Statistické modelování a automatická analýza přirozeného jazyka (morfologie, syntax, překlad)

Jan Hajič: ÚFAL & CKL MFF UK, Malostranské nám. 25, 118 00 Praha 1, Česká republika (hajic@ufal.mff.cuni.cz)

Abstract

Statistical modeling is now the prevailing method using in automatic procedures of analysis of a natural language. Such an analysis can be performed at various levels, from phonetics to semantics. Two levels of representation are described: a morphological one and a syntactic one that is further subdivided into a surface syntax and deep syntax (tectogrammatics). The role of linguistically annotated corpora will be stressed as a necessary prerequisite for any supervised machine learning algorithms, showing examples from the Prague Dependency Treebank (PDT) being developed at Charles University, Prague. A possible application of some of the tools created during (and thanks to) the development of the PDT will be shown, namely, a machine translation system translating from Czech to Slovak.

1. Úvod

Automatická **analýza** přirozeného jazyka[1] počítačem vyžaduje - koneckonců jako každý problém, který řešíme - rozdělit práci na několik menších, dobře definovaných podproblémů, které pak řešíme (pokud možno) nezávisle. V oblasti zpracování přirozeného jazyka se mluví o tzv. rovinách popisu (a zpracování) jazyka. Tyto roviny jsou uspořádány zdola nahoru (pro účely analýzy jazyka), od roviny nejjednodušší (zabývající se ortografií či akustickou stránkou věci) po rovinu nejsložitější, rovinu významu. Každá rovina má své jednotky popisu, definice vztahů na této rovině, a navazuje bezprostředně na rovinu nižší a vyšší. Obvykle se hovoří o pěti až šesti rovinách (akustika/ortografie, fonetika, fonologie, morfologie, syntax, sémantika), ale často se (například z praktických důvodů) některé roviny slučují dohromady (např. při zpracování textu je rovina ortografická a fonetická téměř vždy sloučena, často i s rovinou fonologickou). Syntax a sémantika rovněž úzce souvisí a ne náhodou se **analýza** na strukturní úrovni často nazývá syntakticko-sémantická, přičemž se zde opět slučují dvě roviny. Naopak, někdy je výhodné (nebo technicky lépe proveditelné) vložit mezi morfologii a syntax ještě jednu rovinu, a to rovinu tzv. povrchové syntaxe. V zahraničních pojetích se obvykle setkáváme jen se dvěma rovinami, a to rovinou morfologickou a povrchově-syntaktickou.

V tomto příspěvku budeme mluvit jednak o rovině morfologické, která v našem případě zahrnuje všechny roviny nižší, s výjimkou té části roviny ortografické, která se zabývá identifikací slov a interpunkce, a jednak o rovině syntaktické, a to jak o její povrchové podobě, tak i o tzv. hloubkové syntaxi, která se zabývá reprezentací jazykového významu. Nebudeme však zde tyto rovinu popisu jazyka rozebírat z lingvistického pohledu, nýbrž se zaměříme na to, jak se tyto roviny promítají do práce s textovými korpusy, zejména pro účely jejich **anotování** a následného **automatického zpracování**.

V poslední části příspěvku popíšeme jednu zajímavou aplikaci, systém automatického **překladu z češtiny do slovenštiny**, který (trochu překvapivě) funguje velmi dobře i přesto, že **analýza** jazyka je v něm omezena vlastně jen na rovinu morfologickou.

2. Morfologická analýza a značkování textu

V úvodu jsme řekli, že v našem pojetí **morfologická** (tvaroslovná) **analýza** spojuje všechny nižší roviny až k rovině tradičně nazývané morfématická. Nezabývá se však prvotním zpracováním textu, kterému se v počítačové analýze nemůžeme vyhnout, a to tzv. **tokenizací**. **Morfologická analýza** tedy vstupuje do hry až v okamžiku, kdy ve vstupním textu jsou identifikována slova, mezery, interpunkce, a pokud možno i začátky a konce vět. Jakkoli triviální se tento úvodní problém může zdát, není tomu tak; již jen definice toho, co to je "slovo"[2] je někdy nejasná: je *byl-li*, *pracovals*, *technicko-hospodářský* nebo *naň* jedno slovo, nebo dvě? Je *New York* nebo *Kostelec n./Č. lesy* jedno slovo, nebo dvě (resp. pět slov)? Obvykle se volí nějaký relativně dobře definovatelný kompromis. Zdá se, že z hlediska dalšího zpracování je vhodné v nejasných případech za slovo brát jednotku co nejkratší. V každém případě ale tokenizace není vlastní součástí morfologické analýzy, v této kapitole tedy předpokládáme, že tokenizace je již dokončena a jednotka zpracování pro morfologickou analýzu je tedy již jednoznačně určena[3]. Tento předpoklad je i z praktického hlediska nepříliš omezující, neboť většina existujících textových **korpusů** je tokenizována, jako např. pro nás důležitý Český národní korpus (Čermák 2001).

2.1. Co je to morfologická analýza?

Na střední škole se učí, že úkolem morfologické analýzy slova[2] je určit morfologické kategorie danému slovu v textu příslušné. Pro člověka je tato definice přijatelná, a koneckonců každý z nás na oné střední škole nakonec nějak uspěl. Při počítačovém zpracování je však situaci třeba definovat a popsat mnohem přesněji.

Především je třeba jasně rozlišovat mezi morfologickou kategorií a její hodnotou. **Číslo** je morfologickou kategorií, **singulár** (jednotné číslo) její hodnotou. V češtině a slovenštině je možno rozlišovat mnoho kategorií, v našem systému jich používáme celkem 13: slovní druh, slovní "poddruh", rod, číslo, pád, přivlastňovací rod, přivlastňovací číslo, osobu, čas, slovesný rod, negaci, stupeň a variantu. Hodnotami jsou např. čísla 1 až 7 pro české pády, "aktivní" a "pasívní" pro slovesný rod, atd. Nejbohatší kategorií je slovní poddruh, který má celkem 75 možných hodnot, nejvíce z nich pro zájmena.

Pozornému čtenáři jistě neunikne, že v seznamu kategorií není nejen kategorie vzoru (vzor má v systému pouze pomocnou úlohu, a je zcela nepotřebný pro navazující analýzu jazyka), ale ani např. kategorie způsobu; důvod je však prostý: **morfologická analýza** v našem systému pracuje bez ohledu na kontext, tj. zpracovává **izolovaně** vždy jen jedno slovo (slovní tvar). Tím "odsouvá" řešení některých problémů na pozdější dobu, a jakkoli je to z lingvistického pohledu bolestné, je tento přístup (vyplývající z dělení popisu a zpracování jazyka na jednotlivé roviny) jediný možný, neboť umožňuje nemíchat dohromady věci, které k sobě nepatří a byly by tudíž těžko formalizovatelné a zpracovatelné. Ze stejných důvodů je nutno brát kategorii slovesného času jako kategorii příslušnou k analyzovanému participiu (*pracoval*), nikoli k celému analytickému tvaru (který v uvedeném příkladu může být jak času minulého *pracoval jsem*, tak i času přítomného v podmiňovacím způsobu *pracoval bych*).

