Pravopis starších textů versus textový korpus k lingvistickým analýzám

František Martínek

1. Úvodem

Jako základ projektu *Lingvistická analýza českých humanistických textů* je vytvářen vyvážený elektronický korpus tištěných textů z let 1500–1620. Korpus, který by měl v budoucnosti čítat kolem 600 000 slovních tvarů, bude určený k lingvistickým analýzám, hlavně pro morfologická, syntaktická a lexikologická zkoumání, a to primárně při výuce studentů bakalářských a magisterských programů bohemistiky.[1] Texty se pro účely korpusu transkribují, nadto se v transliteraci (v závorkách se značkami s písmenem *e*, srov. níže) zaznamenávají všechny sporné jevy – zápisy umožňující dvojí interpretaci, ať už je příčinou problému tisková chyba, nejasnost v grafice či některém jazykovém plánu, nebo nutnost emendace.[2] Pro transkripci jsme se rozhodli na základě přesvědčení, že transliterace brání efektivnímu prohledávání textů elektronickými nástroji (srov. Kučera 1998: 306n.).[3]

Ediční příprava textů pro korpus se až na několik drobných odchylek řídí postupy diachronní složky Českého národního korpusu (ČNK). Odchylky jsou dány skutečností, že zatímco v diachronní složce ČNK se řešení některých nejasných a sporných věcí nechává na lemmatizaci, u našeho korpusu z časových důvodů nepočítáme s „celoplošnou“ lemmatizací, ale i tak chceme zajistit, aby se i méně poučený uživatel

1 O žánrovém složení korpusu viz Martínek 2009: 461 a Martínek – Ocelák (v tisku).

2 K užití značek <e> a </e> zde a dalších značek níže (u ukázky editovaného textu) srov. článek Martina Stluky v tomto sborníku.

3 Děkuji P. Koskovi za poznámku v diskusi, že při transliteraci by se objevily problémy podobné těm, které je nutné řešit při transkripci. Řečeno jinak a nikoliv jeho slovy, je ošidné transliteraci preferovat a vnímat ji jako mnohem přesnější podání jazyka příslušné památky.

dostal ke všem výsledkům, které hledá. To však ani v nejmenším neznamená, že bychom některé jevy nenáležitě upravovali.

Následující text předkládá k diskusi několik edičních problémů, jež se objevily při budování uvedeného korpusu a jež se přitom vážou k pravopisu, respektive grafice českých tisků z uvedeného období. Je nutné předeslat, že mívají několik možných řešení a je obtížné posoudit jejich vhodnost izolovaně, bez zřetele k celku. Týkají se jednak grafických a pravopisných jevů, které se do transkripce ani do transliterace nepromítnou (viz 2.), jednak takových pravopisných a hláskoslovných případů, jež v paralelní transliteraci zachováváme (viz 3.). Po jejich rozboru postup transkripce teoreticky odůvodňujeme a zaměřujeme se na obecnější otázky, jak zacházet s jevy na pomezí pravopisu a hláskosloví (viz 4.).

Předtím si na příkladu ukážeme, jak vypadá editovaný text:

<a>Abraham z Gynterrodu
<t>Cyri paedia</t>
<r>1605</r>
<s>1a</s> <k> <o>Cyri paedia</o>, hodnověrná starožitná historia o chvalitebném ve všelikých <e>wewsselikých</e> knížecích <e>Knjžetcých</e> ctnostech vycvičení a zvedení, o slavných skutcích, vítězných válkách a právě heroitském <l>heroický</l> šlechetném životu Cýra Staršího, prvního monarchy perského.
V níž se vypisuje, jakým způsobem <l>způsob</l> Pán Bůh monarchí <l>monarchie</l> kaldejskou <l>chaldejský</l> skrze Daria (…) *a Cýra* (…) *na médský <e>Medský</e> a perský národ přenesti <l>přenésti</l> a monarchí <l>monarchie</l>médskou <e>Medskou</e> a perskou zříditi a utvrditi ráčil.* (…) *</k>*
Léta Páně 1605. <e>M. DC. V.</e>
(XenCyr1605 1a)

Značky *<a>*, *<t>* a *<r>* zaznamenávají autora (popř. překladatele), titul a rok vzniku díla (tisku), *<s>* uvádí foliaci, *<k>* označuje titul, *<e>* pak transliterace sporných míst (srov. výše), *<l>* uvádí tzv. lemma (srov. 3.) a *<o>* vyznačuje tzv. citátové slovo (srov. 2.4.1).

2. Co se transkribuje bez upozornění

S ohledem na výše uvedené účely představovaného korpusu není v transkripcích nezbytné zachovávat podle původních tisků některé jevy povahy grafické a pravopisné.[4] Jde především o následující:

4 Ke grafice a pravopisu humanistických textů srov. Porák 1983.

2.1 Běžné tiskařské zkratky jako *etc*., *dobré*^{o}, *p'gemný* m. „příjemný“. Naproti tomu tam, kde je zkrácena větší část slova, uvádíme také transliteraci, např. u *S*^{o} m. „svatého“.

2.2 Grafické varianty literek. V přepisech nerozlišujeme *s* a *ſ*, dvojí *r* či dvojí *b* a ani hláskový (resp. v některých případech už pouze grafický) rozdíl mezi grafémy *l* a *ł*.

2.3 Distribuce malých a velkých počátečních písmen. Vzhledem k účelu korpusu není rozlišování podstatné. Velká písmena zachováváme např. u náboženských pojmů.

2.4 Dvojí typ písma – fraktura a antikva. Rezignujeme na záznam odlišných typů písma, humanistického a novogotického, a to i v případech těch slov přejatých, u kterých je kmen slova tištěn antikvou a zakončení frakturou (srov. obdobné doporučení J. Vintra – 1998: 343, bod Střídání druhů písma).

