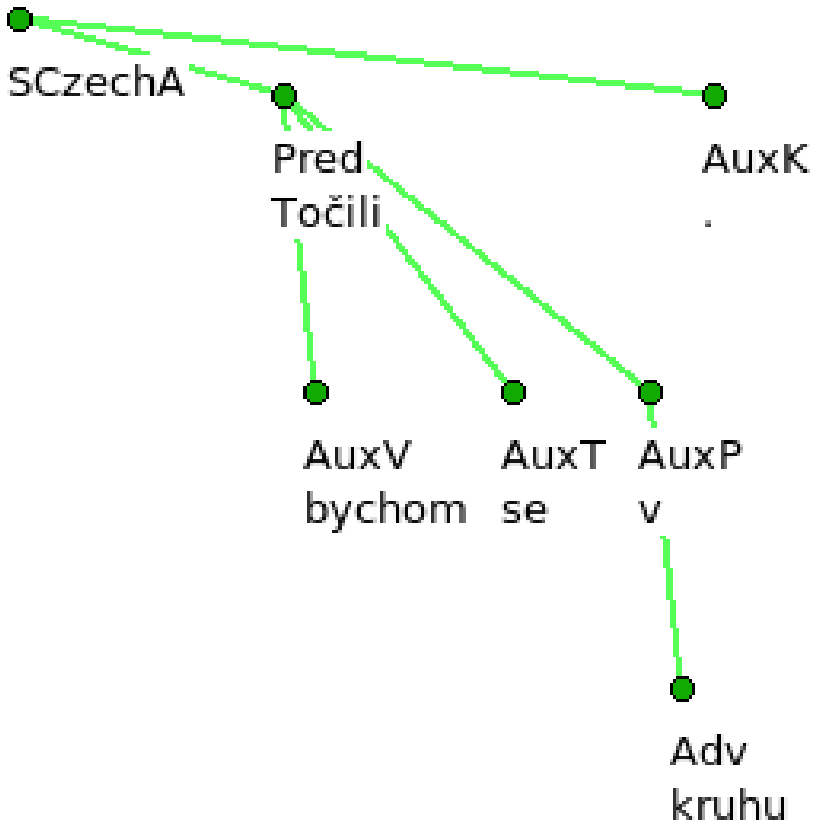
Outline of the talk

Input text processing:

