ČASOPIS PRO MODERNÍ FILOLOGII 97, 2015, Č. 1, S. 70–77

Obsah a značkování diachronního korpusu češtiny[1]

Boris Lehečka (Praha)

OPEN ACCESS

THE CONTENT AND ANNOTATION OF THE DIACHRONIC CORPUS OF CZECH
The paper discusses what kind of content and annotation should be included in the diachronic corpus of Old Czech. Based on his analysis of the current state of DIAKORP and the Old Czech Text Bank the author suggests solutions for how to treat the critical apparatus, foreign words in historical Czech texts and contemporaneous or later marginal or interlinear notes. He also discusses some aspects of the methodology of statistics computation in the diachronic corpus.

KEYWORDS
diachronic corpus, corpus manager, vertical format, frequency

KLÍČOVÁ SLOVA
diachronní korpus, korpusový manažer, vertikální text, frekvence

1. VYMEZENÍ POJMŮ[2]

Označením **diachronní korpus češtiny** budu myslet soubor elektronicky zpracovaných autentických textů od počátků psané češtiny do počátků národního obrození, tj. zhruba od konce 13. století do konce století 18. Jinými slovy se jedná o korpus staročeského a středněčeského období. Problémy, o nichž příspěvek pojednává, a možnosti jejich řešení vycházejí jednak ze znalosti DIAKORPu, který tvoří veřejně přístupnou součást diachronní složky Českého národního korpusu[3], jednak ze zkušenosti s prací se staročeskou textovou bankou, která zpřístupňuje texty staročeského období prostřednictvím korpusového manažeru na stránkách Vokabuláře webového[4]. Oba zmíněné jazykové zdroje používají jako vstupní formát pro korpusový manažer[5] tzv. vertikálu, která kromě strukturních značek obsahuje jednotlivé

1 Vznik příspěvku byl podpořen projektem Ministerstva kultury ČR č. DF12P01OVV028 „Informační technologie ve službách jazykového kulturního bohatství".

2 Děkuji recenzentům za jejich připomínky a návrhy na opravy. Zůstaly-li v článku nějaké chyby, jsou pouze mým vlastním dílem.

3 Kučera, K. — Stluka, M. (2011): DIAKORP: Diachronní korpus, verze z 30. 11. 2014. Praha: Ústav Českého národního korpusu FF UK. Dostupný z: <http://www.korpus.cz>.

4 *Staročeská textová banka* [on-line]. Verze z 24. 6. 2014 [citováno ze dne 30. 11. 2014]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupná z <http://vokabular.ujc.cas.cz/banka.aspx>.

5 Zdrojové kódy korpusového manažeru KonText, využívaného v Českém národním korpusu, jsou k dispozici na stránce < https://bitbucket.org/ucnk/kontext>. Zdrojové kódy programu Corpus Manager, využívaného pro zpřístupnění staročeské textové banky, jsou k dispozici na stránce <http://sourceforge.net/projects/corpman/>.

tokeny a ke každému tokenu může přiřadit další informace (jako je lemma, morfologické značkování, emendační poznámky ap.) pomocí údajů oddělených tabulátorem.

...	<folio cislo="2v">
což	<odstavec>
jest	...
v	v
tobě	tobě
,	dobří
jest	násilé
ješitnost	trpie
a	a
vnitřnieho	vše
duchu	,
vešdajšie	což [emendováno z: cžož]
<e>	jest
wesdassie	v
</e>	tobě
tesknost	,
,	jest
</str>	ješitnost [emendováno z: gieſſytnoſt]
<str id="5">	a
dúš	vnitřnieho
nejisté	duchu
skonánie	vezdajší
...	...

Ukázka zdrojového textu vertikály pro Pasionál, Knihy o životech svatých. Vlevo zpracování v DIAKORPu[6] (rukopis z 2. poloviny 14. stol., Knihovna Národního muzea; sign. III D 44), vpravo ve staročeské textové bance[7] (prvotisk z roku 1476, Knihovna Národního muzea; sign. 25 C 1). Podrobnější informace o přípravě a značkování textů pro jednotlivé korpusy viz Stluka (2010) a Černá — Lehečka (2012).

