
ERRKORP

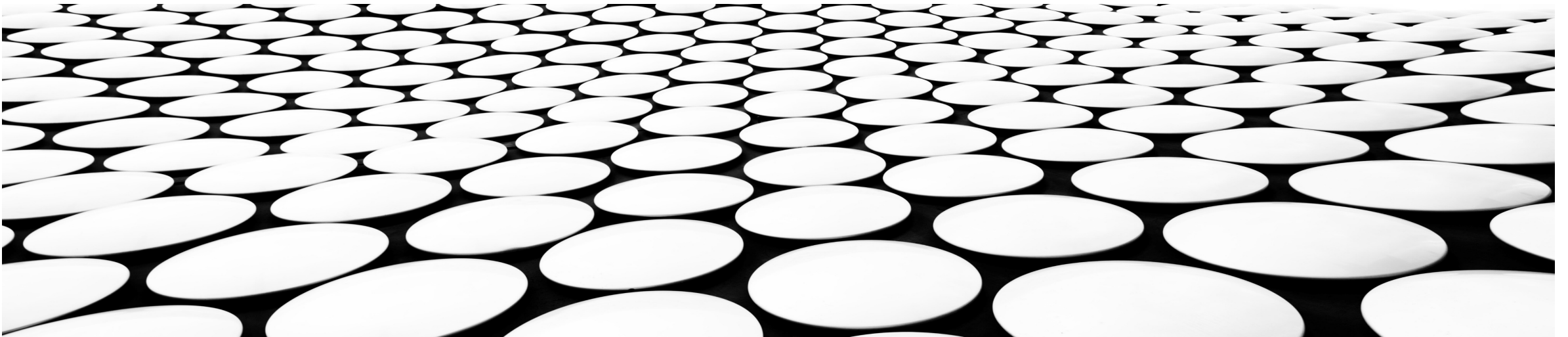
AKVIZIČNÝ KORPUS PÍSANÝCH TEXTOV CUDZINCŮV UČIACICH SA SLOVENČINU AKO CUDZÍ JAZYK

PROJEKT APVV-19-0155 „JAZYKOVÉ CHYBY V SLOVENČINE AKO CUDZOM JAZYKU NA BÁZE AKVIZIČNÉHO KORPUSU“

MGR. MICHAELA MOŠAŤOVÁ, PHD.

MGR. HELENA ĽOS IVORÍKOVÁ

STUDIA ACADEMICA SLOVACA, FILOZOFICKÁ FAKULTA UNIVERZITY KOMENSKÉHO V BRATISLAVE



RIEŠITEĽSKÝ KOLEKTÍV PROJEKTU

Riešiteľský kolektív sa skladá:

- z pracovníčok oddelenia [Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v Bratislave](#) (K. Gajdošová, J. Levická, J. Mášik, K. Rausová; do roku 2020 aj L. Klimová, M. Šimková)
- z pracovníčok a spolupracovníčok [Studia Academica Slovaca Filozofickej fakulty Univerzity Komenského v Bratislave](#) (J. Pekarovičová, M. Mošaťová, H. Ľos Ivoríková, P. Kollárová) a kolegýň z partnerských inštitúcií
- [Filozofickej fakulty Prešovskej univerzity v Prešove](#) (M. Ivanová, M. Kyseľová, M. Imrichová)
- a [Filozofickej fakulty Univerzity Mateja Bela v Banskej Bystrici](#) (A. Gáľisová, L. Urbancová).

(Zdroj: <https://korpus.sk/errkorp.html>)

PILOTNÁ VERZIA (errkorp.pilot)

- **zverejnená od augusta 2022** na: <https://bonito.korpus.sk/>
- má veľkosť **137 704** textových jednotiek, **507** zdrojových textov,
- plne lematizovaná aj morfológicky anotovaná
 - vonkajšia anotácia textov (informácie o študentovi a texte)
 - vnútorná anotácia textov (1. chyby na úrovni jedného segmentu v rámci jedného slova, 2. chyby na úrovni jednej morfémy alebo jedného slova, 3. chyby na úrovni slovného spojenia v rámci jednej vety, 4. chyby na úrovni textu)
- obsahuje ručnú anotáciu chýb aj s ich opravami,
- viac ako **50 osobitných značiek** podľa vypracovaného anotačného tagsetu, dostupného na stránke projektu (https://korpus.sk/errkorp_inannot.html).