2.4.1 K citátovým slovům. J. Vintr také uvádí následující, pro nás problematičtější ediční zásadu: „Citátová slova (často tištěna antikvou uvnitř českého textu vysázeného frakturou) ponecháváme v originálním pravopise.“ (t.) Nechceme-li převádět citátové slovo z originálního pravopisu, máme k dispozici kód <*o*>. Vintrovo vyjádření však nesmíme otočit a odvozovat z něj zásadu, že slova tištěná antikvou jsou citátová.

Bez ohledu na odlišení typů písma, fraktury a antikvy, se nyní věnujme vymezení citátových slov. Ukažme si nejprve příklad slova neadaptovaného:

> „…***a to moře Řekové a Latiníci jmenují <o>Oceanum</o>***, *rozumějíce skrz to samé toliko pravé to moře největší a nejhlubší*…“ (MünstKozm1554 7b, zvýraznění F. M.)

Ne vždy máme k dispozici podobně jednoznačný signál citátovosti, jako je pokyn, že se slovo v dané formě užívá v cizím jazyce. Často je jedno slovo v textu užito několikrát a parametry citátového jednou splňuje, podruhé postrádá. V následující ukázce má slovo *paralell* dvakrát latinskou, dvakrát českou deklinační koncovku:

> „*Z těchto cirkuluov, kteří* ***<o>Paralelli</o>*** *slovou, berou se dnuov krátkosti, dlouhosti a jejich srovnání a nejednostejnosti. Neb když dvě města leží pod jedním* ***paralelem <e>Paralellem</e>****, tehdy den a noc přes celý rok na tom místě jednostejní budou.* (…) *s Mohúčem <l>Mohuč</l> okolo padesáte* ***paraleluov <e>Paralelluov</e>*** (…) *podál jsou. Však tak, aby se* ***<o>Paralelli</o>*** *od stupně k stupni počítali, a tento počet slove širokost zemí a krajin* (…).“ (t. 19b)

Někdy problematizuje citátovost sémantické kritérium:

> „…*a tak z těchto řečí na otázky o duši největší* ***<o>Definitiones</o>*** *se bráti a* ***odpovědi*** *dávány býti mohou*…“ (t. 8a)

Slovo *definitiones* je podle ortografických, morfologických apod. kritérií jednoznačně citátové, ale podle sémantického kritéria začíná být jeho citátovost sporná, protože ve větě stojí ve vztahu k českému slovu blízkého významu *odpovědi*.[5]

V jiných případech jsou naopak přejatá slova plně zapojena do českého jazykového systému. Shrňme, že jednoduché vodítko pro vymezení citátového slova nelze poskytnout, zejména vzhledem k existenci mnoha typů přechodných případů.[6] Citátovost patrně souvisí s jakýmsi grafickým obrazem cizích (resp. přejímaných) slov: jejich latinský či německý pravopis byl písařům a tiskařům znám, a nadto byl zřejmě hodnocen jako prestižnější. Často se můžeme setkat např. s vlivem cizí grafiky na slovo již částečně zdomácnělé (na podobný problém narážíme i u vlastních jmen; srov. příklady v odd. **3.2**).[7] Užití antikvy místo fraktury v tisku ovšem není kritériem pro určení, zda je dané slovo citátové. Abychom citátovost dostatečně postihli, museli bychom podrobněji, na škále, rozpracovat postup adaptací původních cizích, citátových slov do domácího jazykového systému,[8] navíc se zohledněním užívání více jazyků v dobové komunikaci (k tomu srov. Čejka 1998).

2.5 Pravopisné varianty. V korpusu splynou jednotlivé obvyklé pravopisné varianty reprezentující jeden foném, např. w*ijera* a *wiera* bude přepsáno jako *viera*, *wijra* a *wjra* jako *víra*, *naylepſſij* a *naglepſſij* jako *najlepší* apod.[9]

5 K významovému a formálnímu začleňování cizích slov do domácího jazykového systému srov. Němec 1968: 14–15. K citátovým slovům v humanistické češtině srov. Michálek 1973: 78–80 a 84n. Širší pojetí citátových slov zastává monografie *Česká lexikologie* (Filipec – Čermák 1985: 121).

6 Srov. i nejednotné a poměrně vágní vymezení citátových slov v literatuře citované v předchozí poznámce.

7 Čelíme zde vlastně podobnému problému jako u značení kvantity. Stojí-li cizí či přejaté slovo tištěné antikvou v nominativu, nemůžeme podle tisku zcela spolehlivě rozhodnout, zda má být chápáno jako citátové. Stejně tak si u předpokládané dlouhé samohlásky, jíž však označení délky chybí, obvykle nemůžeme být jisti, zda jde o tiskařovu nedůslednost, nebo odraz stavu jazyka, ať už má jakékoliv příčiny.

8 Inspiraci nabízejí texty J. Furdíka (např. 2008: 68–70) a M. Ološtiaka (např. 2004) pracující s pojmy akceptační a interlingvální motivace a rozpracovávající jednotlivé způsoby adaptace.

9 Toto pravidlo se netýká zápisu souhlásky digrafem, který naopak zaznamenáváme v transliteraci, např. *zlato <e>zlatto</e>*.

2.6 Původní interpunkce v tiscích. Její zachování by podle našeho názoru bylo nekompatibilní s transkripcí.[10] Proto ji měníme na syntakticko-logickou, avšak s určitým zřetelem k dobovým odchylkám v členění textů: při oddělování přechodníkových konstrukcí, anakolutů nebo při psaní čárek mezi dvěma spojkami trpíme „novočeskou nejistotou“.[11] Otazníky zachováváme jen tam, kde neruší novočeskou interpunkci. Nezavádíme uvozovky.

