Original research article
UDC 811.161.1`367.625
DOI 10.21618/fil2531153p
COBISS.RS-ID 142752513

Olja Perišić[1]
Università degli Studi di Torino
Dipartimento di Lingue e Letterature straniere e culture moderne

Ranka Stanković
Università di Belgrado
Facoltà di Ingegneria Mineraria e Geologia

IT-SR-NER: IL CORPUS PARALLELO SERBO-ITALIANO PER L'APPRENDIMENTO DEL SERBO COME LINGUA STRANIERA[2]

Abstract: *Il tema del contributo verte sulle possibilità di impiego nella glottodidattica del corpus parallelo* It-Sr-NER *sviluppato nel 2022 dall'infrastruttura europea CLARIN nel contesto del progetto "Bridging gaps". Il suo valore risiede nell'annotazione per il Named Entity Recognition (NER), con particolare attenzione alla classe dei toponimi (inclusi i* pluralia tantum*) e dei nomi propri di persona. Il progetto è stato realizzato in collaborazione da un team di esperti provenienti dall'Università di Torino e dalla Società per le Risorse e le Tecnologie Linguistiche (JeRTeh) di Belgrado. Si tratta del primo corpus parallelo per questa combinazione linguistica, comprendente 10.000 frasi estrapolate dalle opere delle letterature serba e italiana, sia classiche che moderne. Le frasi sono state allineate per agevolare l'esplorazione dei dati e la traduzione delle singole parole nel contesto. L'importanza dei corpora paralleli nelle ricerche linguistiche è ormai nota (Doval e Sánchez, 2019), anche se si rileva una carenza di queste risorse per le lingue meno diffuse. Nel contributo verranno esposte le potenziali applicazioni glottodidattiche del corpus in questione nella didattica del serbo come lingua straniera: a livello iniziale, per l'apprendimento della ricca morfologia della lingua serba e, a livello intermedio e avanzato, per la ricerca dei* lexical gaps*, ossia delle parole che non hanno un traducente in una delle lingue esaminate.*

[1] olja.perisic@unito.it

[2] Il presente contributo è stato scritto da Olja Perišić, che ha curato i paragrafi 1, 3 e 4, e da Ranka Stanković, che ha curato il paragrafo 2.

L'obiettivo è dimostrare che questa tipologia di corpora può costituire un'alternativa valida agli strumenti tradizionali e digitali di apprendimento, nonché un mezzo esclusivo per alcune tipologie di ricerca.

Parole chiave: *corpus parallelo, italiano L1, serbo LS, NER, glottodidattica, traduzione.*

1. Introduzione

Questo contributo si propone di illustrare le potenzialità e le possibili applicazioni didattiche del corpus parallelo[3] *It-Sr-NER*, sviluppato nel 2022 nell'ambito del progetto "Bridging gaps" dell'infrastruttura CLARIN, in collaborazione tra l'Università di Torino e JeRTeh di Belgrado.

Le prime ricerche sui corpora paralleli compiute durante gli anni Novanta sottolineano il loro potenziale soprattutto per quanto riguarda l'ambito della teoria della traduzione, in particolare nella (ri)definizione del concetto di equivalenza. Spostando l'attenzione dal significato all'uso, grazie all'approccio corpus-based la nozione di equivalenza può essere sostituita con quella di pattern d'uso, favorendo il passaggio da regole legate al testo sorgente a categorie descrittive (Baker, 1993). Si evidenzia specialmente la possibilità, caratteristica di questo tipo di approccio, di osservare contemporaneamente un numero elevato di testi insieme alle loro traduzioni e, in questo modo, cogliere l'uso dei "norms" (Toury, 1980), ossia le scelte traduttive compiute con regolarità dai traduttori in determinati contesti linguistico-culturali, in sincronia e diacronia. Questa interconnessione di studi linguistici e traduttivi, resa possibile dall'uso dei corpora paralleli, è stata evidenziata negli studi successivi da Salkie (2002). Al di là delle traduzioni letterali, l'autore riconosce il valore delle cosiddette "inventive translations", che suddivide in due categorie: quelle più creative e quindi rare, da ammirare per la capacità del traduttore di proporre soluzioni originali, e quelle che, mantenendosi più vicine al valore semantico delle parole, si ripetono con una certa regolarità e possono essere analizzate in modo sistematico.

Sempre negli anni Novanta avvengono le prime sperimentazioni nella glottodidattica con i testi paralleli presenti sui supporti CD-ROM (Zanettin, 1994). Tale approccio si dimostra da subito efficace in quanto porta al miglioramento delle competenze testuali e traduttive degli studenti. In uno dei primi volumi dedicati

[3] Un corpus parallelo è un insieme di testi identici tradotti in due o più lingue, con segmenti (generalmente frasi) allineati per consentire il confronto diretto. Tale risorsa facilita ricerche bilingui, permettendo di analizzare traduzioni e corrispondenze lessicali o sintattiche tra le lingue coinvolte.

all'uso dei corpora multilingui nella glottodidattica gli autori Botley et al. (2000) sottolineano due principali requisiti che un corpus parallelo deve soddisfare per essere efficace e utile agli studenti. Da una parte deve essere sufficientemente esteso e rappresentativo del registro linguistico su cui si intende lavorare per fornire un numero sufficiente di campioni di dati validi. Dall'altra parte deve mantenersi accessibile e risultare di semplice utilizzo. Queste prime riflessioni e applicazioni hanno rivelato il potenziale dei corpora paralleli non solo come un'alternativa più potente e diversificata rispetto ai dizionari tradizionali, ma anche come un mezzo efficace per assistere gli studenti nella produzione di testi autentici e culturalmente specifici. Nello stesso tempo, in uno dei primi convegni dedicati ai corpora paralleli[4], Trousterud (2002) evidenziò la scarsa diversificazione dei generi testuali nei corpora paralleli e ne denunciò l'insufficienza per le lingue meno diffuse[5]. In quell'occasione lo studioso aprì il suo intervento con una frase che tuttora risulta attuale: "Corpus linguistics is dominated by English, and the work on parallel corpora is no exception" (Trousterud, 2002:p.111). Oggi, nonostante sia ormai riconosciuta e consolidata l'importanza dei corpora paralleli nelle ricerche linguistiche (Doval e Sánchez, 2019), si nota ancora una certa carenza di queste risorse (Granger, 2018), soprattutto per le lingue meno diffuse. Due sono i temi principali quando si parla dei corpora paralleli: il primo, di aspetto più tecnico, verte attorno alla loro progettazione e costruzione, il secondo, di natura più pratica, riguarda le caratteristiche e l'applicabilità di tali strumenti in quanto prodotti finiti e pronti da utilizzare anche da chi è meno propenso all'uso delle tecnologie (Doval e Sánchez, 2019). Il presente contributo affronterà entrambe le questioni.

Nella prima parte verranno presentate le caratteristiche del primo corpus parallelo per la combinazione linguistica serbo-italiano, ossia *It-Sr-NER*[6] (Perišić et al., 2023). Si tratta di un corpus innovativo per l'annotazione NER (Named Entity Recognition) che permette di identificare e classificare entità nominate come toponimi e antroponimi. Nella seconda parte si rifletterà invece sui possibili usi di questo corpus nell'insegnamento del serbo come lingua straniera, con particolare attenzione ai toponimi e agli antroponimi estrapolabili grazie a questa annotazione.

Questi elementi, spesso assenti nei dizionari bilingui, vanno considerati come vere e proprie voci enciclopediche. Inoltre, vista la natura letteraria dei testi in questione e la loro ricchezza in termini di lessemi culturalmente specifici, sarà

[4] Symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999.

[5] Nel contributo l'autore si riferisce alle lingue scandinave.

[6] Il nome completo del progetto è: "It-Sr-NER: Servizi web per il riconoscimento, il collegamento e la mappatura delle entità denominate (NER - Named Entity Recognition)".

possibile beneficiare dell'osservazione di quegli elementi che nel passaggio da una lingua all'altra non hanno un traducente (*lexical gaps*) e che presentano una sfida a livello lessicografico a causa dell'asimmetria tra i due sistemi linguistici.

