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

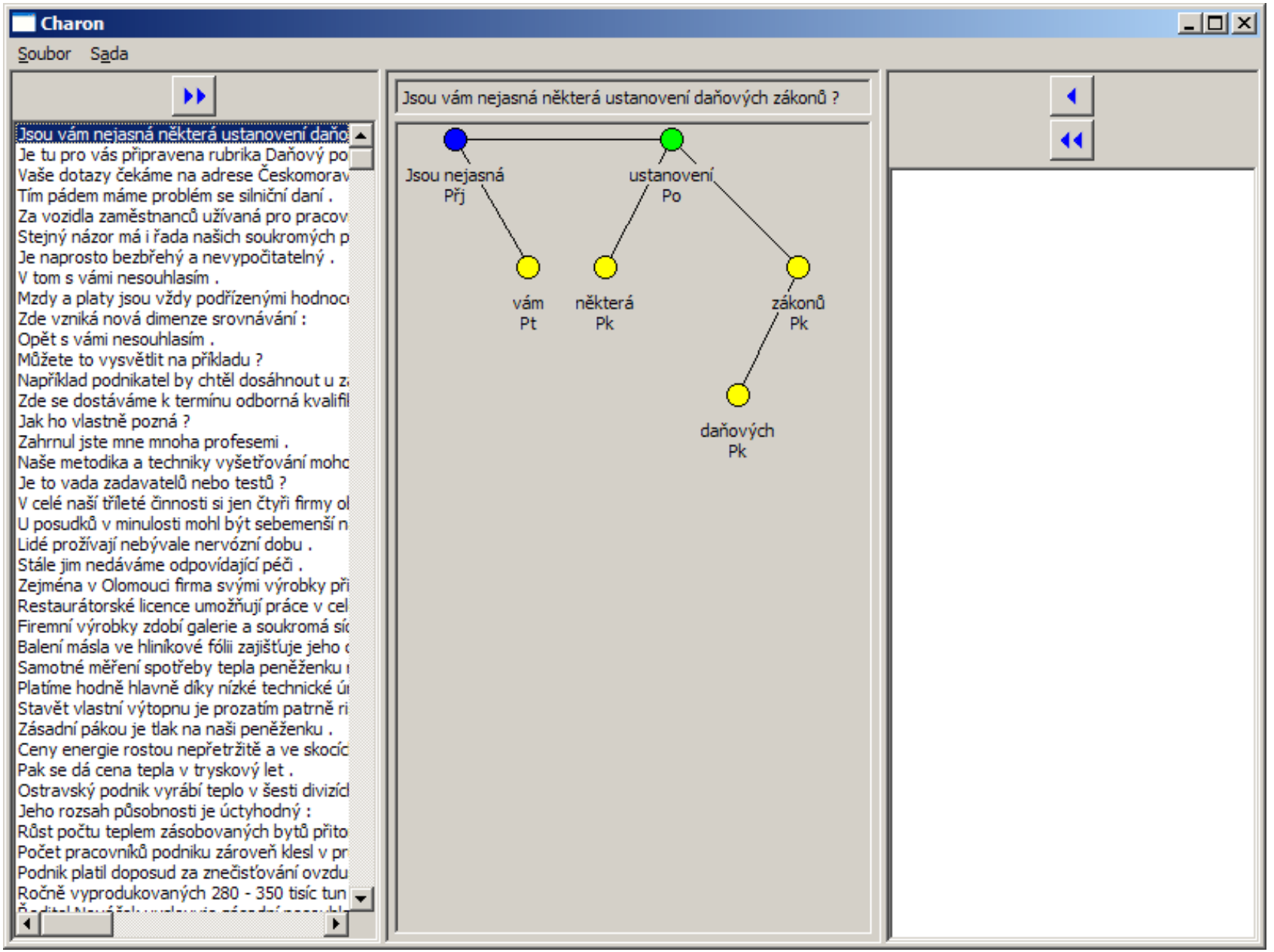
- vysokoúrovňový jazyk s řadou pojistek proti omylům programátora
- přenositelnost
- přítomnost knihovny SWT

Implementace

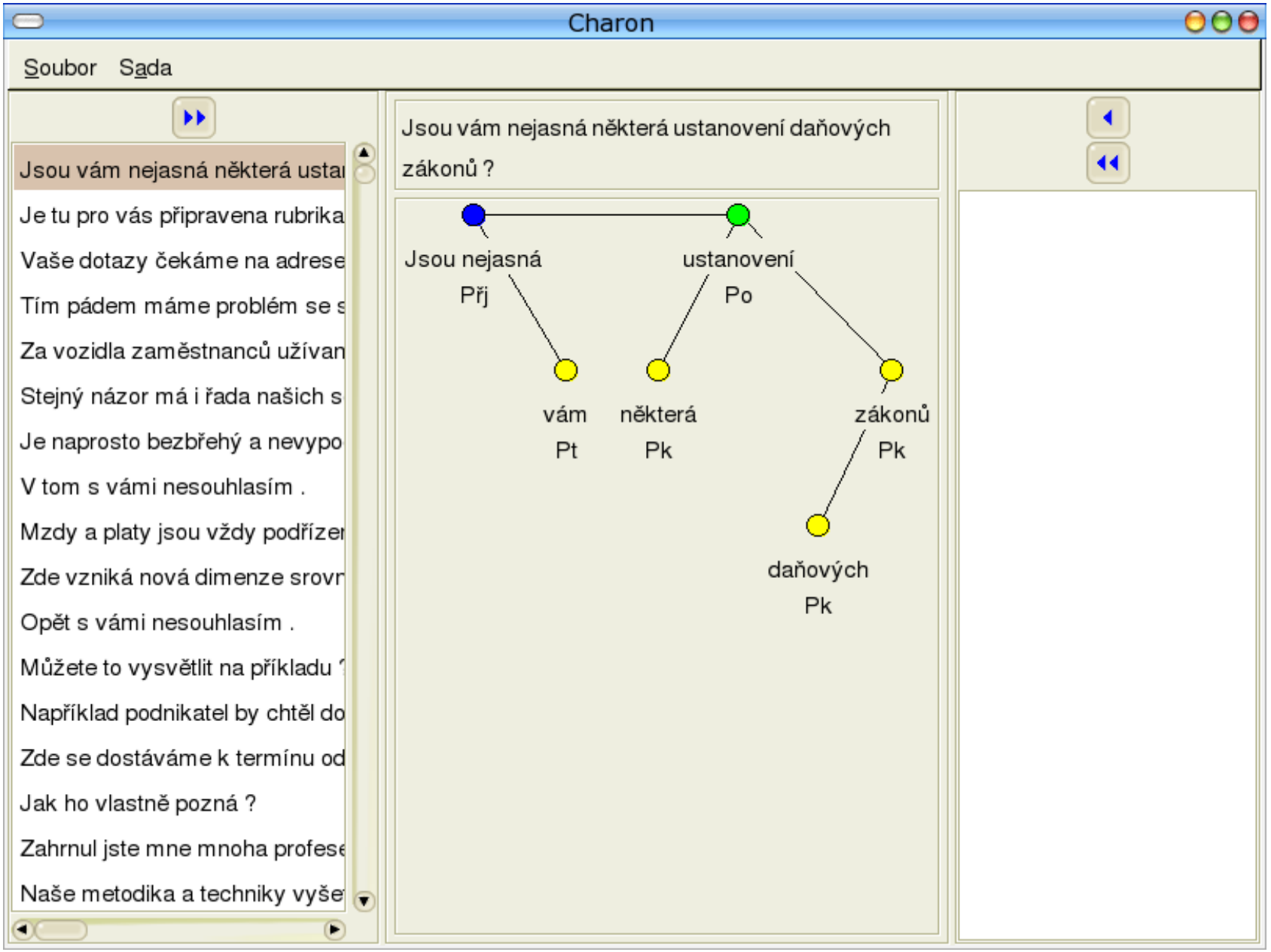
SWT

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

Implementace



Implementace

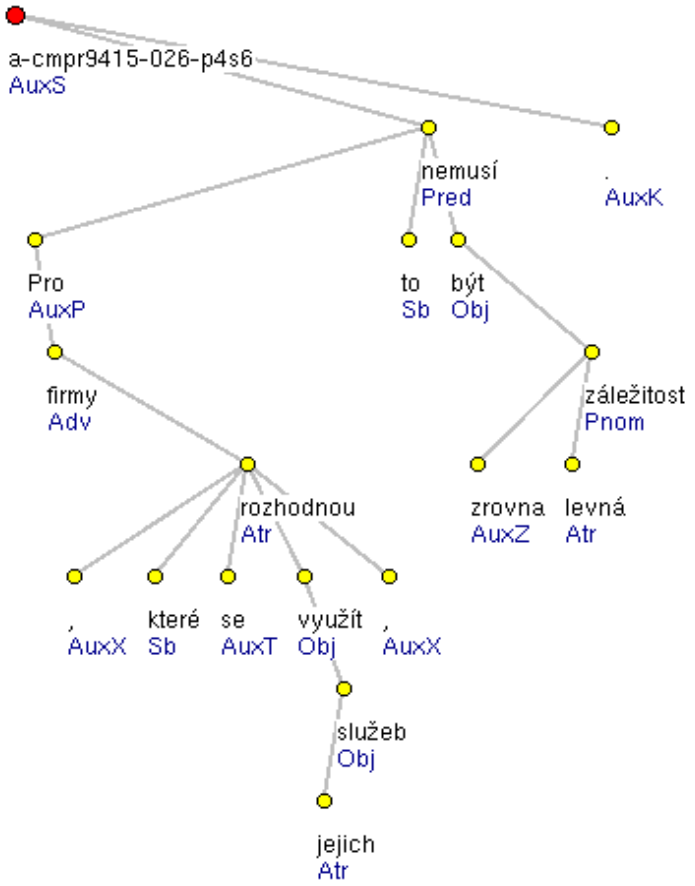


2. Filtrování vět

Filtrování vět

Souvětí (souřadná i podřadná)

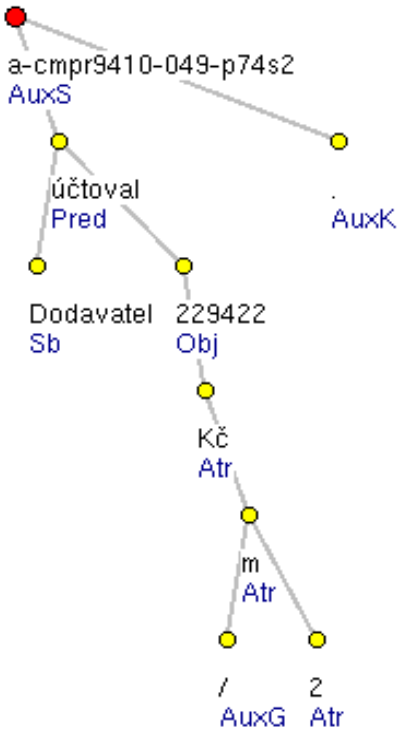
Pro firmy, které se rozhodnou využít jejich služeb, to nemusí být zrovna levná záležitost.



