JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH
FILOZOFICKÁ FAKULTA
ÚSTAV ROMANISTIKY

BAKALÁŘSKÁ PRÁCE

PLURÁL SUBSTANTIV NA –CIA/-GIA V ITALŠTINĚ – KORPUSOVÁ STUDIE

Vedoucí práce: doc. PhDr. Jan Radimský, Ph.D.

Autor práce: Lenka Tomanová

Studijní obor: Italský jazyk

Ročník: 4.

2017

Prohlašuji, že svoji bakalářskou práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice dne 9. května 2017

……

podpis

Ráda bych poděkovala vedoucímu mé bakalářské práce doc. PhDr. Janu Radimskému, Ph.D. za vstřícné, trpělivé vedení a cenné rady a připomínky. Poděkování patří také mojí rodině za všestrannou podporu.

ANOTACE

Plurál substantiv na –cia/-gia v italštině – korpusová studie

Předkládaná bakalářská práce se zabývá pravopisným kolísáním při tvoření množného čísla substantiv končících na –cia/-gia. Hlavním cílem práce je především získat přehled o aktuálních tendencích v této oblasti a o četnosti výskytu jednotlivých zakončení. Úvodní část práce je zaměřena na popis dané problematiky a souhrn několika doporučení uváděných v odborné literatuře především italských jazykovědců. Následující část zahrnuje teorii o korpusové lingvistice a její historii, o korpusech a jejich tvorbě či využití. Na tuto část navazuje samotná korpusová studie, ve které jsou zaznamenány veškeré nalezené výsledky z dostupných online korpusů. Tyto výsledky jsou pak v úplném závěru zhodnoceny a porovnány s jednotlivými pravopisnými pravidly.

Klíčová slova: substantiva na cia/-gia, plurál, korpus, korpusová studie

ANNOTATION

Number inflection of Italian nouns in -cia/-gia - a corpus based study

The presented bachelor thesis occupies with ortographical fluctuation in case of plural formation of Italian nouns in –cia/-gia. The main aim of the thesis is primarily to gain knowledge about current tendencies in this field and about occurrence frequency of each pural ending.The initial part is focused on description of the issue and a sum of several recommendatios that are presented in specialized literature, mainly of italian linguists. The next part includes theory about corpus linguistics, its history, about corpus itself, its creation and using. The last part is applied to corpus study in which all of the found results from corpus available online are recorded. These results are evaluated and compared with orthographical reccomendations in the end of the thesis.

Key words: nouns in –cia/-gia. plural, corpus, corpus based study

OBSAH

ÚVOD

Pro mou bakalářskou práci jsem si vybrala téma, které se zabývá tvořením množného čísla u italských substantiv končících na -cia/-gia, přičemž součástí zadání je i korpusová studie. Hlavním cílem práce je zjistit, jaké jsou současné tendence v pravopisu těchto substantiv, zhodnotit, nakolik se shodují s hlavními pravopisnými doporučeními a na základě tohoto pak popsat aktuální úzus v této problematice.

Práce je zaměřena především na oblast korpusové lingvistiky, a to zejména proto, že považuji tuto oblast jazykovědy za velmi aktuální, přínosnou (nejen svému oboru) a současně v ní vidím i budoucnost celé lingvistiky, neboť informační technologie jsou čím dál tím více na vzestupu.

Co se týče struktury, práce je členěna do třech hlavních kapitol. První kapitola je zaměřena čistě na problematiku plurálu podstatných jmen zakončených na –cia/-gia. Samozřejmostí bylo nejprve popsat obecné zásady tvoření plurálu u běžných podstatných jmen ženského rodu a na základě toho osvětlit princip a důvod, kvůli kterému ke kolísání dochází. V návaznosti na to byl popsán proces palatalizace hlásek „c“ a „g“ jakožto historický důvod vedoucí k nejasnostem v plurálovém zakončení substantiv na –cia/-gia. V závěru této kapitoly bylo uvedeno několik pravopisných doporučení, od nejužívanějšího pravidla předcházející hlásky až po pravidla, která vychází z původu jednotlivých podstatných jmen. Druhá a třetí kapitola je zaměřena čistě na teorii korpusové lingvistiky a korpusu. V úvodu je popsána historie korpusové lingvistiky s podrobnějším zaměřením na kritiku Noama Chomského a vůbec první korpusy. Dále jsou charakterizovány znaky a typy korpusů a jejich využití, základní přístupy v oblasti korpusového výzkumu. Podstatnou částí této kapitoly je část zaměřená na tvorbu korpusů, která zahrnuje charakteristiku korpusových manažerů, popis a praktickou ukázku konkordancí, osvětlení principu značkování či popis v korpusech často zkoumaného jevu – kolokací. V závěrečné kapitole byly v několika tabulkách zaznamenány výsledky vyhledávání vybraného vzorku substantiv ve čtyřech korpusech, které byly zvoleny tak, aby pokryly různorodé oblasti čerpání jazykových dat. Tyto nalezené hodnoty pak byly porovnány s pravopisnými doporučeními z první kapitoly a na základě toho analyzovány a vyhodnoceny v závěr popisující aktuální tendence ve volbě plurálového zakončení substantiv na –cia/-gia.

Dosavadní poznatky v oblasti pravopisného kolísání zahrnují jen několik pravidel, která jsou však pouze doporučující povahy. Navíc se zde začíná projevovat i tendence k eliminaci veškerých „i“ v koncovkách množného čísla, která nejsou nutná k zachování významu slova a tak ve své podstatě „zbytečná“. Naopak v oblasti korpusové lingvistiky došlo zejména v posledních desetiletích k rozsáhlému vývoji a tedy i k nárůstu získaných poznatků, a to zejména díky obrovskému rozvoji v oblasti výpočetní techniky.

Jako metoda zkoumání byla zvolena analýza obsahu literatury a publikací především předních italských lingvistů, které se zabývají principem pravopisného kolísání a pojednávají o pravidlech, podle kterých je možno se řídit. Mimo takto zaměřených publikací byla použita také literatura a zdroje zabývající se korpusovou lingvistikou a korpusy, a online etymologický slovník sloužící ke zjištění původu slov pro možnost porovnání s některými pravopisnými doporučeními, která z původu slov vycházejí. V závěru pak bylo využito samotných online dostupných korpusů k vyhledávání četnosti výskytu obou zakončení množného čísla.

Před téměř dvaceti lety řekl Noam Chomsky v jednom z rozhovorů jako reakci na otázku, jaký je jeho pohled na moderní korpusovou lingvistiku, že neexistuje.[1] Následující roky však dokázaly nejen to, že korpusová lingvistika existuje a je obrovským přínosem pro širokou škálu oborů, ale také to, že je bezpochyby oborem budoucnosti, který si díky stále větším možnostem v oblasti výpočetní techniky a vůbec díky celkovému technologickému rozvoji zajistí uznání ve všech směrech.

[1] MAIR, Christian – HUNDT Marianne. *Corpus Linguistics and Linguistic Theory, Vydání 1999.* Rodopi, 2000, s.5. ISBN 9042014938.

1. Teoretická část

1.1 Princip pravopisného kolísání

1.1.1 Obecná pravidla tvoření plurálu substantiv ženského rodu

Abychom mohli vymezit problematiku pravopisného kolísání v množném čísle podstatných jmen na –cia nebo –gia, je důležité nejprve definovat obecné zásady pravidelného tvoření plurálu substantiv ženského rodu.

Italská substantiva ženského rodu se (až na výjimky) vyznačují buďto koncovou samohláskou -a nebo -e. Vzhledem k tomu, že zásady tvoření množného čísla nejsou nijak komplikované, mohou být přehledně znázorněny v následující tabulce.

singulár: -a, -e	plurál: -e, -i
bors**a**	bors**e**
class**e**	class**i**

Jak je patrné z tabulky, je-li substantivum zakončeno na -a, v množném čísle je nositelem koncovky -e, pokud je substantivum v singuláru zakončeno koncovou samohláskou -e, v množném čísle se mění na samohlásku -i.

1.1.2. Tvoření plurálu substantiv na -cia/-gia

Končí-li však substantivum na –cia nebo –gia, může docházet k pravopisnému kolísání v množném čísle v důsledku připojení plurálové koncovky -e. V takovém případě musíme rozlišit, zda je -i- v koncovce přízvučné či nikoliv.

Je-li koncovkové -i- přízvučné, pravidla pro tvoření plurálu jsou jasně daná a k žádným výkyvům v pravopisu nedochází. Není totiž možné jej eliminovat, neboť i v množném čísle musí stále plnit funkci nositele přízvuku a tím pádem i délky, která musí být zachována. Ve tvaru jednotného čísla se tedy taková koncovka vyslovuje jako [ˋʧia], příp. [ˋʤia], v plurálu pak [ˋʧi:e], příp. [ˋʤi:e].

K výkyvům však dochází v situaci, kdy je -i- v koncovkách –cia, příp. –gia nepřízvučné. Ve tvaru jednotného čísla samohláska -i- mezi souhláskou „c“, resp. „g“ a koncovkovou

samohláskou –a totiž plní funkci grafického symbolu, který označuje palatalizaci předcházející souhlásky, která by se jinak ve spojení s jinými samohláskami (a, o, u), vyslovovala jako velára, tedy [k] či [g]. V tomto případě, tedy ve spojení hlásek v koncovkách –cia/-gia, však podle pravidel výslovnosti vyslovujeme [ʧa], příp. [ʤa]. Pokud tedy z takto zakončených slov vytváříme podle výše uvedeného pravidla množné číslo, vznikají zakončení –cie/-gie, ve kterých však samohláska -i- již z hlediska výslovnosti neplní žádnou funkci. Spojení hlásek c/g + e je totiž podle pravidel výslovnosti samo o sobě vyslovováno jako [ʧe], příp. [ʤe] a samohláska -i- zde tak pro výslovnost není nutná jako v případě tvaru jednotného čísla, kde koncovou samohláskou bylo –a.

Otázka, zda „přebytečné" -i- zachovat, či jestli by mělo odpadávat, tedy vyvstává pouze v případech, kdy je -i- nepřízvučné. Názory na tuto problematiku se mnohdy liší a pravidla v podstatě nejsou pevně stanovena, pokud se tedy nejedná o výrazy, kde by odstranění samohlásky „i" mohlo zaměnit význam slova – v takových případech je nutné -i- zachovat, bez ohledu na jakákoliv doporučení. Příkladem pro tento jev může být substantivum *ferocia* [fe`roʧa]. Vzhledem k tomu, že je zde nositelem přízvuku druhá slabika, jak je patrné z transkripce, nebylo by zde nutné pro vytvoření plurálového tvaru „i" v koncovce zachovávat. Takový plurálový tvar [fe`roʧe] by však mohl být zaměnitelný s adjektivem *feroce* [fe`roʧe], a proto je zde nutné „i" neeliminovat a v koncovce ho ponechat pro zachování a rozlišení významu.

1.2 Historické důvody

Zhruba do poloviny minulého století se pravopis starších textů řídil podle etymologického pravidla. V 50. letech se však postupně rozšířilo a zakotvilo tzv. pravidlo předcházející hlásky.[2] Ve své podstatě bylo toto pravidlo přelomem, neboť nevyžaduje jakoukoliv znalost původu slov a je tak snadno použitelné i pro širokou veřejnost. Volba plurálové koncovky se totiž odvíjí pouze od hlásky, která předchází koncovce –cia nebo –gia. V případě, že koncovce předchází samohláska, plurálový tvar nese zakončení zachovávající „i", pakliže koncovka navazuje na souhlásku, „i" se v koncovce množného čísla eliminuje. Jako příklad lze uvést dvojice substantiv *l'acacia, la ciliegia* a *la provincia, la spiaggia*. První dvojice podstatných slov nese v množném čísle koncovky

[2] http://www.treccani.it/enciclopedia/cia-gia-scia-plurale-dei-nomi-in_%28La_grammatica_italiana%29/

–cie a –gie, neboť koncovky jednotného čísla navazují na samohlásku (sg. *l'acacia* → pl. *le acacie*; sg. *la ciliegia* → pl. *le ciliegie*), zatímco druhá dvojice tvoří množné číslo za pomoci koncovky s eliminovaným „i" (sg. *la provincia* → pl. *le province*; sg. *la spiaggia* → le *spiagge*).

Samotná problematika však vychází již z daleko dřívějšího procesu palatalizace souhlásek *c* a *g*.

1.2.1. Palatalizace *c* a *g*

Podle pravidel výslovnosti klasické latiny se slova jako například *cēra* a *gēlu* vyslovovala se zadopatrovou okluzívou (|c| = neznělou, |g| = znělou) tak, jako ve slově *canis*, tedy [kera] a [ghelu]. Ale palatální samohlásky *e* a *i* vedly k ovlivňování předcházející souhlásky. Tendence vyslovovat veláry *k* a *g* jako palatály před palatálními samohláskami se projevila již ve 3. století.