In questo lavoro verrà utilizzato il concetto di similarità tra le lingue (Chesterman, 2007) e verranno applicati rispettivamente le categorie di similarità convergenti (percepite) per l'analisi contrastiva nella didattica delle lingue e di similarità divergenti (create) per la didattica della traduzione. Le similarità convergenti si riferiscono a somiglianze percepite dagli apprendenti tra due lingue, mentre quelle divergenti riguardano differenze emerse nel processo traduttivo. Dal punto di vista pratico, per gli studenti principianti sono previste attività di riconoscimento dei toponimi e degli antroponimi attraverso l'annotazione NER: a livello morfosintattico, si propone di procedere con la riconduzione delle forme flesse al lemma (e viceversa) e, a livello ortografico, con la trascrizione fonetica. Per gli studenti con un livello intermedio o avanzato vengono invece consigliate attività sui corpora paralleli attraverso l'osservazione contemporanea dei testi originali e delle loro traduzioni partendo dal valore semantico delle parole per arrivare al loro adattamento al contesto funzionale e comunicativo della lingua di arrivo. Nelle sezioni seguenti verranno presentati alcuni esercizi che si possono svolgere sul corpus parallelo italiano-serbo in oggetto e verranno evidenziate le potenzialità didattiche soprattutto in riferimento all'annotazione NER per entrambi i livelli di apprendimento. I risultati del contributo possono trovare applicazione nell'ambito glottodidattico, ma anche in quello traduttivo e lessicografico.

2. Caratteristiche del corpus *It-Sr-NER*

Il corpus *It-Sr-NER* è stato implementato all'interno della Call "Bridging Gaps" organizzata nel 2022 da CLARIN, la più importante infrastruttura per le tecnologie linguistiche a livello europeo[7] (Perišić et al., 2023). Ai fini della ricerca un team composto di membri dell'Università di Torino e della Società per le Risorse e le Tecnologie Linguistiche JeRTeh di Belgrado ha sviluppato alcuni servizi web per il riconoscimento e l'annotazione delle entità denominate (*NE - Named Entities*) come, per esempio, i nomi di persona, di luogo, di organizzazioni, gruppi etnici, eventi e opere d'arte. Il lavoro è stato svolto su testi monolingui e bilingui paralleli in 24 lingue, con un caso studio incentrato su testi paralleli in italiano e serbo. L'obiettivo principale che il bando si poneva era il riconoscimento dei toponimi in

[7] https://www.clarin.eu/

testi non strutturati[8] e il collegamento automatico con le banche dati che contengono informazioni sulle geolocalizzazioni delle entità riconosciute. In particolare, i servizi sviluppati risultano di grande interesse in quanto permettono il collegamento delle entità denominate (*NEL - Named Entity Linking*) con il database Wikidata, offrendo sia la possibilità di geolocalizzazione che l'identificazione delle località riconosciute e la loro visualizzazione su una mappa.

Il corpus parallelo *It-Sr-NER*, ideato e compilato nell'ambito di questo progetto, comprende due set di dati con segmenti allineati (principalmente frasi), estrapolati da numerosi romanzi: il primo costituito da 1000 annotazioni corrette manualmente[9], l'altro da 10.000 entità denominate annotate automaticamente, entrambi pubblicati in diverse serializzazioni e livelli di annotazione[10]. I romanzi delle scrittrici e degli scrittori italiani rappresentati nel corpus sono: *Il nome della rosa* di Umberto Eco, *Le avventure di Pinocchio* di Carlo Collodi, *Storia di chi fugge e di chi resta. L'amica geniale* di Elena Ferrante, e *Uno, nessuno e centomila* di Luigi Pirandello. Il corpus include anche cinque romanzi di scrittori serbi: *Anikina vremena* e *Na Drini ćuprija* di Ivo Andrić, *Nečista krv* di Borisav Stanković, *Opštinsko dete* di Branislav Nušić, e *Bašta, pepeo* di Danilo Kiš. Inoltre, il corpus include le traduzioni italiane e serbe del romanzo di Jules Verne, *Le Tour du monde en quatre-vingts jours*, ricco di entità denominate in forma di toponimi adatti a dimostrare il compito principale del progetto, cioè l'annotazione e il linking di tali entità.

Dopo la parallelizzazione (allineamento), i testi sono stati convertiti nel formato TMX (Translation Memory eXchange) utilizzando il programma ACIDE per la creazione di corpora paralleli (Obradović et al., 2008). Ogni segmento in italiano e in serbo è stato numerato e associato al corrispondente segmento nell'altra lingua, contrassegnato con l'attributo "xml:lang". Il corpus *It-Sr-NER* è stato reso disponibile tramite l'infrastruttura CLARIN e il VLO (Virtual Language Observatory)[11], mentre i servizi sono accessibili sia su VLO che sulla piattaforma

[8] Per "non strutturato" si intende un testo privo di etichette strutturali (come quelle per capitoli, paragrafi, frasi, discorso diretto ecc.) e di annotazioni relative a entità nominate, tipi di parole o lemmi; vale a dire, un testo "grezzo".

[9] La correzione manuale delle annotazioni automatiche, che inevitabilmente contengono errori dipendenti dalla qualità del modello utilizzato, è necessaria per garantire la massima accuratezza (ad esempio, per l'addestramento di modelli futuri) o per valutare il livello di precisione dell'annotazione automatica confrontando il testo con un campione corretto a mano.

[10] Con "serializzazioni" si intendono i formati di registrazione dei dati annotati (ad es. XML, TSV, CONLL, TTL); i "livelli di annotazione" variano invece dagli strati base con segmenti testuali a quelli arricchiti con etichette XML per entità nominate, fino a includere tipi di parole e lemmi.

[11] https://vlo.clarin.eu/record/https_58__47__47_hdl.handle.net_47_20.500.11752_47_OPEN-981_64_format_61_cmdi?1&count=2&index=0&q=It-sr-ner

Language Resource Switchboard[12], come diversi servizi per spaCy[13]. Il corpus include una versione bilingue con testi allineati, singole versioni monolingui ed entità denominate annotate automaticamente[14]. La ricerca nel corpus parallelo *It-Sr-NER*, supportata da un'estensione morfologica e semantica per la lingua serba, è possibile anche attraverso la biblioteca digitale *Bibliša*[15] (Stanković et al., 2017). Un'altra edizione del corpus parallelo, in cui sono annotati le categorie grammaticali (POS tagging) e i lemmi, è disponibile per condurre analisi linguistiche.

La parte serba del corpus è stata annotata con l'Universal POS tagset utilizzando il modello "BEaST" (Stanković et al., 2022), mentre la lemmatizzazione è stata effettuata con l'ausilio di dizionari morfologici elettronici per la lingua serba (Vitas e Krstev, 2012). La parte italiana del corpus è stata annotata utilizzando il modello spaCy per la lingua italiana, con l'uso dello schema di annotazione UD (Universal Dependencies)[16], convertito dall'Italian Stanford Dependency Treebank (Bosco et al., 2014). Per il riconoscimento delle entità denominate (NER) nei testi serbi è stato utilizzato il modello Jerteh-355-tesla[17] che rappresenta una versione modificata del modello Jerteh-355 (Škorić, 2024). Il modello è stato ulteriormente addestrato per il compito di riconoscimento delle entità denominate, raggiungendo F1-score di circa 0,96 sul set di dati sottoposti al test (Stanković et al., 2024). Per i testi in italiano è stato utilizzato il modello "it_core_news_sm-3.4.0", integrato nello strumento spaCy[18], che ha ottenuto F1-score dell'86% sul set per il testing. Dopo l'annotazione automatica, è stato impiegato INCEpTION (Klie et al., 2018) per la correzione manuale e il collegamento delle entità denominate. Delle 7 categorie totali, i tre tipi più comuni erano: persone (PERS), località (LOC) e organizzazioni (ORG).

