

STYX

**Pražský závislostní korpus
jako cvičebnice jazyka českého**

Ondřej Kučera

Obsah

- 1. Úvod** (motivace, PDT, PML, implementace)
- 2. Filtrování vět**
- 3. Transformace stromů**
- 4. STYX:** FilterSentences, Charon, Styx
- 5. Závěr, budoucnost**

1. Úvod

Motivace

- děti dnes počítače běžně používají
 - hry, surfování, chat, psaní, kreslení
- proč by nemohly na počítači určovat slovní druhy nebo větné členy?

Budování cvičebnice

Ručně

- nesmírně pracné
 - vybrat (vymyslet) věty
 - anotovat je
- značně omezené množství vět
- často příliš jednoduché věty neodpovídající reálnému používání jazyka

Budování cvičebnice

Automaticky

- máme-li k dispozici anotovaná data
- práce s výběrem vět a anotací je již hotova
- korpusová data reprezentují skutečné používání jazyka
- množství vět odpovídá velikosti korpusu
- PDT

Prague Dependency Treebank

- anotován na čtyřech rovinách (slovní, morfologická, analytická, tektogramatická)
- vnitřní formát: PML (Prague Markup Language) – postaven na XML

Prague Markup Language

- každá ze čtyř rovin anotována v samostatném souboru
- data provázána pomocí identifikátorů slov, vět, odstavců
- závislostní syntax vyjádřena přímo vztahy XML uzlů

PDT vs. školní syntax

- anotačními pravidly PDT lze zpracovat libovolnou větu
⇒ filtrování vět
- analytická rovina PDT se v mnoha ohledech značně liší od školní syntaxe
⇒ transformace analytických stromů

Implementace

Java

- vysokoúrovňový jazyk s řadou pojistek proti omylům programátora
- přenositelnost
- přítomnost knihovny SWT

Implementace

SWT

- Standard Widget Toolkit
- poskytuje nativní vzhled grafických prvků
- rychlost

Implementace

Charon

Soubor Sada

Jsou vám nejasná některá ustanovení daňových zákonů ?

Jsou vám nejasná některá ustanovení daňových zákonů ?

Je tu pro vás připravena rubrika Daňový poradce. Vaše dotazy čekáme na adrese Českomoravský úřad pro daňové záležitosti. Tím pádem máme problém se silniční daní. Za vozidla zaměstnanců užívaná pro pracovní účely stejný názor má i řada našich soukromých podnikatelů. Je naprosto bezbřehý a nevypočitatelný. V tom s vámi nesouhlasím. Mzdy a platy jsou vždy podřízenými hodnotami. Zde vzniká nová dimenze srovnávání: Opět s vámi nesouhlasím. Můžete to vysvětlit na příkladu? Například podnikatel by chtěl dosáhnout u daně z příjmu. Zde se dostáváme k termínu odborná kvalifikace. Jak ho vlastně pozná? Zahrnul jste mne mnoha profesemi. Naše metodika a techniky vyšetřování mohou být vada zadavatelů nebo testů? V celé naší tříleté činnosti si jen čtyři firmy omloují. U posudků v minulosti mohl být sebemenší nedostatek. Lidé prožívají nebývale nervózní dobu. Stále jim nedáváme odpovídající péči. Zejména v Olomouci firma svými výrobky při restaurátorské licenci umožňuje práci v celkové tichosti. Firemní výrobky zdobí galerie a soukromá sídla. Balení másla ve hliníkové fólii zajišťuje jeho ochranu. Samotné měření spotřeby tepla peněženku i v zimě. Platíme hodně hlavně díky nízké technické úrovni. Stavět vlastní výtopnu je prozatím patrně nejvýhodnější. Zásadní pákou je tlak na naši peněženku. Ceny energie rostou nepřetržitě a ve skocích. Pak se dá cena tepla v tryskový let. Ostravský podnik vyrábí teplo v šesti divizích. Jeho rozsah působnosti je úctyhodný: Růst počtu teplem zásobovaných bytů přitom. Počet pracovníků podniku zároveň klesl v průběhu let. Podnik platil doposud za znečišťování ovzduchu. Ročně vyprodukoval 280 - 350 tisíc tun oxidu uhličitého.

```
graph TD; A[Jsou nejasná Přj] --- B[vám Pt]; A --- C[ustanovení Po]; C --- D[některá Pk]; C --- E[zákonů Pk]; E --- F[daňových Pk];
```

Implementace

Charon

Soubor Sada

Jsou vám nejasná některá ustanovení daňových zákonů ?

```
graph TD; A[Jsou nejasná Pt] --- B[ustanovení Po]; B --- C[vám Pt]; B --- D[některá Pk]; B --- E[zákonů Pk]; E --- F[daňových Pk]
```

Jsou vám nejasná některá ustanovení daňových zákonů ?

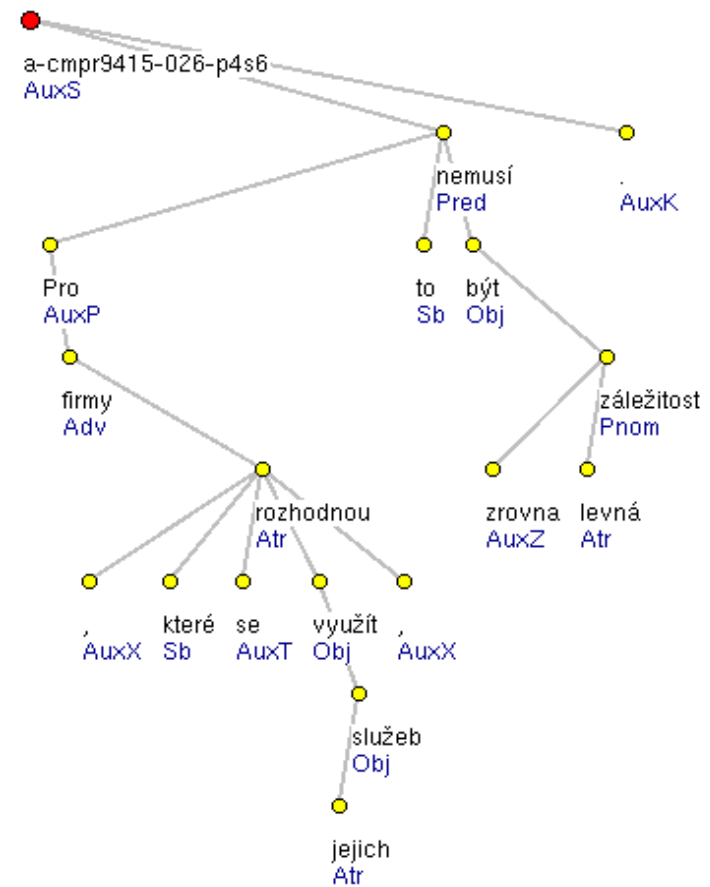
Je tu pro vás připravena rubrika
Vaše dotazy čekáme na adrese
Tím pádem máme problém se s
Za vozidla zaměstnanců užíván
Stejný názor má i řada našich s
Je naprosto bezbřehý a nevypo
V tom s vámi nesouhlasím .
Mzdy a platy jsou vždy podřízen
Zde vzniká nová dimenze srovn
Opět s vámi nesouhlasím .
Můžete to vysvětlit na příkladu ?
Například podnikatel by chtěl do
Zde se dostáváme k termínu od
Jak ho vlastně pozná ?
Zahrnul jste mne mnoha profes
Naše metodika a techniky vyše

2. Filtrování vět

Filtrování vět

Souvětí (souřadná i podřadná)

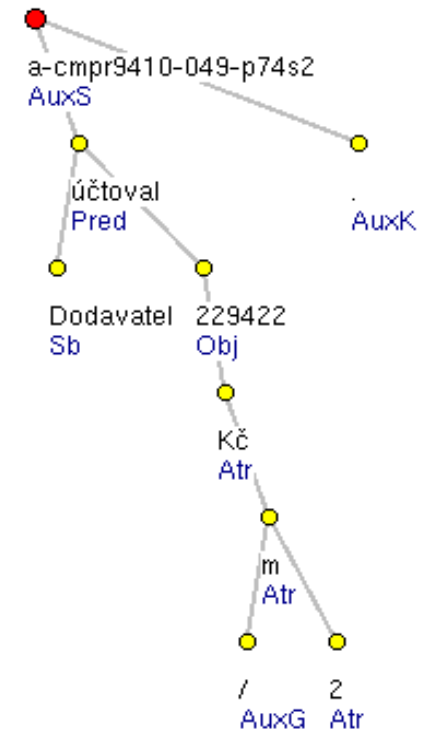
Pro firmy, které se rozhodnou využít jejich služeb, to nemusí být zrovna levná záležitost.