Vzhledem k tomu, že **morfologická analýza** pracuje s jednotlivými slovy z textu izolovaně, bez ohledu na kontext, tak se na rozdíl od úloh řešených na střední škole nezabývá ani jednoznačnou identifikací hodnot morfologických kategorií. Pochopitelně, ani nemůže: bez větného kontextu není možno mezi jednotlivými možnostmi vůbec vybírat. Problémem jednoznačného určení hodnot morfologických kategorií se zabývá tzv. značkování, ke kterému se vrátíme za chvíli.

Pro počítačové zpracování se zavádí tzv. množina morfologických značek (**tagset**). Každá značka shrnuje hodnoty morfologických kategorií pro jeden slovní tvar. Pro vlastní zpracování se používá několik typů notací, z nichž nejrozšířenější je notace tzv. **poziční**. V této notaci se každé kategorii přiřadí pozice ve značce, a každé hodnotě jeden znak, který se zapisuje na příslušnou pozici. Slovní druh je tedy např. na první pozici, a jeho hodnoty jsou reprezentovány např. znaky **N** (pro podstatné jméno, noun), **A** (pro adjektivum), atd. Hodnoty pro daný slovní tvar irelevantních kategorií jsou označeny speciálním znakem, obvykle pomlčkou. Např. tedy pro obyčejné podstatné jméno rodu mužského neživotného ve 4. pádě jednotného čísla v pozičním systému s 15 kategoriemi má příslušná značka tvar **`NNIS4-----A----`** (první pozice je slovní druh (**N**), druhá slovní poddruh (zde **N**), třetí rod (**I** pro mužský neživotný, masc. inanim.), čtvrtá číslo (**S** pro singulár), pátá pád (**4** pro akuzativ), atd. (**A** na jedenácté pozici specifikuje, že dané slovo není negováno příslušnou předponou).

Co tedy (počítačová) **morfologická analýza** vlastně dělá? Po výše uvedeném výčtu toho, co nedělá, by se zdálo, že nedělá téměř nic; samozřejmě, že tak tomu není. **Morfologická analýza** pro každý slovní tvar určí všechny možnosti kombinací hodnot morfologických kategorií, které danému tvaru vůbec mohou příslušet. Že i to je obrovská pomoc pro další zpracování, je vidět z prostého číselného srovnání: zatímco všech možných značek (kombinací hodnot morfologických kategorií) je v našem systému pro češtinu přes 4400, průměrný počet značek po morfologické analýze je menší než 5 (na jedno slovo v běžném textu).

Počítačová **morfologická analýza** však musí řešit ještě jeden problém, na zmiňované střední škole probíraný pouze okrajově: tzv. problém lematizace. Lematizace určuje pro každý slovní tvar jeho základní podobu (obvykle tvar, ve kterém slovo najdeme ve slovnících). Ani lematizace není obecně při zpracování izolovaného slova jednoznačná. Navíc je nutno rozlišovat mezi slovy, která jsou v základním tvaru homonymní - např. *stát* (jako státní útvar) a *stát* (jako sloveso). Počítačová lematizace proto ještě navíc tato slova rozlišuje a jednoznačně identifikuje (např. připojením číselného indexu k základnímu tvaru slova, např. *stát-1*, *stát-2* atd.).

Formálně tedy můžeme popsat morfologickou analýzu jako matematickou funkci, která posloupnosti znaků (písmen) jazyka přiřazuje množinu možných výsledků, složených vždy z dvojic <lema,značka>:

Ma(f) --> { <l,t>; l □ L, t □ T },

kde f z A+ je slovní tvar složený z písmen abecedy A analyzovaného jazyka (např. *stát*), L je množina identifikací lemat (obvykle ve formě řetězce nějakých znaků, považovaného ovšem za nedělitelný) v daném případě bude jedním z možných výsledků např. *stát-1*), a T je množina značek používaná pro daný jazyk (jako např. **`NNIS4-----A----`**; opět jde o řetězec znaků, považovaný z hlediska definice za atomický).

Prakticky **morfologická analýza** pracuje s (tokenizovaným) textem, v dohodnutém formátu, a na výstupu je tentýž text obohacený o lemata a morfologické značky (obr. 1 a 2).

```
<f cap>Pekař

<f>peče

<f>housky

<D>

<d>.
```

Obr. 1 Vstup do morfologické analýzy - tokenizovaný text

```
<f cap>Pekař<MMl>pekař<MMt>NNMS1----A----

<f>peče<MMl>péci<MMt>VeYS------A----<MMt>VB-S---3P-AA---

<f>housky<MMl>houska<MMt>NNFP1-----A----<MMt>NNFP4-----A----
<MMt>NNFS2-----A----

<D>

<d>.<MMl>.<MMt>Z:-------------
```

Obr. 2 Výstup z morfologické analýzy (zjednodušeno)

2.2. Proces morfologické analýzy

Morfologická analýza, jejíž definici jsme uvedli v předchozí sekci, je ovšem realizována v počítači nikoli jako matematická funkce, ale jako výpočetní procedura. Jako základní datová struktura slouží pro daný přirozený jazyk jeho **morfologický slovník**, který je používán vlastním **algoritmem** morfologické analýzy (v zásadě pak již na jazyce nezávislým). Způsobů, jak efektivně provádět morfologickou analýzu, se používá několik (Koskenniemi 1983, Mohri 1998), my zde popíšeme náš systém "přímé" analýzy. Ten potřebuje ke své práci jednak morfologický slovník, a samozřejmě i příslušný algoritmus, který vlastní morfologickou analýzy na základě slovníku realizuje. Na základě stejného slovníku pak může probíhat i i **morfologická** syntéza, o té se ale zmíníme až v sekci 4 o strojovém překladu.

2.2.1. Morfologický slovník

Morfologický slovník obsahuje ke každému lematu informaci o **kmeni** slova (v našem případě, kvůli sloučení nejnižších rovin popisu jazyka do jedné, je za kmen slova považována ta část slova, která se při ohýbání nemění), a o přípustných koncovkách. Množina koncovek tvoří **vzor**. U každé koncovky je navíc informace o tom, které značky (kombinace hodnot morfologických kategorií) jí pro daný vzor odpovídají.

Příkladem vzoru je např. následující množina koncovek a jejich značek:

```
"" NNIS1-----A----, NNIS4-----A----

"u" NNIS2-----A----, NNIS3-----A----, NNIS6-----A---1

"e" NNIS5-----A----

"ě" NNIS6-----A----

"em" NNIS7-----A----

"y" NNIP1-----A----, NNIP4-----A----, NNIP5-----A----, NNIP7-----A----

"ů" NNIP2-----A----

"ům" NNIP3-----A----

"ech" NNIP6-----A----
```

Tento vzor je v našem systému označen **hd2x**. Tedy k lematu *stát-1* bude v morfologickém slovníku uveden kmen "stát" a vzor **hd2x**.

Pro každý vzor je dále ve slovníku uvedeno, zda připouští negaci slova pomocí předpony "*ne-*" (tj. negaci) a u každé koncovky dále informace o tom, zda připouští připojení předpony "*nej-*" (stupňování).

