ČASOPIS PRO MODERNÍ FILOLOGII 107, 2025, Č. 1, S. 7–18

Pasti dat: srovnatelnost dat jazykových korpusů

Markus Giger – Jana Kocková (Praha)

DATA TRAPS: COMPARABILITY OF LANGUAGE CORPUS DATA
Despite the apparent unambiguity of data provided by corpora, the data reflect different composition of the corpora, different conceptions of the synchronic period of a given language, different linguistic traditions, different orthography and other factors. We focus on the most common reasons affecting the comparability of data in parallel corpora, such as unequal lemmatization, tagging and tokenization, and illustrate them with examples from Czech, German and Russian. For example, when comparing Russian and Czech verb forms and lemmas, the data provided by the corpora are not comparable, because in Russian, unlike in Czech, the reflexive and non-reflexive forms are assigned to different lemmas and the verb lemma includes participles, whereas the corresponding Czech forms are tagged as adjectives, in accordance with Czech philological tradition. The differing approaches to tokenization are also reflected in the overall size of the corpus, indirectly affecting the comparability of relative frequencies.

KEYWORDS
corpora; comparative linguistics; tagging; data comparability; corpus balance

KLÍČOVÁ SLOVA
korpusy; komparativní lingvistika; tagování; srovnatelnost dat; vyváženost korpusů

DOI
https://doi.org/10.14712/23366591.2025.1.1

1. KORPUS JAKO ZDROJ DAT

Korpusy, a především paralelní korpusy, jsou dnes stabilní součástí kontrastivního jazykového výzkumu. Kromě cíleného získávání jazykového materiálu umožňují využívat i řadu statistických údajů. Základním údajem je frekvence na různých úrovních jazyka, od hlásek po syntagmata. Frekvence umožňuje odlišit centrální a okrajové jazykové jevy, ale využívá se i ve výzkumu syntagmatických jevů (frazeologie, kolokabilita), lexikalizace nebo identifikace terminologie. V kontrastivním jazykovém výzkumu umožňuje zejména doložit mezijazykové rozdíly v situacích, kdy se daný jev vyskytuje v různých jazycích, ale jeho četnost se zásadně liší. Korpusy obvykle poskytují nejen údaj o absolutní frekvenci jevu, ale i o relativní frekvenci, tj. frekvenci vzhledem k celkové velikosti korpusu.[1] Tato data jsou však ovlivněna celou řadou faktorů (srov. také Stefanowitsch, 2020), které různým způsobem mění jejich srovnatelnost napříč jazyky. U jednoduchých výzkumných

1 Např. počet výskytů na milion slov (i.p.m.) v ČNK, NKRJa, DeReCo IDS Mannheim, korpusech Aranea, nebo ARF (průměrná redukovaná frekvence) v ČNK.

zadání je snazší odhalit interferenci dat nehledě na složitost vlastního korpusového dotazu. U komplexních komparativních zadání často není možné bez důkladné znalosti výchozího značkování v daných jazycích interference odhalit (např. pro srovnání frekvence sloves v různých slovanských jazycích je nutné se vypořádat s různou strukturou slovesného lemmatu — (ne)zahrnutí participií, reflexivnosti atd.). Řadu interferencí lze kompenzovat, pokud si je lingvista dané problematiky vědom. Často ovšem dochází k tomu, že v komplexnějších zadáních není daná interference zřejmá. Níže volíme příklady, které jsou elementární, ale díky tomu názorné.

Rozdíly mezi korpusovými daty je možné shrnout do tří základních oblastí: rozdíly způsobené odlišnou strukturou korpusů (viz část 2), odlišnou strukturou jazyků (viz část 3) a rozdíly způsobené různými lingvistickými tradicemi a přístupy (viz část 4). Jednotlivé kategorie spolu často souvisí a překrývají se.

2. ODLIŠNOSTI KORPUSŮ

Rozdíly v datech způsobené odlišnou strukturou korpusů se týkají jak dat získaných z jednojazyčných korpusů, tak i z korpusů paralelních.[2] U jednojazyčných korpusů je zásadním faktorem vždy složení textů v korpusu,[3] srovnatelnost složení korpusů je ovšem vždy problematická. Určitým vodítkem je procentuální složení podle žánrů. Ale ani podobná žánrová struktura nezajišťuje podobné složení korpusů. Pojetí žánrů se často liší pod vlivem národních tradic, dostupných textů a dalších faktorů: např. ruský korpus NKRJa a ukrajinský korpus Grac v.18 mají jako samostatný žánr religiózní texty a memoárovou literaturu; ve slovenském korpusu SNK jsou memoáry zahrnuty do beletristiky, v ČNK do odborné literatury. V této souvislosti bývá zmiňována tzv. vyváženost a reprezentativnost korpusu, nicméně i tyto pojmy se mohou podstatně lišit.[4] Reprezentativní korpus by měl zahrnovat všechny variety daného jazyka (častěji jeho psané části). Vyváženost korpusu se nejčastěji odvozuje od produkce, nebo recepce textů, zjednodušeně řečeno od toho, kolik textů se v určitém žánru vydává nebo „konzumuje“. Obě kritéria jsou již sama o sobě obtížně objektivně stanovitelná, zvláště v době, kdy se většina textů vydává a čte v elektronické podobě. Kromě toho se i korpusy s podobně stanovenými makroskupinami textových typů mohou značně lišit v mikro-