2. ÚČEL DIACHRONNÍHO KORPUSU ČEŠTINY

Ve svém příspěvku se zaměřuji na dvě hlavní témata: 1) co všechno ze staro- a středněčeského pramene by se mělo objevit v korpusovém zpracování; 2) v jaké podobě by se to v něm mělo objevit. Množství a rozsah dat uváděných ve zdrojových datech korpusu předurčuje, k čemu všemu a jakým způsobem bude možné taková data využít. Bude-li například k dispozici elektronický text v transliterované podobě, bude možné zkoumat pravopisný systém daného pramene; bude-li kupříkladu v korpusu

6 Za laskavé poskytnutí tohoto materiálu děkuji Martinu Stlukovi z Ústavu Českého národního korpusu FF UK.

7 Za možnost pracovat s materiály staročeské textové banky děkuji pracovníkům oddělení vývoje jazyka Ústavu pro jazyk český AV ČR.

zpřístupněno slovníkové dílo s cizojazyčnými ekvivalenty, nabízí se možnost sledovat korelace mezi výchozím a cílovým jazykem. Způsob využití takových dat není pak dán pouze technickými prostředky, tj. kvalitami a funkcemi korpusového manažeru, ale i způsobem zpracování vstupních jazykových dat: díky lemmatizaci bude například možné snadněji vyhledat všechny tvary jednoho lemmatu.

Výše uvedená témata zastřešuje jedna základní otázka: K čemu (všemu) má korpus textů historické češtiny sloužit? Obecnou odpověď v tom smyslu, že takový korpus slouží ke komplexnímu výzkumu historické češtiny v pohledech synchronních i diachronních, je ovšem potřeba podrobit dalším otázkám: Půjde jenom o výzkum hláskosloví, morfologie, lexika a syntaxe? Nebylo by možné využít korpus i pro studium ortografie či korelace s cizojazyčnými ekvivalenty v případě jednotlivých slov, jako je tomu u slovníků, nebo korelace mezi různými zpracováními, tj. různými prameny téže literární památky? Bylo by žádoucí studovat rovněž vztah jinojazyčné předlohy a originálu? Toto jsou jistě relevantní oblasti výzkumu historické češtiny a podle mého názoru tvoří diachronní korpusy vhodnou materii pro takto široce zaměřený výzkum, přičemž korpusové manažery by měly poskytovat vhodné nástroje pro kladení otázek badateli i pro získávání údajů pro formulování odpovědí na tyto otázky.[8]

3. TVORBA DIACHRONNÍHO KORPUSU ČEŠTINY

Diachronní korpusy mají v korpusové lingvistice výjimečné postavení: na jednu stranu je jejich pramenná základna — na rozdíl od korpusů současného jazyka — ustálená, nevznikají žádné nové texty, takže by bylo možné vytvořit všezahrnující korpus. Podle některých zjednodušených představ jde z hlediska autorského zákona o díla volná, takže není potřeba při zařazení těchto děl do korpusu řešit autorská práva. Jak se v tomto ohledu mají věci doopravdy?

Za prvé: pomineme-li prvotisky, tvoří 99 % veškeré staročeské textové produkce rukopisy, z nichž každý je ze své podstaty jedinečný originál. Z hlediska autorského práva se jedná o díla volná, na druhou stranu jsou tyto rukopisy většinou majetkem nějaké instituce, která na ně může uplatnit tzv. reprodukční práva, tj. práva na zveřejnění fotokopií originálu. Pro období střední češtiny existuje daleko více tisků, které jsou k dispozici v několika desítkách či stovkách exemplářů, přesto však existuje i nemalé množství jedinečných rukopisů. Jak zmiňované tisky, tak zejména rukopisy jsou opět v majetku nějaké instituce, která rozhoduje o jejich zpřístupnění v originální, nebo spíše v digitalizované podobě.