Filtrování vět

Grafické symboly (AuxG)

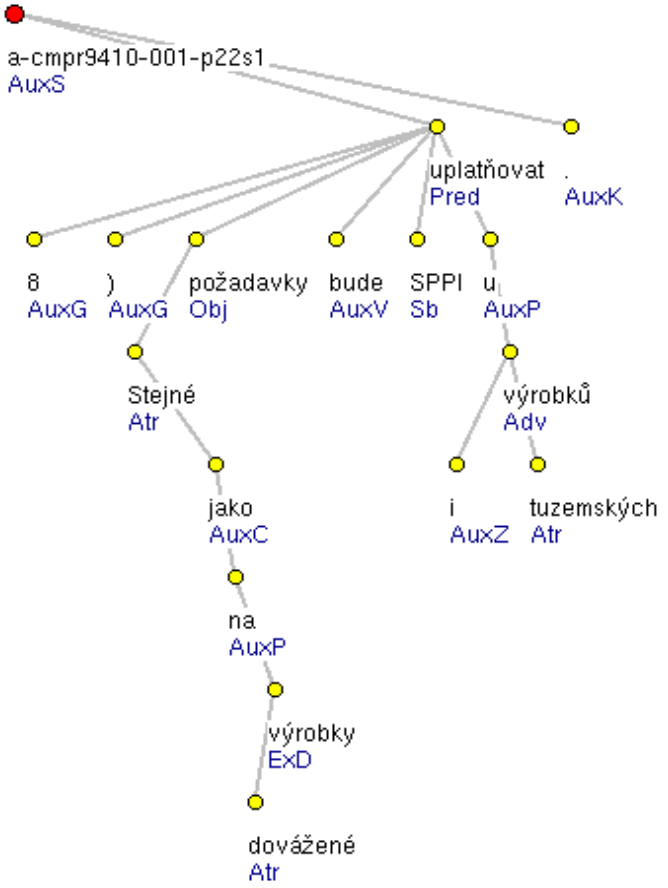
Dodavatel účtoval 229422 Kč / m².



Filtrování vět

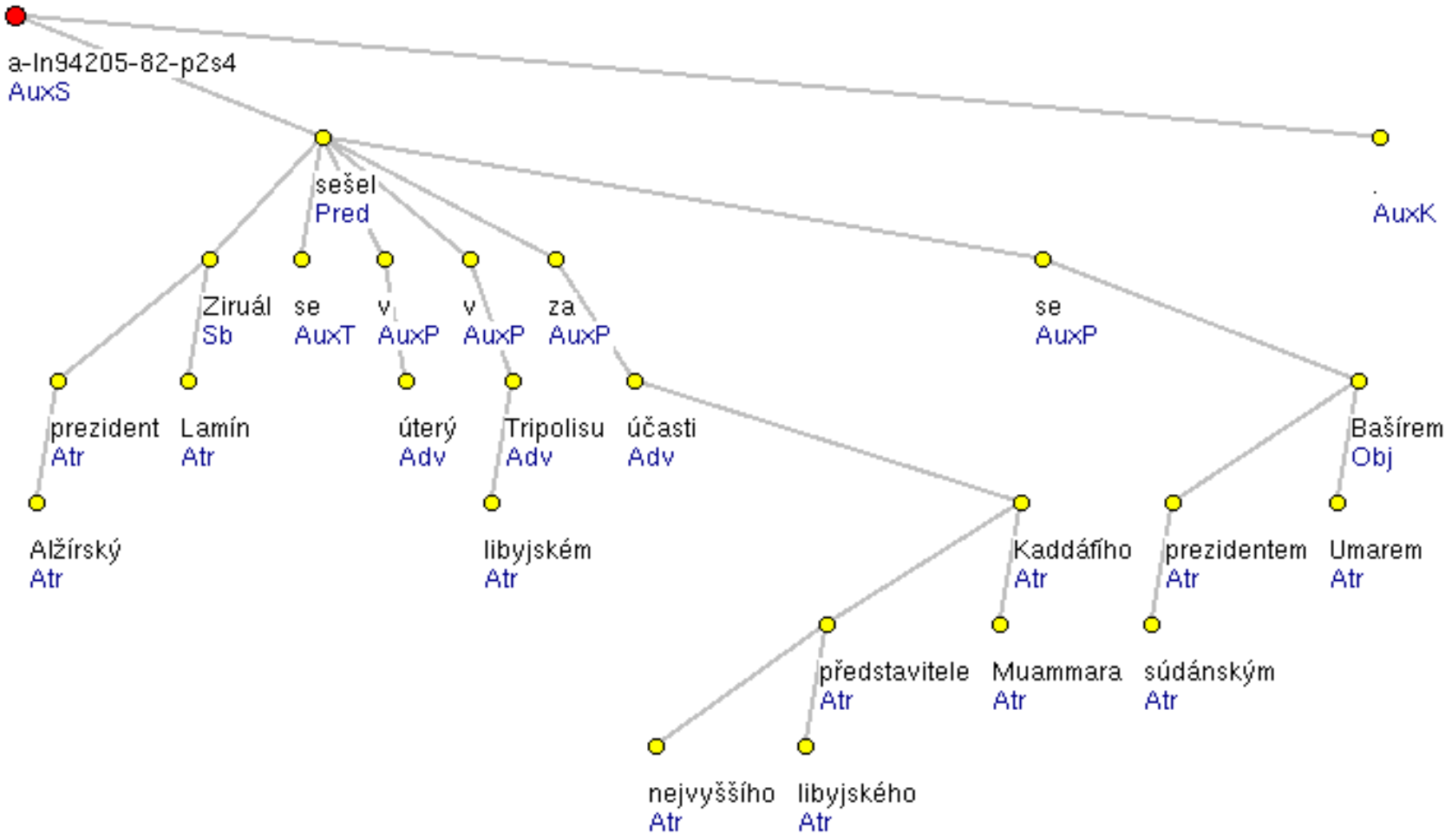
Elipsy, aposice (ExD, Apos)

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



Filtrování vět

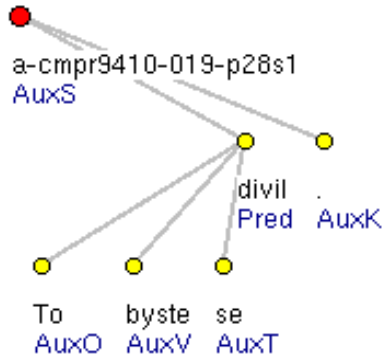
Příliš dlouhé a krátké věty



Filtrování vět

AuxO (nadbytečný (odkazovací, emotivní) element)

To byste se divil.



Filtrování vět

Filtrování v číslech

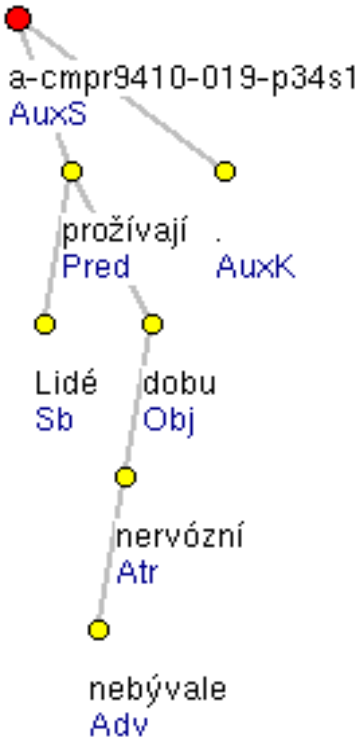
- výchozí počet vět: 49442
- po aplikaci sedmi filtrů: 11705
- zachováno tedy 23,7% vět

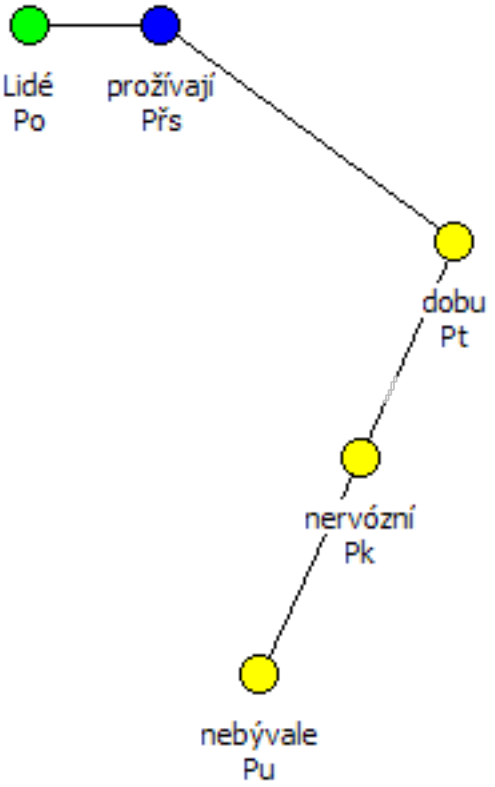
3. Transformace stromů

Transformace

Sb (podmět)

- „povýšit“ na úroveň přísudku

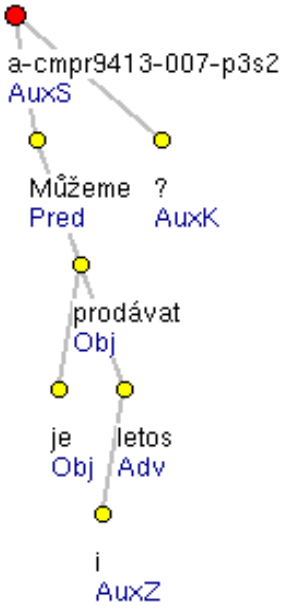


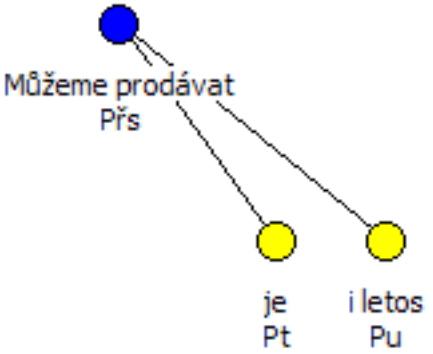


Transformace

Obj (předmět)

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

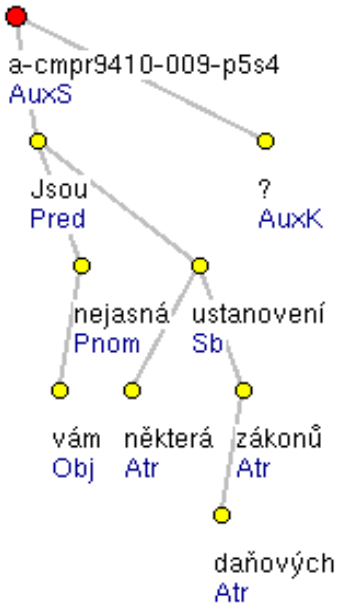


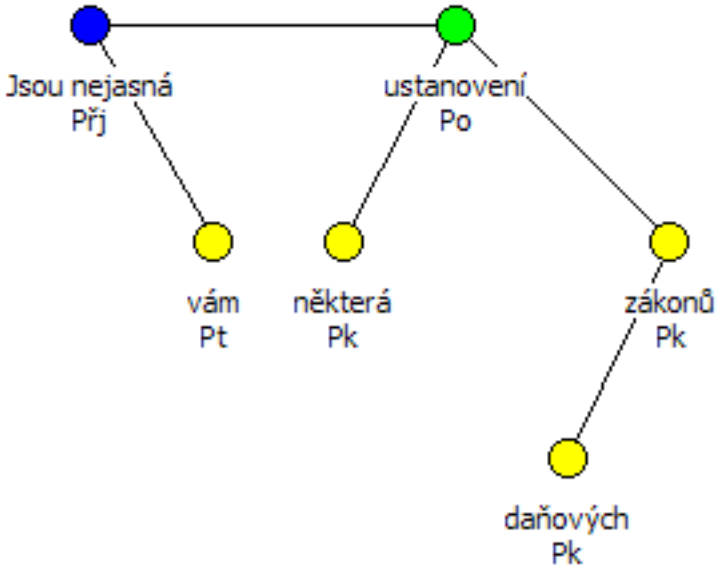


Transformace

Pnom (jmenná část přísudku)

