V předcházejícím rozhovoru jsme se zabývali především otázkou, nakolik je genetický zápis obdobou přirozeného lidského jazyka, který nelze pouze mechanicky („počítačově“) interpretovat, ale jemuž je třeba rozumět. V následující kapitole problém otočíme a podíváme se na něj z úplně jiné strany. Nemohou lidskému jazyku nějak rozumět i počítače? Kromě souvisejících filozofických otázek si podrobněji ukážeme způsoby fungování jazyka a konkrétní metody a systémy, které se s ním snaží automatizovaně pracovat. Zaměříme se hlavně na techniky počítačového překladu.
Vztah mezi biologickým a počítačovým světem se ovšem jako červená nit potáhne i dalšími kapitolami této knihy.

10.
Jak nám počítače rozumějí
Rozhovor s Petrem Strossou

V dobách počátků počítačů zhruba před padesáti lety se všeobecně předpokládalo, že systémy schopné rozumět lidskému jazyku jsou doslova na obzoru. Překladatelé měli brzy přijít o práci. Babylonská rybka (univerzální překladač z románu Douglase Adamse Stopařův průvodce po Galaxii*) však dosud neexistuje, překladové slovníky i další obdobné nástroje mají prozatím jen omezené použití. Jak je možné, že se lidský jazyk počítačovému zpracování tak vytrvale vzpírá?*

Jako reakce se objevil opačný přístup, který lidský jazyk chápal jako něco spíše „esoterického“, co se počítačové logice zcela vzpírá. Automatické zpracování lidského jazyka se však stalo mezitím v mnoha oborech realitou. Došlo k tomu přitom poměrně nenápadně, spíše trpělivou prací lingvistů a programátorů než nějakým radikálním objevem. Na naše otázky o úspěších a mezích počítačového zpracování textu odpovídá RNDr. Petr Strossa, CSc., specialista na počítačovou lingvistiku z katedry informačního a znalostního inženýrství Fakulty informatiky a statistiky pražské VŠE.

Jak vlastně fungují dnešní systémy pro automatický překlad?
Pole počítačového překladu je poměrně široké a vejde se do něho leccos. Optimistické prognózy z poloviny minulého století o tom, že aplikace v dohledné době zvládnou vše, včetně překladů beletrie, se nicméně nepotvrdily. Počítač na poli překladatelských služeb dnes slouží spíše jako pomocný nástroj, který urychluje práci člověka – a vzhledem k tomu, že čas kvalitního překladatele je drahý, představuje přinejmenším nezanedbatelný ekonomický přínos. Každopádně bychom však spíše než o počítačovém překladu měli dnes v mnoha případech hovořit o počítačem podporovaném překladu. Stejně tak je většinou vhodnější vyhýbat se slovu „automatický" a raději hovořit o „poloautomatických" nástrojích, popřípadě nástrojích dílčí automatizace, neboť programy stále vyžadují lidskou asistenci toho či onoho druhu.

Jaký pokrok udělaly překladové systémy od svého vzniku?
Nejjednodušší a historicky také nejstarší jsou systémy označované většinou jako tzv. první generace, překládající víceméně metodou „slovo za slovo". Výsledkem je jen zřídkakdy plně srozumitelný text, pročež se dnes většinou v této souvislosti používá termín indikativní překlad. Ani takový výsledek však nelze považovat za zbytečný. Jistě se dá využít alespoň pro základní orientaci a na jeho základě lze například rozhodnout, zda textu věnovat další pozornost, respektive ho nechat přeložit kvalifikovaným znalcem jazyka.

Úpravou metody překladu „slovo za slovo" je metoda, kterou bychom mohli označit „fráze za frázi". Je zajímavé, že tuto metodu lze dost úspěšně aplikovat všude tam, kde mají překládané dokumenty pevně danou, víceméně formalizovanou strukturu a současně používají limitovanou slovní zásobu. Odborně se tu hovoří o tzv. omezených, popřípadě řízených jazycích. Může jít například o obchodní dopisy nebo o meteorologické či burzovní zpravodajství.