Mluvíme-li o diachronním korpusu, jeho podoba musí být samozřejmě textová, nikoli ve formě digitálních obrázků. Tím se dostáváme k převodu originálů do digitální podoby: na rukopisy a ani na většinu tisků ze sledovaného období nelze uplatnit moderní způsoby digitalizace, zejména optické rozpoznávání znaků (OCR). V posledních

8 Vedle podrobné anotace dat a vytváření sofistikovaných korpusových manažerů, které by se snažily předjímat potenciální zájmy badatelů, nabízí se ještě možnost poskytnout zájemcům o výzkum zdrojová korpusová data, která by pak mohli obohatit o vlastní potřebné údaje a využít k jejich zpracování specializované nástroje.

letech se objevují také snahy o počítačové rozpoznávání rukopisných záznamů (HTR — Handwritten Text Recognition)[9]. S rozpoznáním rukopisů psaných historickou češtinou zatím nemáme žádné zkušenosti a výstupy programů OCR nejsou použitelné. V každém případě budou výstupy takového rozpoznávání znaků k dispozici v tzv. transliterované formě, tj. v podobě, která zachovává pravopisný systém originálu.

Tím se dostáváme k současnému úzu v české bohemistice, a to ke zpřístupňování textů psaných historickou češtinou v transkripci, která využívá moderní pravidla českého pravopisu (tj. jejich repertoár grafémů, s některými výjimkami, a pravidla pro zachycení tvarů slov) pro vystižení hláskoslovné podoby editovaného textu. Tento přístup má některé výhody, zejména sjednocuje grafickou podobu téhož výrazu v různých pravopisných systémech. Na druhou stranu neumožňuje tento způsob právě výzkum historických pravopisných systémů češtiny moderními metodami. Takovému výzkumu by bylo možné pomoci přípravou transliterované formy a paralelní podoby transkribované, kterou lze v případech, kdy je pravopisný systém originálu relativně stabilní, generovat automaticky pomocí nízkého počtu (zhruba do 50) formálních transkripčních pravidel. Samozřejmě editor textu musí takto automaticky generovaný přepis zkontrolovat. V oddělení vývoje jazyka ÚJČ jsme pro tento účel vyvinuli program Transcriptorium[10], který jsme vyzkoušeli na některých tiscích, naposledy např. na Vusínově slovníku z 1. poloviny 18. století.

Při převodu textů do elektronické podoby je nejdůležitější přínos editora, který na historický text aplikuje své znalosti historické češtiny. Vzhledem k tomu, že je takto připravená edice jedinečná, jedná se z hlediska autorského zákona o autorské dílo, přičemž autorem je editor a jeho souhlas, respektive souhlas nositele majetkových práv je potřeba pro zařazení takového textu do korpusu.

4. EMENDACE

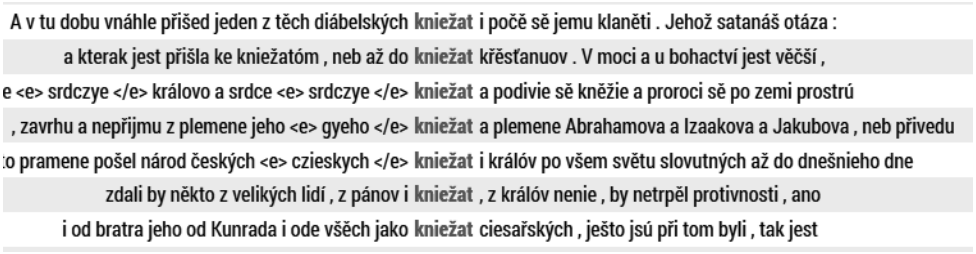
Pokud editor v prameni narazí na porušený text, například v důsledku rozlité tekutiny či řádění hlodavců, nebo na písařskou, popř. tiskařskou chybu, snaží se takové místo emendovat, tj. uvést na správnou, bezchybnou, nebo alespoň na velmi pravděpodobnou podobu, aby byl zachován původní záměr autora nebo jazykový systém pramene. Tím se však narušuje autenticita, tj. podoba, v níž byl pramen reflektován čtenáři; byť i čtenáři mohli při recepci provádět mentálně adekvátní úkony snažící se o nápravu porušeného textu do adekvátnější podoby.[11]

Stavovská čest velí editorům upozornit na případy, kdy opravili, doplnili, emendovali porušené místo; v tištěné produkci tak většinou činí u častějších oprav a úprav souhrnně v ediční poznámce, u jednotlivostí pak v poznámkovém aparátu. Jak se s tím vypořádávají DIAKORP a staročeská textová banka? Každý po svém:

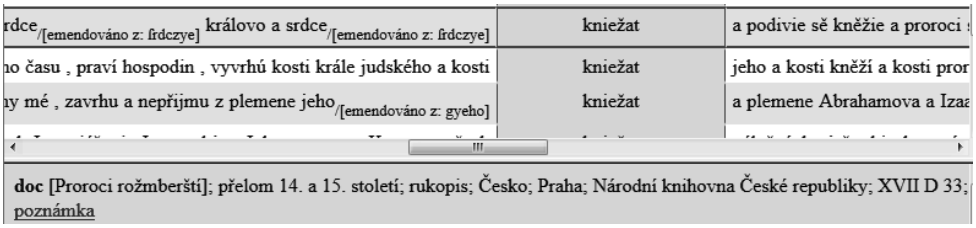
9 Viz např. projekt tranScriptorium (http://transcriptorium.eu).

10 Podrobnější informace o programu a soubory ke stažení jsou dostupné na stránkách Vokabuláře webového <http://vokabular.ujc.cas.cz/moduly/nastroje/transcriptorium/>.

11 Při této příležitosti se sluší zmínit, že autenticita je narušena už tím, že je text převeden z původní pravopisné podoby do novočeského pravopisného systému.



OBRÁZEK 1: Zachycení emendace v DIAKORPu (setříděno podle pozice 1L)

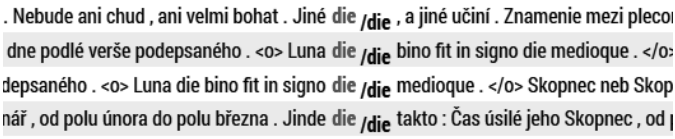


OBRÁZEK 2: Zobrazení emendace ve staročeské textové bance (ve formě komentáře k emendovanému výrazu)

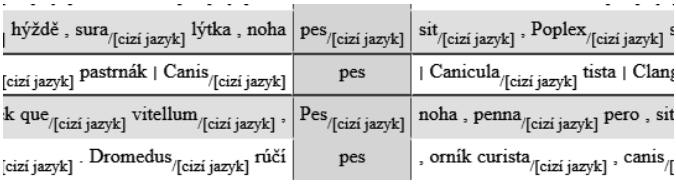
DIAKORP na tyto případy upozorňuje pomocí dvojice kódů, která se uživatelům, ale i korpusovému manažeru jeví jako základní text, tedy tokeny (tj. <e> a </e>), i když jde o metadata, tedy informace přidané editorem k původnímu textu. V důsledku toho se navyšuje počet pozic, respektive slov, daného korpusu, a to nejen o počet samotných značek, ale i o tokeny uvedené mezi těmito značkami. Navíc tyto údaje ovlivňují např. řazení v konkordanci (viz obrázek 1 s doklady setříděnými podle pozice 1L). Staročeská textová banka tyto informace uvádí v komentáři k tokenu, takže údaj není na první pohled patrný, ale zobrazí se na vyžádání (viz obrázek 2). I když zatím staročeská textová banka neumožňuje třídit podle kontextových pozic, údaj o emendaci řazení neovlivní, neboť není na rozdíl od DIAKORPu součástí hlavního textu, tedy na pozici tokenu.

5. CIZOJAZYČNÝ TEXT

Pozornost si v případě diachronních korpusů dále zaslouží cizojazyčný text, který bývá součástí textu staročeského. Jelikož diachronní korpusové texty vznikají plně v režii editora, dochází k jednoznačnému, byť „ručnímu" rozlišení česko- a cizojazyčné části pramene již v rané fázi přípravy. DIAKORP na tyto „nečeské" výrazy upozorňuje pomocí párové značky <o>, která je opět součástí hlavního textu, tedy na úrovni tokenu, a dochází i zde k problémům s řazením a počtem tokenů jako v případě emendace (viz obrázek 3). Staročeská textová banka zvolila opět sobě vlastní řešení, a to uvedení informace ve speciální značce vztažené k tokenu (viz obrázek 4). Z pohledu badatele však ani jedno řešení není dokonalé. V obou případech totiž



