STYX

Prague Dependency Treebank as an exercise book of Czech

Ondřej Kučera

Contents

1. **Introduction** (motivation, PDT, implementation)
2. **Filtering sentences**
3. **Transformations of trees**
4. **STYX:** FilterSentences, Charon, Styx

Motivation

- children of today use computers regularly
  - games, web surfing, chatting, writing, drawing
- why couldn't they parse sentences or determine parts of speech?

Building an exercise book

Manually

- extremely hard
  - choose (make up) the sentences
  - annotate them
- considerably limited number of sentences
- often too simple sentences not reflecting the real usage of the language

Building an exercise book

Automatically

- if we have annotated data
- the work of choosing the sentences and annotating them is already done
- the data in corpus reflect the real usage of the language
- the number of sentences corresponds to the size of the corpus
- PDT

Prague Dependency Treebank

- annotated on four layers (word, morphological, analytical, tectogrammatical)
- inner data format: PML (Prague Markup Language) – based on XML

PDT vs. school syntax

- annotation rules of PDT allow to process any sentence
  ⇨ filtering sentences
- Analytical layer of PDT differs from the school syntax in many ways
  ⇨ transformations of analytical trees

Filtering sentences

Filtering in numbers

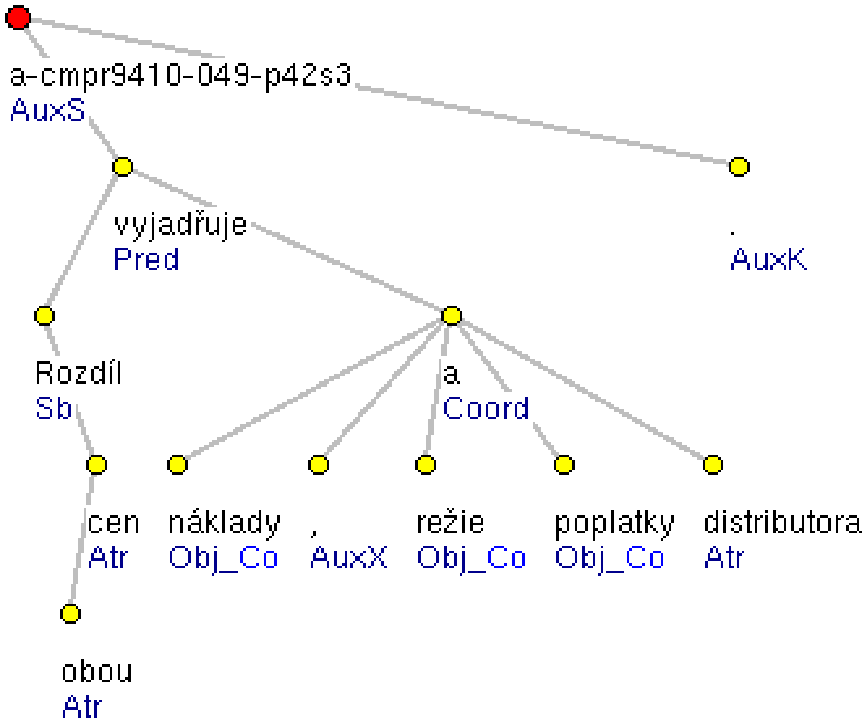
- nine different filters
- starting number of sentences: 49,442
- after application of the filters: 11,705
- about 23.7% sentences kept

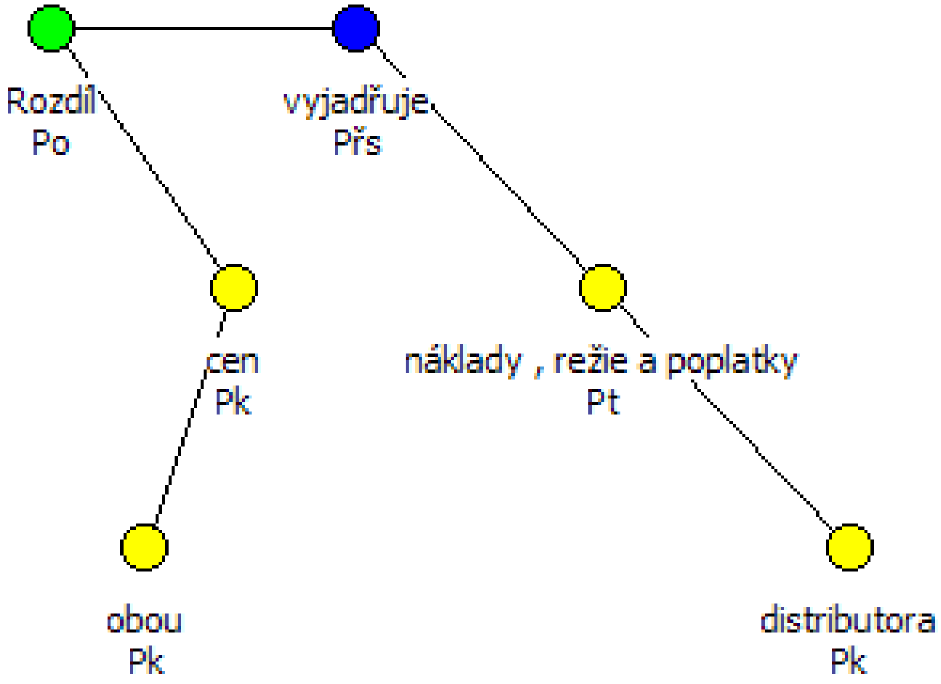
Transformation of trees

- three basic transformations
- particular transformations consist of
  - combining of the three basic transformations
  - rules for modification of the syntactic functions