Interpunkci textů 16. století popsal D. Šlosar (1964; 1966; 1999). Je zpravidla trojstupňová: zahrnuje značky pro tečku, dvojtečku a čárku (ať v podobě i dnes užívané čárky, nebo virgule). Kladení čárek se řídí pauzovým principem (tzv. výdechová interpunkce). Obvykle se neoddělují krátké vložené věty (např. vztažné), větné členy se zdůrazňují pomocí interpunkce,[12] východisko a jádro je rozděleno čárkou, častěji než dnes se klade čárka před *a*. Naše doklady plně dokumentují Šlosarova zjištění.

Interpunkci je nadřazena obecnější otázka: členění textu a hranice větných celků. Zaznamenáváme základní logické i fyzické členění textů (kapitoly a jejich nadpisy, odstavce, stránkování, marginálie apod.), avšak ne již např. dělení tisku na řádky.

V edici je problematické respektovat (tedy reprodukovat) vedle členění v rámci věty (srov. výše) také původní členění tisku na věty: ve starších textech nebývaly větné celky jednoznačně rozčleněny a nová věta se např. mohla těsně připojovat k předešlému textu vztažným zájmenem (srov. Kolár 1993: 204).[13] V některých editovaných textech neodpovídají signály konce věty jejich logickým hranicím.

3. Co se transkribuje a ponechává v připojené transliteraci

Nemá smysl zastírat, že při rozhodnutích o tom, které z jevů signalizovat značkami a zaznamenat v transliteraci, je nutné vést ostrou hranici kontinuem případů – fuzzy množinou. Názorně to ukazuje např. již

10 Srov. požadavek formulovaný J. Kolárem (1993: 204), nechť editor stanoví syntaktické členění textu – nejprve na věty, pak na nižší celky – a podle něj dosadí novočeskou interpunkci.

11 K obecné neurčitosti obdobných pokynů, jež však obecně neruší jejich využitelnost, srov.: „Sjednocujeme tedy interpunkci převáděním **v podstatě** na dnešní grafiku (…), interpunkčních znamének užíváme co nejméně, **nerušíme však interpunkci pramene, není-li zcela pochybená**.“ (*StčS. Úvodní stati* 1968: 45, bod 8.15, zvýr. F. M.)

12 D. Šlosar píše, že „pomocí interpunkce se relativně osamostatňují větné členy, které nemají těsné vztahy k ostatním větným členům (zejména příslovečná určení)“ (1966: 167).

13 Při posuzování, jak je text členěn na souvětí a souvětí na věty, je třeba mít na paměti zákonitosti humanistické periody, zaváděné do českých textů, jak píše Kolár, od 2. pol. 16. stol. (1993: 204).

zmíněná skutečnost, že *dobré*^{o} přepíšeme bez vyznačení jako *dobrého*, *So* jako *svatého* naopak se zachycením transliterace. V některých případech jsme se rozhodovali mezi dvěma „stejně dobrými“ možnostmi, avšak vždy kontrolovatelně a na základě dostatečného množství příkladů, především podle poznatků přepisovačů jednotlivých textů. Právě jejich poznatky podstatně přispěly ke zpřesnění pravidel pro užívání edičních značek.

Korpus byl zpřehledněn zachycováním tzv. lemmat, s jejichž poznamenáváním se původně nepočítalo. Nejde přitom, jako u jiných historických slovníků nebo korpusů, o rekonstrukci hláskových podob k určitému datu ani o kompletní lemmatizaci, ale o výběrové pomocné přiřazení jak předpokládaných dobových, tak novočeských forem k těžko předvídatelným podobám nebo tvarům slov za účelem snadnějšího vyhledávání v korpusu.[14] K zaznamenání lemmat slouží značky s písmenem *l*.

Nyní uvedeme tři příklady grafických a hláskoslovných jevů, které vedle transkripce zachováváme podle tisku v připojené transliteraci.

3.1 Hranice slov

Prvním z jevů, který důsledně zanecháváme i v transliteraci, jsou transkripční **úpravy hranic slov proti podobám v tiscích**. Týká se to i některých jevů značně pravidelných – např. psaní příklonného *bych*, *bys* atd. dohromady s předcházejícím slovem nebo zčásti i psaní předložek dohromady se slovem následujícím –, které jsou v nových edicích systematicky měněny a zároveň z nich sotva odvodíme podstatnější poznatky. Slibnější pro další výzkum grafiky však je paralelní transliterace u následujících skupin případů při psaní odchylném od dnešního úzu:[15]

– psaní jiných příklonek než *bych* atd. dohromady s předcházejícím slovem, nebo zvlášť: *-s*, *-ť* a další příklonky píšeme v transkripci dohromady s předcházejícím slovem; u *-koli* píšeme případy jako *kdykoli* v transkripci dohromady, případy jako *kterýž by koli* zvlášť;

14 Tento postup také výrazně usnadňuje případnou úplnou lemmatizaci, pokud by se s texty pracovalo i v dalších projektech.

15 Je třeba zdůraznit, že užitá formulace „při psaní odchylném od dnešního úzu“ nemá naznačit, že podle takto zpracovaného korpusu není možné statisticky zkoumat hranice slov, že bychom zde tedy zachycovali jen vybrané jevy. Ve všech ostatních případech, tedy pokud hranice slov proti tisku neměníme, je totiž zřejmé, zda bylo slovo tištěno dohromady, nebo zvlášť, už z transkripce.

– psaní (potenciálních) příslovečných spřežek: zápisy *y hned*, *w čas*, *na proti*, *w niweč* atd. transkribujeme jako *ihned*, *včas*, *naproti*, *vniveč* a připojujeme transliteraci; u variant typu *při tom*/*přitom* ponecháváme rozhodnutí na editorovi, z přepisu je však vždy zřejmé, zda byl původní zápis s mezerou či dohromady;
– psaní složených (resp. vícečlenných) spojovacích výrazů: při transkripci formujících se spojovacích výrazů (srov. novočeské spojky jako *jestliže*, *protože*, *takže*, *avšak* atd.) zpravidla respektujeme hranice slov podle památky, pokud zápis nekolísá;
– oddělené psaní částí kompozit: srov. *k snažné její žádosti přesvatosvaté <e>Přeswato Swaté</e>* a úsporný zápis *spolubratrův a sestr <e>spolu Bratrůw a Sestr</e>*, kde vyjádříme, že i u druhého substantiva jde o kompozitum, přidáním lemmatu *<l>spolusestra</l>*.