Filtrování vět

Grafické symboly (AuxG)

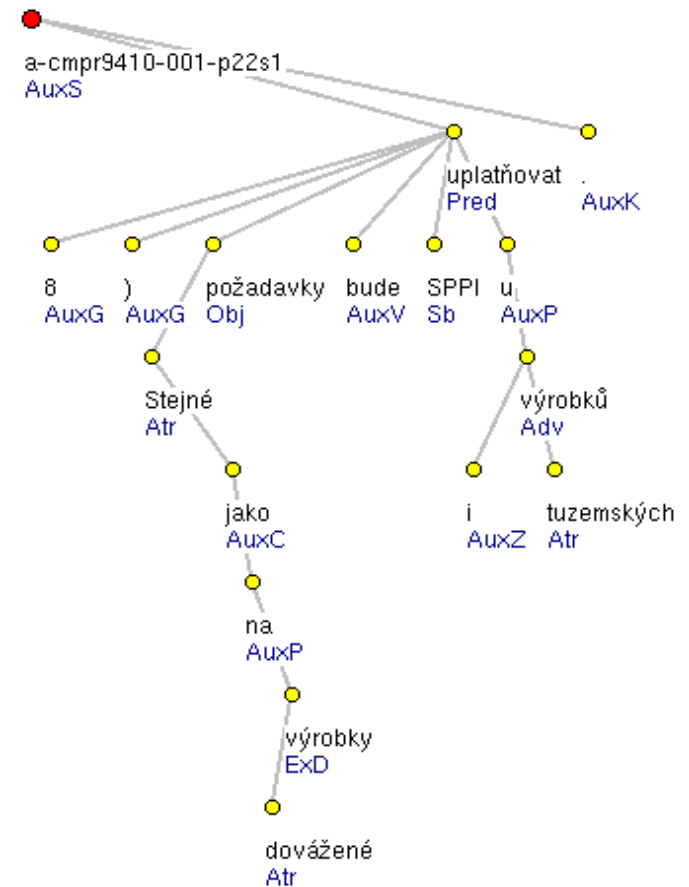
Dodavatel účtoval 229422 Kč / m².



Filtrování vět

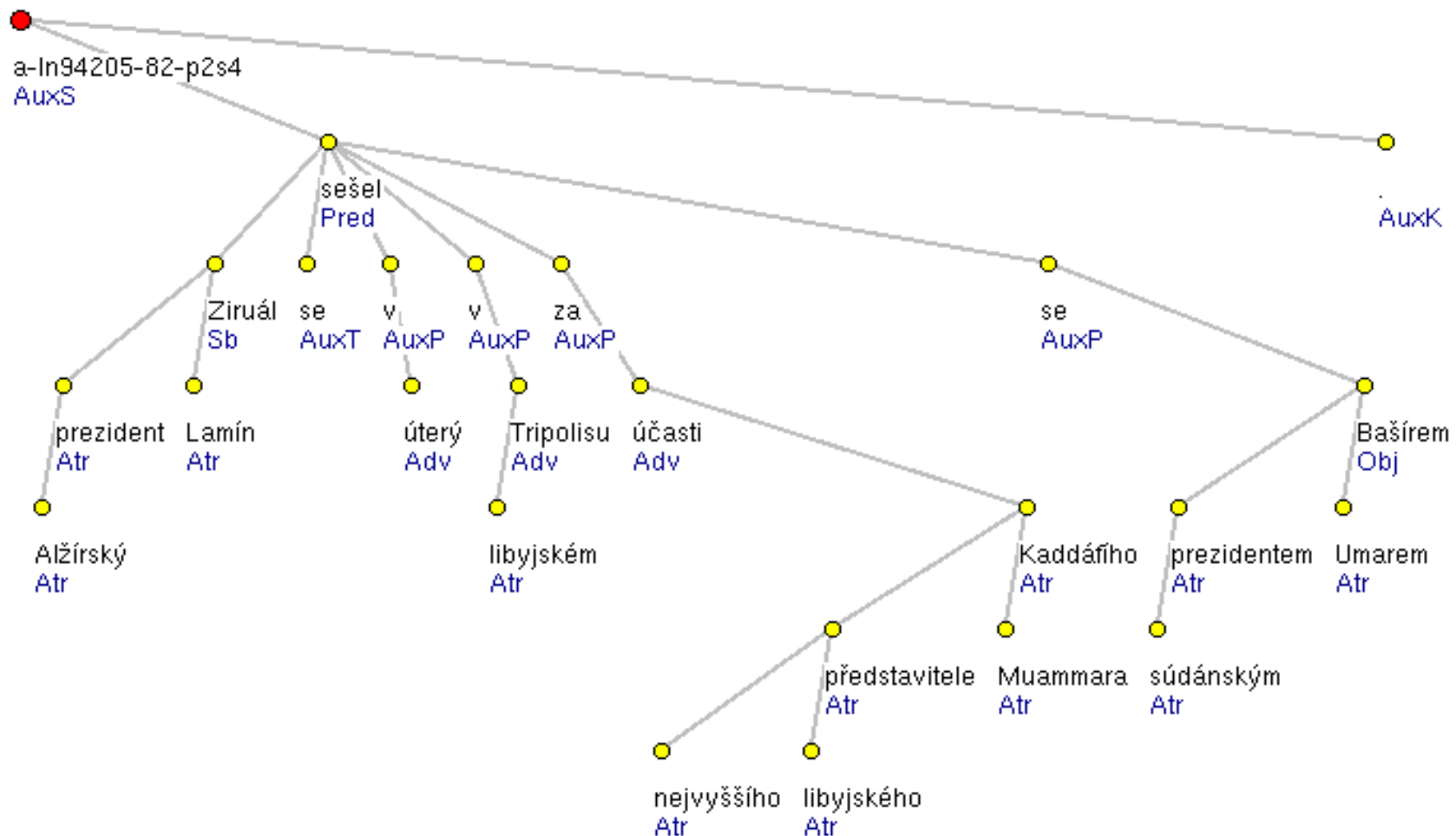
Elipsy, aposice (ExD, Apos)

8) *Stejně požadavky jako na dovážené výrobky bude SPPI uplatňovat i u výrobků tuzemských.*



Filtrování vět

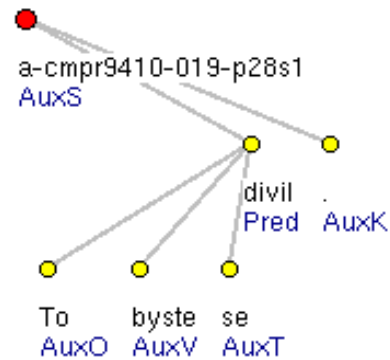
Příliš dlouhé a krátké věty



Filtrování vět

AuxO (nadbytečný (odkazovací, emotivní) element)

To byste se divil.



