Patrice Pognan
PLIDAM - INALCO[1], Paříž
ÚFAL - MFF UK[2], Praha
patrice.pognan@mff.cuni.cz

УДК 811.162:811.162.3
https://doi.org/10.18485/slavistika.2018.22.1.9
оригинални научни рад
примљено 07.03.2018.
прихваћено за штампу 14.05.2018.

VYPOČÍTATELNOST ČEŠTINY VE SROVNÁNÍ S OSTATNÍMI ZÁPADOSLOVANSKÝMI JAZYKY

Cílem této práce je dokázat, že češtinu lze automaticky analyzovat beze slovníku na základě lingvistických prvků a jejich pravidelnosti. V takovém systému je nejdůležitější modul morfologické analýzy, který je mimo jiné založen na poznávání cizích slov řecko-latinského původu a hlavně na poznávání přídavných jmen. Dále porovnáváme vypočítatelnost češtiny s vypočítatelností jiných západních i jihozápadních slovanských jazyků.

Po zkoumání vypočítatelnosti češtiny i dalších západoslovanských jazyků se výhodně uplatnila myšlenka obrátit pozornost na fonologický vývoj této skupiny jazyků. Je pozoruhodné, že pravidelnost fonologických jevů je značně vyšší než se dalo očekávat.

Klíčová slova: vypočítatelnost, automatická analýza češtiny, automatické generování slovanských lexémů.

Our purpose is to show that automatic analysis of Czech without referencing a dictionary and simply on the basis of linguistic elements and their regularity is possible. In such a system, the module of morphological analysis is most important. It includes in particular the recognition of neo-Latin loan words and the recognition of adjectives. We then compare the calculability of the Czech language to that of other western and southwestern Slavic languages.

We also study the diachronic phonology of western Slavic languages. It is surprising to notice that the regularity of historical phonological phenomena is significantly higher than expected.

Keywords: calculability, automatic analysis of Czech language, automatic generation of Slavic words.

Češi s oblibou věnují velkou pozornost reáliím. Tento zájem je podporován hojnými publikacemi o přírodě, atlasy hub, motýlů, brouků či ploštic... Konkurují jim jedině encyklopedie o hradech a zámcích a různé knihy o češtině - všední i nevšední - v literární nebo v obecné podobě. Proto je tento článek se zvláštním pohledem cizince na češtinu věnován té části české veřejnosti, která chce svůj vlastní jazyk znát lépe. Zdá se, že určité znalosti češtiny zmizely z popředí kolektivního vědomí, jak dokazuje již citát Dobrovského, který uvedu dále. Na druhé straně nové metody lingvistiky (a hlavně práce s počítačem) přinášejí nové poznatky.

Dlouhodobá zkušenost v oboru automatické analýzy jazyků a zvláště na úrovni fonologie a morfologie umožnila objevení zajímavých rysů češtiny. V první řadě je to její mimořádná pravidelnost. Toto poznání není založeno jen na vlastní práci a na dlouhých průzkumech mluvnic a slovníků, ale také na mnohaleté spolupráci s pražskou skupinou P. Sgalla, E. Hajičové a J. Panevové. Zvlášť se osvědčila spolupráce na Mozaice se Zdeňkem Kirschnerem. Tato spolupráce přinesla další zkušenosti, které jasně ukazují na pravidelnost systému. Jedná se především o:

- automatickou vypočítatelnost češtiny v synchronii (automatická analýza)
- srovnávací synchronní vypočítatelnost západoslovanských jazyků

[1] Institut National des Langues et Civilisations Orientales.
[2] Matematicko-Fyzikální Fakulta Univerzity Karlovy.

- výuku historické mluvnice češtiny v kontextu západoslovanských jazyků
- určení západoslovanského lingvistického systému.

Lingvistický systém západoslovanských jazyků (sever a jih):

Lingvistický systém chápeme jako souhrn všech synchronních i diachronních jevů dané skupiny příbuzných jazyků.

Lingvistický systém je možné zobrazit různými způsoby. Matice je vhodná pro didaktické účely, databáze pro lingvistický výzkum, především pro lexikografii. Databáze se různým potřebným strukturám přizpůsobuje nejlépe. Do naší databáze západoslovanských jazyků se bez jakýchkoliv úprav vešly slovinština a chorvatština.

