e-ISSN 2280-6792

EL.LE
Vol. 11 – Num. 2 – Luglio 2022

Il corpus RU_SEAH
La lingua russa per la comunicazione specializzata nel settore dell'architettura e delle costruzioni

Maria Chiara Ferro
Università degli Studi «G. d'Annunzio», Chieti-Pescara, Italia

Abstract The compilation of corpora of specialised communication in professional fields is still little-explored for Russian language. In this context, this article presents RU_SEAH, a Russian language corpus for specialised communication in Architecture and Construction, realised in the project *Sharing European Architectural Heritage: Innovative language teaching tools for academic and professional mobility in Architecture and Construction*, developed since 2020 at the Department of Modern Languages, Literatures and Cultures of the University "G. d'Annunzio" of Chieti-Pescara.

Keywords Corpus linguistics. Russian language. Russian corpora. Architecture and construction. Specialised communication.

Sommario 1 Introduzione e scopo del saggio. – 2 I corpora di lingua russa. Breve rassegna. – 3 Il corpus RU_SEAH. – 3.1 Il progetto SEAH: finalità e metodo. – 3.2 I corpora linguistici nel progetto SEAH. – 3.3 RU_SEAH: compilazione e struttura. – 3.4 RU_SEAH: funzionamento. – 3.5 RU_SEAH: un esempio di applicazione didattica. – 4 Conclusioni e prospettive.

Edizioni Ca'Foscari

Peer review

Submitted 2022-05-21
Accepted 2022-07-15
Published 2022-07-29

Open access

Citation Ferro, M.C. (2022). "Il corpus RU_SEAH. La lingua russa per la comunicazione specializzata nel settore dell'architettura e delle costruzioni". *EL.LE*, 11(2), 245-266.

DOI 10.30687/ELLE/2280-6792/2022/02/006

1 Introduzione e scopo del saggio

La linguistica dei corpora, sviluppatasi fin dagli anni Sessanta del secolo scorso in ambito anglosassone (cf. McEnery, Wilson 2001, 22-3)[1] e definita nel 1992 addirittura una «rivoluzione» (Rundell, Stock 1992), ha iniziato ad essere applicata alla lingua russa a partire dalla fine del XX-inizio XXI secolo (Reznikova, Kopotev 2005; Zacharov 2013). Da allora numerosi sono gli archivi digitali interrogabili e annotati creati per il russo, anche grazie alla collaborazione di studiosi italiani, ai quali si deve, ad esempio, l'implementazione del corpus parallelo italiano-russo (Noseda 2018; Biagini, Bonola, Noseda 2019). Tuttavia, la compilazione di corpora rivolti allo studio della comunicazione specializzata in lingua russa rimane una frontiera ancora poco esplorata.

In tale contesto, questo saggio intende presentare RU_SEAH, un corpus di lingua russa per la comunicazione specializzata nel settore dell'architettura e delle costruzioni (AC), nato in seno al progetto multilingue *Sharing European Architectural Heritage: Innovative Language Teaching Tools for Academic and Professional Mobility in Architecture and Construction* (SEAH), avviato nel 2020 presso il Dipartimento di Lingue, Letterature e Culture Moderne dell'Università «G. d'Annunzio» di Chieti-Pescara.[2]

Al fine di apprezzare novità e potenzialità della nuova risorsa linguistica, a un rapido excursus dei principali corpora di lingua russa esistenti (§ 2) seguiranno notizie essenziali sul progetto SEAH (§ 3.1), volte a contestualizzare il ruolo svolto dai corpora al suo interno (§ 3.2); saranno poi illustrate le modalità di compilazione, la struttura (§ 3.3) e il funzionamento (§ 3.4) del corpus e infine una delle applicazioni didattiche (§ 3.5) di RU_SEAH nei moduli linguistici del progetto. L'utilità della risorsa in ambito di didassi RKI (*Russkij kak inostrannyj*, 'Russo come lingua straniera') in generale, nonché i suoi possibili impieghi in ordine alle ricerche di carattere linguistico, lessicografico e semantico meritano, invece, di essere trattati estesamente in un saggio futuro.

1 Si pensi al *Brown Corpus* (*Brown University Standard Corpus of Present-Day American English*) negli Stati Uniti – disponibile oggi online in una versione ampliata caricata su Sketch Engine (`https://www.sketchengine.eu/brown-corpus/`) – e al *London-Lund Corpus* nel Regno Unito (`https://www.nb.no/sprakbanken/en/resource-catalogue/oai-clarino-uib-no-london-lund/`).

2 Il progetto SEAH è coordinato da Mariapia D'Angelo, mentre Sara Piccioni e Maria Chiara Ferro coordinano rispettivamente le sezioni relative allo sviluppo dei corpora specialistici (*Language Corpora Design, Creation and Distribution – French, German, Italian, Russian and Spanish for Architecture and Construction*) e dei materiali glottodidattici (*Developing and Testing Innovative Language Learning Materials – French, German, Italian, Russian and Spanish for Architecture and Construction*). Al progetto partecipano le seguenti università: Università «G. d'Annunzio» di Chieti-Pescara (Italia, ateneo capofila), Bordeaux Montaigne (Francia), Politecnico di Madrid (Spagna), Polotsk State University (Bielorussia) e Masaryk University (Repubblica Ceca).

2 I corpora di lingua russa. Breve rassegna

Agli albori della creazione di archivi interrogabili di testi in lingua russa si situa la realizzazione di liste e dizionari di frequenza delle parole russe, che ha interessato gli studiosi negli anni Cinquanta (Josselson 1953), Sessanta (Shteinfeld 1963), e Settanta (Zasorina 1977) del XX secolo. Avanguardistici per l'epoca, tali lavori non sempre risultano rappresentativi del russo attuale, in ragione dei profondi cambiamenti che la lingua ha subito nell'ultimo trentennio e che sono meglio registrati da strumenti più recenti, quali, ad esempio, quelli prodotti nell'ambito del progetto KELLY (Kilgarriff et al. 2014a, 127)[3] o il dizionario di frequenza sviluppato da S. Sharoff, E. Umanskaya, J. Wilson (2013). Un primo corpus di testi russi[4] compare negli anni Ottanta – con circa un ventennio di ritardo rispetto ai corpora sviluppati nel mondo anglosassone – all'interno del *Computer Fund of the Russian Language* (Shaikevich 1997), base per la creazione del *Uppsala Russian Corpus*,[5] contenente testi pubblicistici e testi rappresentativi del dominio letterario.

Tra la fine del XX e l'inizio del XXI secolo si assiste allo sviluppo del *Tübingen Russian Corpus* (1999), all'interno del progetto *Linguistische Datenstrukturen. Theoretische und empirische Grundlagen der Grammatikforschung* condotto dall'Università di Tübingen, del *Komp'juternyj korpus tekstov russkich gazet konca XX veka* (Corpus informatizzato di testi di giornali russi della fine del XX secolo), presso l'Università Statale di Mosca (2000), dell'*Helsinki Annotated Corpus* (HANCO, 2001), realizzato all'Università di Helsinki (Reznikova, Kopotev 2005).[6] Quasi in contemporanea presso l'Accademia delle Scienze russa (RAN) viene avviato il progetto *Filologija i informatika* in seno al quale nasce il *Nacional'nyj korpus russkogo jazyka*[7] (NKRJa) (Corpus nazionale della lingua russa), reso disponibile in *open access* a partire dal 2003-04 (Sičinava 2005; Plungjan et al. 2005) e che fino a oggi ha conosciuto progressive implementazioni. Il NKRJa rappresenta lo strumento principe per quanti si occupano di *corpus linguistics* applicata alla lingua russa, in ragione della sua consistenza – allo stato attuale supera il miliardo di parole – e dell'ampiezza delle varietà diamesiche, diastratiche, diafasiche e diacroniche lì ricompre-

3 http://kelly.sketchengine.co.uk/.