Původní výslovnost latinských velárních okluzív se zachovala v sardštině v regionu Logudoro. Zde se tedy můžeme setkat s výslovností jako například *kentu* (*cento*), *nuke* (*noce*), *ghirare* (*girare*), *ghelare* (*gelare*).[3]

V severní Itálii je postupu místa artikulace dosaženo až do oblasti alveo-dentální (zubodásňové): [k] a [g] před samohláskami *e* a *i* se transformují (v tomto pořadí) na dásňovou afrikátu neznělou [ts] a dásňovou afrikátu znělou [dz], které většinou následně přecházejí do odpovídajících hlásek sykavých – neznělé [s] a znělé [z]. Tento vývoj |ke| / |ge| → |tse| / |dze| → |se| / |ze|[4] tedy v podstatě znamená, že problematika popisovaná touto prací se netýká oblastí v severní Itálii, neboť pravopisné kolísání onoho plurálového zakončení vychází právě z palatalizace hlásek „c" a „g" před samohláskami „e" a „i".

[3] MARAZZINI, Claudio. *La lingua italiana. Profilo Storico.* Il Mulino, 2002, s. 163. ISBN 978-8815086754.
[4] DARDANO, Maurizio. *Manualetto di linguistica italiana.* Vyd. 2. Bologna: Zanichelli, 1996, s. 116. ISBN 978-8808088567.

Toskánsko	k → [ʧ] g → [ʤ]
severní Itálie	k → [ts] → [s] g → [dz] → [z]
Sardinie	k → [k] g → [g]

1.3 Pravopisná doporučení v jednotlivých pramenech

V souvislosti se vším výše uvedeným je důležité v další části zmínit několik pravidel různých autorů, kterými je možno se řídit při volbě koncovky plurálového tvaru. Žádné z těchto pravidel není závazné a jsou tedy pouze doporučující povahy. Volba žádného z nich není chybná a je zcela libovolná, pouze s výjimkou případů, kde by mohlo dojít k záměně významu, jak bylo zmíněno v části o tvoření plurálu jmen na –cia a –gia.

Sylva Hamplová ve své Mluvnici italštiny v této věci uvádí, že se samohláska „i“ zachovává ve všech případech, kdy je toto „i“ přízvučné, tak jak bylo zmíněno výše. Tedy například ve slovech *la farmacia – pl. le farmacie, la scia – pl. le scie, la bugia – le bugie.*

U slov na –cia/-gia s přízvukem na kmeni se řídí pravidlem předcházející hlásky. Pokud hlásce -c-, resp. -g- předchází samohláska, -i- se zachovává a v plurálu tak vznikají koncovky -cie a –gie, například:

la camicia – pl. le camicie

l'acacia – pl. le acacie

la ciliegia – pl. le ciliegie

la valigia – pl. le valigie

Pokud však souhlásce -c-, resp. -g- předchází souhláska, -i- se eliminuje a v množném čísle se vytvářejí koncovky –ce a –ge, například:

la provincia – pl. le province

la goccia – pl. le gocce

la spiaggia – pl. le spiagge

la scheggia – le schegge

Hamplová se však také zmiňuje o vzrůstající tendenci k eliminaci „nadbytečných“ -i-. I přes uvedené pravidlo se tedy u moderních autorů můžeme setkat s výrazy *valige, ciliege* apod. Zároveň však upozorňuje, že své opodstatnění má -i- tam, kde by mohlo dojít k záměně významu, například v těchto výrazech:

camicia (košile) – *camicie* X *c**a**mice* (bílý plášť)

audacia (odvaha) – *audacie* X *audace* (odvážný)

ferocia (krutost) – *ferocie* X *feroce* (divoký)

reggia (královské sídlo) – *reggie* X *regge* (drží)[5]

Tento princip tvoření plurálu identicky popisují i Patota, a Dardano a Trifone, jen s užitím rozdílných příkladů. Patota uvádí:

la fiducia – le fiducie

la pronuncia – le pronunce

la pancia – le pance

la pioggia le piogge[6]

Dardano a Trifone kromě již uvedených zmiňují také:

la socia – le socie

la boccia – le bocce

la frangia – le frange

Dardano a Trifone se však také ve své publikaci detailněji zabývají oním pravidlem předcházející hlásky. Podle nich se nejedná o pevně stanovené pravidlo, ale spíše o jakousi pomůcku, která zjednodušuje pravidlo etymologické. To se sice opírá o historické základy, ale jeho aplikace a použití je mnohem složitější a náročnější, neboť vyžaduje znalost původu slov. Slova učeného (ne-lidového) původu, tedy především slova z latiny a řečtiny, podle tohoto pravidla -i- v plurálu zachovávají, zatímco slova, která vzešla z jazyka lidového tato -i- eliminují. Obě kritéria, etymologické i ono zjednodušené, však mohou ve skutečnosti bez problémů koexistovat. Případů rozporů a rozdílů mezi těmito pravidly je totiž jen velmi málo, naproti tomu existuje velké množství těch vzájemně shodných. Pokud někdo dobře ovládá latinu, může se podle Dardana a Trifoneho uchýlit k využívání etymologického pravidla, avšak užití toho zjednodušeného rozhodně není chybou. Stejně jako podle Hamplové, také dle jejich názoru je naopak vhodné myslet na nejrůznější kolísání v písemném projevu mnoha současných autorů a na tendenci moderního pravopisu k eliminaci „nadbytečných“ -i- a rovněž upozorňují na případy, kde je nezbytností -i- zachovat, aby se tak zabránilo záměně významu.

[5] HAMPLOVÁ, Sylva. *Mluvnice italštiny: Grammatica italiana*. Vyd. 1. Praha: LEDA, 2004, s. 74. ISBN 80-7335-041-6.

[6] PATOTA, Giuseppe. *Grammatica di riferimento dell'italiano contemporaneo*. Novara: Garzanti Linguistica, 2006, s. 44. ISBN: 9788848040310.

Dodávají však, že dobří spisovatelé mohou v určitých případech tohoto jevu využít, aby dali čtenáři možnost vynaložení vlastního úsilí při snaze o pochopení obsahu a významu kontextu daného textu.[7]

Serianni ve své publikaci Grammatica italiana tvrdí, že -i-, které má v singuláru diakritický význam, je vlastně v množném čísle pouze jakýsi „grafický zbytek“. Je tedy nepodstatné, zda píšeme *ciliegie* či *ciliege*, protože výslovnost je u obou tvarů identická [ʧi'ljɛʤe]. Optimálním řešením by bylo eliminovat veškerá -i- z takovýchto forem plurálů, to by však bylo v rozporu se v současnosti zakotvenými zvyky spisovatelů. Serianni dále pojednává o dvou základních pravopisných kritériích, podle nichž je možné se řídit:

A) První kritérium může být definováno jako historicko-etymologické (obhajované obzvláště A. Camillim v Camilli-Fiorelli 1965: 171-174), které stanovuje zachování -i- ve všech slovech, která byla vytvořena z latinského základu s -ci- nebo -gi-, například:

acacia (← lat. *acăciam)* → *acacie*
audacia (← lat. *audāciam)* → *audacie*
pertinacia (← lat. *pertināciam)* → *pertinacie*

Naproti tomu -i- by mělo být odstraněno v substantivech, ve kterých je palatální hláska [ʧ] a [ʤ] historickým výsledkem jiného základu, například:

camicia (← lat. *camīsiam)* → *camice*
bragia (← germ. **brasiam)* → *brage*
ciliegia (← **cereseam)* → *ciliege*
cupudigia (← **cupidĭtiam)* → *cupidige*

Camilliho vedly k přesvědčení o zachování -i- v slovech z přímého latinského základu především dva důvody: 1) možnost obnovení výslovnosti etymologického -i- jakožto poetické přehlásky nebo pro latinizovanou výslovnost; 2) přisouzení určitého stupně fonetické důležitosti -i- na základě jeho zachování v určitých oblastech jižní Itálie.

B) Druhé kritérium definuje Serianni jako empirické a zároveň je touto publikací upřednostňováno a doporučováno. V podstatě se jedná opět o již zmiňovaný princip předcházející hlásky, Serianni však obě skupiny (skupinu substantiv

[7] DARDANO, Maurizio – TRIFONE, Pietro. *La nuova grammatica della lingua italiana*. 1. vyd. Bologna: Zanichelli, 1997, s. 177-178. ISBN 8808104265.

zakončených v plurálu na -cie/-gie + skupinu zakončenou -ce/-ge) svým způsobem dále rozděluje do dvou sérií:

I. *l'acacia → le acacie, la camicia → le camicie*
la bambagia → le bambagie, la ciliegia → le ciliegie

II. *la provincia → le province, la goccia → le gocce*
la spiaggia → le spiagge, la frangia → le frange

První dvojice má vždy oporu v etymologickém základu s -ci- a -gi- (*acacia, bambogia, provincia, spiaggia)*, druhá nikoliv (*camicia, ciliegia, goccia, frangia*): jak je patrné v porovnání jednotlivých dvojic substantiv, neexistuje zde rozdíl v chování mezi sérií etymologickou a neetymologickou (tedy např. mezi *acacie* (etym. základ) a *camicie* (neetym. základ) – stejné zakončení -cie).[8]

Jistě stojí za povšimnutí také pravopisný rozdíl mezi kritérii u slov *camicia* a *ciliegia*. V případě kritéria podle Camilliho totiž dochází k eliminaci -i- a výsledný plurálový tvar je tedy *camice* a *ciliege*, naopak u Serianniho kritéria je -i- zachováno, tvary množného čísla jsou tedy *camicie* a *ciliegie*.

[8] SERIANNI, Luca – CASTELVECCHI, Alberto. *Grammatica italiana: italiano comune e lingua letteraria*. 1. vyd. Torino: UTET, 1989, s. 135-136. ISBN 8802041547.

2. Korpusová studie – korpusová lingvistika

Pro správné pochopení a zpracování korpusové studie je nezbytné si vymezit související pojmy „korpus“ a „korpusová lingvistika“.

2.1 Úvod do korpusové lingvistiky

„Korpusová lingvistika je ta část a podoba lingvistiky, která studuje jazyk prostřednictvím korpusů a od nich odvozené metodologie.“[9]

Zkoumá jak teorii a principy využívání korpusů, tak jejich praktické použití. V souvislosti s rozvojem umělé inteligence se práce s korpusy často využívá v nově vzniklých oblastech studia jazyka, avšak své uplatnění nacházejí také v oblastech tradičních.[10] Od jiných jazykovědných směrů se ovšem liší jen důsledným využíváním jazykových dat, velkým rozsahem těchto dat a novou metodologií.[11]

Nejedná se přímo o studium určitého konkrétního aspektu jazyka, zabývá se spíše postupy a metodami pro studium jazyka, a proto může být korpusová lingvistika charakterizována spíše jako metodologie.[12]

V rámci tohoto nového korpusově-lingvistického přístupu se texty chápou jako rozsáhlé produkty jazykového systému, díky kterým dochází k poznání obecnějších jazykových pravidel.[13]

Mimo to, že se korpusová lingvistika zabývá shromažďováním, správou, archivací či klasifikací dat, jejím úkolem je také systematické studium dat, nikoli jen výběrový či náhodný a spekulativní zájem o jednotlivost bez ohledu na množství dat a celek.[14]

---

[9] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 9-10. ISBN 80-7106-861-6.
[10] ČERMÁK, František. *Jazyk a jazykověda: přehled a slovníky*. Vyd. 4., V Karolinu 2., dopl. Praha: Karolinum, 2011, s. 108. ISBN 978-80-246-1946-0.
[11] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[12] MCENERY, Tony a Andrew HARDIE. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012, s.1. Cambridge textbooks in linguistics. ISBN 978-0-521-54736-9.
[13] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[14] ČERMÁK, František. *Jazyk a jazykověda: přehled a slovníky*. Vyd. 4., V Karolinu 2., dopl. Praha: Karolinum, 2011, s. 110. ISBN 978-80-246-1946-0.

2.2 Historie korpusové lingvistiky

Obor korpusové lingvistiky je vcelku novým odvětvím jazykovědy, který vznikl teprve v nedávné době, a to především díky rozvoji výpočetní techniky, informačních technologií a počítačů, bez kterých by poznání jazyka v takovémto rozsahu nebylo nikdy možné. Právě softwarové nástroje poskytují možnost třídit, klasifikovat a vyhodnocovat jazyková data v tak obrovském množství.[15] Kořeny korpusové lingvistiky však sahají mnohem dále než do 2. pol. 20. století.

2.2.1 Období tzv. „rané korpusové lingvistiky"

„Raná korpusová lingvistika" je termín užívaný pro veškerou korpusově založenou práci, která byla vytvořena před nástupem Chomského, což zahrnuje období do konce 50. let 20. století. Již v té době využívali lingvisté – představitelé strukturalismu pro svůj výzkum metody, které by se dnes daly označit jako „corpus-based", čili korpusově-založené, i když termín „korpusová lingvistika" v té době ještě nebyl přímo užíván.[16]

Ačkoliv je korpusová lingvistika záležitostí zejména posledních desetiletí, nápad shromažďování textů pro účely zkoumání jazyka sahá až do středověku. V té době lidé vytvářeli seznamy slov spolu s kontextem, ve kterém se dané slovo ve vybraném textu vyskytovalo, nebo vytvářeli seznamy nejčastěji se vyskytujících slov tím, že počítali výskyt slov v jednom textu, případně v souboru více textů.

