ČASOPIS PRO MODERNÍ FILOLOGII 100, 2018, Č. 1, S. 60–79

OPEN ACCESS

Diachronní korpusová lingvistika a španělština: současný stav a problémy*

Zuzana Krinková (Praha)

HISTORICAL CORPUS LINGUISTICS AND SPANISH: THE STATE OF THE ART AND CURRENT PROBLEMS

The first aim of the article is to address major problems of current historical corpus linguistics such as representativeness in genre, place and time, transcription of historical texts, etc. The second goal is to introduce the reader to traditional and innovative historical corpora of Spanish, focusing on their characteristics, advantages and limitations.

KEYWORDS

Spanish, corpus linguistics, historical corpora, diachrony

KLÍČOVÁ SLOVA

španělština, korpusová lingvistika, diachronní korpusy, diachronie

1. ÚVOD

Korpusová lingvistika bývá považována za disciplínu stále ještě relativně mladou, zato však velmi rychle a dynamicky se rozvíjející. Její vznik a rozmach je úzce provázán s rozvojem moderních počítačových a digitálních technologií a internetu[1], díky nimž lze v současné době v kteroukoli dobu a z jakéhokoli počítače konzultovat na korpusech obrovské množství dat[2], v mnoha případech navíc i lemmatizovaných

* Tento článek vznikl za podpory projektu Univerzity Karlovy Progres č. 4, *Jazyk v proměnách času, místa, kultury*.

1 Vznik korpusů první generace se datuje sice již do šedesátých let 20. století, nicméně hlavní rozmach korpusové lingvistiky je svázán s rozšířením počítačů v 90. letech. Do tohoto období spadá vznik korpusů druhé generace, oproti první generaci podstatně rozsáhlejších, k nimž se řadí i španělský referenční korpus CREA (*Corpus de referencia del español actual*) a diachronní korpus CORDE (*Corpus diacrónico del español*). Třetí generace korpusů (po roce 2000) je spjata zejména s rozšířením internetu a digitalizace a využívá mj. i texty dostupné na webu. Typově sem spadá např. synchronní korpus španělštiny CORPES XXI (*Corpus del Español del Siglo XXI*) čerpající mj. i z blogů na webu. Podrobněji o historii španělské korpusové lingvistiky pojednává Rojo (2016).

2 Bez ohledu na možnosti digitálního zpracování dat v uplynulých dvou či třech desetiletích je však třeba připomenout, že empirické jazykové bádání založené na studiu konkrétních textů má mnohem delší tradici. Konkrétní a reálné textové (nebo též — ve strukturalistickém pojetí — parolové) vzorky se tradičně využívaly i v minulosti zejména v lexikografii, v určité míře i v gramatickém popisu jazyka. V současné době je využití korpusů při tvorbě slovníků a gramatik současného jazyka takřka obligátní.

a anotovaných na morfologické i syntaktické úrovni. Zároveň došlo v uplynulých letech i ke značné diverzifikaci korpusů a vedle rozličných korpusů současného jazyka vznikla i řada korpusů zachycujících jeho minulá stádia[3].

Dá se říci, že v současné korpusové lingvistice, ať už synchronního či diachronního zaměření, rozeznáváme několik relativně nezávislých subdisciplín, které z důvodu odlišného zaměření sledují různé cíle a potýkají se s odlišnými problémy[4].

První takovouto subdisciplínou je samotná výstavba korpusu, která se zabývá všemi kroky vedoucími od prvotního výběru a získávání primárních textů až ke konečné prezentaci daného prohlížeče, který uživateli umožňuje korpusová data konzultovat. Články a studie vznikající na toto téma jsou relativně četné, často psané samotnými autory korpusů či jejich úzkými spolupracovníky. Jejich cílem je především nový či teprve vznikající korpus podrobně představit a v neposlední řadě i vyzdvihnout jeho přednosti (zejména oproti jiným korpusům).

Druhá subdisciplína je spjata na jedné straně s informatikou (zabývá se např. automatickým rozpoznáváním textu, značkováním apod.), na straně druhé pak s matematickým a statistickým zpracováním konkrétních dat. I tyto studie jsou četné a pro ostatní uživatele korpusu mnohdy užitečné, neboť poskytují návod, jak získaná data správně vyhodnocovat. V mnoha případech se jedná o studie kritické, upozorňující na úskalí konkrétních korpusů a chyby v jiných studiích z pohledu statistického vyhodnocování dat.

Třetí subdisciplína spadající do korpusové lingvistiky má nejblíže k samotnému jazyku jako předmětu svého hlavního zájmu. Autoři těchto ryze lingvistických studií zkoumají rozličné konkrétní jazykové jevy, přičemž své poznatky primárně zakládají

3 V diachronní rovině bylo tradiční filologické bádání ze své podstaty vždy založeno na zkoumání písemných památek. Filologické zkoumání jednoho konkrétního dokumentu se však od korpusového zkoumání metodologicky liší. Hlavní rozdíl spočívá v tom, že na korpusu se nezkoumají celé texty, nýbrž pouze úryvky textů. To na jedné straně přináší výhodu zkoumat velké množství dat najednou (a nahlédnout tak uceleněji do *langue*), na druhou bývají korpusová data vytržena z širšího kontextu, následkem čehož může dojít k jejich zkreslené interpretaci. Nelze tedy říci, že by korpusová lingvistika zcela nahradila tradiční filologické metody, spíše je doplňuje. V historické lingvistice zároveň existuje i metoda, která nestaví výhradně na dochovaných textech: v jazykové rekonstrukci totiž naopak doložené, písemně zaznamenané tvary často chybějí (v případě španělštiny se jedná zejm. o nedoložené tvary vulgární latiny či nejstarších fází iberské románštiny).

4 Příkladem — srov. Kabatek (2016, s. 3) — může být na jedné straně přání uživatele korpusu pracovat s co největším množstvím volně dostupných dat, perfektně zpracovaných po lingvistické stránce, což znamená nejen lemmatizaci a anotaci, které u diachronních korpusů vykazují specifické problémy, ale i filologicky věrohodný přepis starých textů. Na straně druhé se tvůrce korpusu potýká s limity po stránce technické, časovými, personálními a finančními možnostmi, je omezován autorskými právy apod. Toto ovšem neznamená, že dané subdisciplíny nejsou provázané. Tvůrce korpusu se přirozeně snaží co nejvíce naplnit očekávání uživatelů, ve většině případů se navíc tvůrce sám často ocitá i v roli uživatele a publikuje lingvistické studie založené na svém korpusu.

na datech získaných z korpusů. Kvalita studií tohoto typu, které jsou zastoupeny asi v nejhojnějším počtu, je značně různorodá[5].

Diskusní platformu pro výše zmíněnou problematiku tvoří samozřejmě zejména prestižní mezinárodní časopisy věnované buď přímo korpusové lingvistice, např. *Corpora, Journal of Corpus Linguistics*, případně problematice související obecně s digitalizací dat (např. *Digital Humanities*) a zpracováním dat (např. španělský časopis *Procesamiento del lenguaje natural*). Četné články zabývající se španělskou či latinskoamerickou korpusovou problematikou lze najít též např. v chilském časopise *Revista de lingüística teórica y aplicada* a v dalších lingvisticky či filologicky orientovaných časopisech španělské i jiné provenience. Diachronně zaměřený je španělský časopis *Scriptum Digital*, který se zabývá teoretickými a metodologickými aspekty digitalizace a všemi výše zmíněnými subdisciplínami v diachronní perspektivě. O korpusové lingvistice v souvislosti se španělštinou se pojednává též na četných konferencích a kolokviích. Z těch zaměřených na diachronii jmenujme alespoň jedno kolokvium pořádané již tradičně na španělských i jiných evropských univerzitách: *Congreso Internacional de Corpus Diacrónicos en Lenguas Iberorrománicas* (Palma de Mallorca 2007, Barcelona 2011, Curych 2014, Alcalá 2016)[6].

Z oblasti korpusové lingvistiky vyšlo v uplynulých letech také množství knižních publikací, nejčastěji (avšak nejen) sborníkového charakteru, z nichž řada příspěvků se věnuje i korpusům španělštiny a diachronii[7]. S korpusy v drtivé většině případů pracují i autoři nově vznikajících gramatik současného jazyka[8] a slovníků[9]. Souhrnná historická mluvnice založená na diachronním korpusu však dosud nevznikla[10].