La classe PERS comprende nomi, cognomi, soprannomi e le loro combinazioni (di persone reali e personaggi di fantasia, inclusi dèi e santi). Nella parte italiana del corpus sono stati riconosciuti per esempio Pietro, Lila, Nino, Mehmed-pascià,

[12] https://switchboard.clarin.eu/tools

[13] Si tratta di: "spaCy NEL", "spaCy bilingual NEL (for TMX)", "spaCy bilingual NER (for TMX)", "spaCy bilingual NER and Geoparsing (for TMX)", "spaCy bilingual NER and NEL (for TMX)", "spaCy monolingual NER", "spaCy monolingual NER and Geoparsing" e "spaCy monolingual NER and NEL".

[14] Ulteriori informazioni e dati sono disponibili nel repository It-Sr-NER sulla piattaforma GitHub https://github.com/jerteh/It-Sr-NER/tree/main/corpus

[15] http://biblisha.jerteh.rs

[16] Le Universal Dependencies (UD) sono un framework per l'annotazione coerente della grammatica (parti del discorso, caratteristiche morfologiche e dipendenze sintattiche) attraverso diverse lingue umane, https://universaldependencies.org/

[17] https://huggingface.co/Tanor/sr_pln_tesla_j355

[18] https://spacy.io/models/it#it_core_news_sm

Michele Solara, Nino Sarratore, Nikola Glasinčanin, mentre nella parte serba Lina Čerulo, Arsa, Marko, Sofka.

Un certo numero di errori si verifica quando i nomi di persona sono erroneamente identificati con località (1), con una parola scritta in maiuscolo all'inizio di una frase (2), oppure a metà frase (3):

(1) devojkom iz <pers>Čajniča</pers>

[ragazza di Čajnice[19]]

(2) <pers> Najbednija </pers>i najtragičnija

[La più misera e più tragica]

(3) marljivo radeći, <pers>Vašu</pers> domovinu privede srećnijoj budućnosti

[di guidare la Vostra patria verso un futuro più felice, operando in pace]

La classe LOC indica continenti, paesi, regioni, centri abitati, oronimi, corsi d'acqua, nomi di corpi celesti e località urbane. Nella parte italiana del corpus sono state riconosciute località come: Scuola Normale di Pisa, piazza dei Martiri, Milano, Nova Varoš, Sarajevo, mentre nella parte serba: Napoli, Ulica Mecokanone, Trg Garibaldi, Drina, Turska. Errori di riconoscimento si riscontrano in espressioni come:

(1) Veoma mlada se udaje za <loc> Stefana Karačija </loc>

[Sposa giovanissima Stefano Carracci]

(2) <loc>Ćorkan</loc>se zagledao u devojku

[il Guercio si innamorò di una ragazza]

(3) šta je moglo biti između neobične<loc>Krnojelčeve kćeri</loc>i ovoga

[cosa fosse potuto accadere tra la strana figlia di Krnojelac e quel Mihailo]

La classe ORG viene utilizzata per nomi di aziende, partiti politici, istituti di istruzione, squadre sportive, ospedali, musei, biblioteche, hotel, caffè e luoghi di culto. Nella parte italiana del corpus sono stati riconosciuti, ad esempio: Banca dell'Agricoltura, HOTEL ZUR, Lotta Continua, Corriere della Sera. Questa categoria è caratterizzata da risultati di riconoscimento molto più deboli rispetto a persone e località, con numerosi errori di annotazione anche nel corpus italiano. Ad esempio, in italiano sono stati erroneamente riconosciuti come ORG termini quali: *galateo*, *quell'uscio*, *Rogatica* (cittadina bosniaca). Nella parte serba del corpus sono stati riconosciuti solo 28 casi, tutti errati. Per esempio Ubedih sebe da više voli<org>društvo</org>devojke [Mi convinsi che stava più volentieri con la

[19] Nel testo italiano manca il segno diacritico sulla seconda "c".

ragazza], così come la sovrapposizione con le categorie PERS e LOC. La causa è rappresentata dalle forti differenze tra i testi annotati e quelli su cui il modello è stato addestrato. Per questo motivo le ricerche attuali sono orientate al miglioramento del riconoscimento proprio di questa categoria.

Il collegamento delle entità riconosciute con Wikidata è stato effettuato per un piccolo campione, principalmente per i personaggi dei romanzi (PERS) e le località (LOC) che rappresentano i luoghi in cui si svolge l'azione. La procedura ha comportato l'inserimento nei Wikidata delle informazioni per le quali non esistevano voci (personaggi di alcuni romanzi). Nell'immagine (Fig. 1) sono mostrate le frasi 201-204 del corpus, sia in serbo che in italiano. Il segmento è tratto dal romanzo di Elena Ferrante e nel pannello si può vedere che "Điljolu" in serbo e "Gigliola" in italiano sono contrassegnati come persona con l'etichetta PERS e sono collegati alla voce di Wikidata che rappresenta il personaggio del romanzo Gigliola Spagnuolo (trascritta in serbo come Điljola Spanjolo), con il link alla voce corrispondente (https://www.wikidata.org/wiki/Q122507460). Allo stesso modo, "Napulj" in serbo e "Napoli" in italiano sono contrassegnati come località con l'etichetta LOC, che è collegata alla voce corrispondente su Wikidata (https://www.wikidata.org/wiki/Q2634).

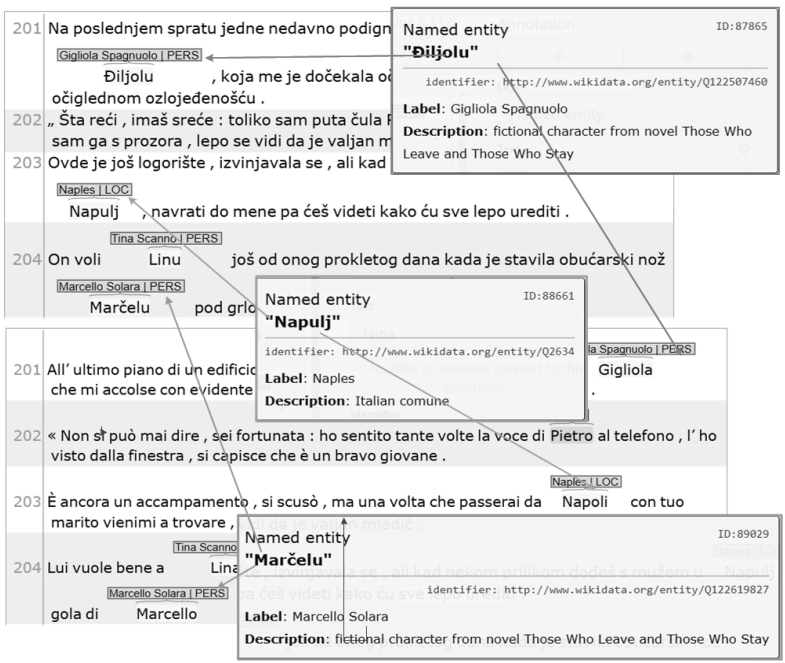

Fig. 1. Esempio di annotazione nell'INCEPTION con collegamento ai Wikidata.

Le risorse sviluppate nell'ambito di questo progetto sono aperte e disponibili per ricercatori, insegnanti e studenti, e sono particolarmente significative per coloro che sono interessati alla lingua italiana in Serbia e alla lingua serba in Italia. Il corpus e i servizi web possono inoltre essere utilizzati nei diversi gradi di istruzione in Croazia, Montenegro, Bosnia ed Erzegovina grazie al policentrismo della lingua serbo-croata. I modelli che supportano i servizi web sono in continuo aggiornamento e le nuove versioni mostrano migliori prestazioni (Ikonić Nešić et al., 2024).

La procedura sviluppata può inoltre essere applicata ad altri corpora monolingui e paralleli, fornendo linee guida per i futuri processi di pubblicazione di corpora multilingui su internet. Inoltre, il lavoro esplora il collegamento delle tecnologie dei Dati Linguistici Aperti Collegati (Linguistic Linked Open Data) sviluppate per l'elaborazione del linguaggio naturale con i dati utilizzati nelle Digital Humanities. Dalle ricerche relative all'interoperabilità semantica è emersa una versione del corpus parallelo in forma di dati collegati (linked data). Il concetto di interoperabilità semantica, applicato ai dati collegati e ai corpora paralleli, garantisce che i dati scambiati tra i sistemi abbiano significati chiaramente definiti, facilitando così una comunicazione chiara e una comprensione accurata[20].