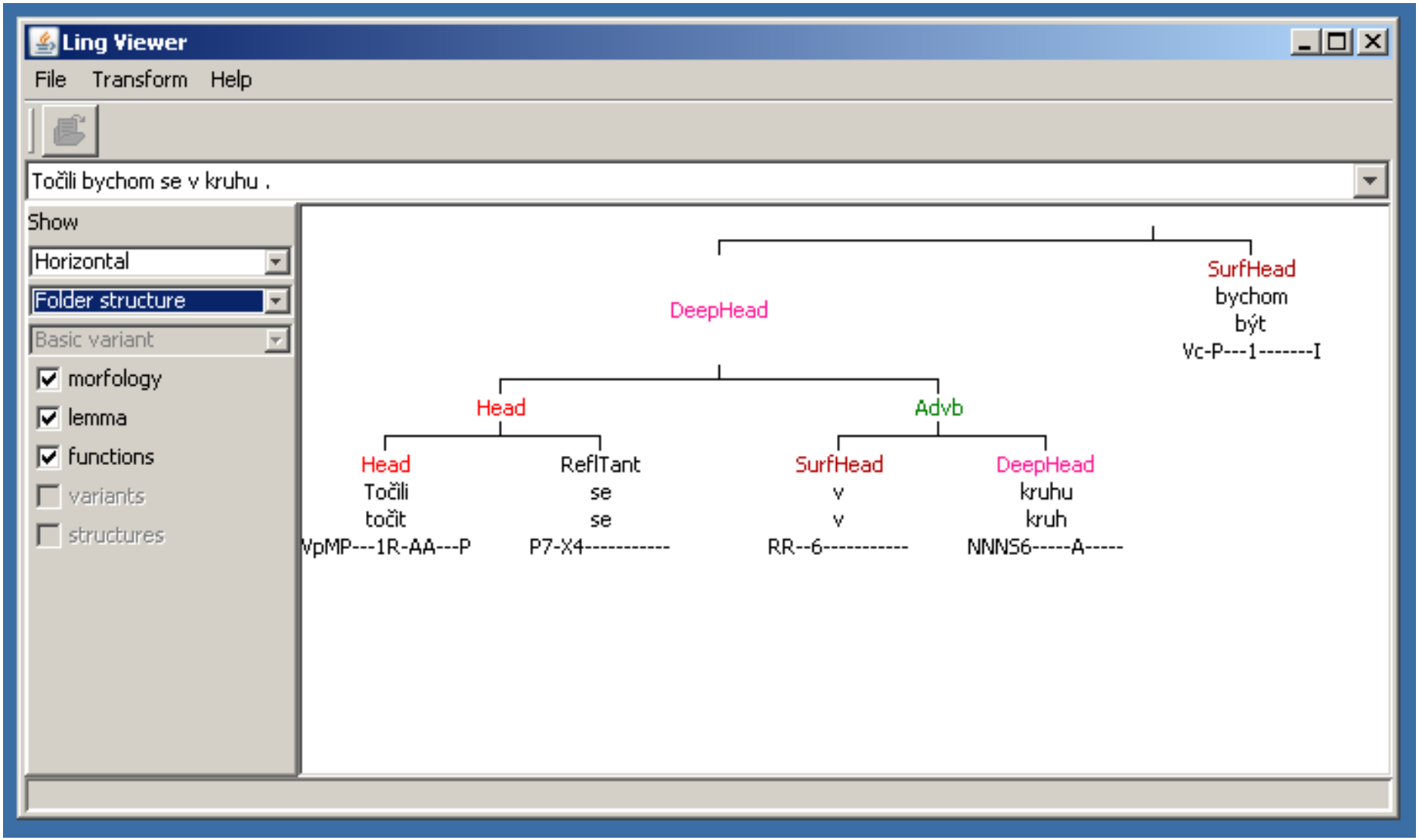
- tokenization and sentence segmentation → sentence on the **graphemic/word level**
- morphological analysis and disambiguation (rule-based disambiguation, MorČe, collocation module) → sentence on the **morphological level**
- McDonald et al.'s parser (or more parsers with a voting scenario) applied to the disambiguated sentence; McDonald et al.'s parser can be parameterized
- automatic correction of the parse still in `tectomt` format
- conversion of the corrected parse from the `tectomt` format to ours + modifications:
  - phenomena that in a dependency tree require arbitrary decisions: constructions with auxiliary verbs, coordinated constructions, lists
  - **disjunction** accounting for structural ambiguities expressed by combined functions **AttrAdv**, **ObjAdv**