- spojit s rodičem

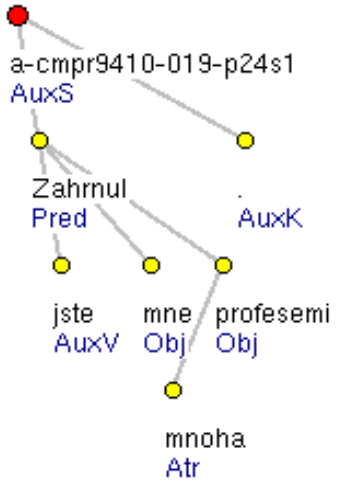


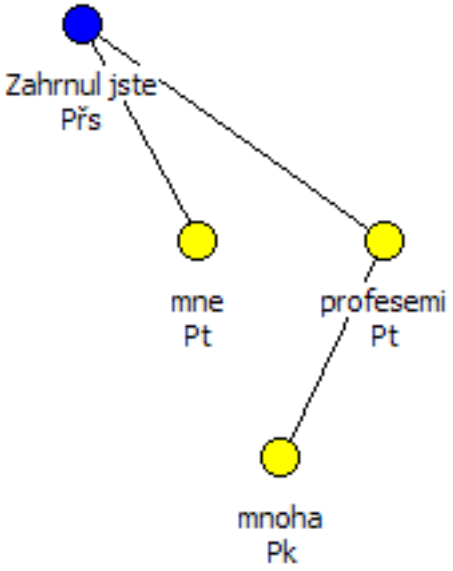


Transformace

AuxV (pomocné sloveso být)

- spojit s rodičem

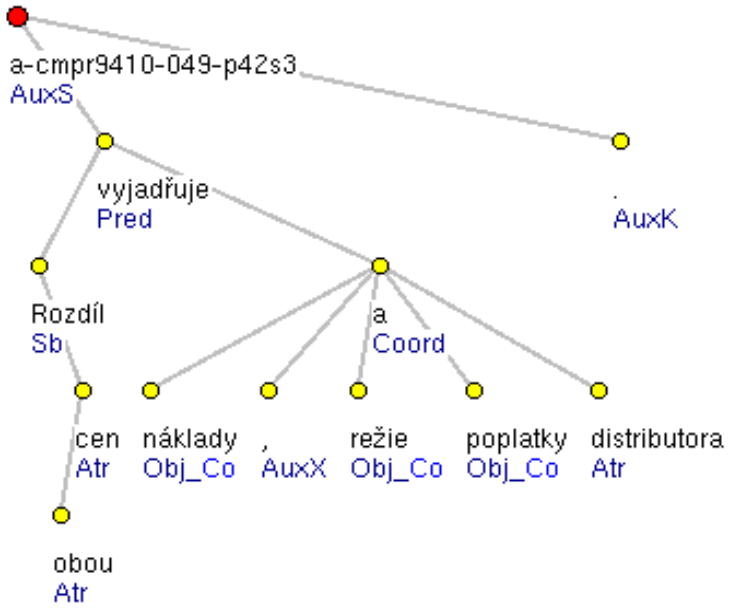


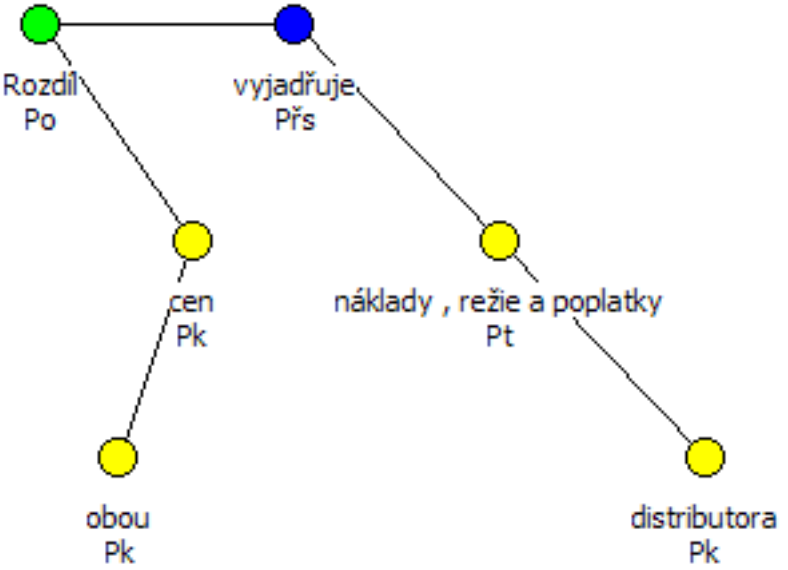


Transformace

Coord (koordinační uzel)

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

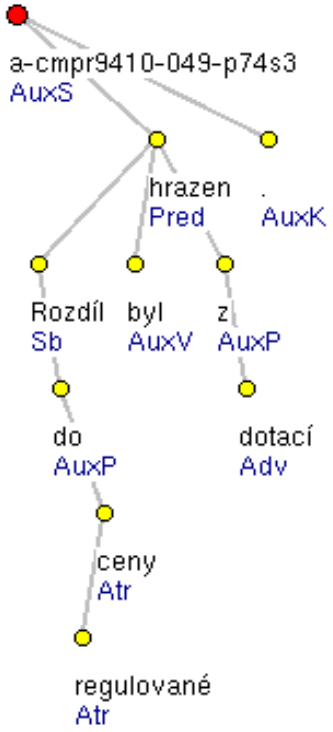


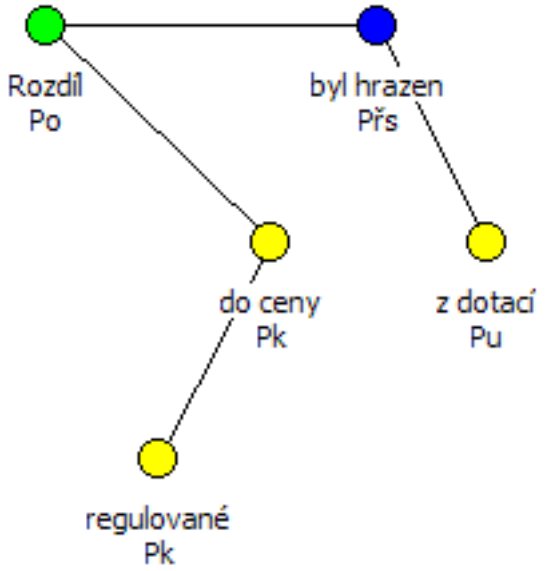


Transformace

AuxP (předložká primární, části předložky sekundární)