Filtrování vět

Filtrování v číslech

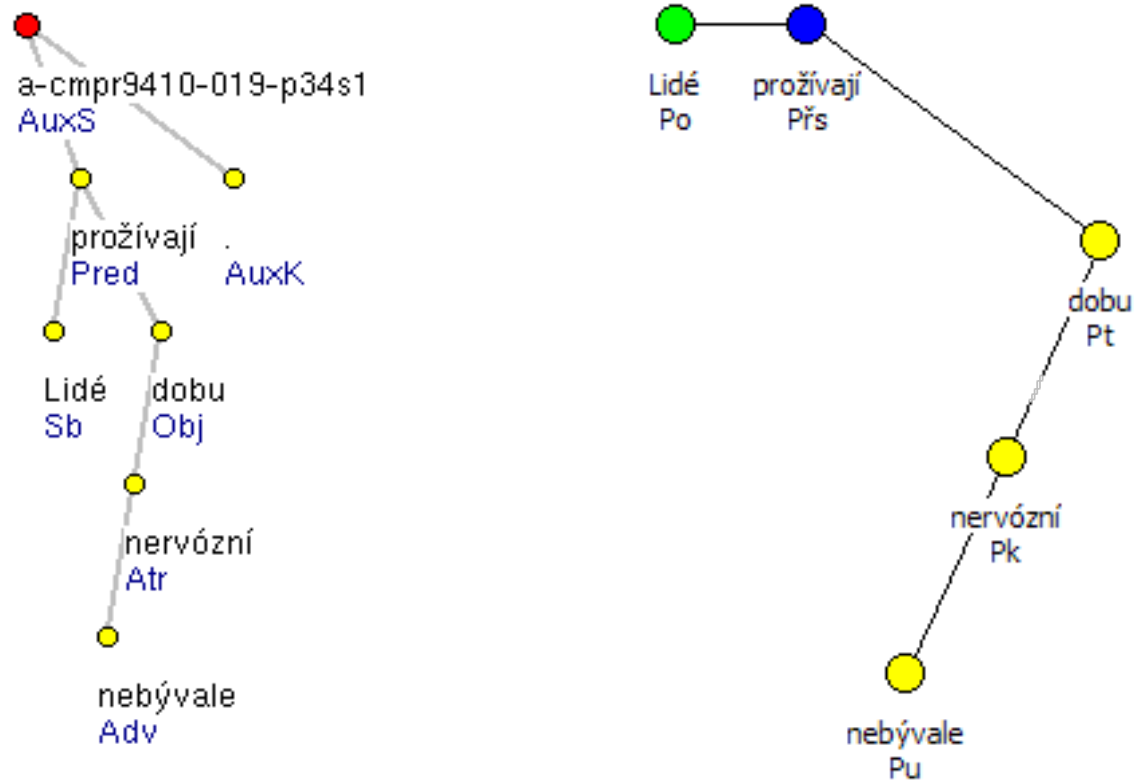
- výchozí počet vět: 49442
- po aplikaci sedmi filtrů: 11705
- zachováno tedy 23,7% vět

3. Transformace stromů

Transformace

Sb (podmět)

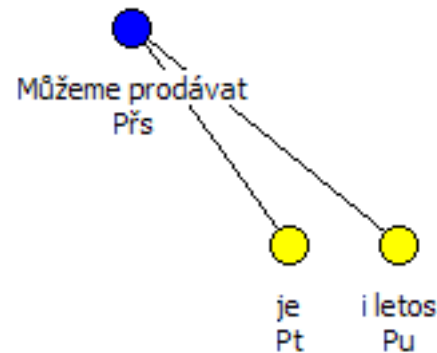
- „povýšit“ na úroveň přísudku



Transformace

Obj (předmět)

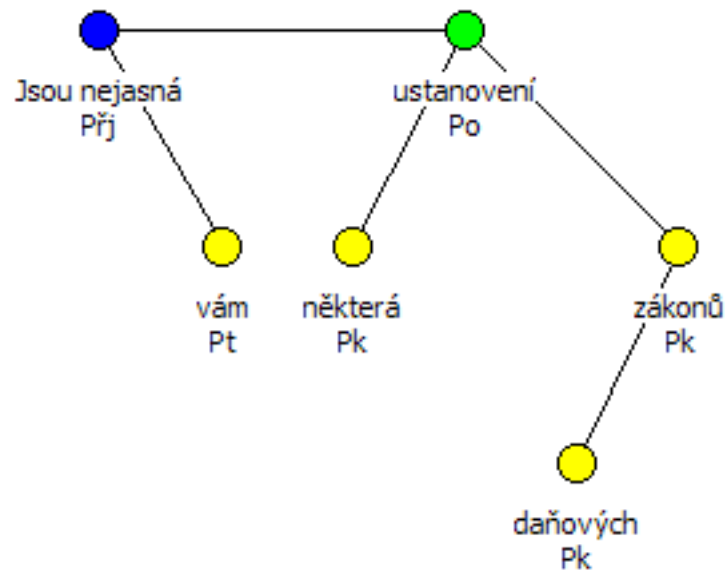
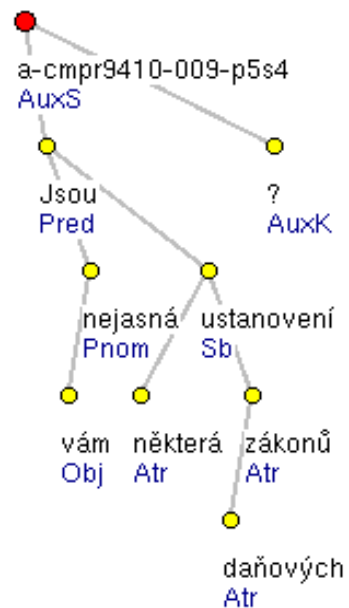
- je-li to infinitivní předmět a rodič je modální sloveso, spojit s rodičem



Transformace

Pnom (jmenná část přísudku)

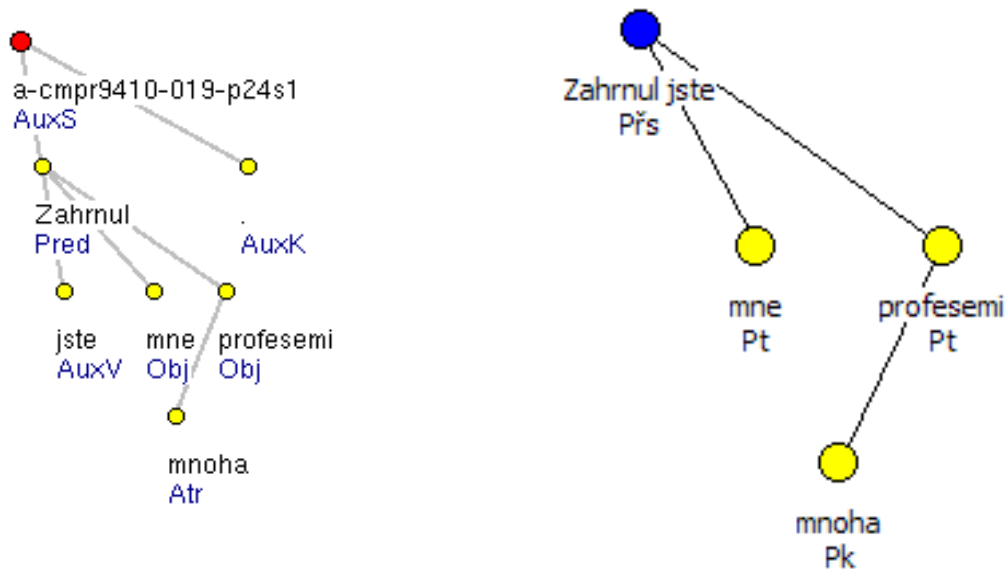
- spojit s rodičem



Transformace

AuxV (pomocné sloveso být)

