

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

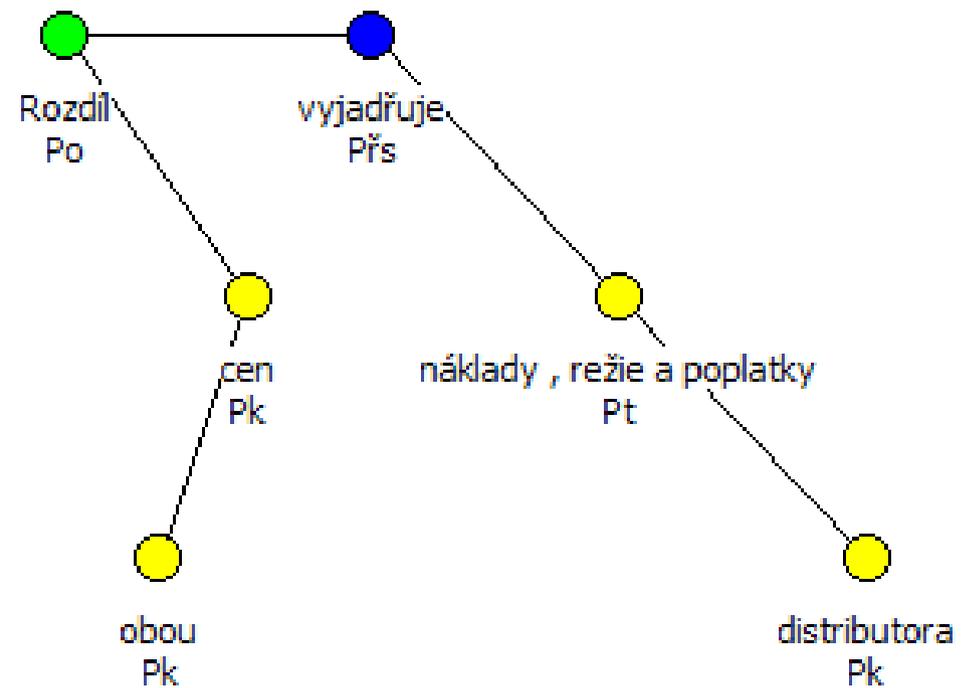
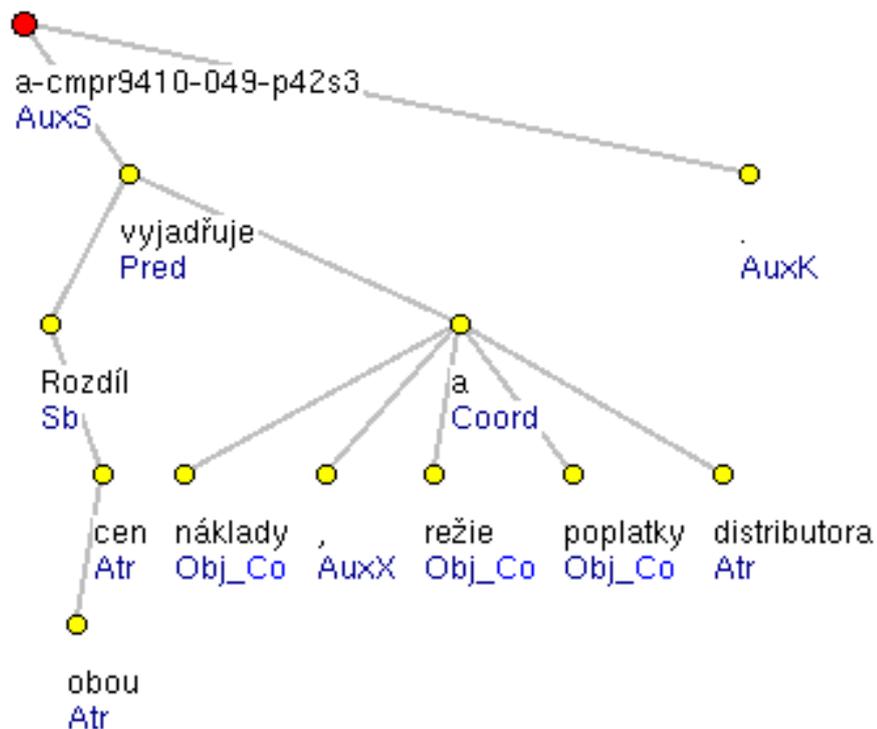
- nine different filters
- starting number of sentences: 49,442
- after application of the filters: 11,705
- about 23.7% sentences kept