2 K rozdílné struktuře paralelních korpusů více v části 3.

3 Jakým způsobem ovlivňuje složení korpusu výsledná data, můžeme dobře ilustrovat na nejfrekventovanějších vlastních jménech v korpusu: v ČNK InterCorp v16ud jsou nejfrekventovanějšími vlastními jmény *John, Jack, David*. Je to způsobeno vysokým podílem překladů z angličtiny a filmových titulků. V ruské části paralelního korpusu InterCorp v16ud je to *Гарри, Джон, Джек*, přitom v základním korpusu NKRJa je to *Иван, Николай, Александр*. Pro srovnání nejfrekventovanějších lemmat v češtině a ukrajinštině viz také Kocková a Sytar (2024).

4 Srov. například definici NKJP: „Balance means that none of text types covers more than half of the corpus“ (Dąbrowska, 2013, s. 30).

struktuře: např. zařazení memoárů, filmových titulků, religiózní literatury apod. Pojetí vyváženosti korpusů se tedy v praxi značně liší a zpravidla se zaměřuje na žánry, nikoli na zastoupení textů z hlediska časového rozložení. Například korpus americké angličtiny MASC se prezentuje jako vyvážený, protože obsahuje od každého žánru stejný podíl (5–6 %). Podobný přístup, ale s jinou strukturou zvolili tvůrci „centrálního" korpusu (Kernkorpus) DWDS a vyvážené korpusy ČNK[5] a SNK (např. syn2020, prim-10.0-public-vyv), které obsahují podobný podíl čtyř, respektive v ČNK tří psaných textových typů (beletristika, noviny, vědecké odborné texty, užitné texty; resp. beletristika, oborová literatura, publicistika). NKRJa stanovuje vyvážené složení pro každou epochu a zaměřuje se na texty, které odrážejí jazykový standard spisovného jazyka doby. Francouzský korpus CRFC (dosud není dostupný) prezentují Siepmann a kol. (2015, s. 8) jako vyvážený i z hlediska mluveného a psaného jazyka: stejný díl mají tvořit mluvené a „pseudomluvené" texty (titulky) a psané texty.

Druhým faktorem, který není na první pohled zřejmý, je různé pojetí synchronního období. Stanovit synchronní období nelze univerzálně pro všechny jazyky. Korpusy často vycházejí z vnějších událostí, které se považují za zásadní pro vývoj daného jazyka (konec druhé světové války, vznik samostatného státu, změna politického uspořádání apod.). Může jít ovšem i o jiné důvody, které nemusí být vždy zřejmé: britský korpus BNC zahrnuje texty od roku 1975; americký MASC texty od roku 1990; základní ruský korpus NKRJa zahrnuje období od roku 1700, tj. od Petrovského období. ČNK zahrnuje do synchronního korpusu texty od roku 1945, SNK od roku 1955. Často bývají do korpusů zahrnována i starší díla, která jsou považována za zásadní pro daný jazyk.[6]

3. ODLIŠNÁ STRUKTURA JAZYKŮ

Kromě složení samotného korpusu je při práci s frekvencemi výskytu (ale i s počtem lemmat apod.) nutné vzít v úvahu řadu faktorů, které získaná data ovlivňují. Názorně můžeme takovou interferenci demonstrovat na srovnání frekvence výskytu lemmat sloves v češtině, ruštině, slovenštině a němčině z paralelního korpusu InterCorp v16[7] (vždy v odpovídajících kombinacích č.–ru., č.–n. atd.):

5 Podíl jednotlivých typů textů se v ČNK podstatně liší v jednotlivých řadách vyvážených korpusů; v SYN2020 má beletrie podíl 15 %, v SYN2005 40 % (k vyhodnocení reprezentativnosti korpusů řady syn viz Křen 2012).

6 V BNC jsou to některé texty od roku 1964, podobně také v centrálním korpusu DWDS: „Kerncorpus must contain a considerable amount of influential and important literature." (Geyken 2007, s. 24) V ČNK zahrnuje řada SYN v3 až 13 i „zásadní" díla vydaná před rokem 1945 (např. *Osudy dobrého vojáka Švejka*, vyd. 1921–23).