I přes slabší znalost těchto dvou jazyků jsem tento fakt tušil opíraje se např. o Starostinovu klasifikační hypotézu[3]:

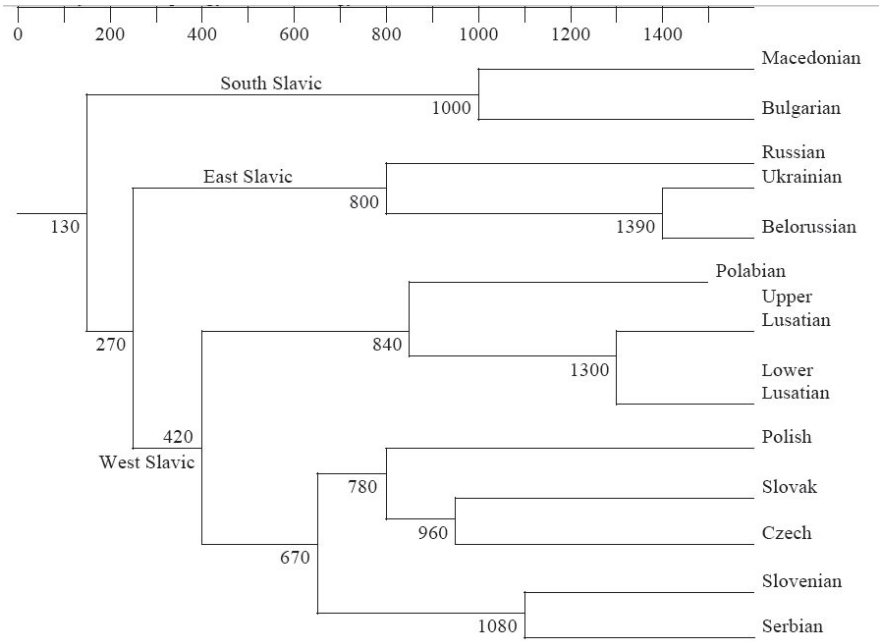

Fig. 1 - Klasikace slovanských jazyků podle Starostina.

Lze dodat, že ani pohled Meilleta na slovanské jazyky není v rozporu s touto hypotézou[4].

[3] cituje Blažek, V. « On the internal classification of Indo-European languages: survey », Linguistica online, 2005.

[4] « Il y a 3 groupes. Le groupe occidental : tchèque, sorabe, polabe et polonais, est net et se reconnaît à toute une série de particularités communes. Le grand et le petit russe forment une unité plus nette encore et font l'effet de deux formes d'un seul et même dialecte. Les parlers méridionaux dits bulgares, serbo-croates et slovènes ***ont en commun peu d'innovations identiques tout à***

Algoritmy morfologické analýzy slovanských jazyků a hlavně češtiny jsou založeny na automatickém rozpoznávání lingvistických forem. U takového postupu je největší překážkou víceznačnost. Proto hledám všechny možnosti, abych ji odstranil.

Je třeba vyhledávat jevy, které jsou stoprocentně jisté (třeba s tím, že bude malý seznam výjimek). Na základě těchto výsledků, pokud jsou bez „šumu“, totiž bez chyb, se mohou vydedukovat i další a další.

Proto také dělíme lingvistický systém na dvě části, na nominální část a na verbální část (jak se to dělá v tradici semitských mluvnic), což už značně redukuje víceznačnost.

Mezi všeslovanské rysy patří vid a dělení hlásek na tvrdé a měkké.

Pro verbální část lingvistického systému je příznačná přítomnost nesložených minulých časů (imperfektum a aorist (lužičtina, srbochorvatština)), složeného budoucího času s příčestím činným (slovinština i polština) i supina (dolní lužičtina, slovinština).

V počtu zajímavých rysů nominální části se nachází duál, souhrn sedmi pádů, dělení maskulina na 3 podtypy (mužský rod životný osobní, životný a neživotný) a zajímavé soužití dlouhých i krátkých (jmenných) koncovek u přídavných jmen.

Oba typy koncovek existují v obou skupinách západoslovanských jazyků (severních i jižních). V jižních jazycích převládají krátké formy (jmenné formy), v severních dlouhé formy[5]. Tím, že si čeština a slovenština zachovaly délku ve psaní, jsou dlouhé formy pro vypočítatelnost zvlášť výhodné (viz níže). V češtině doslova „otevírají“ přístup k automatické analýze na základě rozpoznávání forem.

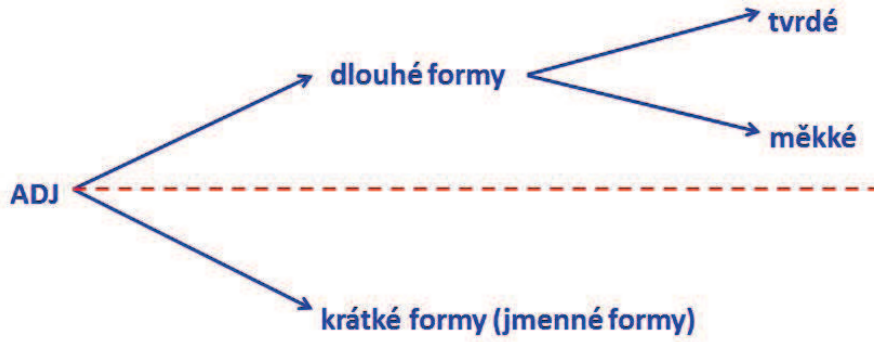

Fig. 2 - Koncovky přídavných jmen ve slovanských jazycích.

fait caractéristiques, mais se laissent **aussi *grouper ensemble sans violence*** ». Meillet: Le slave commun, introduction, p. 3. Honoré Champion, Paris, 1965.