4 Per maggiori informazioni sui singoli corpora disponibili per la lingua russa si vedano, insieme ai riferimenti bibliografici indicati nelle note seguenti, la sezione dedicata agli «altri corpora» all'interno del NKRJa (https://ruscorpora.ru/new/corpora-other.html); Reznikova, Kopotev 2005; Zacharov 2013.

5 https://snd.gu.se/en/catalogue/study/ext0071.

6 https://ruscorpora.ru/new/sbornik2005/04reznikova.pdf.

7 https://ruscorpora.ru/new/index.html.

se. Insieme al corpus generale (*osnovnoj korpus*), che include testi in prosa dei secoli XVIII-XXI, vi si trovano i corpora pubblicistico, dialettale, della lingua parlata, poetico e multimediale, ben 17 corpora paralleli bi- o multilingue (Dobrovol'skij, Kretov, Šarov 2005) e infine i corpora sintattico, didattico e della storia dell'accento russo. Ulteriori elementi apprezzabili del NKRJa sono la quantità e la qualità dei metadati accessibili attraverso l'interfaccia online e le numerose possibilità di interrogazione al suo interno, che non trovano analoghi negli altri corpora di lingua russa, quali ad esempio quello incluso nei *Leeds University Corpora*,[8] quello realizzato all'interno del progetto *Avtomatičeskaja obrabotka tekstov* (Elaborazione automatica del testo)[9] o quello accessibile previa registrazione sulla piattaforma Sketch Engine.[10]

Una menzione a parte meritano i cosiddetti corpora 'didattici', tra i quali ricordiamo il *Russkij učebnyj korpus* (Corpus didattico russo),[11]che raccoglie testi orali e scritti composti da apprendenti di russo come lingua straniera (LS), e il *Korpus russkich učebnych (akademičeskich) tekstov* (KRUT; *Corpus of Russian Student Texts*, CoRST),[12] collezione di testi inerenti diverse discipline di studio (economia, sociologia, scienze politiche, diritto, psicologia, giornalismo, linguistica, storia, filologia, logistica, matematica, filosofia), redatti da studenti universitari e utili tra il resto per osservazioni relative all'analisi degli errori; sempre tra i corpora didattici vanno poi menzionati quelli che pongono gli studenti come destinatari, ad esempio ruSKELL,[13] ideato appositamente per facilitare l'apprendimento del russo come LS.

Tra i corpora della lingua parlata si segnalano «*Rasskazy o snovidenijach» i drugie korpusa svučajuščej reči* («Narrazioni di sogni» e altri corpora del linguaggio parlato),[14] contenente la trascrizione della narrazione di racconti per l'infanzia, e il corpus ORD (*Odin Rechevoj Den'*, 'Una giornata di discorso') sviluppato presso l'Università Statale di San Pietroburgo (Asinovsky et al. 2009) e che, come dice il nome, mira a riprodurre il linguaggio quotidiano.

Una serie di altri corpora sono dedicati a varianti dialettali del russo, e precisamente: la banca dati *Ėlektronnye bazy dannych po russkim narodnym govoram* (Banche dati elettroniche delle parla-

8 http://corpus.leeds.ac.uk/ruscorpora.html.
9 http://www.aot.ru/index.html.
10 https://www.sketchengine.eu/.
11 http://web-corpora.net/RLC.
12 http://web-corpora.net/learner_corpus.
13 https://www.sketchengine.eu/russian-skell-corpus/.
14 http://spokencorpora.ru/.

te popolari russe),[15] accessibile in formato STARLING, creata presso l'istituto di lingua russa dell'Accademia delle Scienze russa (IRJa RAN), la *Ėlektronnaja biblioteka russkich narodnych govorov* (Biblioteca elettronica delle parlate popolari russe)[16] e il *Mul'timedijnyj korpus dialektnych tekstov Ust'janskogo rajona Archangel'skoj oblasti* (Corpus multimediale di testi dialettali del distretto Ust'janskij della regione di Archangel'sk).[17]

Infine, particolare attenzione è rivolta alla digitalizzazione e catalogazione dei testi antichi e medievali, come dimostrano i numerosi corpora diacronici, tra i quali menzioniamo qui RRuDi (*A Russian Diachronic Online Corpus*),[18] realizzato presso l'Università di Regensburg in collaborazione con l'Università Humboldt di Belino; il corpus di testi agiografici compilato presso l'Università di San Pietroburgo (SKAT – *Sankt-Peterburskij korpus agiografičeskich tekstov*, 'Corpus pietroburghese dei testi agiografici');[19] il corpus *Manuskript*[20] approntato a Udmurt; i corpora di paleoslavo composti a Helsinki (*Corpus Cyrillo-Methodianum Helsingiense: Corpus of Old Church Slavonic Texts*);[21] e il recentissimo (2020) *The World Wide Web portal for the study of Cyrillic and Glagolitic manuscripts and early printed books*.[22]

Come si vede, in poco più di due decenni, la linguistica dei corpora applicata alla lingua russa ha raggiunto risultati apprezzabili, dotandosi di validi strumenti per l'indagine *corpus-based* della lingua standard odierna – nelle sue varietà diamesiche – e passata, nonché dei linguaggi letterario e pubblicistico; anche l'indagine delle varietà dialettali del russo può avvalersi di alcune banche dati di riferimento. Il territorio della compilazione di corpora rivolti allo studio della comunicazione specializzata in lingua russa, invece, inizia solo recentemente e solo in minima parte a essere esplorato: all'interno dei *Leeds University Corpora*, ad esempio, è compreso il *Russian Business Corpus*, che fotografa il russo economico-aziendale; inoltre – come ben spiega Ryčkova (2014) – nei testi pubblicistici, sovente inclusi nell'inventario delle fonti dei vari corpora della lingua russa,

15 https://www.ruslang.ru/krylov_dialect.

16 https://dialekt.corpus.tatar/.

17 https://philology.hse.ru/variety/dialects.

18 https://www.slawistik.hu-berlin.de/de/member/meyerrol/subjekte/rrudi#autotoc-item-autotoc-2.

19 http://project.phil.spbu.ru/scat/page.php?page=project.

20 http://mns.udsu.ru/.

21 https://metashare.csc.fi/repository/browse/corpus-cyrillo-methodianum-helsingiense-corpus-of-old-church-slavonic-texts-source/a6cfa17437cb11e29c67005056be118e94602a335d6f4ea0af5025eb1e8034a7/.

22 https://www.obshtezhitie.net/.

a seconda dell'argomento trattato può comparire terminologia settoriale specifica di svariati ambiti professionali. Tuttavia, la presenza di lessico specializzato all'interno dei documenti è cosa differente da un corpus appositamente compilato per restituire le pratiche discorsive proprie dei generi testuali più tipici di una determinata *discourse community*, in ordine alle caratteristiche morfosintattiche, pragmatiche e agli aspetti linguistico-retorici tipici di quel comparto disciplinare o produttivo.