Co následovalo po systémech první generace?
Hlavní trend bychom asi obecně mohli vyjádřit heslem „více gramatiky". Lidé samozřejmě dávno vědí, že když se například v české větě

použije nějaké slovo ve čtvrtém pádě, má věta trochu jiný význam, než když je v ní totéž slovo v prvním pádě a nějaké jiné ve čtvrtém. (Věty „Pavel viděl Petra“ a „Pavla viděl Petr“ říkají každá něco jiného, přestože se skládají ze stejných slov, a dokonce ve stejném pořadí.) Musíme si ovšem uvědomit, že ani gramatická analýza slovního tvaru vytrženého z kontextu nemá obvykle mnoho nadějí na úspěch. Jak se vlastně dá přijít na to, že „Petra“ je čtvrtý pád jména „Petr“ a ne druhý pád téhož jména nebo první pád příslušného ženského jména? Navíc, i kdyby se na to nějak přišlo, přínos samotného rozpoznání pádu pro překlad obecně není příliš velký, protože cílový jazyk může mít jiný systém pádů (třeba němčina má proti českým sedmi čtyři), eventuálně nemusí mít vůbec žádné. Je tedy rozhodně třeba navázat na morfologickou analýzu ještě analýzou syntaktickou – analyzovat stavbu věty jako celku, identifikovat v ní jednotlivá slova jako podmět, přísudek a předmět apod. Teprve na tomto základě má smysl pustit se do vlastního překládání jednotlivých slov (a jejich syntaktické funkce pak musí být vyjádřeny adekvátním způsobem v cílovém jazyce, například anglická věta musí začínat podmětem), ale i tak se mohou vyskytnout různé další problémy.

Jak je to s přeložitelností vlastních slov?

Nejednoznačnost překladových ekvivalentů představuje samozřejmě další problém.

Například ke slovu „poskytovat“ nám dobrý česko-anglický slovník nabídne tyto možné překlady: give, provide, lend, render, grant, allow, afford, accord, extend, accomodate, furnish, supply, yield... Je jasné, že všechny tyto překlady se nehodí do každého daného kontextu, ovšem má-li být posouzení kontextu vůbec kombinatoricky zvládnutelné, je třeba si opět uvědomit, že ve skutečnosti je rozhodující jediný syntaktický vztah (vazba) našeho slovesa ve větě, a to jeho předmět („co je poskytováno“). Tak například „poskytovat zdroje“ se asi přeloží trochu jinak než „poskytovat služby“, zato na tom, kdo, komu a kdy něco poskytuje, nejspíš při překladu slovesa „poskytovat“ příliš nezáleží. (Předmět přitom nemusí být zdaleka vždy slovo následující ihned za slovesem. Věta může třeba znít: „Poskytujeme našim přátelům výhodné půjčky.“)

Zvláště závažný může být problém překladu slov (neboli lexikálního transferu), překládá-li se mezi jazyky výrazně rozdílných kultur. Zde totiž často nejde zdaleka jen o kontext vymezující použitelnost určitých překladových ekvivalentů, ale i o to, že cílový jazyk třeba vůbec nemá pojem bezprostředně odpovídající použitému pojmu vstupního jazyka. Buď zde takový pojem neexistuje vůbec, nebo zde existují pouze pojmy s výrazně odlišnou rozlišovací schopností.

Budeme-li chtít například překládat z češtiny do čínštiny větu, ve které se vyskytuje slovo „strýc", narazíme velmi pravděpodobně na fakt, že v celém dostupném kontextu není žádným způsobem uvedeno, zda jde o otcova staršího bratra, otcova mladšího bratra, matčina bratra, manžela otcovy sestry či manžela matčiny sestry, protože Češi většinou nepovažují za potřebné tuto informaci uvádět. Jenže čínština má pro každou z vyjmenovaných kategorií zvláštní slovo. Naštěstí se dnes přece jen většinou snažíme o překlad mezi jazyky s dost velkým společným kulturním zázemím (v rámci euroamerické civilizace), kde se tento problém nemusí projevovat tak silně.

Nakolik se dá říct, že překladové systémy „rozumějí" textu, se kterým pracují?

V překladových nástrojích druhé generace se zpravidla úloha porozumění omezuje na už zmíněnou syntaktickou analýzu, tj. rozbor určitých druhů vazeb mezi jednotlivými větnými členy, popřípadě hierarchické stavby celé věty. Je to v podstatě něco podobného, co jsme dělali na základní škole při takzvaném větném rozboru.