Transformations of trees

Example





Implementation

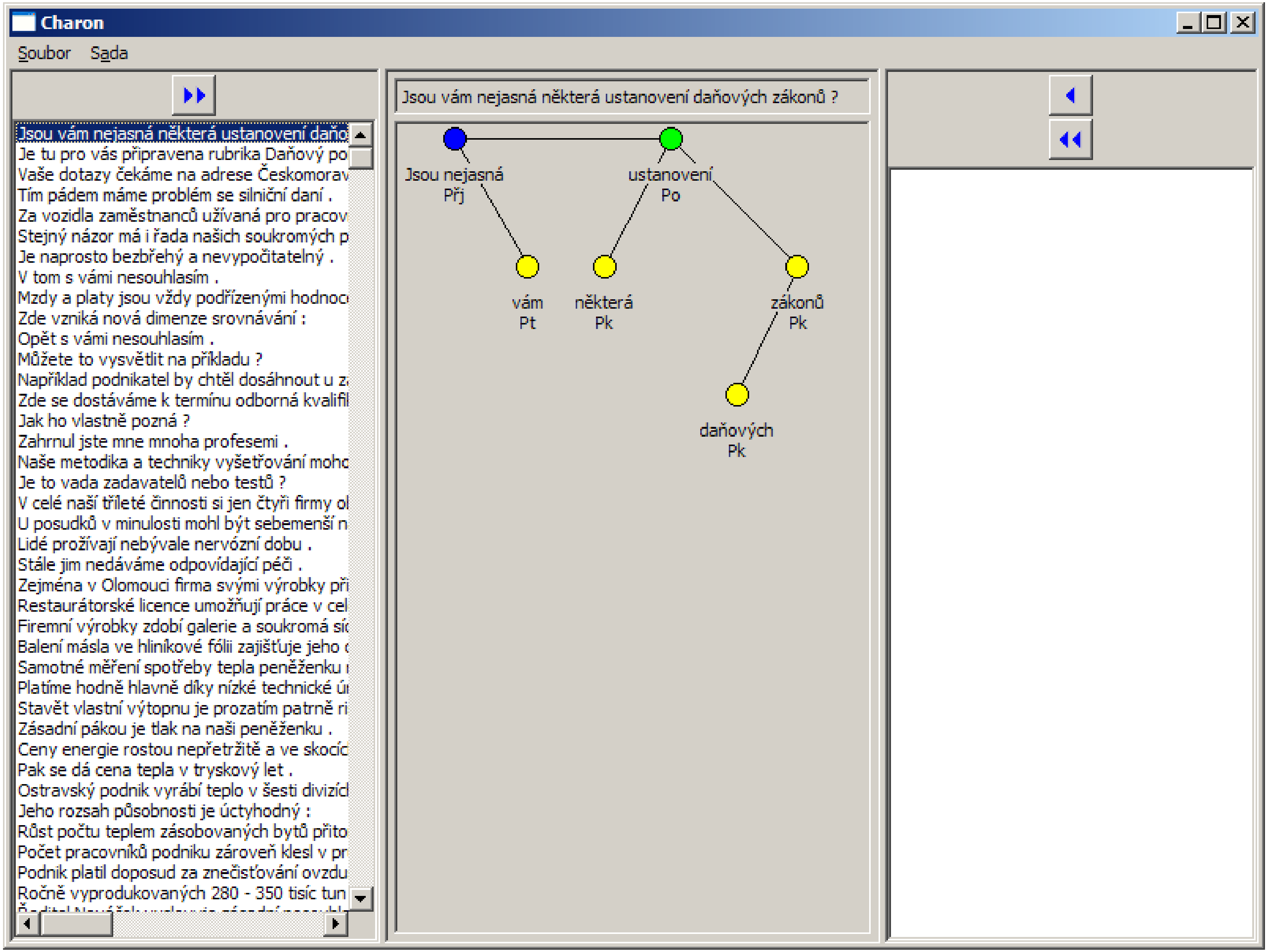
Java

- high-level language with number of mechanisms protecting programmers “against themselves”
- portability
- presence of SWT library

SWT

- Standard Widget Toolkit
- provides native look and feel of graphical user interface
- speed