Pro velmi nepravidelná slova jsou pak ve slovníku uvedenu všechny jejich tvary i s příslušnými značkami.

2.2.2. Algoritmus morfologické analýzy

Tzv. "přímá" **analýza** slovních tvarů je založena na vyčerpávající analýze slova z hlediska možné segmentace na kmen a koncovku (případně i předpony *ne-* a *nej-*). Pro každou takto získanou dvojici kmene a koncovky je nutno ověřit, zda se ve slovníku vyskytuje jak kmen, tak i koncovka a zda kmen i koncovka náleží ke stejnému vzoru. Všechny dvojice lemat (příslušných ke kmeni/kmenům) a značek (nalezených ve slovníku u příslušných koncovek) jsou pak prohlášeny za výsledek morfologické analýzy. Podrobněji o v současnosti používané morfologické analýze češtiny viz (**Hajič**, 2001).

Příkladem může být slovo (slovní tvar) *housky*. Toto slovo je možno rozdělit na kmen *housky* + nulovou koncovku, nebo na *housk* + *y*, nebo na *hous* + *ky*, atd. až k *h* + *ousky* (kmen nulové délky se nepřipouští). Z těchto možností nakonec bude správná jen možnost *hous* + *ky*, neboť ve slovníku je neměnná část základu (zde jen *hous*, neboť 2. p. mn. čísla je *hous+ek*). Koncovky *y*, *sky*, a nulová koncovka jsou sice ve slovníku koncovek uvedeny také, ale kmen *housk (hou)* je nepřipouští (resp. nejsou uvedeny v seznamu koncovek pro vzor příslušný danému kmeni).

Modernější systémy používají pro jádro systému morfologické analýzy aparát konečných automatů, resp. v kombinaci s fonologií aparát tzv. sekvenčních strojů (konečných převodníků)[4]. Prvním takovým systémem byla tzv. "Two-level morphology" (Koskenniemi, 1983), následovníky pak Xerox Language Tools (XLT, zpracována je i čeština, viz (Skoumalová, 1997)), a v poslední době je volně k dispozici univerzální soubor nástrojů pro konečné automaty a převodníky (nejen pro morfologii, či spíše snad použitelný i pro morfologii!) FSM od AT&T Research (Mohri et al. 1998). Je však nutno podotknout, že v dnešní době už vnitřní struktura (implementace) morfologického analyzátoru nehraje prvořadou roli - důležitá je spíše udržovatelnost a rozšiřitelnost systému.

2.3. Značkování (zjednoznačňování morfologické analýzy)

Značkování (anglicky poněkud nevhodně nazývané "Part-of-Speech tagging") je v rámci popisu a zpracování jazyka pomocí rovin jakýsi "krok stranou": snažíme se totiž na úrovni morfologické analýzy o něco, co alespoň teoreticky přísluší až rovině syntaktické (ať už povrchové nebo hloubkové). Nicméně je to problém velmi praktický, jehož výsledky jsou použitelné ve třech směrech: jednak jako (zatím) finální krok při značkování korpusů pro lexikografické účely, dále jako krok výrazně zrychlující syntaktickou analýzy (byť do ní vnáší jistou míru chyb, jak uvidíme dále), a v neposlední řadě i pro některé aplikace, které mohou s výhodou využít i jen částečnou jazykovou analýzu (např. pro vyhledávání v elektronických slovnících, pro vyhledávání informací obecně, a dokonce i pro strojový překlad pro blízké flektívní jazyky - viz dále sekce č. 4).

Značkování již může využít pro zjednoznačnění výstupu morfologické analýzy (na rozdíl od ní samé) kontext, ve kterém se analyzované slovo nachází. Dnes se téměř výhradně používají pro značkování metody statistické, založené na strojovém učení. Počítač se tedy naučí, že po určitých předložkách následují jen některé pády, že na začátku věty nalezneme nejspíše pád první než jakýkoliv jiný, nebo že slovo *při* je téměř vždy předložka, jen velmi málokdy tvar slova pře, a téměř nikdy rozkazovací způsob od slovesa *přít* (a k tomu se, doufejme, naučí i to, kdy jde přeci jen o (soudní) *při*).

Jak se však může počítač takovou věc naučit? Potřebuje k tomu (alespoň v dosud nejúspěšnějších metodách) předem **ručně** označkovaný korpus. Takový korpus je samozřejmě velmi pracnou záležitostí; pro spolehlivé naučení, kdy procento chyb klesá (pro češtinu) pod 5%, bylo třeba označkovat přes 1.5 miliónu výskytů slov v textu (přitom každé zdvojnásobení tohoto počtu přinese jen několik desetin procenta zlepšení, a jistou hranici úspěšnosti zřejmě nelze překročit vůbec). Označkované korpusy jsou proto velmi cenným zdrojem lingvistických informací (nejen pro automatické strojové učení, ale samozřejmě i pro vyhodnocování jiných metod, použitých pro značkování). Příkladem takových korpusů jsou např. Brown Corpus (první značkovaný korpus na světě z konce 60. let), Penn Treebank (Marcus 1993), a pro češtinu čerstvě vydaný Pražský závislostní korpus (**Hajič** et al. 2001b).

Učení z ručně označkovaného korpusu (takovému korpusu se říká **trénovací data**) může probíhat několika způsoby. Velmi jednoduchý a účinný (a dosud prakticky nepřekonaný) je postup, při kterém se spočítají relativní četnosti značek následujících po dvojici bezprostředně předcházejících značek v textu (takový způsob se nazývá **HMM tagging**: viz (Church 1992, Hladká 1994, Mírovský 1999, Hladká 2000, **Hajič** et al. 2001a). Pro každou dvojici značek (tzv. **historii**) se tak vytvoří menší či větší tabulka, ve které jsou uvedeny relativní četnosti značek po ní následujících v trénovacích datech. Jakkoli je tento systém lingvisticky jasně

neadekvátní, značkování založené na efektivním algoritmu aplikace těchto tabulek (virtuálně rozšiřujícím délku historie (kontextu) na mnoho slov na obě strany od analyzovaného slova) na kontinuální text (Jelinek 1998) dává velmi dobré výsledky: pro angličtinu se dosahuje i méně než 3% chyb na prakticky libovolném textu, pro češtinu pak okolo 5%.

Pro češtinu vyvíjíme při její bohatosti značek ještě jeden systém (**Hajič** 2001), který, jak doufáme, přiblíží úspěšnost značkování angličtině. Tento systém je založen na individuálním "předpovídání" hodnot jednotlivých morfologických kategorií. Statisticky, automaticky vybraná vhodná "pravidla" (*features* neboli *rysy*) se ohodnotí váhami (opět zcela automaticky v procesu učení z předem ručně označkovaných dat). Takto ohodnocená "pravidla" se pak používají v procesu automatického značkování tak, že se pro každou hodnotu spočítá její pravděpodobnost v daném kontextu, a výsledná značka je pak "kompromisem", neboť se pochopitelně vybírá pouze mezi značkami nabídnutými morfologickou analýzou. Tato metoda je nyní stejně úspěšná jako výše uvedená metoda HMM taggingu, potřebuje však méně statistických dat při vlastním značkování (avšak je velmi náročná v průběhu učení na čas výpočtu).