A další perspektivy?

Někteří odborníci soudí, že budoucnost může patřit snad jedině překladovým systémům založeným na obecných metodách umělé inteligence, jejichž báze znalostí budou obsahovat jak znalosti o jazycích, mezi kterými se překládá (to znamená jejich slovníky, gramatiku a sémantiku), tak znalosti o světě, jehož se týkají překládané texty. Jedině tak lze skutečně modelovat proces porozumění, který zřejmě běžně probíhá v lidské hlavě.

Kdybyste měl uvést několik příkladů, jak se lidský jazyk vzpírá počítačům, co byste vybral?

Jedním z největších problémů je obecně homonymie. Vyřešit všechny případy homonymie v textu je nejsložitějším úkolem každého typu automatického zpracování textu. Homonymní mohou být různě velké úseky textu (různě složité výrazy), bereme-li je samy o sobě; vyřešit homonymii pak zpravidla znamená „celkově" porozumět většímu úseku textu, který obklopuje náš homonymní výraz.

PROBLÉM ZÁVORKOVÁNÍ

Přirozený jazyk se dá přirovnat k „algebře", ve které například výraz A/B/C může znamenat stejně dobře (A/B)/C i A/(B/C), ale co opravdu znamená, závisí mimo jiné na hodnotách A, B a C. Možná by se dalo v analogii pokračovat zhruba v tom smyslu, že „aktuální význam výrazu A/B/C závisí i na důvodu, proč se vlastně tento výraz počítá".

Pro ilustraci dva konkrétní příklady. Výrazu „regulace chlazení termostatem" každý technik rozumí jednoznačně, ale proč? Protože ví, že termostat je nástroj regulace, ale nikoli nástroj, který by sám chladil. V naší algebraické analogii tedy jde o výraz „(regulace/chlazení)/termostatem", nikoli „regulace/(chlazení/termostatem)".

Naproti tomu výraz „pozorování úniku kouře oknem" jednoznačně uzávorkovat nelze, dokud nezjistíme něco víc o situaci, kterou tento výraz popisuje: uniká někde kouř oknem, nebo to někdo oknem pozoruje?

Ukázkovou hříčkou z této kategorie jsou pak „zmatené" věty, ve kterých vůbec nedokážeme od sebe rozpoznat jednotlivé větné členy a slovní druhy. Česká věta „ženu holí stroj" může mít celkem tři naprosto odlišné významy podle toho, které ze tří slov je tu míněno jako sloveso.

Ještě častější jsou podobné bizarnosti v angličtině, která na rozdíl od češtiny prakticky neohýbá slova. Klasickým příkladem je věta „Time flies like an arrow", která může znamenat „Časové mouchy mají rády šíp", „Časuj mouchy jako šíp" i „Čas letí jako šíp". Správně je samozřejmě poslední možnost. K tomu však můžeme dospět pouze za předpokladu, že víme, že:

- ani mouchy ani šípy nejsou slovesy a nemají ani žádné časovací zařízení, takže časovat je nedává smysl;
- mezi mnoha druhy much, pokud je známo, nerozlišujeme žádnou odrůdu much časových.

Jak si má s podobnými hádankami poradit nebohý automatický systém?

Homonymie má samozřejmě různé formy, z nichž každá působí jiný typ problémů – může existovat na úrovni slov nebo celých vět. Například ve větě „Autobus předjíždí tramvaj“ může být jak slovo „autobus“, tak slovo „tramvaj“ interpretováno jako podstatné jméno v prvním nebo čtvrtém pádě. V tomto případě je homonymní celá věta. Kdybychom však převedli přísudkové sloveso do minulého času a vytvořili tak větu „Autobus předjížděl tramvaj“, situace se poněkud změní: obě podstatná jména („autobus“ i „tramvaj“) jsou sice sama o sobě stále stejně homonymní, pokud jde o pád, ale syntaktické pravidlo shody podmětu s přísudkem tentokrát (spolu s faktem, že každé z podstatných jmen je jiného rodu) pomůže určit, že „autobus“ je podmět, a tedy nutně v prvním pádě, zatímco „tramvaj“ jako předmět je nutně v pádě čtvrtém.