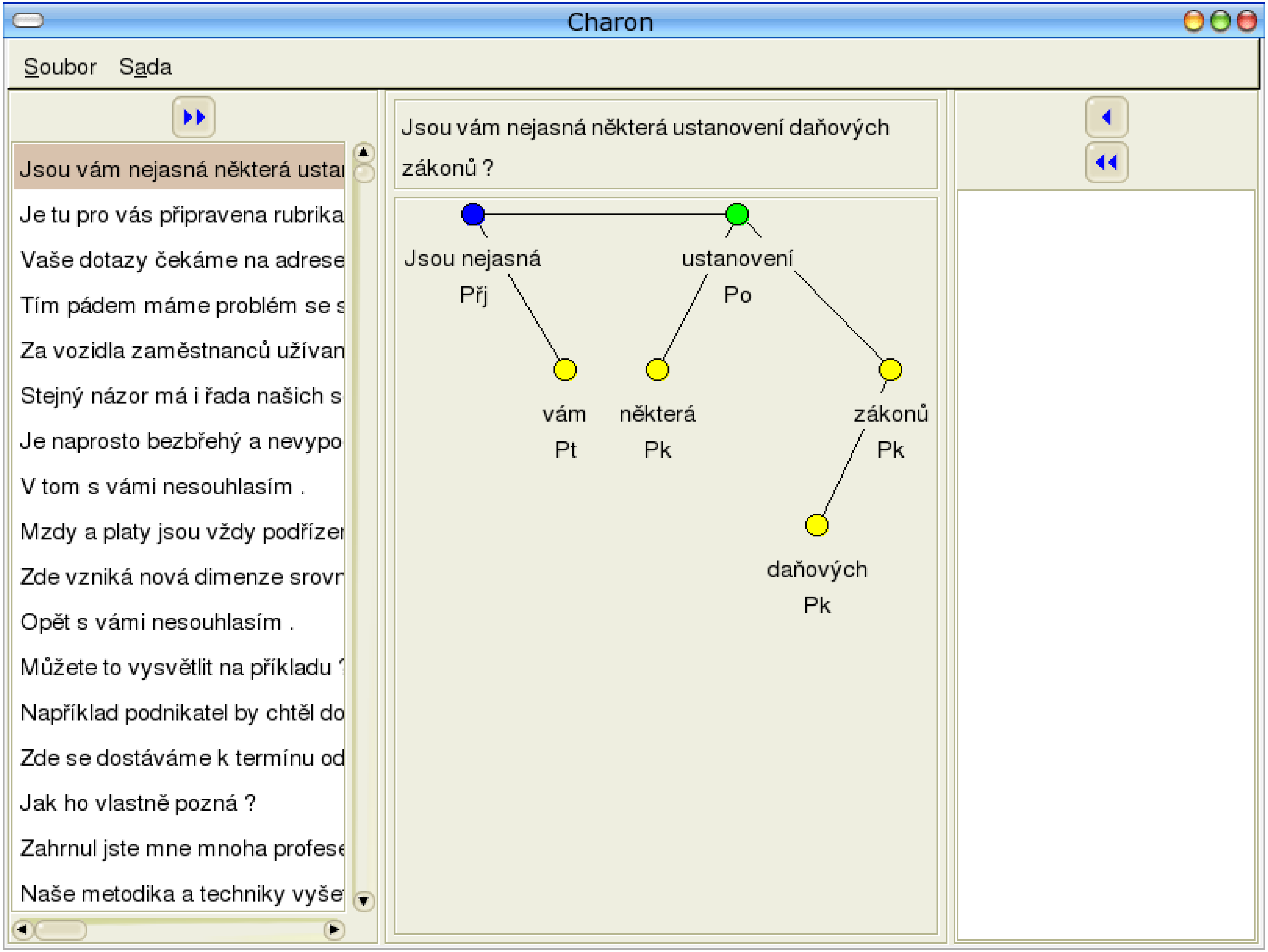
Implementation



Jsou vám nejasná některá ustanovení daňových zákonů ?

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 moho
Je to vada zadavatelů nebo testů ?
V celé naší tříleté činnosti si jen čtyři firmy o
U posudků v minulosti mohl být sebemenší n
Lidé prožívají nebývale nervózní dobu .
Stále jim nedáváme odpovídající péči .
Zejména v Olomouci firma svými výrobky př
Restaurátorské licence umožňují práce v ce
Firemní výrobky zdobí galerie a soukromá s
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 r
Zásadní pákou je tlak na naši peněženku .
Ceny energie rostou nepřetržitě a ve skoc
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 p
Podnik platil doposud za znečisťování ovzdu
Ročně vyprodukovaných 280 - 350 tisíc tun