OBRÁZEK 3: Vyhledání staročeských prvků v DIAKORPu zároveň s cizojazyčnými



OBRÁZEK 4: Zachycení cizojazyčných prvků ve staročeské textové bance

korpusové manažery implicitně vyhledávají v celém korpusu textů bez rozlišení jazyka. Tj. při hledání výrazu *pes* se nalezne jak staročeské označení pro domácí zvíře, tak latinský ekvivalent staročeského slova „noha". Zatímco však ve staročeské textové bance lze pomocí kombinace atributů vyhledávání omezit na český nebo cizojazyčný text ("pes" and jazyk!= "\[cizí jazyk\]" nebo "pes" and jazyk == "\[cizí jazyk\]"), v DIAKORPu je možné takovou podmínku formulovat, pouze pokud známe pozici cizojazyčného výrazu vzhledem k celému úseku. Např. dotaz [word!="<o>"] [word="luna"] zajistí, že se do výsledků vyhledávání cizojazyčný výraz *luna* nedostane, ale pouze v tom případě, kdy je na začátku cizojazyčného úseku; bude-li token *luna* na druhé nebo nějaké z následujících pozic, nelze tímto dotazem jeho vyřazení z výsledků vyhledávání zajistit. Obecné doporučení tedy zní: cizojazyčný text by měl být explicite označen pomocí atributu, aby jej manažery mohly z vyhledávání vyřadit. Ve výchozím nastavení by mělo být vyhledávání v cizím jazyce zakázáno, teprve po zásahu uživatele by mělo být možné omezit hledání na cizojazyčný text nebo toto omezení zrušit a vyhledávat bez rozdílu jazyka.

6. PŘÍPISKY MLADŠÍ RUKOU

Obdobný problém jako u cizího jazyka je potřeba řešit u tzv. mladších přípisků, které se v textu objevily dlouho po vzniku základního textu pramene. Není-li takový přípisek nijak označen, stane se nedílnou součástí hlavního textu a zkreslí tím informace o jazyce daného období. DIAKORP i staročeská textová banka takové přípisky ve zdrojových datech pro korpus vypouštějí. Tím se na jednu stranu zachovává autenticita textu v období jeho vzniku, narušuje se však podoba, v níž byl pramen reflektován pozdějšími čtenáři. Pokud se mají mladší přípisky stát součástí korpusu, bude nutné označit jednotlivé připsané výrazy dobou svého vzniku, aby na ně bylo možné aplikovat filtrační kritéria doby vzniku. Vezmeme-li v úvahu časové hledisko vzniku různých vrstev his-

torického textu, tedy např. základního textu a přípisku o 100 let později, měly by se přípisky stát součástí výsledků (ale také statistik) pouze v tom případě, že časové omezení subkorpusu nebo dotazu zahrnuje také období vzniku, popř. prosté existence přípisků. Z hlediska implementace se opět jako vhodné jeví uvádět časový údaj o vzniku přípisku u jednotlivých tokenů a přizpůsobit korpusový manažer, aby s touto eventualitou počítal. Toto řešení ovšem také znamená, že pokud časové omezení nebude zahrnovat období vzniku základního textu, objeví se ve výsledku pouze samotné přípisky, které mohou být útržkovité.

7. PŘÍPISKY SOUDOBOU RUKOU

Obdobný problém bude nutné řešit u soudobých přípisků, tj. takových, které vznikly ve stejné, popř. časově ne tak vzdálené době jako základní text pramene. Pokud tyto přípisky netvoří organickou součást textu (např. v případě doplnění vynechaných slov), musí editor rozhodnout o umístění přípisku (podle smyslu nebo podle místa připsání), což může v některých případech vést k narušení větné stavby a k vytváření neobvyklých kolokací (se základním textem). I tady se jako řešení nabízí použití metainformace u jednotlivých pozic, které reprezentují text přípisku. Složitost problému přípisků, zejména pokud jde o kontinuitu větné stavby, však vyžaduje další diskusi a spolupráci na straně editorů, programátorů korpusových manažerů a badatelů.