La collaborazione realizzata nel progetto *It-Sr-NER* è proseguita portando alla creazione di un corpus italiano-serbo significativamente più grande, *SerbItaCor3*, che comprende 243 opere letterarie con 664.044 segmenti paralleli. La parte italiana del corpus contiene 11,1 milioni di parole, mentre la parte serba ne conta 10,4 milioni. Collegando la lingua serba e quella italiana, questo corpus apre nuove possibilità per la comunicazione interculturale, contribuendo al miglioramento delle ricerche e degli studi linguistici (Moderc et al., 2023).

3. La proposta di applicazioni di un corpus annotato per NER nella didattica del serbo come LS: riflessioni ed esempi concreti

3.1. Livello base

Tra le principali difficoltà che gli studenti incontrano nelle fasi iniziali dell'apprendimento del serbo come lingua straniera vi è la comprensione dei meccanismi che regolano la morfologia flessiva. In particolare, risulta complesso

[20] La trasformazione del corpus parallelo annotato con entità denominate e collegato a Wikidata è stata convertita in linked data, specificamente in grafi RDF (Resource Description Framework) utilizzando il formato NLP Interchange Format (NIF) (Stanković et al. 2024). I file NIF prodotti sono disponibili su https://llod.jerteh.rs/ItSrNIF/. Possono essere interrogati utilizzando query SPARQL sul punto di accesso http://fuseki.jerteh.rs/#/dataset/ItSrNIF/query.

riconoscere le forme flesse e ricondurle al loro lemma di base, passaggio essenziale per la consultazione efficace dei dizionari bilingui. Questo problema può in parte essere superato con l'uso dei dizionari morfologici, come il dizionario monolingue Hrvatski jezični portal[21] per la variante croata o il dizionario multilingue Serboverb[22] (con le traduzioni dal serbo verso inglese, francese, italiano, spagnolo, portoghese, tedesco)[23], i quali prevedono la ricerca a partire dalle forme flesse. Ciò che si conferma problematico, soprattutto per uno studente alle prime armi, è il riconoscimento di alcuni nomi propri come antroponimi e toponimi (*pluralia tantum* inclusi), dato che si tratta di nozioni enciclopediche spesso assenti nei volumi lessicografici bilingui. La difficoltà per l'apprendente risiede anche nel fatto che alcuni generi grammaticali condividono la stessa desinenza, per esempio neutro singolare in *-e* e femminile plurale (*Nezuke*[24]) o maschile e neutro che condividono la maggior parte delle desinenze[25] al singolare e al plurale. Il tagging e l'annotazione NER in questo caso consentono una ricerca morfologica su diversi livelli.

Nelle fasi iniziali di apprendimento si può partire dai testi in italiano e fare una ricerca CQL (*Corpus query language*[26]) per i toponimi, resa possibile dall'annotazione NER[27] (Fig. 2).

ore e figlio di un ricco industriale di<loc>	**San Giovanni a**	</loc><loc>Teduccio</loc>. </tu> <t
triale di<loc>San Giovanni a</loc><loc>	**Teduccio**	</loc>. </tu> <tu domain=""> Ci ritira
'era grande differenza tra il rione e<loc>	**Napoli**	</loc>, il malessere scivolava dall' u
ain=""> Centro di cibernetica della<loc>	**Statale di Milano**	</loc>,<loc>Centro sovietico</loc>p
i della<loc>Statale di Milano</loc>,<loc>	**Centro sovietico**	</loc>per l' applicazione dei calcolat
n è il rione a essere malato , non è<loc>	**Napoli**	</loc>, è il globo terrestre , è l' unive

Fig. 2. CQL query <loc/>, ricerca dei toponimi, *It-Sr-Ner*.

[21] https://hjp.znanje.hr/

[22] https://serboverb.com/

[23] Si tratta di un dizionario dei verbi che si può consultare come dizionario morfologico e offre la traduzione dei verbi nelle lingue elencate.

[24] L'apprendente dovrà decidere se trattarlo come *pluralia tantum* (u Nezukama) o come neutro singolare (u Nezuku).

[25] Per questo motivo le grammatiche per stranieri di solito inseriscono questi due generi nella prima declinazione (si veda Klajn, 2007, Mrazović, 2009).

[26] Il Corpus Query Language (CQL) è un codice utilizzato per definire criteri di ricerca complessi che non possono essere eseguiti tramite i controlli standard dell'interfaccia utente. Questi criteri possono includere parole o lemmi, ma anche tag e altri attributi, tipi o strutture di testo. È possibile impostare condizioni per token opzionali o ripetizioni di token (la definizione è stata adattata dal glossario Sketch Engine, https://www.sketchengine.eu/guide/glossary/).

[27] Il corpus è consultabile all'indirizzo https://noske.jerteh.rs/

Successivamente, la stessa ricerca può essere effettuata con i testi paralleli. Questo approccio consente di analizzare non solo le specificità morfologiche legate alla declinazione di questi toponimi nella lingua serba, ma anche la loro trascrizione fonetica (Ischia > Iskija) e le regole ortografiche, come l'uso delle maiuscole nei nomi composti (piazza dei Martiri > Trg mučenika), osservando i fenomeni in entrambe le lingue (Tab. 1).

a piazza dei Martiri	na Trgu mučenika
scuola Normale di Pisa	fakulteta u Pizi
Ischia	Iskiji
in Germania	u Nemačku
a Barano	u Baranu
San Giovanni a Teduccio	San Đovanija u Teduču

Tab. 1. Alcuni esempi della traduzione dei toponimi, *It-Sr-Ner*.

La ricerca può essere approfondita attraverso l'osservazione degli aggettivi che caratterizzano determinati toponimi nel corpus monolingue (Fig. 3). A questo punto gli studenti possono osservare le diverse opzioni traduttive soprattutto per approfondire la riflessione sulle traduzioni funzionali che, superando la semantica delle parole, sono profondamente legate al contesto e alle scelte individuali dei traduttori (Tab. 2).

nain=""> Già allora , la verde e	**veloce Drina**	, fiume montano , " che si i
atto una pasticceria rinomata in	**tutta Napoli**	grazie pure alla bravura di
eliminando quel traghetto sulla	**lontana Drina**	, dove la miseria e tutte le
a linea della Drina ; questa è la	**nuova Serbia**	. </tu> <tu domain=""> Sol
nte scorra e dilaghi non solo la	**verde Drina**	, ma l' intero paesaggio ar
i vedeva portato in esilio , nella	**lontana Germania**	, separato per sempre dall
edificare , qui sulla Drina , una	**nuova Istanbul**	. </tu> <tu domain=""> Gli
</pers>ha potuto attraversare l'	**intera Bosnia**	, altro che raggiungere sol
tagne , né avrebbe visto quella	**famosa America**	, né sarebbe riuscito a cre

Fig. 3. CQL query [tag="ADJ"]<loc/>, *It-Sr-Ner*.

ista Drina, zelena i plahovita	la verde e veloce Drina
po čitavom Napulju	in tutta Napoli
dalekoj Drini	lontana Drina
tako je sve ovo Srbija	questa è la nuova Serbia
zelena Drina	verde Drina
daleku Nemačku	lontana Germania
novi Stambol	nuova Istanbul
Bosnu [...] svukoliku	intera Bosnia
tu Ameriku	quella famosa America

Tab. 2. Alcuni esempi di traduzione degli aggettivi seguiti dai nomi, *It-Sr-Ner*.