- děti „pohltit“ k sobě





Transformace

Ostatní analytické funkce

- žádná transformace nebo transformace analogická k těm předchozím

4. STYX:
FilterSentences, Charon, Styx

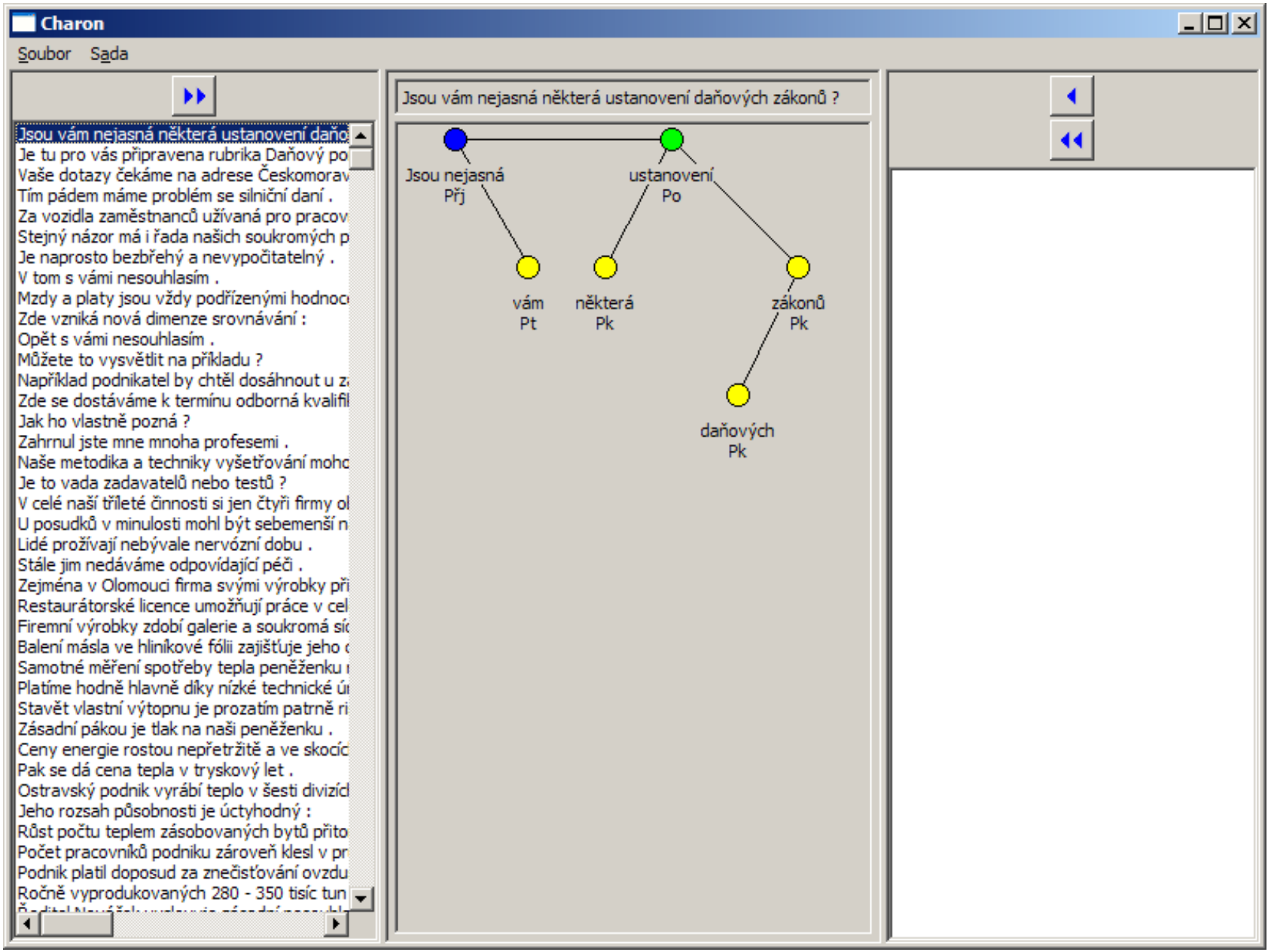
FilterSentences

- 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

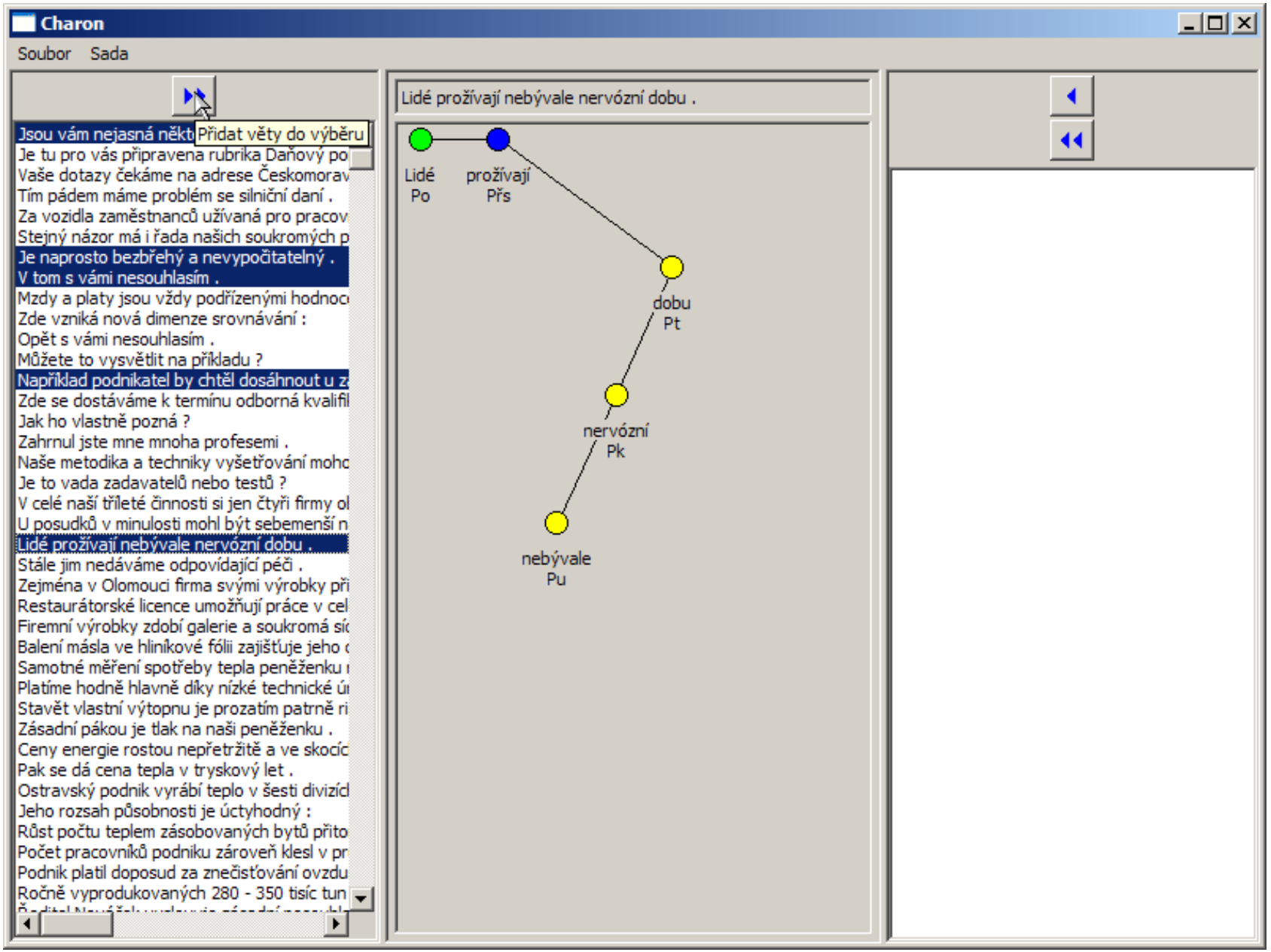
Charon

- „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

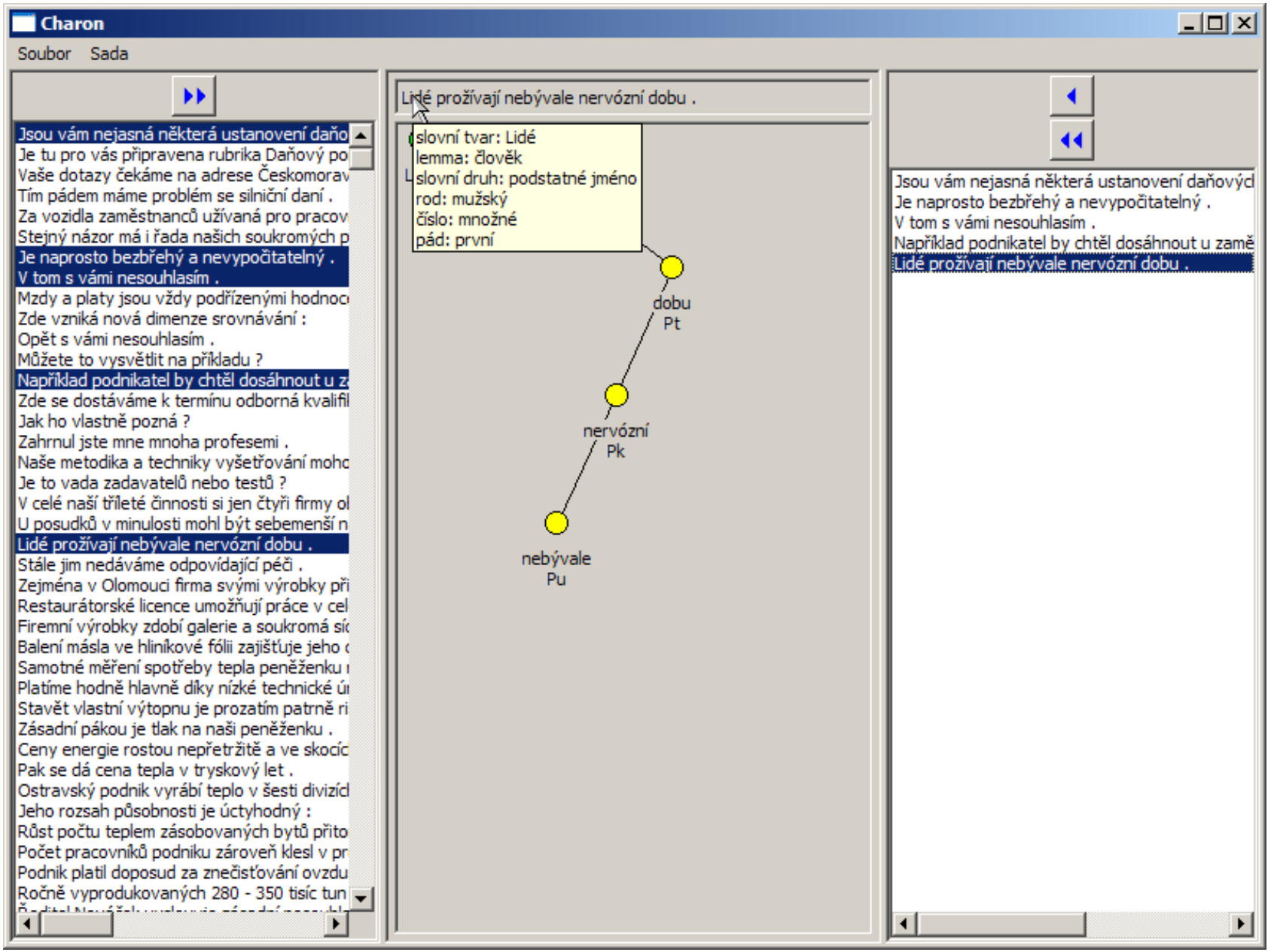
Charon


Charon
Soubor Sada
Jsou vám nejasná některá ustanovení daňo
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 péči .
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
Platíme hodně hlavně díky nízké technické ú
Stavět vlastní výtopnu je prozatím patrně ri
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 znečisťování ovzdu
Ročně vyprodukovaných 280 - 350 tisíc tun
Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná Přj
ustanovení Po
vám Pt
některá Pk
zákonů Pk
daňových Pk

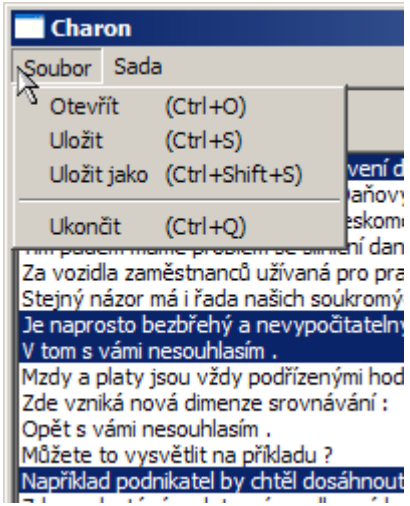
Charon

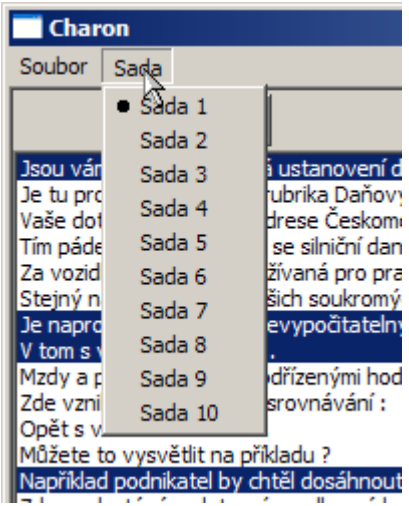

Charon
Soubor
Sada
Přidat věty do výběru
Jsou vám nejasná někt
Je tu pro vás připravena rubrika Daňový po
Vaše dotazy čekáme na adrese Českomorav
Tím pádem máme problém se silniční daní .
Za vozidla zaměstnanců užívaná pro pracov
Stejný názor má i řada našich soukromých p
Je naprosto bezbřehý a nevypočitatelný .
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 péči .
Zejména v Olomouci firma svými výrobky při
Restaurátorské licence umožňují práce v cel
Firemní výrobky zdobí galerie a soukromá síd
Balení másla ve hliníkové fólii zajišťuje jeho
Samotné měření spotřeby tepla peněženku
Platíme hodně hlavně díky nízké technické úr
Stavět vlastní výtopnu je prozatím patrně ri
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 znečisťování ovzdu
Ročně vyprodukovaných 280 - 350 tisíc tun
Lidé prožívají nebývale nervózní dobu .
Lidé
Po
prožívají
Přs
dobu
Pt
nervózní
Pk
nebývale
Pu

Charon


Charon
Soubor Sada
Jsou vám nejasná některá ustanovení daňo
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 kvalifil
Jak ho vlastně pozná ?
Zahrnul jste mne mnoha profesemi .
Naše metodika a techniky vyšetřování moho
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 péči .
Zejména v Olomouci firma svými výrobky při
Restaurátorské licence umožňují práce v cel
Firemní výrobky zdobí galerie a soukromá síd
Balení másla ve hliníkové fólii zajišťuje jeho
Samotné měření spotřeby tepla peněženku
Platíme hodně hlavně díky nízké technické úi
Stavět vlastní výtopnu je prozatím patrně ri
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 divizíd
Jeho rozsah působnosti je úctyhodný :
Růst počtu teplem zásobovaných bytů přito
Počet pracovníků podniku zároveň klesl v pr
Podnik platil doposud za znečisťování ovzdu
Ročně vyprodukovaných 280 - 350 tisíc tun
Lidé prožívají nebývale nervózní dobu .
slovní tvar: Lidé
lemma: člověk
slovní druh: podstatné jméno
rod: mužský
číslo: množné
pád: první
dobu
Pt
nervózní
Pk
nebývale
Pu
Jsou vám nejasná některá ustanovení daňovýd
Je naprosto bezbřehý a nevypočitatelný .
V tom s vámi nesouhlasím .
Například podnikatel by chtěl dosáhnout u zamě
Lidé prožívají nebývale nervózní dobu .