PDT representation

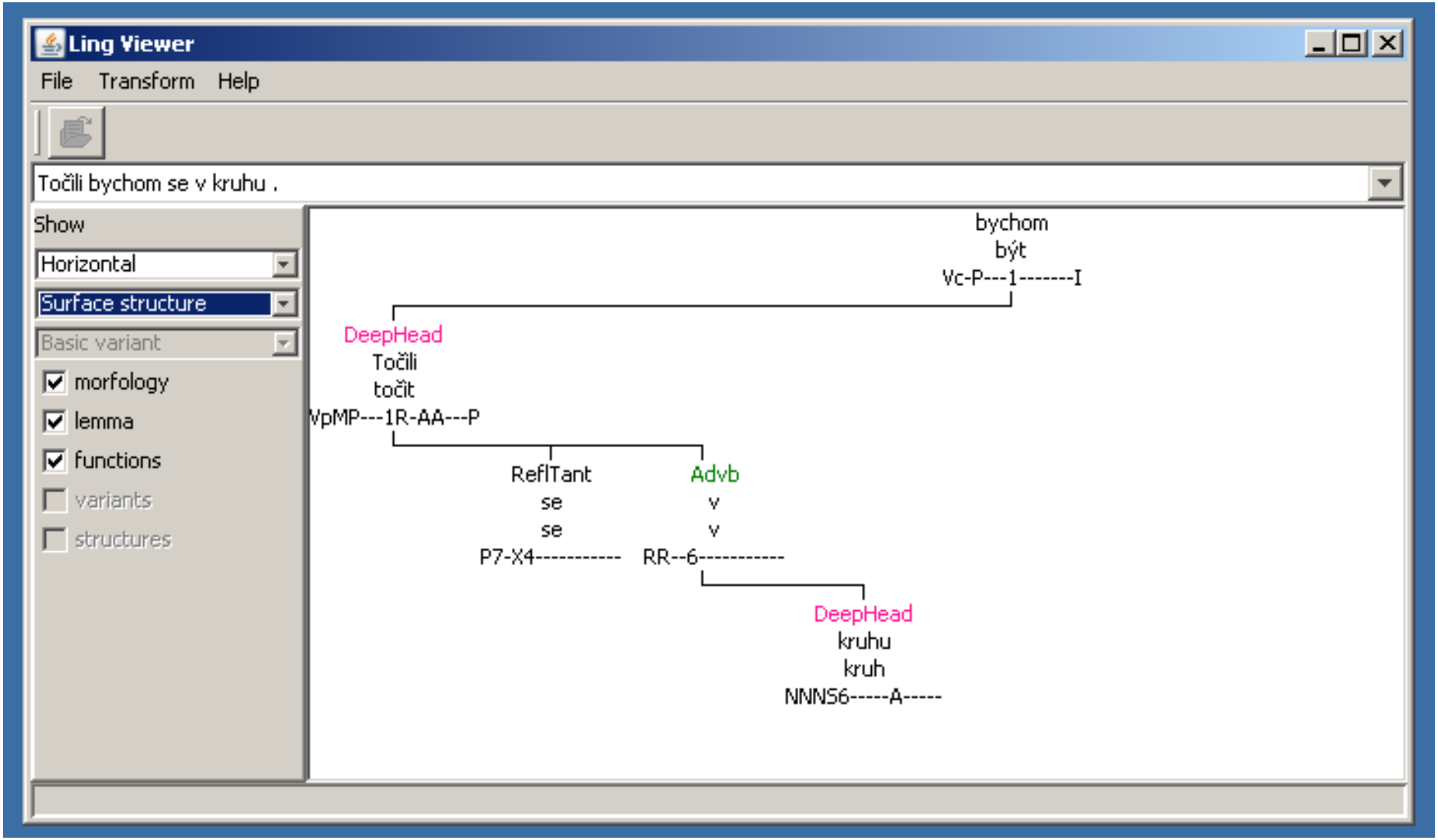
(7) Točili bychom se v kruhu.
Turn would$_{1st,pl}$ REFL in circle.