Una simile ricerca può essere effettuata per gli antroponimi (Fig. 4) nel corpus monolingue.

ende che la sua amica d' infanzia ,<pers>	Lina Cerullo	</pers>, solo da lei chiamata Lila , è spa
sa , dove conosce e si fidanza con<pers>	Pietro Airota	</pers>, e dalla pubblicazione di un roma
alla miseria , si mette al servizio di<pers>	Michele Solara	</pers>, che a un certo punto lo manda i
li diventare come la professoressa<pers>	Galiani	</pers>, mi ero appropriata di toni suoi e
centrò invece su certe formule che<pers>	Nino	</pers>aveva usato marginalmente , ma
zza dei Martiri e che era il padre di<pers>	Gennaro	</pers>, un bambino che non aveva mai
bene che non era vero) , chiese a<pers>	Nino	</pers>e a Tarratano di pronunciarsi , ed
enza del capofamiglia , il professor<pers>	Guido Airota	</pers>, « un uomo veramente eccezion
"> Anni addietro era successo che<pers>	Stefano	</pers>, dopo il matrimonio , si confidass
, mi raccontasse dei problemi con<pers>	Lila	</pers>, ma lo aveva fatto senza mai acc

Fig. 4. CQL query <pers/>, ricerca dei nomi propri di persona, *It-Sr-Ner*.

Per esempio, agli studenti potrebbe essere proposto di analizzare le differenze morfologiche e ortografiche tra i vari antroponimi nelle due lingue, attraverso l'osservazione delle declinazioni e delle trascrizioni fonetiche (Tab. 3).

Stefano Carracci	Stefana Karačija
Pietro Airota	Pjetra Ajrotu
Michele Solara	Mikelea Solare
Carmen Peluso	Karmen Peluzo
Manuela Solara	Manuelu Solaru
Marisa Sarratore	Marizom Saratore

Tab. 3. Alcuni esempi della traduzione dei nomi propri di persona, *It-Sr-Ner*.

Un ulteriore esercizio può prevedere l'osservazione della frequenza con cui si presentano alcuni nomi e la successiva individuazione degli antroponimi più comuni nelle opere letterarie in entrambe le lingue (Fig. 5).

1	☐	Sofka	96
2	☐	Lila	96
3	☐	Nino	81
4	☐	Enco	48
5	☐	Nina	47
6	☐	Vilijem	40
7	☐	Nikola	39

Fig. 5. Risultati delle frequenze nella ricerca dei nomi propri di persona, *It-Sr-Ner*.

In seguito, la ricerca di un nome proprio può essere allargata per consentire l'osservazione non solo della sua forma nel contesto monolingue, ma anche la traduzione attraverso i testi paralleli nel corpus parallelo (Sofka - Sofia; Enzo - Enco; Gennaro - Đenaro) (Fig. 6).

Sama <pers> **Sofka** </pers> uvek se sa jezom i strahom sećala toga što je , ili još kao i dete , od svoje babe , matere i ostalih tetaka i strina mogla da načuje , kad su one bile nasamo i mislile da ih niko neće čuti , još manje se bojale da će to <pers> Sofka </pers> , onako mala , razumeti , a kamo li upamtiti ; ili , što je i sama , kad je odrasla , svojim očima videla .	La stessa <pers> Sofia </pers> ricordava sen potuto sentire fin da bambina , dalla nonna , d quando erano sole e pensavano che nessuno minimamente che <pers> Sofia </pers> , così rammentare ; oppure ciò che essa stessa , già
Sama <pers> Sofka </pers> uvek se sa jezom i strahom sećala toga što je , ili još kao i dete , od svoje babe , matere i ostalih tetaka i strina mogla da načuje , kad su one bile nasamo i mislile da ih niko neće čuti , još manje se bojale da će to <pers> **Sofka** </pers> , onako mala , razumeti , a kamo li upamtiti ; ili , što je i sama , kad je odrasla , svojim očima videla .	La stessa <pers> Sofia </pers> ricordava sen potuto sentire fin da bambina , dalla nonna , d quando erano sole e pensavano che nessuno minimamente che <pers> Sofia </pers> , così

Fig. 6. La traduzione dei nomi propri di persona nel corpus parallelo, *It-Sr-Ner*.

Per esempio, risulta culturalmente e linguisticamente interessante la trasposizione in italiano, non sempre uniforme, dei cognomi femminili con il suffisso *-ka* (Ristićka, Sokolovićka) formati a partire dalla forma maschile (Ristić, Sokolović) per indicare, nella comunicazione informale, una donna sposata. Agli studenti può essere fatto notare come, nel caso di Ristićka, il cognome venga riportato in originale perdendo in questo modo il suo tratto semantico tipico. Questo esercizio permette loro di riflettere sulle sfumature semantiche e culturali legate alla traduzione di nomi propri. Più mirata sembra, invece, la traduzione del cognome Sokolovićka, che diventa "una Sokolović":

[lemma=".*ćka" & tag="PROPN"]
ali on je od svoje majke, koja je bila Sokolovićka > ma da sua madre, che era una Sokolović

All'interno di questo gruppo di antroponimi, particolarmente interessanti risultano i nomi orientali di origine turca, per esempio quelli con il titolo "aga"[28], che possono essere estrapolati grazie all'annotazione NER (Fig. 7).

CQL query [lemma=".*aga"]within<pers/>,

	Lemma	Frequency
1 ☐	Abidaga	8
2 ☐	Ibraga	2
3 ☐	Gonzaga	1
4 ☐	Mujaga	1
5 ☐	Huseinaga	1

Fig. 7. Risultati della ricerca dei nomi propri di persone contenenti il titolo "aga", *It-Sr-Ner*.

Agli studenti può essere proposto di osservare, tramite l'analisi parallela, come nel caso dei cognomi Abidaga e Huseinaga la trascrizione in italiano preveda l'uso del trattino per alcuni di essi (Abid-aga, Husein-aga). Questo esercizio permette loro di riflettere sulle possibili ragioni di tale scelta, indagando se dipenda dallo stile di un traduttore o di un editore, o se rivesta un significato diacronico.

3.2. Livello intermedio/avanzato

Una delle sfide nell'insegnamento di livello intermedio/avanzato consiste nella traduzione dei lessemi culturalmente specifici, i quali possono presentare diversi gradi di asimmetrie (polisemia, *lexical gaps* ecc.). Questa difficoltà si supera solo in parte con l'uso dei dizionari bilingui in quanto in molti casi, soprattutto per quanto riguarda i vuoti lessicali, tali elementi rimangono esclusi o trattati in modo insufficiente dalle fonti lessicografiche (Perišić Arsić, 2020). Utilizzando

[28] Si nota che tra i risultati è rientrato anche il cognome italiano Gonzaga, che non sarà preso in considerazione. La presenza di risultati "contaminati" aiuta l'apprendente a comprendere come funzionano le ricerche corpus-based e come affinarle per ottenere migliori risultati rispetto alla domanda di ricerca.

la metodologia corpus-based, basata sui corpora monolingui, si può in parte superare questo problema, anche se il procedimento richiede un addestramento iniziale mirato e può risultare *time-consuming* (Tognini Bonelli, 2000). In una ricerca su corpora paralleli contenenti testi letterari si possono osservare nello stesso momento il testo originale e la sua traduzione: ciò consente non solo di trovare tutti i traducenti di una parola, ma anche di osservare porzioni di testo più ampie. A titolo di esempio verrà esposta la ricerca svolta sulla parola *kasaba*[29], frequentemente presente nelle opere di Ivo Andrić. Si tratta di una parola che può essere considerata un *lexical gap* a livello semantico in diacronia e, per questo, spesso menzionata nei contributi dedicati alla traduzione delle opere del premio Nobel iugoslavo. Per esempio, Avirović (2003) in un suo studio analizza le due versioni italiane del romanzo *Il ponte sulla Drina* tradotte rispettivamente da Bruno Meriggi[30] e Dunja Badnjević[31]. Nella prima versione, la parola *kasaba* è stata resa con *cittadina*, mentre nella seconda, il lessema è stato omesso, sostituito semplicemente con *Višegrad*, facendo così perdere la sfumatura semantica di "piccola cittadina provinciale orientale" che il termine originale porta con sé. L'autrice sottolinea la differenza fra le due scelte e, mentre reputa appropriata la traduzione di *kasaba* con "cittadina", esprime il proprio disappunto per l'omissione di una parola così significativa per il romanzo e per la poetica di Ivo Andrić, sottolineando inoltre come questa scelta impedisca eventuali rimandi nel testo.