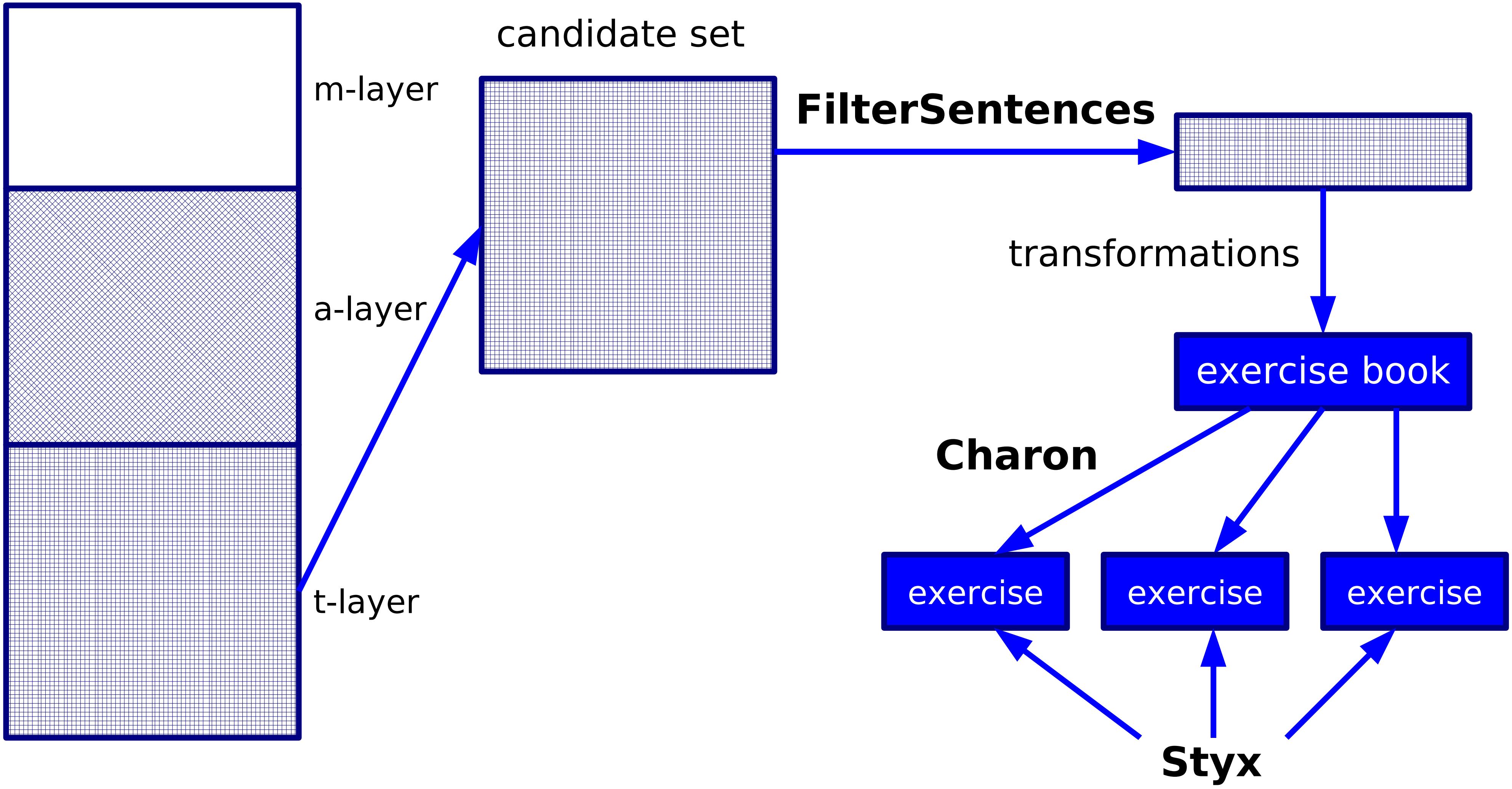
Implementation



Jsou vám nejasná některá ustai
Je tu pro vás připravena rubrika
Vaše dotazy čekáme na adrese
Tím pádem máme problém se s
Za vozidla zaměstnanců užívan
Stejný názor má i řada našich s
Je naprosto bezbřehý a nevypo
V tom s vámi nesouhlasím .
Mzdy a platy jsou vždy podřízer
Zde vzniká nová dimenze srovr
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 profese
Naše metodika a techniky vyše

Jsou vám nejasná některá ustanovení daňových zákonů ?