7 Není-li uvedeno jinak, jsou příklady v tomto článku převzaty z paralelního korpusu InterCorp v16.

	lemma (čeština)	frekvence lemmatu	lemma (zarovnaný jazyk)	frekvence lemmatu
1	*učit*	15 757	ru. *учить*	8 060
2	*koupat*	1 138	ru. *купать*	108
3	*vstát*	12 794	n. *aufstehen*	5 533
			n. *auferstehen*	771
4	*představit*	25 847	n. *vorstellen*	21 600
5	*odhadnout*	2 154	n. *einschätzen*	1 411
6	*moci*	399 296	sk. *môcť*	299 259

TABULKA 1. Frekvence lemmat v paralelním korpusu InterCorp v16

Frekvence uvedené v tabulce 1 neodrážejí skutečně různý výskyt daných sloves v uvedených jazycích (záměrně jsme volili příklady, které jsou silně ekvivalentní ve zvolených jazycích), ale jsou ovlivněny odlišnou strukturou jazyků, která se odráží nejen v pravopise, ale sekundárně i v tokenizaci a lemmatizaci korpusů. Především u komplexnějších dotazů dochází k tomu, že je často výsledek mylně interpretován ve smyslu různosti jazyků, nikoli jako důsledek jiné strukturace korpusů. Typické příčiny odlišností můžeme zhruba shrnout do následujících skupin: reflexivnost, odděleně psané části slov (např. odlučitelné předpony v němčině), negace a kompozice (ke kompozitům viz blíže 4.2).[8]

3.1. VLIV REFLEXIVNOSTI

U prvních dvou příkladů v tabulce 1 je příčinou odlišné frekvence reflexivnost. Zatímco v ruštině je vyjádřena reflexivním postfixem, který se píše dohromady se slovesem, reflexivní slovesa jsou tokenizovaná spolu s tímto postfixem a mají samostatné lemma, v češtině se jedná o volný prvek, sloveso a reflexivní prvek jsou tokenizovány obvykle samostatně (s výjimkou některých kondenzátů jako *ses, sis*) a reflexivní slovesa spadají do jednoho lemmatu spolu s nereflexivním slovesem. Frekvence výskytu slovesa *učit* tedy zahrnuje reflexivní i nereflexivní tvary, frekvence výskytu slovesa *учить* pouze nereflexivní tvary. Podobná situace jako v češtině je i v němčině a dalších jazycích s volným reflexivním prvkem (např. *mýt / mýt se*; *waschen / sich waschen* je z hlediska korpusu jedno lemma).

3.2. VLIV ODLUČITELNÝCH ČÁSTÍ SLOVA

Příčinou odlišné frekvence může být také odlučitelný prefix — viz řádky 3–5 v tabulce 1. Ve většině korpusů jsou do německého slovesného lemmatu zahrnovány pouze tvary, u kterých není předpona odloučena (srov. lemma *aufstehen*: *aufstehen, aufgestanden, aufzustehen, aufstand, aufsteht, aufstehe, aufstehst, aufstanden, aufstünde,*

8 Mezi další faktory patří například stupňování adjektiv a adverbií, různé pojetí předložek a spojek atd.

aufstandest).[9] Stejným způsobem musíme počítat například se zkreslením frekvencí předložek, adverbií (*wiedersehen / wir sehen uns wieder*), substantiv (*teilnehmen / ich nehme teil*). Například předložka *před* má v paralelním německo-českém korpusu frekvenci 174 897 výskytů, ekvivalent *vor 336 268 výskytů*.[10]

3.3. VLIV NEGACE

Podobně jako reflexivita se do tokenizace promítá i negace, ovšem ne vždy ovlivňuje frekvenci lemmat. Například lemma českých sloves v korpusech ČNK zahrnuje i negativní tvary. Pokud tedy porovnáváme frekvenci lemmat například s ruštinou, nebo němčinou, kde se negace píše odděleně, nebudou data zkreslena — lemma *číst* zahrnuje tvary *čtu/nečtu, čteš/nečteš…*; ruské lemma *читать* zahrnuje i negativní tvary, protože se negace píše odděleně: *читаю* / (*не*) *читаю*; obdobně pro německé lemma *lesen*: *ich lese* (*nicht*). Jiná situace nastane při srovnání češtiny se slovenštinou, která má jedno lemma pro kladné tvary a jedno lemma pro záporné tvary, tj. lemma *môcť* a *nemôcť* (srov. také řádek 6 v tabulce 1). Do pojetí negace se tak promítají i různé lingvistické tradice (viz také část 3), které ovlivňují nejen národní korpusy, ale přecházejí i do paralelních korpusů.[11]

Ovšem i v jazycích, jako je ruština nebo němčina, se nestejná tokenizace projeví při hledání tvarů. Rozdíl je evidentní, pokud srovnáme frekvenci výskytu ruského imperativu *говори* (13 649 výskytů) a českého *mluv* (1 944 výskytů), případně *říkej* (674 výskytů). Frekvence v ruštině zahrnuje pozitivní i negativní tvary (*не говори*), v češtině nikoli (*nemluv* je jiný token než *mluv*). Odděleně psanou a značkovanou negaci má většina evropských jazyků (tj. například frekvence *Sprich*! zahrnuje i *Sprich nicht*! atd.).