Limitando la nostra attenzione ai settori professionali chiamati in causa per la divulgazione e comunicazione del patrimonio storico-artistico e culturale – che risulta di estrema attualità nella politica culturale e linguistica europea[23] – un progetto scientifico che ha condotto alla creazione di un corpus di una particolare varietà diafasica della lingua russa è da considerarsi il *Russkij Korpus «Leksika kul'turnogo nasledija»* (Corpus russo «Lessico dei beni culturali»).[24] Realizzato nel 2020 da V. Rossi e N. Žukova, sotto la direzione di M. Garzaniti, nell'ambito del progetto *Lessico multilingue dei beni culturali* promosso dal Dipartimento di Lingue, Letterature e Studi interculturali dell'Università di Firenze a partire dal 2013, il *corpus LBC russo* riflette il linguaggio storico-artistico e delle arti al fine di favorire la conoscenza all'estero del ricco bagaglio culturale italiano, e fiorentino in particolare (Rossi, Garzaniti, Zhukova 2020). I tratti di maggiore pregio che a nostro avviso lo caratterizzano sono in primo luogo la scelta di comporre una banca dati rappresentativa delle variazioni semantiche che il linguaggio di riferimento registra in prospettiva diacronica, elemento assolutamente necessario quando si ha a che fare con testi che descrivono monumenti, reperti e opere di epoche e tradizioni diverse; di conseguenza, la seconda cifra notevole risulta l'oculata selezione di testi autorevoli che compongono quel corpus.

In un settore professionale contiguo si colloca RU_SEAH, che passiamo ora a presentare.

23 Insieme alle possibilità offerte in tal senso dai partenariati estesi del *Programma ERASMUS+ K2* (https://erasmus-plus.ec.europa.eu/it/programme-guide/part-b/key-action-2), si pensi al programma culturale *Europa Creativa 2021-2027* (http://www.europacreativa-media.it/europa-creativa) da poco avviato.

24 http://corpora.lessicobeniculturali.net/ru/.

3 Il corpus RU_SEAH

3.1 Il progetto SEAH: finalità e metodo

SEAH è un progetto multilingue finanziato dall'UE nel quadro del *Programma Erasmus+ K203*, intrapreso nel 2020 e finalizzato alla realizzazione di risorse linguistiche open access in cinque lingue straniere (francese, italiano, russo, spagnolo, tedesco), ideate appositamente per sopperire alle necessità formative degli studenti in mobilità e delle categorie professionali coinvolte con committenti stranieri, pubblici e privati nel campo dell'AC.[25] Per rispondere alle istanze di internazionalizzazione e di condivisione dei saperi che l'Unione Europea promuove anche attraverso gli scambi universitari, infatti, la salvaguardia del multilinguismo appare un elemento cruciale, pena il rischio di omologazione e appiattimento delle specificità e ricchezze dei diversi contesti e tradizioni geografico-culturali.

Nella prima fase del progetto si è pervenuti alla compilazione di corpora specialistici comprendenti testi rappresentativi di alcuni sottodomini nel campo dell'AC nelle suddette lingue, tappa propedeutica alla seconda fase, che prevede l'elaborazione di moduli linguistici online per l'apprendimento del linguaggio accademico dell'AC nelle cinque lingue. Tali risorse sono disponibili sul sito di SEAH (`https://www.seahproject.eu/`) tramite una *open educational resources platform*.

Nel definire il quadro teorico e i fondamenti metodologici che stanno alla base di SEAH, come spiega Mariapia D'Angelo (Piccioni, D'Angelo, Ferro 2021), coordinatrice del progetto, la *corpus linguistics*, che fonda l'implementazione dei corpora, si allea con le ricerche condotte – sia sul versante degli studi descrittivi (Swales 1990; Bathia 1993) che su quello delle applicazioni didattiche – nell'ambito della *genre analysis*, la quale, indagando le modalità con le quali i testi riflettono e al contempo formano le comunità che ne fanno uso, costituiscono un prezioso strumento per l'individuazione dei bisogni comunicativi dei discenti non nativi (Hyland 2014). Altra cifra caratteristica dell'orizzonte teorico-metodologico del progetto è rappresentata dagli studi sulla didattica dei linguaggi della comunicazione specializzata: mutuando per le cinque lingue di SEAH la distinzione tra *English for General Academic Purposes* (EGAP) e *English for Specific Academic Purposes* (ESAP), operata alla fine del XX secolo, le risorse linguistiche del progetto, al fine di corrispondere alle reali esigenze linguistiche degli apprendenti, non sono elaborate sulla base di una rigida sequenza di contenuti imposta da sillabi astratti

25 Per una presentazione esaustiva del progetto SEAH e una panoramica generale sulla progettazione e realizzazione dei corpora si rimanda a Piccioni, D'Angelo, Ferro 2021.

(Hyland 2016) – che contemplano l'introduzione dei linguaggi specialistici e dei generi del discorso accademico soltanto previo raggiungimento del livello B2 nella lingua standard – bensì a partire da testi autentici prodotti per la comunicazione accademica disciplinare e dei linguaggi professionali nel settore dell'AC, quali risultano i materiali selezionati per comporre i corpora.

3.2 I corpora linguistici nel progetto SEAH

All'interno del progetto i corpora sono stati creati appositamente in virtù delle molteplici funzioni che possono assolvere nell'ambito di una didassi della lingua della comunicazione specializzata condotta attraverso materiali *corpus based*, e, di conseguenza, sono calibrati in base alle esigenze di apprendimento e ai bisogni linguistici dei discenti che saranno i fruitori delle *open education resources*: architetti e ingegneri. In particolare, i corpora SEAH sono bilanciati sia orizzontalmente, cioè contemplano una rosa di argomenti diversi e rilevanti nell'ambito dell'AC, sia verticalmente, ovvero includono materiali di registro diverso a seconda del grado di specializzazione dei partecipanti alla comunicazione, da quelli prodotti da studiosi e professionisti e rivolti a colleghi e studenti, a quelli pensati per la divulgazione del sapere financo al pubblico generale (Berruto 1987).

I domini, i generi testuali e le pratiche comunicative che costituiscono un ostacolo alla mobilità internazionale sono stati individuati attraverso una *need analysis* condotta nella fase di progettazione consultando docenti nei Dipartimenti di AC, insieme a professionisti e docenti di lingua impegnati nell'insegnamento del linguaggio specializzato dell'AC.

I corpora SEAH rappresentano perciò anzitutto grandi raccolte di testi rappresentativi del linguaggio specializzato dell'AC, prodotti da docenti e professionisti in contesti comunicativi reali, dalle quali si possono attingere materiali da impiegare come testi *pivot* nelle unità didattiche che si auspica risultino motivanti, in quanto percepite dagli apprendenti come significative per il proprio percorso formativo. In seconda istanza, in linea con i principi del *Data-Driven Learning* (Johns 1991) applicato all'insegnamento dei linguaggi accademico-professionali (Lee, Swales 2006), l'uso guidato dei corpora da parte degli utenti è inteso a sviluppare una maggiore consapevolezza dei tratti propri del linguaggio disciplinare di riferimento (Flowerdew 2015). L'interrogazione dei corpora permette il recupero di unità terminologiche difficilmente reperibili in materiali di consultazione della lingua generale, nonché l'estrazione e analisi di strutture rilevanti dal punto di vista lessicale, morfo-sintattico e retorico-pragmatico nel settore dell'AC. La consultazione dei corpora durante l'apprendimento della LS permette altresì, osservando il comportamento dei

lemmi nel contesto della frase, di acquisire deduttivamente le regole della LS, innescando dei meccanismi di apprendimento che il discente potrà impiegare in autonomia nell'acquisizione di altre unità lessicali e delle strutture morfosintattiche in cui esse funzionano.