Tento článek si klade dva hlavní cíle. V první části vytýčíme určitý teoretický rámec, do nějž zasadíme možnosti a limity korpusové lingvistiky obecně a zejména

5 V našem příspěvku se budeme z důvodu omezeného rozsahu zabývat především první výše zmíněnou subdisciplínou korpusové lingvistiky, neboť výstavbu korpusu (zejména co se týče výběru textů) považujeme za zcela zásadní. O druhém a třetím okruhu, zejména o statistickém zpracovávání dat z historických korpusů a možnostech a limitech diachronních korpusových studií, pojednáme v jiném článku.

6 Z třetí konference vzešla publikace *Lingüística de Corpus y Lingüística Histórica Iberorrománica*, ed. J. Kabatek (2016).

7 Za všechny uveďme alespoň ryze diachronně orientovanou publikaci *New Methods in Historical Corpora* (P. Bennett — M. Durrell a kol., 2013), jejíž příspěvky pocházejí ze stejnojmenné konference pořádané v Manchesteru v r. 2011, a dále kolektivní publikaci zaměřenou na jazykovou variantnost španělštiny *Working with Spanish Corpora* (Parodi, 2007).

8 Jako příklad uveďme alespoň dvě gramatiky využívající kromě jiných zdrojů i španělský referenční korpus CREA. První je rozsáhlá *Nueva gramática de la lengua española* (2009–11) vydaná Španělskou královskou akademií (RAE). Druhým a pro českého čtenáře užitečným příkladem je *Mluvnice současné španělštiny* (Zavadil — Čermák, 2010).

9 Na tomto místě se nabízí jako příklad zejména zcela recentní *Nuevo diccionario histórico del español (NDHE)*, online slovník diachronního zaměření využívající mj. i korpusy CORDE a CREA, který je možné konzultovat na stránkách RAE (http://www.rae.es/recursos/diccionarios/nuevo-diccionario-historico). Slovník není v současnosti kompletní.

10 Podrobněji o problematice historické lingvistiky založené na korpusu pojednává Kabatek (2013).

pak korpusové lingvistiky diachronní. Ta se totiž potýká s několika specifickými problémy, jež nejsou vlastní korpusové lingvistice synchronní (zaměřené na současný jazyk). Druhá část článku si klade za cíl poskytnout čtenáři aktuální přehled a stručnou charakteristiku jednotlivých diachronních korpusů španělštiny, které jsou v současnosti volně dostupné na Internetu, a upozornit na jejich klady a nedostatky[11].

2. OBECNÉ PROBLÉMY DIACHRONNÍ KORPUSOVÉ LINGVISTIKY

2.1 VYVÁŽENOST A REPREZENTATIVNOST KORPUSU

Asi nejvíce diskutovaným tématem v oblasti diachronní korpusové lingvistiky posledních let je reprezentativnost dat v korpusu a otázka vztahu mezi korpusovými daty a vývojem jazyka[12]. Reprezentativnost korpusu je však obecným problémem, který řeší i synchronní korpusová lingvistika, neboť korpus je vždy jen soubor vybraných vyprodukovaných textů, tedy soubor jazykových vzorků, nikoli jazyk jako takový v jeho celistvosti[13].

Ve srovnání s diachronními korpusy jsou však synchronní korpusy současného jazyka v určité výhodě: oproti minulým staletím vzniká v současnosti celkově více psaných textů (nabízí se tedy mnohem širší výběr „reprezentativních" textů), navíc je možné díky moderním technologiím zachytit také orální jazyk[14].

11 Naším původním záměrem bylo demonstrovat výhody a omezení jednotlivých korpusů též na příkladech konkrétních jazykových studií. Z důvodu omezeného rozsahu článku však uvedeme jen velmi základní údaje. Podrobněji se jednotlivými typy studií, které lze s větším či menším úspěchem realizovat na korpusech, a konkrétními příklady budeme zabývat v jiném článku.

12 Srov. též Kabatek (2016, s. 3).

13 V této souvislosti je třeba si mj. uvědomit, že drtivá většina všech vyprodukovaných textů/promluv každého jazyka je orálního charakteru, zatímco psaný jazyk tvoří i v současnosti menšinu. Orální jazyk je zároveň nejčastějším nositelem jazykových inovací, jeho význam je tudíž v jazykovém popisu zcela nesporný. I přes tuto skutečnost je orální složka zastoupena v korpusech v poměrově mnohem menší míře vzhledem ke skutečnosti (např. ve španělském referenčním korpusu CREA tvoří orální složka pouhých 10 %). V diachronních korpusech orální produkce samozřejmě zcela chybí.

14 Většina orální složky zastoupené v korpusech je nicméně formálního či poloformálního charakteru (televizní debaty, rozhovory, přednášky apod.). Nejběžnější neformální promluvy mezi rodinnými příslušníky či přáteli jsou stále zastoupeny zcela minimálně. Na druhou stranu však orální složka může být nepřímo obsažena, jak je všeobecně známo, v beletrii či divadelních hrách v podobě promluv jednotlivých postav (což je jistě dobrá zpráva i pro diachronní korpusovou lingvistiku) a dále také do určité míry i ve spontánním psaném projevu. V tomto kontextu je využití internetových blogů, diskuzí apod. jistě velmi žádoucí. V diachronních korpusech může obdobnou úlohu plnit např. soukromá korespondence.

Enrique-Arias (2012, s. 96) uvádí zásadní tezi týkající se zdánlivého paradoxu výstavby velkých „reprezentativních" diachronních korpusů: diachronní korpus by měl být na jedné straně heterogenní (tj. obsahovat texty různých autorů, žánrů, stylů, dialektů), na druhé straně by však měl být homogenní v tom smyslu, že jednotlivá chronologická období by měla být mezi sebou srovnatelná (v ideálním případě v počtu tokenů i v poměrovém zastoupení různých typů textů)[15]. Splnění tohoto nároku kladeného na diachronní korpus se pojí s celou řadou problémů, o nichž stručně pojednáme níže.

2.1.1 ROZDĚLENÍ TEXTŮ PODLE ŽÁNRU

Synchronní i diachronní korpusy zpravidla pracují s rozdělením textů podle žánrů. Objemy textů daných žánrů jsou vůči sobě obvykle určitým způsobem poměrově zastoupeny. Rozdělení na žánry může samo o sobě představovat problém (některé texty mohou být obtížně zařaditelné, jedná se o žánry na pomezí apod.), bylo by však jistě na druhou stranu nežádoucí (právě i kvůli již zmíněnému relativně vyváženému poměrovému zastoupení, o nějž se tvůrci korpusu snaží), aby klasifikace textů byla v korpusu příliš podrobná, nepřehledná či roztříštěná. Různé žánry se od sebe mohou lišit svým stylem, neboť vycházejí z různých textových tradic[16]. V synchronních a diachronních korpusech se žánrové zastoupení textů či alespoň jejich poměr většinou liší. Např. poezie, která je jazykově specifická (četné metafory, neobvyklá slovní spojení, netypický slovosled, fonetické odchylky z důvodu zachování rýmu apod.) bývá v synchronních korpusech zastoupena ve velmi malé míře či vůbec[17]. Odlišná je však situace v diachronních korpusech, neboť i beletrie a další narativní žánry byly ve středověku často veršované a mnohdy je právě tato literatura v diachronních korpusech relativně četně zastoupena.

Diachronní korpusová lingvistika se navíc potýká s problémem srovnatelnosti a proměnou jednotlivých žánrů v čase. Jak poznamenává Enrique-Arias (2012:96), je otázkou, do jaké míry je metodologicky správné srovnávat např. jazyková data ze středověkých kronik s romány 19. století, byť se v obou případech jedná o narativní žánry.

15 Většina diachronních korpusů obsahuje různě velké vzorky dat, ať už je proměnnou čas, žánr či jiná veličina. S touto skutečností si zpravidla poradí příslušné použité statistické metody. Problémem však zůstává, že v mnohých korpusech nenajdeme přesné informace o velikosti dat.

16 Podrobněji o problematice rozdělení textů na žánry a od nich se odvíjejících textových tradicích viz Kabatek (2013, s. 16–19).