Charon

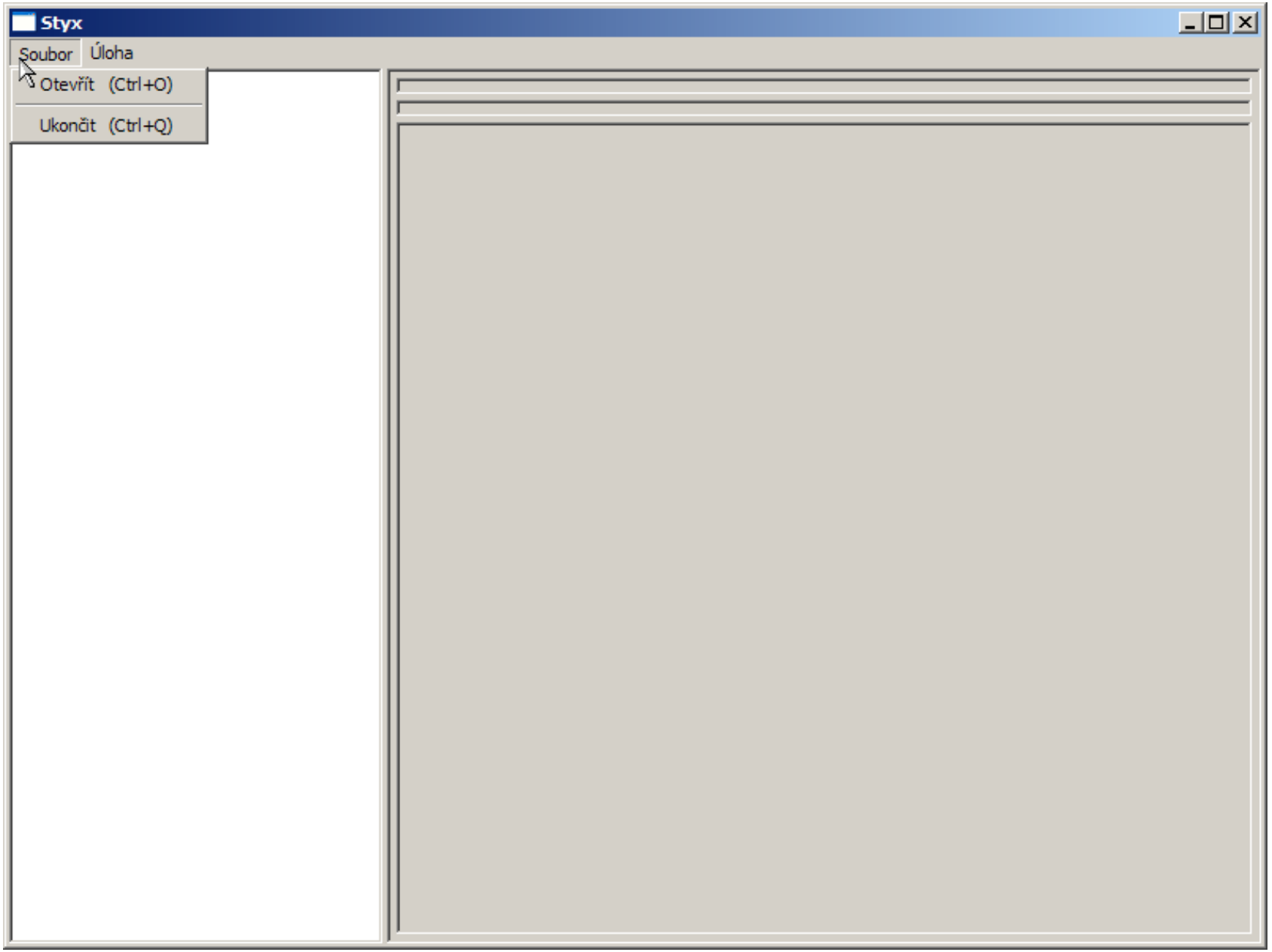

Charon
Soubor
Sada
Otevřít (Ctrl+O)
Uložit (Ctrl+S)
Uložit jako (Ctrl+Shift+S)
Ukončit (Ctrl+Q)
Za vozidla zaměstnanců užívaná pro pra
Stejný názor má i řada našich soukromý
Je naprosto bezbřehý a nevypočitatelný
V tom s vámi nesouhlasím .
Mzdy a platy jsou vždy podřízenými hod
Zde vzniká nová dimenze srovnávání :
Opět s vámi nesouhlasím .
Můžete to vysvětlit na příkladu ?
Například podnikatel by chtěl dosáhnout


Charon
Soubor
Sada
Sada 1
Sada 2
Sada 3
Sada 4
Sada 5
Sada 6
Sada 7
Sada 8
Sada 9
Sada 10
Můžete to vysvětlit na příkladu ?
Například podnikatel by chtěl dosáhnout

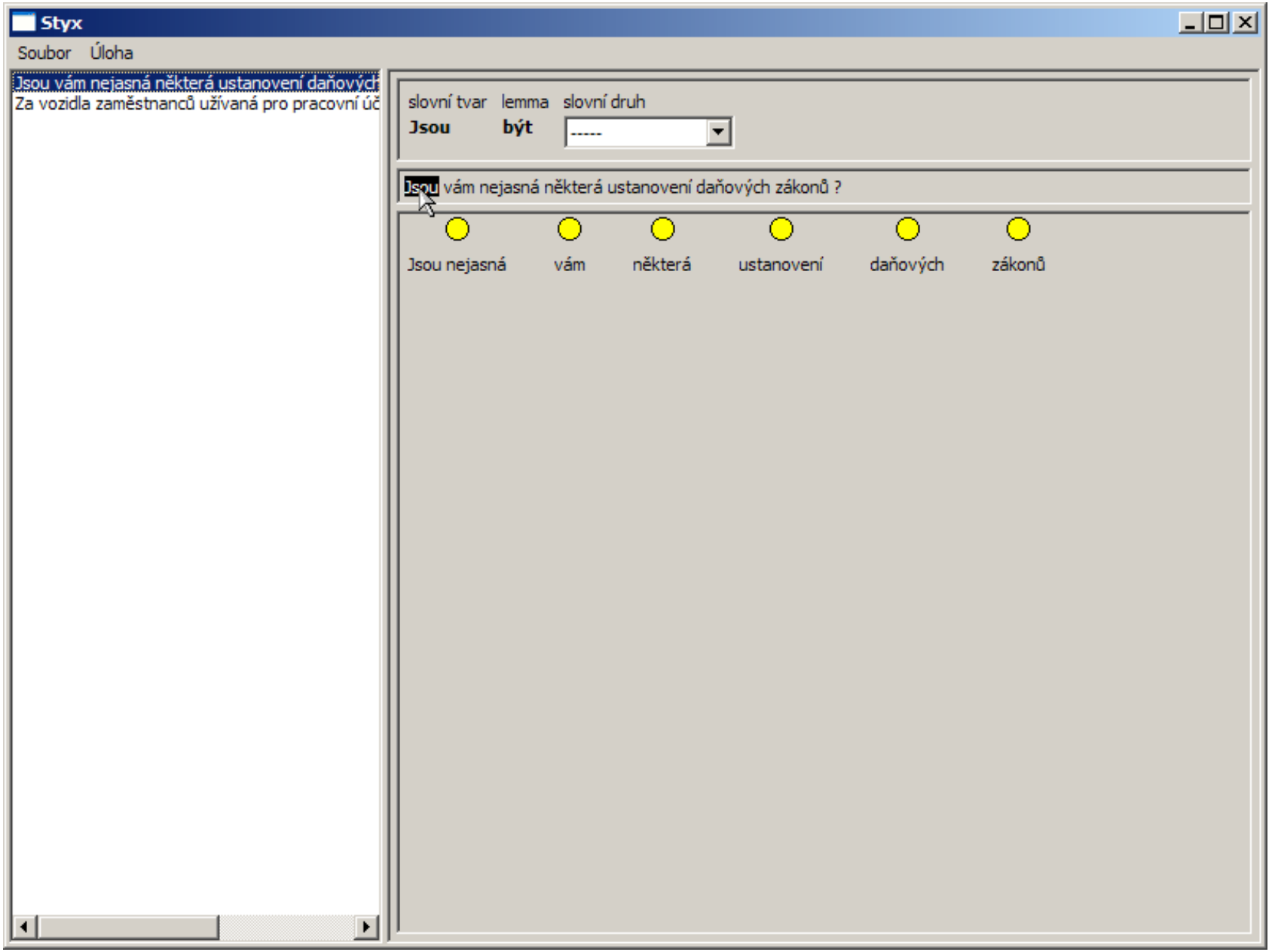
Styx

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

Styx

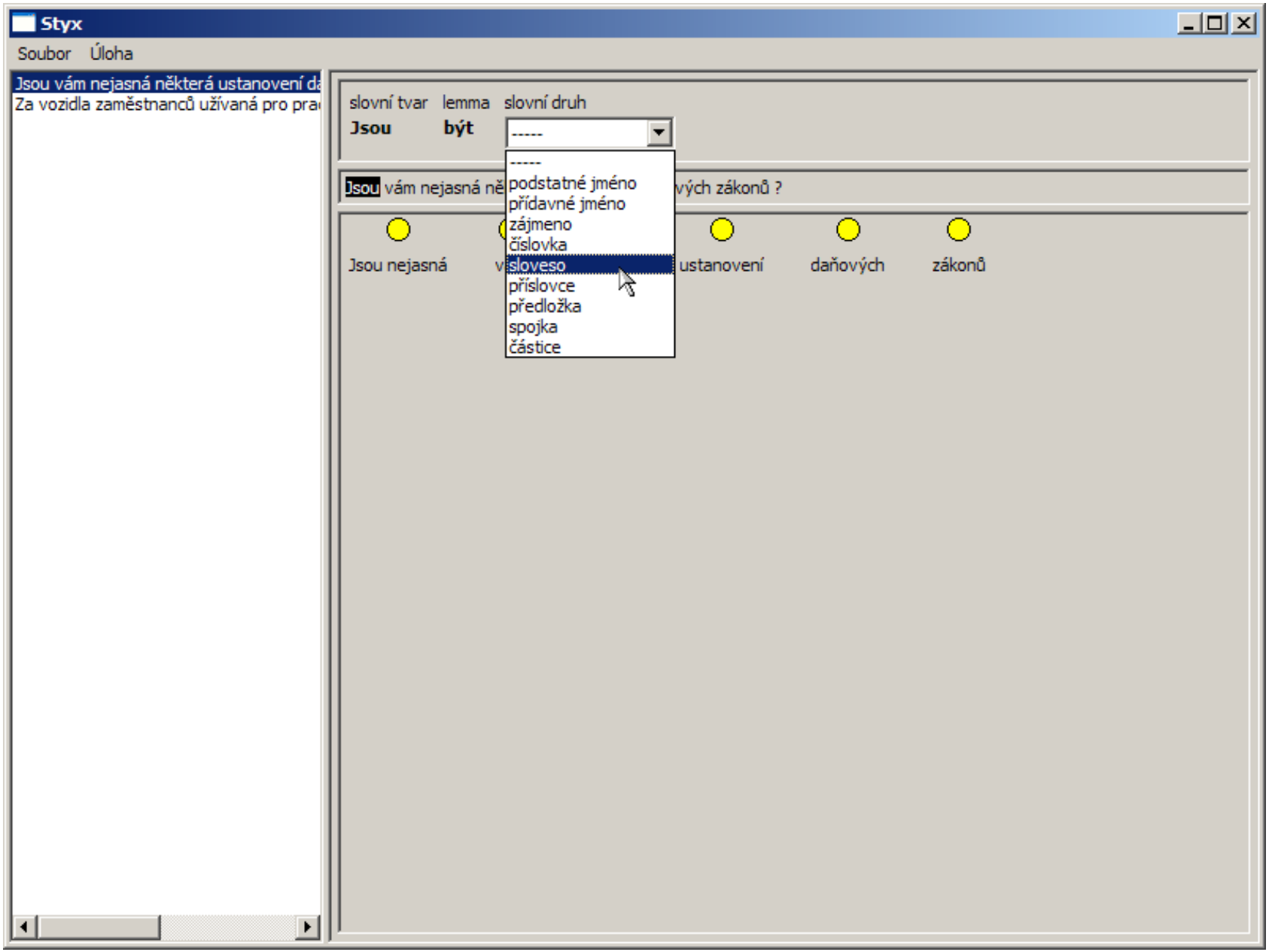


Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení daňových
Za vozidla zaměstnanců užívaná pro pracovní úč
slovní tvar
lemma
slovní druh
Jsou
být