«Existují 3 skupiny. Západní skupina (čeština, lužičtina, polabština a polština) je jednoznačně vyhraněná a vyznačuje se celou řadou společných rysů. Velkoruština a maloruština tvoří ještě silnější jednotu a vypadají jako dvě podoby téhož dialektu. Jižní dialekty označované jako bulharština, srbochorvatština a slovinština zcela příznačné společné novinky téměř postrádají. Lze je však také nenásilně seskupit».

[5] V češtině známe krátké formy ve funkci přísudku „jsem hotov, jsi hotova? Je nemocen“. Toto použití je řídké. To, co je důležité, je „druhotné“ použití v situaci, kde už neexistuje jiná forma než střední rod v singuláru typu „sucho“ („such“ a „sucha“ chybějí). Tato forma může pak nabývat funkce příslovce a dokonce podstatného jména. V češtině se rozmáhá zhoubný vliv aglutinace, která jazykový systém ničí. Místo „na sucho“ a „po suchu“ se příliš často píše „nasucho“ a „posuchu“. Tyto formy se pak nenávratně propadají do hlubin přísloví.

Pravidelnost češtiny je mnohem větší, než si Češi zpravidla myslí. Čeština dokonce zaujímá ve skupině západoslovanských jazyků zvlášť výhodnou pozici oproti ostatním příbuzným jazykům.

Této pravidelnosti je právě možno využít pro zpracování české morfologie. To znamená, že se dají určit morfosyntaktické hodnoty podle formy (automatické rozpoznávání forem). To je to, čemu říkám vypočítatelnost[6].

Cílem mé práce je dokázat, že češtinu lze automaticky analyzovat beze slovníku na základě lingvistických prvků a jejich pravidelnosti. V takovém systému je nejdůležitější modul morfologické analýzy, který je mimo jiné založen na poznávání cizích slov řecko-latinského původu (rozlišení cizích a domácích slov je pro tento druh analýzy velmi důležité) a hlavně na poznávání přídavných jmen. Tento poslední fakt je spojený s tím, že přídavné jméno má vždy „dlouhou“ koncovku, to znamená koncovku s dlouhou samohláskou nebo s dvojhláskou „ou“. Délka je nepochybně nejdůležitější prvek vypočítatelnosti.

Struktura modulu automatické morfologické analýzy

Automatické rozpoznávání cizích slov (řecko-latinský původ)
umožňuje « výpočet » přídavného jména odvozeného od substantiva na -ičnost
Seznam předložek a spojek
Substantiva na -ost
umožňují automatické nalezení kořenů přídavných jmen
Podstatná jména slovesná
Tvrdá přídavná jména
+ odvozená od příčestí trpných
+ přídavná jména na -telný
Měkká přídavná jména
+ odvozená od přechodníků
+ přídavná jména účelová
Substantivní odvozování (použití přípon)
např. přípony označující místo.

Fig. 3 - Struktura modulu morfologické analýzy

[6] Slovo „vypočítatelnost“ jsem odvodil od slov „vypočíst“ a „výpočet“, což mi vyhovuje mnohem víc než název „předvídatelnost“, jenž stejnou úroveň vědecké přesnosti sémanticky neobsahuje.

Automatické rozpoznávání cizích slov:

I letmý popis modulu automatického rozpoznávání cizích slov umožní zachytit jeden z hlavních principů přístupu beze slovníku:

jednak se používají „pozitivní“ pravidla určená na vyhledávání všeho, co odpovídá dané otázce

jednak se používají „negativní“ pravidla, jež mají určit všechno, co bezpečně nemůže být to, co se hledá.

Použití „negativních“ pravidel značně redukuje procento nejistoty při vyhledávání. Axiomatická zásada takového postupu je nepřítomnost „šumu“, což znamená, že výsledky nesmějí obsahovat sebemenší chybu. Díky tomu se výsledky mohou stát základem dalších odvozených výpočtů, třeba na úrovni odvozování.

V případě modulu cizích slov učinkují pozitivně na příklad:

cizí grafémy

„g“: inte**g**rovaný, embar**g**o

„ó“: t**ó**n, m**ó**da, cirh**ó**za, myk**ó**za, skler**ó**za, vir**ó**za

„f“: **f**inále, eu**f**onie

„x“, „q“, „w“: e**x**il, a**x**ióm

dvojhlásky

všechny diftongy kromě „ou“: **eu**fon**ie**

samohlásky na začátku slova

„a“: **a**kce, **a**xióm, **a**plikace (ve všech slovanských jazycích je jen pár domácích gramatických slov začínajících na „a-“. V češtině jsou to slova jako „a“, „ať“, „asi“, „ale“, „aby“, ...“avšak“, „alespoň“, ... Všechna ostatni slova jsou cizího původu)

„e“: **e**ufonie, **e**xil, **e**mbargo

„i“: **i**ntegrovaný

zachování grafické formy (francouzských) nosovek

po (i zjednodušené) morfématické analýze platí

vzor „samohláska – N – souhláska“: **int**egrovaný, k**onst**an**t**a

vzor „samohláska – „M“ – {B/P/F}“: **emb**argo, l**ymf**a, n**ymf**a

použití cizích předpon, přípon a segmentů.