Problém negace se týká i jiných slovních druhů, u nichž bývá tato problematika ještě komplikovanější. V češtině jsou negovaná adjektiva součástí jednoho lemmatu spolu s pozitivními adjektivy (lemma *jasný* zahrnuje *jasného, nejasného* atd.), stejně tak adverbia (lemma *jasně* zahrnuje i *nejasně*). V ruštině mají negované tvary samostatné lemma (*ясный* a *неясный*). V němčině (i angličtině) mají pozitivní a negativní tvary samostatné lemma (n. *klar, unklar;* ang. *clear, unclear*), ale zároveň jsou atributivní i neatributivní tvary zahrnuty do jednoho lemmatu. Tedy n. *klar, klare, klares*

9 Pokud vyhledáváme tvary lemmatu jednoho slovesa, nebo určité skupiny sloves, lze vyhledat většinu tvarů s odloučenou předponou pomocí rozšířeného dotazu, např. pro německý korpus DeReKo: &herausfordern oder (&fordern /+s0 heraus). Tento způsob ovšem není možné efektivně využít u širších dotazů, například pokud srovnáváme frekvenci skupiny transitivních a intransitivních sloves (*eintreten — betreten, beantworten — antworten, bedenken — nachdenken, ankleiden — bekleiden* atd.). Opět platí, že čím komplexnější dotaz, tím je složitější zkreslení dat odhalit a vyrovnat, pokud si není lingvista daného zkreslení vědom.

10 V korpusu InterCorp v16 může být sice *vor* anotováno i jako část slovesa (PART:Verb), ale anotace je velmi chybová. V některých případech je *vor* anotováno jako adverbium.

11 Čeština je anotována stejným způsobem v ČNK i v paralelním slovensko-českém korpusu SNK a rusko-českém korpusu NKRJa. Stejně tak slovenština má v SNK i v paralelním korpusu ČNK dvě lemmata pro kladné a záporné slovesné tvary.

atd. jsou součástí jednoho lemmatu odpovídajícího českému adjektivnímu (*jasný*) i adverbiálnímu (*jasně*) lemmatu, české lemma ovšem na rozdíl od německého zahrnuje i negativní tvary (*jasný, jasná, nejasný, nejasná* atd.).

4. ODLIŠNÉ LINGVISTICKÉ TRADICE

Strukturu korpusů ovlivňují výrazně i různé lingvistické tradice a historické a pragmatické důvody při vytváření lingvistické anotace. Tyto vlivy jsou často provázány se strukturou jazyka, pravopisem, ale i školní praxí daného jazykového areálu (viz lemmata negovaných tvarů v části 3). V řadě případů jsou tyto vlivy očividné pro lingvisty vycházející z dané tradice, ale velmi obtížně odhalitelné pro lingvisty vycházející z jiné tradice. Tradiční přístup v rámci jednoho jazyka se projevuje na různých navzájem propojených rovinách: v terminologii, morfosyntaktické anotaci, v různých ojedinělých jevech a pragmatických rozhodnutích.

4.1. TERMINOLOGIE A MORFOSYNTAKTICKÁ ANOTACE

Tabulka 2 ukazuje frekvence výskytu sloves v paralelním korpusu. Ruská slovesa mají jednoznačně vyšší frekvenci.[12]

české lemma	**frekvence**	**ruské lemma**	**frekvence**
účinkovat	467	действовать	13 265
účastnit	2 217	участвовать	7 812
rozvíjet	852	развиваться	2 839

TABULKA 2. Frekvence výskytu slovesných lemmat

Důvodem je odlišné složení slovesného lemmatu v ruštině a češtině. Ruské lemma zahrnuje všechny neurčité tvary slovesné, tedy i participia s adjektivními koncovkami (*причастия*). Konkrétně lemma *развиваться* zahrnuje 19 tvarů, nejvyšší frekvenci mají tvary *развивающиеся, развиваться, развивается, развивающимся*. Lemma slovesa *rozvíjet* má 27 tvarů, protože zahrnuje kladné i záporné tvary (např. *nerozvíjejí*). Zahrnuje ovšem také frekvence zvratných i nezvratných tvarů, nicméně nezahrnuje participia (např. *rozvíjející* (*se*)). Pokud tedy srovnáváme frekvenci těchto sloves, dostáváme nesrovnatelná data, která jsou ovlivněna různým pojetím neurčitých tvarů slovesných a různou lemmatizací reflexivních tvarů. Zařazení participií mezi adjektiva je v rámci slovanských jazyků spíše výjimkou: například slovenština a slovinština mají participia značkována jako samostatnou kategorii (i v ČNK), v ruštině, polštině jsou přímo součástí slovesného lemmatu.

12 Frekvence výskytu je vyšší i v případě slovesa *участвовать* a *развиваться,* kde bychom na základě zjištění uvedených v části 2 očekávali naopak nižší výskyt, protože frekvence českého ekvivalentu zahrnuje i zvratné tvary (*účastnit, účastnit se*; *rozvíjet, rozvíjet se*).