STYX:
FilterSentences, Charon, Styx



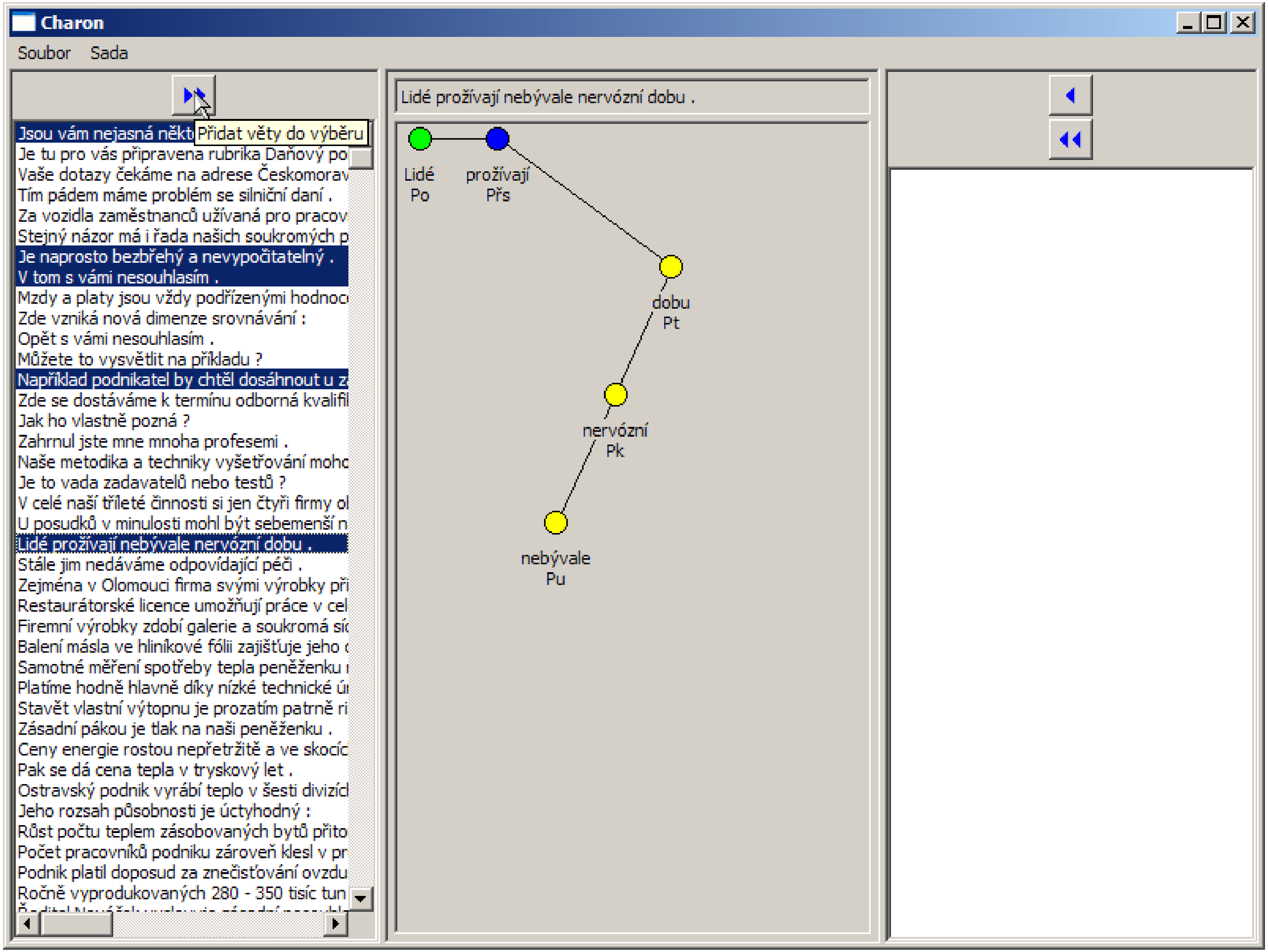
FilterSentences

- used for applying the filters
- reads data in PML format
- each sentence is tested by a filter
- output data contains the sentences that the filter kept
- output again in PML format

Charon

- “administrative” program
- loads all sentences available
- the user selects sentences that he or she wants to have in the exercise
- in the end the user saves the exercise