Metody rané korpusové lingvistiky jsou využívány v množství lingvistických odvětví. Badatel sbírá a analyzuje přirozeně se vyskytující data za účelem popsat a zdokumentovat např. jazykové změny nebo dokázat lingvistické hypotézy. Hlavní myšlenkou je, že popis jazyka je záležitostí objektivního faktu a nikoliv subjektivní spekulace. Objektivní materiály sesbírané z reálného života mohou poskytnout odpovědi na všechny lingvistické otázky.[17]

[15] https://nlp.fi.muni.cz/cs/UvodDoKorpusoveLingvistiky
[16] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus1/1FRA1.HTM
[17] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 5. ISBN 9783638373777.

2.2.2 Korpusově založené práce do konce 50. let 20. století

Studie, které vznikly v raném období počátku korpusové lingvistiky, byly pro tento obor velkým přínosem. Hrály totiž velkou roli v jeho vývoji, neboť vytvořily základy pro myšlenku, která pak byla v následujících desetiletích přezkoumána a vylepšena.[18]

Zhruba mezi léty 1876-1926 probíhal výzkum založený na zkoumání dětského jazyka za pomoci deníků vytvořených rodiči. Tyto primitivní korpusy jsou dodnes využívány jako zdroj normativních dat v jazykovědném výzkumu.

Již na konci 19. století v roce 1897 využil Kading korpus o obsahu 11 milionů německých slov k porovnání frekvenční distribuce písmen a pořadí písmen v němčině. Svým rozsahem se korpus vyrovnal korpusům dnešní moderní doby.

Korpus byl využit mimo jiné také ve výzkumu založeném na pedagogice cizího jazyka, těmito lingvisty byli Fries and Traver (1940), a Bongers (1947). Kenndey (1992) poznamenal, že korpus a pedagogika cizího jazyka měly v první polovině 20. století silnou spojitost, neboť slovní zásoba využívaná pro studenty pocházela často právě z korpusů.[19]

V 30. letech 20. století prováděla Pražská lingvistická škola kvantitativní studie (především češtiny, angličtiny a ruštiny) frekvence výskytu určitých gramatických procesů nebo různých slovních druhů.[20]

2.2.3 Kritika korpusové lingvistiky. 50. léta 20. století – Noam Chomsky a další.

Ke konci 50. let 20. století někteří, zejména Noam Chomsky, začali pochybovat, zda jsou nálezy a zjištění o podstatě jazyka a práce na korpusech opravdu užitečné. Jeho vliv způsobil změnu v lingvistickém paradigmatu tehdejší doby, neboť v následujících dvou desetiletích se korpusová lingvistika stala velmi nepopulární.

Noam Chomsky zapříčinil přechod od empirismu k racionalismu. Empiristický přístup je založen na analýze externích dat, jako jsou sebrané texty a korpusy. Racionalistická teorie

[18] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 6-7. ISBN 9783638373777.
[19] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus1/1FRA1.HTM
[20] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 7. ISBN 9783638373777.

je založena na uměle vytvořených behaviorálních datech a vědomých introspektivních úsudcích.

Chomsky je zastáncem názoru, že není potřeba jakýchkoliv externích dat od doby, co lingvisté dokáží pracovat také se znalostmi o jazyce uvnitř své mysli. Chomského teorie generativní gramatiky jsou založeny na přesvědčení, že mluvčí jazyka dokáže vytvořit nekonečné množství slov a vět, i když disponuje pouze omezeným souborem pravidel. To je důvod, proč by lingvistika měla popisovat spíše schopnost jazyka než jeho výkon.[21]

Schopnost (competence) popisuje Chomsky jako tichou a internalizovanou znalost jazyka. Naproti tomu **výkon (performance)** je vnější známka jazykové schopnosti, která může být ovlivněna různými faktory.[22]

Chomsky popisuje korpus jako soubor projevených výroků. Korpusy jsou výkonovými daty, a proto také pouhým návodem k vytvoření lingvistické schopnosti. Proto Chomsky tvrdí, že lingvista se musí spoléhat spíše na introspekci než pozorování jazyka.[23] Kritikové tohoto názoru však argumentovali tím, že přirozeně se vyskytující data jsou pozorovatelná a prokazatelná, zatímco introspektivní úsudky nikoliv. Sampson (1992) tvrdil, že introspektivní data jsou umělá a tím pádem i vzdálená prokazatelným datům, která můžeme nalézt v korpusu.[24]

Jak již bylo zmíněno, lingvisté se domnívali, že jazyk je konečný, Chomsky naproti tomu tvrdil, že jazyk je nevyčíslitelný, tedy nekonečný. Konečný korpus nemůže být podle Chomského reprezentativní pro nekonečný jazyk a musí tak být zkreslený a neúplný. Tvrdí, že jakýkoli korpus bude zkreslený, a to z toho důvodu, že se v něm např. neobjeví věty, které jsou nepravdivé nebo nezdvořilé.[25] Podstatným argumentem pro teorii nekonečného jazyka je také velké množství možností pro tvorbu věty, ať už lexikálních či syntaktických kombinací.[26]

[21] WONNER, Bernadette. The development of corpus linguistics to its present-day concept. GRIN Verlag, 2005, s.8-9. ISBN 9783638373777.
[22] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus1/1FRA1.HTM
[23] WONNER, Bernadette. The development of corpus linguistics to its present-day concept. GRIN Verlag, 2005, s. 9. ISBN 9783638373777.
[24] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus1/1FRA1.HTM
[25] WONNER, Bernadette. The development of corpus linguistics to its present-day concept. GRIN Verlag, 2005, s. 9. ISBN 9783638373777.
[26] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus1/1FRA1.HTM

Pozdější zesílená kritika si kladla otázku, zda by korpusová metodologie byla nejlepší i za předpokladu, že by byl jazyk konečným nástrojem. Fillmore (1992), který pokládal otázku, proč vyhledávat v korpusu o obsahu milionů slov pro skutečnosti, které nám mohou být dostupné díky introspekci. Korpusová lingvistika byla považována za nekreativní a pasivní. Naproti tomu introspekce dokáže podat informace rychleji a přesněji.

Vedle Chomského byl kritikem také Abercrombie (1965), který poukazoval na pragmatický problém ve zpracovávání dat. V 50. letech nebyla práce s velkým korpusem snadná, protože schopnost počítačových technologií ještě nebyla dostupná, a proto Abercrombie popsal korpusově založený přístup jako přístup složený z pseudo-technik. Podle něj je třeba mnoho analytiků, kteří budou vyhledávat v korpusu pouze svýma očima. Proto je tato metoda zpracování dat přirozeně velmi náchylná k chybám a zároveň časově náročná a drahá.

Kritika Chomského a Abercrombieho byly koncem 50. let velmi důležitými výroky, i přesto, že byly velmi přehnané a nejsou již aktuální. Abercrombieho pohled na „pseudo-techniky" v dnešní době také rozhodně neplatí, neboť počítače zcela nahradily člověka ve zpracování dat. Výzkum korpusů je v dnešní době daleko rychlejší přesnější a zároveň levnější. [27]

2.2.4 Vznik korpusové lingvistiky a první korpusy

Teoretický vznik korpusové lingvistiky můžeme datovat do 50. let 20. století. V té době američtí lingvisté (Harris, Hill) došli k názoru, že pro dostatečný popis určitého jazyka a jeho gramatických jevů, je nutný právě korpus, který byl chápán a charakterizován jako:[28] *dostatečně velký soubor přirozeně se vyskytujících jazykových dat.*[29]

První korpusy a s tím i korpusová lingvistika jakožto samostatná uznávaná věda vznikaly v anglosaském prostředí. Právě Velká Británie byla zemí, kde vznikla jedna z nejdůležitějších a nejvýznamnějších korpusových disciplín, a to korpusová

[27] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 10. ISBN 9783638373777.
[28] https://nlp.fi.muni.cz/cs/UvodDoKorpusoveLingvistiky
[29] Tamtéž.

lexikografie. Téměř veškeré nové slovníky, které v současnosti v Británii vycházejí, jsou totiž vytvořeny na základě korpusových dat.[30]

Pojetí nově vzniklého oboru korpusové lingvistiky navazovalo ve svých názorech na jazyk na strukturalismus, a proto také byla vnímána jako jeho pokračování. V dnešní době se však rýsuje spíše jako metodologie získávání informací z jazyka, která nahrazuje staré způsoby, které začaly být nedostačující. Počátky jazykového korpusu vycházely z touhy po lepším poznání jazyka a snaze porozumět tomu, co jazyk sděluje a jakým způsobem.[31]

V novém pojetí však korpusová lingvistika začala vznikat až počátkem 60. let minulého století, o což se zasadili američtí lingvisté Quirk a Francis, ale také český lingvista H. Kučera. V roce 1960 začal Quirk pracovat na tzv. **Survey of English Usage** (SEU), v rámci kterého se počítalo i s analýzou a zpracováním mluvené podoby jazyka. [32]

SEU ovšem nebyl počítačově orientován, a tak lze za jeden prvních korpusů v dnešním slova smyslu považovat spíše **Computation Analisis of Present-Day American English**, počítačový korpus současné americké angličtiny, známý také jako **Brown Corpus**, který však zpracovával pouze psané texty. O jeho vznik se zasloužil právě Čech H. Kučera a Američan N. Francis.

Tím dalším korpusem, který stál na úplném počátku, byl soubor nahrávek a přepisů o obsahu cca 1 milion slov ve 200 vzorcích, který byl později digitalizován a pak znám jako **LOB Corpus**[33], neboli **Lancaster-Oslo/Bergen Corpus**.

V roce 1975 Jan Svartvik spolu se svým týmem vytvořil **London-Lund Corpus**, o obsahu cca 500 000 slov, ve kterém spojili sílu Brown Corpusu a SEU, což v praxi znamenalo, že mluvená část korpusu SEU byla převedena do elektronické podoby.

[30] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[31] ČERMÁK, František, ed. *Korpusová lingvistika Praha 2011*. Praha: NLN, Nakladatelství Lidové noviny, 2011. Studie z korpusové lingvistiky, sv. 15, s. 11-12. ISBN 978-80-7422-115-6.
[32] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 13-14. ISBN 9783638373777.
[33] ČERMÁK, František, ed. *Korpusová lingvistika Praha 2011*. Praha: NLN, Nakladatelství Lidové noviny, 2011. Studie z korpusové lingvistiky, sv. 15, s. 12. ISBN 978-80-7422-115-6.

V 80. letech druhá generace korpusů masivně překonala rozsah generace první. Je představována korpusem **Birmingham Language Corpus** s asi 30 miliony slovy, jehož autorem je John Sinclaire.

Jako třetí generace jsou pak nazývány korpusy, jejichž rozsah dosahuje stovek milionů slov. K takovým patří například **Bank of English**, nazývaný také jako „monitorovací korpus“. Od jeho spuštění v roce 1991 jsou neustále přidávána nová data. V roce 1996, tedy po pěti letech od spuštění, již obsahoval 320 milionů slov[34], v dnešní době je již za hranicí 500 milionů slov. Díky tomuto korpusu také vzniklo v Birminghamu významné slovníkové nakladatelství *Cobuild*.[35] Dalším takovým korpusem je **British National Corpus**, tedy Britský národní korpus. Jedná se však o konečný korpus o obsahu 100 milionů slov psaného i mluveného jazyka. Za zmínku jistě stojí také **International Corpus of English**, který má dvacet center po celém světě, přičemž každý z nich se zabývá jednou národní variantou angličtiny.[36]

Nejrozsáhlejší rozvoj korpusové lingvistiky však můžeme zaznamenat teprve v posledních letech, což nepochybně souvisí i s vývojem a zdokonalováním počítačové techniky a informačních technologií, stejně jako tomu bylo na úplném počátku korpusové lingvistiky.[37]

[34] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 14-15. ISBN 9783638373777.
[35] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[36] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 14-15. ISBN 9783638373777.
[37] https://nlp.fi.muni.cz/cs/UvodDoKorpusoveLingvistiky

3. Korpusová studie - korpus

Ve staré latině bylo slovo „corpus“ označením pro „tělo“, a to ať už se jednalo o oblast historickou, archeologickou, medicínskou, vojenskou nebo diplomatickou. Avšak i v dnešní době představuje význam tohoto slova[38] *„větší celek nebo soupis a představu jakéhosi těla či tělesa“.*[39]

Podle Františka Čermáka je korpus ucelený soubor textů, který je sestavený s přihlédnutím ke svému cíli reprezentativním způsobem, je obhospodařovaný počítačově a zpracováván souborem korpusových metod.[40]

Takzvané lingvistické korpusy však můžou být také definovány jako sbírky (obvykle velkých rozsahů) textů mluveného i psaného původu, které vznikly ve skutečných komunikačních situacích (např. záznamy novinových článků) a jsou uchovávány v elektronické podobě.[41]