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

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 243, 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. Zde se dostáváme k termínu odborná kvalifikace. Jak ho vlastně pozná? Zahrnul jste mne mnoha profesemi. Naše metodika a techniky vyšetřování mohou být. Je to 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. Firemní výrobky zdobí galerie a soukromá sídla. Balení másla ve hliníkové fólii zajišťuje jeho ochrannou. Samotné měření spotřeby tepla peněženku i. Platíme hodně hlavně díky nízké technické úrovni. Stavět vlastní výtopnu je prozatím patrně neúnosné. 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. Podnik platil doposud za znečišťování ovzduchu. Ročně vyprodukoval 280 - 350 tisíc tun oxidu uhličitého.

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

Implementation

Charon

Soubor Sada

Jsou vám nejasná některá ustanovení daňových zákonů ?

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

Jsou nejasná
Přj

ustanovení
Po

vám
Pt

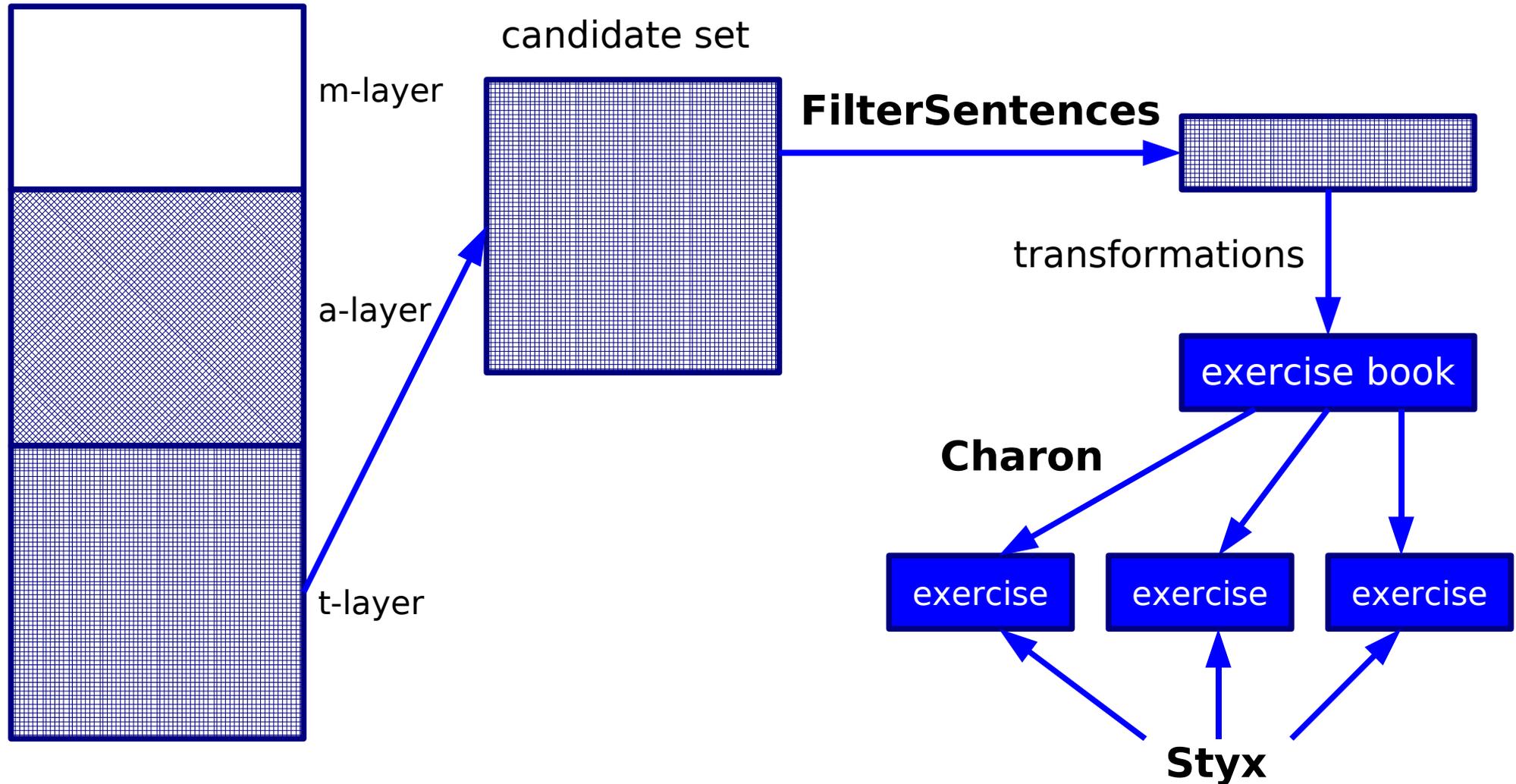
některá
Pk

zákonů
Pk

daňových
Pk

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

The screenshot shows the Charon application window. On the left, a text document is open with several lines highlighted in blue. A mouse cursor is over a button labeled "Přidat věty do výběru". On the right, a