Striktně lingvisticky bychom měli rozlišovat mezi skutečnou homonymií (tj. situací, kdy dva různě utvořené výrazy vypadají stejně) a jevem, který se obvykle nazývá polysémie neboli mnohoznačnost jednoho výrazu: slovo „třída“ označuje řadu různých věcí – třeba školní třídu nebo kategorii v teorii množin – ale stále jde o jedno slovo jako jednotku jazykového systému. V počítačovém překladu působí ovšem homonymie i polysémie potíže velmi podobného druhu.

Zkusme teď přejít k dalšímu typu nástrojů pro automatické zpracování textu, k systémům s překladovou pamětí. Je to slepá ulička, nebo naopak způsob, jak mnoho problémů šikovně obejít?

Na první pohled se může zdát, že překladová paměť je metoda velmi primitivní; jde o obyčejnou hrubou sílu využívající rychle rostoucí paměť a výkon současných počítačů. Podíváme-li se ale na věc z trochu jiného úhlu, můžeme si naopak položit otázku, zda tato „nová cesta automatizace překladu“, totiž prostě hledání analogií s něčím, co už máme v paměti, není vlastně věrnějším obrazem přirozeného lidského přístupu k překládání než všechny modely založené na nějakých exaktních gramatikách. Sdělením (alespoň v mateřském jazyce) obvykle nerozumíme na základě jejich analýzy, ale proto, že jsme se už s podobným užitím slov někdy setkali.

Rozvoj systémů s překladovou pamětí umožnila či přímo vyvolala situace, kdy mnohé firmy a instituce už mají ve svých archivech velké objemy starších textů i s jejich překlady do různých jiných jazyků, a přitom nové texty, jejichž překlady jsou zadávány, často neobsahují zase tak velké množství opravdu nových informací. Představme si například příručky a prospekty k určitým výrobkům nebo službám. Výrobky se modernizují, což vyvolává nutnost neustálé aktualizace průvodních tištěných materiálů, nicméně podstatné funkce, které je třeba popsat, zůstávají stejné – a tedy i velká část textů zůstává nezměněná, nebo alespoň téměř nezměněná.

Zde mají systémy s překladovou pamětí své nezastupitelné místo, je ale patrné i jejich omezení. Třeba při překladech z angličtiny do češtiny, tedy jazyka se složitým ohýbáním slov, vyžaduje výsledek těchto systémů ještě lidskou korekturu.

S počítačovými překlady je to tedy stále ještě poněkud ošidné. Co další lingvistické schopnosti současných programů? Asi to nekončí kontrolorem pravopisu v textovém editoru...

Zmínil bych třeba kontrolu stylistickou. Jakkoli se styl (sloh) obvykle považuje za umění nadřazené prosté schopnosti používat gramatiku, ukazuje se, že v automatické korektuře stylistických chyb se kupodivu dá snáze dosáhnout vyšší míry úspěšnosti než v korektuře gramatické. Někdy to dokonce vypadá tak, že právě korektura slohu je skutečnou silnou stránkou nástroje, který je z komerčních důvodů nazýván gramatický korektor.

Je pravda, že automatizovaná stylistická korektura obvykle nezahrnuje takové prvky, jako je například správná věcná návaznost jednotlivých vět, nemluvě třeba o celkové výstavbě textu s logicky odlišitelným úvodem, jádrem sdělení a závěrem. Přesto lze automaticky testovat velké množství jevů, které mohou být označeny jako stylistické chyby a jejichž odstranění je pro kvalitu, čitelnost a čtivost textu velmi užitečné.

Hodně může pomoci už hodnocení jednotlivých slov (popřípadě některých jejich tvarů) vzhledem k zaměření textu. Automatický korektor například může upozornit uživatele, jenž není

úplně zběhlý v jazyce, kterým píše obchodní dopis, že výraz jako „dát si“ je spíše hovorový a do obchodního dopisu nevhodný. K tomu stačí, aby všechna slova ve slovníku byla zařazena do určitých stylových tříd a aby uživatel měl možnost vybrat si z nabídky systému, že to, co píše, má být obchodní dopis (čímž nastaví míru vhodnosti jednotlivých stylových tříd slov). Pokud by korektor navíc zahrnoval vhodně koncipovaný slovník (tezaurus), mohl by i navrhnout náhradu – například spojení „stanovit termín“ místo „dát si termín“.