Kromě čistě statistických přístupů uvažujeme rovněž o možné kombinaci s metodami "nestatistickými", tj. tradičně lingvistickými, které především pracují s ručně vytvořenými pravidly s komplexními podmínkami. Tato pravidla použitá samostatně vykazují poměrně malou úspěšnost z hlediska počtu víceznačností, které jsou schopny řešit, avšak jsou poměrně přesná (v případech, které řešit umějí). Systém pak pracuje tak, že tato "lingvistická" pravidla jsou aplikována nejdříve, čímž se víceznačnost zredukuje, aniž by byly odstraněny správné varianty, a pak "statistická", tj. automaticky naučená "pravidla" zjednoznačňování dokončí (**Hajič** et al. 2001a).

Jako konkrétní příklad uveďme opět větu *Pekař peče housky*. Funguje-li disambiguace správně, na základě vstupu z obr. 2 obdržíme následující výstup (obr. 3), ve kterém je pro každé vstupní slovo už jen jedna značka a jedno lema:

```
<f cap>Pekař<MDl>pekař<MDt>NNMS1----A----

<f>peče<MDl>péci<MDt>VB-S---3P-AA---

<f>housky<MDl>houska<MDt>NNFP4-----A----

<D>

<d>.<MMl>.<MMt>Z:-------------
```

Obr. 3 Zjednoznačněný výsledek morfologické analýzy

U slova *Pekař* nebylo nutno rozhodovat o ničem, neboť již bylo jednoznačně určeno morfologickým analyzátorem[5]. Slovo *peče* je samozřejmě v této větě v přítomném čase a 3. osobě (nikoli jako přechodník!), a *housky* jsou zde ve 4. pádě množného čísla.

3. Syntaktická závislostní analýza

Jakkoli je **morfologická analýza** a (morfologické) značkování zajímavé a užitečné, nedotýká se přímo struktury věty. Z hlediska skladby věty potřebujeme zjišťovat, která slova

jsou ve vztahu gramatické **závislosti**: řídící slovo je "důležitější", ve větě jej obvykle nelze vynechat bez narušení gramatické skladby věty, a obyčejně určuje většinu gramatických kategorií slova závislého (např. na základě shody).

Přímo zjišťovat skladbu věty je však velmi obtížné: důvodem jsou kromě již známé nejednoznačnosti jazyka i (ne-li více) např. elipsy (slova ve větě vynechaná, byť z hlediska významu a standardní definice syntaxe nezbytná), konstrukce bez slovesa, koordinace a apozice, parenteze (vsuvky) apod. Proto jsme se rozhodli vložit mezi rovinu morfologickou a syntaktickou rovinu tzv. analytickou, která zhruba odpovídá rovině povrchové syntaxe známé z jiných teoretických přístupů. Pracujeme tedy se dvěma syntaktickými rovinami: rovinou **analytickou**, a rovinou vlastní syntaxe, tzv. rovinou **tektogramatickou** (Sgall et al. 1986).

3.1. Analytická rovina syntaxe

Na analytické rovině se reprezentace věty zachycuje závislostním stromem[6] s vrcholy, případně i hranami ohodnocenými jedním nebo několika atributy. Ke každému slovu z analyzované věty (token, tj. i interpunkce) přísluší právě jeden vrchol závislostního stromu. Závislostní vztahy jsou určeny hranami takového stromu, a hodnoty příslušné k jednotlivým hranám určují (povrchově) syntaktickou funkci závislého uzlu vzhledem k uzlu řídícímu. Hodnotami u vrcholů jsou pak dva údaje: příslušné lemma (pro interpunkci se definuje jako identické s původní formou interpunkce) a **morfologická** značka (soubor značek - tagset - se rovněž vhodně rozšiřuje kvůli interpunkci, podobně jako na rovině morfologické). Pro lepší čitelnost se u každého vrcholu zaznamenává i původní tvar daného slova (ačkoli jeje lze jednoznačně vyvodit z lematu a morfologické značky) a je zde i řada dalších, technických a pomocných atributů. Z technických důvodů se rovněž hodnota hrany (tj. povrchově-syntaktická funkce závislého slova) uvádí u závislého uzlu.

Jako příklad lze uvést jednoduchou větu *Kominík vymetá komíny* (obr. 4).

Obr. 4 Analytická reprezentace věty *Kominík vymetá komíny*.

Vidíme, že *Kominík* je podmětem věty (**Sb**), *vymetá* je predikát (řídící sloveso hlavní věty, Pred), a *komíny* je předmět (**Obj**). Závěrečná interpunkce podle zásady co slovo (token) to vrchol stromu je rovněž přítomna, a to s funkcí **AuxK** (speciální funkce pro koncová interpunkce).

Lze tedy říci, že analytická rovina je velmi podobná tomu, co jsme se všichni učili na základní a střední škole, snad s výjimkou postavení podmětu (podmět není na stejné úrovni jako predikát) a přítomnosti všech slov z věty (to se týká nejen interpunkce, ale samozřejmě i předložek, spojek, pomocných a sponových sloves atd.).

Účelem analytické anotace jako předstupně k rovině tektogramatické (sekce 3.2.) je zachytit základní závislostní vztahy, označit pomocná slova a jejich vztah k jiným jednotkám na této rovině (i když jistě nejde o skutečnou závislost v obvyklém smyslu), označit elipsu, pospojovat koordinované a aponované členy věty, označit vsuvky apod.

Podmínka, že každému slovu ze vstupního textu odpovídá právě jeden vrchol závislostního stromu, není náhodná. Umožňuje totiž vytvořit relativně efektivní nástroj pro

automatickou povrchově-syntaktickou analýzu vět přirozeného jazyka (tj. v našem případě češtiny). Obecný postup je zde podobný jako při morfologickém značkování (sekce 2.3.): používají se primárně statistické metody založené na strojovém učení parametrů (pravděpodobností), používaný pravděpodobnostní model pro takovou analýzu je však mnohem komplikovanější. V našem případě používáme analyzátor (Collins 1997) adaptovaný pro češtinu na letním Workshopu na Johns Hopkins University v roce 1998 (**Hajič** 1998), který dokáže správně určit kolem 80% všech závislostí v testovacím textu.

Pochopitelně i pro učení syntaktického analyzátoru jsou třeba trénovací data (tj. ručně syntakticky anotovaný korpus). Práce na ručním syntaktickém anotování je mnohem náročnější než obdobná práce na zjednoznačňování morfologickém, a to jak z hlediska softwarové přípravy (anotovací nástroje musí pracovat s grafickým obrázkem analyzovaného stromu, tak, jak jsou na to lingvisté-anotátoři zvyklí), přípravy pokynů pro anotování (s trochou nadsázky lze říci, že jsme museli přepsat, či snad explicitně dopracovat povrchovou syntax češtiny, viz **Hajič** et al. 1997), i vlastní anotovací práce. Pro češtinu jsou taková data obsažena na CDROM **Pražský závislostní korpus** (**Hajič** 1998, **Hajič** et al. 2001b), spolu se všemi potřebnými nástroji na (ruční) syntaktické anotování korpusu. Na tomto CD je anotováno téměř 1.5 miliónu slov (asi 90 tisíc vět) na analytické rovině.