4.1.1. POJETÍ VERBÁLNÍHO LEMMATU

Na příkladu participií je možné demonstrovat problematiku srovnávání dat u komplexnějších dotazů. Systém participií v ruštině a češtině je z velké části symetrický, ale v některých bodech i odlišný: symetrická je zejm. existence aktivního participia přítomného (typu ru. *делающий*, č. *dělající*)[13] a pasivního participia minulého (typu ru. *деланный, сделанный*, č. *dělaný, udělaný*). Rozdíly se týkají zejména imperfektivních pasivních participií (ru. *делаемый* nemá v češtině analogon, naopak funkce a frekvence českého typu *dělaný* značně převyšují ruský okrajový typ *деланный*) a minulých aktivních participií (ru. *делавший, сделавший* je téměř neomezeně produktivní a má značnou frekvenci v textu, č. *udělavší* je sice z hlediska typů výrazně produktivní, má však velmi nízkou frekvenci, a český typ *přišlý, zežloutlý* nemá v ruštině vůbec analogon).[14]

Zároveň chápe ruská a česká lingvistická tradice tyto tvary různě: zatímco v ruské tradici se o nich obyčejně pojednává v rámci tvarosloví (*делающий, делаемый, делавший, сделавший* a *деланный, сделанный* tedy patří vesměs ke slovesům *делать* a *сделать*), tak v české tradici jsou tyto derivace součástí slovotvorby.[15] Tyto tradice se odrážejí i v korpusech, jak bylo uvedeno výš: konkordance všech tvarů slovesného lemmatu v ruském korpusu obsahuje i participiální tvary, konkordance vzniklá hledáním slovesného lemmatu v českém korpusu však nikoliv, s výjimkou tvarů s jmennými koncovkami.

Tyto různé interpretace mají však dopad na další gramatické kategorie slovesa, s nimiž jsou participia spjata. Týká se to zejm. kategorie slovesného vidu: nejen, že česká participia (kromě pasivního participia minulého se jmenným tvarem) nejsou součástí odpovídajících lemmat, ale protože jsou považována za derivovaná adjektiva, také nejsou tagována s ohledem na vid. Tím vznikají při hledání imperfektivních a perfektivních slovesných tvarů pro ruštinu a češtinu konkordance s různým obsahem, který je jen částečně dán různými jazykovými systémy, částečně však různou lingvistickou tradicí a z ní pramenícím tagováním:

13 Srov. k některým kvantitativním a funkčním aspektům Giger (2020).

14 Srov. Giger (2010) k českému aktivnímu minulému participiu na *-vší*, Giger (2015) k českému aktivnímu minulému participiu na *-lý* a Giger (2021) k českému participiálnímu systému souhrnně.

15 S výjimkou jmenných tvarů typu *dělán* a *udělán*, které jsou považovány spolu s *l*-ovým tvarem typu *dělal, udělal* jako jediné v bohemistické tradici za **příčestí**. Příčestí je tedy v české bohemistické tradici tvar, který nemůže stát v atributivní pozici, tedy zcela odlišně od ruského pojmu **причастие**, který označuje tvar stojící primárně v atributivní pozici, a jen některé tyto tvary mohou stát i v pozici predikativní. Srov. Izotov (1993, zejm. s. 9–10). K českým participiím v rámci slovotvorby srov. např. MČ 1 (1986: 321–326), VAGSČ 1 (2018: 836–855).

tvary	ruština	čeština
tvary imperfektivního vidu celkově	12 553 890	21 549 762[16]
aktivní participium přítomné typu *делающий / dělající*	86 517	*139 584*[17]
pasivní participium přítomné typu *делаемый*	14 290	—
aktivní participium minulé na *-(в)ший / -(v)ší*	10 562	—[18]
pasivní participium minulé ipf. vidu (jmenné tvary) typu *делан / dělán*	1 975	36 733
pasivní participium minulé ipf. vidu (složené tvary) typu *деланный / dělaný*	1 388	?[19]
tvary perfektivního vidu celkově	9 853 353	8 826 745
aktivní participium minulé na *-(в)ший / -(v)ší*	27 888	*1 320*
pasivní participium minulé (jmenné tvary)	374 356	256 673
pasivní participium minulé (složené tvary)	124 227	?[20]

TABULKA 3. Frekvence vidových a participiálních tvarů v InterCorp v16 (údaje v *kurzivě* nejsou zahrnuty do celkových počtů vidových tvarů, protože nejsou tagovány s ohledem na vid)

K tomu se přidává ještě několik dalších důležitých dílčích prvků: tak hrají např. pro český participiální systém nemalou úlohu aktivní participia minulá na *-lý* jako *přišlý, vzniklý, zmrzlý* (srov. Giger 2010; 2015; 2021, s. 340–343). Podobně jako pasivní participium minulé v korpusu SYN nemají vlastní tag. Z hlediska vidu jsou sice vesměs perfektivní, ale frekvence lze zjistit jenom pro jednotlivá participia (v ČNK vždy samostatná lemmata): např. *vzniklý* 364, *zemřelý* 449, *padlý* 887 (InterCorp v16). Co se týče pasivních participií minulých v češtině, tak ani z poměru mezi imperfektivními a perfektivními participii se jmennými tvary nelze jednoznačně usoudit na poměr

16 Výrazně větší počet imperfektivních tvarů v češtině souvisí primárně s tím, že čeština má řadu imperfektivních modálních sloves, která ruština nemá (*muset, mít, smět*), nemá některá perfektivní modální slovesa a slovesa pohybu, která ruština má (*смочь, пойти, поехать*), ale hlavně má přítomné tvary slovesa *být*, které jsou v ruštině nulové, a to jako sponového a existenčního slovesa stejně jako slovesa pomocného v préteritu a kondicionálu (*dělal jsem, dělal bych*), které v ČNK jsou tagovány zvlášť a jako imperfektivní. Srov. k některým aspektům Bláha (2020), zejm. s. 248–249.