(Zdroj: <https://korpus.sk/errkorp.html>)

PILOTNÁ VERZIA

- Pohlavie: žena (395), muž (112)
- Vek študenta: 10+ (70), 20+ (357), 30+ (31), 40+ (17), 50+ (7), 60+ (6), 70+ (1), neznáme (18)
- Krajina, z ktorej študent pochádza (spolu 37 krajín): Belgicko (2), Bielorusko (11), Bosna a Hercegovina (3), Brazília (1), **Bulharsko (48)**, **Čína (35)**, Fínsko (1), Francúzsko (15), Grécko (1), Gruzínsko (4), Holandsko (3), Chorvátsko (21), Irán (2), Izrael (3), Japonsko (5), Kanada (1), **Kórejská republika (38)**, Litva (2), Lotyšsko (3), Maďarsko (1), Nemecko (21), Nórsko (3), **Poľsko (47)**, Portugalsko (4), Rakúsko (8), Rumunsko (21), **Rusko (41)**, Slovinsko (15), Spojené kráľovstvo (1), Spojené štáty (1), Spojené štáty (1), **Srbsko (32)**, Španielsko (12), Švajčiarsko (2), Švédsko (4), Taliansko (23), Turecko (7), **Ukrajina (65)**
- Jazyková úroveň študenta v slovenskom jazyku: A1.1 (50), A1.2 (63), A2.1 (33), A2.2 (46), B1 (126), B2 (143), C1 (38), C2 (8) ____ **A (192), B (269), C (46)**

(Zdroj: Ročná správa o riešení projektu za rok 2021)

PRVÁ VERZIA

- 1.063 zdrojových textov, 349 000 tokenov
- zverejnená bude na jar 2023
- Pohlavie študenta:
 - žena (2151), muž (786)
- Vek študenta:
 - 10+ (680), 20+ (1959), 30 + (118), 40+ (90), 50+ (27), 60+ (8), 70+ (2), neznáme (53)
- Materinský jazyk študenta (43):
 - albánčina (16), angličtina (14), arabčina (148), azerbajdžančina (17), bieloruština (31), bosniančina (6), **bulharčina (435)**, čeština (1), čiernohorčina (1), **čínština (269)**, fínčina (3), francúzština (24), gréčtina (3), gruzínčina (35), hebrejčina (9), holandčina (3), chorvátčina (86), japončina (5), kazaština (3), kórejčina (56), litovčina (2), **maďarčina (169)**, moldavčina (11), mongolčina (6), nemčina (66), nórčina (3), pašťčina (19), perzština (77), **poľština (360)**, portugalčina (5), rumunčina (97), **ruština (322)**, slovinčina (21), srbčina (55), svahilčina (70), španielčina (19), švédčina (4), taliančina (30), tatárčina (3), turečtina (14), **ukrajinčina (365)**, uzbečtina (12), vietnamčina (42)
- Jazyková úroveň študenta v slovenskom jazyku:
 - A1.1 (121), **A1.2 (606)**, A2.1 (339), A2.2 (284), **B1 (571)**, **B2 (674)**, C1 (321), C2 (21)

(Zdroj: Ročná správa o riešení projektu za rok 2022)

Slovenský národný korpus

errkorp-pilot Popis: errkorp-pilot Corpus of Texts of Students Learning Slovak as a Foreign Language Velkost: 137 704 pozícií Výskyty: 1 986

Hľadané **substword** 1 986 (14 422,24 na milión) | ARF: 1 145,44

Strana 1 z 100 [Prejsť](#) [Ďalšia](#) | [Posledná](#)