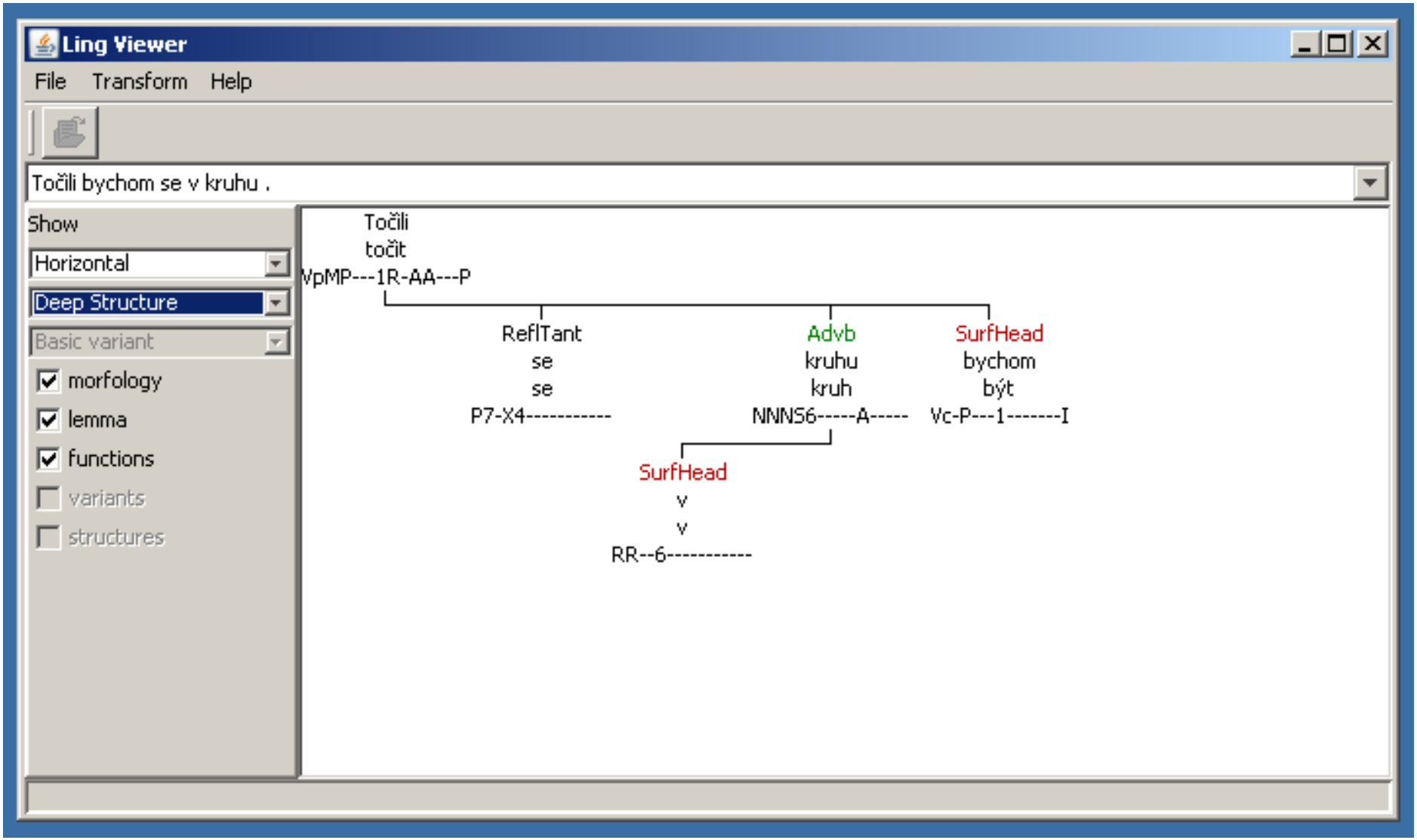
Internal structure



Surface structure



Deep structure



Outline of the talk

Main assets

- Exploitation of existing approaches and tools, we cannot develop the treebank from scratch
- multilayer stand-off annotation
- linguistically motivated corrections of the results of the tools used
- entirely automatic processing

Comparison with analytical layer of PDT

- Single annotation capturing syntactic core, as little theoretical bias as possible but:
  - various interpretations using underspecification, disjunction
  - export options into customizable formats
  - various visualizations of the data

Hajič, J. (2006).
Complex Corpus Annotation: The Prague Dependency Treebank.

In M. Šimková, editor, Insight into the Slovak and Czech Corpus Linguistics, pages 54–73, Bratislava, Slovakia. Veda.

Hajič, J., Hajičová, E., Panevová, J., & Sgall, P. (1998).
Syntax v Českém národním korpusu.
Slovo a slovesnost, **59**(3), 168–177.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997).
An annotation scheme for free word order languages.
In Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97, Washington, DC.