3.2. Tektogramatická rovina syntaxe

Naším cílem však není zastavit se na rovině povrchové syntaxe. Připravujeme proto anotaci na rovině **tektogramatické**, kde se používá jiný repertoár závislostních funkcí (které označují význam, nikoli jen povrchový vztah), odpadají vrcholy s pomocnými slovy a částečně i interpunkcí, a navíc přibývají na povrchu vypuštěné, leč z významového hlediska přítomné elipsy. Navíc zde přibývá označení koreference a aktuální členění. Příklad věty anotované na této rovině je na obr. 5.

Obr. 5 Anotace věty na tektogramatické rovině

Bližší popis tektogramatické roviny je možné nalézt z teoretického hlediska v (Sgall et al. 1986, Petkevič 1995), a pak přímo ve formě příručky pro anotátory (Hajičová et al. 2000)

Tektogramatická rovina je jakýmsi mezičlánkem mezi lingvistickou analýzou a další analýzou sémantickou, logickou, analýzou textu apod., vedoucí ke skutečnému porozumění přirozenému jazyku. Předpokládáme, že pro češtinu dokážeme na této rovině anotovat řádově obdobný počet vět jako na rovině analytické (cca 60 tisíc) do konce r. 2004.

4. Strojový překlad mezi blízkými jazyky

4.1. Základní idea, možná zjednodušení

Ačkoli reprezentaci věty na tektogramatické rovině, jak byla popsána v předchozí sekci, považujeme za hlavní formální nástroj k popisu lingvistického významu, který by měl být jádrem každé aplikace vyžadující porozumění přirozenému jazyku, někdy se obejdeme s prostředky mnohem jednoduššími.

Takovou aplikací je například strojový překlad mezi velmi blízkými jazyky, jako je čeština a slovenština, a jistě by se našly další příklady (někdy nejde jen o blízké jazyky, ale může jít i o varianty jednoho jazyka, ať už pravopisné, nářeční apod.). Pro jazyky vzdálenější (jako např. čeština a ruština, viz (**Hajič** et al., 1987)) je otázka složitější: je jasné, že syntax je v jistých okamžicích potřebná, avšak není jasné, zda chyby, jichž se nutně v syntaktické analýze dopustíme, vyváží tuto výhodu.

I při zjednodušené analýze ve strojovém překladu mezi češtinou a slovenštinou (podrobněji viz (**Hajič** et al. 2000)) zachováváme tradiční scénář strojového překladu (obr. 6).

Transfer

Analýza

Syntéza

Zdrojový jazyk **Cílový jazyk**

Obr. 6 Obecné schéma strojového překladu

4.2. Tři fáze překladu: analýza, transfer, syntéza

Při analýze zdrojového jazyka se jednotlivé věty analyzují bez ohledu na to, do kterého jazyka se překládá. Buduje se reprezentace věty vhodná pro fázi transferu ("vlastního překladu"). Ve složitých systémech touto reprezentací může být hloubková syntaktická reprezentace, nebo dokonce logická struktura užité věty, avšak v našem zjednodušeném případě bude touto analýzou pouze **analýza morfologická**, zjednoznačněná pomocí statistického modulu (taggeru, viz sekce 2).

Transfer pak bude zcela deterministický proces, který nahradí každé zdrojové (české) slovo (resp. jeho lemma) jeho cílovým (slovenským) ekvivalentem, a českou morfologickou značkou značkou slovenskou (ve většině případů bude tato značka zcela stejná, nebo jen formálně odlišná). V některých případech však musíme slovenskou značku poněkud zobecnit, neboť se výjimečně mění rod substantiva, zvláštní varianta koncovky v určitém pádě se překládá standardně, apod.

Ve fázi syntézy (generování) se pak ze slovenských lemat a slovenských morfologických značek vytvoří výsledná forma slovenského slova. Na závěr se pak doplní velká písmena

podle pravidel pravopisu a věta se zformátuje, případně se do textu vloží zpět původní formátování české věty, bylo-li v ní přítomno.

4.3. Analýza

Ve fázi analýzy proběhne tokenizace textu (pokud již vstupní text není takto zpracován, což obvykle není), uchování formátovací informace (to je důležité např. tehdy, je-li původní text např. v HTML, RTF a chceme původní formátování pokud možno zachovat), a převedení do jednotného formátu pro další zpracování, což je SGML formát obdobný uchovávání textů v ČNK, neboť se pochopitelně používají podobné nástroje (morfologie, tagger, atd.). Proběhne i identifikace hranic vět, a speciálně se označí úseky, které je třeba překládat (na rozdíl od např. formátovacích značek).

Příklad:

Věta *Transakce slouží k zobrazení zamčených záznamů v databázi.* bude po tokenizaci a převodu do SGML vypadat takto:

```
<s id="/disk1/home/hajic/f/projects/data/SMALL.tmq-p1s37">
<f>Transakce
<f>slouží
<f>k
<f>zobrazení
<f>zamčených
<f>záznamů
<f>v
<f>databázi
<D>
<d>.
```

Obr. 7 Tokenizovaný vstup do systému překladu

SGML značkou <s> jsou označeny hranice vět, <f> označuje slova, <d> interpunkci, a <D> je značka pro **ne**přítomnost mezery.

Po tokenizaci se text zpracuje morfologickou analýzou a značkovačem (taggerem) (viz sekce 2). Na výstupu bude u každého slova uvedeno lemma a tag po zjednoznačnění (jen připomínáme, že zjednoznačnění probíhá na základě kontextu, a to kontextu v české větě). Tato část systému je posledním krokem ve zjednodušeném systému překladu, neboť další

analýza (syntaktická) již v systému není. JE tedy možné říci, že morfologické zjednoznačnění je jádrem lingvistické analýzy celého systému překladu. Tato fáze je zároveň zcela nezbytná, neboť i když čeština a slovenština mají prakticky shodnou syntax, liší se výrazně právě ve slovníku a morfologii (v paradigmatech), a ve z toho vyplývajících typech homonymie. Např. české slovo *zobrazení* z výše uvedené věty může být použito ve 27 různých morfologických interpretacích, řada z nich pak má různý slovenský překlad: *zobrazení, zobrazenia, zobrazenie, zobrazeniu* atd. Překlad "(slovní) tvar za tvar" je tak (i kdybychom vyřešili technické obtíže s milióny slovních tvarů, které by musely být ve slovníku takového systému) tedy evidentně není možný.