2015-03-12-Srbirf	<err type="comma1"> , </err> a želanie <err type="word1_dep"> o </err><err type="word1"> končenom </err><err type="substword"> riešení
2015-03-12-Srbirf	je momentálne najväčšou prekážkou , ale dúfame v <err type="substword"> otváranie
2015-03-12-Srbirf	má veľký vplyv na kapitál v SR . <err type="substword"> Kludne
2015-03-12-Srbirf	<err type="char1 char1"> OEBSom OEBS </err> . Zahraničná politika SR sa v <err type="substword"> minulých
2015-03-12-Srbirf	. Nikto sa nesmie vzpierať proti vláde , <err type="substword"> pretože
2019-07-02-Zhejke	spolu s priateľom </err> . <err type="substword_dep"> V Na </err><err type="dep"> jare jar </err><err type="substword"> sa
2018-11-00-Hrjidu	, ako a čo jeme a žijeme </title> <err type="substword"> Ešte
2018-11-00-Hrjidu	boli <err type="quantsuf0"> používané používané </err> na liečbu , krásu a <err type="substword"> údržbu
2018-11-00-Hrjidu	používané </err> na liečbu , krásu a <err type="substword"> údržbu zachovanie </err><err type="substword"> mládeže
2018-11-00-Hrjidu	<err type="quantsuf0"> myslíme myslíme </err> lepšie , <err type="quantbase0"> citíme cítime </err> sa lepšie <err type="substword"> ak
2018-11-00-Hrjidu	sebavedomie . Dobré vieme , že človek s <err type="substword"> kontrolovaným
2018-11-00-Hrjidu	vysokým zdravým veľkým </err> sebavedomím <err type="quantsuf0"> vyzerá vyzerá </err><err type="dep"> atraktívnejšia atraktívnejšie </err> a dostupnejšie <err type="substword"> na
2018-11-00-Hrjidu	zdravý život pre zdravý život je </err><err type="seccongr quantsuf0"> dôležité dôležitý </err><err type="substword"> pohybovanie
2018-11-00-Hrjidu	, </err> musí sa pohybovať . Z <err type="morph"> pohyby pohybu </err><err type="substword"> nedostávame
2018-11-00-Hrjidu	životný štýl je štýl na celý život . <err type="styl"><err type="substword"> Kvôli
2018-05-20-Hrdymi	Myslím si <err type="comma0"> , </err> že jeho význam je <err type="substword"> blízko
2018-05-20-Hrdymi	od <err type="quantbase1"> malička malička </err><err type="caron0 quantsuf1"> učíť učiť </err> deti pracovať a <err type="substword"> vedieť
2018-05-20-Hrdymi	samozrejme <err type="comma0"> , </err> sú aj iné prípady , <err type="substword"> kedy
2019-03-00-Roesgo	ktorá sa ku mne pripojí , ktorá ma <err type="substword"> dokončí
2016-05-31-Sridnu	</err> a prečo niektorí na dedine ? To <err type="substword"> si

Strana 1 z 100 [Prejsť](#) [Ďalšia](#) | [Posledná](#)

Hľadané **peniaz** 82 (595,48 na milión) | ARF: 29,75 

Strana z 5 [Prejsť](#) [Ďalšia](#) | [Posledná](#)

pe="order"> **stále mu** | **mu stále** </err> dávajú **peniaze** . Pravdepodobne to dieťa vyrastie , stane sa
môžeme prirovnať k zločinu (najmä keď tie **peniaze** <err type="quantsuf0"> **mame** | **máme** </err>) . Stáva sa <err ty
="comma0"> | , </err> že dobrovoľníci zbierajú **peniaze** pre smrteľne chorých ľudí . Vtedy <err type="order"> **radšej dať**
eň . </doc><doc> <title> <err type="phrase"> **Či** **peniaze** **kazia charakter** | **Kazia peniaze charakter** </err> ? </title> Ná
:"phrase"> **Či peniaze kazia charakter** | **Kazia** **peniaze** **charakter** </err> ? </title> náš charakter sa formuje celý
ch <err type="quantsuf0"> **veci** | **vecí** </err> sú **peniaze** . Ako <err type="quantbase0"> **vplyvajú** | **vplývajú** </err> na spr
na dve najväčšie skupiny - takých , ktorým **peniaze** " <err type="char1|congr"> **stúpily** | **stúpili** </err> do hlavy " a tal
stúpili </err> do hlavy " a takých , ktorí **peniaze** považujú iba za uľahčenie života . Pre tých
stword_dep"> **do** | **v** </err><err type="secdep"> **peniazi** | **peniazoch** </err> **sa** | **záľuba v peniazoch sa často** </err> spá
p"> **do** | **v** </err><err type="secdep"> **peniazi** | **peniazoch** </err> **sa** | **záľuba v peniazoch sa často** </err> spája aj
ep"> **peniazi** | **peniazoch** </err> **sa** | **záľuba v peniazoch sa často** </err> spája aj s workoholizmom) .
e="quantsuf1"> **Druhý** | **Druhý** </err> skupine **peniaze** iba dovoľujú niektoré veci , napríklad <err type="quantbase0"> **ro**
type="quantsuf1"> **vyzeráť** | **vyzerat'** </err> . **Peniaze** by nemali byť prvoradé , mali by byť
<err type="quantbase0"> **služia** | **slúžia** </err> **peniazom** " <err type="comma0"> | , </err> neexistujú . </doc><doc> <title
z <err type="word0"> | **je** </err> ľahšie zarobiť **peniaze** a dosiahnuť úspech a <err type="quantbase0|quantsuf0"> **sučasne**
vne na internet a väčšina firiem už neinvestuje **peniaze** do iných zdrojov reklamy . Z jednej strany
áve zarobili na tom čo najviac <err type="dep"> **peniaze** | **peňazi** </err> . Preto aj keď je umiestnenie reklamy
</err> , ale mal som <err type="dep"> **peniazi** | **peniaze** </err> . Tak neboli <err type="quantbase0|morph"> **problema** | ,
sh"> **členu** | **člena** </err> jednej organizácie za **peniaze** alebo <err type="phrase"> **nejakú inú priazeň iného jednotlivca**
ord1"> **kamkoľvek** | </err> budete potrebovať **peniaze** . <err type="substword"> **Keď** | **Keďže** </err><err type="word1">