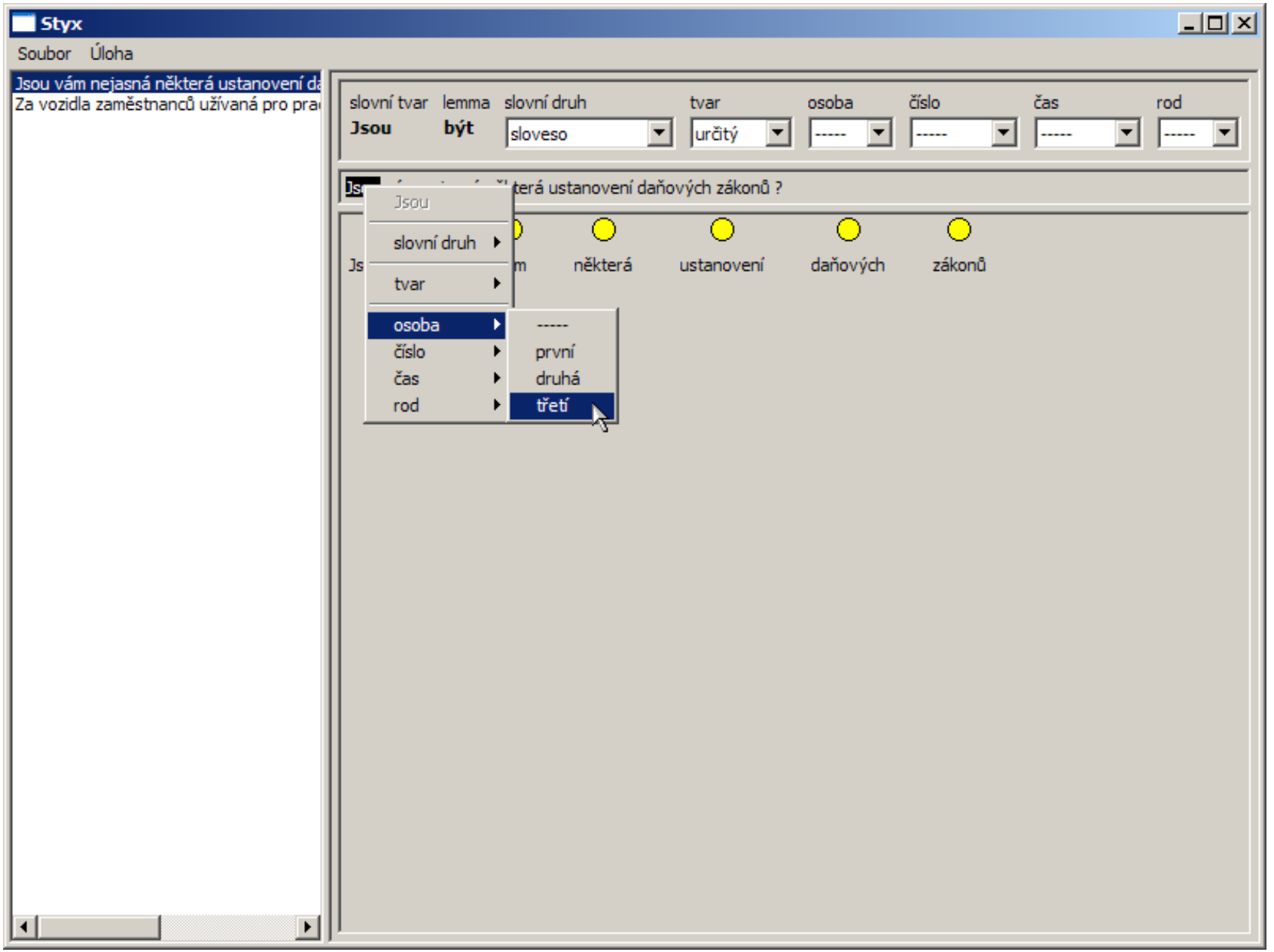
Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
vám
některá
ustanovení
daňových
zákonů

Styx

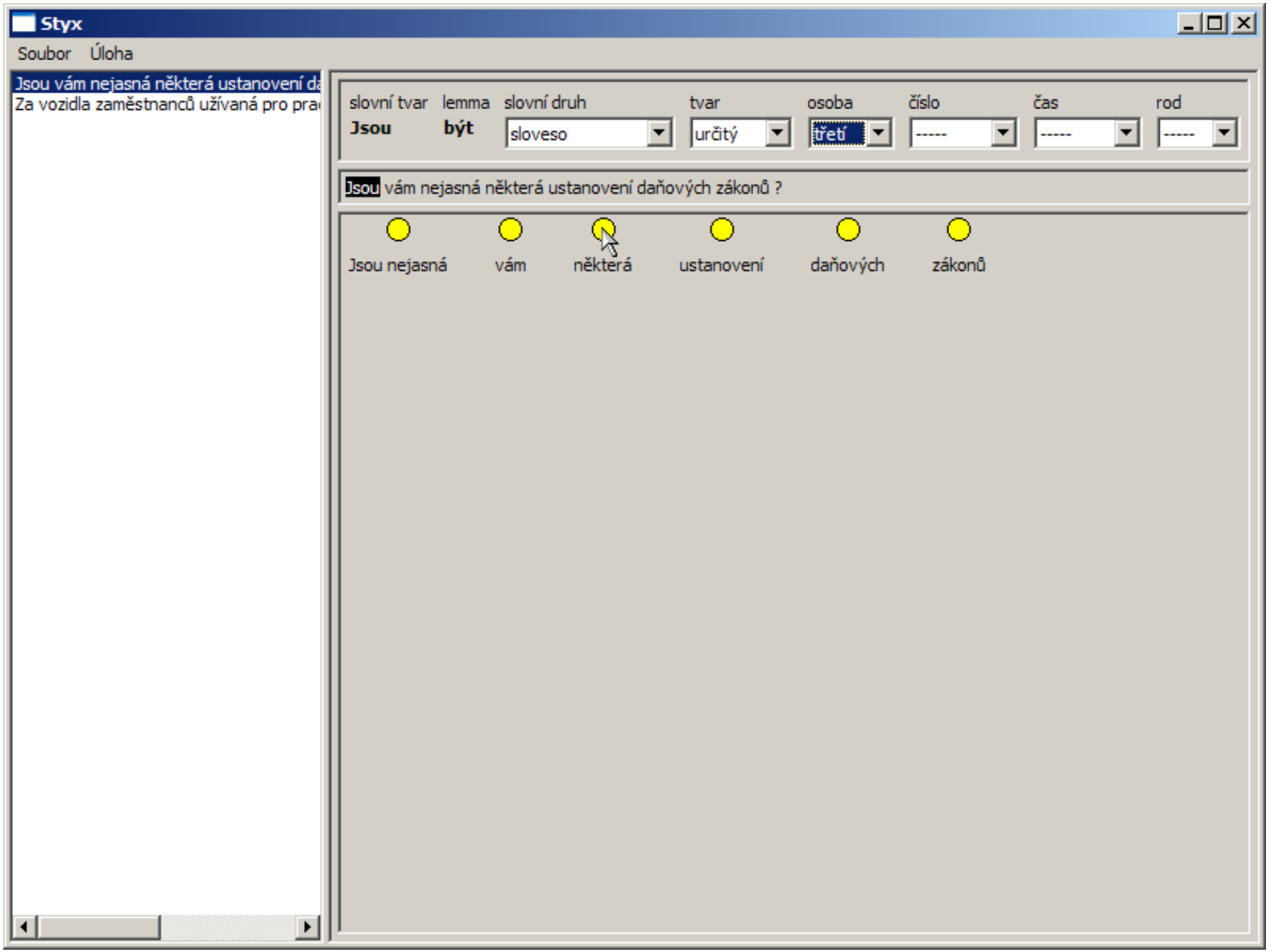

Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
lemma
slovní druh
Jsou
být

-----
podstatné jméno
přídavné jméno
zájmeno
číslovka
sloveso
příslovce
předložka
spojka
částice
Jsou vám nejasná ně
vých zákonů ?
Jsou nejasná
ustanovení
daňových
zákonů