Účinnost těchto postupů je daná tím, že většinou každé slovo obsahuje více příznačných jevů:

Na druhé straně je možno aplikovat negativní pravidla. Čeština má na příklad dva zákony o měkčení, kde je důležitá posloupnost souhlásky a samohlásky (opačná

posloupnost nemá žádný význam). Tvrdý prvek se vždy mění na měkký, ať je to samohláska nebo souhláska:

1. $C_D + V_D \Rightarrow C_D + V_D$
2. $C_M + V_M \Rightarrow C_M + V_M$
3. $C_D + V_M \Rightarrow C_M + V_M$ palatalizace
4. $C_M + V_D \Rightarrow C_M + V_M$

První zákon (pravidlo 3) označuje změkčení tvrdé souhlásky před měkkou samohláskou. Jedná se o tzv.palatalizaci. Negativní použití zákona, to jest zjištění, že k změkčení podle pravidla zrovna nedošlo, umožňuje určení zajímavé hodnoty: neznalost domorodých zákonů dokazuje, že se jedná o „cizince“, totiž o slovo cizího původu. Ku příkladu slova „**r**ezervace“, „**h**isto**r**ie“ a „ko**k**etovat“ jsou označena jako cizího původu z toho důvodu, že písmeno „r“, písmena „h“ a „r“ a konečně druhé „k“ ve slově „koketovat“ nejsou změkčena. Člověk si těžko uvědomuje, že kdyby byla ta slova prošla českou fonologickou evolucí, zněla by jako: „řezervace“, „zistoří“ a „kocetovat“!

Druhý a méně známý zákon (pravidlo 4) označuje měkčení v opačném případě: změkčení tvrdé samohlásky po měkké souhlásce. Negativní použití zákona umožňuje správnou analýzu určitých slov. Slovo „c**u**kr“ je dobře analyzováno jako cizí slovo, i když má stejnou šablonu jako „řekl“, „pekl“, „tekl“,... neboť v dnešní češtině nemůže být tvrdá samohláska po měkké souhlásce. Vzhledem k tomu, že je toto slovo cizího původu, nezměnilo se podle českých historických fonologických pravidel „u“ na „i“.

Negativního přístupu se dá také používat v případě obojetných souhlásek. Vychází to z faktu, že kdysi bývaly tyto souhlásky měkké. Jinde než v omezeném seznamu vybraných slov se obojetné souhlásky chovají i dnes jako měkké. Takže výskyt obojetné souhlásky s „y“ místo „i“ označuje cizí původ slova : „p**y**roceram“, „s**y**stem“ nebo „s**y**stém“...

Česká morfologie

Čeština je flexivní jazyk, což by mělo znamenat, že se bude funkce slov nacházet v koncovkách. Na úrovni morfologie je naším největším přínosem poznatek, že se morfologické hodnoty mohou vyskytnout nejen na konci slova, ale také na začátku v pozici první předpony a, což je mnohem překvapivější, uprostřed slova jako samohláska kořene vázaná s určitou koncovkou :

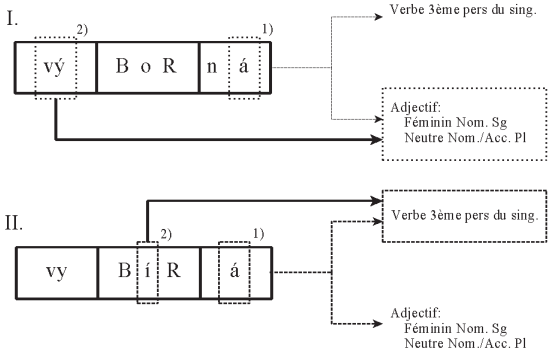

Fig. 4 - morfologické hodnoty jsou všude ve slově

Proč se čeština hodí k automatickému rozboru podle formy?

Historická mluvnice češtiny nás informuje o přirozeném fonologickém vývoji jazyka a o lidských zásazích.

Existují tři hlavní momenty lidských zásahů:

nejdřív vytvoření hlaholice Konstantinem

pak zavedení diakritického pravopisu Husem

a konečně vliv obrození.

V hlaholici, která je zřejmě prvním písmem Slovanů, si fonémy a grafémy jednoznačně odpovídají. Tato vlastnost v podstatě zůstala ve všech jazycích psaných cyrilicí.

Zavedením Husova diakritického pravopisu se čeština k této původní vlastnosti vrátila (až na výjimku „ch“ a neuznávaného „dž“, které našlo právo azylu teprve ve slovenštině) a nabyla fonologického rázu a to mnohem dříve než třeba korejština a turečtina (obě taky pod vlivem lidského zásahu).