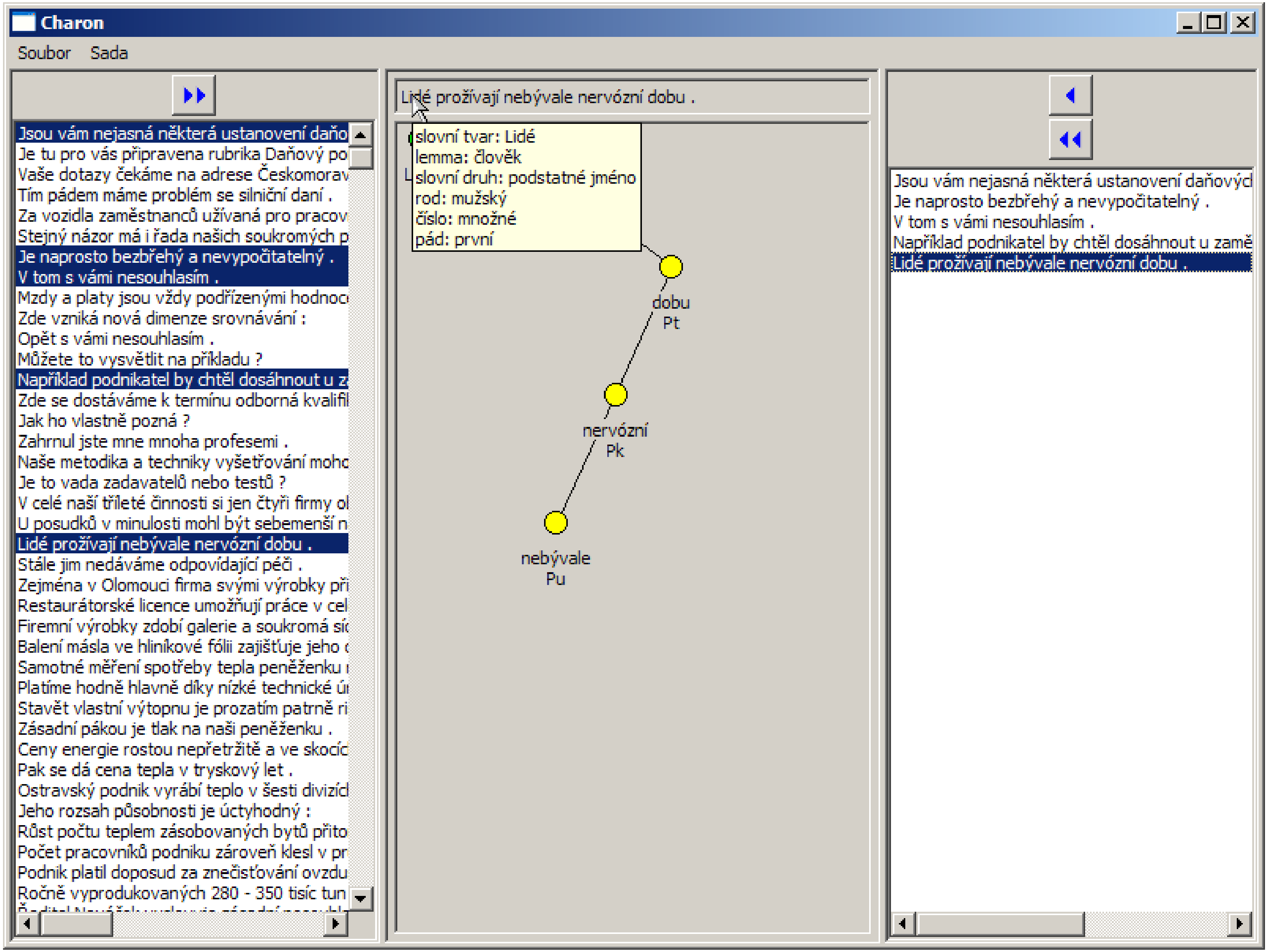
Charon



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 mohc
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 pří
Restaurátorské licence umožňují práce v cel
Firemní výrobky zdobí galerie a soukromá síc
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 divizícl
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 vyprodukovaných 280 - 350 tisíc tun

Lidé prožívají nebývale nervózní dobu .

Charon



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 moho
Je to vada zadavatelů nebo testů ?
V celé naší tříleté činnosti si jen čtyři firmy o
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

Lidé prožívají nebývale nervózní dobu .

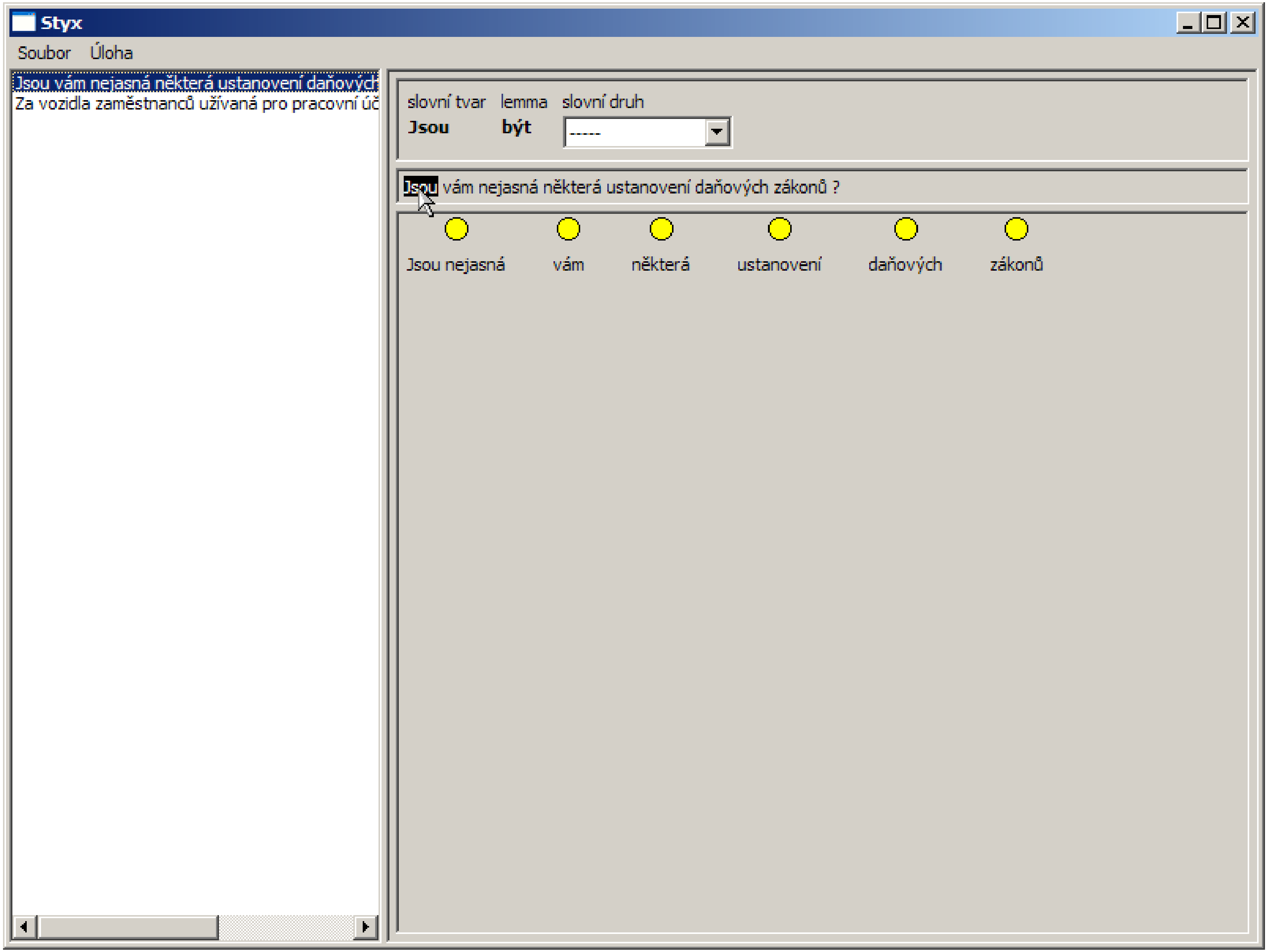
slovní tvar: Lidé
lemma: člověk
slovní druh: podstatné jméno
rod: mužský
číslo: množné
pád: první