3.2 Zápis přejímek

Další jev zachycovaný v paralelní transliteraci může sloužit jako model pro zacházení s grafikou i jiných skupin slov. Jde o přepis **přejímek** z různých jazyků nebo alespoň částečně **adaptovaných vlastních jmen**.[16] Zde se shodujeme s Vintrovými zásadami (1998: 342n.) a pravopis upravujeme podle dnešního úzu, např: *kancelář <e>kancellář</e>*, *Kristofor <e>Krystoffor</e>*, *Kartágo <e>Karthago</e>* nebo *Noremberk <e>Norembergk</e>*. Tím se však dostáváme do rozporu s praxí řady dostupných edic; mnoho edic především z 19. stol. si totiž u obdobných případů libuje v zachovávání nepodstatných pravopisných jevů. Ponechávají-li editoři původní pravopis výše uvedených slov, prozrazují, že tyto jevy nezařazují mezi pravopisné a zacházejí tedy s dobovým pravopisem heterogenně. Naopak jiné edice, především novější, je zase nadbytečně upravují a tím nivelizují podstatné jazykové jevy.

3.3 Značení kvantity

Posledním jevem, na který upozorňujeme, je značení **samohláskové kvantity**. Je jasné, že při každé změně samohláskové kvantity oproti tisku je nutné zaznamenat původní zápis v transliteraci, abychom mohli pracovat se všemi výskyty příslušného slova (slovní čeledi, odvozovacího typu apod.). Zůstává však otázka, kde je opodstatněné počítat

16 K přepisu německých přejímek do češtiny doby střední srov. text Kateřiny Rysové v tomto sborníku.

s kolísáním, tedy které jevy zařadit už do transkripce a neupozorňovat na ně transliterací.[17] Pozorujme následující příklady:

– kvantita předposlední samohlásky v troj- a víceslabičných substantivech typu *kázaní*: v tiscích 16. stol. se projevuje pomalé a postupné vyrovnávání směrem k zakončení na *-ání* bez ohledu na kvantitu třetí slabiky od konce;
– dlouhé *í* v nom. pl. mask.: *doktoří*;
– jiná než dnešní spisovná kvantita samohlásky v tvarech osobních a přivlastňovacích zájmen: *k ním*, *jejích*, *náše*; naopak krátké *i* v instrumentálu zájmen včetně ukazovacích: *s nim*, *s tim jejim*;
– krácení kořenné samohlásky v infinitivech se slabičnou, ale i s neslabičnou předponou;

Typy s jednotlivými kmenotvornými příponami se chovají různě. Kolísání podrobně popsal J. Gebauer (1958: 75n.), včetně analogických vyrovnávání, a posbíral k němu množství dokladů. Náš korpus textů prohlubuje novými doklady poznání o vývoji tohoto jevu: např. u sloves se zavřeným kmenem délka při odvozování původně zůstávala (*nésti – přinésti*), avšak v dokladech z textů ze 16. století se krátká kořenná samohláska vyskytuje velmi často (srov. i některé Gebauerovy doklady).

– komparativy a superlativy adverbií: *raději* a *radějí*.

Ve většině uvedených příkladů jde o kvantitu nekořenné samohlásky; jejich původní kvantitu je výhodné ponechat – kromě mnohem podstatnějších, jazykových důvodů – i proto, že to ve většině případů neovlivní vyhledávání v korpusu, neboť se nezmění podoba kořene. Jinak je tomu u zájmen a u sloves. Zachování kolísání u zájmen (příklady viz výše) odůvodňujeme tím, že členy početně omezených zájmenných paradigmat lze snadno vygenerovat s krátkými i dlouhými samohláskami; u sloves máme na zřeteli pravidelnost a rozšíření tohoto jevu v tehdejší češtině stejně jako jeho zbytky v češtině dnešní (srov. *dát – podat*).

U jiných případů kolísající kořenné kvantity uvádíme odchylnou podobu zpravidla pouze v transliteraci. V transkripci ji ponecháváme tehdy, pokud se jedná o velmi četnou podobu v celém korpusu (především u *i*/*í*: četné *líbost* proti řídkému *libost*, dvojí podoby *míle*/*mile*, *kníha*/*kniha*; také např. *náděje*/*naděje*) nebo v určité památce či malé skupině památek (pak ji v těchto textech zachováváme a lemmatem

17 S otázkou značení kvantity se vyrovnával i *Staročeský slovník* a vymezil následující skupiny případů (byť existují i další, ne tak jasné skupiny): a) Neznačí-li pramen kvantitu, je v jasných případech doplněna. b) U dublet je ponechána pouze značená délka. c) U sporných případů je připojena i podoba v prameni (popř. edici). Viz *StčS. Úvodní stati* 1968: 48 (bod 8.28).

upozorňujeme na očekávanou podobu: srov. např. *litost*, *poličky* ‚facky', *řimsa* aj.). Tímto způsobem je zajištěno, že i méně poučení uživatelé najdou „neočekávané" podoby, protože ty nebudou v korpusu „schovány", a výsledky jejich práce budou reprezentativní.[18]

4. Diskuse

4.1 Teoretický horizont prováděné transkripce

Podle čeho obecně posuzujeme sporné jevy? Jako materiál mohou sloužit především příklady z bodu 3.3 – značení samohláskové kvantity v tiscích a jeho interpretace. Jak při úvahách o jazykovém systému, tak při zpracování textů je výhodné odlišit tři skupiny jazykových jevů – jevy povahy **individuální, kolektivní a systémové**:[19]

a) **Individuální odchylka** bude v transkripci opravena, ale v transliteraci zaznamenána. Ke sporným případům se můžeme snadno vrátit a rozhodnout o nich až po shromáždění většího počtu dokladů, což podporuje rychlost zpracování.

b) **Kolektivní zvláštnost** je jev vlastní výrazné skupině autorů či památek (srov. výše *litost* – bude připojeno lemma *lítost*). Bude zaznamenána v transkripci, v případě potřeby na ni bude upozorněno lemmatem.

c) Popis **systémových jevů** je věcí kompletní lemmatizace korpusu a jeho případného dalšího lexikografického zpracování. To však již nebude náplní projektu.