Obrození mělo mimo jiné kladný vliv na obnovení několika přípon a na jejich sémantické okleštění. Barešova práce (Bareš 1970) ukazuje, že na příklad původní množina slov, jež končila na „krátkou samohlásku - dlo“ byla sémanticky okleštěna: slova jako „přístavidlo“ přešla do množiny zakončené na „-iště“ a změnila vlastní formu na „přístaviště“, slova typu „kružidlo“ se změnila na slova typu „kružítko“ a slova jako „hnojidlo“ na slova typu „hnojivo“. Z toho vyplývá, že dnes množina končící na „-dlo“ sémanticky označuje stroje, které dělají to, co vyjadřuje kmen: „čerpadlo“ je stroj, který čerpá, „vozidlo“ stroj, který vozí... Existuje pár výjimek: ojedinělá slova jako „divadlo“, „umývadlo“, a jinak slova (zhruba 10 až 12), která označují části těla živočichů a která jsou většinou pomnožná: „mluvidla“, „rodidla“, „tykadla“, „kusadla“, „makadla“, „chapadlo“, „chodidlo“,...

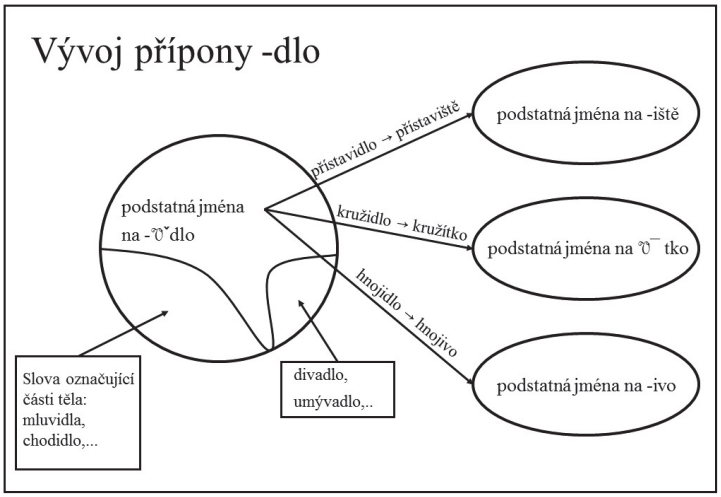

Fig. 5 - vývoj přípony -dlo

Shodou okolností měla také čeština přirozený fonologický vývoj, který nám teď poskytuje výhodný podklad pro automatickou analýzu beze slovníku. Zvlášť důležité jsou tyto jevy:

metateze likvid

změna „g“ ⇨ „h“

stahování (kontrakce)

změna dvojhlásky v jednu hlásku (monoftongizace):

změna „ie“ ⇨ „í“

změna („ó“ ⇨) „uo“/ „uó“ ⇨ „ů“

V češtině je délka hojná díky stahování a monoftongizaci. Připomeňme, že přídavná jména (kromě jmenných forem) mají vždy dlouhou koncovku. Má ji také značná část sloves.

Délka umožňuje uplatnění algoritmu morfologické analýzy, který nachází morfosyntakické hodnoty nejen na konci slova tam, kde se to očekává, ale také na začátku slova, kde předpony s dlouhou samohláskou označují (kromě zhruba 10 až 12 výjimek, např. „záviset“) neverbální kategorie: slova obsahující takovou předponu (pokud je první předpona nalevo nebo je jediná) nemohou být slovesa. Ještě mnohem překvapivější je, že samohláska „í“ v kořeni ve spojení s koncovkou „-á“ jednoznačně ukazuje na nedokonavé sloveso páté třídy (zatím jsem našel pouze tři výjimky) [viz Pognan, 1999]. Slovo „výborná“ dostane kategorii přídavného jména díky předponě „vý“, která označuje, že se o sloveso jednat nemůže. Zato slovo „vybírá“ dostane gramatém slovesa díky „í“ v kořeni „BíR“ ve spojení s koncovkou „á“.

Vypočítatelnost v západních slovanských jazycích:

Vypracování srovnávacího systému vypočítatelnosti vychází ze zkušeností automatické analýzy češtiny a slovenštiny. Vzorem bude Marvanův systém pro stahování (Marvan 2000). Prozatím uvádíme pouze nástin potřebných parametrů označených hodnotou od 0 do 1.

Délka

čeština	1	ý / y
slovenština	0,5	rytmický zákon: fekálna cisterna pravidlo ⇨ pravidlá
polština	0	y / y
dolní lužičtina		
horní lužičtina		

Fig. 6 - délka

Díky délce mohou být určité koncovky plně jednoznačné. Na příklad koncovka „ý“ označuje vždy přídavné jméno mužského rodu v prvním pádě singuláru, případně ve čvrtém ve spojení s neživotným podstatným jménem. Jazyky, které délku nemají, takovou rozlišovací hodnotu postrádají (polština, dolní a horní lužičtiny).