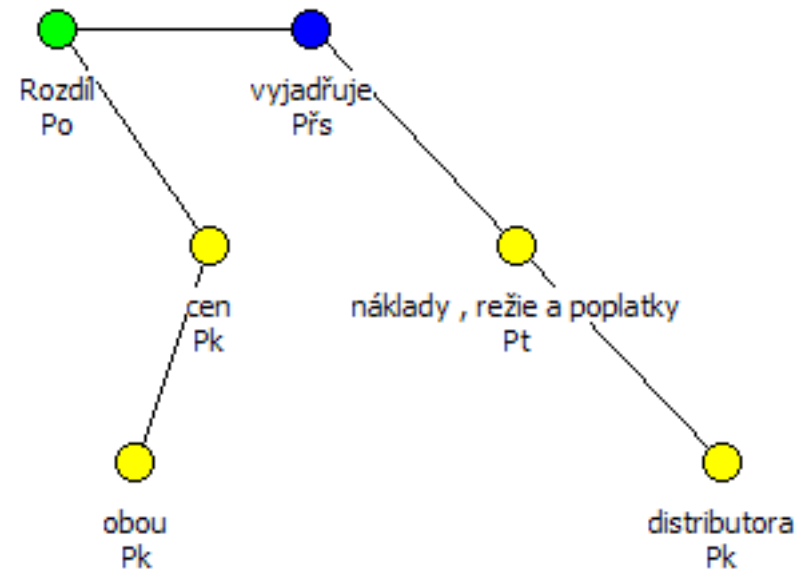
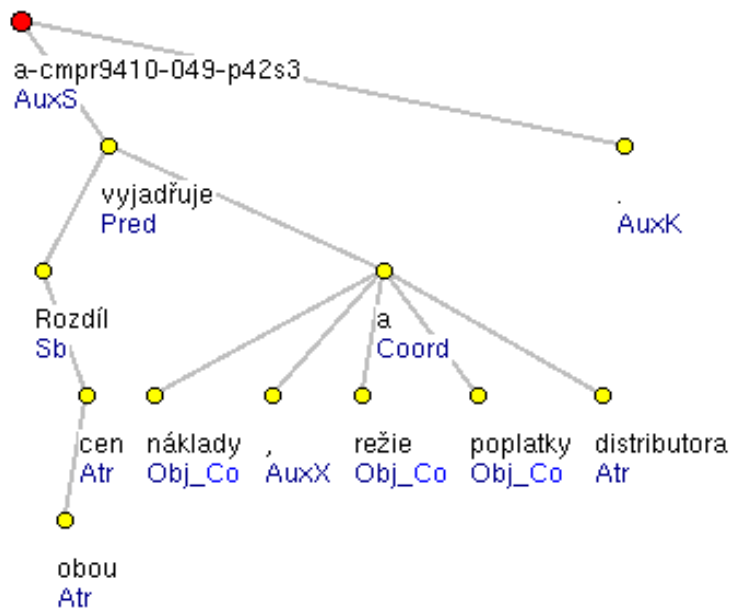
- spojit s rodičem



Transformace

Coord (koordináční uzel)

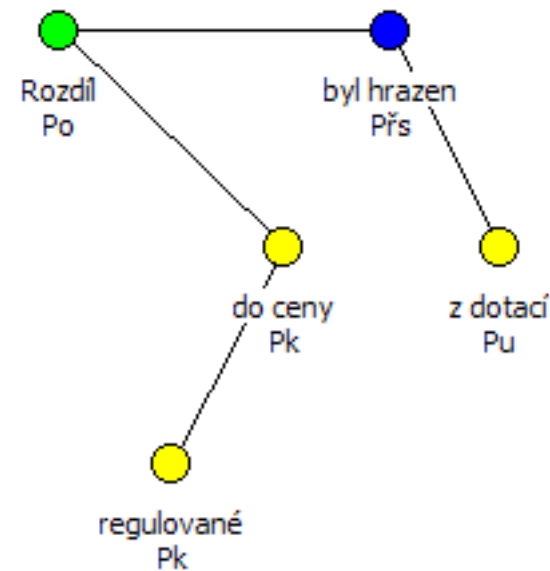
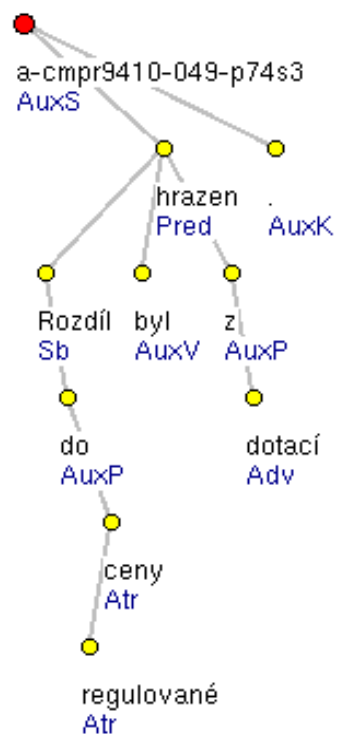
- koordinované děti (is_member) a čárky „pohlitit“ k sobě



Transformace

AuxP (předložka primární, části předložky sekundární)

- děti „pohlit“ k sobě



Transformace

Ostatní analytické funkce

- žádná transformace nebo transformace analogická k těm předchozím

4. STYX:

FilterSentences, Charon, Styx

FilterSentences

- program sloužící k aplikaci filtrů
- načítá data ve formátu PML
- každou větu otestuje filtrem
- výstupem jsou ty věty, které filtrem prošly
- výstup opět ve formátu PML

Charon

- „administrační“ program
- uživateli (administrátorovi) načte všechny věty, které jsou k dispozici
- uživatel vybírá věty, ze kterých chce složit cvičení
- na závěr si cvičení uživatel uloží

Charon

Charon

Soubor Sada

Jsou vám nejasná některá ustanovení daňových zákonů ?

Jsou vám nejasná některá ustanovení daňových zákonů ?

Je tu pro vás připravena rubrika Daňový poradce. Vaše dotazy čekáme na adrese Českomoravská 246, Brno. Tím pádem máme problém se silniční daní. Za vozidla zaměstnanců užívaná pro pracovní účely. Stejný názor má i řada našich soukromých podnikatelů. Je naprosto bezbřehý a nevypočitatelný. V tom s vámi nesouhlasím. Mzdy a platy jsou vždy podřízenými hodnotami. Zde vzniká nová dimenze srovnávání: Opět s vámi nesouhlasím. Můžete to vysvětlit na příkladu? Například podnikatel by chtěl dosáhnout u daně z příjmu. Zde se dostáváme k termínu odborná kvalifikace. Jak ho vlastně pozná? Zahrnul jste mne mnoha profesemi. Naše metodika a techniky vyšetřování mohou být užitečné. Je to vada zadavatelů nebo testů? V celé naší tříleté činnosti si jen čtyři firmy omlouvaly. U posudků v minulosti mohl být sebemenší nedostatek. Lidé prožívají nebývale nervózní dobu. Stále jim nedáváme odpovídající péči. Zejména v Olomouci firma svými výrobky při restaurátorské licenci umožňuje práci v celkové klidnosti. Firemní výrobky zdobí galerie a soukromá sídla. Balení másla ve hliníkové fólii zajišťuje jeho ochranu. Samotné měření spotřeby tepla peněženku i v zimě. Platíme hodně hlavně díky nízké technické úrovni. Stavět vlastní výtopnu je prozatím patrně nejvýhodnější. Zásadní pákou je tlak na naši peněženku. Ceny energie rostou nepřetržitě a ve skocích. Pak se dá cena tepla v tryskový let. Ostravský podnik vyrábí teplo v šesti divizích. Jeho rozsah působnosti je úctyhodný: Růst počtu teplem zásobovaných bytů přitom. Počet pracovníků podniku zároveň klesl v průběhu let. Podnik platil doposud za znečišťování ovzduchu. Ročně vyprodukoval 280 - 350 tisíc tun oxidu uhličitého.

```
graph TD; A[Jsou nejasná Přj] --- B[vám Pt]; A --- C[ustanovení Po]; C --- D[některá Pk]; C --- E[zákonů Pk]; E --- F[daňových Pk];
```


Charon

The screenshot shows the Charon application window. The title bar reads "Charon" and the menu bar contains "Soubor" and "Sada".