ANOTAČNÝ TAGSET (VÝBER)

Chyby na úrovni jedného segmentu v rámci jedného slova

Značka	Vysvetlenie	Priklady
cap0	chýbajúce veľké písmeno	ministerstvo financií SR (Ministerstvo financií SR)
cap1	redundantné veľké písmeno	Minister financií (minister financií), Európska Únia (Európska únia)
caron0	chýbajúci mäkčeň	student (študent), vecera (večera)
caron1	redundantný mäkčeň	več (vec), d'esať (desať), d'ieťa (dieťa), nechaj (nechaj)
comma0	chýbajúca čiarka	Ahoj Eva (Ahoj, Eva); Nebol tu tak neviem (Nebol tu, tak neviem)
comma1	redundantná čiarka	V prípade zlého počasia, pôjdeme do múzea (V prípade zlého počasia pôjdeme do múzea)
defdiacr	tri a viac chýb v diakritike (napr. „čtetové písanie“)	paťnast (pätnást'), dolezity (dôležitý)
defword	nezrozumiteľné slovo (tri a viac zámen grafém v slove s výnimkou chybných diakritiky) alebo slovo v inom jazyku	samuslina (zmrzlina), bybarený (vybavený), uniwersytet (univerzita), Rakousko (Rakúsko), dort (torta), těžký (ťažký)
dot0	chýbajúca bodka	13 januára 2016 (13. januára 2016)
dot1	redundantná bodka	v roku 2014. tu postavili budovu (v roku 2014 tu postavili budovu)
char0	chýbajúca graféma v slove	prtože (pretože), s študentom (so študentom)
char1	redundantná graféma v slove	súčastne (súčasne), pekingský (pekinský)
charmata	zámena grafém v slove	Sbrsko (Srbsko), horovím (hovorím)
hyph	chyba pri zápise spojovníka (chýbajúci alebo redundantný spojovník)	československý slovník (česko-slovenský slovník), rímsko-katolícky (rímskokatolícky)
quantbase0	chýbajúca kvantita v tvarotvornom základe slova alebo v neohybnom slove	krasny (krásny), stretavali (stretávali), vstavali (vstávali), prinos (prínos), vystavba (výstavba), krajín (krajín), miliárd (miliárd), devätnasť (devätnásť)
quantbase1	redundantná kvantita v tvarotvornom základe slova alebo v neohybnom slove	výsoký (vysoký), hrád (hrad), prénos (prenos), výhodiť (vyhodiť), po nemecký (po nemecky)

ANOTAČNÝ TAGSET (VÝBER)