17 Toto ovšem neznamená, že by korpusový výzkum jazyka poezie byl z lingvistického hlediska nezajímavý či irelevantní. Svědčí o tom např. i recentní projekt *Korpus českého verše* (http://versologie.cz/v2/web_content/corpus.php?lang=cz) realizovaný versologickým týmem na Ústavu pro českou literaturu AV ČR. Pro výzkum literárního jazyka, ať už poezie nebo prózy, je korpus nástrojem přirozeně nejvhodnějším. Z tohoto pohledu se tedy může nízké poměrové zastoupení poezie ve velkých referenčních synchronních korpusech jevit jako poněkud nešťastné. Nicméně je třeba dodat, že existují či v současnosti vznikají (v českém i španělském prostředí) specializované autorské korpusy. Při současném boomu vzniku rozličných specializovaných korpusů lze dle našeho názoru předpokládat, že v dohledné budoucnosti jistě vznikne i korpus zaměřený na španělskou poezii.

2.1.2 DIASTRATICKÁ A DIATOPICKÁ VARIANTNOST

V ideálním případě by měl být každý korpus aspirující na reprezentativnost jazyka v určitém období či napříč časem vyvážený také z pohledu diastratického a diatopického. Tato podmínka však bývá splněna jen velmi vzácně a údaje o geografickém či sociálním původu (i jakékoli další sociolingvistické parametry) o mluvčích či autorech textů se v korpusu obvykle nevyskytují či jsou velmi obecné[18].

Nabízí se otázka, do jaké míry jsou takové údaje pro korpusovou lingvistiku vůbec relevantní. Obecně lze říci, že zatímco některé jazykové jevy (zejména z pohledu synchronního zkoumání) jsou na diatopické a diastratické proměnlivosti nezávislé, jiné jevy naopak v tomto směru vykazují variantnost.

Synchronní lingvistika je v tomto ohledu opět v nemalé výhodě, neboť v případě potřeby může vazbu určitého jevu na diatopické či diastratické parametry ověřit jinými metodami (terénním výzkumem mezi mluvčími apod.). Zároveň mohou relativně snadno vznikat i menší korpusy různých současných sociolektů či lokálních mluv orálního charakteru[19]. Také lze konstatovat, že např. současná psaná španělština, která tvoří podstatnou část velkých korpusů, je do značné míry standardizována a z pohledu diatopického vykazuje relativně malou variantnost.

Poněkud odlišná situace je v diachronní lingvistice, která je metodologicky odkázána pouze na zkoumání psaných textů. Zejména ve vzdálenější minulosti zaznamenáváme oproti dnešku mnohem větší jazykovou variantnost, která je ve většině případů vázána právě na geografický původ pisatele textu. Během jazykového vývoje obvykle se stává, že některé z jazykových fenoménů vykazujících variantnost začínají postupem času převažovat a posléze se prosadí na úkor ostatních, které vyjdou z užívání. Úkolem diachronní lingvistiky by mělo být nejen konstatovat, ve kterém časovém období k určité jazykové změně došlo, ale pokud možno též odhalit i příčinu dané jazykové změny. Ta sice někdy tkví v samotném jazykovém systému, velmi často však souvisí s vnějšími okolnostmi jazykového vývoje, např. historickými událostmi (obecným příkladem prosazení určitých lokálních variant na úkor jiných může být změna hlavního města apod.[20]). Diatopický parametr a jeho vyváženost v diachronním korpusu je proto z našeho pohledu velice podstatný[21].

18 Synchronní korpus CREA ani diachronní korpus CORDE např. neuvádějí jiný geografický údaj než zemi, kde text vznikl.

19 Některé takové orální korpusy (COVJA, ACUAH aj.) jsou zahrnuty do CREA. Z dalších španělských specializovaných korpusů různých sociolektů jmenujme např. COLA (*Corpus oral de lenguaje adolescente*), ALCORE (*Corpus para el estudio del lenguaje juvenil*), COJEM (*Corpus Oral Juvenil del Español de Mallorca*), Val.Es.Co (*Valencia — Español Coloquial*). Korpusy současných sociolektů a lokálních mluv přesahují téma tohoto článku, zájemce nalezne více informací např. v Rojo (2016).

20 V případě španělštiny je asi nejznámějším a nejtypičtějším příkladem vnější příčiny jazykových změn *reconquista* (zpětné vytlačování Maurů z Pyrenejského poloostrova), která začala na severu Pyrenejského poloostrova a postupovala směrem na jih. *Reconquista* má za následek prosazení kastilštiny v oblastech původně nekastilských a narušení jazykového kontinua na Pyrenejském poloostrově.

Problémem mnohých starých textů obsažených v korpusech je, že jejich autorství je sporné, některé texty byly opisovány postupně i více autory a geografický původ textů tak lze určit právě jen na základě jazykových rysů daného dokumentu.

Je však velmi pozitivní, že v posledních letech vznikají ve španělském jazykovém prostředí i diachronně zaměřené korpusy, které diatopický a diastratický faktor neopomíjejí[22].

2.1.3 ČASOVÁ ROVINA

Diachronní korpus by měl v ideálním případě poskytnout vyvážený obraz jazyka napříč časem. Jako relativně triviální, v praxi však bohužel často zanedbaný, se proto jeví požadavek na podobný počet tokenů v jednotlivých časově vymezených obdobích. V mnohých diachronních korpusech se kvantita textů v různých obdobích značně liší. Statistické metody si s rozdílnou velikostí vzorků dat sice zpravidla snadno poradí, nemohou však řešit nulový výskyt hledaného jevu, který je důsledkem nedostatečně zastoupeného období. Mnohé korpusy navíc neposkytují údaje o počtech tokenů v jednotlivých časových obdobích.

2.2 PŘEPIS STARÝCH TEXTŮ

Diachronní korpusová lingvistika se musí vypořádat i dvěma dalšími problémy zásadního rázu, které nejsou přítomné u lingvistiky synchronní. Jedná se o přepis a dataci starých textů. Při přepisu starých textů (zejména středověkých manuskriptů) se objevují specifické obtíže: texty např. obsahují značné množství rozličných ligatur a zkratek, pravopis se může i v rámci jednoho díla lišit, často chybí interpunkce, některé skupiny slov mohou být psány dohromady atd. Texty navíc mohou být místy poškozené, neúplné či nečitelné. Všechny tyto aspekty mohou následně vést k případným chybám v transliteraci[23]. V dokumentu se také mohou objevit chyby písařů či kopistů[24].

V ranější fázi moderní korpusové lingvistiky se pro diachronní korpusy nejčastěji využívala buď literární díla, o nichž se předpokládalo, že jsou pro vývoj jazyka stěžejní (ve španělštině např. *Cid*, *El conde Lucanor*, *La Celestina*, *Quijote*), nebo vydané texty ne-

21 Diatopická nevyváženost korpusu může zkreslit výsledky diachronního výzkumu jazykových změn i z hlediska časového.

22 Takovým je např. korpus CODEA a mnohé jiné, menší specializované korpusy. Pojednáme o nich ve třetím oddílu.

23 V tomto článku chápeme termín transliterace jako přepis psacích písmen manuskriptu do písma tiskacího. Příkladem takového chybného přepisu (cf. Kabatek, 2016, s. 6) je dle údajů v CORDE časný italianismus *mafia* (objevuje se již v 16. století), který je však ve skutečnosti chybným přepisem slova *maña*. Přestože se již jedná o známý případ, v CORDE i v korpusu Marka Daviese (*Corpus del español*) je tento tvar stále přítomný. V recentním slovníku *Nuevo Diccionario Histórico*, který zahrnuje i texty z CORDE, je uvedený příklad již opraven.

24 Chyby, jichž se dopouštěli písaři při opisu děl, jsou jiného rázu než chyby v transliteraci, obvykle jsou relativně snadno rozpoznatelné a nijak neodrážejí stav jazyka dané epochy. Naproti tomu fenomény jako anakoluty, které se rovněž mohou vyskytnout v manuskriptech, mohou mít lingvisticky vypovídající hodnotu. Podrobněji o tomto problému viz Kabatek (2013, s. 10–12).

literární, které byly již k dispozici. Tento přístup s sebou přinášel mnohé problémy: slavné literární texty nemusejí být po jazykové stránce typickými představiteli své doby, v některém období bylo k dispozici hodně materiálu jen jednoho určitého typu (zákoníky), což mělo za následek žánrovou nevyrovnanost. Ve všech případech se jednalo o kritická vydání děl, tj. texty byly již určitým způsobem přepsány a upraveny[25]. Korpusy tak z tohoto hlediska obsahovaly texty různorodé povahy, neboť jednotlivá kritická vydání starých textů se mohla i značně lišit po stránce editorských úprav.