Left Panel (Text Document):

Jsou vám nejasná někto Přidat věty do výběru
Je tu pro vás připravena rubrika Daňový po
Vaše dotazy čekáme na adrese Českomorav
Tím pádem máme problém se silniční daní .
Za vozidla zaměstnanců užívaná pro pracov
Stejný názor má i řada našich soukromých p
Je naprosto bezbřehý a nevypočitatelný .
V tom s vámi nesouhlasím .
Mzdy a platy jsou vždy podřízenými hodnoc
Zde vzniká nová dimenze srovnávání :
Opět s vámi nesouhlasím .
Můžete to vysvětlit na příkladu ?
Například podnikatel by chtěl dosáhnout u z
Zde se dostáváme k termínu odborná kvalifi
Jak ho vlastně pozná ?
Zahrnul jste mne mnoha profesemi .
Naše metodika a techniky vyšetřování moh
Je to vada zadavatelů nebo testů ?
V celé naší tříleté činnosti si jen čtyři firmy ol
U posudků v minulosti mohl být sebemenší n
Lidé prožívají nebývale nervózní dobu .
Stále jim nedáváme odpovídající péč .
Zejména v Olomouci firma svými výrobky při
Restaurátorské licence umožňují práce v cel
Firemní výrobky zdobí galerie a soukromá sí
Balení másla ve hliníkové fólii zajišťuje jeho
Samotné měření spotřeby tepla peněženku i
Platíme hodně hlavně díky nízké technické ú
Stavět vlastní výtopnu je prozatím patrně ni
Zásadní pákou je tlak na naši peněženku .
Ceny energie rostou nepřetržitě a ve skocíc
Pak se dá cena tepla v tryskový let .
Ostravský podnik vyrábí teplo v šesti divizíc
Jeho rozsah působnosti je úctyhodný :
Růst počtu teplem zásobovaných bytů přito
Počet pracovníků podniku zároveň klesl v pr
Podnik platil doposud za znečišťování ovzdu
Ročně vyprodukuje 280 - 350 tisíc tun
č. dílně. Nová šifra. ...

Right Panel (Diagram):

Lidé prožívají nebývale nervózní dobu .

The diagram is a semantic network with nodes and edges:

- Green node: Lidé (Po)
- Blue node: prožívají (Přs)
- Yellow node: dobu (Pt)
- Yellow node: nervózní (Pk)
- Yellow node: nebývale (Pu)

Connections: Lidé (Po) - prožívají (Přs) - dobu (Pt) - nervózní (Pk) - nebývale (Pu)

Charon

The screenshot shows the Charon software interface with a text document on the left and a morphological analysis window in the center. The text document contains several paragraphs of text, with the sentence "Lidé prožívají nebývale nervózní dobu ." highlighted. The morphological analysis window displays the following information:

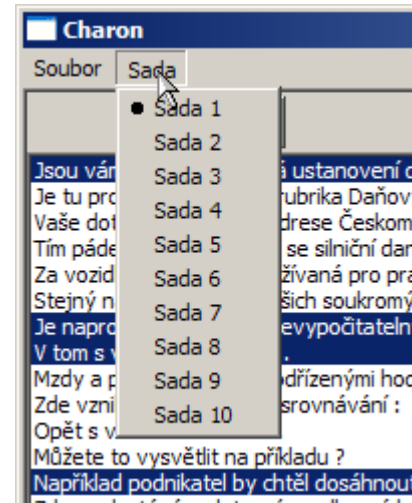
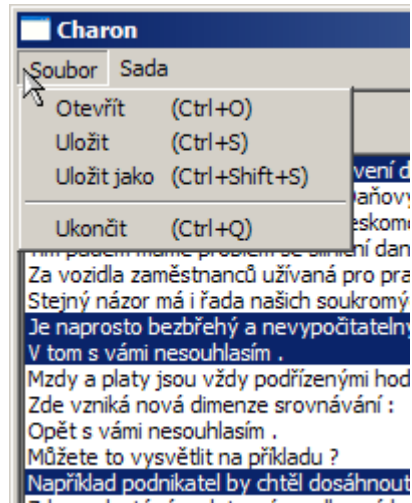
- slovní tvar: Lidé
- lemma: člověk
- slovní druh: podstatné jméno
- rod: mužský
- číslo: množné
- pád: první

The analysis window also shows a morphological tree diagram with three nodes connected by lines:

- Top node: dobu (Pt)
- Middle node: nervózní (Pk)
- Bottom node: nebývale (Pu)

The right side of the interface shows a scrollable area with the same text as the left panel, with the highlighted sentence "Lidé prožívají nebývale nervózní dobu ." visible at the top.