Analizzando il racconto *I tempi di Anika*, le autrici Đukanović e Polovina (2018) riflettono sulla traduzione dello stesso lessema *kasaba* in sloveno (*mesto*), francese (*bourgade*, *cité*, *ville*, *marche*) e inglese (l'uso della parola originale in corsivo), e su come l'uso di turchismi conferisca al testo una precisa dimensione regionale, spaziale e temporale, arricchendolo di un forte carattere culturale. Esse sottolineano come la scelta traduttiva incida sul valore semantico della parola, che in molti casi è metaforico e non si riferisce semplicemente alla località in senso stretto, bensì alla collettività, cioè ai cittadini. Troviamo esempi simili anche in questo corpus:

> i sve što je u to vreme uzbuđivalo kasabu i svet u njoj > e tutto quello che, al suo tempo, aveva commosso la kasaba e la gente che ci viveva.
> Kasaba nije prestajala da govori > La kasaba non finiva più di parlare

[29] *Kasaba* è una parola di origine turca che significa: cittadina, paesino, borgo (gradić, varošica, palanka), *Rečnik sprskoga jezika*, Novi Sad, Matica srpska, 2011. Questo lessema è oggi quasi in disuso, se non con precisi intenti stilistici, quasi sempre con una connotazione negativa.

[30] I. Andrić, *Il ponte sulla Drina*, trad. di B. Meriggi, Milano, Mondadori, 1995.

[31] I. Andrić, *Romanzi e racconti*, trad. di O. Badnjević, Milano, Mondadori, 2001.

Un possibile esercizio sulla parola in questione consiste nell'estrazione di dati quantitativi sulla sua frequenza nel corpus (*lemma*: 291 occorrenze). L'elevato numero di occorrenze in forma originale (217) suggerisce l'importanza di questo lessema e la scelta di una strategia di straniamento[32] da parte dei traduttori. Agli studenti potrebbe essere proposto di osservare le frequenze e le variazioni dei traducenti. Per facilitare la ricerca, evitando di controllare manualmente tutte le occorrenze della parola, nella seconda query è possibile interrogare il corpus parallelo escludendo direttamente la forma originale. È sufficiente inserire il primo traducente individuato (*kasaba*) e selezionare l'opzione *does not contain* (Fig. 8) per aprire le concordanze, visualizzando esempi di altri traducenti utilizzati (*kasaba*, *città*, *cittadina*...), compresi eventuali omissioni e adattamenti testuali.

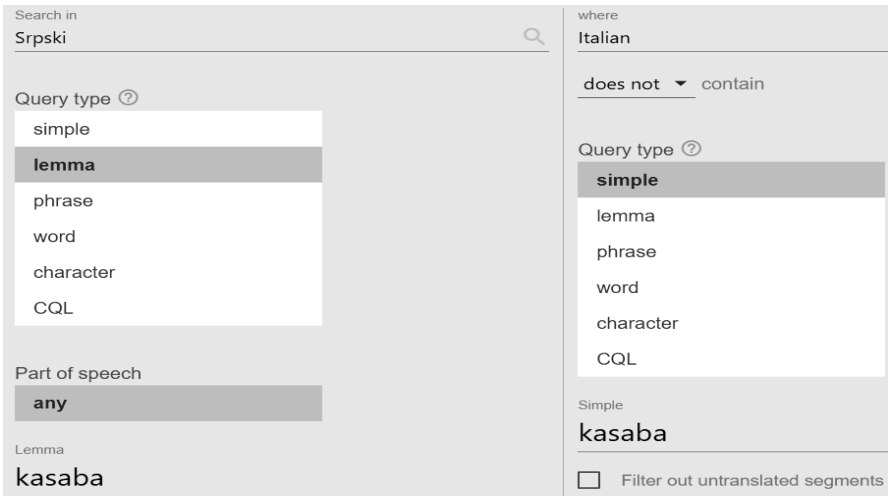

Fig. 8. Ricerca che esclude il traducente "kasaba", *It-Sr-Ner*.

Per comprendere meglio il significato e l'uso della parola *kasaba* si potrebbe chiedere agli studenti di analizzare gli aggettivi che la precedono. Tale analisi consente di esplorare il valore semantico della parola nel suo uso reale e di identificare, attraverso i suoi pattern di utilizzo, le connotazioni (positive o negative) che potrebbero non essere registrate nei dizionari. Utilizzando la query

[32] La strategia di straniamento (*foreignization*) consiste nel mantenere nella traduzione elementi linguistici e culturali propri del testo di partenza preservandone l'alterità e l'identità culturale. È contrapposta alla strategia di addomesticamento (*domestication*), che invece mira ad adattare il testo alla cultura e alle convenzioni linguistiche della lingua di arrivo, rendendolo più naturale per il lettore. Si veda: Venuti, L. (1995), *The Translator's Invisibility: A History of Translation*, London-New York: Routledge.

[tag="ADJ"] [lemma="kasaba"], si ottengono i seguenti collocati in ordine di frequenza:
cela (intera) (9), *bosanska* (bosniaca) (2), *zamrla* (addormentata) (2), *višegradska* (di Višegrad) (2), *ustajala* (remota) (1), *zlosrećna* (sfortunata) (1), *ugašena* (al buio) (1), *zabačena* (periferica) (1), *razasuta* (sparpagliata) (1), *nesrećna* (infelice) (1).

In questo modo gli studenti avranno l'opportunità di osservare come le diverse connotazioni della parola emergano a seconda del contesto, e di riflettere sulle implicazioni di tali scelte traduttive. I collocati consentono di cogliere il valore che lo scrittore dà a questa parola e che la rende ancora oggi utilizzabile con specifiche intenzioni comunicative.

Per ottenere risultati più dettagliati, agli studenti potrebbe essere chiesto di analizzare il corpus *SerbItaCor3*, la versione ampliata del corpus *It-Sr-Ner*, osservando le occorrenze di questa parola nelle diverse opere (Tab. 4).

Ivo Andrić, *Na Drini ćuprija*	646
Ivo Andrić, *Anikina vremena*	106
Mešo Selimović, *Derviš i smrt*	84
Ivo Andrić, *Nemirna godina*	48
Ivo Andrić, *Mila i prelac*	34
Ivo Andrić, *Ljubav u kasabi*	20
Ivo Andrić, *Travnička hronika*	18
Ivo Andrić, *Svadba*	16
Ivo Andrić, *Ćorkan i Švabica*	15
Ivo Andrić, *Osatičani*	14
Ivo Andrić, *Priča o vezirovom slonu*	12
Ivo Andrić, *Na obali*	8
Ivo Andrić, *Most na Žep*i	2
Svetislav Basara, *Srce zemlje*	2
Ivo Andrić, *Šala u Samsarinom hanu*	1

Tab. 4. Frequenze della parola "kasaba", *SerbItaCor3*.

In questo modo si potrà notare che anche Svetislav Basara, uno scrittore contemporaneo serbo, usa lo stesso lessema nella sua opera *Srce zemlje* (Il cuore della terra)[33] con precisi intenti stilistici:

[33] Svetislav Basara, *Il cuore della terra: saggio sul soggiorno di Nietzsche a Cipro*, trad. di Helena Kaloper e Stefania Giancane, Nardò, Besa, 2012.

U vreme Ničeovog prispeća u najveću kiparsku luku, Famagusta je uspavana orijentalna kasaba, izjutra obavijena kužnim maglama iz okolnih močvara [...]
Nei tempi dell'arrivo di Nietzsche nel più grande porto cipriota, Famagosta non è che un'addormentata kasaba orientale, al mattino avvolta nella nebbia sollevatasi dalle vicine paludi [...]