Jsou vám nejasná některá ustanovení daňovýc
Je naprosto bezbřehý a nevypočitatelný .
V tom s vámi nesouhlasím .
Například podnikatel by chtěl dosáhnout u zamě
Lidé prožívají nebývale nervózní dobu .

Styx

- exercise book itself
- user loads an exercise previously created and saved in Charon

Styx

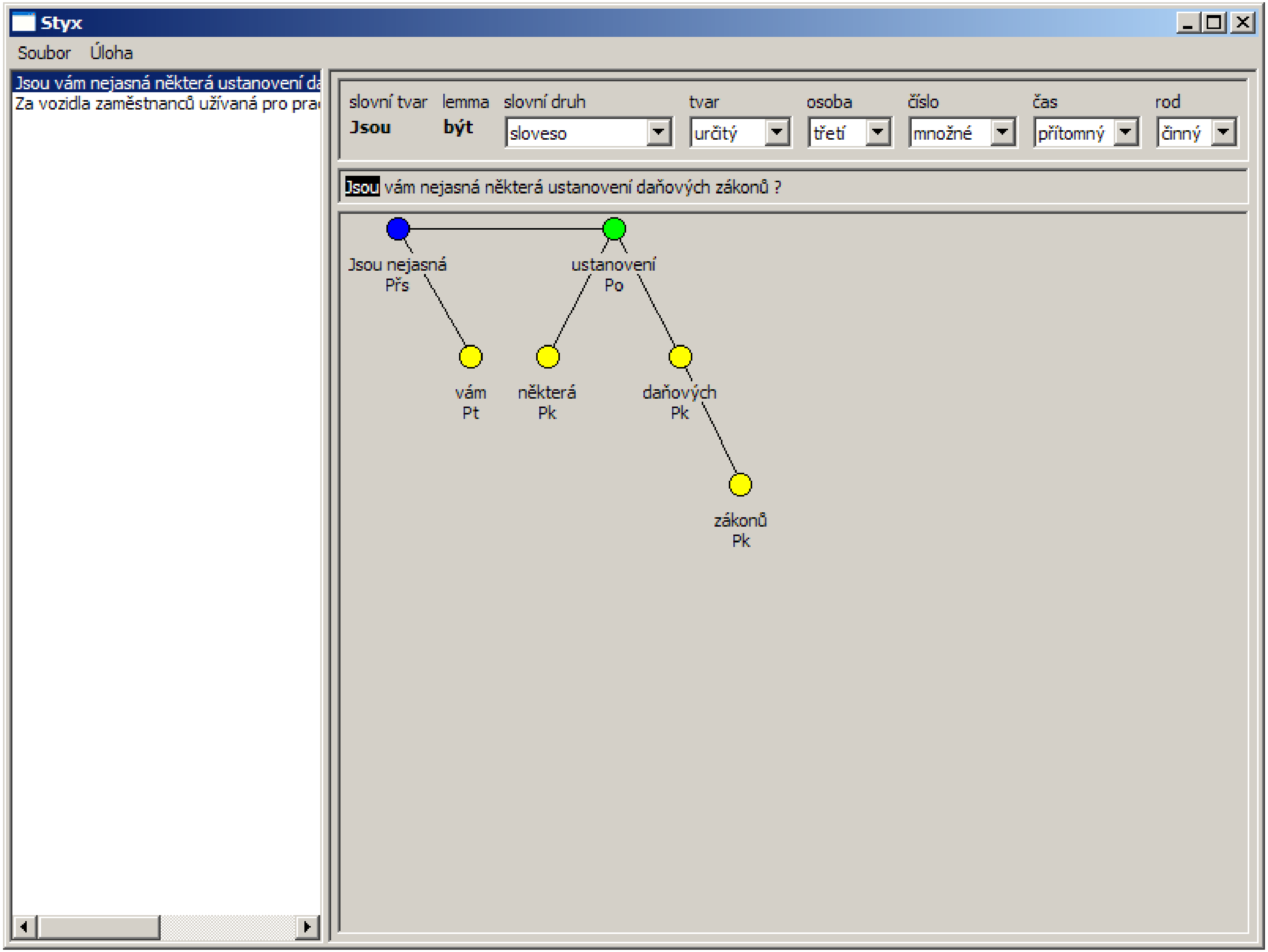


Jsou vám nejasná některá ustanovení daňových
Za vozidla zaměstnanců užívaná pro pracovní úč