Právě konfrontace různých dobových textů mezi sebou umožňuje zařadit jednotlivé jevy do příslušných skupin a toto zařazení odůvodnit.

4.2 Zachycení individuálních odchylek

K našemu přístupu se může vyskytnout tato námitka: Pokud v základním textu něco měním, zdánlivě nějaký jev odstraňuji, nivelizuji, popř. nadbytečně sjednocuji (srov. závěr bodu 3.3).[20] Je však nutné si uvědomit, že počítačové zpracování textu dává možnost zachytit text na více úrovních, a tím mizí problém, **zda** např. odchylně vyznačenou kvantitu zachytit, či ne – rozumí se samo sebou, že ano –, a mění se na otázku, **na jaké úrovni** ji zaznamenat (srov. 4.1, dříve to bylo možné přímo v textu nebo v ediční poznámce).

Při počítačovém zpracování vkládáme do textu značky a do nich údaje umožňující alternativní rozhodnutí, ale také informace o členění textu, o jeho původu apod. Text, který vytváříme, má tedy z pohledu lingvisty dvě roviny, přičemž v počítačovém souboru jde o rovinu jedinou.[21] Optimální příprava textu by spočívala ve vícerovinné html anotaci, jak ji používá např. řezenský diachronní korpus RRuDi (srov. Meyer 2005), ale takový postup není z časových důvodů možný.

Právě „nepředpojaté" počítačové zpracování textu by zdánlivě mělo garantovat adekvátní zachycení odchylek, např. v kvantitě, na rozdíl od dřívějších přístupů, které se vyznačovaly určitou (někdy i implicitní) mírou ohledu na dnešní nebo předpokládaný dobový stav jazyka. Takové naděje bývají vkládány do současných velkých jazykových korpusů, prezentovaných, na rozdíl od dřívějších excerpcí, jako neomylný zdroj k poznání příslušného jazykového stavu, ačkoliv jde zpravidla o korpusy značně omezené rozsahem, o korpusy psaných, nikoliv mluvených textů, a především o korpusy jazyka, na který se anotacemi (a v případě historických textů pochopitelně i transkripcemi) aplikují předem známá pravidla a zákonitosti, aniž by se přitom příliš přemýšlelo o pravidlech a zákonitostech jiných, o nichž není ve chvíli anotace nic známo. Tyto naše pochybnosti přitom nevybízejí k pohrdání novými přístupy, ale snaží se upozornit na fakt, že pohled každého badatele je determinován tím, co o jazyku, jeho systému a jednotlivých rovinách už ví. Tuto věc je vhodné si uvědomovat a nezavrhovat dostupné popisy jazyka,[22] byť by byly metodologicky nevyhovující. Na druhou stranu však není možné dosavadní poznatky absolutizovat, jak ukážeme dále.

18 Děkuji recenzentovi abstraktu za poznámku, že kvantita ve starší češtině byla objektivně nejednotná. Že naše řešení není s jeho tvrzením v rozporu, se pokusíme ukázat v následující kapitole.

19 Tímto rozdělením modifikujeme třírovinný model L. Lemnitzera (1997: 66n.) vypracovaný pro jiné účely – pro oddělení tvarů, podob a lemmat slov v synchronním korpusu. Podrobněji o tomto rozdělení viz Martínek – Ocelák (v tisku).

20 Např. Josef Vintr (1998: 344) požaduje, aby se v edicích kvantita a obdobné jevy zachovávaly podle pramene.

21 Účelné prohledávání takového „jednorovinného" textu umožňuje program MonoConc. Dostupné z <http://www.monoconc.com>.

22 Východiskem k poznání systému dobového jazyka jsou vedle gramatik slovníky (Jungmann 1989/1835–1839), a především lexikální databáze (Nejedlý et al. 2010).

4.3 Kolísání, zachycení kvantity a představy o starém jazyce

Diskusi o kolísání ve starším jazyce uvedeme citátem, jehož autorem je J. Kolár:

> „Na jazyk děl staré české literatury je třeba při ediční práci pohlížet jako na živý organismus, nástroj k sdělení myšlenky, neorganizovaný ještě nějakou obecně přijatou normou. Proto není přípustné jazyk starých děl normalizovat ve významotvorných složkách; kolísání je jev historicky zdůvodněný a odstraňovat jej příklonem k určitému jednotnému tvaru či způsobu by znamenalo zásah do autentického stavu jazyka, jeho nenáležité přizpůsobování vědeckým poznatkům atp.“ (1993: 200)

K tomuto citátu, který v obecné rovině jistě vyhovuje, lze mít při pokusu o jeho aplikaci následující dvě námitky: a) Není zcela zřejmé, co se myslí významotvornými složkami jazyka, a kde tedy leží hranice, za níž jsou úpravy kolísání nepřípustné. b) Co se myslí nenáležitým přizpůsobováním jazyka vědeckým poznatkům? Můžeme uvažovat i o „přizpůsobování jazyka předpokládané pohodlnosti čtenářů“? Nebo šlo o konflikt s odstraňovateli „gramaticky chybných“ podob, kteří usilovali o jazykovou vybroušenost vydávaných památek podle norem češtiny zlatého věku? (Mluvíme-li o vydávání starších textů pro lingvistické účely, není se dnes již třeba s puristickými upravovateli vypořádávat.)