17 Všechny tvary jsou imperfektivní — perfektivní aktivní participia přítomná jsou velmi vzácná a automatický anotátor je nerozeznává (srov. Štícha 2008, Giger — Kocková, 2024).

18 Všechny tvary jsou perfektivní — imperfektivní aktivní participia minulá na *-(v)ší* jsou velmi vzácná a automatický anotátor je nerozeznává (srov. Giger — Kocková, 2024).

19 Pasivní participium minulé se složenými tvary nelze v InterCorp v16 vyhledat (je tam anotováno jako běžné adjektivum). Tzv. verbtag zavedený od verze SYN v2020 umožňuje tyto tvary vyhledat a udává přitom velmi vysokou frekvenci (i.p.m. 8 529,53), výrazně vyšší než všechny ostatní zde diskutované participiální tvary dohromady. Tvary ale nejsou anotovány s ohledem na vid, čili nelze je vyhledat podle vidu.

20 Viz pozn. 19.

mezi oběma vidy u odpovídajících participií se složenými tvary, srov. *psán* 608 / *napsán* 2 010 (1 : 3,3), *psaný* 827 / *napsaný* 1 936 (1 : 2,3), *zadržován* 95 / *zadržen* 590 (1 : 6,2), *zadržovaný* 83 / *zadržený* 373 (1 : 4,5), *zakládán* 2 / *založen* 1 499 (1 : 749,5), *zakládaný* 3 / *založený* 2 814 (1 : 938) atd.[21] (InterCorp v16).

4.1.2. OBOUVIDOVÁ SLOVESA

I obouvidová slovesa se vyznačují v češtině a ruštině paralelami a rozdíly: paralelní je princip, že se sémanticky vhodná přejatá slovesa často integrují nejdříve jako biaspektuální, např. č. *inscenovat*, ru. *инсценировать*. Časem k nim — pokud to odpovídá jejich sémantice — vznikají příznakově perfektivní prefigované deriváty, srov. č. *zinscenovat*, ru. *синсценировать*. Proces je přitom relativně složitý, původně biaspektuální sloveso zůstává leckdy i nadále biaspektuální, ačkoliv už existuje perfektivní prefigovaný derivát, nový derivát bývá někdy považován za nespisovný, někdy se používá jenom v některých funkcích či dílčích významech základního původně biaspektuálního slovesa, někdy vznikají konkurenční tvary s různými prefixy atd. (srov. pro ruštinu např. Черткова и Чанг 1998).

Vedle tohoto typu existuje typ ruských přejatých sloves s přízvukem na sufixu *ова-* (*организовать*), která jsou po původní integraci jako biaspektuální časem interpretována jako perfektivní a vznikají k nim sufigované příznakově imperfektivní deriváty (*организовывать*) (Jászay, 1999; Glovinskaja, 2010, s. 191). Tento typ čeština nezná; jedná se tedy opět o systémový rozdíl mezi oběma jazyky.[22] Kromě toho však pozorujeme mezi ruskou a českou částí InterCorp v16 i rozdíl v tagování: česká část používá značku pro biaspektuální slovesa (124 685 výskytů), ruská část danou značku (ačkoliv je v manuálu uvedena)[23] reálně nepoužívá (o dokladů, všechny jednotlivé výskyty biaspektuálních sloves jsou přiřazeny buď k imperfektivnímu, nebo perfektivnímu vidu). Když srovnáme počty imperfektivních a perfektivních tvarů z tabulky 3, tak by bylo třeba v české části korpusu skupinu biaspektuálních sloves uvést zvlášť (neobjeví se ani mezi imperfektivními ani mezi perfektivními slovesy), v ruské části se naopak ztrácí, je rozdělená mezi imperfektivními a perfektivními slovesy a nelze ji vyhledat automaticky. Na druhé straně je ve dvojici dokladů *Газеты стали тогда печатать целые циклы статей и организовывать письма читателей / Noviny začaly tehdy otiskovat cykly článků a organizovat dopisy čtenářů* ruské sloveso *организоывать* příznakově imperfektivní, zatímco české *organizovat* se mezi imperfektivy neobjeví (má značku ‚biaspektuální'). Kontext (fázová slovesa *стать*, resp. *začít*) přitom v obou jazycích připouští jen imperfektivní vid.

21 Důvody mohou souviset s různými funkcemi jmenných a složených tvarů, zejm. v dějovém pasivu, ale i se stylistickými faktory. Obojí se nemusí u každého slovesa tvořícího pasivní participium projevit stejně.