Chyby na úrovni slovného spojenia v rámci jednej vety

Značka	Vysvetlenie	Príklady
congr	chyba v zhode (kongruencii)	ten centrum (to centrum), zaujímavý esej (zaujímavá esej), veľa ľudí boli (veľa ľudí bolo), rámy je vodotesný (rámy sú vodotesné), povedala som, aby sa vrátil (povedala som, aby sa vrátila)
dep	chyba v dependencii (v syntaktickej závislosti inej ako zhoda)	hľadám sestra (hľadám sestru), káva s mlieko (káva s mliekom), deväť študenti (deväť študentov), človek, ktorý tu čakám (človek, ktorý tu čakajú)
neg	chyba v negácii	nie budem (nebudem), nikto prišiel (nikto neprišiel), nie som bol (nebol som), ne máme (nemáme)
order	porušenie rytmického alebo gramatického princípu slovosledu	veľmi mi tam sa páčilo (veľmi sa mi tam páčilo), v texte sa nachádza veľa mien podstatných (v texte sa nachádza veľa podstatných mien)
phrase	chyba v ustálených spojeniach a vo frazémach	nemá rozprávania (nie o čom), hádzať šošovicu na stenu (hádzať hrach na stenu), ľúbim chodiť do kina (rád chodím do kina)
seccongr	sekundárna chyba v zhode (kongruencii)	idem do filozofickej fakulty (ideme na filozofickú fakultu)
secdep	sekundárna chyba v dependencii prejavujúca sa v tvare substantíva, resp. pronomína	ideme do fakulty (ideme na fakultu), idem do lekára (idem k lekárovi)
space	chýbajúca alebo redundantná medzera medzi slovami	na príklad (napríklad), Akosa voláš (Ako sa voláš)
substword_dep	zámena slova (zvyčajne prepozície alebo konjunkcie) v rámci dependencie	bývame z Bukurešti (bývame v Bukurešti), idem na kaviareň (idem do kaviarne)
word0_asp	chýbajúca gramatická morféma pri aspekte	ako sa mi po skončení školy podarí (bude dariť)
word1_asp	redundantná gramatická morféma pri aspekte (vo futúre)	budem skončiť (skončím), budem napíšem (napíšem)
word0_congr	chýbajúce slovo v rámci gramatického tvaru	Peter, kde bol? (Peter, kde si bol?), opýtal by sa ma, prečo by neprišla (opýtal sa ma, prečo by som neprišla)
word1_congr	redundantné slovo v rámci gramatického tvaru	keby sú prišli (keby prišli), oni sú plakali (oni plakali)
word0_dep	chýbajúca prepozícia alebo konjunkcia v rámci dependencie	zaujímam sa politiku (zaujímam sa o politiku), nahnevane neho pozrel (nahnevane na neho pozrel), povedala som jej vrátiť sa
word1_dep	redundantné slovo (zvyčajne prepozícia) v rámci dependencie	zelená evokuje na prírodu, zaoberať sa o módu
und	gramaticky korektná, ale nezrozumiteľná časť textu	aby sme poslali k mieru, priateľstvu a láske; mladosti je až zmenených

LITERATÚRA:

- Šebesta, K. – Bedřichová, Z. – Šormová, K. – Štindlová, B. – Hrdlička, M. – Hrdličková, T. – Hana, J. – Petkevič, V. – Jelínek, T. – Škodová, S. – Poláčková, M. – Janeš, P. – Lundáková, K. – Skoumalová, H. – Sládek, Š. – Pierscieniak, P. – Toufarová, D. – Richter, M. – Straka, M. – Rosen, A.: CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014. Praha: Ústav Českého národního korpusu FF UK, 2014. Dostupný z WWW: <http://www.korpus.cz>
- Granger, Sylviane – Dagneaux, Estelle – Meunier, Fanny – Paquot, Magali: International Corpus of Learner English v2 (Handbook + CD-Rom). Louvain-la-Neuve: Presses universitaires de Louvain, Louvain-la-Neuve 2009.
- Kotková, Radomila: Čeština nerodilých mluvčích s mateřským jazykem neslovanským. Praha: Univerzita Karlova, Nakladatelství Karolinum, 2017.
- Reznicek, Marc – Lüdeling, Anke – Krummes, Cedric – Schwantuschke, Franziska – Walter, Maik – Schmidt, Karin – Hirschmann, Hagen – Andreas, Torsten (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Dostupný z WWW: <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/view>. [2019-03-18]
- RLC. Russian Learner Corpus. Dostupné z WWW: <http://web-corpora.net/RLC/> [2019-03-18]
- Corder, Stephen Pit: Error Analysis and Interlanguage. Oxford: Oxford University Press, 1981. 120 s.
- James, Carl: Errors in Language Learning and Use. London – New York: Longman, 1998. 304 s.
- Šebesta, K. – Škodová, S. a kol. Čeština – cílový jazyk a korpusy. Liberec, 2012, s. 13 – 33.
- Dostupné z WWW: <http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-cs.pdf>
- Kilgariff, Adam – Baisa, Vít – Bušta, Jan – Jakubíček, Miloš – Kovář, Vojtěch – Michelfeit, Jan – Rychlý, Pavel – Suchomel, Vít (2014). The Sketch Engine: ten years on. Lexicography, 1: 7-36. Dostupný z WWW: <http://www.sketchengine.eu>.