Kolár vyhroceně formuluje potřebu zachování kolísání. Neodkládejme však následující pochybnosti: Co když je předpokládané kolísání v hláskosloví také záležitostí grafických zvyklostí a důsledkem omezených technických možností tiskaře, odkázaného na sadu literek? (K diskusi na toto téma srov. např. Valášek 2000: 153–155.) Nakolik je možné považovat různé podoby v bezprostředním sousedství za projevy variantnosti, jako je tomu třeba v dnešní obecné češtině, kde se spisovné a obecněčeské koncovky vedle sebe „snesou“, a nakolik za nedůslednost tiskaře?

Pokusme se tyto otázky konkretizovat opět na příkladu samohláskové kvantity. Kolár o ní píše:

> „U tištěných knih od počátku 16. století (…) je nutno brát kriticky v úvahu každou vyznačenou kvantitu jako projev obecné tendence k sílícímu respektování tohoto jevu a vycházet přitom ze zkušenosti, že každé vyznačené kvantitě je třeba věnovat kritickou pozornost, avšak její nevyznačení ještě neznamená, že na příslušném místě kvantita nebyla.“ (1993: 203)[23]

Od připomenuté kritické pozornosti není daleko k přehnanému respektování **každé** kvantity v tisku vyznačené. Jako memento takového přístupu nabízíme citaci z článku V. Flajšhanse o listu Řehoře Hrubého z Jelení z r. 1513:

> „…nikde není označena dlouhou slabika taková, kterou bychom musili míti za krátkou. Ovšem o slabikách krátkých důkaz vždy neplatí: tak na př. Hrubý píše důsledně krátce: n*a*děje, ned*o*věra, poč*a*tek, př*í*čina, př*í*hodný, *u*častný, v*y*borný atd., ale odvozovati z toho pro krátkost těchto slabik nelze ničeho; možno, že délka pouhou náhodou neznačena.“ (1893: 383)[24]

Čteme krajní názor, že důležité je pozitivní označení kvantity, zatímco chybějící označení není dostatečným argumentem, že příslušná samohláska byla krátká, protože je mohla způsobit vada tisku, nedůslednost tiskaře, ne tak velký význam kvantity, jako má dnes, nebo dokonce náhoda. O náhodě při pozitivní signalizaci délky se neuvažuje. Narazili jsme na limit tehdejších vědeckých poznatků. Flajšhans má v hlavě jasnou představu o jazyce 16. století a list Řehoře Hrubého ji nemůže zviklat. Svými poznatky ospravedlňoval jednoznačná rozhodnutí a důsledné ediční úpravy na problémových místech. Limity takové transkripce pro poznání stavu jazyka jsou zřejmé, ale limity transkripce by byly obdobné, pokud by se řídila jen **dnes** známými poznatky o jazykovém systému.

Domníváme se, že pro transkripci starších textů nestačí běžná poučka, že se grafické jevy přizpůsobují dnešnímu psaní, a jevy jazykové se naopak měnit nesmí. Existuje totiž pásmo přechodných jevů na pomezí pravopisu a jazyka, jejichž status budeme moci zpřesnit až po popsání rozsáhlejšího textového materiálu, a příslušná slova proto

23 Značení kvantity v rukopisech se zde Kolár nevěnuje.

24 Chápeme samozřejmě „polysémii“, přesněji vágnost pojmů *kvantita*, *délka*, *krátkost*, způsobenou jejich různým chápáním, jejich vztahováním k jiným okolnostem, odpovídajícím cílům popisu příslušného badatele, působícího nadto v jiné epoše. Otázku *krátkosti* Flajšhansovým způsobem vyřešit lze, budeme-li zkoumat krátkost etymologickou, případně snad i systémovou, ale sotva v případě, budeme-li chtít popsat češtinu příslušného autora či období.

zaznamenáváme i v transliteraci.[25, 26] Zacházení s velkým množstvím pramenů 16.–17. století a elektronický přístup k nim snad také umožní nalézt argumenty pro tvrzení o působení grafických zvyklostí zpětně na jazyk (srov. něm. pojem *Leseaussprache*).

5. Závěr

5.1 Metoda a výhledy

Uvedli jsme důvody, proč je vhodné používat přístup diachronní složky Českého národního korpusu. Respektujeme ho a na základě konkrétních zkušeností ho rozvíjíme. I tam, kde interpretujeme (jak jsme viděli, v našem přístupu jsou **v základním textu** odstraněny některé – podle našeho dosavadního stavu poznání – nepravidelné či nepřevládající kvantity), pokaždé umožňujeme se ke sporným místům vrátit a rekonstruovat je na základě připojené transliterace. Snad tedy budou jevy jako kvantita lépe uchopitelné, a to i statisticky, když budou přístupné větší korpusy textů, u nichž budou zachyceny **všechny**, ne jenom vybrané, popř. nápadné případy. (Z popisu nápadných jevů bývají podezřelí dřívější badatelé, ale ani dnes není takový přístup nebo srovnávání s mylnou představou o normách tehdejšího jazyka výjimkou; srov. 4.3.) Naše řešení je možností, která stojí mezi krajnostmi „vše nechat“ (tj. v důsledku: transliterovat) a „vše opravit“ (kolísání bychom nepovažovali za významotvorné), a odpovídá hlavnímu účelu korpusu – snadnému prohledávání.[27]

Postupujeme tak, že co uvedeme do poznatků, tj. spolehlivě popíšeme a uvedeme ve známost, to můžeme v transkripci ponechávat, případně s přidáním „lemmat“; naopak o čem jsme zatím nebádali, můžeme prozatím v transkripci sjednocovat a nechávat v transliteraci.[28] Již při

25 Toto pásmo přechodných jevů může být širší, než se na první pohled zdá, a proto v transliteraci zaznamenáváme i některé problémy, jejichž řešení velmi pravděpodobně známe; srov. výše v 3.1.