L'uso del lessema *kasaba* non si limita qui a una semplice attribuzione geografica, ma contribuisce a creare un'atmosfera che richiama un immaginario orientale e suggerisce un senso di immobilità e arretratezza, rafforzato dalla descrizione di Famagosta come "addormentata" e avvolta da "nebbie pestilenziali". In questo modo la parola acquisisce un valore simbolico che va oltre la semplice definizione di luogo in quanto evoca uno scenario culturale e storico ben preciso.

Si può chiedere inoltre agli studenti di riflettere sulla scelta delle traduttrici di optare, nello stesso volume, per una traduzione che si allontana notevolmente dal significato semantico della parola:

Onda dolazi krivudava linija razdvajanja prekrivena mrklim mrakom, pa otuda naovamo ulice severne Nikozije u kojima čkilje slabašne svetiljke, poput onih u balkanskim kasabama [...]
Poi, dopo un'appena visibile curva, flebile linea di frontiera sommersa nella fitta oscurità, ecco le strade di Nicosia Nord con i loro fiochi lampioni che fumano debolmente come se si trovassero in oscure taverne balcaniche [...]

In questo caso, la scelta di tradurre *kasaba* con *oscure taverne balcaniche* si discosta dal significato letterale della parola, per cui gli studenti potrebbero discutere se le traduttrici abbiano cercato di evocare un'atmosfera suggestiva e decadente, attraverso il carattere cupo della scena descritta. Un'ulteriore riflessione riguarda l'uso della strategia di addomesticamento con un intento di adattamento culturale che rende la scena descritta più immediata e comprensibile per un lettore italiano, sacrificando la precisa connotazione semantica di questo lessema.

Un'ulteriore osservazione merita la parola derivata "kasabalija"[34] presente per ben 39 volte con il significato "l'abitante della kasaba". La parola non solo rappresenta un vero e proprio *lexical gap*, ma rivela il suo tratto tipico legato alla scrittura di Ivo Andrić dimostrabile attraverso la ricerca nel corpus *SerbItaCor3* che attesta 101 occorrenze nelle cinque opere di Andrić (la maggior parte in *Il ponte sulla Drina*) e solo due occorrenze in un'opera di Mešo Selimović (Fig. 9).

[34] Grazie alla query [lemma=kasaba.*] si può risalire all'aggettivo possessivo *kasabalijski* (55).

	Doc.Autor_originala	Doc.Naslov_originala	Frequency
1	Andrić Ivo	Na Drini ćuprija	90
2	Andrić Ivo	Anikina vremena	4
3	Andrić Ivo	Nemirna godina	4
4	Andrić Ivo	Travnička hronika	2
5	Selimović Meša	Derviš i smrt	2
6	Andrić Ivo	Osatičani	1

Fig. 9. Frequenze della parola "kasabalija" nelle diverse opere, *SerbItaCor3*.

Le traduzioni attestate confermano che si tratta di un *lexical gap* che ha richiesto diversi procedimenti traduttivi: l'uso del sintagma (1), l'omissione (2, 3) o il ricorso a parafrasi che richiedono una traduzione funzionale (4).

1. *kasabalije* > *gli abitanti della kasaba* / *gli abitanti di Višegrad* / *gli abitanti della città*
2. *tu su svi postajali opet kasabalije* > *tutti tornavano ad essere quelli che erano sempre stati*
3. *izgovarao je mnogi kasabalija* > *dicevano molti*
4. *Fehim Bahtijarević je samo po majci kasabalija.* > *la madre [...] è nativa di Višegrad.*

Potrebbe essere interessante proporre agli studenti di riflettere sulle diverse strategie adottate nella traduzione del lessema *kasabalija*, confrontando le soluzioni scelte dai traduttori e valutandone l'efficacia. A partire da questo esercizio essi possono individuare altre parole culturalmente connotate che non trovano un corrispondente diretto in italiano e discutere le possibili strategie di traduzione (adattamento, mantenimento del termine originale, parafrasi ecc.). In questo modo, gli apprendenti possono sviluppare una maggiore consapevolezza sulle sfide della traduzione culturale e sulla necessità di adattare il testo al contesto linguistico e culturale di arrivo.

4. Conclusioni

La presente ricerca si poneva due obiettivi principali: da una parte illustrare la creazione del primo corpus parallelo serbo-italiano e, dall'altra, dimostrare il suo potenziale applicativo nell'insegnamento.

La realizzazione del corpus *It-Sr-NER*, frutto della collaborazione tra l'Università di Torino e la società JeRTeh di Belgrado nell'ambito del progetto "Bridging gaps"

dell'infrastruttura CLARIN, ha colmato una significativa lacuna nelle risorse glottodidattiche disponibili per questa combinazione linguistica. In particolare, nella didattica della lingua serba un corpus con l'annotazione NER (*Named Entity Recognition*) si rivela particolarmente utile per molteplici applicazioni, stratificate per livello di competenza linguistica.

Per gli studenti principianti, questa tipologia di corpus può offrire un valido supporto nell'apprendimento di toponimi e antroponimi, spesso trascurati negli strumenti didattici e lessicografici tradizionali. L'analisi di questi lessemi permette di affrontare sistematicamente le peculiarità morfologiche, fonetiche e ortografiche della lingua serba come, per esempio, le complesse regole di trascrizione fonetica e il riconoscimento della forma base (lemma) a partire dalle forme flesse presenti nel contesto.

Sul versante della didattica della traduzione che di solito viene trattata con gli studenti di livello intermedio o avanzato, il corpus in questione si dimostra particolarmente utile nell'analisi dei lessemi culturalmente specifici, come si è cercato di dimostrare con l'esempio della parola *kasaba*. Attraverso un approccio quantitativo si può confermare la centralità di questo *lexical gap* nelle opere del premio Nobel Ivo Andrić, mentre un'analisi qualitativa può evidenziare e confermare le sue sfumature semantiche e contestuali. L'espansione del corpus con *SerbItaCor3* permette di estendere l'analisi alla letteratura contemporanea e di rivelare eventuali evoluzioni nell'uso di questo lessema e dei suoi derivati.

Si auspica che i risultati della ricerca presentata possano fungere da modello metodologico per l'insegnamento del serbo come lingua straniera e che i corpora presentati possano nelle future ricerche confermare la loro natura polivalente, in grado di adattarsi efficacemente sia agli scopi didattici che a quelli della ricerca linguistica, traduttologica e lessicografica.

Bibliografia

1. Avirović, Lj. (2003) Il ponte di Andrić collega uomini e cose: sulla traduzione di Ivo Andrić in Italia. *Comunicare. Letterature lingue*. 3, 377–388.
2. Baker, M. (1993) Corpus Linguistics and Translation Studies. Implications and Applications. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds), *Text and Technology. In Honour of John Sinclair*. Philadelphia-Amsterdam, John Benjamins, pp. 233–250.
3. Bosco, C., Cosi, P., Dell'Orletta, F., Falcone, M., Montemagni, S. & Simi, M. (eds). (2014) *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014: 9-11 December 2014*. Pisa, Pisa University Press.