Slovenština má také délku, ale dlouhá koncovka u podstatných jmen středního rodu tvrdého vzoru v plurálu a rytmický zákon její účinnost omezují. V uvedeném příkladě koncovka „na“ přídavného jména „fekálny“ má stejnou formu jako koncovka -na u podstatných jmen ženského rodu. Pomocí přípon jako je „álny“, je možno zachytit jen podmnožinu toho, co by mohla zachytit koncovka „ný“. Ani dlouhá koncovka „á“ není ve slovenštině jednoznačná, protože může označovat současně přídavná jména (pokud se koncovka nezkracuje pod vlivem rytmického zákona) a první a čtvrtý pád v množném čísle substantiv středního rodu („pravidlo“ ⇨ „pravidlá“).

Změna „g“ na „h“

horní lužičtina	1	ratarske graty
čeština		integrovaný hotel Gomel
slovenština		*výjimky*: mozog
polština	0	
dolní lužičtina		

Fig. 7 - Změna „g“ na „h“

Tato změna je typická pro horní lužičtinu, češtinu a slovenštinu. Umožňuje nepochybnou identifikaci cizích slov, která obsahují „g“. Malý seznam výjimek ve slovenštině je zanedbatelný.

Dvě samohlásky za sebou

čeština	0,9	kromě „ou“
slovenština	0,5	kromě „ou“, „ia“, „ie“, „iou“, „iu“

Fig. 8 - dvě samohlásky za sebou

Čeština má větší účinnost tím, že má pouze jednu domácí dvojhlásku. Pokud není dvojhláska na morfematickém švu, označuje cizí původ slova. Přesně kvůli tomu rozvíjeme automatické morfematické dělení slov. Situaci v jiných jazycích jsem ještě neprobral.

Samohláska na začátku slova a protetické „j“ a „w“

horní lužičtina	1	ja a agronom je e energija ji i integrować wo o organ wu u unija
čeština	0,6	ja a apatie je e embargo ji i ideologie

Fig. 9 - Samohláska na začátku slova a protetické „j“ a „w“

Tato fakta byla už dávno známá. Dobrovský ve své publikaci „*Dějiny české řeči a literatury*“ z roku 1792 napsal: „Slovan nechává samohlásku **a** na začátku slova zřídka kdy, **e** pak nikdy bez joty. Říman říká **est**, Slovan **jest**“. Prostudování slovníků češtiny ukázalo, že čeština překrývá také začáteční „i“ protetickým „j“, což nebývá ve slovenštině (je tam „istota“, zatím co v češtině je „jistota“). Existuje zhruba 10 až 12 výjimek pro začáteční „a“, asi 2 pro začáteční „e“ a „i“.

Horní lužičtina poskytuje lepší výsledky než čeština, protože má protetické „w“ před „o“ a „u“ („wokno“ místo „okno“ a „wučbnica“ místo „učebnice“). Z toho vyplývá, že žádné hornolužické slovo nemůže (až na neprovedenou metatezi typu „ert“) začít samohláskou. V horní lužičtině začáteční samohláska vždy označuje slovo cizího původu.

Metateze likvid

horní lužičtina	0,5	kruwa ert
čeština	1	kráva ret
slovenština		krava
polština		krowa
dolní lužičtina	?	krowa

Fig. 10 - metateze likvid

Metateze likvid proběhla plně v češtině. Češi říkají „Labe“, Francouzi a Němci „Elbe“. Situace je méně jasná v horní lužičtině, kde jsou slova s metatezí „kruwa“ a ostatní bez metateze „žałza“, „ert“ (znamená „pusa“) tam, kde jsou v češtině slova „žláza“ a „ret“. Díky metatezi se mohou pak uplatnit šablony typu „samohláska {R | L} souhláska“ pro vyhledávání cizích slov. Nemilosrdně ukazuje na cizí původ slov už dávno zdomácnělých jako „barva“ (z němčiny „Farbe“), „malba“ (z německého slovesa „malen“) nebo „larva“ (viz „larve“ ve francouzštině).

Vypočítatelnost diachronního slovanského systému:

Po zkoumání vypočítatelnosti češtiny i dalších západoslovanských jazyků se výhodně uplatnila myšlenka obrátit pozornost na fonologický vývoj této skupiny jazyků. Je samozřejmé, že existuje určitý počet výjimek, ale velmi pozoruhodný poznatek je, že pravidelnost fonologických jevů je značně vyšší než jsem očekával.

Program, který jsem implementoval, pečlivě sleduje krok za krokem výklad „historické mluvnice češtiny" podaný Lamprechtem, Šlosarem a Bauerem (1986). Je velmi zajímavé, že program, aby správně vygeneroval slovanské lexémy, musí sledovat takové pořadí, v jakém jsou fonologické jevy chronologicky popsány v této mluvnici.

Fonologické jevy prvního období vývoje (od praslovanštiny do konce X. století) jsem zvlášť pečlivě rozebral, protože tyto jevy jsou pro slovanské jazyky vůbec nejzávažnější. Z toho vyplývá, že vyžadují nejpracnější programování. Běžný didaktický výklad zde nepostačí, např. při zpracování nosovek se musí taky brát v úvahu typ přízvuku / tónu.