Per quanto concerne la lingua russa le risorse linguistiche sono state sviluppate dall'unità dell'Ateneo pilota,[26] in collaborazione con l'unità di ricerca dell'Università Statale di Polock in Bielorussia.[27] L'apporto del gruppo di lavoro bielorusso è stato determinante sia nella fase di ideazione che in quella di realizzazione del corpus RU_SEAH, in modo particolare per il reperimento di materiali pedagogico-didattici, quali consegne d'esame, piani di studio, elaborati di fine corso degli studenti, incluse alcune registrazioni, che sono stati attinti dall'archivio di quella università, grazie alla disponibilità dei colleghi dei dipartimenti di Architettura e Design e di Ingegneria.

3.3 RU_SEAH: compilazione e struttura

Nella fase di progettazione del corpus le aree tematiche individuate come maggiormente significative nella consultazione con gli *stakeholder* che hanno collaborato alla compilazione di RU_SEAH sono risultate: storia e teoria dell'architettura, progettazione architettonica e design, architettura sostenibile. I criteri fondamentali che hanno determinato la scelta di questi temi sono anzitutto la rilevanza nel comparto professionale di riferimento e, in considerazione delle verosimili conoscenze del *target group*, la presenza nei piani di studio universitari. L'inclusione di materiali sull'architettura sostenibile è motivata anche dall'attualità di questo campo d'indagine e dagli sviluppi teorici e pratici che lo stesso sta conoscendo negli ultimi anni. La limitazione a soli tre ambiti va ricondotta ai limiti temporali imposti dal progetto, e non esclude la possibilità di futuri ampliamenti del corpus, sia attraverso l'inclusione di altre aree tematiche dell'AC, sia tramite l'ampliamento dei topic a settori disciplinari contigui (ad esempio beni culturali, archeologia sul fronte storico-artistico, i diversi comparti dell'ingegneria civile sul versante più tecnico, e così via).

I testi selezionati a partire da questi argomenti appartengono ai quattro domini (*domain* nell'interfaccia di interrogazione del corpus) contemplati in tutti i corpora SEAH, e precisamente il Dominio acca-

26 Composta da Maria Chiara Ferro e Natalia K. Guseva.

27 Composta da Svetlana M. Ljasovich, Natalia G. Apanasovich, Natalia N. Nester, Volha V. Zhurauskaya. La scelta di questo partner è stata dettata, oltreché dal bilinguismo presente in Bielorussia, dai requisiti del bando della call Erasmus+, che prevedevano pregressi rapporti di collaborazione formalizzati (attraverso accordi Erasmus o convenzioni internazionali) con gli atenei partecipanti.

demico (ACCAD), che include testi in cui ricercatori e/o studenti presentano i risultati delle proprie ricerche alla comunità scientifica; il Dominio didattico (DID), che comprende testi prodotti da ricercatori o docenti e rivolti a studenti; il Dominio professionale (PROF), i cui testi di riferimento sono prodotti da professionisti del campo dell'AC rivolti ai colleghi e, in seconda battuta, a studenti e ricercatori/docenti; e infine il Dominio divulgativo (DIV), che identifica testi prodotti da ricercatori/docenti e/o professionisti e rivolti al pubblico generale.

Nella fase di lavorazione, i testi individuati sono stati raccolti, classificati e immagazzinati in formato .doc o .docx. I testi in .pdf sono stati convertiti e successivamente 'ripuliti' per garantirne l'integrità formale, la qualità e la rilevanza ai fini del progetto (ad esempio sono stati eliminati intestazioni e piè di pagina, tabelle, illustrazioni, bibliografia non utile, ecc.). Anche nel caso dei testi orali si sono rese necessarie operazioni di *post-editing* della trascrizione automatica ottenuta tramite la funzione Trascrizione della versione online di Word (Microsoft 365). I file sono stati archiviati, classificando ogni testo in base a 10 metadati (*ID*, *Language*, *Mode*, *Domain*, *Genre*, *Title*, *Author type*, *Source*, *Size*, *URL*), utili nella fase di consultazione del corpus per il recupero automatico di testi e alla selezione di subcorpora per ricerche mirate.

Come illustra Sara Piccioni, responsabile dell'architettura generale dei corpora SEAH, l'ultima fase di realizzazione del corpus è stata l'annotazione e indicizzazione dei testi, tramite il software di gestione di corpora Sketch Engine (Kilgarriff et al. 2014b):

> queste operazioni hanno permesso di corredare i corpora di lemmatizzazione e annotazione morfosintattica, attribuendo a ciascuna parola un'etichetta che ne identifica il lemma (o forma base di riferimento), la categoria morfosintattica cui appartiene e ulteriori informazioni morfologiche (genere e numero per sostantivi e aggettivi, tempi e persone verbali per i verbi, ecc.). Utilizzando le funzionalità di creazione dei corpora di Sketch Engine, lemmatizzazione e annotazione morfosintattica sono state realizzate […] con MULTEXT-East (Erjavec 2012; 2017) per la lingua russa. (Piccioni, D'Angelo, Ferro 2021, § 2)

Il corpus RU_SEAH compilato manualmente e linguisticamente annotato è liberamente accessibile al link `https://corpora.unich.it/sito/seah-corpora-it.html`.

La dimensione di RU_SEAH ammonta a 1.240.746 *token* (d'ora in poi: 't.'), i documenti caricati sono 170, appartenenti a 15 dei 22 generi testuali selezionati nella fase di progettazione dei corpora SEAH;[28]

28 L'elenco completo delle tipologie testuali corrispondenti a ciascuno dei quattro domini di riferimento è pubblicato in Piccioni, D'Angelo, Ferro 2021.

tutte e quattro le tipologie d'autore contemplate in SEAH (accademico, professionista, media, studente) sono rappresentate. Quanto al canale di ricezione (*mode* nell'interfaccia, si veda la teoria del registro in Halliday, Hasan 1985), sono stati inclusi sia testi scritti che testi orali, nella misura indicativa del 79% e 21% rispettivamente. Se nell'architettura generale dei corpora SEAH lo sbilanciamento a favore dei testi scritti si deve al rilievo che la comunicazione scritta ha in contesti accademici e in seconda istanza agli oneri del processo di lavorazione del testo orale ai fini del suo inserimento nel corpus nelle fasi di trascrizione, controllo e pulizia della bozza ottenuta con software *speech to text*, nel caso della lingua russa si sono rivelate determinanti sia la composizione del *target group* che la tipologia di linguaggio in questione. In particolare, va considerato che il livello presunto delle conoscenze pregresse degli studenti e professionisti cui sono rivolti i moduli di SEAH è il B1, cioè un livello *low-intermediate*, mentre la realizzazione orale del discorso accademico in lingua russa presenta un maggiore grado di complessità rispetto a quello scritto, potendo includere marche colloquiali o espedienti linguistico-testuali di carattere espressivo-emozionale (Markova 2016, 111).