slovní tvar	lemma	slovní druh
Jsou	**být**	-----

Jsou vám nejasná některá ustanovení daňových zákonů ?

Styx



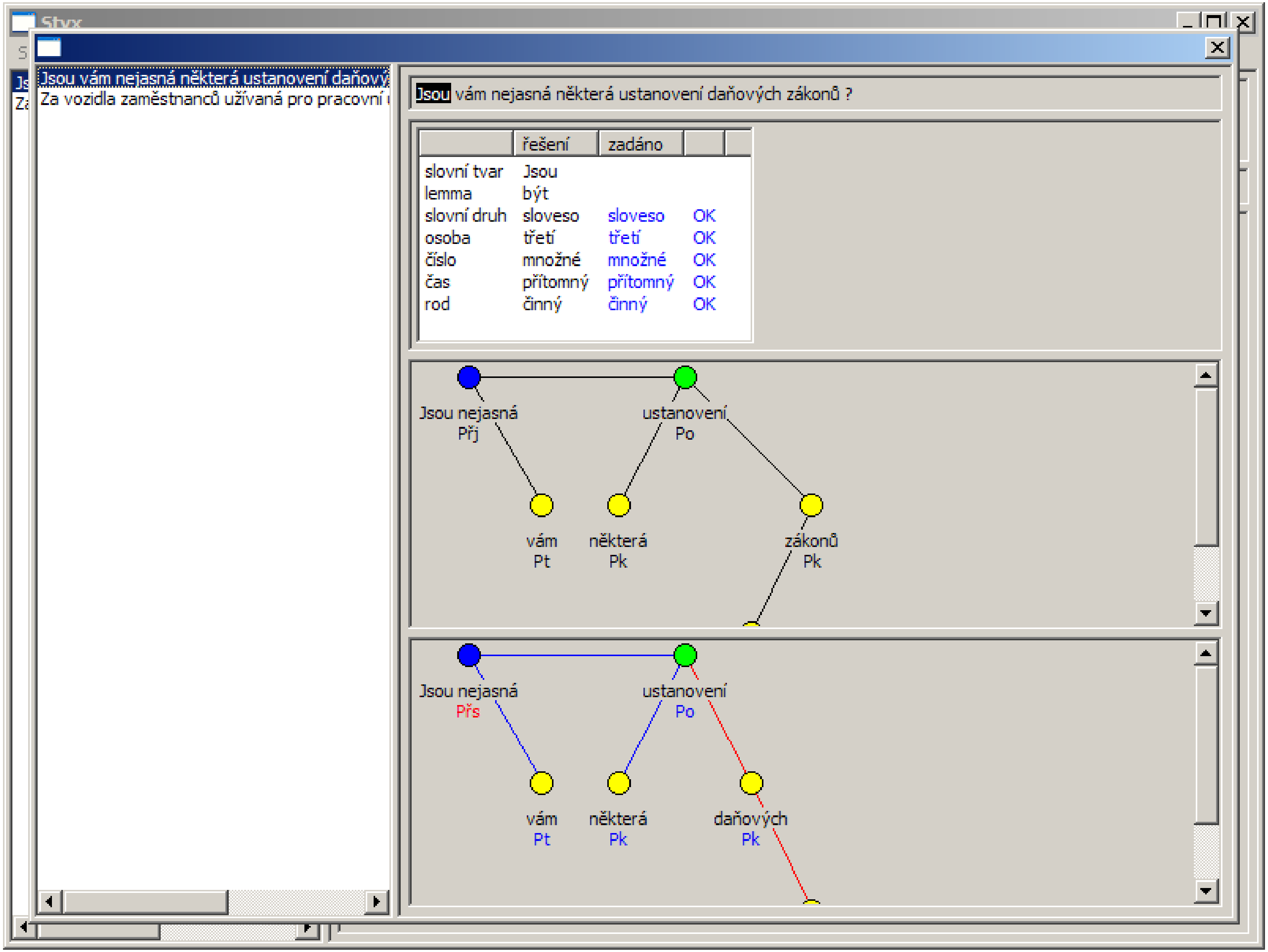
Jsou vám nejasná některá ustanovení da

Za vozidla zaměstnanců užívaná pro pra

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

Jsou vám nejasná některá ustanovení daňových zákonů ?

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	

Questions

and perhaps some answers...