K PROJEKTU A K TAGSETU:

- MOŠAŤOVÁ, Michaela – GAJDOŠOVÁ, Katarína: *Annotations in the corpus of texts of students learning slovak as a foreign language (ERRKORP)*. Jazykovedný časopis, 2019, roč. 70, č. 2, s. 345-357. (Spoluautorka K. Gajdošová)
- MOŠAŤOVÁ, Michaela: *Jazykové chyby v slovenčine ako cudzom jazyku v pripravovanom akvizičnom korpuse*. In: *Studia Academica Slovaca* 47. Bratislava: Univerzita Komenského, 2018, s. 344 – 350;
- MOŠAŤOVÁ, Michaela: *Anotovanie chýb v korpuse písaných textov študentov učiacich sa slovenčinu ako cudzí jazyk*. In: *Studia Academica Slovaca* 48. Bratislava: Univerzita Komenského, 2019, s. 322 – 338;
- MOŠAŤOVÁ, Michaela: *Hraničné prípady pri anotovaní jazykových chýb v písaných textoch študentov učiacich sa slovenčinu ako cudzí jazyk*. In: *Studia Academica Slovaca* 49. Bratislava: Univerzita Komenského, 2020, s. 113 – 124.

DOTERAJŠIE PUBLIKAČNÉ VÝSTUPY (VÝBER):

- IVANOVÁ, Martina: *Pojem gramatickosti a negramatickosti pri hodnotení slovosledu (chyby v slovoslednej realizácii enklitik v slovenčine ako cudzom jazyku)*. Slovenská reč, 2022, roč. 87, č. 1, s. 110 – 126.
- IVANOVÁ, Martina: *Pravdepodobne si myslím, že niekedy v roku 2020. Pragmatické markery v slovenčine ako v cudzom jazyku*. Studie z aplikovanej lingvistiky, 2022, roč. 13, č. 2, s. 35 – 64.
- IVANOVÁ, Martina: *Myslím, že to nikdy nebude jednoduché. Epistemické a evidenčné markery v písaných textoch slovenčiny ako L2*. In: Studia Academica Slovaca 51. Prednášky 58. letnej školy slovenského jazyka a kultúry. Bratislava: Univerzita Komenského v Bratislave, 2022, s. 164 – 184.
- GAJDOŠOVÁ, Katarína: *Čo možno nájsť v pilotnej verzii akvizičného korpusu ERRKORP?* In: Varia 30. Zborník príspevkov z 30. kolokvia mladých jazykovedcov (Bratislava 3. – 5. 11. 2021). Eds. K. Gajdošová – N. Kolenčíková. Rec. Anna Gálisová, Andrej Závodný. 1. vyd. - Bratislava: Slovenská jazykovedná spoločnosť pri Jazykovednom ústave Ľudovíta Štúra SAV, v. v. i., 2022, s. 88 – 94.
- ĽOS IVORÍKOVÁ, Helena: *Ako sa dá vymedziť relevantnosť opráv chýb cudzincov učiacich sa SakoCJ (so zameraním na chyby typu substword)?* Zborník z medzinárodnej interdisciplinárnej vedeckej konferencie pri príležitosti 45. výročia výučby slovenčiny na Sliezskej univerzite v Katoviciach *Horizonty slovakistiky: dávne inšpirácie a nové výzvy* (17. – 18. máj 2022), v tlači.
- IMRICHOVÁ, Mária: *Propriálny subsystém slovenčiny s poukazom na problémy ortografie a jeho výučba na hodinách slovenčiny ako cudzieho jazyka*. In: Súčasná slovakistika – súradnice a vektory. Ed. V. Panajotov et al. Sofia: Univerzitné vydavateľstvo „St. Kliment Ochridski“, 2022, s. 355 – 366.
- IVANOVÁ, Martina – KYSEĽOVÁ, Miroslava – GÁLISOVÁ, Anna: *Tá robota sa sama neurobí. Zvratné pasívum a zvratné deagentívum v slovenčine a jeho osvojovanie poľskými hovoriacimi*. Zborník z medzinárodnej interdisciplinárnej vedeckej konferencie pri príležitosti 45. výročia výučby slovenčiny na Sliezskej univerzite v Katoviciach *Horizonty slovakistiky: dávne inšpirácie a nové výzvy* (17. – 18. máj 2022), v tlači.
- MOŠAŤOVÁ, Michaela – KOLLÁROVÁ, Petra – ĽOS IVORÍKOVÁ, Helena: *Negácia v písaných textoch študentov učiacich sa slovenčinu ako cudzí jazyk*. In: *Slovenčina (nielen) ako cudzí jazyk III. Zborník príspevkov z medzinárodnej vedeckej konferencie konanej 1. – 2. februára 2021 na FiF UK v Bratislave*, v tlači.