V současnosti se v mnoha případech však dokumenty vybírají a připravují již přímo za účelem jejich zahrnutí do korpusu. Důležitou roli tu sehrává postupující digitalizace archivů, jež umožňuje snadný přístup k rozličným starým textům.

Dá se říci, že dnešní iberorománská diachronní korpusová lingvistika se čím dál tím více přiklání k edičním kritériím proklamovaným sítí CHARTA[26], podle nichž se korpus již neomezuje na jediné vydání starého textu, nýbrž nabízí vedle kritického vydání (upraveného na základě jednotných kritérií) i paleografickou verzi s transliterací manuskriptu a dále též fotografický náhled originálního dokumentu, který umožňuje badateli ověřit správnost přepisu.

2.3 DATACE STARÝCH TEXTŮ

Závažný problém, s nímž se potýkají zejména dva velké diachronní korpusy CORDE a *Corpus del Español*[27], představuje datace starých textů. Mnohdy se zaměňuje datum vzniku existujícího či domnělého originálu s datem vzniku manuskriptu či datem knižního vydání textu. Mnohá stará díla jsou navíc zachována jen v pozdějších opisech[28]. Nejasná kritéria pro dataci textů mohou být posléze příčinou chybných závěrů řady studií a zavést diachronní bádání do slepé uličky.

3. PŘEHLED A CHARAKTERISTIKA DOSTUPNÝCH DIACHRONNÍCH KORPUSŮ ŠPANĚLŠTINY[29]

V současné době existuje celá řada diachronních korpusů zaměřených na evropskou či mimoevropskou španělštinu. Pro větší přehlednost je zde rozdělíme do pěti skupin.

25 Vybrané kritické edice starých děl byly následně skenovány a ručně upravovány. Kvůli zachování koherence textu byl např. vypuštěn poznámkový aparát.

26 Red Internacional CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos), podrobná kritéria se nacházejí zde: http://files.redcharta1.webnode.es/200000023-de670df5d6/Criterios%20CHARTA%2011abr2013.pdf

27 Kvalitativně ještě závažnější je z tohoto pohledu datace děl v rozsáhlém diachronním korpusu GoogleBooks.

28 Za všechny uveďme dva příklady: *Cid* v CORDE nese dataci pol. 12. století, byť originál z této doby není zachován. Stejně tak *Calila e Dimna* se datuje do 13. století, i když oba dochované opisy pocházejí z 15. století.

29 O diachronních korpusech dalších jazyků Pyrenejského poloostrova v tomto článku nepojednáváme, zájemce odkazujeme na následující webové stránky: CICA (*Corpus Informati-*

První skupinu tvoří dnes již tradiční CORDE a *Corpus del Español*. Jedná se o největší diachronní korpusy španělštiny, které lze v jistém smyslu považovat za referenční. Svou výstavbou spadají ještě mezi korpusy starší generace. Do druhé skupiny jsme zařadili korpusy paralelní. Třetí a čtvrtá skupina korpusů (nejčastěji velmi recentních) využívá všeobecného rozmachu digitalizace archivních dokumentů, které jsou za účelem začlenění do korpusu pečlivě vybírány a dále zpracovávány podle jednotných kritérií. Poslední pátou skupinu pak tvoří (rovněž recentní) korpusy využívající texty z webu (mezi ně patří i rozsáhlé korpusy Google Books).

3.1 VELKÉ REFERENČNÍ KORPUSY

3.1.1 CORDE (CORPUS DIACRÓNICO DEL ESPAÑOL)[30]

Asi nejznámějším a zároveň největším a nejstarším diachronním korpusem je CORDE vytvořený Španělskou královskou akademií (RAE) již koncem 90. let. Korpus CORDE, v současnosti již uzavřený, obsahuje texty rozličných žánrů a témat datované od 8. století do roku 1975 (celkem zhruba 250 milionů tokenů[31]). Texty pocházejí z různých oblastí, kde se v současnosti hovoří či v minulosti hovořilo španělsky. Geografická provenience textů je omezena na údaj o (současném) státu. Dle údajů na stránkách korpusu pochází 74 % textů ze Španělska. Z hlediska časového je korpus rozložen nerovnoměrně v tomto poměru: 21 %[32] pochází z období do r. 1492 (*Edad Media*), dále 28 % je věnováno Zlatému věku (*Siglos de Oro*, zde konkrétně 1493–1713), 51 % je z moderního období (*Época contemporánea*, zde 1714–1975). Podrobnější údaje o počtu tokenů nejsou k dispozici. Jak již bylo uvedeno výše, i samotná datace jednotlivých textů je místy problematická. Problematický je také různorodý přepis jednotlivých textů. Cenná je v tomto směru zcela recentní rozsáhlá studie od Rodrígueze Moliny a Octavia de Toledo y Huerta (2017), v níž autoři hodnotí dokumenty obsažené v CORDE z pohledu jejich datace, kvality a věrohodnosti.

Korpus CORDE není lemmatizovaný ani anotovaný, což značně komplikuje studium morfologie a syntaxe. Kromě konkrétních slovních tvarů (a jednoduchých logických kombinací) lze hledat i části slov doplněné hvězdičkou, nicméně v tomto posledním případě, pokud je počet nalezených konkordancí příliš velký, se nezobrazí výsledek. Hledání je možné filtrovat podle zemí, žánrů, chronologie (podle letopočtů), zároveň je možné omezit hledání jen na jednoho konkrétního autora či dílo.

tzat del Català Antic), URL: http://www.cica.cat/, COTAGAL (Corpus de Textos Antiguos de Galicia), který je součástí sítě CHARTA, URL: http://ilg.usc.es/es/proxectos/corpus-de-textos-antiguos-de-galicia-cotagal, UNESP (Córpus Diacrônico do Português), URL: http://www.cdp.ibilce.unesp.br/, Corpus do Português, URL: http://www.corpusdoportugues.org/.

30 URL: http://corpus.rae.es/cordenet.html

31 Na stránkách korpusu (http://corpus.rae.es/ayuda_c.htm) je uveden i údaj 125 milionů slov (tokenů), tento údaj je však nejspíše zastaralý.

32 Není uvedeno, zda se daný poměr vztahuje k počtu tokenů či textů, pravděpodobnější se však jeví první možnost.

Výsledky hledaných výrazů se zobrazují formou konkordancí a jsou také statisticky zpracovány v tabulkách. Korpusové rozhraní je intuitivně snadno pochopitelné a uživatelsky přívětivé.

I přes některé evidentní nedostatky zůstává korpus CORDE zejména pro svou velikost, žánrovou pestrost a relativní geografickou rozmanitost důležitým zdrojem dat v mnoha diachronně zaměřených korpusových studií.

3.1.2 CDE (CORPUS DEL ESPAÑOL)[33]

Dalším hojně využívaným velkým korpusem je diachronní *Corpus del Español* (z roku 2002), který se v mnoha ohledech příliš neliší od CORDE. Obsahuje přes 100 milionů tokenů (přes 20 tisíc textů). Texty pocházejí z období od 13. do 20. století včetně a do korpusu byly přidány z různých zdrojů. Každý text je přiřazen k příslušnému století, hledání nelze filtrovat dle přesných letopočtů. Přesný počet textů i tokenů v každém příslušném století[34] je uveden na webových stránkách. Texty z 20. století se dále dělí podle žánrů. Po stránce filologického zpracování (tj. nejednotné edice a datace textů) je tento korpus ještě problematičtější než CORDE.

Velký podíl na oblíbenosti tohoto korpusu má jistě skutečnost, že na rozdíl od CORDE je *Corpus del Español* částečně lemmatizovaný i anotovaný na morfosyntaktické úrovni[35]. Realizace studií morfosyntaxe je tak nepochybně usnadněna, nutno však podotknout, že zejména v historické části korpusu není lemmatizace ani anotace zdaleka kompletní.

Korpus v současnosti funguje ve dvou uživatelských rozhraních (oproti CORDE poněkud méně přehledných), starším z roku 2008 a novém z roku 2016, kdy byl korpus dále mnohonásobně rozšířen o současné texty pocházející z webových stránek z různých španělsky hovořících zemí.