Poměrně velmi snadná a pro řadu tematických oblastí textů užitečná je rovněž automatická detekce určitých, často používaných víceslovných obratů, kterým by bylo lepší se vyhnout, protože jsou:

• vágní (v podstatě ničím nepřispívají k jádru sdělení – jako třeba „více či méně“);

• zbytečně rozvláčné (lze je úspěšně nahradit jedním slovem – například „vzít v úvahu“, „v případě, že“);

• redundantní (říkají dvakrát totéž – jako třeba spojení „v případě, pokud“).

Vzhledem k tomu, že detekce takových výrazů se patrně musí opírat o tabulku (slovníček), kde jsou tyto výrazy vyjmenovány a klasifikovány, zdá se být vcelku snadné doplnit i funkci nabídky jejich interaktivní opravy: u vágního výrazu může systém nabíd-

SEATTLE ZA LETU K MARSU

Dokážete jednoznačně porozumět větě „Kosmická loď fotografovala Seattle za letu k Marsu“? Pouze z logiky věci tušíme, že tím, co letělo k Marsu, nebyl nejspíš Seattle – města obvykle nikam nelétají. K jednoznačnému porozumění textu je tedy třeba přinejmenším jisté znalosti reálií, jen z gramatiky jazyka to prostě není možné.

Jak by měl program poznat správnou strukturu věty, pokud jí věcně nerozumí? Nabízí se šalamounská odpověď: program správnou strukturu poznat nemusí, ale měl by zjistit, že se zde skrývá přinejmenším potenciální dvojsmysl. Při překladu by pak mohl nalézt podobně dvojsmyslnou formulaci v cílovém jazyce, například anglicky: „The spaceship photographed Seattle flying to Mars“. Vtip spočívá v tom, že pokud je pro čtenáře originál ve skutečnosti jednoznačně srozumitelný, pak bude totéž asi platit i pro takto vytvořený překlad.

nout jeho vypuštění, u rozvláčného nebo redundantního výrazu náhradu jedním slovem, které najde ve své tabulce.

Další poměrně snadno kontrolovatelný prohřešek proti slohové správnosti představuje příliš velká hloubka rozvíjení určitou stále stejnou kategorií doplnění. Podstata prohřešku je, alespoň předpokládám, vidět právě ve výše uvedené definici.

Styl textu mohou nepřímo pomoci vylepšit jeho nejrůznější statistické analýzy. Uživatel může být například upozorněn, že velmi často používá určité slovo na začátku věty. Označovány mohou být také jednotlivé – z hlediska určeného stylu – nadprůměrně dlouhé věty.

Čím se zabývají počítačoví lingvisté orientující se primárně na češtinu?

Jednou věcí, kterou je asi třeba se zabývat neustále, je upozorňování různých jiných odborníků (například programátorů všech možných systémů), že nějaké problémy související s češtinou vůbec existují. Vezměme si například abecední řazení různých seznamů. Je pravda, že česká norma abecedního řazení patří suverénně k nejsložitějším na světě a některá její pravidla by si možná v souvislosti s potřebami počítačového zpracování dat zasloužila trochu provětrat. Nicméně dokonalé nebo téměř dokonalé počítačové implementace této normy už existují. Přesto se stále ještě každou chvíli potýkáme s různými aplikačními programy generujícími seznamy jako například „Cepl, Choděra, Cileček...“

Stejně tak je zřejmě stále poměrně dost málo zažitý fakt, že ohýbání slov, jakým disponuje čeština, poměrně značně komplikuje aplikaci obecně známých metod indexování a vyhledávání textů (většinou pocházejících z anglicky mluvící části světa). Představte si, že chcete například pomocí některého z celé řady českých plnotextových (fulltextových) vyhledávačů nabízejících se na internetu najít něco na téma „sběrný dvůr“. Můžete být úspěšní, může ovšem nastat situace, kdy žádné takové stránky nenajdete... Pokud je tomu tak prostě proto, že žádné takové stránky neexistují, samo o sobě by to stále ještě nebyl důvod k přemýšlení. Možná ale zato existují stránky jiné, ve kterých se v nějaké souvis-

losti zmiňují „sběrné dvory“, komentuje se „umístění sběrných dvorů“, jen tak mimochodem se sděluje, čeho byste se měli zbavovat „ve sběrném dvoře“ – a podobně. Všechny takové texty by pro vás mohly být zajímavé – a koneckonců právě od toho je tu plnotextový vyhledávač, aby vám našel to, co hledáte, ať už je to v textech dokumentů schováno kdekoliv.