22 Počty výskytů těchto sloves jsou v InterCorp v16 většinou skromné: *арестовывать* 354, *организовывать* 224, *образовывать* 55, *основывать* 44, *согласовывать* 33, *реализовывать* 27; všechna další odpovídající slovesa mají jen jeden nebo dva doklady.

23 https://nl.ijs.si/ME/Vault/V4/msd/html/msd-ru.html

4.2. ODLIŠNÁ TOKENIZACE

Hranice slov a definice samostatného slova jsou specifické pro jednotlivé jazyky. Definice slova se sekundárně promítá do tokenizace daného jazyka a ovlivňuje různým způsobem vyhledávání i frekvenční data — v tomto případě je ovlivněna i absolutní i relativní frekvence výskytu, protože definice slova ovlivňuje i celkový počet tokenů v korpusu. Tuto problematiku lze demonstrovat na kompozitech. Způsob psaní kompozit je většinou výsledkem úzu, liší se mezi jazyky, ale často není jednotný ani u různých typů kompozit v rámci jednoho jazyka (srov. č. *dechberoucí, Sazka aréna, učení se*), ani u stejných typů kompozit v rámci jednoho jazyka (ru. *видео-игра, видеоигра*).[24]

Způsob pravopisu, respektive tokenizace se sekundárně promítá do absolutní frekvence: srov. frekvence lemmatu v paralelním korpusu: č. *průmyslový* 9 470, n. *industriell* 2 459; č. *letní* 1 572, n. *sommerlich* 45. Rozdíly ve frekvenci tak mohou odrážet skutečné rozdíly mezi jazyky, tj. vyjádření pomocí kompozita v němčině a syntagmatu v češtině: č. *průmyslové zboží,* n. *Industriegüter*; č. *průmyslové země,* n. *Industriestaaten.* Mohou však odrážet pouze rozdílný pravopis: n. *Video* 3 426, ang. *video* 5 618 (*Videospiele, video games*); n. *rot*, ang. *red* (*Rotwein, red wine; rothaarig, red-haired*).

4.3. ODLIŠNOSTI VZNIKLÉ V PRŮBĚHU ROZVOJE KORPUSŮ

S prodlužující se dobou existence korpusů a stále komplexnějším objemem i rozsahem nabízených funkcí stoupá i množství kroků, které nebyly provedeny zcela předvídatelně, nebo jsou nesystémové. Tyto kroky jsou často odhalitelné jen náhodně. Uvedeme dva příklady. Tagování ČNK vychází z logiky, že verbální adjektiva (participia) se tvoří od přechodníku, minulé tvary aktivní se tvoří od přechodníku minulého, který je definován jako dokonavý tvar. Z toho plyne, že „minulá verbální adjektiva" musí být dokonavá. Imperfektivní tvary jsou označeny jako neznámé (viz pozn. 18). Nicméně v korpusu najdeme i tvar imperfektivní *tvořivší*, tagovaný jako „minulé verbální adjektivum". Pravděpodobně jde o tvar zadaný ručně. Druhý příklad má širší dopad. Zadáme-li do ruské části InterCorpu například dotaz: [lemma="Европа"], zjistíme, že má v korpusu 0 výskytů. Je to způsobeno tím, že všechna lemmata, včetně proprií se v ruské části paralelního korpusu píší s malým písmenem.[25] V InterCorpu v16ud se značkováním pomocí UD jsou už i odpovídající ruská lemmata psaná velkým písmenem.

5. ZÁVĚR

Korpus je bezpochyby nedílnou součástí moderního lingvistického výzkumu. Při využití korpusových dat je však nutné zohlednit otázku srovnatelnosti dat a jejich

24 Příklady jsou převzaty z paralelního korpusu InterCorp v16.

25 NKRJa zahrnuje do jednoho paradigmatu tvary psané velkým i malým písmenem *(лабрадор, Лабрадор)*. Lemmata psaná výhradně malým písmem jsou výjimkou i v rámci InterCorpu: lemma psané pouze malým písmenem má jen ruština a portugalština.

validity pro danou problematiku. Korpusová data jsou ovlivněna nejen odlišnou strukturou jazyků, ale i odlišnou strukturou korpusů a morfosyntaktické anotace, lingvistickými tradicemi a dalšími faktory. Reálné odchylky dat sice často nejsou tak velké, aby výrazně ovlivnily celkový výsledek (například počty participiálních tvarů nezahrnutých do českých slovesných paradigmat, nebo počty výskytů českých biaspektuálních sloves výrazně neovlivní celkové poměry vidových tvarů při srovnání mezi češtinou a ruštinou, srov. např. Bláha 2020). Na druhé straně je zřetelné, že v mnoha ohledech jsou tyto počty nesrovnatelné, protože se zakládají na různém tagování a v důsledku toho obsahují různá dílčí paradigmata.

Práce s korpusy vyžaduje kritický přístup k datům a ověřování výsledku zadaného dotazu. Zároveň se s rozvíjející se strukturou korpusů ukazuje i důležitost metainformací, jak aktuálních, tak i těch, které mapují vývoj korpusů v minulosti.

PRIMÁRNÍ ZDROJE

Aranea Corpora (2020): Jazykovedný ústav Ľudovíta Štúra SAV. http://unesco.uniba.sk/.