3.2 PARALELNÍ KORPUSY

Paralelní korpusy (obvykle vícejazyčné) obsahují stejné texty v různých jazykových verzích vedle sebe. Zpravidla jsou zaměřeny na synchronní srovnávání současných jazyků (převážně v jejich standardní formě). Existují však i výjimky. Jako příklad uveďme TRADI IMTti (XX-XXI) (*Traducción de dialectos del inglés moderno temprano de teatro inglés*)[36], projekt bilingvního překladového korpusu s anglickými dialektálními texty z 16.–17. stol. a jejich španělskými překlady z 20.–21. století. Kontrastivních studií zaměřených na historické texty a jejich překlady (ať už současné či historické) je nicméně zatím stále relativně málo a obvykle se omezují na vybrané literární dílo[37].

33 Korpus byl vytvořen Markem Daviesem z Brigham Young University. URL: https://www.corpusdelespanol.org/.

34 Počet tokenů v jednotlivých stoletých časových úsecích je rozložen nerovnoměrně.

35 Cf. Davies (2002, 2010).

36 Cf. Martínez Magaz (2005). Korpus není v současnosti dostupný na webu.

37 Viz např. studie Santiago del Rey Quesada (2016), Eide (2014).

Do budoucna by však dle našeho názoru mohly mít překladové korpusy historických literárních, případně i neliterárních textů v diachronní korpusové lingvistice své pevné místo.

3.2.1 BIBLIA MEDIEVAL[38]

Paralelní korpusy založené na překladech Bible do různých jazyků nejsou v korpusové lingvistice věcí neznámou[39]. Překlady Bible, které jsou vždy připravovány nanejvýš pečlivě, se k jazykovému srovnávání přímo nabízejí, ve svém celku navíc Bible obsahuje i relativně dost žánrově heterogenních textů.

Přesto je španělský projekt *Biblia Medieval* v současné době ojedinělý svého druhu: jednak je zaměřen na středověké biblické překlady, jednak je korpus velmi dobře zpracován po technické i filologické stránce[40]. Jedná se o korpus obsahující kromě hebrejské verze a latinské Vulgaty také všechny dostupné překlady Bible z 13.–15. století do kastilštiny. Celkem korpus obsahuje zhruba 5 milionů slov. Do budoucna se počítá s normalizovanou ortografií, lemmatizací a morfosyntaktickou anotací.

Příklady typických studií, které lze v současnosti na tomto korpusu s úspěchem realizovat řadu, uvádí Enrique-Arias (2012), jenž vyzdvihuje oproti neparalelním („konvenčním") korpusům zejména srovnatelnost biblických překladů a dále možnost otevřené jazykové analýzy od funkce k formě (není nutné znát předem všechny existující formy vyjádření určité funkce).

3.3 KORPUSY SÍTĚ CHARTA[41]

3.3.1 CHARTA (*CORPUS HISPÁNICO Y AMERICANO EN LA RED: TEXTOS ANTIGUOS*)

Korpus CHARTA je pojat jako globální projekt, který si klade za cíl vytvořit rozsáhlou reprezentativní sbírku španělsky psaných dokumentů z různých geografických oblastí z období mezi 7. a 19. stoletím. Důraz se klade zejména na precizní filologické zpracování dosud nevydaných archivních dokumentů podle jednotných kritérií, přesnou dataci a geografickou lokaci dokumentů. K dispozici je trojí zobrazení textů (faksimilní, paleografický přepis a kritický přepis upravený podle jednotných kritérií).

38 Enrique-Arias, Andrés (koord.): *Biblia Medieval*. URL: http://www.bibliamedieval.es

39 Srov. např. Resnik — Olsen — Diab (1999), Christodouloupoulos — Steedman (2015).

40 Podrobný popis korpusu, jeho jednotlivých částí, kritérií pro přepis textů a manuál k užití je k dispozici na stránkách korpusu. Korpus se v mnoha ohledech řídí kritérii korpusu CHARTA.

41 Síť CHARTA (koordinátor Pedro Sánchez-Prieto Borja, Univerzita v Alcalá) je složena z výzkumných skupin působících v centrech a univerzitách v Evropě, Americe i Asii. Obsahuje části celkem deseti diachronních subkorpusů. O některých z nich bude stručně pojednáno v tomto oddílu.

Jedná se o otevřený korpus, který obsahuje v současnosti 2076 dokumentů (přes 1,3 milionů tokenů) vybraných podle parametrů geografických[42], chronologických a typologických z celkem deseti subkorpusů.

Podrobné parametry korpusu včetně možností hledání jsou uvedeny na webových stránkách. Hledat lze celá slova, varianty slov, části slov a kolokace (na úrovni celého korpusu i jednotlivých dokumentů). Korpus není v současnosti lemmatizovaný ani anotovaný na morfologické a syntaktické úrovni (nelze omezit hledání na určitý slovní druh apod.). Korpus nabízí statistické zpracování výsledků hledaných výrazů formou tabulek i grafů z pohledu chronologického i geografického.

Hlavní nevýhodou je prozatím malá velikost korpusu a geografická nevyváženost. Kritéria pro filologické zpracování dat v korpusu CHARTA jsou však vzorem pro mnohé další nově vznikající korpusy, které využívají digitalizace starých dokumentů.

3.3.2 CODEA (*CORPUS DE DOCUMENTOS ESPAÑOLES ANTERIORES A 1800*)[43]

Současná verze korpusu nese označení CODEA^{+} 2015. Obsahuje 2491 dokumentů (přes 1,4 milionů tokenů) z období mezi 12. a 17. stoletím včetně. Dokumenty pocházejí z různých provincií Španělska. Jedná se o archivní dokumenty různého typu (úřední listiny, závěti, soupisy majetku, smlouvy, dopisy, prohlášení atd.).

Korpus je kompletně lemmatizovaný, umožňuje jednoduché i komplexní hledání, které může být filtrováno podle různých parametrů (data, místa atd.). Výsledky se zobrazují jak formou konkordancí, tak i tabulek a grafů zpracovaných podle různých kritérií (datum, místo, žánr).

Korpus se svým zaměřením hodí ke zkoumání jazykové variantnosti evropské španělštiny. I přes relativně malou velikost (ze všech korpusů sítě CHARTA[44] je tento korpus však nejrozsáhlejší) na něm již lze s úspěchem realizovat výzkum, jak ukazují četné recentní studie, z nichž některé jsou uvedeny na stránkách korpusu.

42 V současnosti korpus obsahuje převážně španělsky psané dokumenty z následujících oblastí: Španělsko (téměř 94 % všech dokumentů), Venezuela (cca 2 %), dále Mexiko, Ekvádor, Portugalsko, Kuba, Kolumbie, Dominikánská republika, Panama, Anglie, Salvador. Nejedná se tedy o korpus vyvážený z diatopického hlediska.

43 Korpus CODEA (URL: http://corpuscodea.es/) je vytvářen výzkumnou skupinou GITHE na Univerzitě v Alcalá. Stejná skupina pracuje také na dalších diachronně orientovaných projektech: ALDICAM-CM (*Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid*) je zaměřen na historickou dialektologii, projekt CODOXIX (*Corpus diacrónico del siglo XIX*) zpracovává španělské manuskripty 19. století (Almeida, 2015)

44 Korpus CODEA tvoří v současnosti největší část (cca 40 %, 799 dokumentů) korpusu CHARTA.

3.3.3 CORHEN (*CORPUS HISTÓRICO DEL ESPAÑOL NORTEÑO*)[45]

Korpus CORHEN se soustřeďuje na soukromou dokumentaci klášterních a jiných fondů ze severního Španělska (kolébky vzniku kastilštiny), a to zejména v období mezi 10. a 13. stoletím[46]. V současné době se jedná se o korpus ve výstavbě, který nyní obsahuje 253 dokumentů (zpracovaných dle kritérií CHARTA) z kláštera San Salvador de Oña.

I přes velký potenciál, který korpus nabízí do budoucna, na něm prozatím kvůli omezenému množství textů lze jen stěží provádět výzkum.

3.3.4 CORDEREGRA (*CORPUS DIACRÓNICO DEL ESPAÑOL DEL REINO DE GRANADA*)[47]

Corpus CORDEREGRA obsahuje sbírku nevydaných historických dokumentů (převážně se jedná o svědecké výpovědi a majetkové inventáře) z území bývalého Granadského království (odpovídající rozlohou dnešním provinciím Granada, Málaga, Almería) z období mezi 1492 a 1833. Korpus v současnosti není anotovaný a nabízí pouze možnost hledání celých slov. Kromě úryvků textu s hledaným výrazem je možné si zobrazit paleografický přepis celých dokumentů. Náhled originálu není k dispozici. Na stránkách korpusu je k dispozici ke stažení kompletní antologie textů korpusu CORDEREGRA[48].