Jak řešit právě popsaný problém? Klasický nástroj, nabízený snad každým vyhledávacím strojem či databázovým systémem, operátor pravostranného rozšíření vyhledávaného výrazu, vám tady příliš nepomůže: pokud chcete najít všechny tvary slova „dvůr“, museli byste zadat něco jako „dv*“ (kde hvězdička vyjadřuje operátor rozšíření) a pak byste vyhledávali mimo jiné všechny texty, kde se zmiňují třeba „dveře“, „dvojčata“, „dvojice“ nebo „dva“.

Navíc v systémech indexujících skutečně obrovské objemy textů za tím účelem, aby uživatel nalézal pokud možno nejrelevantnější dokumenty ke svému požadavku, nejde obvykle jen o to, zda se v textu vyskytuje hledaný výraz, ale také o to, jak významně se tam vyskytuje – což mimo jiné zahrnuje otázku, kolikrát se tam vyskytuje. Odpovědět opravdu smysluplně na tento dotaz znamená dokázat zjistit, že se v textu například vyskytuje třikrát slovo „dvůr“, bez ohledu na to, v jakém pádě a čísle, a ne jenom, že je tam jedenkrát řetězec znaků „dvory“, jedenkrát „dvorů“ a jedenkrát „dvoře“. K tomu slouží nástroj nazývaný lematizátor.

BEZ HACKU A CAREK

Víceznačnost jazyka výrazně narůstá v 7bitové češtině (tedy bez použití diakritiky). Vezměte si například větu zapsanou bez diakritických znamének: „Je rada dolu, kde se tezi med.“ Intuitivně můžeme usoudit, že tato věta je jako celek (skoro) jednoznačná. Pokud ji však budeme chtít analyzovat zdola nahoru, tj. počínaje jednotlivými slovy, zjistíme, že za předpokladu potenciálních diakritik nad libovolnými písmeny je pět ze sedmi slov nejednoznačných.

Dnes už tento problém není tak aktuální jako ještě před několika lety, v elektronické poště i na webových stránkách se stále častěji používá čeština včetně diakritiky.

Na jaké hlavní problémy lze při tvorbě českého lematizátoru narazit a jak se řeší?

To jsme v podstatě právě ukázali. Je sice pravda, že například česká podstatná jména se většinou skloňují prostě pomocí pádových koncovek, takže například lematizovat slovo skloňované standardním způsobem podle vzoru „hrad" znamená pouze odstranit kteroukoli z možných koncovek, ovšem toto základní schéma má řadu modifikací, kde jednotlivé koncovky různě zasahují do části slova vlevo od nich, tzv. kmene slova. Například slovní tvar „kly" je třeba lematizovat na „kel", a podobně „dvoře" na „dvůr". Hlavní problém přitom je, že zde nelze dost dobře formulovat obecně platná pravidla.

Další potíž spočívá opět v homonymii, v tomto případě zejména různých kombinací kmenů a koncovek. Například je-li někde v textu výraz „v tancích", pak bez analýzy kontextu (ale jak širokého a pomocí jakých nástrojů?!) nelze určit, zda je tématem sdělení „tanec", anebo „tank".

Nástroje řešení těchto problémů mohou být různé, v každém případě však lze říci jedno: nejsou právě levné. Jeden z dílčích nástrojů obecně hodných doporučení je například slovník všech slov použitelných v daném jazyce se zakódovanými přesnými vzory ohýbání. Protože se v češtině navíc ohýbají i skoro všechna cizí vlastní jména, je v praxi třeba mimo jiné takový slovník neustále doplňovat, a pokud má být skutečně relativně univerzální, je nutno počítat s rozsahem blížícím se miliónu slov.

Jaký je vlastně hlavní rozdíl mezi přirozeným a formálním jazykem?