Příklad:

Výsledek po morfologické analýze a jejím zjednoznačnění je na obr. 8.

```
<f>Transakce<MDl>transakce<MDt>NNFS1-----A----

<f>slouží<MDl>sloužit<MDt>VB-S---3P-AA---

<f>k<MDl>k-1<MDt>RR--3----------

<f>zobrazení<MDl>zobrazení<MDt>NNNS3-----A----

<f>zamčených<MDl>zamčený<MDt>AAIP2----1A------1A----

<f>záznamů<MDl>záznam<MDt>NNIP2-----A----

<f>v<MDl>v-1<MDt>RR--6----------<D>

<f>databázi<MDl>databáze<MDt>NNFS6-----A----

<D>

<d>.<MDl>.<MDt>Z:------------
```

Obr. 8 Zjednoznačněný výsledek morfologické analýzy

SGML značky <MDl> slouží k označení lematu, <MDt> uvádí morfologickou značku. Slovo *zobrazení* se zde tedy jednoznačně určilo (velmi pravděpodobně díky předcházející předložce *k*, která vyžaduje třetí pád) jako neutrum v dativu; jeho určení jako singuláru pak plyne spíše z faktu, že v obdobných konstrukcích se používá spíše singulár (kontext nic takového nevyžaduje).

Věta je nyní připravena pro vlastní překlad, který nazýváme tradičně transferem.

4.4. Transfer

Ve fázi transferu se nahradí česká lemata slovenskými a značky se rovněž "přeloží" do zobecněné formy, vyhovující rovněž slovenskému systému morfologických značek. V této fázi tedy teprve do hry vstupuje slovenština (dosud se systém zabýval pouze zpracováním

češtiny jako zdrojového jazyka). Podobně teprve zde by se zapojila např. polština, pokud bychom chtěli překládat právě do ní.

Překlad značek lze zařídit poměrně snadno. Překlad je řízen tabulkou, ve které je ke každé české morfologické značce přiřazena jedna nebo více zobecněných slovenských morfologických značek, v prioritním pořadí.

Zobecněnou morfologickou značkou se myslí **morfologická** značka, která není plně specifikována. Modul generování (viz dále sekce 4.5.) je přizpůsoben tak, že za nespecifikovanou hodnotu určité morfologické kategorie (např. rodu) dosadí všechny možnosti, které přicházejí pro dané slovo v úvahu. (V případě více možností vybere první, která zpracováním projde.) V námi používaném pozičním systému se pro nespecifikovanou hodnotu používá znak tečka ('.'). Tedy například **morfologická** značka pro třetí pád (dativ) jednotného čísla (sg.) obyčejných substantiv s nespecifikovaným rodem vypadá takto:

```
NN.S3-----A----
```

Prioritní seznam cílových morfologických značek pak ve spolupráci s modulem generování zajistí, že na výstupu se objeví první vytvořený slovní tvar (za použití morfologické značky s nejvyšší prioritou). Tím se ošetřují jednotně jak případy změny rodu u substantiv, tak i případy, kdy rod je nejednoznačný a je třeba dát přednost rodu použitému v češtině.

Příkladem takového seznamu je např. posloupnost dvou značek:

```
NNNS3-----A---- NN.S3-----A----
```

která říká, že nejprve je třeba zkusit rod střední, ale pokud taková

značka s daným lematem nic nevygeneruje, má se použít libovolný rod.

Prioritní systém spolu s ideou zobecněných morfologických značek umožňuje elegantně a bez dalších zásah' do slovníku řešit i případy, kdy některé gramatické charakteristiky slovenštiny neodpovídají češtině.

Vlastní slovník (tj. překladový slovník lemat) je vytvořen tak, že může zpracovávat i víceslovnou terminologii. Pomocí pravidla "delší vyhraje" pak umožňuje řešit i nejednoznačné případy, kdy ve slovníku je zvlášť uveden jak několikaslovný termín, tak i jeho počátek.

Terminologický slovník však znamená jednu nevyhnutelnou komplikaci: ve flektívních jazycích může být část termínu skloňovaná spolu s řídícím slovem termínu, ale část může být fixní a tedy i ve slovníku uvedená v příslušném pádě (nebo i čísle). Při analýze češtiny však ještě nevíme, a ani nemůžeme vědět, zda určité slovo je součástí nějakého termínu nebo ne, a

proto všechna slova jsou lematizována jednotlivě. Potřebujeme proto, aby slovník obsahoval ve formě lemat i ty části termínů, které nepodléhají ohýbání.

Například termín *daň z příjmů* je třeba ve slovníku lematizovat jako *daň z příjem*, jinak by se v textu nemohlo najít poslední slovo (*příjmů*) termínu.

Abychom vyloučili pracné ruční zpracování slovníku, používáme naprosto stejný morfologický analyzátor a značkovač i pro předzpracování slovníku, a to na obou jeho stranách (české i slovenské). Tím je zaručena naprostá shoda lemat s morfologickými moduly, a to i tehdy, jestliže lemata obsahují nějakou vnější identifikaci, jako např. číslo významu (viz *k-1*, *k* jako předložka).

Transfer tedy vydá posloupnost slovenských lemat s morfologickými značkami; v této posloupnosti už česká slova ani značky nemusí být (obr. 9).

```
<Gil>transakcia<Git>NNFS1-----A----<Git>NN.S1-----A----

<Gil>slúžiť<Git>VB-S---3P-AA---

<Gil>k-1<Git>RR--3----------

<Gil>zobrazenie<Git>NNNS3-----A----<Git>NN.S3-----A----

<Gil>zamknutý<Git>AAIP2----1A----

<Gil>záznam<Git>NNIP2-----A----<Git>NN.P2-----A----

<Gil>v-1<Git>RR--6----------

<Gil>databáza<Git>NNFS6-----A----<Git>NN.S6-----A----

<D>

<Gil>.<Git>Z:------------
```

Obr. 9 Výsledek transferu (vlastního překladu lemat a morf. značek)

SGML značky <Gil> označují slovenské lema, <Git> pak každou slovenskou morfologickou značku, a to jak v případě, že je uvedena značka jediná, tak i v prioritním seznamu.

4.5. Syntéza (Generování)

Vzhledem k tomu, že na české straně je **analýza** ukončena po morfologické analýze a značkování, je i syntéza na slovenské straně výlučně morfologickou (a formátovací) záležitostí. Morfologický generátor (program, jehož funkce je inverzní k funkci morfologického analyzátoru) pak z každého lematu a prioritního seznamu zobecněných značek vytvoří posloupnost slovenských slov v odpovídajících formách.

Morfologický generátor slovenštiny používá stejná data jako morfologický analyzátor slovenštiny použitý pro předzpracování slovníku pro transfer (viz sekce 4.4.), automaticky zkonvertovaný pro efektivní vyhledávání mezi kmeny, vzory a koncovkami pro účely morfologické syntézy.

V našem příkladu je tedy výsledkem morfologické syntézy věta na obr. 10.

```
<Gef>transakcia

<Gef>slúži

<Gef>k

<Gef>zobrazeniu

<Gef>zamknutých

<Gef>záznamov

<Gef>v

<Gef>databáze

<D>

<Gef>.
```

Obr. 10 Výsledek překladu do slovenštiny

Po závěrečném formátování pak dostaneme konečný výsledek *Transakcia slúži k zobrazeniu zamknutých záznamov v databáze.*

4.6. Použití v praktických systémech

Strojový překlad sám o sobě nemá valnou praktickou hodnotu, není-li použit ve vhodně koncipovaném softwarovém systému, ať už pro malé "domácí" nebo on-line použití, nebo pro profesionální překlad ve velkém.