45 Korpus CORHEN je vytvářen výzkumnou skupinou GHEN složenou z odborníků několika univerzit. Do korpusu CHARTA bylo zařazeno 168 dokumentů tohoto korpusu. URL: http://corhen.es

46 Skupina GHEN si klade za cíl zrevidovat tradičně přijímanou teorii R. Menéndeze Pidala o vzniku kastilštiny v oblasti staré Kantábrie a severního Burgosu. Tato teorie se zakládá na zkoumání dokumentů z 9.–11. století, které byly srovnávány se stavem španělštiny v 20. století (většina dokumentů této oblasti, zejména z období 12. a 13. století, dosud nebyla prozkoumána).

47 Korpus CORDEREGRA vytváří výzkumná skupina GIHLD složená z odborníků z andaluských univerzit. Korpus CHARTA zahrnuje v současnosti 8 dokumentů (17 304 tokenů) korpusu CORDEREGRA. URL: http://www.corderegra.es

48 Antologie textů je také součástí publikace *El español del reino de Granada en sus documentos (1492–1833). Oralidad y escritura* (Calderón Campos 2015). Tato publikace podrobněji představuje korpus CORDEREGRA a na základě dokumentů v něm obsažených rozebírá fonetické, morfosyntaktické a lexikální aspekty granadské španělštiny. Jak uvádějí autoři projektu na stránkách korpusu, z hlediska výzkumu jazykové historie by mohlo být zajímavé srovnat postupující pronikání kastilského dialektu do tohoto regionu, které se časově shoduje s pronikáním španělštiny do Ameriky. Svědecké výpovědi zachycené v dokumentech jsou zajímavé pro svůj spontánní charakter, obsahují hovorové obraty. Majetkové inventáře nabízejí zajímavá lexikální data z dané doby a lokality.

3.3.5 CODEMA (*CORPUS DIACRÓNICO DE DOCUMENTACIÓN MALAGUEÑA*)[49]

Korpus CODEMA shromažďuje dokumenty různého typu z 16.–19. století, které pocházejí z archivů v Malaze. Dokumenty jsou přepsány v souladu s kritérii CHARTA[50], faksimilní náhledy jsou rovněž k dispozici. Korpus v současnosti není anotovaný a nabízí pouze možnost hledání celých slov. Informace o počtu dokumentů a tokenů není k dispozici.

3.3.6 COREECOM (*CORPUS ELECTRÓNICO DEL ESPAÑOL COLONIAL MEXICANO*)[51]

Korpus obsahuje různé typy archivních dokumentů datované mezi roky 1500 a 1821, které pocházejí z Nového Španělska, Kanárských ostrovů, Antil, Filipín a Pyrenejského poloostrova. Dokumenty jsou přepsány podle kritérií CHARTA. Korpus je vhodný zejména pro výzkum americké španělštiny.

3.4 OSTATNÍ MENŠÍ A SPECIALIZOVANÉ KORPUSY

3.4.1. CORDIAM (*CORPUS DIACRÓNICO Y DIATÓPICO DEL ESPAÑOL DE AMÉRICA*)[52]

Systematické diachronní studium americké španělštiny stálo až do nedávna na okraji zájmu lingvistů. Dosvědčuje to i dosud trvající absence rozsáhlejších publikací zaměřených na jazykový vývoj španělštiny na americkém kontinentu. V posledních letech však zaznamenáváme pozitivní změnu v podobě vzniku četných institucí a projektů zaměřených na shromažďování a zpracovávání historické dokumentace americké španělštiny[53].

Cílem výstavby korpusu CORDIAM bylo zachytit jazykovou historii americké španělštiny, která je ve španělských korpusech i přes nespornou početní převahu mluvčích reprezentována nedostatečně a poměrově podhodnocena. Je složen ze tří subkorpusů — dokumenty, literatura a tisk. Korpus dokumentů je složen výhradně z textů americké provenience datovaných mezi lety 1493–1905. Archivní dokumenty byly vybrány a zpracovány přímo za účelem jejich zařazení do korpusu. Kritéria, jimiž se

49 Korpus zpracovává skupina ARINTA z univerzity v Malaze (URL: http://www.arinta.uma.es). Stejná výzkumná skupina vytvořila zároveň i další diachronně zaměřený korpus DITECA (*Diccionario de Textos Concejiles de Andalucía*), který je zaměřen pouze na lexikum a obsahuje kromě konkordancí (s možností náhledu celých textů) i etymologii slov.

50 Korpus CHARTA obsahuje 140 dokumentů (100 206 tokenů) korpusu CODEMA.

51 Korpus je výsledkem projektu skupiny GEECOM realizovaném na mexické univerzitě UNAM. URL: http://www.iifilologicas.unam.mx/coreecom/presentacion.html

52 Korpus je výsledkem projektu Mexické jazykové akademie, spolupracuje na něm však i řada jiných pracovišť. URL: http://www.cordiam.org/. Podrobné představení projektu a korpusu cf. Bertolotti — Company Company (2014).

53 Cf. Bertolotti — Company Company (2014, s. 132).

řídil přepis manuskriptů, jsou uvedena na webových stránkách. Korpus literatury obsahuje vydaná i nevydaná díla napsaná na americkém kontinentu mezi 16.–19. stoletím. Korpus tisku obsahuje žurnalistické texty publikované v Americe v 18. a 19. století. Všechny texty jsou rozděleny podle žánrů. U každého textu je uvedeno datum a místo jeho vzniku, pohlaví a etnikum autora (pokud je známo) a žánrové zařazení.

Korpus nabízí uživatelsky přívětivé rozhraní. Kromě konkordancí je možné si zobrazit i jednotlivé celé dokumenty (manuskripty v transliterované formě). Hledání je možno zúžit podle vybraných kritérií. Lze hledat konkrétní tvary, lemmata, části slov i kolokace. V současnosti je korpus částečně lemmatizován.

Z hlediska počtu tokenů se jedná o menší korpus (cca 4,5 milionů tokenů), který však usiluje o žánrovou i diatopickou vyváženost.

3.4.2 P.S. (*POST SCRIPTUM. ARQUIVO DIGITAL DE ESCRITA QUOTIDIANA EM PORTUGAL E ESPANHA NA ÉPOCA MODERNA*.)[54]

Korpus P.S. (Post Scriptum)[55] je žánrově omezen na soukromou korespondenci z území Španělska a Portugalska z období mezi roky 1500–1833. Soukromá korespondence, psaná ve většině případů jedinci s omezeným vzděláním, je obzvláště cenná v tom smyslu, že může obsahovat hovorové jazykové varianty, které nejsou zachyceny v dokumentech oficiálnějšího charakteru.

Korpus je rozdělen po stránce jazykové na část španělskou a portugalskou, každá obsahuje 3500 dopisů. Dopisy pocházejí z fondů soudních a inkvizičních tribunálů na území Španělska a Portugalska. Výhodou je, že známe nejen datum a místo vzniku dopisu a jeho situační kontext, ale též podrobná biografická data jejich pisatelů (jméno, příjmení, datum a místo narození, příbuzenský vztah, zaměstnání, stav, náboženské vyznání, vzdělání atd.). Tyto údaje jsou obsaženy u každého dopisu. Autoři dopisů jsou muži, ženy i děti z rozličných sociálních vrstev (kněží, řemeslníci, zloději, vojáci…).

Pro přepis textů se uplatnil relativně konzervativní přístup. Upravila se pouze segmentace slov a sjednotil se pravopis *i*/*j* a *u*/*v*. Kromě tohoto semipaleografického přepisu je k dispozici i normalizovaný přepis (s moderním pravopisem a interpunkcí) a dále fotografický náhled celého dokumentu. Kromě konkordancí (úryvků textů obsahujících hledaný tvar) si uživatel může zobrazit i text celého dokumentu. Korpus by měl být v konečné fázi anotovaný na morfologické i syntaktické úrovni, v současnosti je tato anotace hotova z části.

Z výše uvedeného vyplývá, že korpus se z lingvistického pohledu jeví jako obzvláště vhodný pro studium historické sociolingvistiky a dialektologie[56].