8. FREKVENČNÍ ÚDAJE

V souvislosti se zahrnutím nebo nezahrnutím různých částí původního textu do korpusového zpracování je potřeba pečlivou pozornost věnovat frekvenčním údajům o zastoupení jednotlivých prvků v korpusu a od nich odvozeným statistikám. Jak už bylo řečeno výše, v DIAKORPu například zvolený způsob označování emendací vede k navýšení počtu tokenů o vyloženě nejazykové prvky. Podobně je potřeba řešit problematiku cizojazyčných pasáží v korpusových materiálech: tyto prvky v současné implementaci navyšují počet tokenů a zkreslují tím pádem statistiky českých výrazů. Ve staročeské textové bance je například ve verzi z 24. 6. 2014 celkem 3 701 567 tokenů, z toho je však 52 945 tokenů cizojazyčných. Ze zbývajících 3 648 622 staročeských tokenů pak představuje 594 649 výskytů interpunkce a 386 číslice. Staročeská textová banka tedy obsahuje 3 053 587 staročeských slov, tj. 82,5 % celkového počtu tokenů. Korpus historických textů by měl mít s ohledem na výše uvedené příklady jasně stanovenou metodiku, jaké prvky bude zahrnovat do svých statistických údajů a výpočtů.

9. LEMMATIZACE

Dalším důležitým tématem historických korpusů je lemmatizace, přestože ani jeden ze zmiňovaných zdrojů diachronních textů ji zatím nenabízí. Lemmatizace usnadňuje hledání všech tvarů jednoho lemmatu. U diachronních korpusů hraje význam-

nou roli proměnlivost hláskoslovné podoby velké části lemmat v čase (např. podoby *kóň, kuoň, kůň*). Tento problém, tj. zjednodušení hledání různých hláskoslovných podob téhož lemmatu, řešíme pomocí tzv. hyperlemmatu. Pracovníci ÚČNK počítají s uváděním hyperlemmatu v hláskoslovné podobě, která má co nejblíž současné češtině (tj. *kůň*); pracovníci ÚJČ naproti tomu pro hyperlemma zvolili podobu nejstarší, tj. ve formě k roku 1300 (tedy *kóň*). Obě varianty mají své opodstatnění, a jako vhodné se proto jeví uvádět u tokenu dvě hyperlemmata: historické a moderní.

10. HLÁSKOVÉ ZMĚNY

Poslední bod hodný pozornosti je značkování hláskoslovných změn proběhnuvších na určitém tokenu. Při lemmatizaci, která reflektuje rovněž hláskoslovný vývoj jazykových jednotek, by zejména pro starší vývojové fáze jazyka bylo vhodné zaznamenávat, k jakým hláskoslovným změnám došlo. V tomto případě je potřeba zaznamenat, nejen k jaké změně došlo, ale rovněž na jakém místě, neboť v rámci jednoho výrazu mohla tatáž změna nastat i nenastat, viz např. změna *ó* > *uo* v dokladu *dóstuojenstvie*[12]. Uspokojivé řešení tohoto problému se mi doposud nepodařilo nalézt.

POUŽITÉ KORPUSY

Kučera, K. — Stluka, M. (2011): *DIAKORP: Diachronní korpus, verze 5 z 21. 2. 2011*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z: <http://www.korpus.cz>

Staročeská textová banka [on-line]. Verze z 24. 6. 2014 [citováno ze dne 30. 11. 2014]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupná z <http://vokabular.ujc.cas.cz/banka.aspx>.

LITERATURA

Černá, A. M. — Lehečka, B. (2012): Elektronická příprava a publikace starších českých textů. In: *Čeština v pohledu synchronním a diachronním: stoleté kořeny Ústavu pro jazyk český*. Vyd. 1. Praha: Karolinum, s. 265–269.

Stluka, M. (2010): Zpracování jednotlivých pravopisných systémů v rámci diachronní složky Českého národního korpusu. In: M. Čornejová — L. Rychnovská — J. Zemanová, J. (eds.), *Dějiny českého pravopisu (do r. 1902): sborník příspěvků z mezinárodní konference Dějiny českého pravopisu (do r. 1902) 23.–25. září 2010, Brno, Česká republika*. Brno: Host; Masarykova univerzita (vyšlo 2011), s. 435–448.

Boris Lehečka | oddělení vývoje jazyka Ústavu pro jazyk český AV ČR | Valentinská 1, 116 46 Praha 1
boris@daliboris.cz

12 Doklad pochází z latinsko-českého slovníku z rukopisu ostřihomského z 2. poloviny 15. století; vyskytuje se na straně 61.