4.6.1. Systémy s překladovou pamětí

Nejefektivnější systémy pro profesionální strojový překlad jsou založeny na využití tzv. překladových pamětí. Překladová paměť si pamatuje veškerý již jednou přeložený text (ukládá si vždy dvojici zdrojová věta --> její překlad), a při překladu dalšího, nového textu je schopna porovnat nově překládanou větu s touto pamětí, a nabídnout překladateli překlad, který je u příslušné zdrojové věty uložen. Přitom jednotlivé věty nemusí být zcela identické, mohou se lišit v jednom nebo několika slovech, v číselné hodnotě, interpunkci apod. Efektivnost systémů strojového překladu založených na překladových pamětech pak plyne z faktu, že většina "průmyslově" prováděných překladů se týkají jen málo změněných verzí toho, co již jednou bylo přeloženo (např. příručka k textovému editoru se jistě změní od verze k verzi jen málo, zvlášť při vysoké frekvenci "upgrade" takových softwarových produktů).

Do tohoto systému je velmi jednoduché zapojit strojový překlad tak, že vytvoříme "překladovou paměť" a naplníme ji všemi překládanými větami spolu se strojově vytvořeným překladem. Překladatel pak ke každé jím překládané větě dostane pro něj obvyklým způsobem návrh překladu, jako kdyby daná věta byla již někým v minulosti přeložena. Je samozřejmé, že překladatel musí být varován, že se jedná a strojový překlad, a ne o překlad "lidský". Navíc je třeba zajistit (technickými prostředky), aby tatáž věta, byla-li dříve již přeložena člověkem, dostala při výběru z překladové paměti prioritu před větou přeloženou strojově.

4.6.2. Vícejazyčný překlad

Jednoduchý, rychlý a relativně kvalitní překlad mezi blízkými jazyky pomocí popsané metody vede i k návrhu organizace překladu v případech, kdy z textu v jednom jazyce je třeba vytvořit překlad v mnoha dalších jazycích. To je případ návodů k domácím spotřebičům, příruček k softwarovým systémům, a vůbec ke všem příručkám, které doprovázejí výrobku nebo služby exportované do mnoha různých jazykových oblastí.

Základní schéma je na obr. 11. Z původního jazyka se text přeloží ve vysoké kvalitě (tj. profesionálním překladateli) jen do několika "centrálních" jazyků ("bridge languages"), a z těch se při překladu do jazyků jim blízkých použije automatický překlad (jen s manuální postredakcí).

Obr. 11 Využití "centrálního" jazyka při vícejazyčném překladu

4.7. Výsledky experimentů s překladem do slovenštiny a polštiny, další výhled

Experimenty s úplným systémem překladu z češtiny do slovenštiny jsme prováděli s technickými příručkami pro použití databázového software. Úspěšnost jsme měřili s použitím software pro podporu překladu TRADOS, resp. jeho části která počítá tzv. "match" (souhlas) mezi ručně "dopřeloženou větou" a jejím předchozí variantou (v našem příkladě touž větou přeloženou automaticky). Systém evaluace systému TRADOS je velmi přísný, neboť evaluační systém se v tomto systému používá k určení obtížnosti překladu (obecně platí, že překladatelské firmy účtují podstatně více, je-li shoda s předcházející verzí (tj. v našem případě výsledkem automatického překladu) menší než 90%). Shoda se počítá na základě modifikované Loewensteinovy vzdálenosti (zhruba řečeno, jde o počet editačních zásahů, které je nutno udělat, aby věta byla v "definitivně správné" podobě).

Pro slovenštinu jsme tohoto cíle dosáhli (shoda se pohybovala těsně nad hranicí 90%, pro polštinu jsme však zůstávali na úrovni 75% (na tomtéž textu). Texty použité pro testování byly texty, z nichž byl částečně zapracován překladový slovník, ale např. česká **morfologická analýza** a český značkovač (coby jádro systému) pracovaly na nich nezávisle, tj. testy byly dostatečně realistické a "férové".

Předpokládáme, že systém budeme dále vyvíjet (zejména systém překladu do slovenštiny) jak zvětšováním slovníku, tak i zlepšováním českého značkovače (a morfologie, pochopitelně). Polský systém bude nutno zdokonalit podstatněji, zejména s ohledem na jisté rozdíly v syntaxi - zdá se, že alespoň základní **analýza** jmenných frází bude nutná pro podstatnější vylepšení tohoto systému. Pak by ovšem bylo možno uvažovat i o ruštině, ukrajinštině a dalších jazycích podobně "vzdálených" od češtiny.

5 Závěr

V tomto příspěvku jsme se snažili popsat metody počítačového zpracování dvou klíčových rovin (morfologie a syntaxe) přirozeného jazyka, a také přiblížit možnou aplikaci těchto metod na reálný problém. Ukazuje se, že ač v některých aplikacích je možné použít i analýzu jen částečnou (a to nemluvíme o takových z jazykového hlediska velmi jednoduchých aplikacích, jako je vyhledávání nebo extrakce informací z textu), je jasné, že úplné porozumění vyžaduje analýzu jazyka dost hlubokou. Právě pro tyto účely budujeme jazykové zdroje, jako jsou morfologicky a důkladně syntakticky anotované texty.

Podrobnější informace o budování anotovaných korpusů je možno nalézt na webových stránkách Ústavu aplikované a komputační lingvistiky a Centra Komputační lingvistiky na MFFUK v Praze (http://ufal.mff.cuni.cz a http://ckl.mff.cuni.cz). Pro hlubší studium statistických a pravděpodobnostních metod v lingvistice, které jsou s danou problematikou úzce svázány, lze doporučit zejména publikace (Manning a Schuetze 2001), (Jurafsky a Martin 2000), (Charniak 1998) a (Jelinek 1998). Kompletní materiály k vlastnímu studiu této problematiky jsou umístěny na volně dostupné adrese http://ufal.mff.cuni.cz/hajic/courses/pfl043/0102/syllabus.html. Téměř kompletní bibliografie jak k problematice tvorby anotovaných korpusů, tak jejich zpracování a využití, je pak na již zmíněném CD ROM "Prague Dependency Treebank 1.0" (**Hajič** et al. 2001), ve většině případů s plnými texty článků, příruček a manuálů (a samozřejmě i s kompletními českými korpusy!); kopii dokumentace k tomuto CD je pak možné nalézt i na webu na http://ufal.mff.cuni.cz/pdt.

Literatura

COLLINS, Michael. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the 35th ACL/EACL*. Madrid 1997. s. 16-23.

COLLINS, Michael - **HAJIČ**, Jan - BRILL, Eric - RAMSHAW, Lance - TILLMANN, Christopher. 1998. A Statistical Parser for Czech. In *Proceedings of the 37th ACL*. College Park, MD, USA. s. 505--512.

ČERMÁK, František. 2001. Český národní korpus. Tento sborník.

HAJIČ, Jan. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Hajičová(eds): *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Praha: Karolinum, Charles University Press. s. 12-19.