54 Projekt P.S. je realizován na Lisabonské univerzitě a klade si za cíl systematický výzkum, editaci a historicko-lingvistické studium soukromé korespondence napsané ve Španělsku a Portugalsku v raně moderní době (1500–1833).

55 CLUL (Ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. URL: http://ps.clul.ul.pt.

56 Čtenář může pro inspiraci nahlédnout do seznamu publikací založených na tomto korpusu, který je uveden přímo na webových stránkách (http://ps.clul.ul.pt/es/index.php?action=papers).

3.4.3 CORLEXIN (*CORPUS LÉXICO DE INVENTARIOS*)[57]

Korpus obsahuje rozličné nevydané notářské dokumenty z archivů španělských a několika latinskoamerických (mexických, kolumbijských, bolivijských a uruguayských), zejména inventáře, ocenění a soupisy majetku, závěti apod., které byly přepsány za účelem jejich shromáždění v korpusu. Způsob přepisu není na stránkách korpusu definován, zdá se však, že texty jsou do určité míry upravené (např. segmentace slov, zkratky).

Celkově se jedná o menší korpus (cca 1 milion tokenů), který je primárně zaměřen na lexikum běžného života období Zlatého věku (17. století). Lexikum tohoto typu často není dokumentováno ve velkých korpusech typu CORDE[58], někdy dokonce ani ve slovnících. V korpusu lze hledat celá slova, víceslovná spojení i segmenty písmen (např. slabiky či kořeny doplněné hvězdičkou), k zobrazení jsou však k dispozici pouze celé texty (přičemž hledaný segment není zvýrazněn). Každý text obsahuje informaci o názvu dokumentu, dataci a geografické lokaci.

Celkově lze konstatovat, že korpus CorLexIn představuje potenciál zejména pro výzkum lexika a uplatní se zejména v historické lexikografii. Cenné poznatky může poskytnout i z pohledu historické dialektologie, případně též jazykového kontaktu na úrovni lexikálních výpůjček.[59]

3.5 DIACHRONNÍ KORPUSY GOOGLE BOOKS

V posledních letech vznikají i korpusy využívající jako zdroj dat web. Tyto korpusy bývají velmi rozsáhlé (čítají několik miliard až desítek miliard tokenů) a jsou z pochopitelných důvodů zaměřeny na současný jazyk. Takto rozsáhlé diachronní korpusy až na výjimky v současnosti neexistují.

Speciální nástroj vyhledávače Google Books[60] fungující na bázi n-gramů však nabízí možnost hledání jednotlivých slov či kolokací z diachronní perspektivy[61]. Každé hledané slovo či kolokace se však musí v knihách Google Books vyskytovat alespoň 40×, v opačném případě je pro vyhledávač „neviditelné“[62]. Nástroj se jeví nejvhodnější

57 Korpus začal vznikat v r. 2006. Participují na něm tři španělské univerzity (León, Burgos, Oviedo) ve spolupráci s Institutem Rafaela Lapesy a RAE. Projekt řídí José R. Morala Rodríguez z Leónské univerzity. URL: http://web.frl.es/CORLEXIN.html

58 Cf. Morala Rodríguez (2014).

59 Morala Rodríguez (2014, s. 23–25) uvádí pro ilustraci několik vzorových studií, pro které lze s úspěchem využít tento korpus. Studie mají různé zaměření (slova dosud nezdokumentovaná, formální nebo diatopická variantnost, lexikální inovace, lexikální morfologie).

60 URL: https://books.google.com/ngrams/. Vyhledávač n-gramů Google Books funguje v současnosti v 8 jazycích, kromě angličtiny a španělštiny též v čínštině, francouzštině, němčině, hebrejštině, italštině a ruštině. Španělská verze je z roku 2009.

61 Kromě celých slov lze již v současnosti použít i pokročilejší hledání. Více informací zde: https://books.google.com/ngrams/info.

62 Výsledky nástroje Google Books N-gram viewer tak nejsou totožné s prostým prohledáváním v Google Books.

pro porovnávání frekvence výskytu dvou či více výrazů v určitém časovém období. Výsledek je zobrazen formou grafu.

Propracovanější verzi korpusu využívající části dat z korpusu Google Books a fungující na stejném principu n-gramů vytvořil Mark Davies pro britskou a americkou angličtinu a pro španělštinu. Španělská část jeho *Google Books Corpora*[63] obsahuje 45 miliard slov. Jedná se tak o bezkonkurenčně nejrozsáhlejší diachronní korpus v pravém slova smyslu.

Korpus nabízí více možností hledání oproti vyhledávači Google Books, chronologické zobrazení výsledků je možné po jednotlivých letech či dekádách (v období od r. 1500 do prvního desetiletí 21. století včetně). Konkrétní konkordance sice nelze zobrazit, lze však snadno kliknutím přejít na konkrétní knihu v Google Books. I pro tento korpus platí omezení, že hledaný výraz se musí v korpusu vyskytovat minimálně 40x. Korpus tedy na jedné straně nabízí ke konzultaci obrovské množství dat[64], na straně druhé toto omezení může mít za následek, že některé okrajové výrazy, které se nevyskytují v menších korpusech, zůstanou i tak skryty[65].

Toto však bohužel není zdaleka jediná nevýhoda tohoto korpusu. Pomineme-li skutečnost, že knihy v tomto korpusu nejsou dále nijak tříděny podle žánru, tématu či místa vydání (které samozřejmě není často totožné s původem autora), a dále též nevyváženost počtu tokenů z hlediska časového, největší problém z našeho pohledu představuje datace jednotlivých knih. Týká se to nejen knih, které vyšly ve 20. či 21. století opakovaně (v nezměněné formě) v různých vydáních s víceletým odstupem (v Google Books se často neobjevuje jejich původní, první vydání). Z pohledu diachronního jazykového výzkumu závažnější (a bohužel poměrně četné) jsou však případy, kdy kniha stará evidentně několik set let je automaticky časově zařazena do 21. století, neboť se jedná např. o recentní faksimilní reedici.

Diachronní bádání založené na tomto korpusu tedy nejen nemůže dle našeho názoru odhalit příčinu jazykových změn a variantnosti, ale může být navíc velmi zkreslené i z pohledu čistě chronologického.

63 URL: https://googlebooks.byu.edu/x.asp. Jak je patrné, *Google Books Corpora* Marka Daviese není totožný s vyhledávacím nástrojem Google Books firmy Google, nýbrž se jedná o jeden z korpusů vytvořený na Brigham Young University podobně jako *Corpus del Español*.

64 I přes omezení minimálního počtu výskytů jsou v takto velkém korpusu dohledatelné např. i překlepy u běžnějších výrazů (*amgo* místo *amigo*),

65 Konkrétním příkladem takového okrajového výrazu je dialektální arabismus *alfira* „oleandr", který se ve španělštině běžněji nazývá *adelfa*. Zatímco běžnější název je na Google Books Corpora i v nástroji Google Books N-gram viewer doložen od 18. století (v mnohem menším diachronním korpusu CORDE je však doložen již od 13. století, přičemž je dokumentován i v 16. a 17. století, tedy v období, které je pokryto v Google Books Corpora), název *alfira* není doložen ani v jednom korpusu Google Books (ani v CORDE). V prostém vyhledávání v Google Books se však výraz *alfira* v tomto významu vyskytuje.

4. ZÁVĚR

Korpusová lingvistika se v uplynulých letech stala nepochybně jednou ze stěžejních lingvistických disciplín. Dá se říci, že v současnosti zažívá opravdový rozkvět a minimálně synchronní jazykové bádání je bez korpusu již často nemyslitelné (byť lingvistika disponuje i jinými výzkumnými metodami). Důvodem velké obliby korpusové metody je jednak snadná dostupnost velkého množství korpusových dat a možnost jejich kvantifikace, jednak relativně vysoká kvalita a reprezentativnost současných korpusů.

Jak je patrné z tohoto článku, velkou odezvu nachází korpusová lingvistika i v diachronním výzkumu španělštiny. Svědčí o tom nejen řada nově vznikajících diachronních korpusů evropské i latinskoamerické španělštiny, ale i stále četnější diachronní studie podložené korpusovými daty.