Mám-li odpovědět přijatelně stručně, tak hlavní rozdíl spočívá v tom, že přirozený jazyk se vyvinul jakýmsi ne zcela uvědomělým procesem, během kterého získal jisté vlastnosti, které „fungují" (tj. lidé užívající ten jazyk je respektují, většinou dokonce aniž na to myslí), ale které se teprve ex post jazykovědci snaží popsat v jeho gramatice. (Což se jim daří někdy lépe, někdy hůře, a když se nějaký jev vytrvale vzpírá jejich exaktnímu uchopení, máme v gramatikách pravidla, která ani v řeči ani v písmu nikdo nerespektuje.)

Umělé, formální jazyky mají naopak zpravidla předem danou gramatiku, a protože jsou často vymýšleny přímo s představou počítačového zpracování (třeba takový programovací jazyk by asi jinak ani neměl vůbec smysl, že), jejich autoři cílevědomě směřují k tomu, aby jejich gramatika byla v jistém smyslu snadno počítačově uchopitelná. Z toho vznikla slavná Chomského hierarchie formálních jazyků a odpovídajících typů automatů použitelných k jejich rozpoznávání (tj. rozhodnutí, zda nějaký řetězec znaků patří nebo nepatří do daného jazyka), resp. k analýze struktury daného řetězce znaků (neboli jak byl ten řetězec vytvořen) podle dané gramatiky. Samozřejmě, jakmile byla tato teorie na světě, a možná dokonce už o něco dříve, odborníky zajímala otázka, dají-li se do jejích kategorií nějak rozumně umístit i přirozené jazyky.

Myslím, že nemá příliš smysl, abychom tady podrobněji rozebírali, co je to „jazyk typu 0“, „bezkontextová gramatika“ nebo „konečný automat“. Důležité je spíš to, co z toho vyplynulo pro praxi počítačového zpracování lidského jazyka. Je víceméně dokázáno (k tomu, co to znamená, se ještě vrátím), že obvyklým způsobem strukturované věty přirozených jazyků by mělo být možné syntakticky analyzovat s časovou i paměťovou náročností nanejvýš kubicky závislou na délce vět. Jakými konkrétními nástroji by to bylo nejlepší opravdu dělat, to je téma na celé knihy, a snad na to ani neexistuje jednoznačná odpověď...

Použil jsem výraz „je víceméně dokázáno“. To byla, přestože to tak možná nevypadá, pečlivě uvážená volba vyjádření. Přirozené jazyky se totiž liší od formálních jazyků mimo jiné i tím, že u nich často nelze jednoznačně rozhodnout, co všechno do daného jazyka ještě patří (co je v něm ještě správně) a co už ne. Navíc se i ta místy zamlžená kritéria správnosti, která objektivně fungují v určitém okamžiku, neustále vyvíjejí – i když obvykle velmi pomalu. Mentální struktury, které ve skutečnosti pomáhají lidem správně mluvit a rozumět tomu, co říkají jiní, jsou evidentně poměrně značně pružné – ale to je asi tak všechno, co o nich dnes dokážeme seriózně říct.

Co je co?

DERIVACE – obecně odvození, odvozování. V počítačovém zpracování textů opak lematizace, tedy vygenerování všech možných tvarů (popřípadě odvozenin) slova z jeho základní slovníkové podoby. Programový nástroj, který tuto operaci provádí, se nazývá derivátor.

HOMONYMIE – jev spočívající v tom, že určité dva různé slovní tvary, popřípadě celé jazykové konstrukce, znějí stejně, takže je nelze „na první pohled/poslech" od sebe odlišit. Příbuzným jevem je tzv. polysémie.

LEMATIZACE – proces, kterým je slovnímu tvaru přiřazen jeho základní („slovníkový") tvar (například 1. pád jednotného čísla, pokud jde o podstatné jméno v češtině). Nástroj, který provádí příslušnou operaci, se nazývá lematizátor.

MORFOLOGIE – domácím slovem tvarosloví. Část gramatiky zabývající se ohýbáním slov a jejich odvozováním z jiných slov pomocí předpon, přípon apod. V morfologii češtiny se například zkoumají způsoby, jakými se podstatná a přídavná jména, zájmena a číslovky skloňují, přídavná jména a příslovce stupňují a slovesa časují, a dále způsoby, jakými se například od podstatných jmen odvozují přídavná jména a slovesa (konkrétně například „škola – školní – školit").