Styx

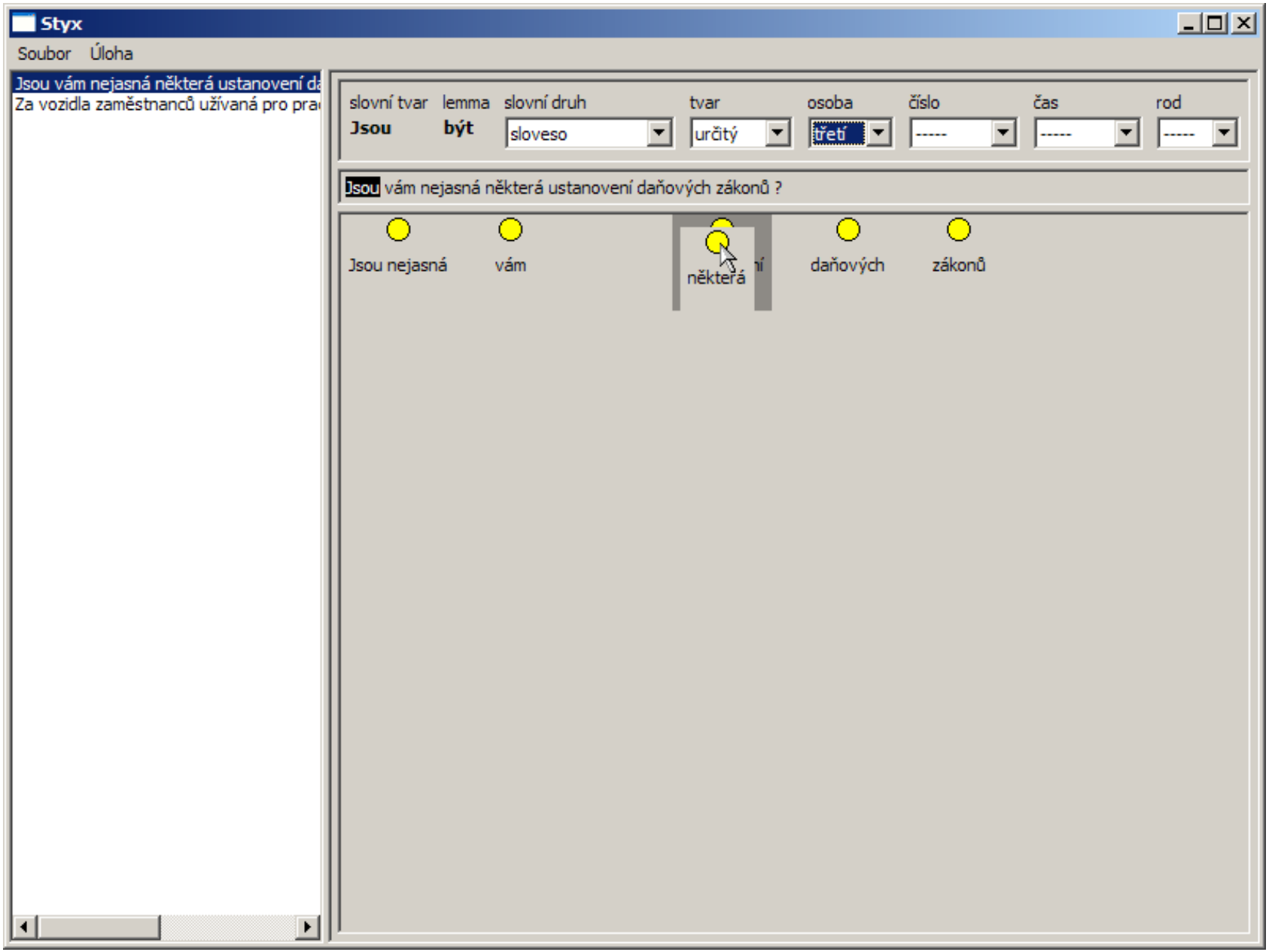

Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
Jsou
lemma
být
slovní druh
sloveso
tvar
určitý
osoba
číslo
čas
rod
Jsou
slovní druh
tvar
osoba
číslo
čas
rod
-----
první
druhá
třetí
některá
ustanovení
daňových
zákonů

Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
lemma
slovní druh
tvar
osoba
číslo
čas
rod
Jsou
být
sloveso
určitý
třetí

Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
vám
některá
ustanovení
daňových
zákonů

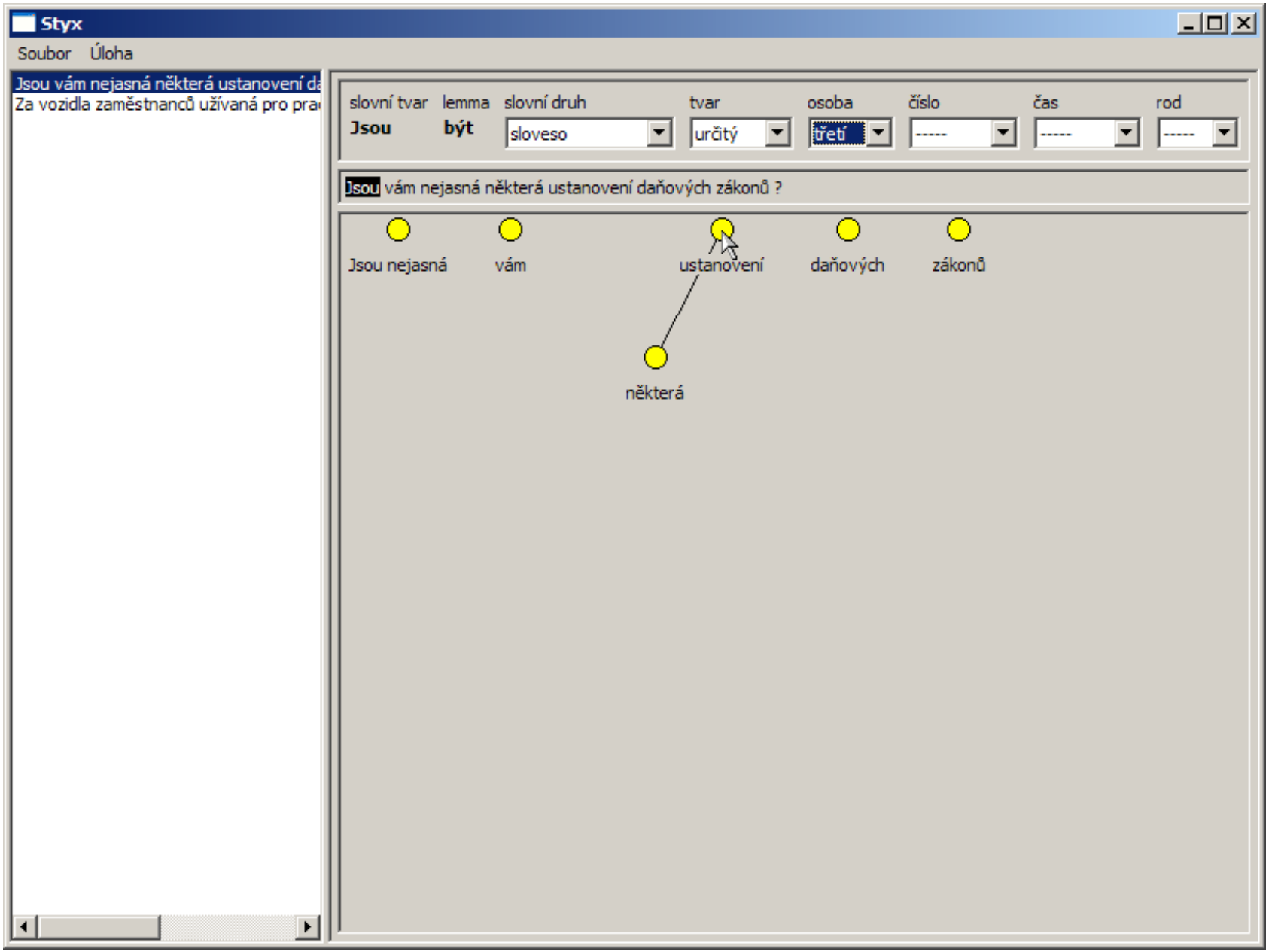
Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
Jsou
lemma
být
slovní druh
sloveso
tvar
určitý
osoba
třetí
číslo

čas
-----
rod

Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
vám
některá
daňových
zákonů

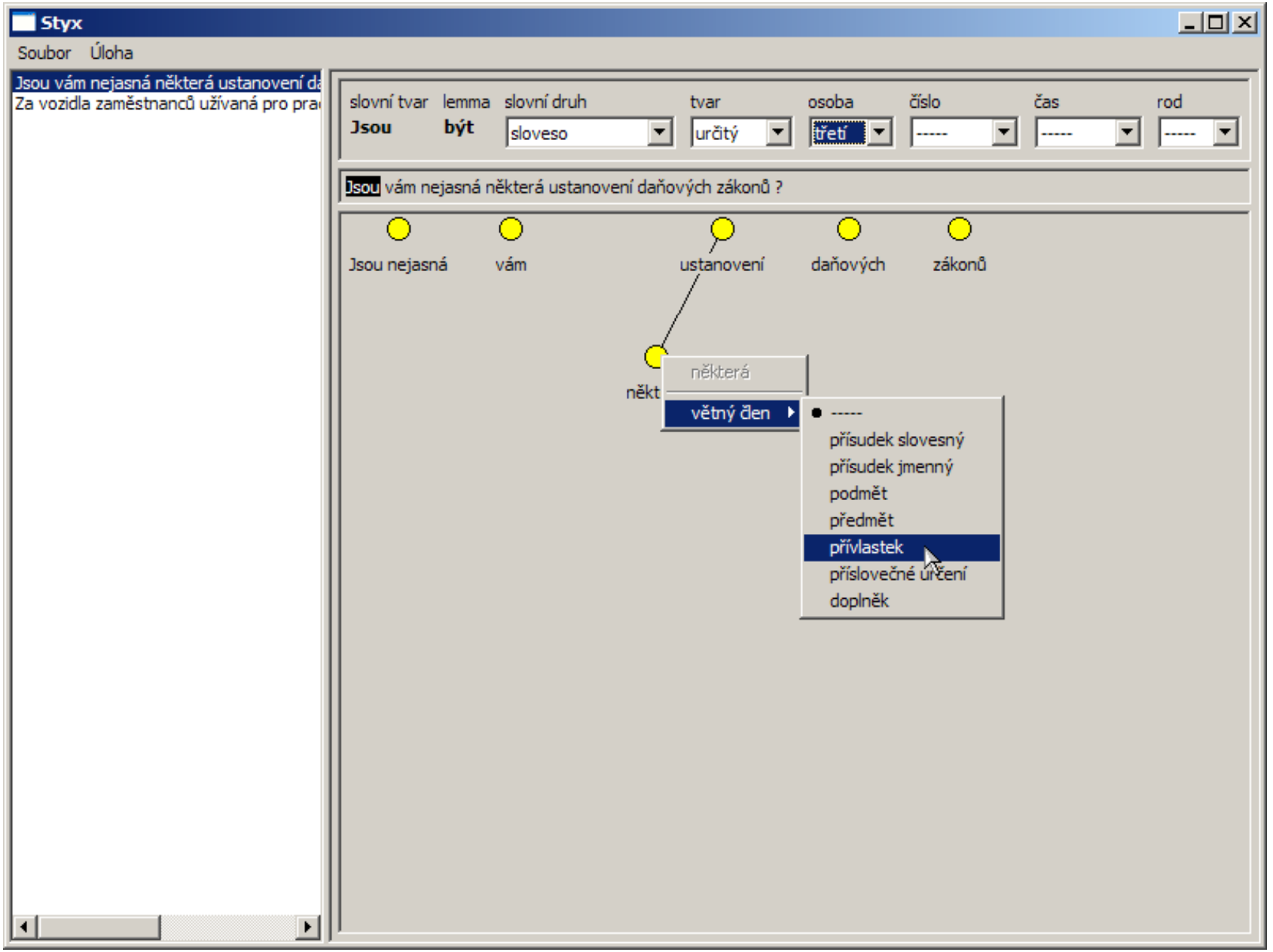
Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
Jsou
lemma
být
slovní druh
sloveso
tvar
určitý
osoba
třetí
číslo

čas
-----
rod

Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
vám
ustanovení
daňových
zákonů
některá