V tomto prvním období jsou rozebrány metateze s „e" a s „o", stahování, vývoj nosovky „ǫ" (s dlouhým stoupajícím přízvukem, s dlouhým klesajícím přízvukem a jako dlouhá nepřízvučná nosovka), vývoj nosovky „ę" (vzor „maso" a vzor „pět"), vypouštění tvrdého lichého jeru, vokalizace tvrdého sudého jeru, vokalizace měkkého sudého jeru a konečně vokalizace měkkého lichého jeru. Při generování se také projevilo, že vypouštění lichého měkkého jeru není tak jednoduché, jak to naznačují didaktické výklady o Havlíkově pravidle. Předcházející souhláska se pod vlivem jeru mění podle pravidel palatalizace.

Pro druhé období (od konce X. století do konce XIV. století) jsou programem vygenerovány: změna „g ⇨ h", přehlásky „'a ⇨ ě" a „'u ⇨ i" a konečně depalatalizace (v češtině a zvlášť ve slovinštině).

Fonologické jevy třetího období (od konce XIV. století do konce XVI. století) se hlavně vztahují k rozdělení češtiny a slovenštiny. Jedná se o diftongizaci „ú ⇨ ou", o monoftongizaci „ie ⇨ í" a vývoj dlouhého slovanského „ó".

Programování historického vývoje západoslovanských jazyků se stalo dobrým nástrojem pro výuku historické mluvnice (zvlášť u cizích studentů) a přineslo několik zajímavých poznatků. Nejpozoruhodnější z nich je stálé „vzdalování" češtiny od společné „československé základny", jak je vidět níže:

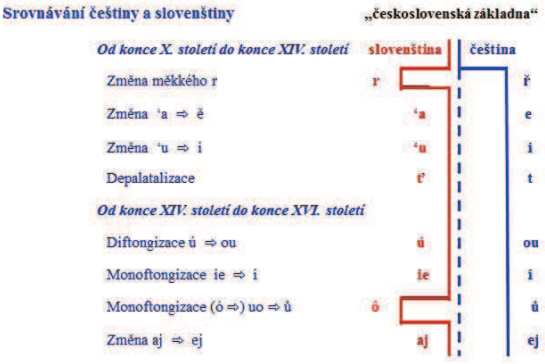

Fig. 8 - srovnávání vývoje češtiny a slovenštiny

Závěrem bych chtěl ukázat, že grafotax češtiny, totiž vnější graf(émat)ická organizace (zřejmě nejvýhodnější ze všech západních slovanských jazyků) je natolik důležitá, že jakákoliv budoucí reforma pravopisu by měla také uvažovat o účinnosti nových změn na úrovni vypočítatelnosti, jež nesmí být narušena. Uvedu příklad:

V dnešní češtině „ó“ vždy označuje slovo cizího původu, protože se původní slovanské „ó“ změnilo v „uo“/ „uó“ (přechodný stupeň), které se dále změnilo v jednoduchou dlouhou hlásku „ú“, zapisovanou jako „ů“ (Pleskalová, 2001). Kdybychom zrušili délku ze slov jako „milión“, „citrón“, „balkón“, dalo by se automaticky poznat jedině slovo „milión“ díky faktu, že má dvě samohlásky za sebou, případně „balkón“ díky tomu, že nedošlo k metatezi. V jiných případech je pravděpodobně vypočítatelnost co do jeho cizího původu nulová.

Literatura

Bareš, Rudolf Die. Nomina auf -dlo. Ein Beitrag zur tschechischen Wortbildung. Meisenheim am Glan: Verlag Anton Hain, 1970.

Blažek, Václav. “On the internal classification of Indo-European languages: survey”. Linguistica ONLINE, 2005
<http://www.phil.muni.cz/linguistica/art/blazek/bla-003.pdf.> 15.05.2018.

Brodda, Benny. “An Experiment with Heuristic Parsing of Swedish”. [In:] First Conference of the European Chapter of the Assocition of Computational Llinguistics, Pise, 1983, 66−73.

Derksen, Rick. Etymological Dictionary of the Slavic Inherited Lexicon. Leiden: Brill, 2008.

Dobrovský,ľJozef. Dějiny české řeči a literatury. Praha: Československý ľspisovatel, 1792 –1951.

Hajičová, Eva, Petr Sgall. “Towards Automatic Understanding of Technical Texts”. Prague Bulletin of Mathematical Linguistics 36, 1981: 5−20.

Havránek, Bohuslav, Alois Jedlička. Česká mluvnice. Praha: SPN, 1960.

Jacquet-Pfau, Christine, Marie-Anne Moreaux. « Motivation et transparence des emprunts gréco-latins en français et en allemand ». [In:] André Clas, Mejri Salah, Taïeb

Baccouche (éd.) La mémoire des mots. Actes du colloque de Tunis. Tunis: AUPELF-UREF, 1998, 587−600. [Universités francophones].

Jamborová-Lemay, Diana. Analyse automatique du slovaque. Etude approfondie du système Linguistique slovaque et sa reconnaissance d’après la forme dans les textes scientifiques et techniques. Application au machinisme agricole. Thèse de doctorat. Paris: CERTAL-INALCO, 2003.