Diachronní korpusová lingvistika se potýká s několika specifickými problémy (mj. jednotný přepis a datace textů, absence orálních dat, vyváženost korpusu z pohledu chronologického, případně i žánrového, diatopického, diastratického atd.), které je nezbytné následně zohledňovat i při interpretaci korpusových dat v konkrétních studiích. Kromě toho je třeba mít neustále na mysli skutečnost, že ani ten největší a nejreprezentativnější korpus nikdy nemůže obsahovat veškeré jazykové jevy, které se mohou vyskytnout v jednotlivých jazykových plánech. Toto však neznamená, že diachronní lingvistika by se měla kvůli výše uvedeným těžkostem korpusů vzdát, nebo se problémy vůbec nezabývat, nýbrž by je měla pojmout jako výzvy, jimž je třeba v rámci možností co nejlépe čelit.

Diachronně zaměřený lingvista zabývající se španělštinou má v současnosti k dispozici relativně velké množství korpusů různé velikosti a zaměření, s nimiž může v případě potřeby s úspěchem pracovat. Stále však zatím platí, že korpusy vyhovující nejnovějším standardům jsou v současnosti poměrně malé, a naopak největší korpusy mají značné nedostatky po stránce technického či filologického zpracování. Protože mnoho korpusů se však nadále vyvíjí, rozšiřuje a zkvalitňuje, zdá se, že i diachronní badatelé budou moci do budoucna počítat s příslibem nových objevů učiněných na základě korpusového bádání, a to nejen z oblasti lexika, ale i ostatních jazykových plánů.

LITERATURA

Almeida, B. (2015): Un corpus documental del siglo XIX: CODOXIX. *Études Romanes de Brno*, 36, 1, s. 11–20.

Bennett, P. — Durrell, M. — Scheible, S. — Whitt, R. J. (ed.) (2013): *New Methods in Historical Corpora*. Tübingen: Gunter Narr Verlag.

Bertolotti, V. — Company Company, C. (2014): El corpus diacrónico y diatópico del Español de América (CORDIAM). Propuesta de tipología textual. *Cuadernos de la ALFAL*, 6, s. 130–148.

Calderón Campos, M. (2015), *El español del reino de Granada en sus documentos (1492–1833). Oralidad y escritura*. Bern: Peter Lang, s. 139–273.

Christodouloupoulos, Ch. — Steedman, M. (2015): A massively parallel corpus: the

Bible in 100 languages. *Language Resources and Evaluation*, 49, s. 375–395.

Contreras Seitz, M. (2017): Lo que cuentan los documentos: para una historia del español de Chile en el período colonial. *Atenea*, 515, s. 173–188.

Davies, M. (2002): Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del Lenguaje Natural*, 29, s. 21–27.

Davies, M. (2010): Creating useful historical corpora: A comparison of CORDE, the Corpus del Español, and the Corpus do Português. In: A. Enrique Arias, *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*. Frankfurt/Madrid: Vervuert/Iberoamericana, s. 137–166.

Eide, K. G. (2014): Studying word order differences in a historical parallel corpus. An example from Old Spanish and Old Portuguese. In: S. O. Ebeling — A. Grønn — K. R. Hauge — D. Santos (ed.): *Corpus-based Studies in Contrastive Linguistics*. Oslo Studies in Language 6, 1, s. 181–199.

Enrique Arias, A. (2012): Dos problemas en el uso de corpus diacrónicos del español: perspektiva y comparabilidad. *Scriptum Digital*, 1, s. 82–106.

Kabatek, J. (2013): ¿Es posible una lingüística histórica basada en un corpus representativo? *Ibero*, 77, s. 8–28.

Kabatek, J. (ed.) (2016): *Lingüística de corpus y lingüística histórica iberrorománica*. Berlin/Boston: Walter de Gruyter.

Martínez Magaz, J. (2005): Traducción, dialecto y alejamiento cronológico. El corpus TRADI IMTti. In: M. L. Romana García (ed.), *II AIETI. Actas del II Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación*. Madrid: AIETI, s. 602–609.

Morala Rodríguez, J. R. (2014): El CorLexIn, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro. *Scriptum Digital*, 3, s. 5–28.

Parodi, G. (ed.) (2007): *Working with Spanish Corpora*. London/New York: Continuum.

Real Academia Española (2009–2011): *Nueva gramática de la lengua española.*

Resnik, P. — Olsen, M. B. — Diab, M. (1999): The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 33, s. 129–153.

Rodríguez Molina, J. — Octavio de Toledo y Huerta, Á. (2017): La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística. *Scriptum Digital*, 6, s. 5–68.

Rojo, G. (2016): Los corpus textuales del español. In: J. Gutiérrez-Rexach, (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, s. 285–296.

Vaamonde, G. (2014): Limitaciones en el uso de corpus diacrónicos del español. Nuevas aportaciones desde el proyecto de investigación Post Scriptum. In: *XXXII Congreso Internacional de la Asociación Española de Lingüística Aplicada* (AESLA), Universidad Pablo de Olavid, Sevilla.

KORPUSY A KORPUSOVÉ SLOVNÍKY

[ALDICAM-CM] GITHE (koord. P. Sánchez Prieto-Borja): *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid.* <http://textoshispanicos.es/index.php?title=P%C3%A1gina_principal>

[*Biblia Medieval*] Universitat de les Illes Balears (koord. A. Enrique-Arias): *Biblia Medieval*. <http://www.bibliamedieval.es>

[CdE] BYU (koord. M. Davies): *Corpus del Español.* <https://www.corpusdelespanol.org/>

[CdP] BYU (koord. M. Davies): *Corpus do Português*. <http://www.corpusdoportugues.org/>

[CDP] UNESP (koord. S. R. Longhin-Thomazi): *Córpus Diacrônico do Português*. <http://www.cdp.ibilce.unesp.br/>

[CHARTA] (koord. P. Sánchez-Prieto Borja): *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http://www.redcharta.es/>
[CICA] ICREA/UAB, UV-IEC, UA-IEC (koord. Torruella, J., Pérez Saldanya, M., Martines J.): *Corpus Informatitzat del Català Antic*, <http://www.cica.cat/>
[CODEA] GITHE: *Corpus de Documentos Españoles Anteriores a 1800*. <http://corpuscodea.es/>
[CODEMA] ARINTA (koord. Carrasco Cantos, I., Carrasco Cantos P.): *Corpus Diacrónico de Documentación Malagueña*. <http://www.arinta.uma.es/>
[CORDIAM] Academia Mexicana de la Lengua (koord. V. Bertolotti — C. Company Company): *Corpus Diacrónico y Diatópico del Español de América*. <www.cordiam.org>
[CORDE] Real Academia Española: *Corpus diacrónico del español.* <http://www.rae.es>
[CORDEREGRA] GIHLD (koord. M. Calderón Campos, Mª. T. García-Godoy): *Corpus diacrónico del español del reino de Granada. 1492–1833*.URL: <http://www.corderegra.es>
[COREECOM] GEECOM (koord. B. Arias Álvarez): *Corpus Electrónico del Español Colonial Mexicano*. IIFL-UNAM <http://www.iifilologicas.unam.mx/coreecom/presentacion.html>
[CORHEN] GHEN (koord. M. J. Torrens Álvares): *Corpus Histórico del Español Norteño*. <http://corhen.es>
[CorLexIn] Universidad de León (koord. J. R. Morala Rodríguez): *Corpus Léxico de Inventarios*, <http://web.frl.es/CORLEXIN.html>
[CORPES XXI] Real Academia Española: *Corpus del Español del Siglo XXI (CORPES)*. <http://www.rae.es>
[CREA] Real Academia Española: *Corpus de referencia del español actual.* <http://www.rae.es>
[COTAGAL] USC, Instituto da Lingua Galega (koord. R. Pichel Gotérrez): *Corpus de Textos Antiguos de Galicia.* <http://ilg.usc.es/es/proxectos/corpus-de-textos-antiguos-de-galicia-cotagal>
[DITECA] ARINTA (koord. Carrasco Cantos, I., Carrasco Cantos P.): *Diccionario de Textos Concejiles de Andalucía*. <http://www.arinta.uma.es/>
[*Google Books Corpora*] BYU (koord. M. Davies): *Google Books Corpora.* <https://googlebooks.byu.edu/x.asp>
[NDHE] Real Academia Española: *Nuevo diccionario histórico del español.* <http://www.rae.es/recursos/diccionarios/nuevo-diccionario-historico>
[P.S.] CLUL: *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt>

Zuzana Krinková | Ústav románských studií, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0001-5000-9204
zuzana.krinkova@ff.cuni.cz