Vzhledem k velkému rozsahu korpusových dat je lze charakterizovat jako[42]:
„1) typická, nenáhodná a věrná ve vztahu k tomu, jak lidé užívají jazyka; dále jako
2) aktuální, resp. skutečně odrážející svou dobu,
3) neselektivní a objektivní,
4) dostatečná a
5) s pomocí počítače snadno získatelná a rychle přístupná“[43]

3.1 Typické znaky korpusů

(1) **Reprezentativnost** [44] : korpusy bývají v současnosti často tzv. reprezentativní. V takových korpusech jsou texty rovnoměrně zastoupeny tak, aby jejich složení odráželo sledovaný žánr, oblast nebo celé univerzum jazyka. Toto rovnoměrné zastoupení

[38] ČERMÁK, František, ed. *Korpusová lingvistika Praha 2011*. Praha: NLN, Nakladatelství Lidové noviny, 2011. Studie z korpusové lingvistiky, sv. 15, s. 11. ISBN 978-80-7422-115-6.
[39] Tamtéž.
[40] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 10. ISBN 80-7106-861-6.
[41] http://www.treccani.it/enciclopedia/corpora-di-italiano_%28Enciclopedia_dell'Italiano%29/
[42] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[43] Tamtéž.
[44] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 17-18. ISBN 9783638373777.

zpravidla vychází z výzkumu proporcí čtenosti různých typů textů během stejného časového období, popř. z míry vypůjčování knih různých žánrů z knihoven, rozšířenosti a nákladu denního tisku apod.[45]

I přes svůj rozsah a velikost je tedy korpus ve své podstatě vzorkem či výběrem. *„Každá kvantitativně vyjádřitelná část korpusu pochází z určitého konkrétního okruhu zdrojů a naopak každý okruh zdrojů má určité oprávnění být v korpusu zastoupen určitým kvantem textů. Při seriózním přístupu ke stavbě korpusu proto nelze volit ani okruhy zdrojů, ani míru jejich zastoupení pouze intuitivně a tím spíše ne náhodně.“*[46]

(2) **Konečná (omezená) velikost**: většina v současnosti fungujících korpusů má pevný počet slov, která obsahuje, jako například LOB nebo Brown Corpus se zhruba 1 000 000 slov, jak již bylo zmíněno v části věnované historii korpusové lingvistiky. Výjimkou je již také zmíněný BoE jakožto monitorovací korpus, jehož rozsah se neustále mění, resp. narůstá.

(3) **Strojově čitelná forma**: je pravděpodobně nejdůležitějším aspektem v definici moderního korpusu, neboť veškeré výzkumy v korpusové lingvistice mohou být prováděny pouze za předpokladu, že je možno pracovat s korpusem prostřednictvím výpočetní techniky.

(4) **Referenční standard**: i když se nejedná o nezbytnou část definice korpusu, má se za to, že korpus vytváří jakýsi referenční standard pro jazykovou různorodost, kterou představuje. To předpokládá jeho širokou dostupnost ostatním výzkumníkům, která je výhodou v tom smyslu, že poskytuje měřítko, podle kterého mohou být měřeny po sobě jdoucí výzkumné práce.[47]

3.2 Typy korpusů

Jazykové korpusy lze klasifikovat podle nejrůznějších kritérií a hledisek, která odrážejí zamýšlený účel a zaměření korpusu. Jejich rozdělení tedy může vypadat následovně:

1. z hlediska počtu jazyků:
    a) jednojazyčný
    b) vícejazyčný
2. z hlediska tématu:

[45] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 10. ISBN 80-7106-861-6.
[46] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[47] WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005, s. 17-18. ISBN 9783638373777.

a) obecný

b) specializovaný

3. z hlediska modu:

a) psaný

b) mluvený

4. z hlediska časového záběru:

a) synchronní

b) diachronní

5. z hlediska „archeologického“:

a) synchronní

b) archivní

6. z hlediska vymezeného účelu různě[48]

Nepatrně odlišnou typologii můžeme nalézt v publikaci Petra Pořízky Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje:

1. korpusy psaného X mluveného jazyka
2. synchronní X diachronní (stav současného jazyka X vývoj jazyka a jeho proměny)
3. monolingvní X srovnatelné X paralelní (korpusy jednoho jazyka X texty vybrané podle stejných kritérií v různých jazycích či různých variantách jednoho jazyka X projekty zdrojových textů a jejich překladů)
4. obecné X specializované (tvořící reprezentativní vzorek jazyka X se zaměřením, např. na určité téma)
5. neanotované X anotované (text bez jakýchkoliv úprav X k textu dodány interpretační, bibliografické či jiné informace)[49]

3.3 Využití korpusů

Zejména pro svůj rozsah a svou univerzálnost jsou korpusy obrovskými zdroji informací a své uplatnění nacházejí nejen v oblasti lingvistiky. Existuje totiž jen málo oborů, které

[48] https://wiki.korpus.cz/doku.php/pojmy:korpus

[49] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s 10. ISBN 978-80-87895-17-7.

je možno studovat jiným způsobem než prostřednictvím jazyka.[50] A tak jsou korpusová data použitelná pro odborníky v řadě oborů:

- psychology [51]

  V psychologickém výzkumu se korpusy využívají při frekvenční charakteristice stimulů nebo během asociačních testů. Docent Václav Cvrček z Ústavu Českého národního korpusu během svého semináře na téma „Dějiny v zrcadle jazyka. Jazykové korpusy – výzkumný nástroj nejen pro jazykovědce." uvádí jako příklad dotazy z psychologického ústavu na kolokace ke slovu „vražda".[52]

- sociology
- sociolingvisty
- odborníky v oblasti masové komunikace a médií
- lexikografy a lingvisty
- překladatele (strojový překlad)
- tvůrce učebnic a referenčních příruček (gramatiky, slovníky)
- v oblasti umělé inteligence (porozumění v přirozeném jazyce, reprezentace znalostí aj.) [53]

3.3.1 Korpusově-řízený a korpusově-založený přístup

V rámci studia využívání korpusů je možné se často setkat s pojmy „corpus-driven approach" a „corpus-based approach". Jedná se ve své podstatě o výrazy, které vyjadřují míru zapojení korpusů do výzkumu.

Corpus-driven approach je přístup, podle kterého se výsledky výzkumu zakládají pouze na korpusech.[54] V tomto přístupu je povinností lingvisty, aby jednotnost dat jako celku a popisný cíl byly naprosto komplexní s ohledem na korpusové důkazy.[55] Korpus samotný by tedy měl být hlavním zdrojem našich hypotéz o jazyce.[56] Veškeré závěry či výroky

[50] http://full.nkp.cz/nkkr/nkkr0403/0403152.html
[51] https://nlp.fi.muni.cz/cs/UvodDoKorpusoveLingvistiky
[52] CVRČEK, Václav. *Dějiny v zrcadle jazyka. Jazykové korpusy – výzkumný nástroj nejen pro jazykovědce.* https://www.ustrcr.cz/prednasky-ustr/cyklus-verejnych-historickych-seminaru-2015/dejiny-v-zrcadle-jazyka/
[53] https://nlp.fi.muni.cz/cs/UvodDoKorpusoveLingvistiky
[54] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 15. ISBN 80-7106-861-6.
[55] MCENERY, Tony a Andrew HARDIE. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012. Cambridge textbooks in linguistics, s. 150. ISBN 978-0-521-54736-9.
[56] http://corpora.lancs.ac.uk/clmtp/1-cb-cd.php

tedy vycházejí čistě z korpusového výzkumu bez jakýchkoliv předpokladů či očekávání. [57]

Corpus-based approach je přístup, který sice korpusu využívá a v podstatě z něj vychází, ale výzkum zároveň doplňuje o jiné zdroje nebo vlastní reflexe.[58] Termín má odkazovat k metodologii, která využívá korpus především k tomu, aby vysvětlila, otestovala a ilustrovala teorie a popisy, které byly formulovány předtím než byly pro studium jazyka dostupné velké korpusy.[59]

V současnosti se někdy hovoří ještě o dalších dvou přístupech:
Corpus-informed approach – neboli korpusem-poučený přístup, který korpusy využívá jen okrajově, např. k vyhledávání jazykových příkladů.
Corpus-assisted approach – neboli korpusem-podporovaný přístup[60], *„který zkoumá proměny významu určitých pojmů či konceptů na základě teoretických východisek a srovnání klíčových slov různých textů v různých časových úsecích“*.[61]

3.4 Tvorba korpusů a práce s korpusy

Nahlédnutí do teorie týkající se tvorby korpusů je důležité stejně jako schopnost v nich vyhledávat, neboť korpusová lingvistika v současnosti jednou z nejdůležitějších metod lingvistického výzkumu.

Vytěžování jazykových dat pro účely vytvoření korpusu je závislé na třech hlavních faktorech[62]:

> *„1. zda, do jaké míry a jakým způsobem byla provedena anotace původních textů*
> *2. jaký softwarový nástroj byl použit k tvorbě databáze a k vyhledávání dat (tj. jaké nabízí prostředky a možnosti vyhledávání a vyhodnocování dat, příp. ukládání výsledků)*

[57] STORJOHANN, Petra. *Corpus-driven vs. corpus-based approach to the study of relational patterns*. Institut für Deutsche Sprache Mannheim, s. 8.
[58] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 15. ISBN 80-7106-861-6.
[59] MCENERY, Tony a Andrew HARDIE. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012. Cambridge textbooks in linguistics, s. 150. ISBN 978-0-521-54736-9.
[60] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 12. ISBN 978-80-87895-17-7.
[61] Tamtéž, s. 12.
[62] Tamtéž, s. 18

3. na tom, jaká je obeznámenost konkrétního uživatele se softwarovým nástrojem a především s možnosti (syntaxí) vyhledávacího jazyka"[63]

Pro to, aby bylo možné získat z korpusů maximální množství informací a poznatků, je třeba využívat jejich vnitřní strukturu a možnosti pro samotné vyhledávání v korpusu, tedy prostředky dotazovacího jazyka. Tím nejčastěji používaným v dnešním době je dotazovací jazyk CQL, neboli Corpus Query Language, a tzv. regulární výrazy, které jsou jeho důležitou součástí. Slouží k definici určitých vzorů, přezdívaných jako vyhledávací masky, které jsou rozpoznány programem korpusu a vyhledávány v textu.[64] Pro regulární výrazy je užíváno různorodých znaků (např. tečka, otazník, kulatá závorka apod.), přičemž každý z nich je nositelem speciálního významu.[65]

3.4.1 Tokenizace a značkování

Tokenizace představuje v tvorbě korpusů prvotní fázi v automatickém zpracování. Jedná se o rozčlenění zpracovávaného textu a definici na samostatné jednotky, které jsou nazývány jako „token".[66] „Token" tak představuje nejmenší jednotku korpusu. Obvykle každá slovní forma a jakákoliv forma interpunkce vytváří samostatnou jednotku „token". Z tohoto důvodu obsahuje korpus více jednotek „token" než slov.[67]

Během zpracovávání jazykových dat je každému výrazu přiřazena určitá zkratka/značka **pos** (Part of Speech) či **tag** (značka).[68] POS tagging je proces připojování poznámek ke každé jednotce „token". Každá taková poznámka nese informaci o slovním druhu a často také morfologické a gramatické informace jako například číslo, rod, pád apod. Nástroj, který automaticky přiřazuje takovéto informace k jednotce „token" je označován jako „tagger" [69], případně automatický morfologický analyzátor.[70]

---

[63] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 18. ISBN 978-80-87895-17-7.
[64] Tamtéž, s. 18.
[65] https://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy
[66] OSOLSOBĚ, Klára. Česká morfologie a korpusy. Karolinum Press, 2014, s. 12-13. ISBN 9788024625621.
[67] https://www.sketchengine.co.uk/my_keywords/token/
[68] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 26. ISBN 978-80-87895-17-7.
[69] https://www.sketchengine.co.uk/itwac-corpus/
[70] OSOLSOBĚ, Klára. Česká morfologie a korpusy. Karolinum Press, 2014, s. 13. ISBN 9788024625621.