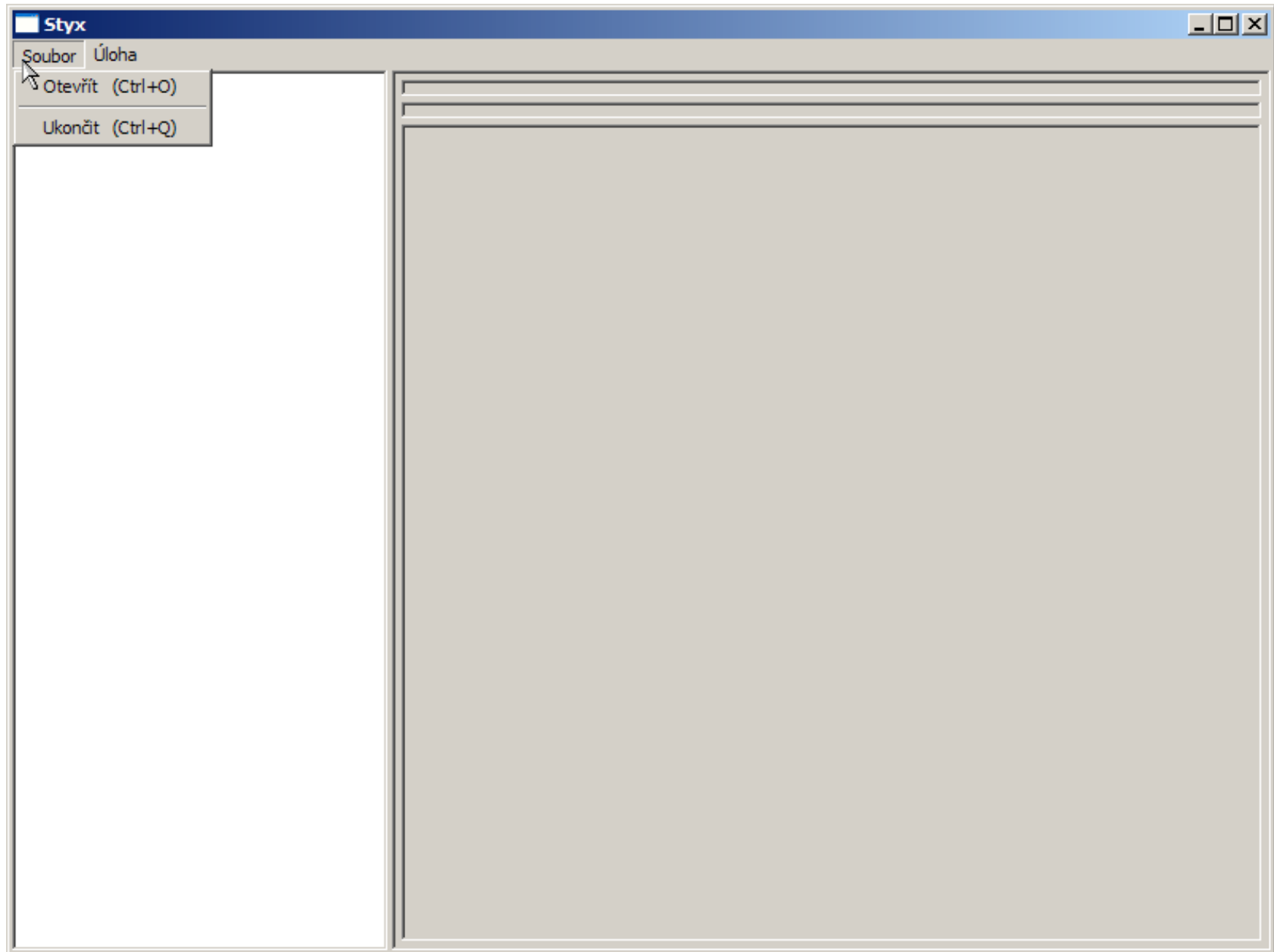
Charon



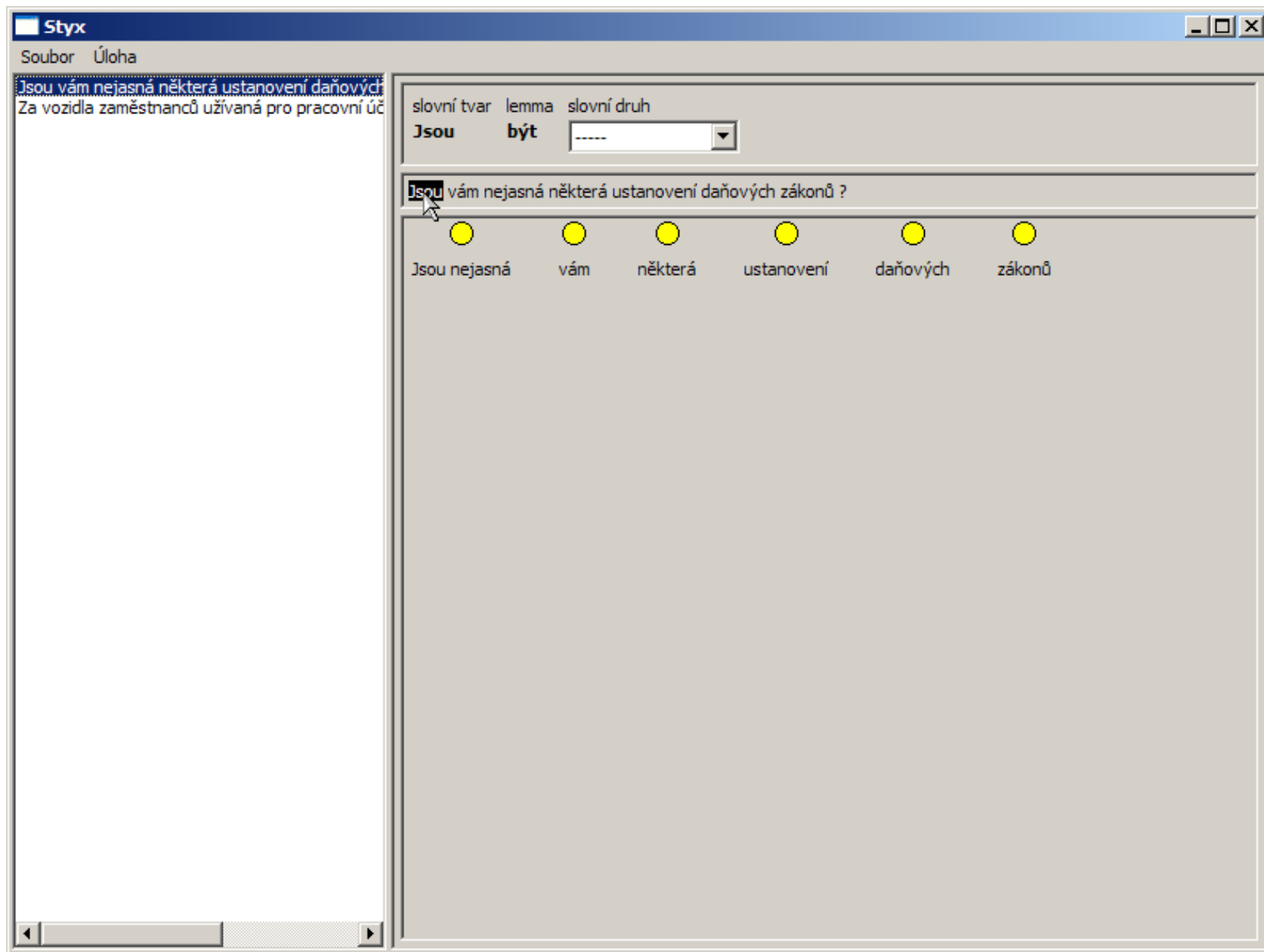
Styx

- vlastní cvičebnice
- uživatel si načte cvičení dříve vytvořené programem Charon