HAJIČ, Jan. 2001. Disambiguation of Rich Inflection (Computational Morphology of Czech). Praha: Karolinum, Charles University Press.

HAJIČ, Jan - ROSEN, Alexandr - SKOUMALOVÁ, Hana. 1987. RUSLAN - systém strojového překladu z češtiny do ruštiny. *Výzkumná zpráva*. Praha: Výzkumný ústav matematických strojů.

HAJIČ, Jan - PANEVOVÁ, Jarmila - BURÁŇOVÁ, Eva - UREŠOVÁ, Zdeňka - BÉMOVÁ, Alla- ŠTĚPÁNEK, Jan - PAJAS, Petr - KÁRNÍK, Jiří. 1997. Anotace na analytické rovině (manuál pro anotátory). *Technická zpráva TR-1997-03*. Praha: ÚFAL MFF UK.

HAJIČ, Jan - BRILL, Eric - COLLINS, Michael - HLADKÁ, Barbora - JONES, Douglas - KUO, Cynthia - RAMSHAW, Lance - SCHWARTZ, Oren - TILLMANN, Christopher - ZEMAN, Daniel. 1998. Core Natural Language Processing Technology Applicable to Multiple Languages. Research Note 37. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA. **`http://www.clsp.jhu.edu`**.

HAJIČ, Jan - HRIC, Jan - KUBOŇ, Vladislav. 2000. Česílko: Machine Translation Between Closely Related Languages. In *Proceedings of the 6th Applied NLP*, Seattle, WA, USA. ACL / MIT Press. s. 7-12.

HAJIČ, Jan - KRBEC, Pavel - KVĚTOŇ, Pavel - OLIVA, Karel - PETKEVIČ, Vladimir. 2001a. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Proceedings of ACL'01*, Toulouse, France. s. 160-167.

HAJIČ, Jan - HAJIČOVÁ, Eva - PAJAS, Petr - PANEVOVÁ, Jarmila - SGALL, Petr - VIDOVÁ HLADKÁ, Barbora. 2001b. The Prague Dependency Treebank 1.0. CDROM. Philadelphia: Linguistic Data Consortium LDC2001T10. ISBN 1-58563-212-0.

HAJIČOVÁ, Eva - PANEVOVÁ, Jarmila - SGALL, Petr. 2000. Anotace na tektogramatické rovině (manuál pro anotátory). *Technická zpráva TR-2000-09*. Praha: ÚFAL MFF UK.

HLADKÁ, Barbora. 2000. Czech Language Tagging. PhD thesis, Praha: ÚFAL MFF UK.

CHARNIAK, Eugene. 1996. Statistical Language Learning. Cambridge: The MIT Press.

CHURCH, Kenneth. 1992. Current Practice in Part of Speech Tagging and Suggestions for the Future. In Simmons(eds), *Studies in Slavic Philology and Computational Linguistics*: In Honour of Henry Kučera. Michigan Slavic Publications. s. 13-48.

CHYTIL, Michal. 1984. Automaty a gramatiky. Praha: SNTL. Matematický seminář, vol. 19.

JELINEK, Frederick. 1998. Statistical Methods for Speech Recognition. Cambridge: The MIT Press.

JURAFSKY, Daniel - MARTIN, James. 2000. Speech and Language Processing. Prentice-Hall.

KOSKENNIEMI, Kimmo. 1983. Two-level morphology. PhD thesis. Technical reports No. 11. Helsinki: Dept. of Linguistics, University of Helsinki.

MANNING, Christopher - SCHUETZE, Heinrich. 1999. Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press.

MARCUS, Mitch - SANTORINI, Beatrice - Marcinkiewicz M. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2), s. 313-330.

MÍROVSKÝ, Jiří. 1999. Morfologické značkování textu: automatická disambiguace. Mgr. Thesis. Praha: MFF UK.

MOHRI, Mehryar - RILEY, Michael - PEREIRA, Fernando C. N. 1998. A Rational Design for a Weighted Finite-State Transducer Library. *Lecture Notes in Computer Science 1436*. Berlin: Springer Verlag.

PETKEVIČ, Vladimír. 1995. A New Formal Specification of Underlying Representations. In *Theoretical Linguistics*, Vol. 21. s. 7--61

SGALL, Petr - HAJIČOVÁ, Eva - PANEVOVÁ, Jarmila. 1986. The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Prague/Netherlands: Academia/Reidel Publishing Company.

SKOUMALOVÁ, Hana. 1997. Czech lexicon by two-level morphology. In R. Marcinkevičiene and N. Volz (eds): *Proceeedings of the 2nd European Seminar of TELRI -- Language Applications for a Multilingual Europe*. Mannheim/Kaunas: IDS/VSU. s. 123--145.

Poznámky v textu

[1] V tomto příspěvku se omezíme na zpracování textu. Rozpoznávání (a syntéza) **mluvené řeči** je sice ve smyslu "porozumění" jazyku podobný problém, avšak tradičně se soustředí zejména na zpracování akustického signálu, a v jistém smyslu - aspoň z dnešního pohledu, s existujícími aplikacemi a systémy v ruce - se na něj lze dívat jako na přídavný krok, ve kterém nejprve převedeme řečené na text, který dále zpracováváme.

[2] "Slovem" se zde myslí slovo to v tom tvaru, ve kterém se v textu vyskytuje, takže *korunou* a *korunami* jsou dvě různá slova.

[3] Mluvíme-li o jednoznačném určení (zde slovních jednotek, tokens), musíme zároveň říci, jak je toto určení realizováno v textu. K tomu se používají dnes už téměř výhradně tzv. **markup jazyky**, definované na základě standardu **SGML**, který je dnes nahrazován jednodušším a pro počítačové zpracování příhodnějším **XML** (jež je svým způsobem podmnožinou SGML). (Známý jazyk pro popis webových stránek, **HTML**, je rovněž specifikován pomocí SGML.) Zjednodušeně lze říci, že každá značka - zde samozřejmě mluvíme o značce v technickém smyslu, nikoli o značce morfologické - má své jméno, a pro účely rozlišení mezi textem a značkami je jednotně ohraničena symboly '<' a '>'.

[4] Podrobněji o konečných automatech a sekvenčních strojích viz např. (Chytil 1984).

[5] Lze ovšem oprávněně namítnout, že slovo Pekař mělo být morfologickým analyzátorem určeno též jako první pád jednotného čísla rodu mužského životného od vlastního jména Pekař. To je samozřejmě nedostatek slovníku, ovšem jen těžko odhalitelný v plném rozsahu.

[6] Strom je matematicky definován jako souvislý acyklický orientovaný graf s jedním kořenem (tj. vrcholem, do kterého nevede žádná hrana). Obvykle se znázorňuje "vzhůru nohama", tj. kořen se kreslí nahoře, a orientace hran se zachycuje pomocí šipek, které vedou shora dolů, od

řídícího k závislému vrcholu. Z technických důvodů se ovšem v elektronické podobě využívá s výhodou toho, že do každého vrcholu (někdy nazývaného podle angličtiny též "uzlem") vede pouze jedna hrana, a směr závislosti se uchovává obráceně - to však nemá žádný vliv na skutečný směr závislosti.