La versione attuale di RU_SEAH presenta uno sbilanciamento a favore dei testi pedagogico-didattici e scientifici, che tuttavia appare in linea con gli scopi generali del progetto SEAH, essendo gli studenti in mobilità i primi destinatari delle risorse create. Nello specifico, come si vede anche dal grafico in figura **1**, il corpus di testi scritti conta 976.283 t. ripartiti tra testi dei domini accademico (monografie, 248.039 t.; articoli scientifici, 105.743 t.; abstract di tesi di dottorato, 111.699 t.), didattico (manuali, 363.482 t.; materiali per l'insegnamento, 18.468 t.; report di progetto 9.548 t.; descrizione di progetto, 11.656 t.), e professionale (materiali di mostre ed esposizioni, 107.648 t.). Il corpus orale consiste invece di 264.463 t., tra testi accademici (discussione di tesi dottorali, 38.091 t.; conferenza, 25.531 t., tavola rotonda, 5.485 t.), divulgativi (lezioni pubbliche, 78.109 t.; podcast, 41.514 t.) e didattici (pitch di progetto, 6.812 t.).

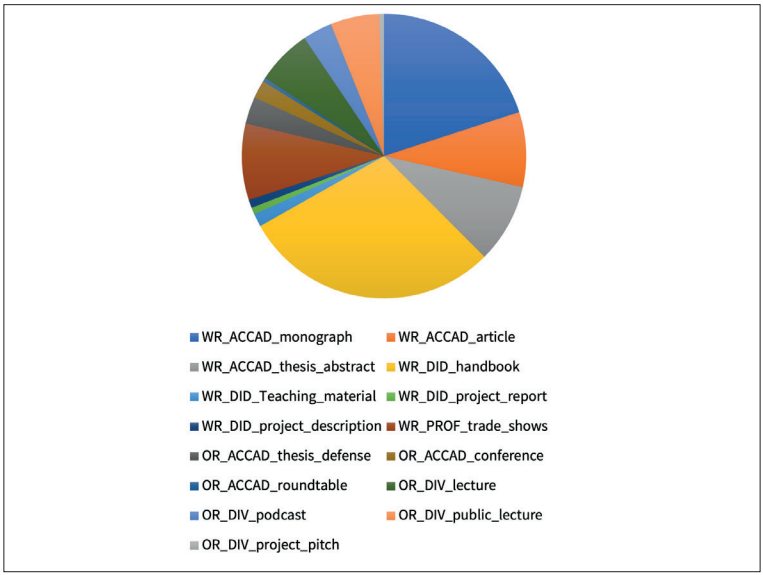

Figura 1 Composizione del corpus RU_SEAH

3.4 RU_SEAH: funzionamento

La piattaforma NoSketch Engine (Rychlý 2007), sulla quale il corpus funziona, è uno strumento di interrogazione *open source* per la gestione dei corpora che supporta diverse funzionalità.

Le principali modalità di ricerca sono tre: *Concordance*, *Wordlist* e *Keywords* **[fig. 2]**. Estremamente utile risulta la guida all'utilizzo dell'interfaccia – si vedano i tutorial videoregistrati e le sezioni *About* – che permettono all'utente di comprendere in autonomia le potenzialità offerte e le possibilità di impiego delle risorse caricate. Nell'ambito del progetto SEAH le funzionalità attualmente implementate per i corpora linguistici sono le prime due.

Lo strumento *Concordance* include due modalità di interrogazione, semplice (*basic* sull'interfaccia) e avanzata (*advanced*), tramite le quali è possibile studiare il comportamento di parole, espressioni e financo frasi visualizzando i risultati nel contesto sotto forma di concordanza; quest'ultima può essere ordinata, filtrata ed elaborata ulteriormente in base a una serie di scelte. Le opzioni di visualizzazione permettono di evincere informazioni aggiuntive come lemmi, attributi, metadati e altre strutture del corpus. La ricerca CQL (*Corpus Query Language*) nella scheda avanzata è utile per ricerche complesse, che necessitano dell'inserimento di formule: impratichendosi nella sintassi necessaria, si possono recuperare, ad esempio, ag-

gettivi che qualificano un sostantivo e viceversa, reggenze verbali, polirematiche, e via dicendo. Nelle figure **3** e **4** riportiamo la schermata di interrogazione per il recupero delle sequenze 'aggettivo+фасад' e 'дом+с+nome' con i relativi risultati.

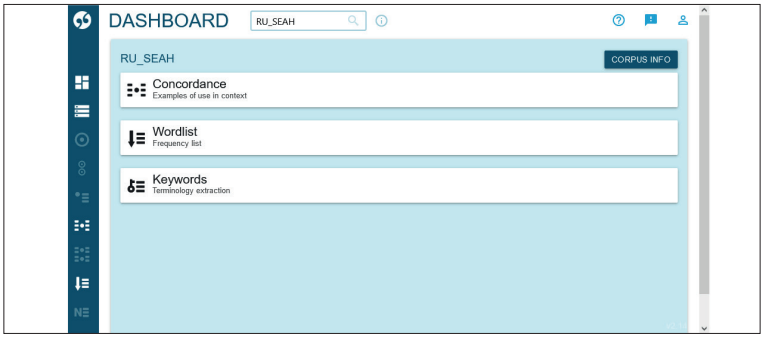

Figura 2 Dashboard del corpus RU_SEAH

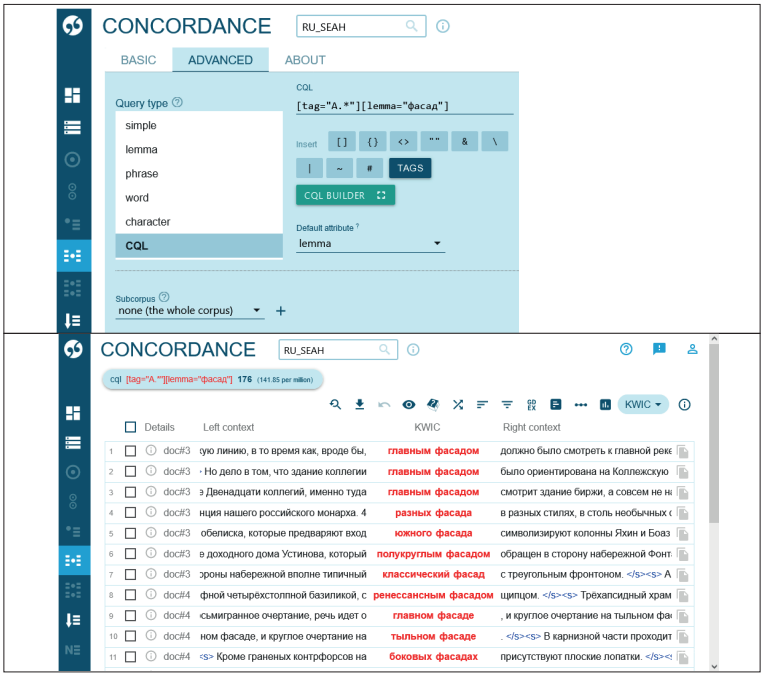

Figura 3 Esempi di interrogazione del corpus RU-SEAH tramite la funzione *Concordance*/*Advanced*/CQL. Sequenza 'aggettivo+фасад' e relativi risultati