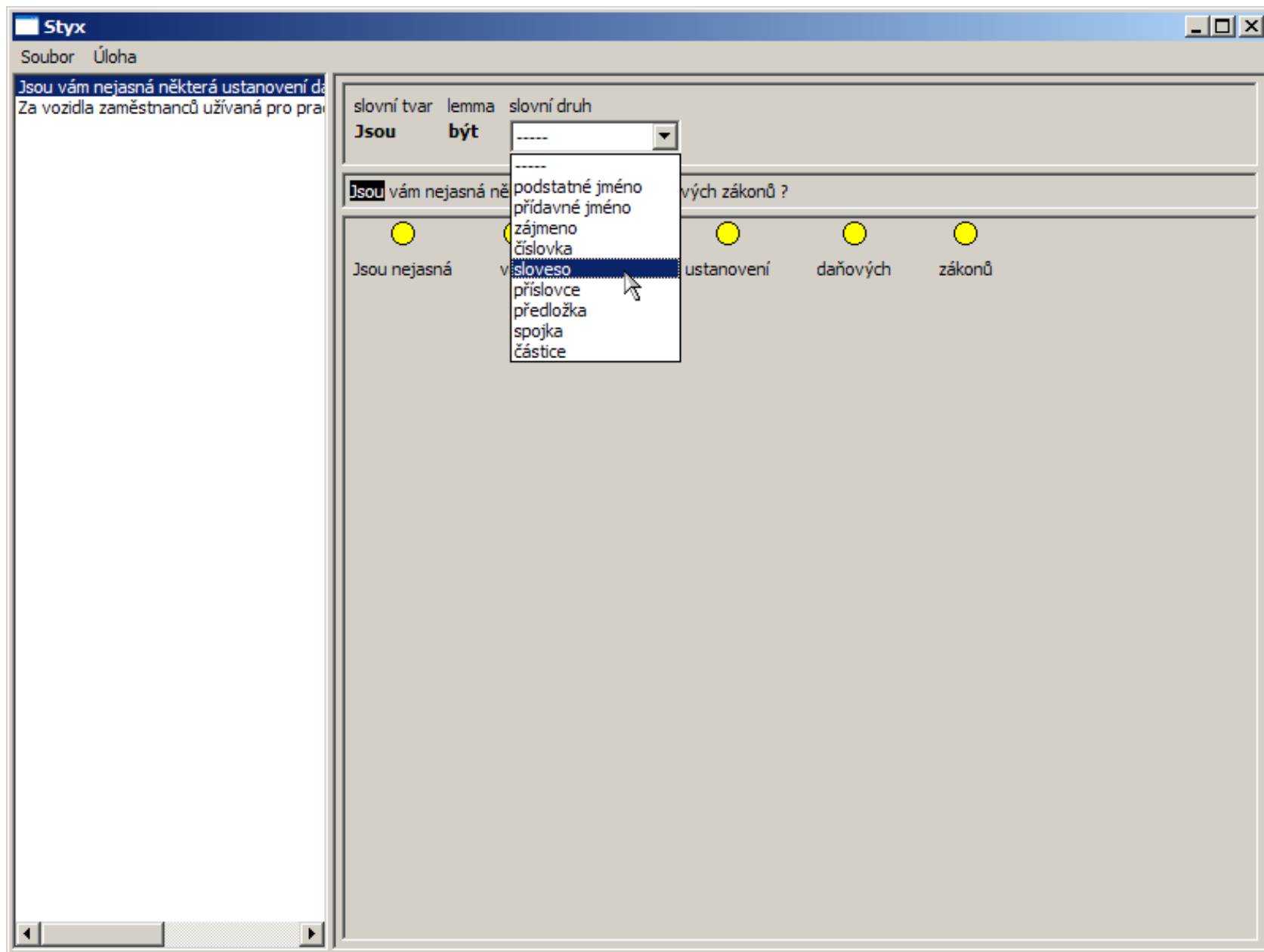
Styx



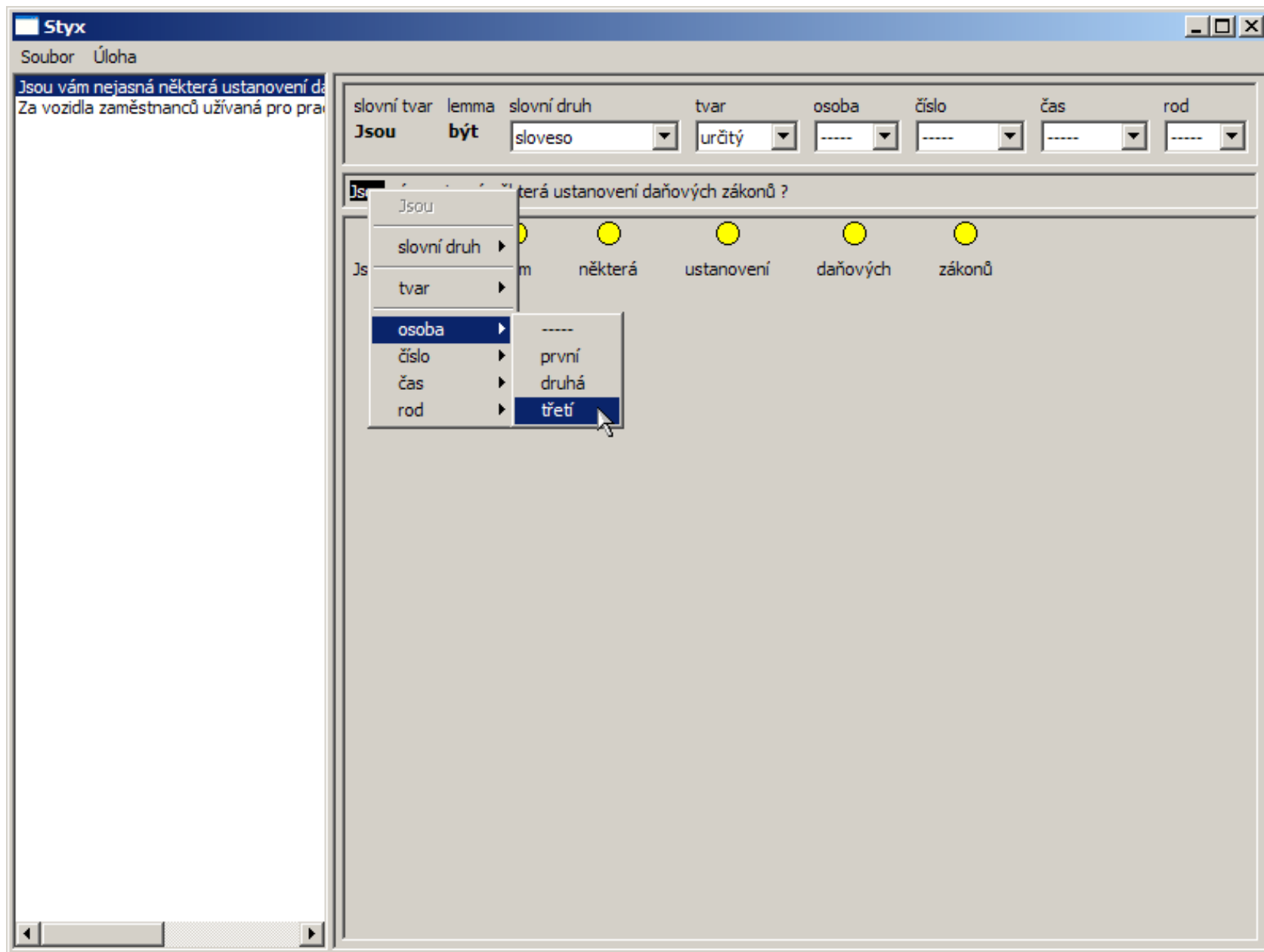
Styx



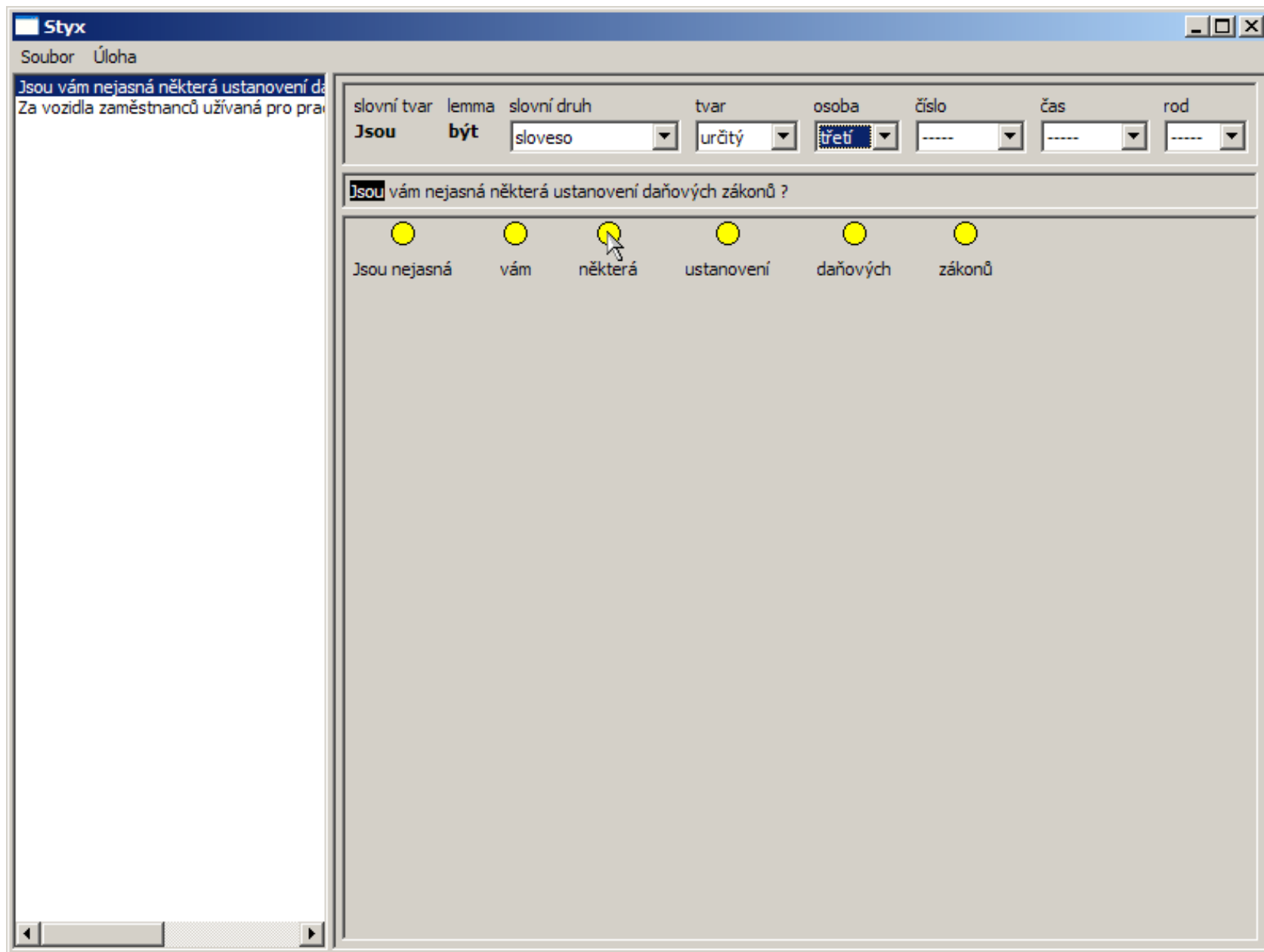
Styx



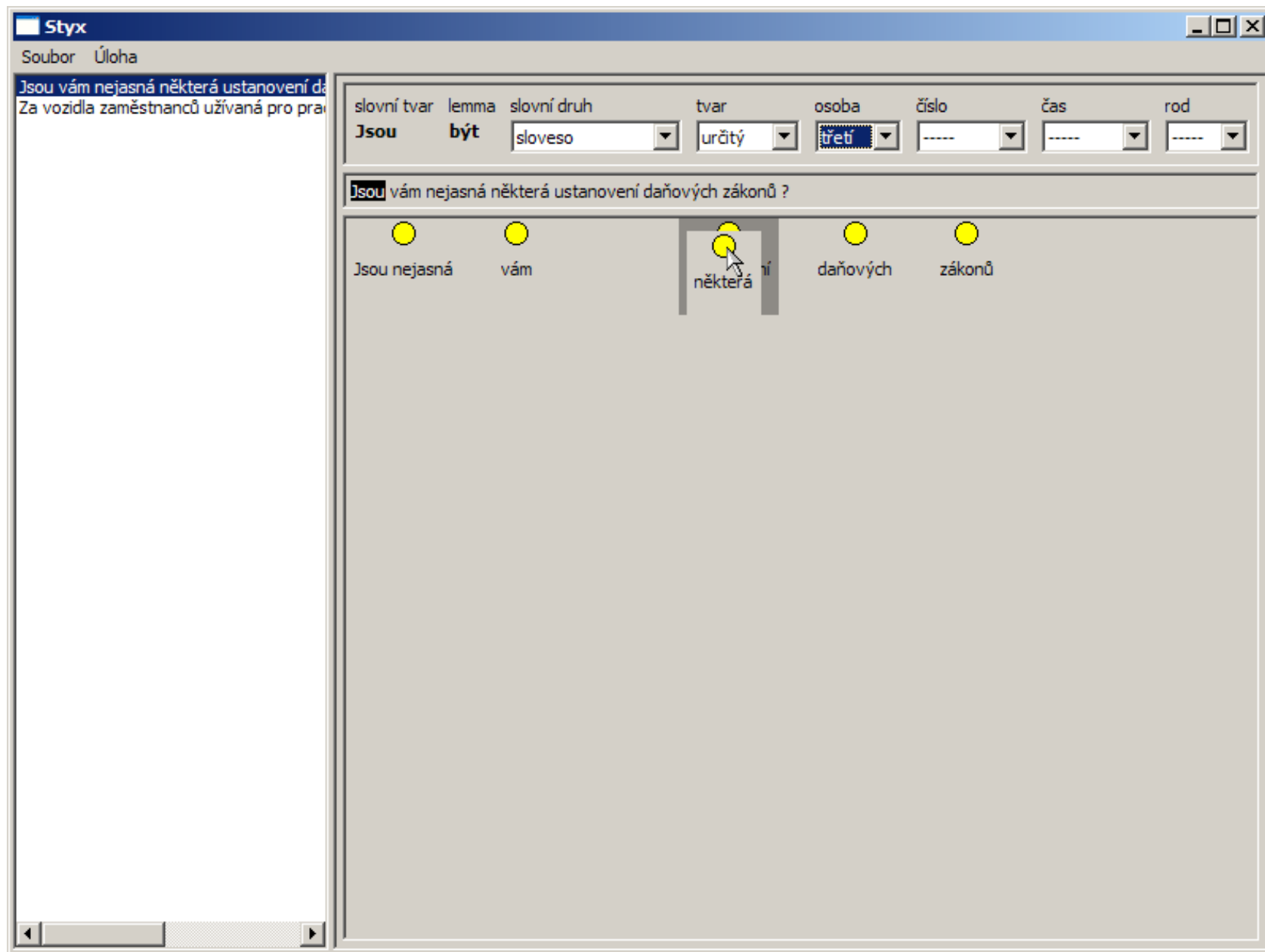
Styx



Styx



Styx



Styx

The screenshot shows the Styx software interface. The window title is "Styx". The menu bar contains "Soubor" and "Úloha". The main text area contains the sentence "Jsou vám nejasná některá ustanovení daňových zákonů?". Below the text, a morphological analysis is displayed for the word "ustanovení". The analysis shows the word "ustanovení" with a yellow dot above it, and a line connecting it to a yellow dot below it labeled "některá". Above the word "ustanovení" are five yellow dots, each corresponding to a word in the sentence: "Jsou", "vám", "ustanovení", "daňových", and "zákonů".

Jsou vám nejasná některá ustanovení daňových zákonů ?

Jsou nejasná vám ustanovení daňových zákonů

některá

slovní tvar lemma slovní druh tvar osoba číslo čas rod
Jsou být sloveso určitý třetí ----- ----- -----

Styx

The screenshot shows the Styx software interface. The window title is "Styx". The menu bar contains "Soubor" and "Úloha". The main text area displays the sentence: "Jsou vám nejasná některá ustanovení daňových zákonů ?". Below the text, a morphological analysis tree is shown for the word "některá". The root node is "některá", which is connected to a child node "někt". A context menu is open over the "někt" node, listing various grammatical categories. The "přívlastek" (adjective) option is highlighted by the mouse cursor. The menu also includes options for "přísudek slovesný", "přísudek jmenný", "podmět", "předmět", "příslovné určení", and "doplňěk".

Styx

Soubor Úloha

Jsou vám nejasná některá ustanovení daňových zákonů ?

Jsou nejasná vám ustanovení daňových zákonů

slovní tvar lemma slovní druh tvar osoba číslo čas rod

Jsou být sloveso určitý třetí ----- ----- -----

některá

někt

-
- přísudek slovesný
- přísudek jmenný
- podmět
- předmět
- přívlastek**
- příslovné určení
- doplňěk

Styx

Styx

Soubor Úloha

Jsou vám nejasná některá ustanovení daňových zákonů ?
Za vozidla zaměstnanců užívaná pro práci

slovní tvar lemma slovní druh tvar osoba číslo čas rod

Jsou být sloveso určitý třetí množné přítomný činný

Jsou vám nejasná některá ustanovení daňových zákonů ?

```