Tato interpretace je druhým krokem ve zpracování jazykových dat, avšak její neúplnost v některých případech oproti přirozenému jazyku dále vede k nutnosti zpětné verifikace těchto případů.[71]

3.4.2 Korpusové manažery

K práci s korpusy je využíváno různých aplikací. Nejčastěji se jedná o tzv. korpusové manažery, jinak nazývané také jako konkordanční nástroje. Korpusové manažery jsou určeny k práci s databází a ke zpracování a vyhledávání v korpusových datech. Tyto nástroje totiž provádějí operace nezbytné pro korpusový výzkum, tedy vyhledávání slova nebo sousloví v daných kontextech, jejich frekvence výskytu, zdrojový text nebo např. vyhledávání kolokací či koligací. Nejvyužívanějšími korpusovými manažery vůbec je The Sketch Engine a jeho novější varianta NoSketch Engine jakožto open source. [72]

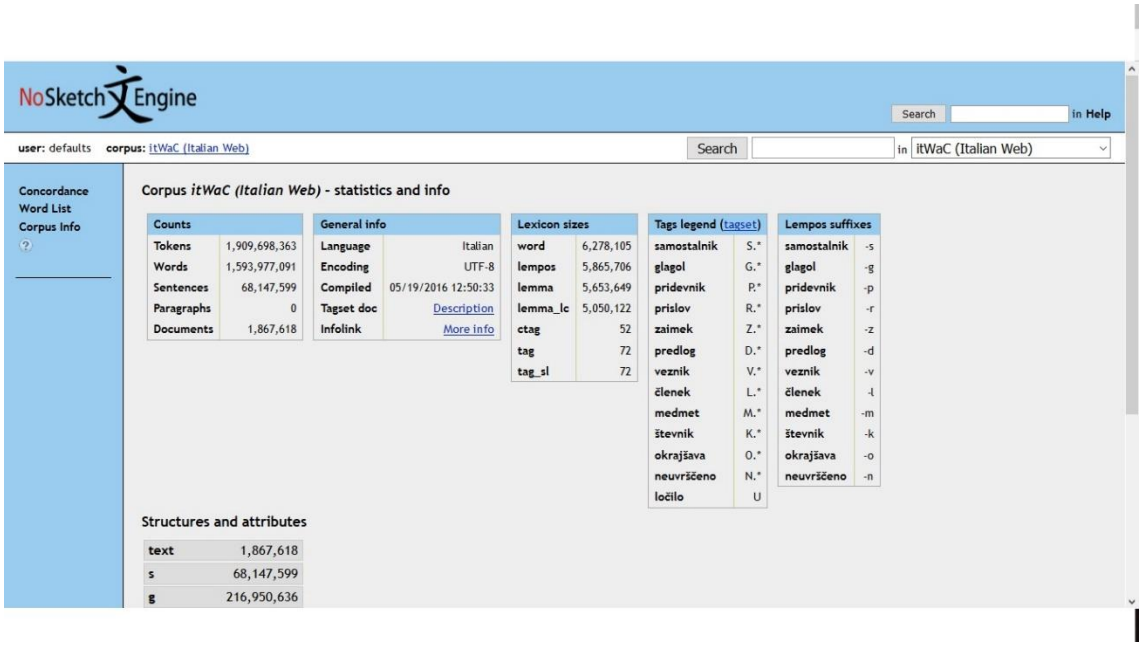
3.4.3 Konkordance

To, co je výstupem korpusového manažeru, je nazýváno jako tzv. konkordance. Jedná se o soupis veškerých výskytů daného vyhledávaného výrazu, v některých případech i spolu s jeho zdrojem či lemmatem. Každý jednotlivý nalezený výraz je zobrazen v řádku, jehož uspořádání je zpravidla tzv. typu KWIC, čili Key Word in Context, ve kterém se dané vyhledávané slovo vyskytuje uprostřed tohoto řádku.[73] Jak již vyplývá ze zkratky KWIC, vyhledávaný výraz je uváděn v rámci svého levého i pravého kontextu. Mimo to bývá zpravidla barevně rozlišen od zbytku textu a zároveň zarovnán pod sebe. Prostřednictvím korpusových manažerů je možné vyhledávat nejen na základě klíčového slova (tedy KWIC) či slovních spojení, ale i podle různých lingvistických kategorií.[74]

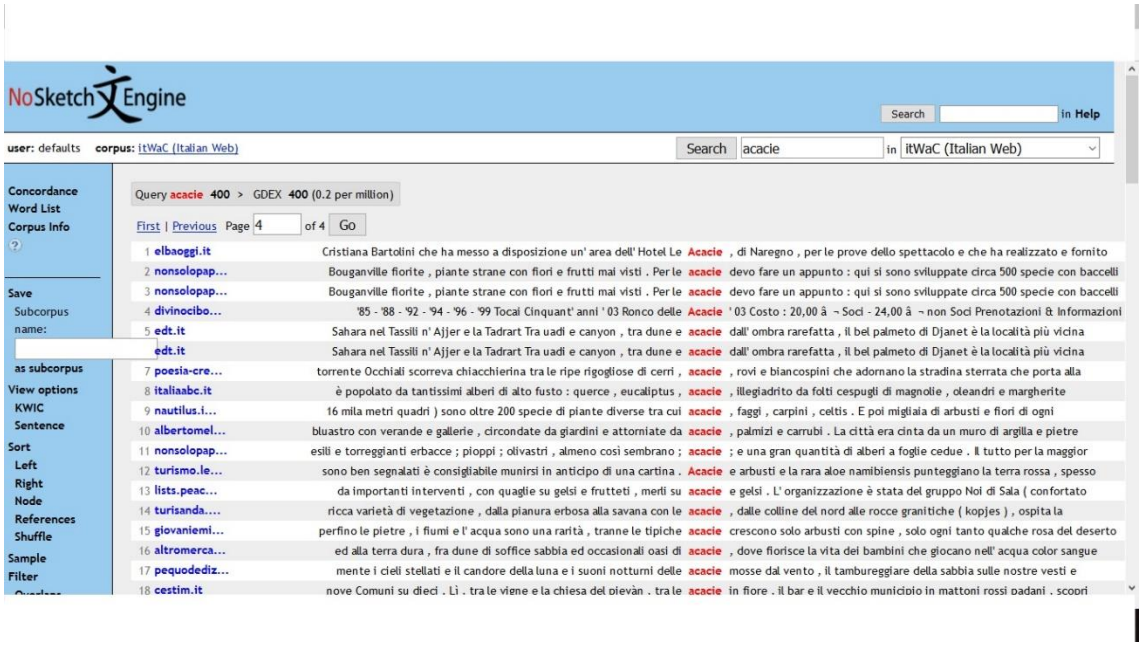
Příklad takového zobrazení je patrný na následujících obrázcích, které byly vytvořeny jako printscreen během vyhledávání v korpusu ItWac jakožto součást praktické části této práce.

---

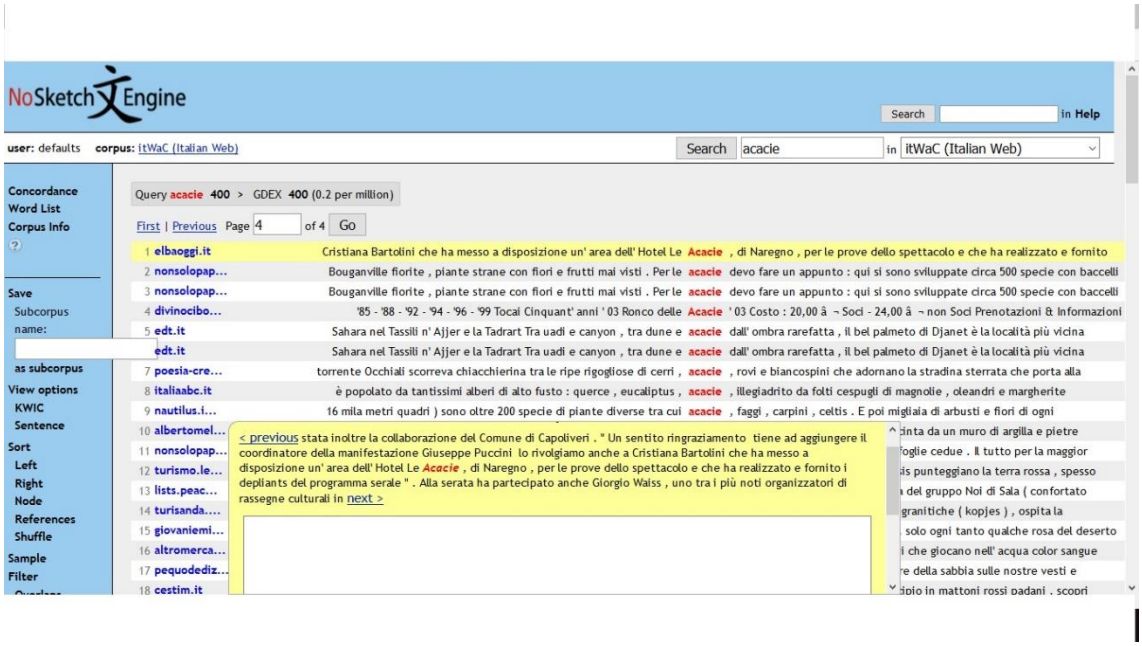
[71] OSOLSOBĚ, Klára. Česká morfologie a korpusy. Karolinum Press, 2014, s. 13. ISBN 9788024625621.
[72] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 18-19. ISBN 978-80-87895-17-7.
[73] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 11-12. ISBN 80-7106-861-6.
[74] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 25 ISBN 978-80-87895-17-7.



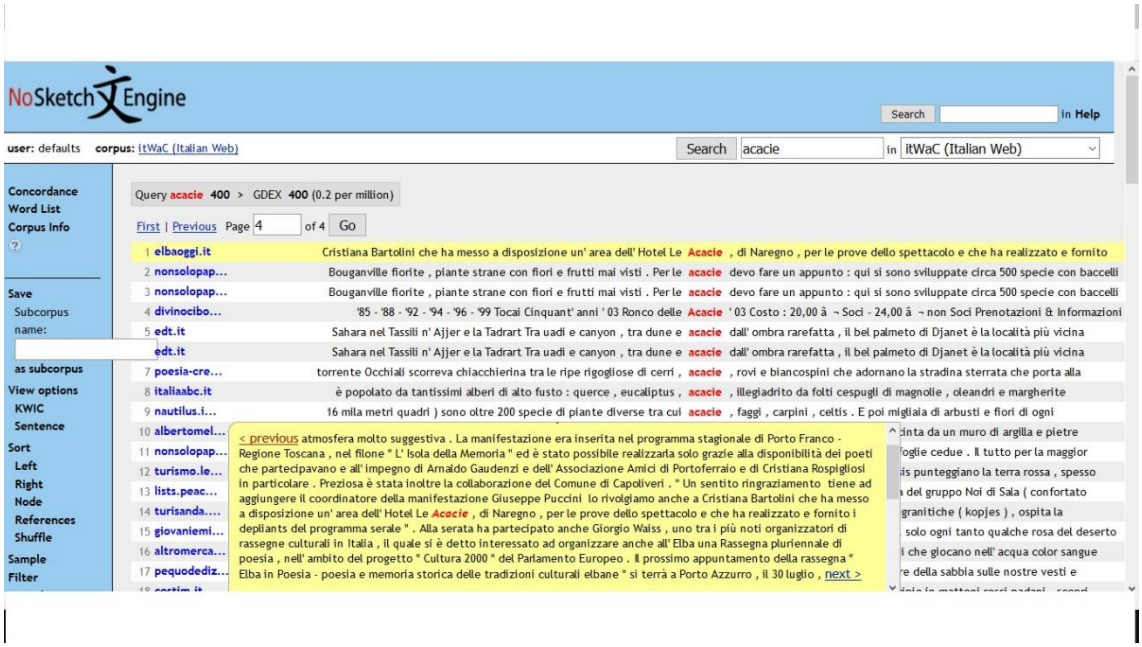
První obrázek je printscreenem úvodní stránky, ze které je mimo jiné patrné, že jako korpusový manažer byl zvolen NoSketch Engine. Vedle vyhledávacího řádku je možno zvolit požadovaný korpus, v tomto případě ItWac.



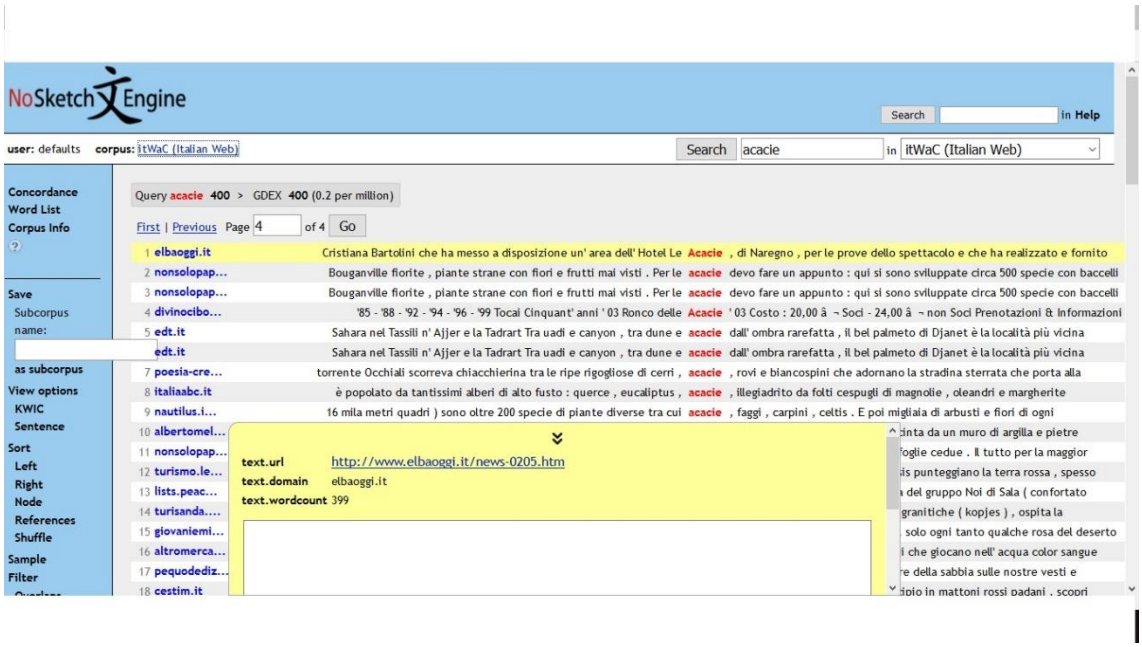
Další obrázek je zobrazením vyhledávání po zadání výrazu „acacie“ do vyhledávacího řádku, tedy tzv. konkordance. Jak již bylo zmíněno, hledaný výraz je odlišen barvou a zarovnán na stejnou úroveň se všemi vyhledanými příklady. Kromě toho je zde patrný také celkový počet nalezených výsledků.



