Hrozí, že čeština zmizí z počítačů

Hospodářské noviny, 27.9.2011, Rubrika: Česko, Strana: 4, Autor: Zuzana Keményová

věda a výzkum Bílá kniha češtiny

Výzkum české lingvistiky pro počítače je na samém začátku Programů, které by rozuměly češtině, je zatím velmi málo V Česku vyšla první Bílá kniha češtiny - zásadní dokument, který mapuje minulý i budoucí vývoj mateřštiny Čechů ve vztahu k moderním technologiím. Ironií je, že kdo umí jen česky, příliš si v ní nepočte. Bílá kniha češtiny zatím existuje jen v angličtině, a to navíc v odborné.

„Potřebovali jsme ji představit před Evropskou komisí a na český překlad ještě nebyl čas. Ale pracujeme na něm, bude hotový do dvou měsíců," ujišťuje Jan Hajič, profesor Ústavu formální a aplikované lingvistiky při Matematicko-fyzikální fakultě Univerzity Karlovy, který se svým týmem studii připravil.

Anglicky psaná Bílá kniha češtiny je nechtěným důkazem svých vlastních zjištění. Práce, která dokumentuje, jak se český jazyk vyrovnal s nástupem počítačů a internetu v posledních dvaceti letech, totiž dospívá k závěru, že čeština je v ohrožení. Alespoň v útrobách počítačových technologií, ze kterých se vytrácí a nahrazuje ji angličtina a další světové jazyky.

„Ještě nikdo neinvestoval do výzkumu češtiny pro počítače v takové míře, aby jí stroje skutečně rozuměly a nepotřebovaly pomoc člověka. Nadnárodní firmy sice investují obrovské prostředky do výzkumu jazykových technologií pro pár světových jazyků, čeština je ale na pokraji zájmu. Je možné, že v budoucnu z počítačového prostředí, včetně například internetu, začne postupně mizet," upozorňuje Hajič.

Statistika pro češtinu nefunguje Bílá kniha upozorňuje, že s češtinou si neumějí poradit především složitější jazykové technologie.

„Například rozpoznávání mluvené češtiny je v úplných začátcích. Hlavní problém je obrovské množství různých forem ohýbání slov a také volný pořádek slov ve větě. To znemožňuje použít statistické modelování řeči, které je založeno na nejčastějších slovních tvarech a spojeních a funguje například pro angličtinu," shrnuje studie.

Pro češtinu neexistují kvalitní překladače ani programy, které dokážou česky psaný text analyzovat a udělat z něj v češtině výtah, což se běžně používá například v anglicky mluvících zemích. Jediná jazyková technologie, která funguje na dobré úrovni, jsou takzvané spelling checkery - jednoduché programy, které upozorňují na gramatické chyby v textových souborech.

Neschopnost počítačů rozumět češtině se projevuje i ve vyhledávání na internetu. „Místní vyhledávače už si osvojily některé části morfologické analýzy, ale jejich kvalita se různí," uvádí Bílá kniha. Vyhledávače pracují na základě indexace, kdy „robot" ověří shody požadovaného dotazu se všemi dostupnými zaindexovanými stránkami. „Způsobů, jak se Google učí česky, je několik a nejvíce se naučí právě přímo z vyhledávání. Například ze dvou po sobě zadaných dotazů ‚hotel Praha' a ‚hotel v Praze' se naučí, že Praha a Praze jsou související pojmy," vysvětluje Vladimír Třebický, vývojář Googlu v Curychu.

„Vytvořit jazykový model češtiny nebo jakéhokoliv jiného jazyka je však náročný úkol," dodává Třebický. Podle autorů Bílé knihy není problém v češtině jako takové, ale v neochotě firem investovat do jazyka s malým počtem mluvčích.

„České firmy nechtějí investovat peníze do složité přípravy dat. Chtěly by už hotový jazykový systém, kterému porozumí jejich počítač," říká Hajič a dodává, že například v anglicky mluvících zemích firmy pomocí řečových automatů nebo analyzátorů textů šetří statisíce dolarů. Když například americký zákazník volá do větší firmy, prvních pár otázek položí stroji, který mu porozumí. Lékař zase popis rentgenů diktuje přímo do svého počítače.

Software v angličtině frčí Výrobci softwaru potvrzují, že o angličtinu v počítačích je v Česku stále větší zájem. „Je to skutečně trend. Je to dáno lepší jazykovou vybaveností mladých profesionálů," říká Marek Svoboda, marketingový manažer softwarové firmy Autodesk. „Firmy mají zájem o programy v angličtině. Mnoho společností totiž pracuje v mezinárodním prostředí, ve kterém je nejčastěji používána angličtina a anglické termíny," dodává Svoboda. „Některé softwary zaměřené například na správce IT nebo vývojáře, jsou dostupné v angličtině i v české verzi. Lidé z těchto oborů ale češtinu obvykle ani nevyžadují a chtějí anglickou verzi," dodává Lukáš Křovák, manažer společnosti Microsoft.

***

Co počítače neumějí

Pořádek slov ve větě Čeština má téměř volný slovosled, počítač má proto problém rozeznat, co je podmět a přísudek a která slova se na sebe vážou.
Ohýbání slov Počítače mají problém rozlišit správné tvary slov podle pádu, čísla a rodu.
Někdy je možné oddělit přídavné jméno od podstatného jména a položit jej kamkoliv ve větě („Vánoční nadešel čas.“).
Web a cizí jazyky 57 procent uživatelů internetu v Evropě nakupuje zboží a služby v internetových obchodech, které nejsou v jejich mateřské řeči.
55 procent uživatelů čte na internetu obsah v jiné než mateřské řeči.
35 procent uživatelů píše v cizí řeči e-maily nebo vkládá texty na web. Zdroj: Evropská ko mise