26 Jak známo, obdobou důsledné paralelní transliterace sporných míst při elektronickém zpracování je v klasické kritické edici kritický aparát.

27 S takto vymezeným účelem korpusu koresponduje označení „korpus edic“, které v diskusi po konferenčním příspěvku použil Pavel Kosek.

28 Připouštíme, že takový přístup předpokládá neustálé návraty k již hotovým textům a zanášení změn na základě nových poznatků, což odpovídá skutečnosti, že nelze poskytnout definitivně správná řešení některých problémů. Přitom se vhodně využívají možnosti elektronického přístupu k textům a zachovávají se lingvisticky relevantní informace, aby je bylo v budoucnu možné využít k lepším řešením. Ideálem na nedosažitelném horizontu pak teoreticky není pouze popis, ale dokonce znalost všech jazykových pravidel, aby bylo v textech nutné opravovat pouze tiskařské chyby.

zpracování byla nalezena řada zajímavých jevů, které byly vytipovány pro podrobnější zkoumání v celém materiálu. V souboru textů bez morfologické anotace je však těžko možné získávat materiál např. k syntaktickým analýzám (anakoluty, slovosled apod.) pomocí jednoduchého vyhledávání. A proto se, i při snaze o korpusově založenou diachronní bohemistiku a zejména gramatiku, budeme muset ještě nějakou dobu spokojit s kombinací nových a klasických pracovních metod.

5.2 Nakolik jsou texty v korpusu připraveny pro edici?

Pro případné edice jsou texty zahrnuté do korpusu připraveny v následujících ohledech: a) jsou spolehlivě přepsány (včetně kontroly přepisů podle zpřesněných zásad); b) zpravidla jsou vybaveny stručnou ediční poznámkou; c) podle transliterací lze snadno doplnit jevy, které je v edici jednoho konkrétního textu potřeba zachovat, a vypracovat kritický aparát; d) k nesrozumitelným jazykovým, především lexikálním jevům lze dodat vysvětlivky díky existenci dalších pramenů, např. *Lexikální databáze humanistické a barokní češtiny* (Nejedlý et al. 2010), navíc byly některé z těchto jevů při přípravách textů zachyceny a osvětleny. Naopak doposud zcela chybějí věcné vysvětlivky k reáliím, zpravidla také k pramenům textů a citacím v nich, jakož i komplexnější literárněhistorické informace o tiscích a biografické informace o jejich autorech, překladatelích, sazečích apod. (srov. Linka 2000), protože jsme je vzhledem k účelu korpusu shromažďovali jen velmi výběrově.

Grantová podpora

Text vznikl v rámci doktorandského grantu *Lingvistická analýza českých humanistických textů* 16 809 uděleného Grantovou agenturou Univerzity Karlovy.

Prameny

MünstKozm1554 Muenster, Sebastian (1554): *Kozmografia česká*. Přel. Zikmund z Púchova. Praha (K5969).

XenCyr1605 Xenofon (1605): *Cyri Paedia. Hodnověrná starožitná historia* (…) *o slavných skutcích* (…) *Cýra staršího*. Přel. Abraham z Gynterrodu. Praha (K17061).

Jungmann, Josef (1989, 1. vydání 1835–1839): *Slovník česko-německý*. Praha. Dostupné též z <www.slownjk.cz>.
Knihopis Digital [on-line]. [citováno 20. 9. 2010]. Filozofický ústav AV ČR, v. v. i. Kabinet pro klasická studia, Praha. Dostupné z < http://www.knihopis.org>.
Manuscriptorium [on-line]. [citováno 12. 10. 2010]. Dostupné z <www.manuscriptorium.com>.
Nejedlý, Petr et al. (2010): *Lexikální databáze humanistické a barokní češtiny* [on-line]. [citováno 12. 10. 2010]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupné z <http://madla.ujc.cas.cz>.

Literatura

Čejka, Mirek (1998): Střídání kódů u Jana Blahoslava a Martina Luthera (Několik poznámek k tzv. kulturní diglosii). *Listy filologické* 121, s. 84–104.
Filipec, Josef – Čermák, František (1985): *Česká lexikologie*. Praha.
Flajšhans, Václav (1893): Ke kvantitě češtiny XVI. století. *Listy filologické* 20, s. 383–390.
Furdík, Juraj (2008): *Teória motivácie v lexikálnej zásobe*. Košice.
Gebauer, Jan (1958, reprint 2. vydání z r. 1907, 1. vydání 1898): *Historická mluvnice jazyka českého 3. Tvarosloví. 2. Časování*. Praha.
Hádek, Karel (1970): Ke kvantitě samohlásek v češtině 17. století. *Listy filologické* 93, s. 44–53.
Kolár, Jaroslav et al. (1993): Textologie a problematika starší české literatury. In Vašák, Pavel et al.: *Textologie. Teorie a ediční praxe*. Praha, s. 184–205.
Kučera, Karel (1998): Diachronní složka Českého národního korpusu: Obecné zásady, kontext a současný stav. *Listy filologické* 121, s. 303–313.
Lemnitzer, Lothar (1997): *Akquisition komplexer Lexeme aus Textkorpora*. Tübingen.
Linka, Jan (2000): Explikatologie? *Listy filologické* 123, s. 157–158.
Lupínková, Naděžda (1971): Obraz kvantity kořenných slabik v raném díle J. A. Komenského. *Listy filologické* 94, s. 41–50.
Martínek, František (2009): Korpus českých textů z období humanismu jako východisko lingvistických analýz. In Ološtiak, Martin – Ivanová, Martina – Gianitsová-Ološtiaková, Lucia, eds.: *Varia XVIII. Zborník príspevkov z XVIII. kolokvia mladých jazykovedcov*. Prešov, s. 457–464.
Martínek, František – Ocelák, Radek (v tisku): Tvorba korpusu k lingvistické analýze humanistické češtiny.
Meyer, Roland (2005): The Regensburg Diachronic Corpus of Russian: A New Source for Linguistic Research (Not Only) on Modality. In Hansen, Björn – Karlík, Petr, eds.: *Modality in Slavonic Languages. New Perspectives*. München, s. 315–336.
Michálek, Emanuel (1973): K charakteristice české slovní zásoby doby střední. *Listy filologické* 96, s. 77–87.
Němec, Igor (1968): *Vývojové postupy české slovní zásoby*. Praha.
Ološtiak, Martin (2004): O interlingválnej motivácii (teoreticko-metodologické a terminologické poznámky). In Imrichová, Mária, ed.: *Slovenčina na začiatku 21. storočia. Na počesť profesora Ivora Ripku*. Prešov, s. 112–122.
Porák, Jaroslav (1983): *Humanistická čeština*. Praha.