Zvýrazněné vyhledávané slovo slouží svým způsobem jako odkaz, jehož prostřednictvím lze zobrazit část reálného textu, ve kterém byl hledaný výraz nalezen.



Díky odkazům „previous“ a „next“ může být zobrazena ještě širší část původního textu.



Na začátku každého řádku s vyhledaným výrazem, vedle označení pořadí v seznamu, je možno prostřednictvím dalšího odkazu získat konkrétní informace o zdroji, tedy v případě ItWac odkaz na webovou stránku, ale kromě toho i počet slov textu, který byl zahrnut do korpusového souboru dat.

3.4.4 Kolokace a koligace

V souvislosti se vším výše uvedeným je důležité zmínit pojem „kolokace“, a to z toho důvodu, že kolokace je jedním z nejčastějších předmětů studia v konkordanci.[75]

„*Termín kolokace je zaveden pro ustálený souvýskyt několika slov, která spolu sémanticky (ne pouze gramaticky) souvisí. “*[76]

Soubor veškerých kolokací daného výrazu tvoří tzv. kolokační paradigma, které je pro tento výraz zcela jedinečné. S tím souvisí další pojem – kolokabilita, tedy schopnost pojit se sémanticky s dalšími výrazy. Takový výraz, který je schopen kolokace, je pak nazýván jako kolokát. Pokud však pro dané slovo existuje pouze jediná varianta kolokátu, pak je označován jako monokolokabilní.[77]

[75] ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006, s. 12. ISBN 80-7106-861-6.
[76] http://oltk.upol.cz/encyklopedie/index.php5/Kolokace
[77] Tamtéž.

Podstatnou zmínkou je také to, že význam celé kolokace přesahuje souhrn významů jednotlivých kolokátů (např. sousloví „cestovní ruch“).

Výrazným prvkem charakteristiky kolokací je jejich ustálenost. Ta je dána frekvencí konkrétní kombinace v korpusu a zároveň schopností výrazu pojit se s dalšími výrazy v korpusu. Tedy čím častěji se kolokace v korpusu vyskytuje, tím je ustálenější, a čím nižší je schopnost pojit se i s dalšími výrazy, tím je kolokace také ustálenější.[78]

Lexikální kombinace lze klasifikovat především na základě opozic systém-text a pravidelnost-nepravidelnost. František Čermák uvádí následující typologii:

„ *A Systémové*	*1 pravidelné*	*a: termínové kolokace (víceslovné termíny)*
		b: propriální kolokace (víceslovná propria)
	2 nepravidelné	*idiomatické kolokace (idiomy a frazémy)*
B Textové	*3 pravidelné*	*a: běžné kolokace (gram-sémantické kombinace)*
		b: analytické kombinace tvarů (analytické formy)
	4 nepravidelné	*a: individuální metaforické kolokace (autorské metafory)*
		b: náhodné kombinace sousední
		c: jiné kombinace (blábol)

Potenciálně k nim lze ještě přiřadit i přechodný typ (mezi A1a a B3a)

C Textově-systémové *běžné kolokace uzuální“*[79]

Koligace – je v některých případech označována jako poddruh kolokace, který však nesleduje lexikálně-sémantické vztahy, ale vztahy mezi lexikem a gramatickou kategorií, případně mezi lexikem a pozicí ve větě nebo v textu. Příkladem mohou být slovesné vidy, u nichž se jeden z nich pojí pouze s některými příslovci, zatímco druhý vid se pojí s ostatními příslovci. Příkladem může být [80] *„ne/možnost koligace* **pořád skáče/*skočí**“*.[81]

[78] https://wiki.korpus.cz/doku.php/pojmy:kolokace
[79] ČERMÁK, František. *Jazyk a slovník: vybrané lingvistické studie*. Praha: Karolinum, 2014, s. 411. ISBN 978-80-246-2660-4.
[80] https://wiki.korpus.cz/doku.php/pojmy:koligace
[81] Tamtéž.

3.4.4.1 Asociační míry (MI-score a T-score)

Asociační nebo také kolokační míry jsou matematické vzorce, které jsou užívány k vyhledávání daných kolokací v korpusu. Výsledek výpočtu udává míru asociace mezi (ve většině případů dvěma) slovy, která tvoří tuto kolokaci. V některých případech může nastat situace, kdy je výsledek záporný, což poukazuje na negativní asociaci mezi vybranými výrazy, tedy že se tzv. „odpuzují".

V praxi se většinou používá kombinace více asociačních měr, protože každá z nich vyhledává kolokace různě silné o různé relativní frekvenci.[82] Mezi ty vůbec nejznámější a nejužívanější patří MI-score a T-score, přičemž MI-score je schopen vyhledávat kolokace s pevnějšími vazbami mezi výrazy, ale méně častým, až téměř náhodným výskytem, zatímco T-score vyhledává kolokace ustálené, ale spíše nevýrazné.[83]

MI-score vzorec je definován jako:

$I(x,y) = log_2 \frac{P(x,y)}{P(x)*P(y)}$,

kde je *P(x)* pravděpodobnost výskytu vyhledávaného slova *x*

P(y) pravděpodobnost výskytu vyhledávaného slova *y*

P(x,y) pravděpodobnost, že se slovo *y* vyskytuje v kontextu slova *x*.

V jazyce ale není přímo známa pravděpodobnost výskytu slov, a tak je vhodné hodnoty nahradit relativní frekvencí slov v korpusu. V takovém případě vzniká nově upravený vzorec, kde *N* představuje velikost korpusu, resp. počet slov korpusu:

$$mi(x,y) = log_2 = \frac{N * f(x,y)}{f(x) * f(y)}$$

[82] relativní frekvence = poměr počtu výskytů slov (= absolutní frekvence) v korpusu a celkové velikosti korpusu

[83] https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry

Pro T-score platí vzorec:

$$T = \frac{\left(f(x,y) - \frac{f(x)*f(y)}{N}\right)}{\sqrt{f(x,y)}},$$

kde je *f* frekvence výskytu vyhledávaného slova

x vyhledávané slovo

y slovo vyhledávané v kontextu slova x

N velikost korpusu, resp. počet slov korpusu.[84]

[84] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014, s. 42. ISBN 978-80-87895-17-7.

4. Korpusová studie – praktická část, zhodnocení

4.1 Charakteristika využitých korpusů italského jazyka

Volba využitých korpusů byla záměrná, a to z toho důvodů, že každý z nich čerpá z textů z odlišných prostředí. ItWac a PAISÀ jsou korpusy webových textů, La Repubblica korpus novinových textů a korpus CORIS byl vytvořen především na základě prozaických textů z 80. a 90. letech. To mě vedlo k hypotéze, že v korpusech ItWac a PAISÀ by mohlo častěji docházet k výskytu „modernějšího" přístupu ve věci plurálového zakončení substantiv na -cia a -gia, tedy k eliminaci „i" a vytvoření koncovky -ce/-ge.

4.1.1. ItWac

V současnosti se jedná o největší korpus dostupný pro italštinu a v rámci ostatních korpusů italského jazyka také o jeden z nejmladších, byl totiž uveden do provozu teprve v roce 2009. Jde o korpus webových textů, který obsahuje více než jeden a půl miliardy slov, za jehož zakladatele je považován lingvista Marco Baroni.[85] Byl vytvořen na Univerzitě v Boloni jakožto součást projektu WaCky, jehož cílem bylo vytvořit korpusy na základě webových textů v různých jazycích.[86]

4.1.2 La Repubblica

La Repubblica je korpus čerpající z italských novinových textů obsahující přibližně 380 milionů jednotek „token". Stejně jako v případě korpusu ItWac, i zde byla pro vytvoření využita metoda pos-tagging prostřednictvím nástroje TreeTagger. V korpusu dochází k označování žánru (zprávy, komentáře), témat (církev, kultura, ekonomika, vzdělání, novinky, politika, věda, společnost, sport, počasí). Zároveň jsou články v korpusu strukturovány do částí nadpis, podnadpis, text, shrnutí. Dostupné jsou také informace o autorovi článku a roku vydání.[87]

[85] http://www.treccani.it/enciclopedia/corpora-di-italiano_(Enciclopedia-dell%27Italiano)/
[86] GOLA, Elisabetta - ERVAS Francesca. *Metaphor and Communication*. John Benjamins Publishing Company, 2016, s. 55. ISBN 9027267588
[87] http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

4.1.3 CORIS

CORIS neboli Corpus di italiano scritto je korpus s obsahem více než 100 milionů slov, jejichž zdrojem jsou především noviny a knihy, ale také texty akademické, právní, administrativní a další. Projekt pro vytvoření korpusu vypracoval R. Rossini Favretti.[88]

4.1.4 PAISÀ

PAISÀ (čili Piattaforma per l'apprendimento dell'italiano su corpora annotati) je korpus čerpající z internetových textů vytvořený se záměrem poskytnout autentický materiál.[89] Finální verze korpusu obsahovala asi 388 000 dokumentů z 1067 různých webových stránek s více než 250 miliony jednotek „token". Veškeré tyto dokumenty pocházely ze září či října roku 2010.[90]

4.2. Vyhledávání v korpusech

Pro vyhledávání v korpusech byl vybrán reprezentativní vzorek čtrnácti substantiv, která byla zmíněna v úvodní teoretické části práce v rámci kapitoly o pravopisných doporučeních v jednotlivých pramenech. Zároveň byla vybrána tak, aby tvar s jednou variantou zakončení množného čísla nemohl být zaměněn se jménem zcela jiného významu. Z tohoto důvodu nebylo do vzorku zkoumaných jmen zařazeno například často užívané substantivum *camicia*., protože *camicia,* v množném čísle se zakončením -cie, je výrazem pro „košili", ovšem *camicia*, která v množném čísle nese zakončení -ce, je označením pro „bílý plášť" (lékařský).

[88] http://www.treccani.it/enciclopedia/corpora-di-italiano_(Enciclopedia-dell%27Italiano)/
[89] http://www.corpusitaliano.it/
[90] LYDING, Verena a kol. *The PAISÀ Corpus of Italian Web Texts*. In: BILDHAUER, Felix – SCHÄFER, Roland. *Proceedings of the 9th Web as Corpus Workshop (WaC-9).* The Association for Computational Linguistics, 2014, s. 37

Veškerá vyhledávání byla zaznamenána do tabulky rozdělené tak, aby přehledně zobrazovala výsledky vyhledávání. Ke každému jednotlivému příkladu jsou tak přiřazeny nalezené hodnoty, které jsou dále rozděleny podle typu zakončení a podle korpusu, ve kterém bylo hledání provedeno. Zároveň pak byly hodnoty pro oba typy zakončení u jednotlivých příkladů a korpusů převedeny na poměry (zaokrouhlené na celá čísla), které poskytují lepší porovnání všech výsledků. Údaje pro jednotlivé příklady ze všech korpusů byly pak spojeny v celkové hodnoty zaznamenané v druhé tabulce zaměřené tak čistě jen na konečné porovnání číselných hodnot bez ohledu na použitý korpus.