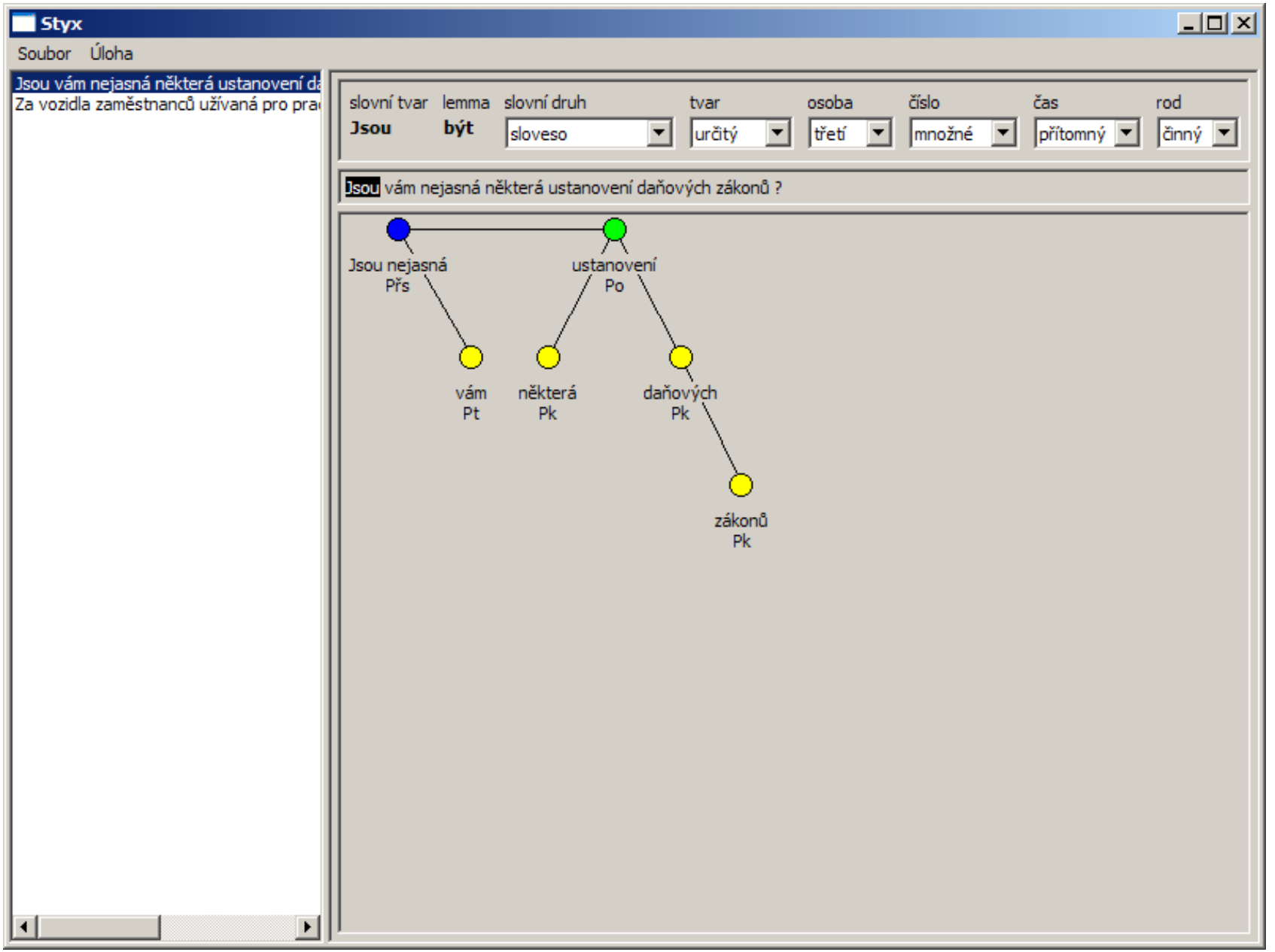
Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
Jsou
lemma
být
slovní druh
sloveso
tvar
určitý
osoba
třetí
číslo

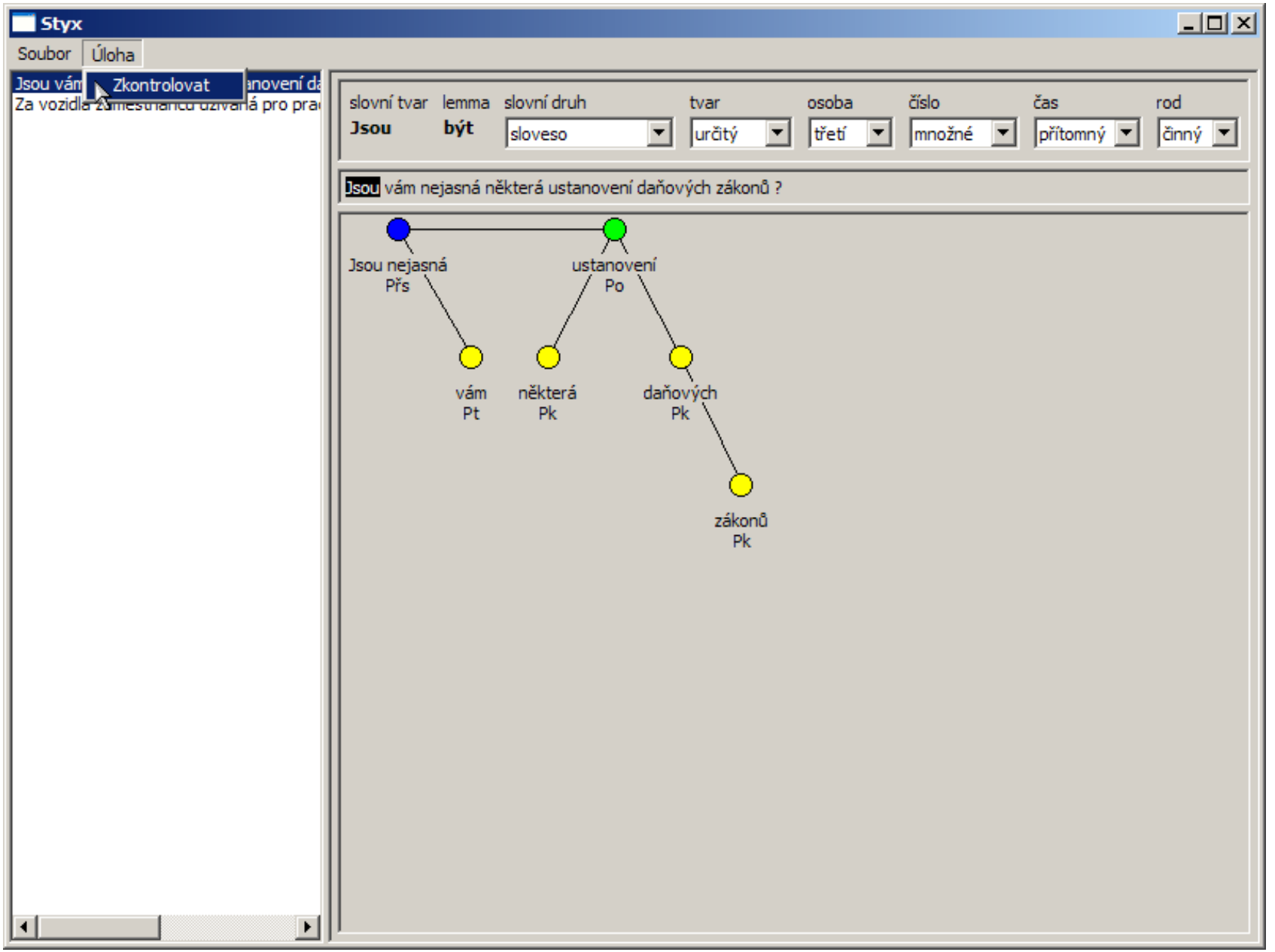
čas
-----
rod

Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
vám
ustanovení
daňových
zákonů
někt
některá
větný člen
-----
přísudek slovesný
přísudek jmenný
podmět
předmět
přívlastek
příslovečné určení
doplněk

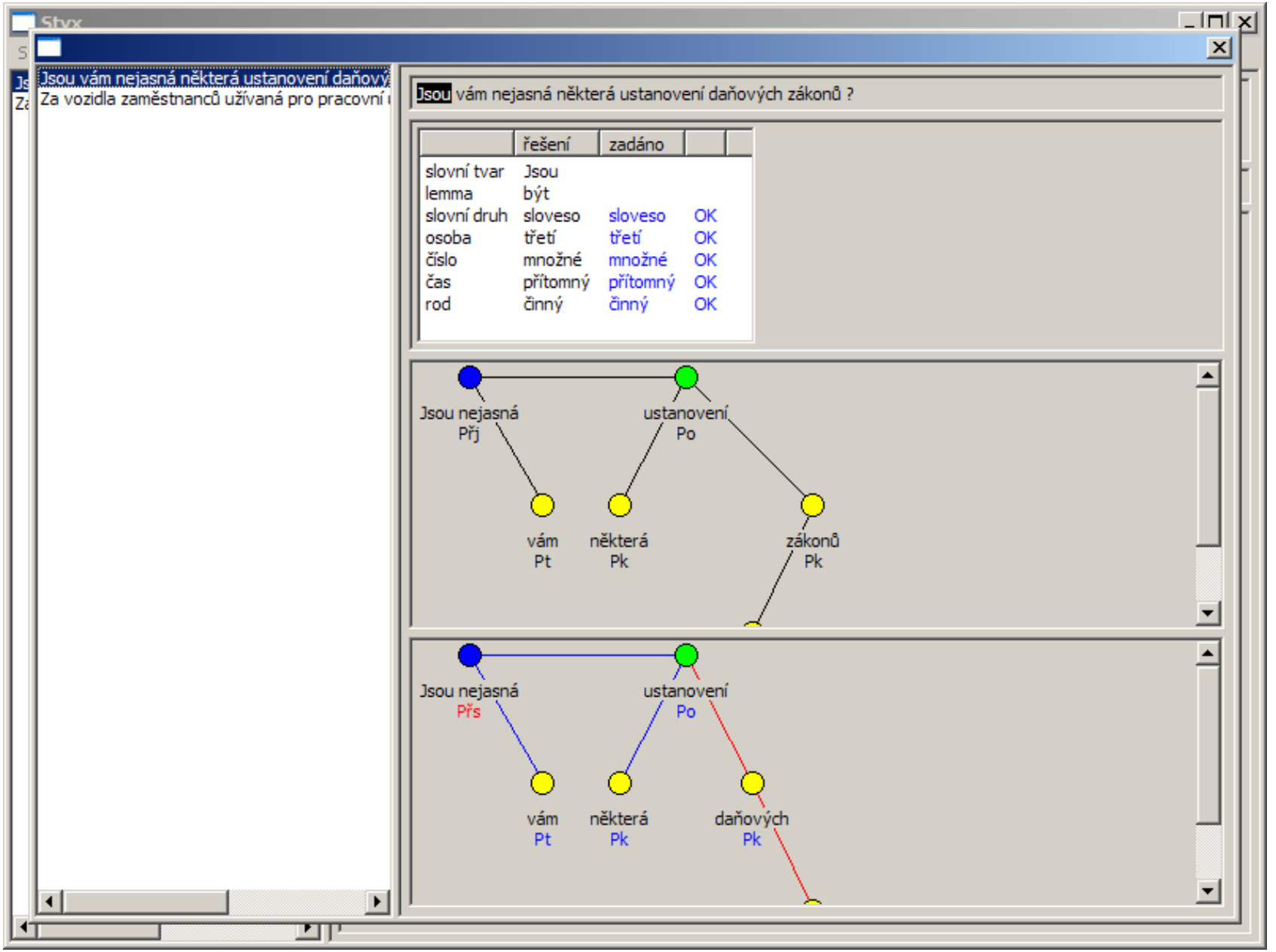
Styx


Styx
Soubor
Úloha
Jsou vám nejasná některá ustanovení da
Za vozidla zaměstnanců užívaná pro pra
slovní tvar
Jsou
lemma
být
slovní druh
sloveso
tvar
určitý
osoba
třetí
číslo
množné
čas
přítomný
rod
činný
Jsou vám nejasná některá ustanovení daňových zákonů ?
Jsou nejasná
Přs
ustanovení
Po
vám
Pt
některá
Pk
daňových
Pk
zákonů
Pk

Styx


Styx
Soubor
Úloha
Zkontrolovat
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
ustanovení Po
vám Pt
některá Pk
daňových Pk
zákonů Pk

Styx


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