Källgren, Gunnar “Parsing without Lexicon: the MorP System”. Berlin: 5th Conference of the European Chapter of The Associarion for Computational Linguistics, pp. 143-148, 1991.

Kirschner, Zdeněk. “MOSAIC. A Method of Automatic Extraction of Technical Terms in Texts”. Bulletin of Mathematical Linguistics 37, 1982.

Kirschner, Zdeněk. “MOSAIC. A Method of Automatic Extraction of Significant Terms from Texts”. Prague: Explizite Beschreibung der Sprache und automatische Textbearbeitung X, 1983.

Komárek, Miroslav. Historická mluvnice česká. Tome I: Hláskosloví. Praha: SPN, 1969.

Lamprecht, Arnošt, Dušan Šlosar, Jaroslav Bauer. Historická mluvnice češtiny. Praha: SPN, 1986.

Mareš, František Václav. Diachronische Phonologie des Ur- und Frühslavischen. Frankfurt am Main: Peter Lang,1999.

Mareš, František Václav. Cyrilometodějská tradice a slavistika. Praha: Torst, 2000.

Marvan, Jiří. Jazykové milénium. Slovanská kontrakce a její český ľzdroj. Praha: Academia, 2000.

Mazon, André. Grammaire de la langue tchèque. Paris: Institut d’Etudes Slaves, 1952.

Meillet, Antoine (édition révisée par André Vaillant). Le slave commun. Paris: Honoré Champion, 1965.

Mistrík, Jozef Moderná slovenčina. Bratislava: SPN, 1983.

Pleskalová, Jana. Stará čeština pro nefilology. Brno: Filozofická fakulta Masarykovy Univerzity, 2001.
Pognan, Patrice. "Une reconnaissance automatique des mots étrangers dans les textes scientifiques. Un essai en langue tchèque". The Prague Bulletin of Mathematical Linguistics 40, 1983 : 31−42.
Pognan, Patrice. "Histoire de l'écriture et de l'orthographe tchèques". Histoire, Epistémologie, Langage 21/1, 1999 : 27−62.
Pognan, Patrice. Introduction aux systèmes d'écriture des langues slaves de l'Ouest (polonais, bas-sorabe, haut- sorabe, tchèque, slovaque). Toulouse: Slavica occitania, 2001.
Pognan, Patrice. "Forme et fonction en analyse automatique du tchèque. Calculabilité des langues slaves de l'Ouest". [In:] Aleksandra Didkiewicz, Izabella Thomas (coord.) BULAG 32 "Les langues slaves et le français: approches formelles dans les études contrastives". Besançon, 2007, 13−33. [Nos revues].
Pognan, Patrice. "Système linguistique et calculabilité des langues slaves de l'Ouest (Nord et Sud): définition d'une décomposition morphématique homogène". [In:] Amr Helmy Ibrahim (éd.) Colloque international "Universalité et grammaire: paradoxe insoluble ou solution matricielle?" Paris: CRL, 2016, 174−181.
Pognan, Patrice, Jarmila Panevová. "Génération automatique de lexèmes slaves à partir de leurs racines historiques. Une des bases de l'enseignement multilingue des langues slaves de l'Ouest (Nord et Sud)". Ljubljana: Linguistica 52, 2013: 59−75.
Schlamberger Brezar, Mojca, Gregor Perko, Patrice Pognan. Les bases de la morphologie du slovène pour locuteurs francophones. Ljubljana: Filosofska Fakulteta, Univerza v Ljubljani, 2015.
Zemb, Jean-Marie. "Comment les mots éclairent les mots. A. approche graphématique. B. approche sémantique. C. approche métaphorique". Résumés des cours et travaux.
Chaire de grammaire et pensée allemandes. Paris: Annuaire du Collège de France,
1996−1997.

Патрис Поњан

ИЗРАЧУНЉИВОСТ ЧЕШКОГ ЈЕЗИКА У ПОРЕЂЕЊУ СА ОСТАЛИМ ЗАПАДНОСЛОВЕНСКИМ ЈЕЗИЦИМА

Резиме

У чланку је изложен мој досадашњи четрдесет петогодишњи рад у области синхронијског и дијахронијског проучавања система прво чешког језика, а после и других западнословенских језика, као и њихове аутоматске обраде.

Различити морфолошки и морфосинтаксички програми анализе чешког језика, засновани на препознавању лингвистичких облика, показују правилност чешког језика и висок степен његове израчунљивости (вероватно највећи од свих словенских језика).

Стечена сазнања обилато су искоришћена у граматици чешког језика и у наставним методама.

Међутим, основни задатак и даље представља аутоматско састављање главних делова речника чешког језика, тј. одреднице и грамема (граматичких вредности одреднице), приликом анализе велике количине текста. Тек би овакав хеуристички поступак без речника потврдио израчунљивост чешког језика.

Кључне речи: израчунљивост, аутоматска обрада чешког језика, аутоматско генерисање словенских лексема