	ItWac		La Repubblica	
	-cie/-gie	-ce/-ge	-cie/-gie	-ce/-ge
l‘acacia	**400**	20	**65**	0
	20:1		---	
la ciliegia	**2 552**	846	**136**	97
	3:1		1,4:1	
la valigia	**4 221**	1,429	**2 087**	475
	3:1		4:1	
la provincia	9 347	**135 901**	444	**4 979**
	1:15		1:11	
la goccia	217	**6 617**	4	**940**
	1:31		1:235	
la spiaggia	741	**19 726**	32	**3 059**
	1:27		1:96	
la scheggia	75	**3 803**	10	**1 444**
	1:51		1:144	
la fiducia	**84**	13	**27**	2
	7:1		14:1	
la pronuncia	125	**6 591**	3	**170**
	1:53		1:57	
la pancia	48	**482**	3	**150**
	1:10		1:50	
la pioggia	324	**10 857**	62	**5 495**
	1:34		1:89	
la socia	**1 770**	14	**131**	1
	126:1		131:1	
la boccia	70	**3 081**	0	**491**
	1:44		---	
la frangia	33	**4 065**	8	**1 516**
	1:123		1:190	

	CORIS		PAISÀ	
	-cie/-gie	-ce/-ge	-cie/-gie	-ce/-ge
l'acacia	**68**	1	**75**	1
	68:1		75:1	
la ciliegia	**236**	58	**309**	36
	4:1		9:1	
la valigia	**816**	126	**418**	142
	7:1		3:1	
la provincia	221	**3 510**	842	**7 829**
	1:16		1:9	
la goccia	18	**1 419**	14	**855**
	1:79		1:61	
la spiaggia	5	**1 088**	48	**2 597**
	1:218		1:54	
la scheggia	2	**501**	18	**511**
	1:251		1:28	
la fiducia	**5**	0	**4**	1
	---		4:1	
la pronuncia	21	**315**	7	**183**
	1:15		1:26	
la pancia	0	**74**	1	**92**
	---		1:92	
la pioggia	34	**821**	68	**1 925**
	1:24		1:28	
la socia	**46**	1	**90**	2
	46:1		45:1	
la boccia	0	**221**	5	**322**
	---		1:64	
la frangia	0	**344**	8	**576**
	---		1:72	

	-cie/-gie	-ce/-ge
l’acacia	608	22
	28:1	
la ciliegia	3 233	1 037
	3:1	
la valigia	7 542	2 172
	3:1	
la provincia	10 854	152 219
	1:14	
la goccia	253	9 831
	1:39	
la spiaggia	826	26 470
	1:32	
la scheggia	105	6 259
	1:60	
la fiducia	120	16
	20:1	
la pronuncia	156	7 259
	1:47	
la pancia	52	798
	1:15	
la pioggia	488	19 098
	1:39	
la socia	2 037	18
	113:1	
la boccia	75	4 115
	1:55	
la frangia	49	6 501
	1:133	

Hned na první pohled je z první tabulky patrné, že u všech uvedených příkladů bylo ve všech využitých korpusech preferované stejné zakončení tvaru množného čísla. Tedy např. pokud byla převážná většina nalezených výsledků tvaru plurálu slova *ciliegia* s koncovkou -gie v ItWac, stejně tak tomu bylo i v ostatních třech korpusech.

3.3 Porovnání s pravopisnými doporučeními v jednotlivých pramenech

Pro větší přehlednost uvádím ve stručnosti znovu veškerá doporučení z úvodní části práce.

Sylva Hamplová v *Mluvnici italštiny* a Patota v publikaci *Grammatica di riferimento dell' italiano contemporaneo* uvádějí pravidlo předcházející hlásky. Tedy pokud koncovce -cia/-gia předchází samohláska, „i“ se v koncovce množného čísla zachovává a vzniká tak zakončení -cie/-gie. Ovšem v případě, kdy koncovce -cia/-gia předchází souhláska, množné číslo je vytvořeno pomocí koncovky -ce/-ge. Hamplová se navíc zmiňuje o vzrůstající tendenci k eliminaci „nadbytečných i“, která nejsou kvůli významu nezbytně nutná.

Dardano a Trifone v *publikaci La nuova grammatica della lingua italiana* uvádějí také pravidlo předcházející hlásky, ale rozšiřují ho navíc o pravidlo etymologické, které je však schopno koexistovat spolu s již zmíněným pravidlem. U etymologického pravidla záleží původu slova. Pokud je slovo ne-lidového původu pocházející z řečtiny nebo latiny, „i“ se zachovává, v opačném případě, tedy pokud je slovo původu lidového, dochází k eliminaci hlásky „i“. Současně zmiňují stejně jako Hamplová tendenci k eliminacím „i“, kromě toho však také možnost využití těchto eliminací jakožto literárního prvku.

Serriani v *Grammatica italiana: italiano comune e lingua letteraria* zmiňuje hned dvě pravidla. První z nich vychází z publikace Camilliho, který prosazuje zachování „i“ ve slovech, jejichž latinský základ zahrnuje -ci-/-gi-. K eliminaci dochází tam, kde je palatální hláska výsledkem jiného základu. Druhé pravidlo je opět pravidlem předcházející hlásky.

l'acacia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky, jedná se o slovo ne-lidového původu, které pochází z latinského *acacia(m)*[91], jež vychází z řeckého *akakía*[84], tedy odpovídá zároveň i pravidlu etymologickému. Vzhledem k tomu, že latinský základ obsahuje -ci-, je potvrzeno i pravidlo podle Serrianiho, resp. Camilliho.

la ciliegia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Jedná se však o slovo pocházející z lidové latiny - *ceresea(m)*[92], a proto by mělo být „i" eliminováno na základě etymologického pravidla, stejně tak jako na základě toho, že latinský základ neobsahuje -ci- a palatální hláska je tak výsledkem jiného základu. Rozpor mezi těmito pravidly a pravděpodobně i vliv vzrůstající tendence k eliminacím hlásky „i" se projevil i ve výsledcích vyhledávání, neboť rozdíl v nalezených číslech u obou zakončení není zdaleka tak velký, jako např. u předcházejícího slova, u kterého došlo ke shodě ve všech zmiňovaných pravidlech.

la valigia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Podle etymologického slovníku je však původ tohoto slova nejistý[93], proto není možné určit, zda slovo vychází z lidového původu nebo jestli jeho základ obsahoval spojení hlásek -gi-. Rozdíl mezi výsledky vyhledávání jsou však zhruba ve stejném poměru jako u předchozího slova *ciliegia*, proto můžeme předpokládat, že zde došlo k prolínání pravidel nebo k projevování vzrůstající tendence k modernizaci a eliminaci „nadbytečného i".

la provincia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Jedná se o slovo latinského ne-lidového původu – *provincia(m)*[94], podle kterého by však „i" mělo být zachováno, stejně tak jako podle Serianniho, resp. Camilliho pravidla, neboť latinský základ je nositelem hláskového spojení -ci-. Opět zde tedy dochází k prolínání pravidel, avšak výsledky vyhledávání jsou vcelku jednoznačné a rozdíly poměrně výrazné, proto zde můžeme uvažovat pouze o vlivu pravidla předcházející hlásky, příp. vlivu vzrůstající tendence k eliminacím.

[91] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/A/AC/acacia.html?q_search=acacia
[92] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/C/CI/ciliegia.html?q_search=ciliegia
[93] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/V/VA/valigia.html?q_search=valigia
[94] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/P/PR/provincia.html?q_search=provincia

la goccia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Jedná se o slovo lidového původu – *guttia(m)*[95], u kterého podle etymologického pravidla dochází k eliminaci „i“. Navíc základ neobsahuje -ci-, a proto je zde potvrzeno i pravidlo podle Serianniho, resp. Camilliho. To dokazují i velmi jednoznačné nalezené hodnoty.

la spiaggia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Etymologický slovník v případě tohoto slova uvádí, že se jedná o slovo odvozené od *piaggia* pomocí předpony *s-*[96]. U slova *piaggia* je pak uvedeno, že se jedná o slovo latinského původu – *plagia(m)*, které vzniklo složením latinského *plaga* a řeckého *plághios*[97]. To však neodpovídá etymologickému pravidlu, které tak v tomto případě není potvrzeno, stejně jako pravidlo podle Serianniho, resp. Camilliho, podle kterého by mělo být „i“ zachováno.

la scheggia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Jedná se o slovo, které pochází z latinského *schidia(m)*, které vzniklo z řeckého *schídion* nebo *schíza* odvozeného od *schízein*[98], a tak by podle etymologického pravidla mělo být „i“ zachováno. To však neodpovídá nalezeným hodnotám. Můžeme však potvrdit pravidlo doporučované Seriannim, resp. Camillim, neboť v základu nenalezneme spojení -gi-.

la fiducia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Substantivum pochází z latinského *fiducia(m)*[99], což potvrzuje etymologické pravidlo, podle kterého se „i“ zachovává. Vzhledem k tomu, že latinský základ je nositelem hláskového spojení -ci-, je potvrzeno i pravidlo podle Serianniho, resp. Camilliho.

la pronuncia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Etymologický slovník uvádí, že substantivum je odvozeno od slovesa *pronunciare*, které pochází z latinského *pronuntiare*[100]. Slovo je ne-lidového původu, a proto by mělo být „i“ v množném čísle zachováno, což však výsledky vyhledávání nepotvrzují. Potvrzeno je však naopak pravidlo, podle kterého se „i“ eliminuje tam, kde je palatální hláska jiného původu než latinského, neboť neobsahuje -ci-.

---

[95] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/G/GO/goccia.html?q_search=goccia
[96] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/S/SP/spiaggia.html?q_search=spiaggia
[97] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/P/PI/piaggia.html?q_search=piaggia
[98] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/S/SC/scheggia.html?q_search=scheggia
[99] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/F/FI/fiducia.html?q_search=fiducia
[100] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/P/PR/pronunciare.html?q_search=pronunciare

la pancia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Substantivum pochází z latinského *pantice(m)*[101], a tak by mělo „i“ zůstat zachováno, což však nalezené hodnoty nepotvrzují. Podle Serianniho, resp. Camilliho by však „i“ mělo být eliminováno, což většinové hodnoty v nalezených výsledcích potvrzují.

la pioggia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Slovo vychází z lidového latinského *ploia(m)*[102], z čehož vyplývá, že by v tomto případě mělo docházet k eliminici „i“ v koncovkách množného čísla, což potvrdily i většinové hodnoty ve výsledcích vyhledávání. Současně s tím bylo potvrzeno i pravidlo podle Serianniho, resp. Camilliho, neboť v latinském základu se nevyskytuje hláskové spojení -gi-.

la socia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Co se týče etymologického původu, ve slovníku nebylo substantivum nalezeno, a proto není možné dále posoudit potvrzení ostatních pravopisných pravidel. Výsledky hledání jsou však velmi jednoznačné a výskytů tvarů plurálového zakončení -ce je jen minimální množství.

la boccia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Podle etymologického slovníku je původ slova nejasný[103], a proto není možné srovnání podle dalších pravidel.

la frangia – výsledky vyhledávání odpovídají pravidlu předcházející hlásky. Etymologický slovník uvádí, že substantivum pochází z francouzského *frange*, které vzniklo z lidového latinského *frimbia(m)*[104]. Na základě tohoto lidového původu by mělo docházet k eliminaci „i“, což většinové výsledky ve vyhledávání potvrdily, stejně tak jako nepřítomnost -gi- v původním latinském základu. I proto ve výsledkové tabulce nalezneme většinové hodnoty pro plurálové zakončení -ge.

Veškerá výše uvedená porovnání nalezených výsledků s pravopisnými pravidly z jednotlivých pramenů jsou pro přehlednost zaznamenána ještě v následující tabulce, která současně zobrazuje i očekávané formy plurálu podle těchto pravidel.

[101] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/P/PA/pancia.html?q_search=pancia
[102] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/P/PI/pioggia.html?q_search=pioggia
[103] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/B/BO/boccia.html?q_search=boccia
[104] http://saperelb-538884594.eu-west-1.elb.amazonaws.com/sapere/dizionari/dizionari/Italiano/F/FR/frangia.html?q_search=frangia

	Pravidlo předcházející hlásky	Etymologické pravidlo (Dardano a Trifone)	Pravidlo podle latinského základu (Serriani, resp. Camilli)	Preferovaný tvar plurálu na základě vyhledávání
l'acacia	le acacie	le acacie	le acacie	**le acacie**
la ciliegia	le ciliegie	le ciliege	le ciliege	**le ciliegie**
la valigia	le valigie	---	---	**le valigie**
la provincia	le province	le provincie	le provincie	**le province**
la goccia	le gocce	le gocce	le gocce	**le gocce**
la spiaggia	le spiagge	le spiaggie	le spiaggie	**le spiagge**
la scheggia	le schegge	le scheggie	le schegge	**le schegge**
la fiducia	le fiducie	le fiducie	le fiducie	**le fiducie**
la pronuncia	le pronunce	le pronuncie	le pronunce	**le pronunce**
la pancia	le pance	le pancie	le pance	**le pance**
la pioggia	le piogge	le piogge	le piogge	**le piogge**
la socia	le socie	---	---	**le socie**
la boccia	le bocce	---	---	**le bocce**
la frangia	le frange	le frange	le frange	**le frange**

3.4 Zhodnocení zjištěných výsledků

Jak již bylo zmíněno v úvodu kapitoly, ve všech využitých korpusech u všech vybraných příkladů byl preferován stejný typ zakončení množného čísla, ačkoliv v některých případech ve velmi rozdílných poměrech. Toto zjištění je důkazem toho, že preference typů zakončení je stejná v jakékoliv oblasti psaného jazykového projevu, ať už je jedná o novinové články nebo webové texty, ačkoliv by se právě u internetových zdrojů dala předpokládat vyšší inklinace k užívání „modernějšího" přístupu a tedy častější eliminace „nadbytečného i" navzdory jakýmkoliv pravidlům. V určitých případech došlo dokonce k tomu, že v rámci jednoho korpusu nebyla varianta se zakončením množného čísla s eliminovaným „i" bez ohledu na pravopisná doporučení nalezena vůbec.

Co se týče srovnávání nalezených výsledků u jednotlivých příkladů s uvedenými pravopisnými pravidly z různých zdrojů, jistě je velmi zajímavým faktem to, že veškerá substantiva z vybraného reprezentativního vzorku se shodují s pravidlem, podle kterého je pravopis koncovky možného čísla řízen v závislosti na hlásce, která předchází koncovku v singuláru –cia/-gia.