BNC: The British National Corpus, version 2 (BNC World) (2001). http://www.natcorp.ox.ac.uk/.

DeReKo: Deutsches Referenzkorpus Mannheim IDS (2022): Leibniz-Institut für Deutsche Sprache. https://cosmas2.ids-mannheim.de/.

Grac v.18: General Regionally Annotated Corpus of Ukrainian, (2017). uacorpus.org.

InterCorp v16: Institut českého národního korpusu FF UK (2022). Praha. http://www.korpus.cz.

Kernkorpus DWDS: Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache, 20. Jahrhundert. (2018): www.dwds.de.

MASC: Manually Annotated Sub-Corpus, The Open American National Corpus. https://anc.org/data/masc/.

NKRJa: The Russian National Corpus (ruscorpora.ru). 2003—2023. ruscorpora.ru.

SNK: Slovenský národný korpus — Bratislava: Jazykovedný ústav Ľ. Štúra SAV. (2022). https://korpus.sk.

SYN v2020: Institut českého národního korpusu FF UK (2020). Prague. http://www.korpus.cz.

LITERATURA

Bláha, O. (2020): Typologické aspekty ruského a českého vidového systému. *Slavia*, 79, s. 245–255.

Dąbrowska, A. (2013): *'National Corpus of Polish' and 'Great Dictionary of Polish': two leading projects of present-day Polish lexicography*. Konferenční příspěvek. http://efnil.nytud.hu/documents/conference-publications/budapest-2012/15-EFNIL-Budapest-Dabrowska-Final.pdf.

Geyken, A. (2007): The DWDS Corpus: a Reference Corpus for the German Language of the 20th Century. In: C. Fellbaum (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press, s. 23–41.

Giger, M. (2010): Příčestí minulé činné na *-(v)ší* v dnešních českých publicistických textech. *Korpus — Gramatika — Axiologie*,1, 2, s. 3–23.

Giger, M. (2015): Subjektová rezultativa v češtině ve srovnání s ruštinou. *Časopis pro moderní filologii*, 97, s. 146–156.

Giger, M. (2020): Několik poznámek k příčestí přítomnému činnému v ruštině a češtině. In: J. Bílková J. — I. Kolářová — M. Vondráček (eds.), *Lingvistika — Korpus — Empirie*. Praha: Ústav pro jazyk český, s. 9–16.

Giger, M. (2021): C. Participia a predikativa. In: F. Štícha a kol., *Velká akademická gramatika spisovné češtiny. II. Morfologie: Morfologické kategorie / Flexe. Část 1*. Praha: Academia, s. 331–352.

Giger, M. — Kocková, J. (2024): Grenzüberschreitungen an der Peripherie: Aspektuelle Funktionen von Aktivpartizipien und Verbalsubstantiven im Tschechischen. *Zeitschrift für Slawistik*, 69, s. 1–26.

Kocková, J. — Sytar, H. (2024): The Most Frequent Lemmas in the Ukrainian and Czech Corpus as a Resource for Foreign Language Learning and Teaching. *Зборник Матице српске за славистику*, 105, s. 369–382.

Křen, M. (2012): *Diachronní srovnání synchronních korpusů*. PhD. dis., FF UK, Praha.

MČ 1. (1986): *Mluvnice češtiny 1. Fonetika, fonologie, morfonologie a morfemika, tvoření slov*. Pod red. M. Dokulila a kol. Praha: Academia.

Štícha, F. (2008): Uzuálnost, funkčnost a systémovost jako kritéria gramatičnosti. K jednomu typu morfologické derivace *(udělajíc — udělající)*. *Slovo a slovesnost*, 69, s. 176–191.

VAGSČ 1. (2018): F. Štícha et al. *Velká akademická gramatika spisovné češtiny I. Morfologie: druhy slov, tvoření slov*. Praha: Academia.

Siepmann, D. — Bürgel, C. — Diwersy, S. (2015): The Corpus de référence du français contemporain (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography*, 30, s. 63–84.

Stefanowitsch, A. (2020): *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press (Language Sciences 7).

Гловинская, М. Я. (2010): Потенциальные глагольные формы. In: Л. П. Крысин, (ред.), *Современный русский язык. Система — норма — узус*. Москва: Языки славянских культур, с. 171–199.

Изотов, А. И. (1993): *Чешские атрибутивные причастия на фоне русских*. Москва: МГУ, Филологический Факультет.

Черткова, М. Ю. — Чанг, П.-Ч. (1998): Эволюция двувидовых глаголов в современном русском языке. *Russian Linguistics*, 22, s. 13–34.

Markus Giger | Ústav východoevropských studií, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 1/2, 116 38 Praha 1
ORCID ID: 0000-0002-9074-7161
markus.giger@ff.cuni.cz

Jana Kocková | Oddělení slavistické lingvistiky a lexikografie, Slovanský ústav Akademie věd České republiky, v. v. i. | Valentinská 1, 110 00 Praha 1
ORCID ID: 0000-0003-0813-089X
kockova.jana@gmail.com