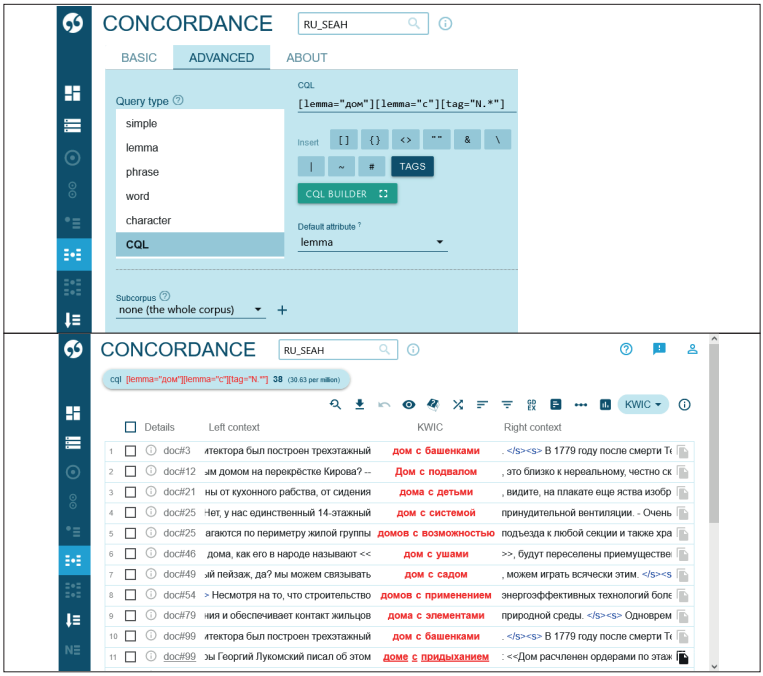

Figura 4 Esempi di interrogazione del corpus RU-SEAH tramite la funzione *Concordance*/*Advanced*/CQL. Sequenza 'дом+с+nome' e relativi risultati

La funzione *Wordlist* consente di generare elenchi di parole e liste di frequenza in base ad una molteplicità di criteri selezionabili all'interno della ricerca base e avanzata:

- nomi, verbi, aggettivi e altre parti del discorso;
- parole che iniziano, finiscono, contengono determinati caratteri;
- singole forme di parole, tag, lemmi e altri attributi;
- o ancora una combinazione delle tre opzioni precedenti.

Ad esempio, la figura **5** riproduce l'interrogazione del corpus tramite la funzione *Wordlist* in modalità di ricerca *advanced* finalizzata ad apprezzare la frequenza d'uso dell'anglicismo сквер rispetto al suo iperonimo russo площадь.

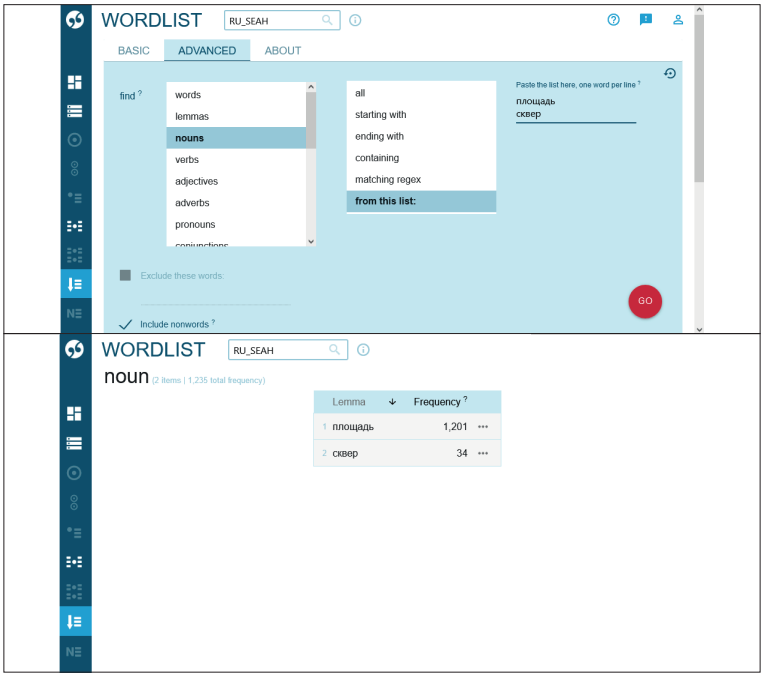

Figura 5 Esempio di interrogazione del corpus RU-SEAH tramite la funzione *Wordlist*. Frequenza dei lemmi площадь e сквер

In entrambe le modalità è possibile effettuare una ricerca sull'intero corpus, oppure su un sub-corpus appositamente selezionato; le variabili contemplate in questa funzione sono quelle definite nella fase di progettazione del corpus **[fig. 6]**, e spaziano dalla possibilità di esplorare un singolo testo, scelto in base all'identificativo (*File ID*) o alla sua fonte (*Source*), a quella di interrogare i testi di uno o più domini (*Domain*), di una particolare tipologia di autore (*Author type*) o di un determinato genere testuale (*Genre*).

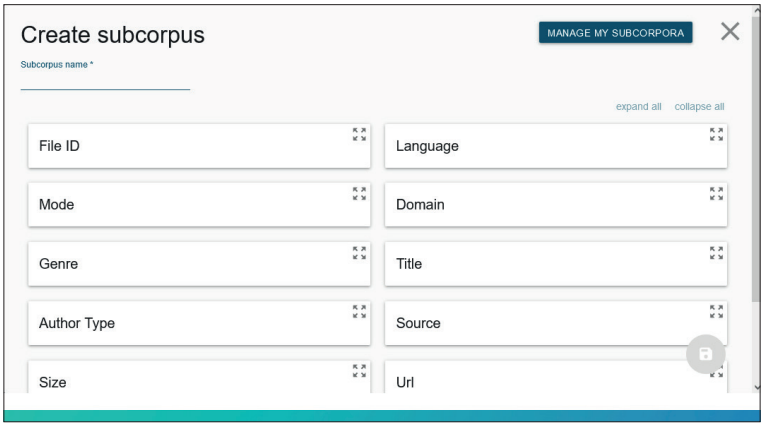

Figura 6 Parametri di creazione di subcorpus per ricerche mirate

3.5 RU_SEAH: un esempio di applicazione didattica

Come si è detto (§ 3.2), nell'ambito del progetto SEAH il corpus russo ha rivestito come funzione principale quella di archivio di materiali testuali autentici e specializzati utili al reperimento di testi *pivot* finalizzati all'elaborazione delle unità didattiche. Oltre a ciò, in considerazione dei bisogni formativi specifici della *target audience* di riferimento, l'impiego di RU_SEAH appare particolarmente utile ai fini dell'arricchimento lessicale, sia in vista dello sviluppo di glossari autonomi[29] sia per quanto concerne l'introduzione della terminologia settoriale. Ecco perché ciascun modulo propone attività *corpus-based*.