Ostatní pravidla, vycházející z původu substantiv, nehrají příliš velkou roli, neboť jen velmi málo lidí má natolik pokročilou znalost latiny na to, aby se mohli těmito pravidly řídit. U některých příkladů je možno pozorovat, že pokud došlo ke shodě ve všech třech pravopisných pravidlech, poměr mezi preferovaným a méně frekventovaným zakončením byl mnohem výraznější než v ostatních případech.

ZÁVĚR

Hlavním cílem předložené práce bylo na základě korpusové studie zjistit, jaký je aktuální úzus v tvoření plurálu substantiv na –cia a –gia a poté tento úzus porovnat s doporučovanými pravopisnými pravidly.

V první části práce jsem uvedla a představila problematiku substantiv končících na –cia a –gia. Tedy na počátku popis pravidelného tvoření množného čísla obecně podstatných jmen ženského rodu a dále pak těch, která jsou nositeli koncovky -cia/-gia. Na tuto část jsem navázala souhrnem pravopisných doporučení několika autorů a popisem důvodu vzniku této problematiky, včetně historického počátku, který nalezneme v procesu palatalizace hlásek c a g.

V následující kapitole jsem se zaměřila na teorii korpusové lingvistiky. Z počátku jsem shrnula vývoj korpusové lingvistiky, v němž velmi význačnou roli hrála kritika Noama Chomského v polovině minulého století. Poté jsem navázala popisem korpusu jako takového se zaměřením na jejich tvorbu, od tokenizace a značkování, přes korpusové manažery a praktickou ukázku konkordancí až po stručné představení kolokací a s nimi spojených asociačních měr.

Závěrečná třetí kapitola je stěžejní částí práce, neboť zahrnuje veškeré výsledky vyhledávání a tabulky četnosti výskytů obou variant zakončení u čtrnácti zvolených podstatných jmen, která byla vybrána z uvedených příkladů v první kapitole práce. Tyto výsledky jsem pak porovnala s doporučenými pravopisnými pravidly uvedenými rovněž v první kapitole. Mimo to jsem uvedla také stručnou charakteristiku jednotlivých využitých korpusů, a to zejména z toho důvodu, že každý z nich čerpá svá jazyková data z odlišných prostředí. Výsledky jsem pak zhodnotila a popsala aktuální tendence ve volbě plurálové koncovky.

Za pomoci vyhledávání ve čtyřech online dostupných italských korpusech bylo zjištěno, že aktuální pravopis zkoumaných podstatných jmen se řídí ve většině případů podle tzv. pravidla předcházející hlásky, které je vcelku novodobou záležitostí, neboť bylo rozšířeno teprve ve druhé polovině minulého století. Do té doby byla užívána pravidla vycházející z původu slov, která jsou však pro dnešní uživatele jazyka jen těžko použitelná, a to z toho

důvodu, že vyžadují velmi pokročilou znalost latiny a znalost původu italských slov vůbec. Do italského jazyka však v nedávné době začala pronikat také tendence eliminovat veškerá nadbytečná koncovková „i“ bez ohledu na jakákoliv pravopisná doporučení, tedy v podstatě jakási modernizace jazyka. Tato tendence se však ve vyhledávání a nalezených výsledcích nijak výrazně neprojevila, z čehož lze usoudit, že uživatelé se raději uchylují k užití již zavedeného snadno použitelného pravidla, než k modernímu jednoznačnému přístupu.

Na úplný závěr lze zkonstatovat, že problematika představená v předkládané práci, je bezesporu velmi aktuálním tématem, a to nejen pro studenty italského jazyka, ale i pro rodilého mluvčího, což je dáno zejména neexistencí jednoho závazného pravidla. Stejně tak aktuální je i oblast korpusové lingvistiky, která je díky novým možnostem v oboru počítačových technologií na vzestupu.

RESUMÉ

Bakalářská práce se zabývá problematikou tvoření množného čísla italských substantiv na –cia a –gia, přičemž součástí práce je i korpusová studie. Hlavním úkolem této práce je popsat na základě výsledků vyhledaných prostřednictvím korpusů aktuální tendence v této oblasti pravopisného kolísání.

Úvod práce vysvětluje celou problematiku, tedy její princip a vznik včetně historického opodstatnění, a hlavní pravopisná doporučení z různorodých pramenů. Následující dvě kapitoly se věnují oblasti korpusové lingvistiky a korpusů. V počátku druhé kapitoly byl shrnut vývoj korpusové lingvistiky, na nějž navazuje část věnující se hlavnímu předmětu korpusové lingvistiky – korpusu. Ve třetí kapitole bylo vymezeno jeho využití, charakterizovány jeho typické znaky a na závěr stručně popsána jeho tvorba, včetně tokenizace, praktického zobrazení konkordancí či charakteristiky kolokací. Závěrečná kapitola pak shrnuje výsledky vyhledávání ve čtyřech online dostupných korpusech, z nichž každý čerpá svá jazyková data z odlišných prostředí. Tyto výsledky pak byly porovnány s doporučenými pravopisnými pravidly z první kapitoly.

Prací bylo zjištěno, že aktuální pravopis plurálu substantiv na –cia a –gia se řídí především podle pravidla, jehož princip je založen na hlásce, která předchází koncovce –cia nebo –gia. Lze usuzovat, že tím hlavním důvodem pro jeho časté využití je snadná aplikace pro širokou veřejnost, ať už rodilé mluvčí, či studenty italského jazyka. Pravidla vycházející z původu slov, která vyžadují pokročilou znalost latiny, jsou pro běžného uživatele jen těžko využitelná, což se projevilo i ve výsledcích vyhledávání.

Ke zpracování práce byla využita obsahová analýza literatury popisující problematiku pravopisného kolísání a pravopisná doporučení, a analýza literatury zabývající se korpusovou lingvistikou a korpusy. V závěrečné části práce bylo využito online dostupných korpusů a online etymologického slovníku.

RESUMÉ

La mia tesi di laurea si dedica a tema di formazione del plurale dei nomi italiani in –cia e –gia, fa parte della tesi anche studio dei corpora. L'obiettivo principale della tesi è descrivere a base dei risultati trovati per il tramite dei corpus la tendenza attuale in questo campo di oscillazione ortografica.

L'introduzione della tesi spiega in totale la problematica, allora il suo principio e l'origine incluso motivazioni storici e principali raccomandazioni ortografici dei fonti svariati. I due capitoli seguenti si dedicano alla linguistica dei corpora e al proprio corpora. All'inizio del secondo capitolo è stato riassunto sviluppo della linguistica dei corpora a cui riallaccia una parte che caratterizza l'oggetto principale del campo – il corpora. Nel terzo capitolo sono state definite i suoi caratteristiche tipiche e dopo sono stato definito e caratterizzato il suo uso e la sua formazione incluso tokenizzazione, dimostrazione pratica delle concordanze o caratteristica delle collocazione. L'ultimo capitolo riassume risultati della ricerca via quattro corpora online disponibili che attingono i dati della lingua dei sorgenti diversi. Questi risultati dopo sono stati comparati con raccomandazioni ortografici del primo capitolo.

È stato scorpito che l'ortografia attuale del plurale dei sostantivi in –cia e –gia segue soprattutto la regola della quale principio è basato su suono che precede le desinenze –cia o –gia. È possibile concludere che il motivo principale per il suo uso frequente é applicazione semplice per una larga parte del pubblico, sia per un madrelingua che per studenti della lingua italiana. Le regole, che sono derivate da origine delle parole e che richiedono avanzata conoscenza del latino, sono per gli utenti comuni utilizzabili solo con grande difficoltà, e questo si è dimostrato anche nei risultati della ricerca.

Per l'elaborazione della tesi sono stata usata analisi della letteratura che descrive la problematica della oscillazione ortografica e racommandazioni ortografici e anche analisi della letteratura che si dedica alla linguistica dei corpora. Nella parte finale sono stati usati i corpora online disponibili e un dizionario etimologico online.

SEZNAM ZDROJŮ A PRAMENŮ

Literatura:

1. ČERMÁK, František. *Jazyk a jazykověda: přehled a slovníky*. Vyd. 4., V Karolinu 2., dopl. Praha: Karolinum, 2011. ISBN 978-80-246-1946-0.

2. ČERMÁK, František, ed. *Korpusová lingvistika Praha 2011*. Praha: NLN, Nakladatelství Lidové noviny, 2011. Studie z korpusové lingvistiky, sv. 15. ISBN 978-80-7422-115-6.

3. ČERMÁK, František (ed.) – BLATNÁ, Renata (ed.). *Korpusová lingvistika - stav a modelové přístupy*. Vyd. 1. Praha: NLN, Nakladatelství Lidové noviny, 2006. ISBN 80-7106-861-6.

4. DARDANO, Maurizio. *Manualetto di linguistica italiana.* Vyd. 2. Bologna: Zanichelli, 1996. ISBN 978-8808088567.

5. DARDANO, Maurizio – TRIFONE, Pietro. *La nuova grammatica della lingua italiana*. 1. vyd. Bologna: Zanichelli, 1997. ISBN 8808104265.

6. GOLA, Elisabetta - ERVAS Francesca. *Metaphor and Communication*. John Benjamins Publishing Company, 2016. ISBN 9027267588

7. HAMPLOVÁ, Sylva. *Mluvnice italštiny: Grammatica italiana*. Vyd. 1. Praha: LEDA, 2004. ISBN 80-7335-041-6.

8. MAIR, Christian – HUNDT Marianne. *Corpus Linguistics and Linguistic Theory, Vydání 1999.* Rodopi, 2000. ISBN 9042014938.

9. MARAZZINI, Claudio. *La lingua italiana. Profilo Storico.* Il Mulino, 2002, s. XX. ISBN 978-8815086754.

10. MCENERY, Tony a Andrew HARDIE. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012. Cambridge textbooks in linguistics. ISBN 978-0-521-54736-9.

11. OSOLSOBĚ, Klára. Česká morfologie a korpusy. Karolinum Press, 2014, s. 13. ISBN 9788024625621.

12. PATOTA, Giuseppe. *Grammatica di riferimento dell'italiano contemporaneo*. Novara: Garzanti Linguistica, 2006. ISBN: 9788848040310.

13. POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014. ISBN 978-80-87895-17-7.

14. SERIANNI, Luca – CASTELVECCHI, Alberto. *Grammatica italiana: italiano comune e lingua letteraria*. 1. vyd. Torino: UTET, 1989. ISBN 8802041547.

15. WONNER, Bernadette. *The development of corpus linguistics to its present-day concept*. GRIN Verlag, 2005. ISBN 9783638373777.

Elektronické zdroje:

1. Il centro per le ricerche di computer della Sezione di Studi in Lingue Moderne per Interpreti e Traduttori dell‘ Università di Bologna [online], dostupné z: http://sslmitdev-online.sslmit.unibo.it/

2. BARONI, Marco. *Corpora di italiano.* L’Enciclopedia italiana Treccani [online], dostupné z: http://www.treccani.it/

3. Centrum zpracování přirozeného jazyka [online], dostupné z: https://nlp.fi.muni.cz/web3/

4. Databáze článků v českých novinách a časopisech Národní knihovny [online], dostupné z: http://full.nkp.cz/

5. Katedra obecné lingvistiky [online], dostupné z: http://oltk.upol.cz/

6. Lancaster University [online], dostupné z: http://www.lancaster.ac.uk/

7. Language Corpus Management and Query System [online], dostupné z: https://www.sketchengine.co.uk/

8. L'enciclopedia De Agostini, dizionari di italiano, strumenti per lo studio e l'approfondimento [online], dostupné z: http://www.sapere.it/

9. Grammatica italiana (2012) L’Enciclopedia italiana Treccani [online], dostupné z: http://www.treccani.it/

10. Příručka ČNK [online], dostupné z: https://wiki.korpus.cz/doku.php

Audiovizuální zdroje:

CVRČEK, Václav. *Dějiny v zrcadle jazyka. Jazykové korpusy – výzkumný nástroj nejen pro jazykovědce.* prezentace dostupná z: https://www.ustrcr.cz/

Korpusové zdroje:

1. A corpus of written Italian [online], dostupný z: http://corpora.dslo.unibo.it/coris_ita.html

2. Corpus Italiano PAISÀ [online], dostupný z: http://www.corpusitaliano.it/

3. Corpus La Repubblica [online], dostupný z: http://dev.sslmit.unibo.it/

4. Italian Web Corpus itWac [online], dostupný z: http://nl.ijs.si/noske/index-en.html

Ostatní zdroje:

1. LYDING, Verena a kol. *The PAISÀ Corpus of Italian Web Texts*. In: BILDHAUER, Felix – SCHÄFER, Roland. *Proceedings of the 9th Web as Corpus Workshop (WaC-9).* The Association for Computational Linguistics, 2014.

2. STORJOHANN, Petra. *Corpus-driven vs. corpus-based approach to the study of relational patterns*. Institut für Deutsche Sprache Mannheim.