4. Botley, S. P., McEnery, A. M., & Wilson, A. (eds). (2000) *Multilingual Corpora in Teaching and Research*. Amsterdam, Rodopi.
5. Chesterman, A. (2007) Similarity Analysis and the Translation Profile. *Belgian Journal of Linguistics*. 21 (1), 53–66, https://doi.org/10.1075/bjl.21.05che
6. Doval, I. & Sánchez Nieto, M. T. (eds). (2019) *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*. Amsterdam, John Benjamins.
7. Đukanović, M. & Polovina, V. (2018) Kulturno-specifična leksika u prevodu Iva Andrića na slovenački, francuski i engleski jezik. In Vraneš, A. (ur.), *Ivo Andrić u našem vremenu: zbornik radova*. Andrićgrad – Višegrad, Andrićev institut, pp. 241–260.
8. Granger, S. (2018) Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution? In Čibej, J., Gorjanc, V., Kosem, I. & Krek, S. (eds), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Ljubljana University Press, Faculty of Arts, pp. 17–24.
9. Ikonić Nešić, M., Petalinkar, S., Škorić, M. & Stanković, R. (2024) BERT Downstream Task Analysis: Named Entity Recognition in Serbian. In Trajanović, M., Filipović, N. & Zdravković, M. (eds), *Disruptive Information Technologies for a Smart Society. ICIST 2024. Lecture Notes in Networks and Systems, vol 860*. Cham, Springer, pp. 333–347, https://doi.org/10.1007/978-3-031-71419-1_29
10. Klajn, I. (2007) *Grammatica della lingua serba*. Beograd, Zavod za udžbenike.
11. Klie, J. C., Bugert, M., Boullosa, B., De Castilho, R. E. & Gurevych, I. (2018) The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In Zhao, D. (ed.), *Proceedings of the 27th international conference on computational linguistics: System demonstrations*. Santa Fe, New Mexico, Association for Computational Linguistics, pp. 5–9.
12. Moderc, S., Stanković, R., Tomašević, A. & Škorić, M. (2023) An Italian-Serbian Sentence Aligned Parallel Literary Corpus. *Review of the National Center for Digitization.* 43, 78–91, https://doi.org/10.5281/zenodo.11203388
13. Mrazović, P. (2009) *Gramatika srpskog jezika za strance*. Sremski Karlovci-Novi Sad, Izdavačka knjižarnica Zorana Stojanovića.
14. Obradović, I., Stanković, R. & Utvić, M. (2008) Integrisano okruženje za pripremu paralelizovanog korpusa. In Tošović, B. (ed.), *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*. Münster, LitVerlag, pp. 563–578.
15. Perišić Arsić, O. (2020) Translating lexical gaps: A contrastive corpus-based analysis. In Matešić, M. & Memišević, A. (eds), *Language and Mind. Proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society*. Berlin, Peter Lang, pp. 93–108.
16. Perišić, O., Stanković, R., Ikonić Nešić, M. & Škorić, M. (2023) It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text. In Erjavec, T. & Eskevich, M. (eds), *Selected Papers from the CLARIN Annual Conference 2022*, pp. 99–110. https://doi.org/10.3384/ecp198

17. Salkie, R. (2002) How can linguists profit from parallel corpora? In Borin, L. (ed.), *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, 22 – 23 April, 1999*. Amsterdam – New York, Rodopi, pp. 111-122.
18. Stanković, R., Krstev, C., Vitas, D., Vulović, N. & Kitanović, O. (2017) Keyword-Based Search on Bilingual Digital Libraries. In Calì, A., Gorgan, D. & Ugarte, M. (eds), *Semantic Keyword-Based Search on Structured Data Sources. IKC 2016. Lecture Notes in Computer Science*. Cham, Springer, pp. 112–123, https://doi.org/10.1007/978-3-319-53640-8_10
19. Stanković, R., Škorić, M. & Šandrih Todorović, B. (2022) Parallel bidirectionally pretrained taggers as feature generators. *Applied Sciences*. 12(10), 5028, https://doi.org/10.3390/app12105028
20. Stanković, R., Ikonić Nešić, M., Perišić, O., Škorić, M. & Kitanović, O. (2024) Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking. In Chiarcos, C., Gkirtzou, K. et al. (eds), *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, Turin, 20-25 May 2024, Torino, ELRA & ICCL, pp. 115–125.
21. Škorić, M. (2024) Novi jezički modeli za srpski jezik. *Infotheca*. 24(1), accepted for publishing.
22. Tognini Bonelli, E. (2000) 'Unità funzionali complete' in inglese e in italiano: verso un approccio corpus-driven. In Bernardini, S. & Zanettin, F. (eds), *I corpora nella didattica della traduzione*. Bologna, Clueb, pp. 153–175.
23. Toury, G. (1980) *In Search of a Theory of Translation*. Tel Aviv, The Porter Institute for Poetics and Semiotics.
24. Trousterud, T. (2002) Parallel corpora as tools for investigating and developing minority languages. In Lars, B. (ed.), *Parallel corpora, parallel worlds*. *Selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April, 1999. *Language and Computers*, 43, 111-122.
25. Venuti, L. (1995), *The Translator's Invisibility: A History of Translation*, London-New York: Routledge.
26. Vitas, D. & Krstev, C. (2012) Processing of corpora of Serbian using electronic dictionaries. *Prace Filologiczne*. 63, 279-292.
27. Zanettin, F. (1994) Parallel words: designing a bilingual database for translation activities. In Wilson, A. & McEnery, T. (eds), *Corpora in language education and research: a selection of papers from Tale 94*, UCREL technical papers, 4. Lancaster, UCREL, pp. 163–180.

Olja Perišić
University of Turin
Department of Foreign Languages and Literatures and Modern Cultures

Ranka Stanković
University of Belgrade
Faculty of Mining Engineering and Geology

IT-SR-NER: SERBIAN-ITALIAN PARALLEL CORPUS FOR LEARNING SERBIAN AS A FOREIGN LANGUAGE

Summary

This paper explores the applications of the *It-Sr-NER* parallel corpus in foreign language teaching. The corpus, developed in 2022 by researchers from the University of Turin and the Society for Linguistic Resources and Technologies (JeRTeh) from Belgrade as part of the "Bridging gaps" project under CLARIN infrastructure, represents the first parallel corpus for the Italian-Serbian language pair. It contains 10.000 sentences extracted from both classical and modern Italian and Serbian literature, which have been aligned to facilitate data exploration and word translation in context. The corpus is annotated for Named Entity Recognition (NER), with special attention to toponyms (including *pluralia tantum*) and personal names. The study addresses three key aspects: the challenges faced by beginners due to Serbian's rich morphology in recognising and connecting named entities with their lemmas, the utility of the corpus for intermediate and advanced learners in studying *lexical gaps* (words without direct translation equivalents), and the potential of NER-annotated parallel corpora as an alternative to traditional bilingual dictionaries and digital tools. The paper demonstrates how this type of corpus can serve as both an effective alternative to conventional resources and a unique tool for specific types of linguistic research.

▸ ***Keywords:*** parallel corpus, Italian L1, Serbian FL, Named Entity Recognition, language teaching, translation.

Оља Перишић
Универзитет у Торину
Департман за стране језике и књижевности и савремене културе

Ранка Станковић
Универзитет у Београду
Рударско-геолошки факултет

IT-SR-NER: СРПСКО-ИТАЛИЈАНСКИ ПАРАЛЕЛНИ КОРПУС ЗА УЧЕЊЕ СРПСКОГ КАО СТРАНОГ ЈЕЗИКА

Резиме

У раду се истражују могућности примјене паралелног италијанско-српског корпуса *It-Sr-NER* у настави српског као страног језика. Ради се о првом паралелном корпусу за ову језичку комбинацију који су 2022. године развили истраживачи са Универзитета у Торину и Друштва за језичке ресурсе и технологије (Јертех) у Београду у оквиру пројекта *Bridging gaps* инфраструктуре за језичке технологије CLARIN. Корпус садржи 10.000 реченица преузетих из класичних и модерних дјела италијанске и српске књижевности које су поравнате (паралелизоване) ради лакшег истраживања података и превођења ријечи у контексту. Корпус је анотиран за препознавање именованих ентитета (Named Entity Recognition, NER) са посебном пажњом усмјереном на топониме (укључујући именице *pluralia tantum*) и антропониме. У раду се истражују могућности употребе овог корпуса у циљу превазилажења изазова са којима се студенти српског језика као страног на почетном нивоу суочавају у препознавању и повезивању именованих ентитета са њиховим основним облицима, усљед богате морфологије српског језика. У исто вријеме истражује се употреба корпуса за ученике средњег и напредног нивоа у проучавању лексичких празнина (ријечи без директних преводних еквивалената). Рад указује да паралелни корпуси са NER анотацијом представљају не само дјелотворну алтернативу традиционалним ресурсима за учење српског језика као страног, већ су и незамјењив алат за одређене врсте лингвистичких истраживања.
▸ ***Кључне ријечи:*** паралелни корпус, италијански као матерњи, српски као страни, препознавање именованих ентитета, глотодидактика, превођење.

Preuzeto: 10. 12. 2024.
Korekcije: 20. 2. 2025.
Prihvaćeno: 10. 3. 2025.