Úkolem morfologické analýzy textu při jeho automatickém zpracování je přiřadit každému slovu textu jeho slovní druh, základní („slovníkový") tvar a informace o tom, v jakém tvaru se nachází v daném místě textu (třebas pád a číslo podstatného jména).

OMEZENÝ (ŘÍZENÝ) JAZYK – specifická podmnožina jazyka s limitovanou slovní zásobou i arzenálem typů obratů. V případě omezeného jazyka jde o omezení objektivně existující, přirozeně plynoucí z omezeného použití, jako například v jazyce meteorologického zpravodajství nebo v jazyce elementární obchodní korespondence. O řízeném jazyce mluvíme tehdy, když určitá omezení uměle stanovíme. Takový přístup se dnes často uplatňuje například při tvorbě uživatelských příruček k různým výrobkům.

PŘEKLADOVÁ PAMĚŤ – potřebujeme-li přeložit nový dokument do určitého jazyka a máme-li k dispozici do jisté míry podobný starší dokument i s jeho překladem, vyznačíme v novém dokumentu odlišnosti od staršího a na odpovídajících místech upravíme starý překlad. Překladová paměť není vlastně nic víc než schopnost zjistit, že určitá část textu byla už jednou přeložena určitým způsobem.

SÉMANTIKA – nauka o významu jazykových výrazů, tedy slov, sousloví, frází i celých vět, potažmo souvětí. Na rozdíl od pragmatiky se sémantika zabývá pouze významem plynoucím ze samotného jazykového systému, tedy z všeobecných zvyklostí užívání slov a gramatických konstrukcí, nikoli významem v různých konkrétních situacích.

SYNTAX – česky řečeno skladba. Část gramatiky zabývající se způsoby, jakými se z jednotlivých slovních tvarů sestavují sousloví, fráze a věty.

Úkolem syntaktické analýzy textu je přiřadit větám (či souvětím) jejich syntaktické struktury neboli označit, jak jsou sestaveny ze skladebných dvojic (například „toto je podmět, toto přísudek a toto přívlastek k podmětu"; v počítačové analýze je ovšem většinou užitečné pracovat s ještě trochu jemnějšími syntaktickými kategoriemi, než na jaké jsme zvyklí z našich základních a středních škol).

Při hledání odpovědi na otázku, jak naučit počítače rozumět lidskému jazyku, jsme (už poněkolikáté v této knize) zavadili o umělou inteligenci. V následujícím textu zaměříme svoji pozornost právě tímto směrem.

11.
Cesty k umělé inteligenci: stroje, testy a zombie
Rozhovor s Jaroslavem Peregrinem

Umělá inteligence je na jednu stranu oborem zcela praktickým, který zahrnuje třeba robotiku nebo tvorbu expertních systémů. Nám však půjde spíše o teoretičtější aspekt věci, respektive obecnou rovinu problému. Kdy můžeme nějaký systém prohlásit za inteligentní?

Soustředíme se na jeden z možných přístupů k této otázce, který představuje Turingův test. Na téma se zkusíme podívat ještě z několika dalších úhlů, na scénu tedy vstoupí Turingovy stroje, Gödelovy věty, paradox Čínského pokoje či všemožné druhy zombií. Na naše otázky odpovídá prof. RNDr. Jaroslav Peregrin, CSc., matematik, vědecký pracovník Filozofického ústavu AV ČR a vedoucí katedry logiky na pražské FF UK. Zabývá se především sémantikou, analytickou filozofií a filozofií logiky.

Jak se díváte na Turingův test po zhruba padesáti letech od chvíle, kdy byl poprvé navržen?

Myslím, že Turingův test rozhodně ani dnes neztrácí na zajímavosti. Pozoruhodné je, že na jedné straně existuje celá řada vědců a filozofů, kteří ho považují za příliš „měkký" – připsat počítači myšlení jen na základě toho, že tímto testem projde, považují za absurdní. Na straně druhé se naše počítače, navzdory fantastickému technologickému pokroku, který se během padesáti let od Turingovy formulace tohoto problému odehrál, ani zdaleka nepřibližují něčemu, co by toho bylo schopno.