graph TD; A((Jsou nejasná Přs)) --- B((ustanovení Po)); A --- C((vám Pt)); B --- D((některá Pk)); B --- E((daňových Pk)); B --- F((zákonů Pk)); E --- G((zákonů Pk));
```

Styx

The screenshot shows the Styx software interface. At the top, there is a menu bar with 'Soubor' and 'Úloha'. Below the menu bar, there are three buttons: 'Jsou vám', 'Zkontrolovat', and 'anovení da'. The main window displays a morphological analysis of the sentence 'Jsou vám nejasná některá ustanovení daňových zákonů?'. The analysis is shown as a tree structure with nodes and edges. The root node is a blue circle labeled 'Jsou nejasná Přs'. It branches into two nodes: a yellow circle labeled 'vám Pt' and a green circle labeled 'ustanovení Po'. The 'ustanovení Po' node branches into two nodes: a yellow circle labeled 'některá Pk' and a yellow circle labeled 'daňových Pk'. The 'daňových Pk' node branches into a yellow circle labeled 'zákonů Pk'. The interface also includes a toolbar with various morphological parameters: 'slovní tvar' (word form), 'lemma', 'slovní druh' (word class), 'tvar' (form), 'osoba' (person), 'číslo' (number), 'čas' (time), and 'rod' (gender). The current settings are: 'Jsou' (word form), 'být' (lemma), 'sloveso' (word class), 'určitý' (form), 'třetí' (person), 'množné' (number), 'přítomný' (time), and 'činný' (gender).

Styx

Soubor Úloha

Jsou vám Zkontrolovat anovení da
Za vozidla Zmesitanciozivená pro pra

slovní tvar lemma slovní druh tvar osoba číslo čas rod
Jsou být sloveso určitý třetí množné přítomný činný

Jsou vám nejasná některá ustanovení daňových zákonů ?

Jsou nejasná Přs
vám Pt
ustanovení Po
některá Pk
daňových Pk
zákonů Pk

Styx

Styx

Jsou vám nejasná některá ustanovení daňových zákonů ?
Za vozidla zaměstnanců užívaná pro pracovní

	řešení	zadáno	
slovní tvar	Jsou		
lemma	být		
slovní druh	sloveso	sloveso	OK
osoba	třetí	třetí	OK
číslo	množné	množné	OK
čas	přítomný	přítomný	OK
rod	činný	činný	OK

Jsou nejasná
Přj

ustanovení
Po

vám
Pt

některá
Pk

zákonů
Pk

Jsou nejasná
Přs

ustanovení
Po

vám
Pt

některá
Pk

daňových
Pk

5. Závěr, budoucnost

Současnost

- systém je v použitelném stavu, nikoliv však v ideálním
- obtížnost cvičení na úrovni deváté třídy ZŠ

Budoucnost

- zaměření se na uživatele
- konfigurovatelnost
- rychlost
- opravy chyb
 - nedostatečné filtry, příliš přísné filtry
 - nedostatečné či chybné transformace

Otázky

a snad i odpovědi...