Nell'unità dedicata all'architettura della Russia contemporanea (*Architektura sovremennoj Rossii*) il testo di partenza, che descrive alcune caratteristiche architettoniche tipiche delle grandi città, reca spesso l'aggettivo городской (della città, cittadino). Si tratta di un attributo di cui il discente di livello B1 verosimilmente conosce il significato e, se anche non lo conoscesse, potrebbe facilmente intuirlo, dato che l'aggettivo è derivato dal sostantivo город (città), uno tra i primi a essere introdotti nella didassi del russo a stranieri. La consegna dell'esercizio chiede di reperire, attraverso l'interrogazione del corpus tramite sintassi CQL, i sostantivi che possono combinarsi con l'aggettivo dato.

29 Primi cenni a questo tema sono contenuti in Piccioni, D'Angelo, Ferro 2021, § 5.1.

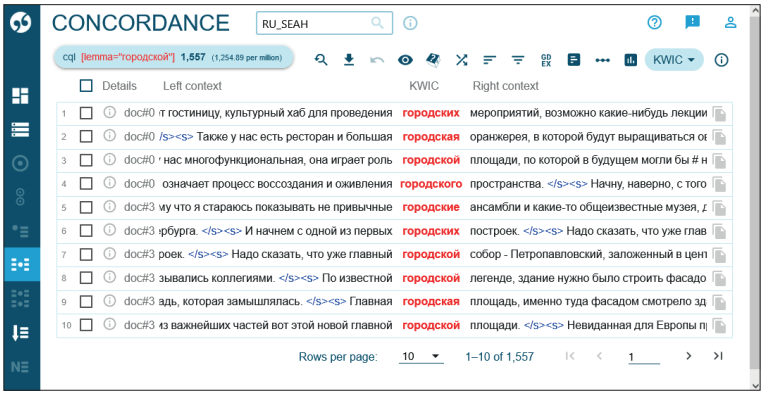

Figura 7 Interrogazione del corpus tramite sintassi CQL per la stringa [lemma=«городской»]; risultati 1-10 (di 1.557)

L'attività si concentra sui primi 40 risultati, nei quali il discente deve individuare il sostantivo che di volta in volta è concordato con l'aggettivo городской e trascriverlo nella sua forma base (cioè al nominativo singolare). La soluzione del *task* prevede la registrazione dei seguenti 23 sostantivi:

> мероприятие, оранжерея, **площадь**, пространство, **ансамбль**, постройка, **собор**, легенда, **архитектор**, инфраструктура, планирование, **район**, **среда**, **центр**, автобус, ферма, **транспорт**, **столовая**, узел, население, **план**, механизм, технология

Questo semplice compito costituisce un'occasione di consolidamento della flessione nominale, obiettivo fondamentale della didassi RKI fino a tutto il livello intermedio (I/B1): nei differenti contesti d'uso, infatti, i sostantivi appaiono in casi diversi dal nominativo, anche al femminile o al neutro, e al plurale, oltre che al singolare. Nello specifico, insieme alle regole principali della flessione del nome sostantivo, sono qui da richiamare alla mente la declinazione in segno debole, sia dei maschili (qui *ansambl'*) che dei femminili (qui *ploščad'*), le desinenze dei neutri in -ie (qui *meoprijatie*, *planirovanie*, *naselenie*) e quelle dei femminili in -ija (qui *technologija*).

L'ultimo *step* dell'attività prevede l'elicitazione dei sostantivi che non risultano propri del discorso sull'AC, ma funzionano anche nella lingua standard, cioè *meroprijatie*, *legenda*, *ferma*, *avtobus* e *naselenie*; per rispondere correttamente lo studente è obbligato e reperire sul dizionario il significato delle parole che non conosce. A questo punto egli avrà individuato 18 sostantivi propri dell'ambito dell'AC; tra questi quelli evidenziati in grassetto nel riquadro – per un totale di 10 sostantivi – sono presenti nel minimo lessicale del livello B1

del TRKI (Andrjušina 2015): come si vede, conclusa l'attività, il discente avrà aggiunto almeno 8 nuovi termini al suo bagaglio linguistico presunto. Inoltre, avrà appreso che i lemmi среда e столовая, registrati nel minimo lessicale in base al loro primo significato – rispettivamente di 'mercoledì' e di 'mensa' –, in contesti diversi valgono l'uno 'ambiente' e l'altro 'sala da pranzo'.

4 Conclusioni e prospettive

Il corpus RU_SEAH rappresenta una risorsa inedita nel panorama dei corpora di lingua russa. Innovativa è in primo luogo la scelta di comporre un archivio interrogabile di testi rappresentativi di un determinato linguaggio professionale, nella fattispecie quello accademico e professionale dell'Architettura e delle Costruzioni, che apre la strada a una nuova frontiera nella linguistica dei corpora applicata alla lingua russa. Come si è detto, l'esigenza di sviluppare strumenti siffatti trova conferma anche nella compilazione del *corpus LBC russo*, che, sebbene originato da presupposti differenti, condivide con RU_SEAH l'attenzione alle peculiarità proprie dei linguaggi di specializzazione. In secondo luogo, nuovo risulta il metodo seguito per la compilazione del corpus presentato, che ha preso le mosse dalla consultazione di esponenti dei comparti professionali e dell'ambito accademico, permettendo una mappatura dei reali bisogni linguistici di architetti e ingegneri, con la conseguente identificazione dei domini, generi testuali e pratiche comunicative da prediligere al fine di costruire uno strumento efficace per la *target audience*.

L'utilità di RU_SEAH è inoltre apprezzabile nella prospettiva di un approfondimento della conoscenza del patrimonio architettonico russo e delle pratiche professionali a esso connesse.

Grazie poi all'interfaccia *user friendly* e alle numerose funzionalità lì incluse, si rivela una risorsa fruibile da parte di una varietà di utenti, quali ad esempio traduttori e interpreti nel campo dell'AC, guide turistiche; analogamente può essere impiegato ben oltre gli scopi del progetto in cui ha visto la luce, ad esempio per la redazione di dizionari, glossari, liste di frequenza, e financo per la creazione di applicazioni di *natural language processing*.

Abbreviazioni

AC	Architettura e Costruzioni
EGAP	English for General Academic Purposes
ESAP	English for Specific Academic Purposes
NKRJa	Nacional'nyj korpus russkogo jazyka
LS	lingua straniera
RKI	Russkij kak inostrannyj
SEAH	Sharing European Architectural Heritage: Innovative Language Teaching Tools for Academic and Professional Mobility in Architecture and Construction
t.	token

Bibliografia

Andrjušina, N.P. et al. (2015). *Leksičeskij minimun po russkomu jazyku kak inostrannomu. Pervyj sertifikacionnyj uroven'. Obščee vladenie* (Minimo lessicale di russo come lingua straniera. Primo livello di certificazione. Livello soglia). Sankt-Peterburg: Zlatoust.

Asinovsky, A.; Bogdanova, N.; Rusakova, M.; Ryko, A.; Stepanova, S.M.; Sherstinova, T. (2009). «The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation of Principles and Annotation». Matoušek, V.; Mautner, P. (eds), *Text, Speech and Dialogue*. Berlin; Heidelberg: Springer-Verlag, 250-7. Lecture Notes in Computer Science 5729. https://doi.org/10.1007/978-3-642-04208-9_36.

Bhatia, V.K. (1993). *Analysing Genre: Language Use in Professional Settings*. London; New York: Longman.

Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo*. Firenze: La Nuova Italia.