diagram titled "Lidé prožívají nebývale nervózní dobu ." shows a network of nodes and edges. The nodes are: "Lidé" (green circle), "prožívají" (blue circle), "dobu" (yellow circle), "nervózní" (yellow circle), and "nebývale" (yellow circle). The edges connect "Lidé" to "prožívají", "prožívají" to "dobu", "dobu" to "nervózní", and "nervózní" to "nebývale". Below the diagram, the words "Po", "Přs", "Pt", "Pk", and "Pu" are listed, corresponding to the nodes.

Charon

Soubor Sada

Přidat věty do výběru

Jsou vám nejasná někte...
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 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 péč .
Zejména v Olomouci firma svými výrobky při...
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 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 vyprodukovanych 280 - 350 tisíc tun...
č...
č...
č...

Lidé prožívají nebývale nervózní dobu .

Lidé prožívají dobu nervózní nebývale

Po Přs Pt Pk Pu

Charon

The screenshot shows the Charon software interface. The main window displays a text document with several paragraphs. A yellow box highlights the sentence "Lidé prožívají nebývale nervózní dobu .". A morphological analysis diagram is overlaid on this sentence, showing the word "lidé" at the top, connected to "dobu" (Pt), "nervózní" (Pk), and "nebývale" (Pu). A tooltip for "lidé" provides the following information:

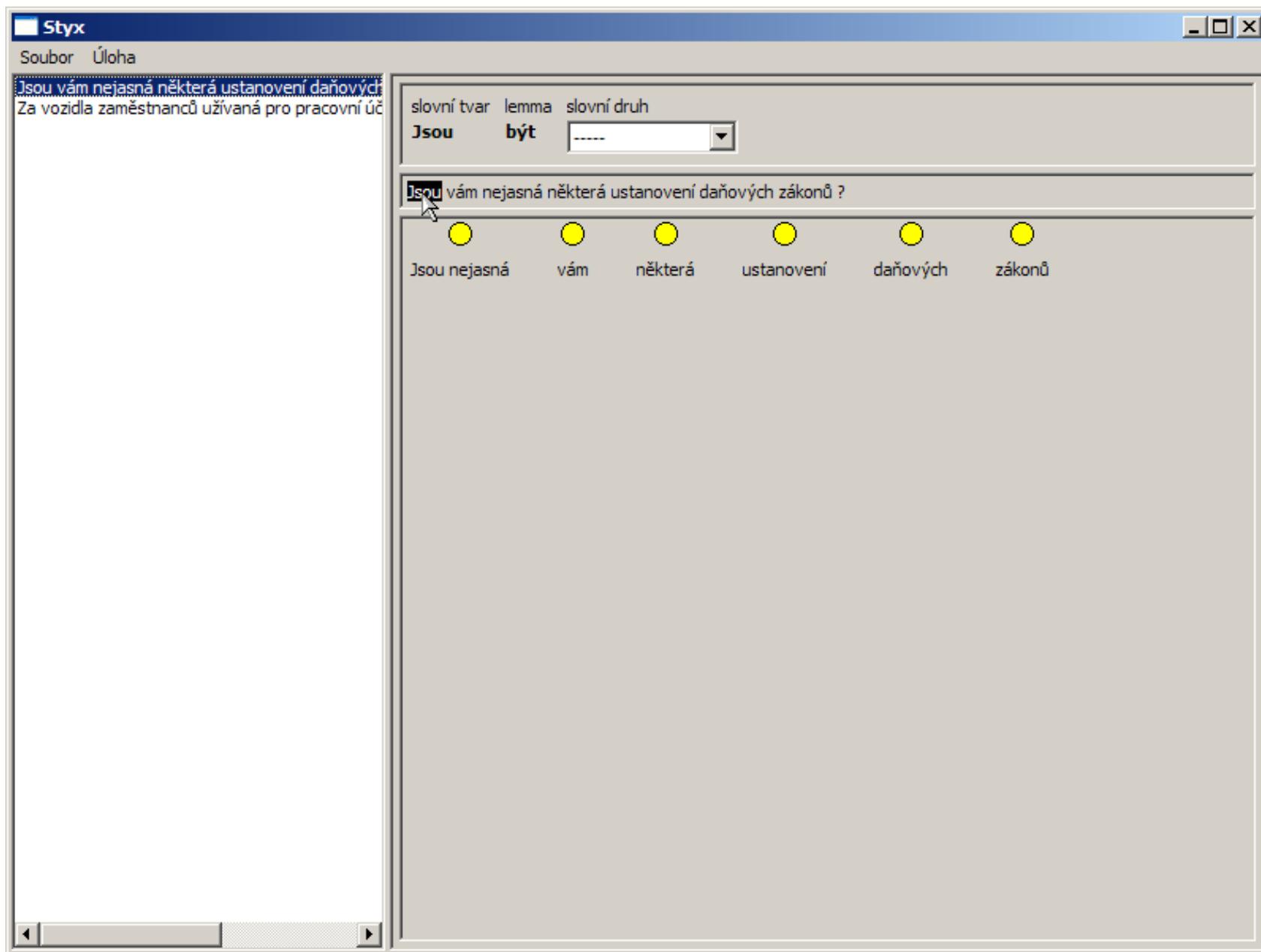
- slovní tvar: Lidé
- lemma: člověk
- slovní druh: podstatné jméno
- rod: mužský
- číslo: množné
- pád: první

The interface also includes a menu bar with "Soubor" and "Sada", and a status bar at the bottom.

Styx

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

Styx



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. On the left, a text editor contains the sentence: "Jsou vám nejasná některá ustanovení daňových zákonů. Za vozidla zaměstnanců užívaná pro pracovní". The main window displays the sentence "Jsou vám nejasná některá ustanovení daňových zákonů ?" and a table of morphological features for the word "Jsou".

	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

Below the table are two parse trees. The top tree shows the initial parse of "Jsou" (Přj) and "ustanovení" (Po) with their children "vám" (Pt), "některá" (Pk), and "zákonů" (Pk). The bottom tree shows the updated parse after the word "daňových" (Pk) has been added to the tree, with "Jsou" (Přs) and "ustanovení" (Po) as parents of "vám" (Pt), "některá" (Pk), and "daňových" (Pk).

Questions

and perhaps some answers...