StčS. Úvodní stati (1968): *Staročeský slovník. Úvodní stati, soupis pramenů a zkratek*. Praha.
Šlosar, Dušan (1964): Poznámky k vývoji české interpunkce v 16. století. *Listy filologické* 87, s. 126–135.
Šlosar, Dušan (1966): Průřez vývojem staročeské interpunkce. *Listy filologické* 89, s. 164–170.
Šlosar, Dušan (1999): Středník. In Zand, Gertraude – Holý, Jiří, eds.: *Tschechisches Barock – Sprache, Literatur, Kultur. České baroko – Jazyk, literatura, kultura*. Frankfurt a. M.–Bern–New York–Paris–Wien, s. 33–42.
Valášek, Martin (2000): Jindy a nyní (K historii vydávání raně novověkých česky psaných textů). *Listy filologické* 123, s. 149–156.
Vintr, Josef (1998): Zásady transkripce českých textů z barokní doby. *Listy filologické* 121, s. 341–346.

Orthography in Older Czech Texts versus a Text Corpus for Linguistic Purposes

The paper discusses some editorial problems with respect to orthography which occurred during building of a linguistic corpus of transcribed Czech texts from 1500 to 1620. In section two, the graphical and orthographical phenomena are pointed out which cannot occur in the transcription. In the next section, some orthographical and phonological phenomena are given which have to be transliterated besides the transcription because its interpretation may contribute to the interpretation of the Czech language development (e.g. vowel quantity, grammaticalization of conjunctions and adverbs, orthography of loanwords). Finally, a classification of language phenomena into individual, collective and systematic classes is supplied and thus theoretical reasons for the transcription method are explained.

Keywords

Corpus – Orthography – Historical Phonology – Vowel Quantity/Length – Transcription – Edition of Older Texts

Author

František Martínek (*1982) graduated in Czech Language and Literature and German Language and Literature at the Faculty of Arts, Charles University in Prague. Since his graduation in 2008 he has been a doctoral student of Czech Language there (the topic of his dissertation is *Analytická verbonominální spojení v humanistické češtině*). Since 2005, he has been working at the Department of Language Development of the Institute of the Czech Language of the Academy of Sciences of the Czech Republic and in 2010 he became an assistant at the Institute of Czech Language and Theory of Communication at the Faculty of Arts, Charles University in Prague. He participated in preparation of Božena Němcová's correspondence series and Jiří Haller's works collection. Apart

from textology and editology he also focuses on diachronic lexicology, lexicography and syntax, working with Humanist and Baroque Czech materials in particular. Moreover, he is interested in German-Czech language contacts and translation analysis.

(Polo)automatická počítačová transkripce

Boris Lehečka, Kateřina Voleková

V pojednáních o transkripci často používané formulace typu „hláska/foném *č* se v prameni zachycuje grafémem *c* nebo digrafem *cz*“ snad dovolují velmi zjednodušeně popsat transkripci jako proces, při němž se z podoby transliterované, odpovídající originálnímu textu, odvodí podoba fonetická a na jejím základě se vytvoří zápis, který odpovídá dnešním pravopisným pravidlům – výsledku se pak také říká transkripce.

Obdobně časté formulace typu „*d* přepisujeme jako *d* nebo *ď*“ nás sváděly k prozkoumání pravidel, jež se při tomto převodu z jedné pravopisné podoby do druhé uplatňují. Tato pravidla jsme se také pokusili formálně popsat a počítačově zpracovat.

Při studiu obecných edičních zásad (např. Daňhelka 1985; Vintr 1998), ale i transkripčních poznámek k jednotlivým edicím se však ukazuje, že na výslednou transkribovanou podobu má vliv daleko více faktorů než jen vztah mezi rovinou grafickou a fonologickou. Jedná se například o původ slova (jinak se přepisují grafémy ve slovech cizího původu a ve slovech domácích), literární útvar originálu (doporučení typu *v časoměrných verších zdvojená písmena ponecháváme*), chybu (tiskaře/písaře) v originálu, frekvenci různých zápisů téhož výrazu v celém textu (což vede k rozhodnutí, že jde buď o odchylku, která se v přepisu ponechá, nebo jde o chybu, která se v přepisu nahradí převažující správnou podobou), jazykovou znalost a zkušenost editora, zachycení identické pasáže v jiném rukopise atp.

Jak je vidět, repertoár faktorů je bohatý a jejich formální zachycení bude komplikované, někdy možná i nemožné. Proto jsme se drželi při zemi a přicházíme zatím se dvěma nástroji. Jeden z nich, kterému pracovně říkáme *Brus*, by měl pomoci při studiu transliterované podoby