Biagini, F.; Bonola, A.; Noseda, V. (2019). «Il corpus parallelo italiano-russo del NKRJa. Progetto di ampliamento, applicazioni e sviluppi». Bragone, M.C.; Bidovec, M. (a cura di), *Il mondo slavo e l'Europa. Contributi presentati al VI Congresso Italiano di Slavistica* (Torino, 28-30 settembre 2016). Firenze: Firenze University Press, 35-45. https://doi.org/10.36253/978-88-6453-910-2.

Dobrovol'skij, D.O.; Kretov, A.A.; Šarov S.A. (2005). «Korpus parallel'nych tekstov: architektura i vozmožnosti ispol'zovanija» (Corpus parallelo: struttura e potenzialità di uso). *Nacional'nyj korpus russkogo jazyka: 2003-2005* (Corpus nazionale della lingua russa: 2003-2005). Moskva: Indrik, 263-96. https://ruscorpora.ru/new/sbornik2005/17dobrovolsky.pdf.

Erjavec, T. (2012). «MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages». *Language Resources and Evaluation*, 46(1), 131-42. http://www.jstor.org/stable/41486069.

Erjavec, T. (2017). «MULTEXT-East». Ide. N.; Pustejovsky, J. (eds), *Handbook of Linguistic Annotation*. New York: Springer, 441-62.

Flowerdew, L. (2015). «Corpus-Based Research and Pedagogy in EAP: From Lexis to Genre». *Language Teaching*, 48(1), 99-116. https://doi.org/10.1017/S0261444813000037.

Halliday, M.; Hasan, R. (1985). *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective*. Geelong: Deakin University Press.

Hyland, K. (2014). «English for Academic Purposes». Leung, C.; Street, B. (eds), *The Routledge Companion to English Studies*. London: Routledge, 392-404. https://www.perlego.com/book/1546122/the-routledge-companion-to-english-studies-pdf.

Hyland, K. (2016). «General and Specific EAP». Hyland, K.; Shaw, P. (eds), *The Routledge Handbook of English for Academic Purposes*. London: Routledge, 17-29. https://doi.org/10.4324/9781315657455.

Johns, T. (1991). «From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning». *English Language Research Journal*, 4, 27-45.

Josselson, H. (1953). *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian*. Detroit: Wayne University Press.

Kilgarriff, A.; Charalabopoulou, F.; Gavrilidou, M. et al. (2014a). «Corpus-Based Vocabulary Lists for Language Learners for Nine Languages». *Language Resources and Evaluation*, 48, 121-63. https://doi.org/10.1007/s10579-013-9251-2.

Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P.; Suchomel, V. (2014b). «The Sketch Engine: Ten Years On». *Lexicography – Journal of ASIALEX*, 1(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9.

Lee, D.; Swales, J. (2006). «A Corpus-Based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-Compiled Corpora». *English for Specific Purposes*, 25(1), 56-75. http://hdl.handle.net/2027.42/88141.

Markova, V.A. (2016). *Stilistika russkogo jazyka. Teoretiko-praktičeskij kurs* (La stilistica della lingua russa. Corso teorico-pratico). Moskva: LENAND.

McEnery, T.; Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Noseda, V. (2018). «La *corpus revolution* russa e il corpus parallelo italiano-russo». *L'analisi linguistica e letteraria*, 26, 115-32. https://publires.unicatt.it/it/publications/la-corpus-revolution-russa-e-il-corpus-parallelo-italiano-russo-s-4.

Piccioni, S.; D'Angelo, M.; Ferro, M.C. (2021). «I Corpora SEAH di comunicazione specializzata nel settore dell'Architettura e delle Costruzioni. Struttura, compilazione e usi». *Linguistica*, 61(2), 97-123.

Plungjan, V.A.; Reznikova, T.I.; Sičinava, D.V. (2005). «Nacional'nyj korpus russkogo jazyka: obščaja charakteristika» (Corpus nazionale della lingua russa: presentazione). *Naučno-techničeskaja informacija*, s. 2, 3, 9-13.

Reznikova, T.I.; Kopotev, M.V. (2005). «Lingvističeski annotirovannye korpusa russkogo jazyka (obzor obščedostupnych resursov)» (Corpora di lingua russa annotati linguisticamente [rassegna delle risorse liberamente accessibili]). *Nacional'nyj korpus russkogo jazyka: 2003-2005*. Moskva: Indrik, 31-61. https://ruscorpora.ru/new/sbornik2005/04reznikova.pdf.

Rossi, V.; Garzaniti, M.; Zhukova [Žukova], N. (2020). «Il corpus LBC russo». Billero, R.; Farina, A.; Nicolás Martínez, M.C. (a cura di), *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali.* Firenze: Firenze University Press, 101-22. https://doi.org/10.36253/978-88-5518-253-9.

Rundell, M.; Stock, P. (1992). «The Corpus Revolution». *English Today*, 8(3), 21-32.

Rychlý, P. (2007). «Manatee/Bonito – A Modular Corpus Manager». Sojka, P.; Horák A. (eds), *Proceedings of Recent Advances in Slavonic Natural Langua-*

ge Processing (RASLAN 2007). Brno: Masaryk University, 65-70. https://nlp.fi.muni.cz/raslan/2007/.

Ryčkova, L.V. (2014). «Otraženie jazykov dlja special'nych celej v SMI Grodnenščiny» (Le lingue per scopi speciali nei mezzi di comunicazione di massa nella regione di Grodno). Rovdo, I.S. et al. (a cura di), *Russkij jazyk: sistema i funkcionirovanie (k 75-letiju filologičeskogo fakul'teta BGU) = sbornik materialov VI Meždunar. Nauč. Konf.* (Minsk, 28-29 okt. 2014g) (La lingua russa: sistema e funzionamento [per il 75° anniversario della Facoltà di Lettere dell'Università Statale della Bielorussia] = materiali del VI convegno scientifico internazionale), parte 2. Minsk: Izd. Centr BGU, 266-71.

Shaikevich, A. (1997). «The Computer Fund of Russian Language». *International Journal of Corpus Linguistics*, 2(1), 163-7.

Shteinfeld, E.A. (1963). *Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka* (Dizionario di frequenza della lingua letteraria russa moderna). Tallin: s.n.

Sharoff, S.; Umanskaya, E.; Wilson, J. (2013). *A Frequency Dictionary of Contemporary Russian: Core Vocabulary for Learners*. London: Routledge.

Sičinava, D.V. (2005). «Nacional'nyj korpus russkogo jazyka: očerk predystorii» (Il Corpus nazionale della lingua russa: tappe di elaborazione). *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy* (Corpus nazionale della lingua russa: 2003-2005. Risultati e prospettive). Moskva: Indrik, 21-30.

Swales, J.M. (1990). *Genre Analysis: English in Academic and Research Setting*. Cambridge: Cambridge University Press.

Zacharov, V. (2013). «Corpora of the Russian Language». Habernal, I.; Matoušek, V. (eds), *Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013* (Plzen, Czech Republic, 1-5 September 2013). Berlin; Heidelberg: Springer-Verlag, 1-13. Lecture Notes in Artificial Intelligence 8082. https://link.springer.com/chapter/10.1007/978-3-642-40585-3_1.

Zasorina, L.N. (a cura di) (1977). *Častotnyj slovar' russkogo jazyka* (Dizionario di frequenza della lingua russa). Moskva: Russkij Jazyk.